Proceedings

of the

2nd International Beilstein Workshop

on

Experimental Standard Conditions of Enzyme Characterizations

March 19th – 23rd, 2006

Rüdesheim/Rhein, Germany

Edited by Martin G. Hicks and Carsten Kettner

Beilstein-Institut zur Förderung der Chemischen Wissenschaften

Trakehner Str. 7 – 9
60487 Frankfurt
Germany

Telephone: +49 (0)69 7167 3211
Fax: +49 (0)69 7167 3219
E-Mail: info@beilstein-institut.de
Web-Page: www.beilstein-institut.de

Impressum

Experimental Standard Conditions of Enzyme Characterizations, Martin G. Hicks and Carsten Kettner (Eds.), Proceedings of the Beilstein-Institut Workshop, March 19th – 23rd 2006, Rüdesheim, Germany.

The Beilstein-Institut and its Editors assume no responsibility for the statements and opinion made by the authors. Registered names and trademarks etc., used in this publication, even in the absence of specific indication thereof, are not to be considered unprotected by law.

Bibliographic information published by *Die Deutsche Bibliothek*
Die Deutsche Bibliothek lists this publication in the *Deutsche Nationalbibliografie*; detailed bibliographic data is available in the internet at http://dnb.ddb.de

ISBN ■

Layout by: Hübner Electronig Publishing GmbH,
Eltville

Cover Illustration: N.N.

Printed by Logos Verlag Berlin
Comeniushof, Gubener Str. 47
10243 Berlin
Tel: +49 (0)30 42 85 10 90
Fax: +49 (0)30 42 85 10 92
Internet: http://www.logos-verlag.de

Preface

The post-genomic era is significantly characterized by a high integration and interdisciplinary of research resources from such diverse fields as computational biology, bioinformatics, functional genomics, structural biology, and proteomics. In this perspective, established biological systems can be comprehensively investigated in terms of interactions of individual or groups of proteins and enzymes as well as the behaviour of collective networks of such interactions. On the other hand, these systems can be re-examined in the light of new results that suggest novel associations between otherwise unrelated pathways and individual proteins.

Modern experimental technologies are providing seemingly endless opportunities to generate massive amounts of sequence, expression and functional data. Continuous advances and improvements have enabled proteome analyses to proceed with increased depth and efficiency. To capitalize on this enormous pool of information and in order to understand fundamental biological phenomena it is essential to collect, organize, categorize, analyse, and share data and results.

However, whilst the large international genome sequencing projects elicited considerable public attention with the creation of huge sequence databases, it has become increasingly apparent that functional data for the gene products, in particular for enzymes, has either limited accessibility or is unavailable. Additionally, although enzyme structural information has been rapidly accumulated in databases, little effort has been invested toward systematic characterization of enzyme functions.

The problem is twofold; deriving data from experimental work is expensive and very time consuming and it is inherently very difficult to collect, interpret and standardize published data since they are widely distributed among journals covering a number of fields, and the data itself is often dependent on the experimental conditions.

For these reasons a systematic and standardized collection of functional enzyme data is essential for the interpretation of the genome information.

The first ESCEC meeting in 2003 resulted in a general agreement that standardization of experiments and methods for enzyme characterization is definitely necessary and in the formation of the STRENDA commission. STRENDA stands for **St**andards for **R**eporting **En**zyme **Da**ta and the commission accompanies the upcoming series of ESCEC symposia.

This 2nd ESCEC symposium provided a platform to discuss a number of checklists worked out and presented by the STRENDA commission. In general, these lists are intended to support the improvement of reporting enzyme data and can be found on the STRENDA website www.strenda.org/documents.

We would like to thank particularly the authors who provided us with written versions of the papers that they presented. Special thanks go to all those involved with the preparation and organization of the workshop, to the chairmen who piloted us successfully through the sessions and to the speakers and participants for their contribution in making this workshop a success.

Frankfurt/Main, July 2007

Carsten Kettner
Martin G. Hicks

Contents

Page

Thermodynamics of Enzyme-Catalysed Reactions

Robert A. Alberty

Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

E-Mail: alberty@mit.edu

Received: 12th April 2006 / Published: 31st August 2007

Abstract

Since the pH is treated as an independent variable in biochemical thermodynamics, the Gibbs energy G does not provide the criterion for equilibrium, but the transformed Gibbs energy G' does. The standard transformed Gibbs energy of formation $\Delta_f G'^0$ of a reactant (sum of species) can be calculated at the desired temperature, pH, and ionic strength if the standard Gibbs energies of formation $\Delta_f G^0$ and standard entropies of formation $\Delta_f H^0$ of the species that make up the reactant are known. BasicBiochemData3 in MathSource provides species properties for 199 biochemical reactants and Mathematica programs for calculating apparent equilibrium constants K' and other transformed thermodynamic properties of enzyme-catalysed reactions are given. This database can be extended, and the number of reactions for which apparent equilibrium constants can be calculated increases exponentially with the number of reactants in the database.

Alberty, R.A.

Database on Thermodynamics of Enzyme Catalysed Reactions

My main interest is in the thermodynamics of reactions catalysed by enzymes, but some of the things I am going to talk about apply to the kinetics of these reactions. There are two ways to discuss the thermodynamics of enzyme-catalysed reactions: (1) with chemical reactions written in terms of species (like ATP^{4-}, $HATP^{3-}$, and H_2ATP^{2-}) using the Gibbs energy G, enthalpy H, and entropy S. (2) with biochemical reactions written in terms of reactants (sums of species like ATP) at a specified pH and using the apparent equilibrium constant K', the transformed Gibbs energy G', the transformed enthalpy H', and the transformed entropy S'. These transformed thermodynamic properties depend on the pH. Both ways are needed by biochemists. Chemical reactions are needed to discuss mechanisms of enzyme-catalysed reactions in terms of species. Biochemical reactions are needed to obtain a broader overview of enzyme-catalysed reactions at specified pH (and perhaps pMg). This is all explained in IUPAC/IUB Recommendations, 1994 [1].

Both chemical equations and biochemical equations are mathematical equations in the sense that chemical equations must balance numbers of atoms of all elements and electric charges. Biochemical equations must balance numbers of atoms of all elements except for hydrogen. Biochemical equations also do not balance electric charge. The reason why biochemical equations do not balance numbers of hydrogen atoms and electric charges is that it is assumed that the pH is held constant during the reaction. This can be done with a pHstat, but biochemists use a buffer to keep the pH nearly constant during an enzyme-catalysed reaction, and then, if they are determining an apparent equilibrium constant, they measure the composition and pH at equilibrium. The equilibrium composition corresponds with this pH, and so this is equivalent to using a pHstat.

Most enzyme-catalysed reactions produce or consume hydrogen ions. This leads to a new thermodynamic property, the change in binding of hydrogen ions $\Delta_r N_H$ in the reaction. $\Delta_r N_H$ will depend on the pH if any reactant has a pK in the pH range of interest (usually between pH 5 and pH 9). This property can be calculated using:

$$\Delta_r N_H = \sum \nu_i' \bar{N}_{Hi} \tag{1}$$

where ν_i' is the stoichiometric number for reactant i and $\bar{N}_{Hi}$ is the average number of hydrogen atoms in reactant i. The prime is needed to distinguish ν_i' from the stoichiometric numbers in the underlying chemical reactions. In making tables of thermodynamic properties of enzyme-catalysed reactions, $\Delta_r N_H$ is usually calculated using:

$$\Delta_r N_H = \frac{1}{RT\ln(10)} \frac{\partial \Delta_r G'^\circ}{\partial \mathrm{pH}} \tag{2}$$

However, it can also be calculated using Equation 1. It should be noted that in the thermodynamics of biochemical reactions, $\Delta_r N_H$ is on the same level as $\Delta_r H'^0$ and $\Delta_r S'^0$ since all three of these properties are calculated by taking partial derivatives of $\Delta_r G'^0$. The average number of hydrogen atoms in a biochemical reactant can be calculated using the binding polynomial [2, 3] for the reactant. This calculation requires the pK values of the reactant in the pH range of interest and the numbers of hydrogen atoms in the species of the reactant. The use of Equation 1 has the advantage over Equation 2 in that $\Delta_f G^0$ values for species in the reaction do not have to be known.

These requirements of the thermodynamics of enzyme-catalysed reactions also apply to the kinetics because the complete steady-state rate equation for an enzyme-catalysed reaction must yield the same equilibrium composition for the reaction as thermodynamics. The apparent equilibrium constant K' for an enzyme-catalysed reaction can be calculated from kinetic parameters; this expression for the apparent equilibrium constant is referred to a Haldane equation. For some mechanisms there is more than one Haldane equation.

The calculations in biochemical thermodynamics are very complicated, but fortunately the application Mathematica® [4] is very convenient for making them. I have developed a database, written in Mathematica and called BasicBiochemData3 [5], that gives the standard Gibbs energies of formation $\Delta_f G^0$ and standard enthalpies of formation $\Delta_f H^0$ of species of biochemical interest at 298.15 K and zero ionic strength for 199 reactants of biochemical interest. Some of these species properties come from the NBS and CODATA thermodynamic tables that deal with chemical species, but for larger molecules of biochemical interest, species properties have to be calculated from experimental measurements of apparent equilibrium constants K' and enthalpies of enzyme-catalysed reactions. The experimental data in the literature has been summarized and evaluated (actually graded A, B, C) by Goldberg and Tewari in six survey papers in *J. Phys. Chem.* Ref. Data (1991 – 1999). They have summarized experimental data on about 500 different reactions involving about 1000 reactants. They have also established a web site on this experimental data [6]. We are indebted to Goldberg and Tewari for assembling all this literature data.

Mathematica is so useful that I have written a second book [3] this time in Mathematica, entitled "*Biochemical Thermodynamics*; *Applications of Mathematica*." This makes it possible to intermingle explanations, programs, and calculations. It has a CD in the back with all the programs, data, and words. All the steps in calculating properties, making tables, and making figures are shown. We usually think of computer programs that calculate numbers, but Mathematica can do more. It can be used to derive equations that are too big to write out by hand.

BasicBiochemData3 provides $\Delta_f G^0$ values for species of 199 biochemical reactants, but $\Delta_f H^0$ are known for species for only 94 of these reactants. This database can be used to calculate standard transformed Gibbs energies of formation $\Delta_f G'^0$ of these 199 reactants at 298.15 K in the pH range 5 to 9 and ionic strengths from zero to about 0.35 M. For the 94 reactants for which enthalpies are known, it is possible to calculate $\Delta_f G'^0$ from 273.15 K to about 313.15 K.

These $\Delta_f G'^0$ have been used to calculate standard transformed Gibbs energies of reaction $\Delta_r G'^0$ and apparent equilibrium constants K' at 298.15 K and 0.25 M ionic strength for 229 enzyme-catalysed reactions [3], but the $\Delta_f G'^0$ can be used for even more reactions. When $\Delta_f H^0$ are known for the species of a reactant, standard transformed Gibbs energies $\Delta_f G'^0$ and standard transformed enthalpies $\Delta_f H'^0$ of reactants can be calculated in the temperature range 273.15 K to about 313.15 K, pH values in the range 5 to 9, and ionic strengths from zero to about 0.35 M. This information has been used to calculate $\Delta_r G'^0$, $\Delta_r H'^0$, $\Delta_r S'^0$ and apparent equilibrium constants K' for 90 enzyme-catalysed reactions.

Since biochemists need thermodynamic properties at various temperatures, pH values and ionic strengths, tables and plots cannot satisfy these needs. Having a file of mathematical functions that give $\Delta_f G'^0$ of reactants does satisfy these needs, and BasicBiochemData3 makes available Mathematica programs and 774 mathematical functions for these properties. These functions can be added and subtracted to obtain changes in thermodynamic properties in biochemical reactions and apparent equilibrium constants. Plots can also be made to show how reaction properties depend on temperature, pH, and ionic strength. BasicBiochemData3.nb contains the functions of temperature, pH, and ionic strength that yield the standard transformed Gibbs energies of reaction for the 90 enzyme-catalysed reactions for which the effects of changing the temperature can be calculated.

It is important to emphasize the importance of ionic strength in the thermodynamics of enzyme-catalysed reactions. According to the Debye-Huckel theory, the logarithm of the activity coefficient of an ion in water is proportional to its charge squared. This means that the ionic strength effect for the species ATP^{4-} is 16 times that for a chloride ion, a huge effect.

I often see MgATP in the biochemical literature, but this is not a species or a reactant. To treat the effect of magnesium ions, a further Legendre transform is required to introduce pMg as an independent variable. People determining apparent equilibrium constants of reactions in the presence of magnesium ions often give the total magnesium concentration, but it is $\mathrm{pMg} = -\log[Mg^{2+}]$, where $[Mg^{2+}]$ is the free concentration, that affects the value of K'. The effects of pMg on the hydrolysis of ATP, ADP, and AMP have been calculated, but this is about the only series for which there is sufficient information about the dissociation of magnesium complex ions. The effect of Mg^{2+} sometimes cancels because both reactants and products bind Mg^{2+}.

Stoichiometry of Enzyme-Catalysed Reactions

In making thermodynamic calculations on biochemical reactions, it is necessary to be very careful about stoichiometry (for example, see Equation 1). It is assumed that when a reactant is made up of species with different numbers of hydrogen atoms, these species are in equilibrium at a specific pH. I think that biochemists understand this pretty well for reactants like ATP, but not for reactants like carbon dioxide and ammonia. Many biochemical reactions are balanced on the web with CO_2 or NH_3, but I do not think this is very appropriate for considering reactions in a living cell where there is no gas phase. In aqueous phases, carbon dioxide is made up of four species: CO_2, H_2CO_3, HCO_3^-, and CO_3^{2-}. I represent this sum of species in the aqueous phase as CO_2tot, for which the transformed thermodynamic properties depend on the temperature, pH, and ionic strength. When CO_2(gas) is replaced with CO_2tot in a reaction equation, a H_2O has to be added on the other side of the reaction to balance oxygen atoms. In the aqueous phase ammonia is made up of NH_3 and NH_4^+. I represent this sum of species by ammonia, for which the transformed thermodynamic properties depend on the temperature, pH, and ionic strength. These comments apply to other gases that dissolve in water and exist in the aqueous phase in different protonated forms.

Hydrogen ions should never appear in balanced biochemical equations because it is understood that the pH is held constant during the approach to equilibrium by adding or removing hydrogen ions. I am not advocating the use of pHstats, but what I am saying is that biochemists interpret determinations of apparent equilibrium constants and enthalpies of reaction as if they were carried out in a pHstat.

The abbreviations NAD^+ and NADH are a problem because this seems to indicate that hydrogen atoms and electric charges are to be balanced on the two sides of the biochemical equation, but they are not. I favour using NAD_{ox} and NAD_{red} instead. These remarks apply to other complicated coenzymes that exist in oxidized and reduced forms.

In my new book I have always written reactions in the direction in which they have apparent equilibrium constants greater than unity at pH 7 and 0.25 M ionic strength. In the 229 reactions in my book for which apparent equilibrium constants are calculated at 298.15 K, 78 are written in the opposite direction from the EC list.

Future Developments Using the Database of Goldberg and Tewari

Many more species data can be obtained from the database surveyed by Goldberg and Tewari [6]. When the apparent equilibrium constant has been determined for an enzyme-catalysed reaction, there is the potential for calculating $\Delta_f G^0$ for the species of a reactant. It is necessary to say „there is the potential“ because the following conditions have to be met: (1) The $\Delta_f G^0$ of all of the species of all of the reactants, but the reactant of interest, are needed. (2) If the reactant of interest has pK values in the range of approximately 5 to 9, these pK values are needed. (3) The experiments have to be carried out carefully and reported accurately. These are pretty demanding requirements.

There is an exception to requirement (1) that should be used sparingly: when there are two reactants in an enzyme-catalysed reaction for which thermodynamic properties are not known, $\Delta_f G^0 = 0$ can be assigned to one species of one of these two reactants. This was done by Alberty and Goldberg (1992) with the ATP series when $\Delta_f G^0$ (adenosine^0) = 0 was adopted as a convention of the thermodynamic table. This made it possible to calculate $\Delta_f G^0$ for all the other species in the ATP series. When this convention is used, the adenosine moiety must appear on both sides of a biochemical reaction. After Boeiro-Goates and coworkers (2001) determined the standard entropy of adenosine(cryst) using the third law method, they were able to calculate $\Delta_f G^0$ (adenosine^0) in aqueous solution with respect to the elements in their reference states, and this changed the $\Delta_f G^0$ of all the species in the ATP series by the same amount. It did not change the apparent equilibrium constants that had been calculated earlier for reactions involving the ATP series. But now it is possible to explore the thermodynamics of the formation of adenosine and adenine all the way back to the elements. The convention that $\Delta_f G^0 = 0$ for one species is especially useful for reactants in oxidoreductase reactions because the oxidized form is on one side of the equation and the reduced form is always on the other side. Similar remarks apply to the use of the convention that $\Delta_f H^0 = 0$ for one of the species of the reactant.

>Enzymes are making it possible to learn about the thermodynamics of large molecules in aqueous solution because they catalyse very specific reactions rapidly. The thermodynamics of these large molecules could never have been determined classically because without catalysts complicated mixtures are obtained.

Factors that Favour the Extension of the Database on Species

In looking ahead to the future of biochemical thermodynamics I want to point out that as species properties of reactants are added to the database, the number of reactions for which apparent equilibrium constants can be calculated increases exponentially. ATP participates in 41 of the 229 reactions for which I have made calculations at 298.15 K, and urea is involved in one. The “average” reactant is involved in about 6 reactions. Thus we can expect that when a new reactant is added to the database, K' can be calculated for about 6 additional reactions. This leads to an exponential increase in the number of enzyme-catalysed reactions for which apparent equilibrium constants can be calculated. Many equilibrium constants that can be calculated are so large that they cannot be measured directly with today's technology.

>Oxidoreductase reactions are a striking example of this. The table of standard apparent reduction potentials of half reactions can be used to calculate apparent equilibrium constants for any pair of half reactions. If the table of standard apparent reduction potentials contains N different half reactions, the number R of different reactions for which K' can be calculated is given by $R = N(N-1)/2$. The current table of standard apparent reduction potentials in BasicBiochemData3 contains 60 half reactions, and so K' can be calculated for $60 \text{ x } 9/2 = 1770$ oxidoreductase reactions. Of course enzymes are not known for all of these reactions, but I am sure that enzymes will be found for more of them.

Another reason for this exponential increase in the number of reactions with known K' is coupling. Transferase reactions couple two oxidoreductase reactions or two hydrolase reactions, and so knowledge of K' for oxidoreductase reactions and hydrolase reactions yield K' for transferase reactions that have not been studied. Lyase reactions are coupled by definition, and so their K' values can be obtained by multiplying the K' for the two or three reactions coupled by the lyase reaction.

Independence of the Reactivities of Some Groups in Large Molecules

Since many reactants in enzyme-catalysed reactions are rather large molecules, the chemical thermodynamic properties of various groups may be nearly independent, especially at zero ionic strength. As an example of this, when Boeiro-Goates and coworkers obtained $\Delta_f G^0$ and $\Delta_f H^0$ of inosine by calorimetric methods, they calculated the $\Delta_f G^0$ and $\Delta_f H^0$ of all the species of ITP, IDP, and IMP on the assumption that the phosphate pK values and the chemical equilibrium constants for phosphatase reactions and nucleosidase reactions are the same in the ITP series as in the ATP series. This does not mean that the standard transformed Gibbs energies of formation of reactants in the two series are different by a constant increment because the pK values of the purine rings in the two series are different. The pK values for the purine rings are 4.68 for ATP, 4.36 for ADP, 3.99 for AMP, compared with 10.09 for ITP, 9.56 for IDP, and 9.63 for IMP.

This kind of reasoning has been applied to put the guanosine triphosphate series, xanthosine triphosphate series, cytidine triphosphate series, uridine triphosphate series, and thymidine triphosphate series in the next version of BasicBiochemData3. This has required the introduction of the conventions that $\Delta_f G^0$(guanosine, 298.15 K, $I = 0) = 0$ and $\Delta_f G^0$(cytidine, 298.15 K, $I = 0) = 0$, but it is not necessary to introduce the conventions that $\Delta_f G^0$(xanthosine, 298.15 K, $I = 0) = 0$, $\Delta_f G^0$(uridine, 298.15 K, $I = 0) = 0$, and $\Delta_f G^0$(thymidine, 298.15 K, $I = 0) = 0$.

I do not want to over-emphasize this idea of exponential growth because a lot of work is required to analyse the current experimental data summarized by Goldberg and Tewari, and unfortunately not many new measurements of apparent equilibrium constants and enthalpies of reactions are being made today.

Transformed thermodynamic properties are also needed in the discussion of protein-ligand binding [7].

Kinetics of Systems of Enzyme-Catalysed Reactions

For a system of enzyme-catalysed reactions, thermodynamics can do two things if the species properties are known for all of the reactants: (1) At a given temperature, pH, and ionic strength, and given concentrations of the reactants in the system, thermodynamics can tell the direction in which each enzyme-catalysed reaction will go. This information is provided by the following calculation of the transformed Gibbs energy of reaction at species concentrations [*i*] of reactants:

$$\Delta_r G' = \Delta_r G'^0 + RT\ln\prod[i]^{\nu_i} \tag{3}$$

where Π indicates a product involving all the reactants. When $\Delta_r G'$ is negative, the reaction goes to the right. (2) Thermodynamics can also be used to calculate the equilibrium concentrations that will be reached at long times. These are the concentrations that will make $\Delta_r G' = 0$ for all the reactions. The equilibrium concentrations cannot be calculated analytically, but the Newton-Raphson method makes it possible to use a computer to iterate to the equilibrium concentrations. Two short programs in Mathematica make it possible to do this by specifying the stoichiometric number matrix, a list of apparent equilibrium constants, and the initial concentrations of the reactants.

I do not think that biochemists have sufficiently appreciated how much thermodynamics can help in understanding the series and cycles of enzyme-catalysed reactions. The steady-state rate law for a reaction can be used to calculate a small change in the concentration of each reactant. This changes the concentrations, and so thermodynamics can be used again to tell the directions of the reactions. When this process is continued with small steps the equilibrium concentrations will be reached. This equilibrium composition can be checked by comparing it with the equilibrium composition calculated using thermodynamics. Mathematica provides NDSolve that yields numerical solutions to systems of differential equations. This program yields interpolation functions for each reactant that can be plotted. In the absence of Michaelis constants and other kinetic parameters, calculations can always be made when concentrations of reactants are low in comparison with Michaelis constants.

webMathematica

I want to close by describing a recent development that promises to make complicated thermodynamic calculations available to biochemists who do not have the Mathematica application in their computer and do not know how to use Mathematica. webMathematica makes it possible to put up a web site that has boxes to fill in with input for a calculation. When the „Compute“ button is clicked, the problem is solved in a server that has Mathematica and BasicBiochemData3, and the numerical result or plot appears. Such a web site can be used to calculate the apparent equilibrium constant for an enzyme-catalysed reaction at the desired temperature, pH, and ionic strength. In the first box, the user would type in the names of reactants and their stoichiometric numbers. In the second box, the user would type in the names of products and their stoichiometric numbers. Then the desired temperature, pH and ionic strength would be typed in. When the „Compute“ button is clicked, the server on the web that has Mathematica and BasicBiochemData3 will make the calculation and present the desired apparent equilibrium constant on the computer screen. Wolfram research has placed about 40 of these html files on the web as examples, and these sites can be run by going to http://www.wolfram.com, clicking on webMathematica. You do not need Mathematica in your computer to do this. At MIT we are working on the html file to make the calculation of apparent equilibrium constants I have just described.

Conclusion

I hope that in the future it will be possible for biochemists to calculate K', $\Delta_r G'^0$, $\Delta_r H'^0$, $\Delta_r S'^0$, and $\Delta_r N_H$ for a much larger number of enzyme-catalysed reactions at desired temperatures, pH values, and ionic strengths. I believe that these properties will contribute to the understanding of both the thermodynamics and the kinetics of individual enzyme-catalysed reactions and networks of reactions.

Acknowledgements

I am indebted to Robert N. Goldberg for many helpful discussions and to the National Institutes of Health for grant 5-R01-GM48348 – 10.

References

[1] Alberty, R.A., Cornish-Bowden, A., Gibson, Q.H., Goldberg, R. N., Hammes, G.G., Jencks, W., Tipton, K.F., Veech, R., Westerhoff, H.V., Webb, E.C. (1994) Recommendations for nomenclature and tables in biochemical thermodynamics. *Pure Appl. Chem.* **66:**1641–1666. Reprinted in *Eur. J. Biochem.* **240:**1–14 (1996).
http://www.chem.qmw.ac.uk/iubmb/thermod/

[2] Alberty, R.A. (2003) *Thermodynamics of Biochemical Reactions*, Wiley, Hoboken, NJ.

[3] Alberty, R.A. (2006) *Biochemical Thermodynamics: Applications of Mathematica*, Wiley, Hoboken, NJ.

[4] Wolfram Research, 100 World Trade Center Drive, Champaign, IL.
http://www.wolfram.com

[5] Alberty, R.A. (2005) BasicBiochemData3
http://library.wolfram.com/infocenter/MathSource/797

[6] Goldberg, R.N., Tewara, Y.B., Bhat, T.N. Thermodynamics of Enzyme-Catalyzed Reaction
http://xpdb.nist.gov/enzyme_thermodynamics/enzyme1.pl

[7] Alberty, R.A. (2003) ProteinLigandProg.
http://library.wolfram.com/infocenter/MathSource)/4808

Standardization and ‘*In Vivo*’-Like Enzyme Activity Measurements in Yeast

Jildau Bouwman*[1], Karen van Eunen*[1], Isil Tuzun[2], Jarne Postmus[3], André Canelas[4], Joost van der Brink[5], P. Alexander Lindenbergh[1], M. Joost Teixeira de Mattos[2], Gertien J. Smits[3], Pascal A.L. Daran-Lapujade[5], Walter M. van Gulik[4], Rob J. van Spanning[1], Josef J. Heijnen[4], Johannes H. De Winde[5], Stanley Brul[3], Klaas J. Hellingwerf[2], Hans V. Westerhoff[1,6], Barbara M. Bakker[1]

***These authors contributed equally**

[1]Faculty of Earth and Life Sciences, Department of Molecular Cell Physiology, Vrije Universiteit Amsterdam, De Boelelaan 1085, 1081 HV, Amsterdam, The Netherlands;

[2]Department of Molecular Microbial Physiology, Swammerdam Institute for Life Sciences, Faculty of Science, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands;

[3]Swammerdam Institute of Life Sciences, Department of Molecular Biology & Microbial Food Safety, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands;

[4]Department of Biotechnology, TU Delft, Julianalaan 67, 2628 BC Delft, The Netherlands;

[5]Industrial Microbiology, Department of Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, The Netherlands;

[6]Manchester Centre for Integrative Systems Biology, University of Manchester, P.O. Box 88, Sackville Street, Manchester M60 1QD, U.K.

E-Mail: jildau@bio.vu.nl

Received: 27thJuly 2006 / Published: 31st August 2007

Abstract

The aim of this study was to standardize the cultivation conditions of *Saccharomyces cerevisiae* and the *in-vitro* enzyme activity assays between different laboratories. Furthermore, the conditions under which the enzyme activity measurements were carried out were adapted such that they would be as close as possible to *in vivo* conditions, thus yielding results which are relevant for Systems Biology. This approach is different from the classical enzymologists' approach which is to optimize for the highest catalytic activity.

Saccharomyces cerevisiae strain CEN.PK113–7D was cultured in aerobic, glucose- limited chemostats under standardized conditions. It was shown that, in accordance with earlier interlab comparisons, the main culture characteristics, including biomass, dry weight, glucose flux, and mRNA levels of glycolytic enzymes were comparable between five different laboratories.

As could be expected, the V_{max} values of the glycolytic enzymes were lower when measured under *in vivo*-like conditions than in optimized assays, but still sufficient to account for the glycolytic capacity of the cells. The addition of a crowding agent (polyethylene glycol) hardly affected the measured enzyme activities.

Introduction

In Systems Biology the question is addressed as to how biological functions emerge from the interactions between the molecular components of the cell. In a project with 6 groups from three different universities we attempt to understand what regulates changes in glycolytic flux in bakers' yeast as a function of time and upon a number of different perturbations. To this end mRNA concentrations, protein concentrations, V_{max} values, metabolite concentrations and fluxes are experimentally determined and the extent to which various processes contribute to the regulation of metabolic flux are quantified with Regulation Analysis [1–3].

To be able to integrate the results from different laboratories into a coherent picture, we have standardized the cultivation and all assay protocols. We have chosen to examine the regulation of glycolysis in yeast, as it is one of the few pathways for which the kinetic properties of the enzymes are known sufficiently to calculate the flux from the enzyme activities. Furthermore, yeast can be cultured under well-defined steady-state and transient conditions. In this article we describe the results of the standardization process, with an emphasis on the enzyme activity assays.

In the project yeast cells are grown in chemostat cultures under well-defined steady-state conditions in terms of pH, temperature, dissolved oxygen concentration, and substrate and product concentrations. The CEN.PK113 – 7D yeast strain was used since its physiology has been well-characterized and it was used successfully in earlier attempts at standardization [4,5].

All groups started from the same CEN.PK113 – 7D stock, freshly obtained from the Euroscarf collection of yeast strains. Cells were grown at a dilution rate of 0.1 h^{-1} in the mineral medium described by Verduyn [6] supplemented with 7.5 g/L glucose as the sole carbon source, because this medium does not contain sodium, which is toxic for CEN.PK113 – 7D. Samples were taken after at least 5 residence times of chemostat cultivation, to ensure that steady-state conditions are satisfied, and not later than 20 residence times, to prevent physiological changes associated with prolonged chemostat cultivation [7,8].

To estimate the *in vivo* enzyme activities in the cell, assays were developed that mimicked the intracellular conditions as closely as possible. Recent data obtained within one of the contributing labs (Orij, R. and Smits, G. J.) show that the cytosolic pH is approximately 7 if the external pH is 5.0. Therefore, the *in vivo*-like assays were performed at pH 7.0. Intracellular potassium concentrations between 50 mM and 200 mM have been reported for yeast [9]. Therefore, the potassium concentration was fixed at 200 mM. Sulfate was added, since it is the main anion in our medium. Therefore, if magnesium was needed for the assay, it was added in the form of magnesium sulfate instead of magnesium chloride. An intracellular phosphate concentration of 7 mM was reported at a cytosolic pH of 7.5 [10]. Therefore, phosphate was added to a concentration of 10 mM. Finally, it was tested as to whether the enzyme activities were affected by a crowding agent (polyethylene glycol (PEG)). Since macromolecular crowding promotes the binding of macromolecules to each other, the activity of enzymes composed of several subunits could be affected by the addition of PEG.

Materials And Methods

Growth conditions

The growth procedures have been described in detail in Van Hoek *et al.* [11]. Shortly, *S. cerevisiae* strain CEN.PK113 – 7D was grown in aerobic glucose-limited chemostat cultures at a dilution rate of 0.1 h^{-1} at 30 °C in defined mineral medium [6] kept at pH 5.0 with 2 M of KOH. The feed medium contained 42 mM of glucose (7.5 g l^{-1}). The chemostats were stirred at a rate of 800 rpm, aerated at 0.5vvm and most equipment was acquired from Applikon (Schiedam, NL).

qPCR

Oligonucleotide primers were designed to amplify an 80 – 120bp amplicon. PDI1 was chosen as an internal standard. Primers were designed using Primer Express software 1.0 (PE Applied Biosystems, Foster City, CA, U.S.A). PCR reactions (20 µl) were set up and run as described by the manufacturer. Briefly, the reactions contained 10 µl SYBR Green

PCR Core Kit (PE Applied Biosystems, Foster City, CA, U.S.A), 20 pmol of each primer (Sigma or Eurogentec, Seraing, Belgium); and 3 µl of cDNA template (equivalent to 1 ng RNA). Amplification, data acquisition, and data analysis were carried out in the ABI 7900 Prism Sequence Detector (once at 2 min, 50 °C; 10 min, 95 °C; and 40 cycles at 95 °C, 15 s; 59 °C, 1 min). The calculated cycle threshold values (Ct) were exported to Microsoft Excel for analysis using the ΔΔCt method [12]. Dissociation curves (Dissociation Curves 1.0 f. software, PE Applied Biosystems, Foster City, CA, U.S.A) of PCR products were run to verify by amplification of the correct product.

Fermentative capacity assay

>Steady-state fluxes were measured for 30 min in a cell suspension kept anaerobic at 30 °C in a setup described by Van Hoek *et al.* [11] for the determination of fermentative capacity, with the modification that the headspace was flushed with N_2 instead of CO_2. Ethanol, glucose, glycerol, succinate, acetate, and trehalose were measured by HPLC (300 mm x 7.8 mm ion exchange column Rezex ROA-organic acid (Phenomenex), with 22.5 mM H_2SO_4 kept at 55 °C as eluent at the flow rate of 0.5 ml min^{-1}).

Enzyme activity measurements

Cell free extracts were prepared by sonication (6 times for 30 s.) with glass beads (250 – 500 µm) on ice water as described by Van Hoek *et al.* [11]. The total protein content of the cell free extract was measured using the Lowry method [13]. The absorbance (750nm) was measured in a Novostar plate reader (BMG Labtech, Germany).

The enzyme assays were carried out on four dilutions of freshly prepared extracts through NAD(P)H-linked assays as described by Van Hoek *et al.* [11], with a Cobas Bio automated analyser for spectroscopic measurements (Roche, Switzerland). Enzyme assays were performed in three different buffers. The first set of activities was measured according to Van Hoek *et al.* [11], in which the buffer content was different for the different enzymes. The second set of enzyme assays was performed in 100 mM K_2SO_4, 10 mM KH_2PO_4, and at a pH of 7.0. In every assay essential components were added if required (NADH, ATP, EDTA, $MgSO_4$, coupling enzymes and substrates). The third set of assays was done in the same conditions as the second assay, but with the addition of 10 % PEG (3350).

Table 1. Enzyme activity protocols according to Van Hoek *et al.*, [11]. *In the '*in vivo*'-like enzyme activities the buffer concentrations were for all enzyme assays 100 mM K_2SO_4, 10 mM KH_2PO_4, 5mM $MgSO_4$ (if needed), and the pH was 7.0. Furthermore, in the '*in vivo*'-like enzyme activities all concentrations were the same as in the Van Hoek protocol.

	ADH	ALD	GAPDH	HXK	TPI	PDC	PFK	PGI	PGK	GPM	PYK	G6PDH
Salt1* **conc**	Glycine 50mM	Tris-HCl 50mM	Triethanolamine 100mM	Imidazol-HCl 50mM	Triethanolamine 100mM	Imidazol-HCl 40mM	Imidazol-HCl 50mM	Tris-HCl 50mM	Triethanolamine 100mM	Triethanolamine 100mM	Cacodylic Acid 100mM	Triethanolamine 100mM
Salt 2* **conc**		KCl 100mM	EDTA 1mM			TPP 0.2mM			EDTA 1mM	2,3dPGA	KCl 100mM	
Salt 3* **conc**			MgSO4 1.5mM	MgCl2 5mM		MgCl2 5mM	MgCl2 5mM	MgCl2 5mM	MgSO4 1.5mM	MgSO4 1.5mM	MgCl2 25mM	
pH*	9	7.5	7.6	7.6	7.6	6.5	7.0	8.0	7.6	7.6	6.2	7.6
Substrate1 **conc**			ATP 1mM	ATP 1mM			F2,6bP 0.1mM		ATP 1mM	ADP 10mM	ADP 10mM	
Substrate2 **conc**	NAD 1mM	NADH 0.15mM	NADH 0.15mM	NADP 1mM	NADH 0.15mM	NADH 0.15mM	NADH 0.15mM	NADP 0.4mM	NADH 0.15mM	NADH 0.15mM	NADH 0.15mM	NADH 0.15mM
Startreagent1 **conc**	Ethanol 100mM	F1,6bP 2mM	3-PGA 5mM	Glucose 10mM	GAP 5.8MM	Pyruvate 50mM	F6P 0.25mM	F6P 2mM	3-PGA 5mM	3-PGA 5mM	F1,6bP 1mM	F1,6bP 1mM
Startreagent2 **conc**							ATP 0.5mM					TPI 75U/ml
Startreagent3 **conc**												ALD 1U/ml
Enzyme1 **conc**		G3PDH 0.6U/ml	PGK 22.5U/ml	G6PDH 1.8U/ml	G3PDH 8.5U/ml	ADH 88U/ml	ALD 0.45U/ml	G6PDH 1.75U/ml	G3PDH 8.0U/ml	PYK 13U/ml	LDH 11.3U/ml	
Enzyme2 **conc**		TPI 1.8U/ml					G3PDH 0.6U/ml			LDH 11.3U/ml		
Enzyme3 **conc**							TPI 1.8U/ml			ENO 2U/ml		

Results

Interlab comparison of culture characteristics

To be able to later integrate all project data in one model, the cultivation conditions should be the same in the different laboratories. We compared the steady-state properties of the standard chemostat cultures, which were performed in four different labs (Table 2). The glucose flux and biomass yield on glucose were similar in all groups. The specific consumed oxygen and produced carbon dioxide however, deviated from the published standard (reference) in group 2 and group 4. This is probably a matter of calibration of our gas-analysing systems, which can be improved. The respiratory quotient (i.e. the ratio of the specific CO_2 production over the specific oxygen consumption) was close to 1 in all cultures, implying that they were all fully respiratory. Accordingly, no ethanol was detected in any of the cultures.

In addition, the levels of a number of glycolytic transcripts were compared between two labs (Fig. 1). For this we subtracted the cycle threshold of the control gene PDI1 and the transcript of interest. This gives a relative estimate of the amount of transcript present in the cell. The cycle threshold represents the number of cycles after which a certain sample exceeds the threshold signal intensity. Thus, when two samples differ by one cycle, the one with the lowest cycle threshold has a two-fold higher mRNA concentration than the other sample. The transcript levels from two different laboratories show a similar pattern. This is in agreement with a more extensive inter-laboratory comparison of transcript levels in samples that were obtained from the same culture conditions and analysed using micro-arrays [4].

These data show that if we standardize the conditions in different laboratories, we end up with a similar culture.

Table 2. Steady-state properties of the aerobic glucose-limited chemostat cultivation in the different laboratories, compared to a published reference [14] (N.D. is not detectable).

	RQ	**qCO2 Produced** (mmol/gDW.h)	**qOJ Consumed** (mmol/gDW.h)	**qglucose Consumed** (mmol/gDW.h)	**qethanol Produced** (mmol/gDW.h)	**Yglucose** (gDW/ gglu)	**C-recov-ery** (%)
Literature	1.0	2.3 ± 0.3	2.8 ± 0.3	1.1 ± 0.0	N.D.	0.49 ± 0.0	98
Group 1	1.1	3.0 ± 0.1	2.3 ± 0.0	1.0 ± 0.1	ND.	0.56 ± 0.0	98
Group2	1.0	2.5 ± 0.0	2.4 ± 0.0	1.0	ND.	0.51	100
Group3	1.0	2.8 ± 0.1	2.7 ± 0.1	1.1 ± 0.1	ND.	0.48 ± 0.0	101
Group4	1.0	2.3 ± 0.2	2.3 ± 0.4	1.1 ± 0.1	ND.	0.48 ± 0.0	94
Group5	1.0	2.3 ± 0.1	2.7 ± 0.1	1.1 ± 0.0	ND.	0.50 ± 0.0	103

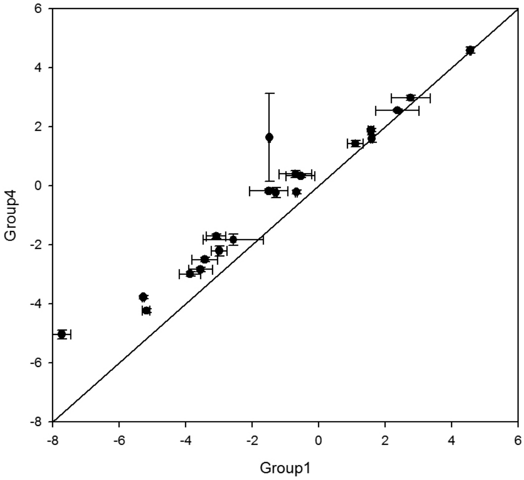


Figure 1. qPCR signal relative to our control gene (PDI1). We subtracted the cycle threshold signal of the control gene (PDI1) from the cycle threshold of our transcript of interest. Comparing these data from different laboratories gives an estimate of the similarity of our fermentors at the transcriptional level.

Enzyme activity assays under in vivo conditions

We analysed the effect of the different buffers on the enzyme activities. For most enzymes the new *in vivo* protocol resulted in lower enzyme activities than the existing optimized protocol (Fig. 2), as should have been expected. Aldolase was the only enzyme for which the enzyme activity increased using the new assays. The activity of TPI was in large excess according to the Van Hoek assay, but the under the *in vivo*-like assay conditions it is of the same order of magnitude as the other enzyme activities. This was probably caused by the addition of phosphate, which is known to be an inhibitor of TPI [15]. Even in the *in vivo*-like assay the activity of TPI was among the highest measured activities, in agreement with the fact that it is close to equilibrium in most cases [16].

Addition of a crowding agent (10% PEG) had hardly any effect on the enzyme activities (Fig 2).

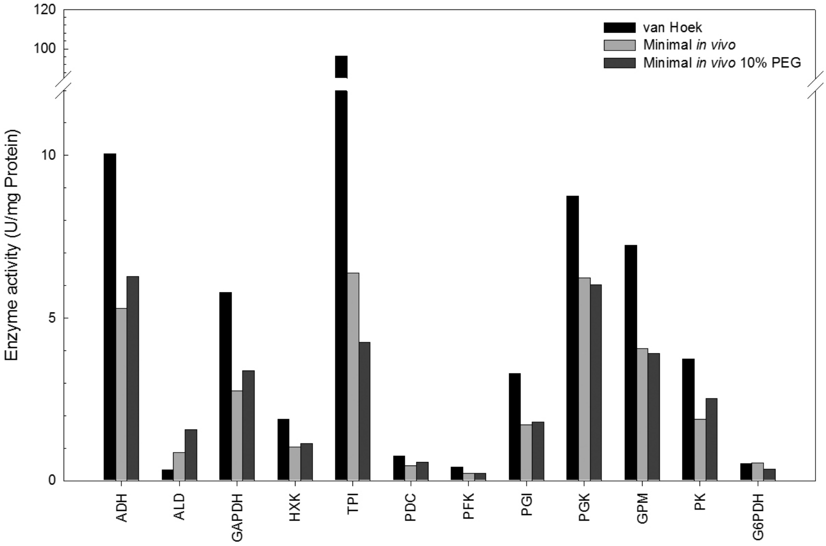


Figure 2. Enzyme activities relative to the amount of total soluble protein. The V_{max} values analysed with the different buffers are shown: Van Hoek (buffer according to [11]) (black columns), minimal '*in vivo*'-like (light grey columns), and minimal '*in vivo*'-like with 10% PEG (grey columns). The V_{max} was measured in the catabolic direction (from glucose to ethanol) except for the enzymes ADH, GAPDH, PGI, and PGK, which were measured in the reverse direction.

The measured V_{max} values should be enough to account for the fluxes measured both under the steady-state conditions and in the fermentative capacity assay. The glucose flux under the steady-state conditions has been shown to be 4.0 mmol h^{-1} $gProt^{-1}$. The ethanol flux in the fermentative capacity measurements was 29.6 mmol h^{-1} $gProt^{-1}$. Taking into account the direction in which the enzyme activities were measured and the branching of glycolysis (roughly the flux through the lower part of glycolysis should be twice as high as the flux through the upper part), the measured enzyme activities in all assays are large enough to account for the calculated fluxes (Table 3).

Table 3. V_{max} values (mmol h^{-1} gProt^{-1}) from Fig. 2 were recalculated in the direction of the flux from glucose to ethanol of the in vivo assay (V_{max} values were obtained in the catabolic direction making use of the equilibrium constants, Michaelis-Menten constants and forward and reverse V_{max} values from literature (ADH [17]), (GAPDH [18]), (PGI [19]), (PGK [20])).

	ADH	*ALD*	*GAPDH*	*HXK*	*PDC*	*PFK*	*PGI*	*PGK*	*GPM*	*PK*
mmol h^{-1} gProt^{-1}	*318.3*	*51.4*	*29.7*	*62.6*	*27.1*	*13.1*	*150.8*	*29.9*	*235*	*113*

Conclusions

This paper describes the standardization of yeast cultivation and analysis for Systems Biology research. A single laboratory will never be able to obtain the large amount of data that are required to understand the cell in terms of the interactions between its molecular components. In view of the quantitative nature of this type of research, it is of vital importance to standardize cultivation conditions and analytical procedures between laboratories.

In this study the same yeast strain and cultivation conditions were used that were applied earlier [14] and obtained comparable results in four different laboratories. We optimized the enzyme activity assays to mimic the *in vivo* conditions as closely as possible. The measured activities could account for the calculated flux in the chemostat cultures and in an off-line fermentative capacity assay.

For these assays the ion content of the cell was estimated from literature. In the near future, we will further improve the method by analysing the intracellular ion concentrations using inductively coupled plasma-atomic emission spectroscopy (ICP-AES) [21] and apply the results in the *in vivo*-like assay conditions.

Acknowledgements

This project was financially supported by the IOP Genomics programme of Senter. The CEN.PK113 – 7D strain was kindly donated by P. Kötter, Euroscarf, Frankfurt.

References

[1] Ter Kuile, B.H., Westerhoff, H.V. (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.* **500:**169–171.

[2] Rossell, S., van der Weijden, C.C., Lindenbergh, A., van Tuijl, A., Francke, C., Bakker, B.M., Westerhoff, H. V. (2006) Unraveling the complexity of flux regulation: a new method demonstrated for nutrient starvation in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A.* **103:**2166–2171.

[3] Daran-Lapujade, P., Rossell, S., van Gulik, W. M., Luttik, M.A.H., de Groot, M.J.L., Slijper, M., Heck, A.J.R., Daran, J.M., de Winde, J.H., Westerhoff, H.V., Pronk, J.T., Bakker, B.M. What regulates glycolytic gene expression? Quantifying transcription, translation, posttranslational and metabolic regulation when yeast adjusts its carbon and energy metabolism. *Submitted.*

[4] Piper, M.D.W, Daran-Lapujade, P., Bro, C., Regenberg, B., Knudsen, S., Nielsen, J., Pronk, J.T. (2002) Reproducibility of oligonucleotide microarray transcriptome analyses. *J. Biol Chem.* **277:**37001–37008.

[5] Van Dijken, J.P., Bauer, J., Brambilla, L., Duboc, P., Francois, J.M., Gancedo, C., Giuseppin, M.L., Heijnen, J.J., Hoare, M., Lange, H.C., Madden, E.A., Niederberger, P., Nielsen, J., Parrou, J.L., Petit, T., Porro, D., Reuss, M., van Riel, N., Rizzi, M., Steensma, H.Y., Verrips, C.T., Vindelov, J., Pronk, J.T. (2000) An interlaboratory comparison of physiological and genetic properties of four *Saccharomyces cerevisiae* strains. *Enzyme Microb Technol.* **26:**706–714.

[6] Verduyn, C., Postma, E., Scheffers, W.A., Van Dijken, J.P. (1992) Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation. *Yeast* **8:**501–517.

[7] Jansen, M.L., Diderich, J.A., Mashego, M., Hassane, A., de Winde, J.H., Daran-Lapujade, P., Pronk, J.T. (2005) Prolonged selection in aerobic, glucose-limited chemostat cultures of *Saccharomyces cerevisiae* causes a partial loss of glycolytic capacity. *Microbiology* **151:**1657–1669.

[8] Wu, L., Mashego, M.R., Proell, A.M., Vinke, J.L., Ras, C., van Dam, J., van Winden, W.A., van Gulik, W.M., Heijnen, J.J. (2006) In vivo kinetics of primary metabolism in *Saccharomyces cerevisiae* studied through prolonged chemostat cultivation. *Metab Eng.* **8:**160–171.

[9] Roomans, G.M., Seveus, L.A. (1976) Subcellular localization of diffusible ions in the yeast *Saccharomyces cerevisiae*: quantitative microprobe analysis of thin freeze-dried sections. *J. Cell Sci.* **21:**119–127.

[10] Den Hollander, J.A., Ugurbil, K., Brown, T.R., Shulman, R.G. (1981) Phosphorus-31 nuclear magnetic resonance studies of the effect of oxygen upon glycolysis in yeast. *Biochemistry* **20:**5871–5880.

[11] Van Hoek, P., Van Dijken, J.P., Pronk, J.T. (1998) Effect of specific growth rate on fermentative capacity of baker's yeast. *Appl. Environ. Microbiol.* **64:**4226–4233.

[12] Spijker, S., Houtzager, S.W., De Gunst, M.C., De Boer, W.P., Schoffelmeer, A.N., Smit, A.B. (2004) Morphine exposure and abstinence define specific stages of gene expression in the rat nucleus accumbens. *FASEB J.* **18:**848**-**850.

[13] Lowry, O.H., Rosebrough, N.J., Farr, A.L., Randall, R.J. (1951) *J. Biol. Chem.* **193:**265–275.

[14] Tai, S.L., Boer, V.M., Daran-Lapujade, P., Walsh, M.C., de Winde, J.H., Daran, J.M., Pronk, J.T. (2005) Two-dimensional transcriptome analysis in chemostat cultures. Combinatorial effects of oxygen availability and macronutrient limitation in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **280:**437–447.

[15] Burton, P.M., Waley, S.G. (1966) The active centre of triose phosphate isomerase. *Biochem. J.* **100:**702–710.

[16] Teusink, B., Westerhoff, H.V. (2000) 'Slave' metabolites and enzymes. A rapid way of delineating metabolic control. *Eur. J. Biochem.* **267:**1889–1893.

[17] Ganzhorn, A.J., Green, D.W., Hershey, A.D., Gould, R.M., Plapp, B.V. (1987) Kinetic characterization of yeast alcohol dehydrogenases. Amino acid residue 294 and substrate specificity. *J. Biol. Chem.* **262:**3754–3761.

[18] Byers, L.D., She, H.S., Alayoff, A. (1979) Interaction of phosphate analogues with glyceraldehyde-3-phosphate dehydrogenase. *Biochemistry* **18:**2471–2480.

[19] Tewari, Y.B., Steckler, D.K., Goldberg, R.N. (1988) Thermodynamics of isomerization reactions involving sugar phosphates. *J. Biol. Chem.* **263:**3664–3669.

[20] Bergmeyer, H.U. (1974) *Methods of Enzymatic Analysis.* Verlag Chemie, Weinheim.

[21] Eide, D.J., Clark, S., Nair, T.M., Gehl, M., Gribskov, M., Guerinot, M.L., Harper, J.F. (2005) Characterization of the yeast ionome: a genome-wide analysis of nutrient mineral and trace element homeostasis in *Saccharomyces cerevisiae*. *Genome Biol.* **6:**R77.

Assay of Enzymes with Insoluble or Unknown Substrates: The Membrane-Bound Quinone Reductases as an Example

Richard Cammack

Department of Biochemistry, King's College London, U.K.

E-Mail: Richard.cammack@kcl.ac.uk

Received: 18th October 2006 / Published: 31st August 2007

Abstract

The conventional assay method for the majority of enzymes envisages a reaction between substrates in aqueous solution. A measurable concentration of product is accumulated over time. This paradigm has served well for the characterization of many enzymes. Variations of the method, often using chromogenic or fluorogenic substrates, have been developed and are widely used for purposes such as clinical diagnosis and screening. There are some metabolically important enzymes for which the only published assay methods use artificial substrates. Some of these are oxidoreductases that use artificial mediators, and are listed in the EC list under EC 1.*x*.99. For computational reconstruction of the metabolism of a cell, however, it is necessary to use kinetic data from assays that reflect the physiological function in the cell, and the physiological substrates. For some oxidoreductases it is known, or considered likely that the acceptors are water-insoluble membrane-bound quinones such as ubiquinone or menaquinone, which present particular problems for measurement of kinetic parameters. Succinate dehydrogenase/fumarate reductase is considered as an example. The oxidoreductases from membranes must be rendered soluble by detergents, which alter their kinetic behaviour. Uncertainty about the way of measuring activity of such enzymes has led to confusion in textbooks and metabolic maps, such as the persistent myth that free FAD is the acceptor for succinate dehydrogenase and related enzymes. New strategies are discussed to measure electron-transfer flux, under conditions that reflect the physiological activity of

membrane-associated oxidoreductases. An example is direct electrochemistry of enzymes adsorbed onto carbon surface. In favourable cases this method is able to observe electron flux both within and through individual enzyme molecules. The kinetic parameters and substrate specificity of membrane-bound oxidoreductases may be obtained in this way.

Conventional Enzyme Assays

The STRENDA initiative aims for the standardization of enzyme assay protocols, with the prospect of simulating the metabolism of the cell by computational reconstruction of reaction pathways. It also underlines the need for the accurate classification of enzymes according to their physiological activities. The obvious first requirement for each enzyme is to know what its substrates are in the pathway. When determining the kinetic parameters, the natural substrates should be used, even if they are unstable, insoluble or expensive to produce. The assay method should reflect the conditions (pH, ionic strength, protein concentrations, etc.) in the cell for which the metabolism is being reconstructed. Ideally the flux through the enzyme should be measured in the presence of other metabolites, which might be allosteric activators or inhibitors, and of any proteins with which the enzyme interacts, all at the physiological concentration. At present, there are some enzymes for which the only published data fall well short of this ideal. Just as some enzymes are easier to assay than others, some enzymes are easier to classify than others. For some, the EC classification is incomplete, and their assay methods use artificial cosubstrates.

The classic enzyme assay, since the work of Michaelis and Menten, is one carried out by an enzyme in dilute solution, with a fixed, relatively high initial concentration of substrate(s) and in the absence of product. The reaction rate is determined by the change of concentration of product P or substrate S, for example by spectrophotometry:

$$\mathrm{v} = d[\mathrm{P}]/dt = -\mathrm{d}[\mathrm{S}]/dt \qquad (1)$$

This approach is popular since it allows the accurate determination of catalytic parameters such as K_m and k_{cat}. As is well documented [1] however it deviates significantly from the situation in the cell where the enzyme substrate and product are in turnover but their concentrations do not change substantially.

The EC classification lends itself most easily to the description of enzyme-catalysed reactions involving up to two substrates, such as an X-transferase (EC class 2):

$$\text{X-donor} + \text{acceptor} = \text{donor} + \text{X-acceptor} \qquad (2)$$

Oxidoreductases
For an oxidoreductase (EC class 1) the reaction can be written:

$$\text{reductant} + \text{oxidant} = \text{oxidized reductant} + \text{reduced oxidant} \qquad (3)$$

the EC number of an enzyme of this type takes the form EC 1.*x*.*y*.*z*, where *x* represents the type of reducing substrate, and *y* the type of oxidant. The class EC 1 can be represented in terms of a matrix of *x* and *y*, as illustrated in Fig. 1, where the vertical (*z*) axis represents the number of enzymes of each type.

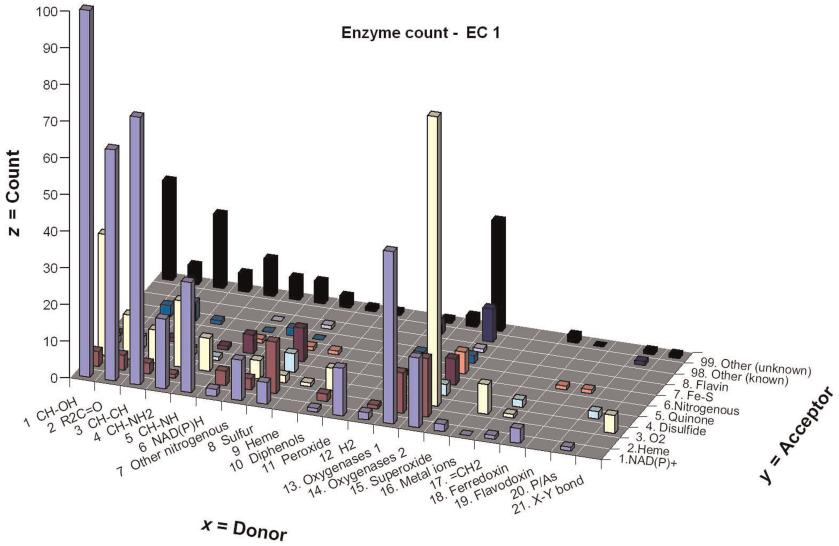


Figure 1. Classification of the oxidoreductases in Class 1 of the EC list. The vertical axis shows the number of enzymes of each sub-subclass, except that the count for EC 1.1.1.1 is truncated; there are currently 288 such enzymes.

Enzyme Assays Using Artificial Substrates

Enzyme assays are among the most widely used of biochemical measurements. They were developed for many other purposes, not just to characterize enzymes or study metabolism (Table 1). Colorimetric enzyme assays using chromogenic or fluorogenic substrates are routinely used. For example, in the pioneering studies of Jacob and Monod [2] on induction of the *lac* operon of *Escherichia coli*, hydrolysis of *o*-nitrophenyl-β-galactoside was used to measure the level of expression of β-galactoside, EC 3.2.1.45. The product nitrophenol has a bright yellow colour in alkaline solution. This well-studied detection system is still used to investigate gene regulation; for example the gene for β-galactosidase is coupled to a promoter of interest, and *E. coli* cells in which the promoter is activated can be observed by the colour reaction. The assay has been refined by the use of substrates such as X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside) which generates an indigo precipitate, and 4-methyl umbelliferyl β-D-galactoside, which releases a fluorescent product [3]. The colorimetric assay for β-galactosidase was found to be effective in the diagnosis and monitoring of Gaucher disease [4]; the enzyme deficiency being measured is in fact for a glucosylceramidase, EC 3.2.1.45.

Enzyme assays are widely used in clinical medicine for diagnosis and monitoring of disease [5]. They are used to measure unusually high or low levels of enzyme, or the presence of enzyme in an inappropriate location, as an indication of tissue damage. For such assays, all that is required is that the method be sensitive, robust, and specific to the enzyme of interest. Other compounds present in the sample should not interfere. It is not important that the assay should reflect the precise activity that the enzyme displays in the living cell. A few examples are listed in Table 1. Sometimes the physiological substrate of the enzyme being measured has not been established; for example alkaline phosphatase in blood is an indicator of bone disease. Cytochemical stains for respiratory enzymes are used in histology, and enzyme activity stains are used to visualize redox enzymes on non-denaturing electrophoresis gels. Vital stains using the reduction of a tetrazolium compound to a coloured formazan are basically tests for oxidoreductases [6]. The ability of an enzyme molecule to turn over many molecules of substrate represents a large amplification factor, which is exploited in enzyme-linked tests such as enzyme-linked immunosorbent assays (ELISA) [7].

Table 1. Examples of enzyme assays used in clinical diagnosis and screening.

Enzyme assay	Test for	Enzyme	EC No
p-nitrophenyl phosphatase	liver function and bone disease	alkaline phosphatase	EC 3.1.3.1
Paraoxonase	antioxidant stress	aryldialkyl- phosphatase	EC 3.1.8.1
peroxidation of 3,3',5,5'-tetra-methylbenzidine	ELISA	peroxidase	EC 1.11.1.7
β-galactosidase or glucosidase	gangliosidosis	glucosylceramidase	EC 3.2.1.45

The use of redox-active dyes and other small molecules as electron acceptors and donors to enzymes derives from some of the earliest research in biochemistry, and was important in the discovery of important cellular processes such as the light reactions of photosynthesis [8] and respiration [9]. It has been known since early in the 20th century that extracts of living tissues can catalyse the oxidation and reduction of compounds such as 2,6-dichloroindophenol and $K_3Fe(CN)_6$ (reviewed by Keilin, [9]), long before the molecular properties of the protein complexes involved were established.

Some reducing compounds in the cell, such as NADH, react poorly with dyes but the process is facilitated by compounds such as phenazine methosulfate that will facilitate the transfer of electrons dyes such as tetrazolium compounds. Such compounds are known as *mediators*. Mediators can also facilitate electron transfer to platinum electrodes, and have been used in the determination of thermodynamic oxidation–reduction potentials of metabolites and proteins such as cytochromes [10, 11]. Mediators have their difficulties when used to study the kinetics of enzymes. The reaction of enzymes with redox mediators is unpredictable, and few systematic studies have been made since the early days [12]. The most effective mediators are usually one-electron carriers that can produce free radicals.

Mediators often react at sites that are not accessible to the physiological substrates; or they may not react at the active sites, for reasons of steric hindrance or electrostatic charge. Nevertheless, assays using the oxidation and reduction of mediator dyes have continued to be widely used.

Table 2. Midpoint potentials of some redox couples, at pH 7, in millivolts *vs* the hydrogen electrode.

Compound	E°, mV
$K_3Fe(CN)_6/K_4Fe(CN)_6$	+420
2,6-dichloroindophenol	+217
Phenazine	+80
Ubiquinone Q/QH_2	+60
Fumarate/succinate	+30
Methylene blue	-11
Menaquinone MK/MKH_2	-60
Indigodisulfonate	-125
$FAD/FADH_2$	-207
$NAD^+/NADH$	-320
Methyl viologen	-440

A considerable number of oxidoreductases were first studied by assay with mediators or artificial donors (Table 2). Succinate dehydrogenase was shown to act with methylene blue [13]. These assays helped to establish the specificity of the enzyme for their substrates and inhibitors, though they obviously could not be used to investigate the kinetics of reaction with the physiological cosubstrate. Some of them were documented in the first list of enzymes [14] which became the EC list. Rather than omit these enzymes from the classification, they were placed in the sub-subclass “99“ [15]. They can be seen in the back row of the chart of EC 1 enzymes (coloured in black in Fig. 1). This sub-subclass included any enzyme which could not be listed anywhere else, including enzymes for which one substrate is uncertain; and enzymes for which the substrates are known, but for which no other subclass in the list is suitable. Eventually the “99“ enzymes should all be deleted or relocated elsewhere in the list. A recent proposal in the list, not yet implemented, is to invoke the classification EC 1.*x*.98.*z* for enzymes where the acceptor is known but for which there is no suitable sub-subclass, and EC 1.97.*y*.*z* for enzymes where the donor is known but for which there is no subclass. Some of the sub-subclass “99“ enzymes proved on further investigation to be degraded or incomplete parts of enzymes, and are being eliminated after further investigation. Others are from organisms that have been little studied. However some are from well-studied organisms, such as *E. coli* and they are particularly interesting as they point to gaps in our knowledge of metabolism, and possibly further complexities in the organization of the cell.

The flavins, FAD and FMN, are redox-active in their own right, and can act as mediators. However apart from a few enzymes (dioxygenases in subclass EC 1.14) that appear to use the free flavin as donor, FMN and FAD form part of an enzyme and are considered as

prosthetic groups. In some enzyme assays FMN or FAD are required to supplement a flavin that is a dissociable prosthetic group in the enzyme. Despite this, some textbooks and metabolic pathways indicate that $FADH_2$ is the product of succinate dehydrogenase.
The reaction may sometimes be found written in the form:

$$\text{succinate} + \text{FAD} = \text{fumarate} + \text{FADH}_2 \qquad (4)$$

which treats FAD as a dissociable substrate. This is usually mentioned in the context of the citrate cycle, in which the soluble products of pyruvate oxidation are described as CO_2, 4 NADH and 1 $FADH_2$. Although this analogy to NAD is superficially attractive, it represents confusion between a prosthetic group (which is part of the enzyme) and a cosubstrate (a cofactor that is a substrate of the reaction catalysed). We now know that the enzyme in mitochondria that oxidizes succinate is a membrane-bound enzyme, succinate dehydrogenase (ubiquinone), EC 1.3.5.1, also known as Complex II of the respiratory chain [16]. This is a four-subunit enzyme, which contains FAD, iron–sulfur clusters and heme (Fig. 2a).

Thus Equation 4 should be written:

$$\text{succinate} + \text{Q} = \text{fumarate} + \text{QH}_2 \qquad (5)$$

The origins of the idea of FAD as an acceptor may date from the 1950 s, when Massey and Singer [17] showed that FAD could act as a mediator with soluble "succinic dehydrogenase"; this was a preparation of the two membrane-extrinsic subunits, still in the list as EC 1.3.99.1, succinate dehydrogenase. These authors did not suggest that $FADH_2$ was the acceptor, which is unlikely for several reasons.

- $FADH_2$ could not dissociate from the enzyme, as it is covalently bound to a cysteine residue in the protein.
- Free $FADH_2$, unlike NADH, is readily oxidized by O_2, producing toxic oxygen radicals.
- The equilibrium of the reaction of Equation 4 would lie in the direction of reduction of fumarate to succinate since the midpoint potential of the fumarate/succinate couple (30 mV at pH 7; Table 2) is more positive than that of FAD/$FADH_2$ (-210 mV).
- FAD is a carrier (as are the iron–sulfur clusters and heme) in the flux of reducing equivalents from succinate to ubiquinone (Fig. 2a).
- Equation 5 is formally a transfer of two hydrogen atoms from succinate to ubiquinone. In fact, the process involves electron transfers. Flavins can be reduced in one-electron steps with the formation of an intermediate semiquinone; hence they act as a transformer between hydrogen- and electron-transfer reactions. Quinones such as ubiquinone are also best considered as electron carriers, the reaction going through the formation of a semiquinone radical; to preserve

charge neutrality, transfer of each electron is usually accompanied by a proton [18].

Classification of Membrane-Bound Enzymes and the Reactions Catalysed

When annotating a metabolic pathway, the aim is to identify the enzyme by the reaction catalysed in a particular step. The identification of an enzyme in the EC list is not quite the same; it depends on the observed substrate preferences of enzymes that have actually been isolated [18, 19]. This distinction is reflected in the use of the term "reaction class" (RC) in the KEGG annotation of genomes [20]. The same EC number may be attached, quite correctly, to different reactions in different organisms or cells. This happens because the enzyme is of broad specificity, but it only encounters a particular substrate in that organism. For example, an alcohol dehydrogenase that oxidizes one alcohol in a particular organism may be indistinguishable from one that oxidizes another alcohol in another organism [21]. An enzyme only needs to be specific enough for the purposes of catalytic efficiency, and to avoid unwanted reactions with other metabolites found in the cell. There will be no evolutionary pressure to avoid reactions with molecules that the enzyme never encounters. The enzyme is only induced in the presence of that particular substrate in that organism; the specificity lies not in the enzyme, but in the regulatory systems that induce its biosynthesis.

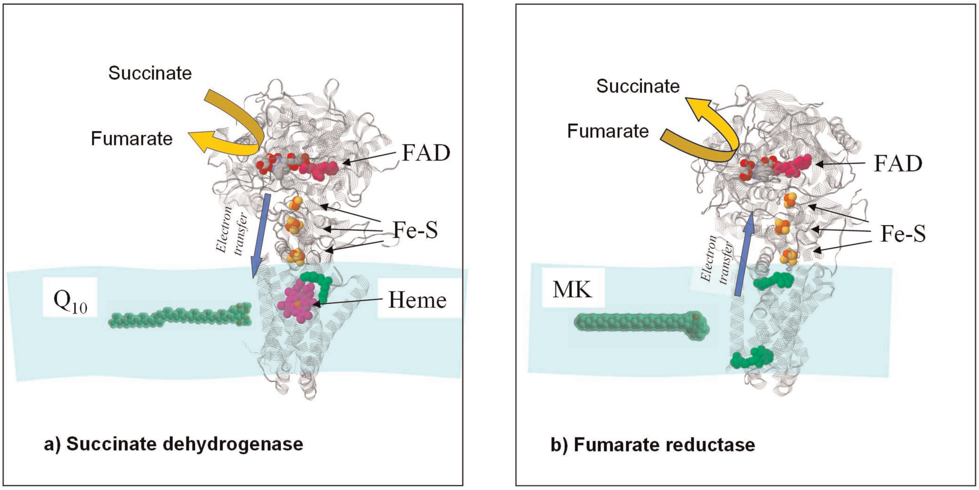


Figure 2. Proposed organization of electron carriers in a) succinate dehydrogenase in the membrane of aerobic *E. coli* cells and b) fumarate reductase in the membrane of anaerobic *E. coli* cells. The membrane is indicated in pale green. The structures were determined from Protein Databank files and 1LOV, respectively, drawn with RasMol version 2.7.2.1. The position of ubiquinone (Q) and menaquinone (MK) are indicated in green, drawn with Chem3D (Cambridgesoft). Two binding sites for the MK headgroup in fumarate reductase are indicated in b).

A simplifying principle of the EC classification, which affects their use in annotation of metabolic pathways, is that the direction of reaction is not considered in allocating a subclass or sub-subclass. Enzymes that catalyse the same reaction in opposite directions will have the same EC class, unless it can be demonstrated that they have different substrate specificity. For example in aerobic conditions the enzyme that oxidizes succinate to fumarate, as in mitochondria, is Complex II, succinate: ubiquinone reductase [16]. Under anaerobic conditions expression of this form of the enzyme is suppressed, and a similar enzyme, fumarate reductase, is expressed, which uses fumarate as an oxidant. Both enzymes are listed as EC 1.3.5.1, and they have a similar molecular architecture (Fig. 2). These enzymes might be classed separately if it could be demonstrated that they are specific for a particular quinone. In fact the quinones present under the two different growth conditions are different, ubiquinone under aerobic conditions, and menaquinone, which has a lower midpoint reduction potential (Table 2) under anaerobic conditions. However so far there have been few cases where it has been possible to demonstrate specificity of membrane-bound oxidoreductases for particular naturally-occurring substrates. This may be due to the difficulty of measuring kinetic parameters of such reactions. There are biophysical methods to do this, for example in photosynthetic reaction centres, where it was shown that the length and structure side-chain of the quinones has a significant influence on the rate of reaction [22].

Within a membrane such as the mitochondrial inner membrane, a “pool“ of quinones such as ubiquinone (Q) or menaquinone (MK) diffuses in this phase, and interacts with specific quinone binding regions of the membrane protein complexes [23]. These quinones have long prenyl chains, and are virtually insoluble in water. They are located in the hydrophobic region between the bilayer leaflets of cell membranes [24]. The quinone/quinol headgroups are somewhat hydrophilic, and tend to orient toward the aqueous layers on either side of the membrane (Fig. 2). The quinones interact with substrates in the aqueous phases by electron transport through the membrane protein complexes [25].

For membrane-bound enzymes that react with water-insoluble quinones, the paradigm for the enzyme assay described above (Equation 1) cannot be readily applied. The amount of quinone is confined to the small volume of the lipid bilayer, so the “initial rate“ of an enzyme reaction will produce a very small amount of product. Because the quinones are virtually insoluble in water, their oxidation and reduction cannot readily be followed by conventional solution methods such as spectrophotometry. In order to study them in solution, detergents are added, so that both the enzyme and substrate are present in the form of detergent micelles. Now, if the oxidation–reduction of the quinone is measured, the kinetics of diffusion in and between micelles is a complicating factor.

Smaller quinone molecules such as menadione or Q_1 are more water-soluble, and may be used instead of the native substrates. However any quinone with a shorter chain length than 5 prenyl units will not partition correctly in the membrane, and so its interaction with the quinone-binding sites may be different [26]. Small quinone molecules such as menadione or Q_1 can act as general mediators, accessing redox centres outside of the membrane bilayer, and transferring electrons inappropriately. They can also react with oxygen to produce reactive oxygen species. A compromise is to use synthetic substrates such as ubiquinone or menaquinone with a decyl side-chain; these artificial mediators have reasonable solubility in water, and partition into the membrane in a similar way to the natural cofactors [27].

Alternative methods are needed to investigate the kinetic properties of enzyme assays with membrane-bound substrates. One way to measure the rates of enzymes with membrane-bound substrates is to couple the reaction to another enzyme, of which the product can accumulate and be more easily measured. An example would be the succinate:cytochrome c reductase activity of the mitochondrial membrane, where the reduction exogenous cytochrome c can be monitored spectrophotometrically. This still assumes however that the binding and dissociation of cytochrome c into the membrane is not rate-limiting.

Protein Film Voltammetry

A considerable number of oxidoreductases, containing redox centres such as heme, flavin, and/or iron–sulfur clusters, have been found to adhere, under suitable conditions, to a carbon electrode in such a way that they transfer electrons [28]. Membrane-associated oxidoreductases appear to work particularly well (Fig. 3). Direct electrochemistry of a film of these proteins provides information about kinetic parameters that is difficult to obtain by other means. When the substrate is present, the electric current is equivalent to the rate of substrate oxidation, *v*. The voltage dependence of the current *i* is equivalent to the dependence on concentration of an electron donor.

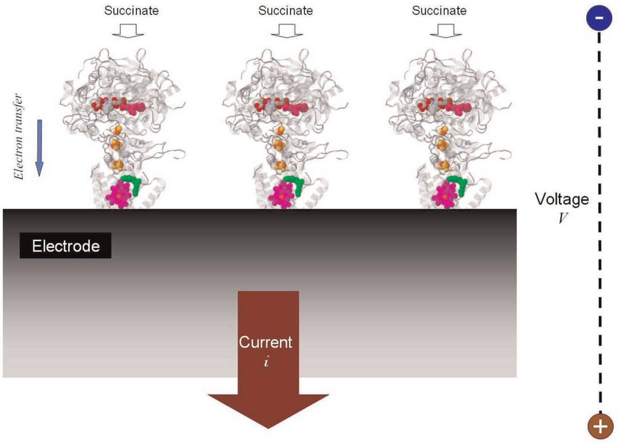


Figure 3. Diagram of succinate dehydrogenase on the surface of a carbon electrode. A monolayer of active molecules is adsorbed, so that succinate from solution can bind to the surface. The electrode is maintained at a voltage *V*, which can be swept, and the current *i* of electrons flowing through the enzyme molecules is measured. In the absence of succinate, the current observed is due to electrons flowing into the flavin, iron–sulfur clusters and heme.

The adsorption of an oxidoreductase onto a membrane, or onto a carbon electrode, transforms the kinetics from homogeneous catalysis to heterogeneous (at a two-dimensional surface). The rate of reaction depends not only on the concentration of substrate, but also on the rate of diffusion of substrate molecules to the surface. This can be studied by use of a rotating disk electrode; the rate of diffusion is proportional to the square root of the rate of rotation, ω [29]. When extrapolated to limiting value of ω, the current i is then proportional to the rate of catalysis by the enzyme. The Koutecky–Levich equation, which describes the quantitative relationship between i and ω, for an enzyme at a rotating electrode, takes a form analogous to the Michaelis–Menten equation, and provides values that are equivalent to K_m and V_{max} for substrate oxidation [29]. The method is very sensitive, needing only a monolayer of enzyme molecules over a surface of a few square millimeters.

Protein film voltammetry, in which the current i is measured as a function of the applied voltage V is swept, makes it possible to examine other features of the enzyme-catalysed reaction. As the applied voltage is swept, the current rises in a „catalytic wave“, usually at the midpoint potential of the substrate, for example the fumarate/succinate potential for succinate dehydrogenase. Cyclic voltammetry, in which the field is repeatedly swept up and down, shows that the reaction was nearly perfectly reversible [29, 30]. However some unusual kinetic properties of the enzymes emerged. Succinate dehydrogenase showed a „diode-like behaviour“ at higher driving potentials, the current decreased, a situation analogous to high substrate inhibition by the reducing agent [31]. However by judicious choice of the conditions of measurement it was possible to measure enzyme-catalysed rates much higher than those observed with artificial electron acceptors. The method can be used to measure the specificity of enzymes with different substrates and inhibitors, and study the effect of parameters such as pH.

In order to determine k_{cat} by this method, it is necessary to calculate the number of protein molecules on the surface that are giving rise to the catalytic current. This may be obtained by voltammetry of the enzyme in the absence of substrate, when catalytic waves can be measured from the redox centres in the protein itself. In the case of succinate dehydrogenase these are identified as flavin, iron–sulfur clusters and heme, for which the oxidation–reduction potentials can be measured.

Cyclic voltammograms of adsorbed enzyme layers containing membrane lipids offer a solution to the problem of determining the specificity of oxidoreductases for membrane-soluble quinone cosubstrates. Electrochemistry has been applied to thin films of ubiquinone [32]. A recent development is the construction of “tethered“ membranes on gold electrodes [33]. These bilayer membranes are connected, both physically and electrically, to the electrode, by a cholesterol tether which allows electron transfer. They can be loaded with protein complexes and quinones, and in favourable cases appear to behave kinetically like the native proteins.

Conclusions

The “99” enzymes represent an area of uncertainty in the description of enzymes. Ultimately they should be removed or transferred to other parts of the enzyme list. Meanwhile they indicate a fertile area for future studies. If the function of an enzyme is not clear, it may indicate interesting new biochemical processes.

For the purposes of metabolic reconstruction, the hydrophobic interiors of membrane bilayers represent separate, mobile compartments in the cell. Membrane-bound quinones such as ubiquinone-10 communicate through the membrane-bound protein complexes. Assays that assume a simple two-substrate, two-product reaction in dilute solution do not apply in such cases. New methods are needed for studying their activities and kinetics.

Membrane-bound oxidoreductases, which are not amenable to conventional solution enzyme assays, may be studied from their reactions at a carbon electrode surface. This makes it possible to examine their reactivity with different substrates, and the thermodynamic and kinetic properties of the redox centres within the enzymes.

Acknowledgements

I thank Fraser Armstrong, Peter Quinn, Peter Rich and Joel Weiner for helpful discussions.

Abbreviations

MK, Menaquinone
Q, Ubiquinone
STRENDA, Standards for Reporting Enzymology Data

References

[1] Kettner, C., Hicks, M.G. (2003) Chaos in the world of enzymes – how valid is functional characterization without methodological experimental data? In: *Experimental Standard Conditions of Enzyme Characterizations.* (Hicks, M.G., Kettner, C., Eds), pp. 1–16, Beilstein Institute, Frankfurt.

[2] Jacob, F., Monod, J. (1961) Genetic regulatory mechanisms in synthesis of proteins. *J. Molec. Biol.* **3:**318–356.

[3] Ockerman, P. A. (1968) Identity of beta-glucosidase b-xylosidase and one of beta-galactosidase activities in human liver when assayed with 4-methylumbelliferyl-beta-D-glycosides studies in cases of Gauchers disease. *Biochim. Biophys. Acta* **165:**59–62.

[4] Beutler, E., Grabowski, G.A. (2001) Gaucher disease. In: *The Metabolic and Molecular Bases of Inherited Diseases* (Scriver, C.R., Beaudet, A.L., Sly, W.S., Valle, D., Eds), Vol. 3, pp. 3635–3668, McGraw-Hill, New York.

[5] Scriver, C.R., Beaudet, A.L., Sly, W.S., Valle, D., Eds (2001) *The Metabolic and Molecular Bases of Inherited Diseases.* McGraw-Hill, New York.

[6] Mosmann, T. (1983) Rapid colorimetric assay for cellular growth and survival – application to proliferation and cyto-toxicity assays. *J. Immunol. Methods* **65:**55–63.

[7] Porstmann, T., Kiessig, S.T. (1992) Enzyme-immunoassay techniques – an overview. *J. Immunol. Methods* **150:**5–21.

[8] Hill, R., Bendall, F. (1960) Function of the 2 cytochrome components in chloroplasts – working hyothesis. *Nature* **186:**136–137.

[9] Keilin, J. (1966) *The History of Cell Respiration and Cytochrome.* Cambridge University Press, Cambridge.

[10] Clark, W.M. (1960) *The Oxidation–Reduction Potentials of Organic Systems.* The Williams and Wilkins Company.

[11] Prince, R.C., Linkletter, S.J.G., Dutton, P.L. (1981) The thermodynamic properties of some commonly used oxidation–reduction mediators, inhibitors and dyes as determined by polarography. *Biochim. Biophys. Acta* **351:**132–148.

[12] Dixon, M. (1971) Acceptor specificity of flavins and flavoproteins .3. Flavoproteins. *Biochim. Biophys. Acta* **226:**269–284.

[13] Lehmann, J. (1930) Information on biological oxidation reduction potential. Measurements in the system: Succinate-fumurate-succinatehydrogenase. *Skand. Archiv Physiol.* **58:**173–312.

[14] Dixon, M., Webb, E.C. (1958) *Enzymes.* London.

[15] IUB (1961) *Report of the Commission onEenzymes.* Pergamon, Oxford.

[16] Cecchini, G., Schroder, I., Gunsalus, R.P., Maklashina, E. (2002) Succinate dehydrogenase and fumarate reductase from *Escherichia coli*. *Biochim. Biophys. Acta-Bioenergetics* **1553:**140–157.

[17] Massey, V., Singer, T.P. (1957) Studies on succinic dehydrogenase .3. The fumaric reductase activity of succinic dehydrogenase. *J. biol. Chem.* **228:**263–274.

[18] Rich, P.R. (1984) Electron and proton transfers through quinones and cytochrome bc complexes (Biochim. Biophys. Acta 86108). *Biochim. Biophys. Acta* **768:**53-end.

[19] Tipton, K., Boyce, S. (2000) History of the enzyme nomenclature system. *Bioinformatics* **16:**34–40.

[20] Kotera, M., Okuno, Y., Hattori, M., Goto, S., Kanehisa, M. (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.* **126:**16487–16498.

[21] Tipton, K., Boyce, S., Mcdonald, A.G. (2004) Extending enzyme classification with metabolic kinetic data: some difficulties to be resolved. In: *Experimental Standard Conditions of Enzyme Characterizations.* (Hicks, M. G., Kettner, C., Eds), Beilstein Institute, Frankfurt.

[22] Moser, C.C., Dutton, P.L. (1987) The effect of ubiquinone tail length on Qb-activity in reconstituted photosynthetic reaction center proteoliposomes. *Biophys. J.* **51:**A124 –A124.

[23] Kröger, A., Klingenberg, M., Schweidl, S. (1973) Kinetics of redox reactions of ubiquinone related to electron-transport activity in respiratory chain. *Eur. J. Biochem.* **34:**358–368.

[24] Hauss, T., Dante, S., Haines, T.H., Dencher, N.A. (2005) Localization of coenzyme Q(10) in the center of a deuterated lipid membrane by neutron diffraction. *Biochim. Biophys. Acta-Bioenergetics* **1710:**57–62.

[25] Rich, P., Fisher, N. (1999) Generic features of quinone-binding sites. *Biochem. Soc. Trans* **27:**561–565.

[26] Rich, P.R., Harper, R. (1990) Partition-coefficients of quinones and hydroquinones and their relation to biochemical reactivity. *FEBS Letts* **269:**139–144.

[27] Rich, P.R., Madgwick, S.A., Moss, D.A. (1991) The interactions of Duroquinol, Dbmib and Nqno with the chloroplast cytochrome-Bf complex. *Biochim. Biophys. Acta* **1058:**312–328.

[28] Leger, C., Elliott, S.J., Hoke, K.R., Jeuken, L.J.C., Jones, A.K., Armstrong, F.A. (2003) Enzyme electrokinetics: Using protein film voltammetry to investigate redox enzymes and their mechanisms. *Biochemistry* **42:**8653–8662.

[29] Sucheta, A., Cammack, R., Weiner, J., Armstrong, F.A. (1993) Reversible electrochemistry of fumarate reductase immobilized on an electrode surface – direct voltammetric observations of redox centers and their participation in rapid catalytic electron-transport. *Biochemistry* **32:**5455–5465.

[30] Leger, C., Heffron, K., Pershad, H.R., Maklashina, E., Luna-Chavez, C., Cecchini, G., Ackrell, B.A.C., Armstrong, F.A. (2001) Enzyme electrokinetics: Energetics of succinate oxidation by fumarate reductase and succinate dehydrogenase. *Biochemistry* **40:**11234–11245.

[31] Sucheta, A., Ackrell, B.A.C., Cochran, B., Armstrong, F.A. (1992) Diode-like behaviour of a mitochondrial electron-transport enzyme. *Nature* **356:**361–362.

[32] Gordillo, G.J., Schiffrin, D.J. (2000) The electrochemistry of ubiquinone-10 in a phospholipid model membrane. *Faraday Discuss.* 89–107.

[33] Jeuken, L.J.C., Connell, S.D., Henderson, P.J.F., Gennis, R.B., Evans, S.D., Bushby, R.J. (2006) Redox enzymes in tethered membranes. *J. Am. Chem. Soc.* **128:**1711–1716.

The IUBMB Recommendations on Symbolism and Terminology in Enzyme Kinetics

Athel Cornish-Bowden

CNRS-BIP, 31 chemin Joseph-Aiguier, B.P. 71, 13402 Marseille Cedex 20, France

E-Mail: acornish@ibsm.cnrs-mrs.fr

Received: 7th June 2006 / Published: 31st August 2007

Abstract

Recommendations on the symbolism and terminology of enzyme kinetics were approved by the International Union of Biochemistry in 1981. They were primarily necessitated by the need for a systematic treatment of reactions of more than one substrate, but some important omissions have subsequently become evident, and a decision is needed as to whether these warrant the preparation of new recommendations, and if so whether these should constitute a complete revision of the entire document, or just the preparation of some new sections.

Introduction

The explosive growth in systems biology in the early years of the 21st century has brought with it a new interest in incorporating kinetic data enzymes into models of metabolism. Enzyme databases have greatly increased in importance, but their work has been severely impeded by the lack of standards for reporting kinetic data. However, the problem is not new: even 50 years ago the newly born International Union of Biochemistry was concerned that in the absence of any guiding authority the nomenclature of enzymology was getting out of hand, and it created the Commission on Enzymes as a remedy. The *Report of the Commission on Enzymes* [1], published in 1961, was mainly concerned with the naming of enzymes, but it also included brief recommendations on the symbols and terminology of enzyme kinetics. In a later reference to these, the 1973 edition of *Enzyme Nomenclature* [2] stated that “obviously, it would be of great advantage if all authors used the same system of symbols in their mathematical equations.” Is this so obvious, however? Is it even true? In this chapter I shall examine how the perceived needs of the subject led to the current recommendations on *Symbolism and Terminology in Enzyme Kinetics* [3], and I shall discuss how well these serve the needs of biochemistry 25 years later.

As long as biochemists were concerned mainly with single-substrate reactions there was little necessity for standardized symbols and terminology. If two different papers used the symbols k_{-1} and k_2 for the same rate constant, or if the same symbol k_2 was used for two different rate constants, only minor confusion was generated. However, the development in the 1950s of serious interest in reactions of two or more substrates introduced new difficulties, because numerous symbols were needed and translation from one system to another was neither obvious nor trivial: a pair of papers would use the same symbol for one quantity, different symbols for another, and the same symbol for two different quantities. Among many examples (see below), the K_{AB} of Bloomfield, Peller and Alberty [4] was the same as K_{AB} of Alberty [5], but their K_A was Alberty's $K_{AB}/_{KA.}$

The *Report of the Commission on Enzymes* [1], published by the International Union of Biochemistry in 1961, made tentative steps towards defining consistent symbols and terminology in enzyme kinetics, but the recommendations were omitted (without any indication of the reasons) from the 1979 edition of *Enzyme Nomenclature* [6]. The problems had not disappeared, however, and in 1978 – 1979 the views of numerous biochemists interested in kinetics were solicited. Following these consultations the International Union of Biochemistry set up a panel to prepare a complete set of recommendations on *Symbolism and Terminology in Enzyme Kinetics*, and these were approved in 1981 [3]. They tried, while taking account of the existing practices in biochemistry, to bring them into closer accord with the *Report on Symbolism and Terminology in Chemical Kinetics* that IUPAC had approved in 1981 [7]. IUB claimed in 1973 that their recommendations of 1961 had been “widely followed” [2], but this assessment was more wishful thinking than fact. Subsequently, the 1981 recommendations [3] have had some influence on biochemical practice but they have by no means been overwhelmingly adopted. Moreover, some important omissions, such as the lack of treatment of reversible reactions, have become especially important with the development of interest in computer modelling of metabolism, added to the importance that they already had for studies of biochemical thermodynamics.

The International Union of Biochemistry and Molecular Biology now needs to decide whether these omissions are sufficiently important to warrant the preparation of new recommendations, and if so whether these should constitute a complete revision of the entire document, or just the addition of some new sections.

Organizations Involved in Making Recommendations

The various bodies that have been involved in making recommendations on enzymes and enzyme kinetics have experienced as many changes in name and abbreviations as most topics in biochemistry itself, and so it may be helpful to list them. The *Commission on*

Enzymes of the International Union of Biochemistry, more often called the *Enzyme Commission*, was created in 1956 and made its report in 1961 [1]. It was then replaced by the IUB *Standing Committee on Enzymes*, which had responsibility for maintaining the nomenclature of enzymes until this was transferred to the IUB *Nomenclature Committee* when this was created in 1977.

Although the Enzyme Commission ceased to exist in 1961, its disappearance went unnoticed by most biochemists and references to it are still made today. Its name survives in the prefix EC used for enzyme numbers in *Enzyme Nomenclature* [8]. Despite the obvious advantages of EC numbers, their use in publications was patchy for many years, as by no means all of the major journals of biochemistry insisted on it. However, the greatly increased importance of computer databases in recent years has brought with it enhanced awareness of the need to identify enzymes unambiguously, and there is now much wider recognition that EC numbers provide the best chance currently available of achieving this. Nonetheless, the thoroughly objectionable practice of referring to enzymes simply as gene products, calling nitrate reductase the product of the *nar* genes, for example, remains common. It is hard to think of any legitimate reason to do this, not only implying that enzymes exist only to express what is recorded in the genome, but also utterly obscure to all but the small circles of researchers who work with the enzymes in question.

The IUBMB has always worked in conjunction with IUPAC in matters of biochemical nomenclature, and until 1977 most aspects of this were in the hands of the IUPAC-IUB Commission on Biochemical Nomenclature. This was reconstituted in 1977 as the IUPAC-IUB Joint Commission on Biochemical Nomenclature, the IUB Nomenclature Committee being created at the same time to deal with topics that IUPAC did not wish to handle (most notably enzyme nomenclature). In practice these two committees have always held joint meetings, with a common Chairman and Secretary. It may be noted that just as biochemists continue to refer to the Enzyme Commission as a living entity more than 40 years after it ceased to exist, they also frequently attribute to IUPAC recommendations that were actually made jointly by IUPAC and IUBMB, or even, like most of recommendations about enzymes, by IUBMB alone.

For about 20 years the International Union of Biochemistry also promoted a Committee of Editors of Biochemical Journals, which had responsibility for maintaining liaison with the nomenclature committees and ensuring that the recommendations made were consistent with current practice.

The names and abbreviations of the various organizations are listed in Table 1. As several of the names are cumbersome and unmemorable they are replaced in the remainder of this article by the abbreviations given in the right-hand column.

Constituton of the IUB Panel of 1981

The panel set up by IUB consisted of seven members, A. Cornish-Bowden, H. B. F. Dixon, K. J. Laidler, J. Ricard, I. H. Segel, S. F. Velick and E. C. Webb, and numerous other biochemists were also consulted. The convener was initially Segel, but he subsequently resigned and was replaced by Cornish-Bowden. Laidler had recently prepared a report for IUPAC on *Symbolism and Terminology in Chemical Kinetics* [7], and the first draft of the IUB recommendations [3] was in fact written by him.

Basic Definitions

The first part of the 1981 document [3] defined various terms of importance in enzyme kinetics, such as *catalysis*, *enzyme*, *substrate* etc. As these excite little controversy they will not be discussed here. One topic that did generate some disagreement, however, was the labelling of generic substrates, products and inhibitors. As long as there was only one of each the traditional use of S for substrate, P for product, and I for inhibitor created no difficulties, but these started to appear with studies of reactions with two or more substrates. Simply adding subscripts, as in S_1, S_2, etc., creates no logical difficulty, but it does add to the typographical complications of a subject already overburdened with subscripts, superscripts, primes etc., and most authors have preferred an alphabetical system with substrates A, B, etc., products P, Q, etc., and inhibitors I, J, etc. The 1981 recommendations used such a system for illustration, apart from using Z, Y etc. for products, as in the well known textbook of Laidler and Bunting [9], rather than P, Q, etc.; however, they emphasized that the essential point is not to try to impose a uniform system for use in all circumstances, but to expect authors to define the symbols they use and to use them consistently.

Table 1

Full name	Period	Abbreviation
International Union of Biochemistry	1955 – 1991	IUB
International Union of Biochemistry and Molecular Biology	1991-present	IUBMB
International Union of Pure and Applied Chemistry	1919-present	IUPAC
IUB Commission on Enzymes	1955 – 1961	EC
IUB Standing Committee on Enzymes	1961 – 1977	(none)
IUPAC-IUB Commission on Biochemical Nomenclature Nomenclature	Until 1977	CBN
IUPAC-IUB Joint Commission on Biochemical Nomenclature	1977 – 1991	JCBN
IUPAC-IUBMB Joint Commission on Biochemical Nomenclature	1991-present	JCBN
IUB Nomenclature Committee	1977 – 1995	NC-IUB
IUBMB Nomenclature Committee	1995-present	NC-IUBMB
IUB Committee of Editors of Biochemical Journals Journals	1955 – 1990	CEBJ

Order of Reaction and Rate Constants

The recommendations on order of reaction likewise produced little disagreement, but this section also dealt with the numbering of rate constants, a topic that had excited extensive discussion among biochemists; indeed, it accounted for about 40% of the total length of the chapter on Symbols of Enzyme Kinetics in the *Report of the Commission on Enzymes* [1]. The essential disagreement was between those who preferred the practice common in chemistry of referring to the forward and reverse rate constants for the first reaction in a sequence as k_1 and k_{-1}, and those who followed what had been long-standing practice in biochemistry of calling them k_1 and k_2 respectively. The Enzyme Commission preferred the former system, but felt that the existence of k_2 in both systems but with different meanings when applied to a simple two-step Michaelis-Menten mechanism was a source of ambiguity, and they proposed prefixing the positive subscripts with + signs, replacing, for example, k_1 by k_{+1}.

This matter had by no means been resolved to general satisfaction in 1981, but the Panel at that time felt that the emphasis in previous discussions had been misplaced. Rather than seeking to impose a universal system that could be used without definition, the essential was for authors to define whatever symbols were most appropriate for their purposes. Within the document itself the first of the systems mentioned was used for illustration, the + signs being treated as unnecessary.

Since 1981 the use of even-numbered indices for reverse reactions has not disappeared from the literature, but it seems to be in the process of doing so. Of 21st century textbooks, only one [10][1] still follows this system; all others known to me [11 – 14] use negative indices. Although the numbers involved are too small to be statistically significant, this is quite different from the case in 1981: at that time, only one [15] of the textbooks known to me followed what were then the recommendations and included + signs, five used negative indices but did not write + signs with positive indices [9, 16 – 19], and five avoided negative indices by using even-numbered indices for reverse steps [20 – 24].

Reactions Involving More Than One Substrate

The discussion of simple Michaelis-Menten kinetics requires no comment here, but matters became more complicated with the consideration of reactions of two or more substrates. In the earliest discussion of two-substrate kinetics known to me, Haldane [31] used numbered binding constants accompanied by the symbols x and y for the two concentrations:

$$\frac{V}{v} = 1 + \frac{K_4}{x} + \frac{K_3}{y} + \frac{K_1K_2}{xy} \tag{1}$$

[1] Although Leskovac [10] incorporated a substantial amount of material from the 2nd edition of my book [14], he did not number the rate constants in the same way.

The subsequent development of the subject by Alberty and others in the 1950 s led to wide variation in symbols: the K_{AB} of Alberty [5] was the same as that of Bloomfield *et al.* [4], but was written as $K_{ia}K_b$ by Cleland [25]; on the other hand, Alberty's K_A was the same as Cleland's K_a, but different from the K_A of Bloomfield *et al.*, despite the fact that Alberty was an author of two of these papers [4, 5]. Even with just three systems to compare there was ample scope for confusion, but in fact by the middle 1960s at least five or six different systems were in widespread use. Of these, the one introduced by Dalziel [26] was quite different from the others: less likely, therefore, to invite ambiguity, but also less easy to be understood by readers unfamiliar with it. It is now rarely used, but in 1981 it was still sufficiently frequent for the IUB Panel to think it worthwhile to include a note on the pronunciation of “Dalziel” (virtually identical to that of the prefix in DL-lactic acid).[2]

In this confusing environment Mahler and Cordes [27] noted the variation in symbols used by different authors in the 1950s and 1960s. As their emphasis was on the symbols used rather than on the way of organizing them into a rate expression, all of the rate equations were written in the same way, as expressions for the reciprocal rate, though they were not all written in this way in the original publications; the same convention is followed here. Alberty [5] wrote:

$$\frac{V_f}{v_f} = 1 + \frac{K_A}{A} + \frac{K_B}{B} + \frac{K_{AB}}{AB} \tag{2}$$

though in another paper [4] he used a different system:

$$\frac{V_{AB}}{v_f} = 1 + \frac{K_{AB}}{K_B}\frac{1}{A} + \frac{K_{AB}}{K_A}\frac{1}{B} + \frac{K_{AB}}{AB} \tag{3}$$

whereas Dalziel [26] wrote:

$$\frac{e}{v_f} = \phi_0 + \frac{\phi_1}{S_1} + \frac{\phi_2}{S_2} + \frac{\phi_{12}}{S_1 S_2} \tag{4}$$

and Cleland [25] wrote:

$$\frac{V_1}{v_1} = 1 + \frac{K_a}{A} + \frac{K_b}{B} + \frac{K_{ia}K_b}{AB} \tag{5}$$

[2] The need for this note was suggested by the fact that the Russian translator of my first book [32] used a footnote to mention that some Soviet authors wrote the name as Dal’tsil, implying that that was incorrect. The translator himself wrote the equivalent of Diil, which was also incorrect.

Despite their concern for this variation, Mahler and Cordes [27] used none of the systems then in use, but introduced a new one of their own:

$$\frac{V_1}{v_1} = 1 + \frac{K_a}{a} + \frac{K_b}{b} + \frac{\bar{K}_a K_b}{ab} \tag{6}$$

Their symbols have subsequently been adopted by essentially no other authors, in part because of the difficulty of printing symbols with overbars, but also because the use of overbars for distinguishing between Michaelis constant and inhibition constants is not obvious at sight but needs to be learned. Rather surprisingly, they did not list the symbols used in the textbook by Dixon and Webb [28, 29], though this was very widely used at the time they were writing:

$$\frac{ke}{v} = 1 + \frac{K'_a}{a} + \frac{K'_b}{b} + \frac{K_a K'_b}{ab} \tag{7}$$

They did, however, refer mysteriously to symbols used by the “Enzyme Commission”, symbols that occur nowhere in the *Report of the Commission on Enzymes* [1]. They are identical to those used *later* by Dixon and Webb [30], though not in the editions of their book [28, 29] that would have been available to Mahler and Cordes while they were writing:

$$\frac{V}{v} = 1 + \frac{K_m^A}{a} + \frac{K_m^B}{b} + \frac{K_s^A K_m^B}{ab} \tag{8}$$

One may surmise that they learned of these symbols from correspondence with Dixon, who had been, as noted previously, the Chairman of the original Commission on Enzymes. As may be deduced from the forms of the equations, Dalziel [26] designated the substrates as S_1 and S_2, but they were designated as A and B by all of the other authors mentioned, including Haldane [31], who, however, wrote their concentrations as *x* and *y* respectively.

Consistent with their attitude to other questions of uniformity, the members of the IUB Panel of 1981 considered that the essential point was not to try to impose a universal system, but to insist on the necessity to define whatever symbols authors choose to use.

For illustrative purposes they used symbols very similar to those of Dixon and Webb [30], but with the substrate indicated by a second subscript rather than by a superscript:

$$\frac{V}{v} = 1 + \frac{K_{\mathrm{mA}}}{a} + \frac{K_{\mathrm{mB}}}{b} + \frac{K_{\mathrm{iA}} K_{\mathrm{mB}}}{ab} \tag{9}$$

They also moved (silently) from italic to roman subscripts, replacing K_{mA} with K_{mA}, and so on. No reason was given for the change, but it agrees with present IUPAC recommendations [33]. It may be explained by the fact that *K* is here an algebraic variable, and should follow the normal mathematical convention of representing such variables by italic symbols. The subscripts, however, are not algebraic variables and should not be printed as if they were. In particular, m is not an index but the first letter of the name Michaelis, and A represents a chemical species, and should not therefore be written in italics either.

Most of the systems listed in these equations fail to distinguish between symbols for chemical species and symbols for their concentrations, even though these are logically distinct: the identity of a chemical species is not the same as its concentration. For example Alberty [5] used *A* both for the first substrate and for its concentration. Most of the early authors made no distinction, but for Dixon and Webb [28] *a* was the concentration of A, and so on. Some authors, such as Laidler and Bunting [9], preferred to use square brackets for concentrations, [A] for the concentration of A, for example.

The recommendations of IUB [3] considered the distinction important, and indicated that square brackets could be used without definition, but recognized that other systems might sometimes be typographically more convenient and were unobjectionable if defined in context. In the discussions within the Panel, some members thought that italic and roman type alone were sufficient to make the distinction (with *A* as the concentration of A), but the majority view was that differences between italic and roman type pass unnoticed by many readers and were thus inadequate to make an important conceptual distinction. *An extended piece of text in italics is, of course, quite obvious*, but an isolated letter *A* is much less obviously different from an isolated roman A. In any case, there are wide variations in what different people consider to be obvious, most simple truths being obvious once they have been pointed out.[3]

Inhibition

The treatment of enzyme in the 1981 document [3] is relatively brief, being mainly directed towards the classification of inhibition types as *reversible* or *irreversible*, as *linear* or *non-linear*, and as *competitive*, *uncompetitive*, *mixed* or *non-competitive*. In view of the great and growing importance of enzyme inhibition in drug development [37], a case could doubtless be made that a more extended treatment is now needed, and this is a question that NC-IUBMB should examine.

The names *competitive* and *uncompetitive* for the two extreme cases of linear inhibition (with effects on the apparent values of the specificity and catalytic constants respectively) are now widely accepted, and there was no support among the members of the Panel for the term *anticompetitive* used, for example, by Laidler and Bunting [9] in their textbook. The major disagreement that existed in 1981 and has still not been resolved is the name that

[3] The current scandal [34] over faked data in the *New England Journal of Medicine* illustrates this well. The same micrograph was used to illustrate results supposedly obtained with two different patients [35], and the similarity between the two panels of the relevant figure is so obvious that one might think it could hardly fail to be noticed even with an inexpert eye. Nonetheless, in reality it did pass unnoticed for several years, not only by the referees and editors, but also by readers of the article; it only came to light after revelation that fraudulent data from the same group had been published in another journal [36].

should be given to the range of intermediate cases in which there are effects on the apparent values of both the specificity and the catalytic constants, and the name, if any, to be given to the special case of this in which the effects on the two constants are equal. Although there was general agreement with Cleland's view that this special case had no particular mechanistic or other importance [38], and therefore had no need for a unique name, there was much less agreement with his view that the name *non-competitive* that had been given to this case for many years could therefore be generalized to encompass the whole intermediate range. The problem with this loosening of the definition is that the restricted meaning was still very widely used, and continues to be, and the shorter term *mixed* (or *mixed-type*) was already available for the general case. The Panel therefore preferred to follow the usage of Dixon and Webb [30], in which *non-competitive* refers to the special case, and *mixed* to the general case.

Nonetheless, the view that the usage of Dixon and Webb is unambiguous has not met with universal agreement. Copeland [37], for example, recently commented as follows: “In my experience, the term mixed-type inhibition can lead to misunderstandings about the physical meaning of the term (e. g., I have had discussions with chemists who have mistakenly believed that mixed-type inhibition must require two inhibitor molecules binding to separate sites on the enzyme); therefore we will use the term non-competitive inhibition in its broader definition to describe any inhibitor that displays affinity for both the free enzyme and the ES complex.” However, this argument appears unconvincing.[4]

Although there has long been agreement that linear inhibition is characterized by two different inhibition constants (for the competitive and uncompetitive components, either of which may be negligible), there has been less agreement about how they should be symbolized. When only one constant is relevant it is normally symbolized K_i, but when both are needed Dixon and Webb [30], for example, used K_i for the competitive inhibition constant and $K^{'}_i$ for the uncompetitive inhibition constant, whereas Cleland [38] used K_{is} and K_{ii} respectively (for $K_{i\,slope}$ and $K_{i\,intercept}$ respectively, referring to the slope and ordinate intercept of a plot of reciprocal rate against reciprocal substrate concentration).

In the 1981 recommendations [3] both of these conventions were considered unsatisfactory, the use of primes being unsystematic and the second subscripts *s* and *i* being derived from a particular type of plot with no necessary relationship to the subject. (With plots of substrate concentration divided by rate against substrate concentration, for example, Cleland's K_{is} refers to the ordinate intercept and K_{ii} to the slope, an inversion of roles that can hardly fail to be confusing.) For these reasons the symbols K_ic and K_iu were recommended for the competitive and uncompetitive components respectively. Although not yet in universal use, these have been widely adopted.

[4] I have encountered biochemists who feel strongly that MoCo is a clear and satisfactory abbreviation for “molybdenum cofactor”, but I doubt whether many chemists would find that a convincing reason to abandon Co as a symbol for cobalt.

Activation

Activation was also dealt with rather briefly in the 1981 recommendations, but two points of nomenclature needed to be addressed. First, it was noted that although classification of activation as linear or non-linear often has the same results as classifying it as *essential* or *non-essential*, exceptions are possible, because in principle essential activation (in which the enzyme has no activity in the absence of activator) can be non-linear (so that the reciprocal rate is not a linear function of the reciprocal concentration of activator).

A more important point was to emphasize that although the different kinds of linear activation are analogous to the familiar classes of inhibition, the name *competitive* cannot be used for the type of activation in which the activator binds only to the free enzyme because there is nothing that can be considered a competition in such a mechanism. Although less obviously objectionable, the terms *uncompetitive* and *non-competitive* were also recommended to be avoided for describing activation. Instead, the names *specific activation* and *catalytic activation* (corresponding to competitive and uncompetitive inhibition respectively) were suggested for effects on the apparent values of the specificity constant and catalytic constant respectively, *mixed activation* being entirely acceptable for the case where both effects are present.

Although the recommendations did not mention it – doubtless wanting to avoid the storm of protest that would have greeted any suggestion of abandoning the term *competitive* altogether – the terms *specific* and *catalytic* could perfectly well be applied to inhibition as well, resulting in an exact correspondence between the terms used in activation and inhibition. However, biochemists in general have been far more interested in inhibition than in activation, and would certainly resist any change to inhibition terminology that was introduced solely with the aim of greater concordance with activation terminology. Nonetheless, in contexts where both activation and inhibition need to be discussed together it is simplest to qualify both as *specific*, *catalytic* or *mixed* [see, e. g., 39].

pH Effects

The discussion of pH dependence in the recommendations of 1981 [3] introduced no new principles or terminology, and was in general based on what was already common practice in the literature. It requires no discussion here.

Pre-Steady-State Kinetics

The discussion of pre-steady-state kinetics in the recommendations of 1981 [3] was rather brief, in part because there was no particular need in enzyme kinetics to depart from normal practice in chemistry, and so the IUPAC recommendations [7] should cover most needs, and in part because the members of the panel were mainly people with experience of steady-state kinetics: preparation of recommendations for pre-steady-state kinetics would require a different panel.

One point that such a panel might wish to consider was brought to my attention by Gösta Pettersson during preparation of the current edition of my textbook [14]: equations for pre-steady-state kinetics typically contain terms of the form $A\exp(-\lambda t)$, where A is a constant known as the *amplitude*, t is the time and λ is a constant with dimensions of reciprocal time: it is the reciprocal of a time commonly symbolized as τ and called the *relaxation time* or the *time constant*, but λ has no generally accepted name of its own. Although it has the dimensions of a first-order rate constant, it is not in general the rate constant of any particular first-order reaction, so terms such as "apparent first-order rate constant" are not only cumbersome but also potentially misleading. Pettersson proposed the name *frequency constant* for λ. Authoritative texts [e. g. 40, 41] typically switch arbitrarily between writing equations in terms of λ and in terms of τ, and often write $1/\tau$ rather than λ. In a well known textbook [42] a table entitled "Physical meaning of the relaxation time, τ" actually tabulates not τ but $1/\tau$.

Reversible Reactions

As noted already, the 1981 recommendations [3] paid very little attention to the reversibility of enzyme-catalysed reactions. However, even in the simplest case of a one-substrate one-product reaction there are points to be taken into account, most obviously that the rate equation cannot be linearized by writing it as an expression for reciprocal rate and that therefore there is no advantage in taking reciprocals at all; some such form as:

$$v = \frac{V^{\mathrm{f}}a - \frac{V^{\mathrm{r}}K_{\mathrm{mA}}}{K_{\mathrm{mP}}}p}{K_{\mathrm{mA}}(1 + p/K_{\mathrm{mP}}) + a} \tag{10}$$

is as simple as one can obtain. The concentrations and Michaelis constants can be represented in the same way as in the irreversible case, but some additional convention is needed to distinguish between the forward and reverse limiting rates, and superscript f and r respectively are used in this example.

Nonetheless, representing the equation like this has some disadvantages, which become more important when one needs to consider more complicated examples, such as equations for reactions with multiple substrates and reactions that do not obey Michaelis-Menten kinetics. Equation 10 obscures at least two points: it fails to illustrate the symmetry of the behaviour with respect to substrate and product, and it fails to separate it into the components – catalytic activity of the enzyme, thermodynamic state of the reaction, degree of saturation of the enzyme – that characterize any enzyme-catalysed reaction. This separation becomes much clearer if we rearrange it into the following form, where K is the equilibrium constant:

$$v = \frac{\frac{V^{\mathrm{f}} a}{K_{\mathrm{mA}}}\left(1 - \frac{p/a}{K}\right)}{1 + \frac{a}{K_{\mathrm{mA}}} + \frac{p}{K_{\mathrm{mP}}}} \tag{11}$$

Here the right-hand factor in the numerator separates the thermodynamic state of the reaction from any properties that the enzyme may have. In particular, as the only term in the equation that can be negative it is the only term that decides the direction in which the reaction will proceed, but as it contains no kinetic information it says nothing about how fast it will do so. When the equation is written in this way the thermodynamic factor is fixed, regardless of mechanistic complexities, but the rest of the equation can be freely modified (as long as no negative quantities are introduced) without violating any thermodynamic constraints.

Non-Michaelis-Menten Kinetics

The section of the recommendations of 1981 [3] in most obvious need of revision is that dealing with reactions that do not obey Michaelis-Menten kinetics. This was partly because discussions of this topic are normally focused on mechanisms and models of cooperativity [e. g. 43 – 45], which were inappropriate topics for extensive discussion in a nomenclature document, and partly because the need for reasonably simple rate equations that could be used in metabolic models for fully reversible reactions [46] was not apparent at that time. For irreversible cases the Hill equation was already widely used as a simple alternative to mechanistically realistic equations that are too complicated to use, and the recommendations made several important points about it. As long as the thermodynamic factor in the reversible case is written as in Equation 11 the equation will remain thermodynamically correct; this important point has not always been realized in discussions of cooperative kinetics in the literature, and equations have sometimes appeared that suggest that non-thermodynamic factors may determine the direction of a reaction.

The Hill equation can be regarded as a variant of the Michaelis-Menten equation in which both the substrate concentration and the half-saturation concentration (not the Michaelis constant: see below) are raised to a power h known as the *Hill coefficient*. In the literature the Hill coefficient had often been written as n, a symbol that invited confusion with the number of binding sites for substrate on the enzyme, or as n_{H} by authors who were aware of the danger of confusion and wished to avoid it; the alternative h, which has become the recommended symbol, was already occasionally found in the literature though it was unusual. The symbol n was definitely discouraged, on account of the danger of confusion noted; n_{H} was not discouraged, but it was noted that it was typographically inconvenient to include a subscript in a symbol that represents an exponent and therefore sometimes needs to be printed as a superscript to a symbol that already has a subscript: $\mathrm{K}^{n\mathrm{H}}{}_{0.5}$, for example, is legible if carefully printed, but less legible than $\mathrm{K}^{h}{}_{0.5}$.

There is one other major point that was noted in the recommendations of 1981 [3], albeit in an unfortunate context (section 4.3 rather than the more appropriate section 11): the Michaelis constant K_m is by definition a parameter of the Michaelis-Menten equation, and has no meaning for non-Michaelis-Menten kinetics. Failure to appreciate this remains commonplace in the literature. To avoid the error one needs to replace any symbol like K_mA with a generic symbol like $a_{0.5}$ that suggests half-saturation without implying any particular kinetic equation.

Taking account of these considerations, a form of the reversible Hill equation that avoids violating thermodynamic constraints would be as follows (as suggested in [46]):

$$v = \frac{\frac{V^\mathrm{f} a}{a_{0.5}}(1 - \frac{p/a}{K})(\frac{a}{a_{0.5}} + \frac{p}{p_{0.5}})^{h-1}}{1 + (\frac{a}{a_{0.5}} + \frac{p}{p_{0.5}})^h} \qquad (12)$$

Notice that this simplifies to Equation 11 when $h = 1$.

Transport Processes, Insoluble Enzymes, etc.

There are several topics that are completely missing from the recommendations of 1981. Although it is widely recognized that the kinetics of transport processes have much in common with the kinetics of enzyme-catalysed reactions, and transporters are quite similar to enzymes, there appears to have no attempt to harmonize terminology in these closely related subjects. Indeed, at the time of writing the IUBMB have not approved any recommendations at all in the area of transport processes. Similarly, even if they do not state it explicitly the 1981 recommendations mainly assume that they are dealing with enzymes in free aqueous solution, and contain no mention, for example, of processes that take place at lipid-water interfaces. In the future the IUBMB will need to consider whether these topics should be dealt with separately, or incorporated into new recommendations about enzyme kinetics.

Acknowledgements

I thank M.L. Cárdenas and H.B.F. Dixon for helpful criticism of earlier drafts of this article.

References

[1] International Union of Biochemistry (1961) *Report of the Commission on Enzymes*. Pergamon Press, Oxford.

[2] International Union of Biochemistry (1973) *Enzyme Nomenclature (1972*). Elsevier, Amsterdam.

[3] Nomenclature Committee of the International Union of Biochemistry. (1982) Symbolism and terminology in enzyme kinetics. *Archs Biochem. Biophys*. **224:**732–740 (1983); *Biochem. J.* **213:**561–571 (1983); *Eur. J. Biochem.* **128:**281–291(1982); *Biochemical Nomenclature and Related Documents.* pp.96–106.(1992) Portland Press, London; electronic version available at: http://www.chem.qmul.ac.uk/iubmb/kinetics/

[4] Bloomfield, V., Peller, L., Alberty, R.A. (1962) Multiple intermediates in steady-state enzyme kinetics. II. Systems involving two reactants and two products. *J. Am. Chem. Soc*. **84:**4367–4374.

[5] Alberty, R.A. (1956) Enzyme kinetics. *Adv. Enzymol*. **17:**1–64.

[6] International Union of Biochemistry (1979) *Enzyme Nomenclature 1978*. Academic Press, New York.

[7] International Union of Pure and Applied Chemistry (1981) Symbolism and terminology in chemical kinetics. *Pure Appl. Chem* **53:**753–771.

[8] International Union of Biochemistry (1992) *Enzyme Nomenclature 1992*. Academic Press, New York

[9] Laidler, K.J., Bunting, P.S. (1973) *The Chemical Kinetics of Enzyme Action.* (2nd Edn) Clarendon Press, Oxford.

[10] Leskovac, V. (2003) *Comprehensive Enzyme Kinetics*. Kluwer Academic/Plenum Publishers, New York.

[11] Copeland, R.A. (2000) *Enzymes.* (2nd Edn) Wiley-VCH, New York.

[12] Bisswanger, H. (2002) *Enzyme Kinetics: Principles and Methods*. Wiley-VCH, Weinheim.

[13] Marangoni, A.G. (2003) *Enzyme Kinetics, a Modern Approach.* Wiley-Interscience, Hoboken, New Jersey.

[14] Cornish-Bowden, A. (2004) *Fundamentals of Enzyme Kinetics.* (3 rd Edn) Portland Press, London.

[15] Cornish-Bowden, A. (1979) *Fundamentals of Enzyme Kinetics* (1st Edn) Butterworths, London.

[16] Wong, J.T.-F. (1975) *Kinetics of Enzyme Mechanisms.* Academic Press, London.

[17] Segel, I.H. (1975) *Enzyme Kinetics*. Wiley, New York.

[18] Roberts, D.V. (1977) *Enzyme Kinetics.* Cambridge University Press, Cambridge.

[19] Fersht, A.R. (1977) *Enzyme Structure and Mechanism.* Freeman, Reading.

[20] Plowman, K. (1972) *Enzyme Kinetics.* McGraw-Hill, New York.

[21] Fromm, H.J. (1975) *Initial Rate Enzyme Kinetics.* Springer-Verlag, Berlin.

[22] Piszkiewicz, D. (1977) *Kinetics of Chemical and Enzyme-Catalysed Reactions.* Oxford University Press, New York.

[23] Ainsworth, S. (1977) *Steady-state Enzyme Kinetics*. Macmillan, London.

[24] Engel, P. (1977) *Enzyme Kinetics*. Chapman and Hall, London.

[25] Cleland, W.W. (1963) The kinetics of enzyme-catalyzed reactions with two or more substrates or products. I. Nomenclature and rate equations. *Biochim. Biophys. Acta* **67:**104–137.

[26] Dalziel, K. (1957) Initial steady state velocities in the evaluation of enzyme-coenzyme-substrate reaction mechanisms. *Acta Chem. Scand.* **11:**1706–1723.

[27] Mahler, H.R., Cordes, E.H. (1966) *Biological Chemistry*. Harper and Row, New York.

[28] Dixon, M., Webb, E.C. (1958) *Enzymes.* (1st Edn) Longmans, London.

[29] Dixon, M., Webb, E.C. (1964) *Enzymes*. (2nd Edn) Longmans, London.

[30] Dixon, M., Webb, E.C. (1979) *Enzymes* (3 rd Edn) Longmans, London.

[31] Haldane, J.B. S. (1930) *Enzymes*. Longmans, London.

[32] Cornish-Bowden, A. (1976) *Principles of Enzyme Kinetics.* Butterworths, London; Russian translation by B.I. Kurganov, Mir, Moscow.

[33] International Union of Pure and Applied Chemistry (1993) *Quantities, Units and Symbols in Physical Chemistry*. p. 5. Blackwell, Oxford.

[34] Curfman, G.D., Morrissey, S., Drazen, J.M. (2006) Expression of concern. *New Engl. J. Med.* **354:**638.

[35] Sudbø, J., Risberg, B, Koppang, H.S., Danielsen, H.E., Reith, A. (2001) DNA content as a prognostic marker in patients with oral leukoplakia. *New Engl. J. Med*. **344:**1270–1278.

[36] Sudbø, J., Lee, J.J., Lippman, S.M., Mork, J., Sagen, S., Flatner, N., Ristimaki, A., Sudbø, A., Mao, L., Zhou, X., Kildal, W., Evensen, J.F., Reith, A., Dannenberg, A.J. (2005) Non-steroidal anti-inflamatory drugs and the risk of oral cancer: a nested case-control study. *Lancet* **366:**1359–1366.

[37] Copeland, R.A. (2005) *Evaluation of Enzyme Inhibitors in Drug Discovery*. Wiley-Interscience, Hoboken, New Jersey.

[38] Cleland, W.W. (1963) The kinetics of enzyme-catalyzed reactions with two or more substrates or products. II. Inhibition: nomenclature and theory. *Biochim. Biophys. Acta* **67:**173–187.

[39] Cárdenas, M.L., Cornish-Bowden, A. (1989) Characteristics necessary for an inter-convertible enzyme cascade to give a highly sensitive response to an effector. *Biochem. J.* **257:**339–345.

[40] Hammes, G.G., Schimmel, P.R. (1970) In: *The Enzymes.* (Boyer, P.D. Ed.), pp. 67–114. Academic Press, New York.

[41] Gutfreund, H. (1995) *Kinetics for the Life Sciences*. Cambridge University Press, Cambridge.

[42] Hiromi, K. (1979) *Kinetics of Fast Enzyme Reactions*. Wiley, New York.

[43] Monod, J., Wyman, J., Changeux, J.-P. (1965) On the nature of allosteric transitions: a plausible model. *J. Molec. Biol.* **12:**88–118.

[44] Koshland, D.E., Jr., Némethy, G., Filmer, D. (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* **5:**365–385.

[45] Cornish-Bowden, A., Cárdenas, M.L. (1987) Co-operativity in monomeric enzymes. *J. Theoret. Biol.* **124:**1–23.

[46] Hofmeyr, J.-H.S., Cornish-Bowden, A. (1997) The reversible Hill equation: How to incorporate cooperative enzymes into metabolic models. *Comp. Appl. Biosci.* (*CA-BIOS*) **13:**377–385.

Discovering Novel Enzymes and Pathways by Comparative Genomics

Valérie de Crécy-Lagard

Department of Microbiology and Department of Microbiology and Cell Science,
University of Florida, P.O. Box 110700, Gainesville, FL 32611 – 0700, U.S.A.

E-Mail: vcrecy@ufl.edu

Received: 31st July 2006 / Published: 31st August 2007

Abstract

Identifying the function of every gene in all sequenced organisms is one of the major challenges of the post-genomic era and is one of the obligate steps leading to systems biology approaches. This objective is far from being reached. By different estimates, over 30 – 50 % of the genes of any given organism are of unknown function, incorrectly annotated or given a broad nonspecific annotation.

Most genome functional annotations programs rely on an homology based approach, using first simple Blast or FASTA scores then more elaborate, sensitive and precise algorithms stemming from the field of protein structure prediction. The inherent limitations of homology based approaches (only similar objects can be identified), has driven the development of non-homology based methods to link gene and function. Integrative genome mining tools that can analyse gene clustering, phylogenetic distribution, or protein fusions on a multi-genome scale have been developed recently. These bioinformatics tools allow the experimental biologist to make predictions on unknown gene function that can be tested experimentally and discover novel enzymes, regulators and transporters that expand our knowledge of metabolism in all species.

Introduction

The availability of nearly four hundred complete genomes (http://www.genomesonline.org/) has changed the way the experimental scientist generates hypothesis and identifies novel enzymes. The computer programming illiterate bench scientist has the unique possibility to link genes and function by combining comparative genomic tools that are freely available, with the experimental tools of physiology, genetics and enzymology. These approaches are leading to the discovery of novel enzymes and pathways of both fundamental and applied interest and also improve the general quality of genome annotations.

Towards a Complete Functional Analysis of Genomes: The Post-Genome Challenge

Identifying the function of every gene in all sequenced organisms is one of the major objectives of the post-genomic era, and one that is driving the development of systems biology [1]. This objective is far from reached as, by different estimates, 30 – 60% of the genes of any given organism have no assigned function [2 – 4]. As more genomes are being sequenced, the number of unknown genes and annotation errors are propagating at an alarming rate, making it increasingly difficult to extract correct functional information. Without a specific functional annotation effort, the genome information generated will become difficult to analyse and greatly underexploited [5].

Mining Genomes for New Enzymes

The availability of genomic sequence from both cultured and non-cultured organisms from diverse environments has had a great impact on the availability of enzymes that are better adapted for biocatalysis (for review see [6]). Also, "biochemical profiling" approaches [7 – 13] have been quite successful in identifying new enzymes [14, 15]. All these methods, however, rely on high throughput protein expression and enzymatic screens, and less labour intensive methods that are also more target specific are clearly needed to fully mine the catalytic potential of genomes. This is especially critical for implementing new biocatalytic activities into industrial processes. As Schmid *et al.* commented, "Future biocatalytic processes generally will not be limited by the available technology or the nature of the substrates and products. Instead, the feasibility of new biocatalytic processes will often be determined by the availability of the biocatalyst..." [16]. An untapped resource of novel catalysts is lying in the thousand of genes of unknown function that are now available at our fingertips if both bioinformatic and experimental methods can be combined to identify them.

Mining Genomes for New Antibacterial Targets

The need for the development of new antibiotics that escape common resistance mechanisms is becoming an acute public health problem. The World Health Organization (WHO) states that "In the race for supremacy, microbes are sprinting ahead" and "Microbial resistance could bring the world to a pre-antibiotic age (http://www.who.int/infectious-disease-report/2000/). The value of using genomics in anti-infective research was recognized early on by both fundamental and applied research enterprises (for review see [17]). Pipelines combining identification of bacterial genes essential for growth or virulence followed by structural efforts have been implemented and leads are starting to trickle out [18]. One major problem in this approach has been that targets identified from genomics approaches are often of unknown function, therefore no assay can be developed to screen for inhibitors.

Homology-Based Functional Analysis

Functional inferences based on comparative sequence analysis are well-established foundations of genomic annotation. The most significant advancements in this field over the last decade are directly related to the dramatic increase in the amount of sequenced genomes, as well as to the development of the robust and sensitive algorithms, such as FASTA, BLAST and their modifications (for the overview, see [19]). Domain analysis and reduction of the protein space via grouping of putative orthologues (such as COGs [20]) play an important role in the projection of functional assignments between diverse species. A significant contribution is provided by research communities focused on the detailed curation efforts of model organism genomes (e.g., *Escherichia coli* (http://bmb.med.miami.edu/EcoGene/EcoWeb), *Bacillus subtilis* (http://genome.jouy.inra.fr/cgi), *Saccharomyces cerevisiae* (http://www.yeast genome.org/index.html). For well studied gene families, in which the initial annotation has been experimentally verified, these homology based methods are quite accurate in predicting function [21]. However, factors such as poor homologies [21], multi-domain proteins [22], gene duplications [21, 23] and non-orthologous displacements [24] all contribute to incorrect or absent annotations that have accumulated and propagated, leading to the current poor functional annotation status of the genomic data [3, 4, 24, 25]. Furthermore, the inherent limitations of homology based approaches (only similar objects can be identified) require the development of non-homology based methods to link genes to function.

Beyond Homology, from Comparative Genomics to Experimental Verification

From gene to function

Systematic approaches such as structural genomics initiatives, systematic interaction mapping or systematic gene disruption combined with phenotypic screenings has led to the elucidation of some gene functions [26, 27]. Nearly 1000 structures have been deposited to date by structural genomics programs in the Protein Data Bank (http://www.rcsb.org/pdb/). However, examples where cellular functions were inferred directly from structural information are rare – in fact there are only a handful [28, 29]. Large-scale deletion mutant libraries have been completed for *S. cerevisiae* [30], *B. subtilis* [31] and in *E. coli* (http://ecoli.aist-nara.ac.jp/). Broad systematic phenotype screens [32] allowed the prediction of a few functions such as a missing histidine biosynthesis gene [33] or the discovery of anabolic and catabolic phosphorylating glyceraldehyde-3-phosphate dehydrogenises [34]. The real power of these libraries lies in using them in specialized screenings: this strategy has been successfully used in *S. cerevisiae* where novel cell cycle genes have been identified [35]. The road from phenotype to cellular function is often long and requires many downstream characterization steps [35]. Recently, "biochemical profiling" approaches consisting of testing the activity of all the proteins of a given genome (in pools or individually) in specific biochemical assays [7 – 13] or testing hundreds of proteins of unknown function in arrays covering a wide range of enzyme activities have been quite successful in correcting annotations or identifying functions of unknown genes [15, 36]. These large-scale efforts have not been as predictive as anticipated, but have been extremely valuable for the community in producing expression clones, mutants, structural and experimental data that can be used to predict and confirm functions as shown below.

From function to gene; Integration of genomic information

Large-scale cross-genomic integrations (such as NCBI [37], EMBL [38], TIGR [39], Uniprot [40]) provide important environments for extracting information from genomes. A dramatic enhancement of the quality and utility of genomic annotations is achieved by combining genome integration with metabolic reconstruction technology (see below). Among the key public resources supporting this approach are KEGG [41] and MetaCyc [42]. In these methods, genes are not analysed individually or as gene families but in a larger multi-genomic context. Additional information, not related to sequence homology, is gathered to help link gene and function (Fig. 1), and include:

- **Metabolic reconstruction:** by placing the genes in the context of the metabolic pathways found in a given organism, one can evaluate the biological relevance of an annotation [43, 44].
- **Clustering data:** Genes of a given pathway have a high probability of being physically linked on the chromosome [45].
- **Protein fusion events:** Genes of the same pathway can be fused to encode multi-domain proteins in some organisms [46].

- **Phylogenetic occurrence profiles or signatures:** Presence/absence patterns of genes (or of set of genes) among genomes can be used to identify candidates for missing genes [47].
- **Shared regulatory sites:** Pathway genes are often regulated by a common protein recognizing a specific DNA sequence [48].
- **Functional and structural genomics:** Efforts provide additional clues to genome interpretation. The rapid increase in the volume and quality of such data, as well as their integration in publicly available repositories is expected to strongly impact gene and pathway analysis. Among the growing number of web-resources are: PDB, the best established collection of protein structures (http://www.rcsb.org/pdb/), SMD for expression data (http://genome-www5.stanford.edu/), DIP for protein–protein interactions (http://dip.doe-mbi.ucla.edu/).

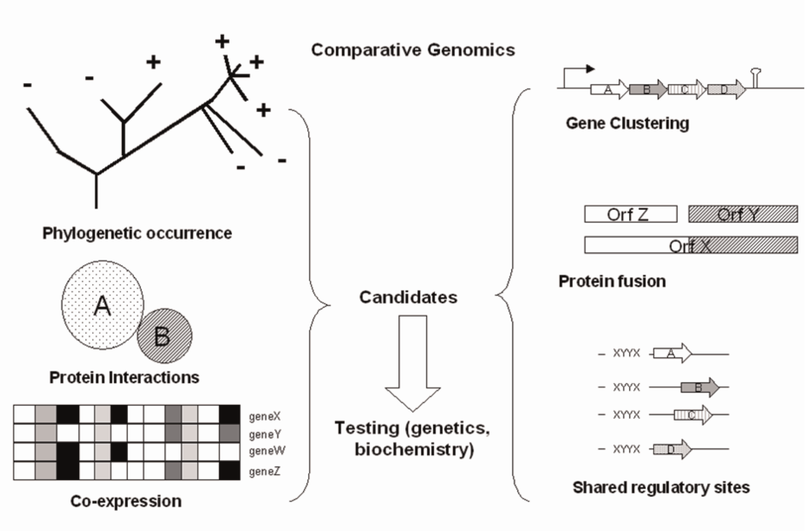


Figure 1. Comparative genomic strategies used to make predictions on gene function.

Early efforts to integrate different types of data to annotate genomes were developed by Koonin and colleagues based on the Cluster of Orthologs Groups (COG) database [49] that lists families of orthologues found in a subset of the sequenced genomes. In the last five years several integrated databases that contain phylogenetic occurrence profiles, clustering or protein fusion data and many combinations of the three have been implemented. These databases are all freely available with web-interfaces and include PhydBac [50], String [51], Microbial Database [52], Genomenet [[41], Plex [53], Cytoscape [54], Metacyc [42] and SEED [55] (Table 1). Genome researchers have built on this multi-tiered approach to help in calling gene function and reducing the number of errors as recently described for the genomes of *Haloarcula marismortui* [56] and *Methylococcus capsulatus* (Bath) [28]. The combination of structural information and comparative genomic methods has led to many robust predictions [1,29].

Table 1. Freely available comparative genomic analysis websites.

Name	Location
Cluster of Orthologous Groups	http://www.ncbi.nlm.nih.gov/COG/
FusionDB and PhydBac	http://igs-server.cnrs-mrs.fr/phydbac/
	http://igs-server.cnrs-mrs.fr/FusionDB/
TIGR-CMR	http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi
	http://www.tigr.org/tigr-scripts/CMR2/GenomeSlicer.spl
STRING	http://dag.embl-heidelberg.de/newstring_cgi/show_input_page.pl
IMG	http://img.jgi.doe.gov/cgi-bin/pub/main.cgi
Cytoscape	http://www.cytoscape.org/
GenomeNet and KEGG	http://www.genome.ad.jp/
Protein Link Explorer (Plex)	http://apropos.icmb.utexas.edu/plex/plex.html
MetaCyc	http://metacyc.org/
SEED	http://theseed.uchicago.edu/FIG/

Using Comparative Genomics to Link Genes to Function

Although the field of comparative genomics is still young, these tools have allowed the genetic characterization of a number of critical metabolic pathways that had eluded scientific inquiry for decades and an estimated 100 gene families have been identified successfully using comparative genomic methods to date [57] (Ross Overbeek, personal communication). For example, predictions based exclusively on phylogenetic occurrence profiles resulted in the identification of the last steps of the non-mevalonate isoprenoid pathway [58]. Protein fusion analysis allowed the identification of missing Coenzyme A biosynthesis genes in *Homo sapiens* [59]. Chromosome clustering analysis revealed a missing fatty acid synthesis gene (target of antibacterial compounds) in *Streptococcus pneumoniae* or missing genes in folate biosynthesis [60, 61]. A combination of approaches was used to identify the diverse NAD recycling pathways of cyanobacteria [62]. A search for regulator sites allowed the identification of many missing thiamine biosynthesis genes [63], metal transporters [64] or decipher the N-acetylglucosamine utilization pathway of *Shewanella oneidensis* [65]. The approach has been particularly productive in discovering missing and novel enzymes in Archaea because of the originality of their metabolic solutions and the recent availability of 30 archaeal genomes [66].

Applying these comparative genomic methods to the field of tRNA modification and coenzyme metabolism has allowed us to identify the function of eight enzyme families, unravelling novel enzyme activities, cases of orthologous displacements, novel pathways and potential drug targets (Table 2).

Table 2. Novel genes and pathways identified using comparative genomics techniques.

Functional Role	Novelty	Verified in	Key evidence
Pantetheine-phosphate adenylyltransferase/ dephospho–CoA kinase [67]		B [67]	Fusion
Carbamoyl-threonnyl-Adenosine syntase [a]	Potential target	E [a]	Occurrence profile/Structure
Wyeosine synthase [68]	Novel enzyme	E [68]	Occurrence profile
tRNA dihydrouridine synthase [69]	Novel enzyme	B [70]	Occurrence profile/operon
Queosine/Archeosine biosynthesis YkvJ	Novel enzyme	B [71]	Occurrence profile/operon
Queosine/Archeosine biosynthesis YkvK	Novel Enzyme	B [71]	Occurrence profile/operon
Queosine/Archeosine biosynthesis YkvL	Novel enzyme	B [71]	Occurrence profile/operon
PreQ0 reductase YkvM	Novel enzyme	B [72]	Occurrence profile/operon
GTP Cyclohydrolase I	Potential Target	B [73]	Occurrence profile/operon

B = Bacteria, E = Eukaryotes . [a] de Crécy-Lagard and collaborators (unpublished results)

Conclusion

This body of work opens the problem of how to name enzymes discovered through comparative genomics methods and give them EC numbers, as in general these enzymes have been very poorly described or were totally unknown. The number of enzymes discovered by these methods is steadily increasing and guidelines for the "gene discoverers" who are often not enzymologists need to be defined.

References

[1] Bonneau, R., Baliga, N.S., Deutsch, E.W., Shannon, P.,Hood, L. (2004) Comprehensive de novo structure prediction in a systems-biology context for the archaea *Halobacterium* sp. NRC-1. *Genome Biol.* **5:**R52.

[2] Siew, N., Azaria, Y., Fischer, D. (2004) The ORFanage: an ORFan database *Nucleic Acids Res.* **32 Database issue:** D281–283.

[3] Brenner, S.E. (1999) Errors in genome annotation. *Trends Genet.* **15:**132–133.

[4] Devos**, D.,** Valencia**, A.** (2001) Intrinsic errors in genome annotation. *Trends Genet.* **17:**429–431.

[5] Roberts, R.J. (2004) Identifying protein function–a call for community action. *PLoS Biol* **2:**E42.

[6] Ferrer, M., Martinez-Abarca, F., Golyshin, P.N. (2005) Mining genomes and 'metagenomes' for novel catalysts. *Curr. Opin. Biotechnol.* **16:**588–593.

[7] Grayhack, E.J., Phizicky, E.M. (2001) Genomic analysis of biochemical function. *Curr. Opin. Chem. Biol.* **5:**34–39.

[8] Gu, W., Jackman, J.E., Lohan, A.J., Gray, M.W., Phizicky, E.M. (2003) tRNA^{His} maturation: an essential yeast protein catalyzes addition of a guanine nucleotide to the 5' end of tRNA^{His}. *Genes Devel.* **17:**2889–2901.

[9] Jackman, J.E., Montange, R.K., Malik, H.S., PPhizicky, E.M. (2003) Identification of the yeast gene encoding the tRNA m^{1}G methyltransferase responsible for modification at position 9. *RNA* **9:**574–585.

[10] Phizicky, E.M., Martzen, M.R., McCraith, S.M., Spinelli, S.L., Xing, F., Shull,N.P., Van Slyke, C., Montagne, R.K., Torres, F.M., Fields, S., Grayhack, E.J. (2002) Biochemical genomics approach to map activities to genes. *Methods Enzymol* **350:**546–559.

[11] Polevoda, B., Martzen, M.R., Das, B., Phizicky, E.M., Sherman, F. (2000) Cytochrome c methyltransferase, Ctm1 p, of yeast. *J. Biol. Chem.* **275:**20508–20513.

[12] Steiger, M.A., Kierzek, R., Turner, D.H., Phizicky, E.M. (2001) Substrate recognition by a yeast 2'-phosphotransferase involved in tRNA splicing and by its *Escherichia coli* homolog. *Biochemistry* **40:**14098–14105.

[13] Xing, F., Martzen, M.R., Phizicky, E.M. (2002) A conserved family of *Saccharomyces cerevisiae* synthases effects dihydrouridine modification of tRNA. *RNA* **8:**370–381.

[14] Kutznetosva, E., Proudfoot, M., Sanders, S.A., Reinking, J., Savchenko, A. *et al.* (2005) Enzyme genomics: Application of general enzymatic screens to discover new enzymes. *FEMS Microbiol. Rev.* **29**(2)**:** 263–279.

[15] Martzen, M.R., McCraith, S.M., Spinelli, S.L., Torres, F.M., Fields, S., Grayhack, E.J., Phizicky, E.M. (1999) A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286:**1153–1155.

[16] Schmid, A., Dordick, J.S., Hauer, B., Kiener, A., Wubbolts, M., Witholt, B. (2001) Industrial biocatalysis today and tomorrow. *Nature* **409:**258–268.

[17] Haney, S.A., Alksne, L.E., Dunman, P.M., Murphy, E., Projan, S.J. (2002) Genomics in anti-infective drug discovery–getting to endgame. *Curr. Pharm. Des.* **8:**1099–1118.

[18] Schmid, M.B. (2004) Seeing is believing: the impact of structural genomics on antimicrobial drug discovery. *Na ure Rev. Microbiol.* **2:**739–746.

[19] Koonin, E.V., Galperin, M.Y. (2002) *SEQUENCE – EVOLUTION – FUNCTION. Computational Approaches in Comparative Genomics*. 488 pp. Kluwer Academic Publishers, Boston.

[20] Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29:**22–28.

[21] Tian, W., Skolnick, J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* **333:**863–882.

[22] Hegyi, H., Gerstein, M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.* **11:**1632–1640.

[23] Gerlt, J.A., Babbitt, P.C. (2000) Can sequence determine function? *Genome Biol.* **1:**REVIEWS 0005.

[24] Galperin, M.Y., Koonin, E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol.* **1:**55–67.

[25] Attwood, T.K. (2000) Genomics. The Babel of bioinformatics. *Science* **290:**471–473.

[26] Mittl, P.R., Grutter, M.G. (2001) Structural genomics: opportunities and challenges. *Curr. Opin. Chem. Biol.* **5:**402–408.

[27] Huynen, M.A., Snel, B., van Noort, V. (2004) Comparative genomics for reliable protein-function prediction from genomic data. *Trends Genet.* **20:**340–344.

[28] Yakunin, A.F., Yee, A.A., Savchenko, A., Edwards, A.M., Arrowsmith, C.H. (2004) Structural proteomics: a tool for genome annotation. *Curr. Opin. Chem. Biol.* **8:**42–48.

[29] Zhang, C., Kim, S.H. (2003) Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.* **7:**28–32.

[30] Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., *et al*. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285:**901–906.

[31] Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., Boland, F., Brignell, S.C., Bron, S., Bunai, K., Chapuis, J., *et al*. (2003) Essential *Bacillus subtilis* genes. *Proc. Natl Acad. Sci. U S A* **100:**4678–4683.

[32] Ogasawara, N. (2000) Systematic function analysis of *Bacillus subtilis* genes. *Res. Microbiol.* **151:**129–134.

[33] le Coq, D., Fillinger, S., Aymerich, S. (1999) Histidinol phosphate phosphatase, catalyzing the penultimate step of the histidine biosynthesis pathway, is encoded by *ytvP* (*hisJ*) in *Bacillus subtilis*. *J. Bacteriol.* **181:**3277–3280.

[34] Fillinger, S., Boschi-Muller, S., Azza, S., Dervyn, E., Branlant, G., Aymerich, S. (2000) Two glyceraldehyde-3-phosphate dehydrogenases with opposite physiological roles in a nonphotosynthetic bacterium. *J. Biol. Chem.* **275:**14031–14037.

[35] Carpenter, A.E., Sabatini, D.M. (2004) Systematic genome-wide screens of gene function. *Nature Rev. Genet.* **5:**11–22.

[36] Kuznetsova, E., Proudfoot, M., Sanders, S.A., Reinking, J., Savchenko, A., Arrowsmith, C.H., Edwards, A.M., Yakunin, A.F. (2005) Enzyme genomics: Application of general enzymatic screens to discover new enzymes. *FEMS Microbiol. Rev.* **29:**263–279.

[37] Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Helmberg, W., Kapustin, Y., Kenton, D.L., Khovayko, O., *et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **34:**D 173–180.

[38] Cochrane, G., Aldebert, P., Althorpe, N., Andersson, M., Baker, W., Baldwin, A., Bates, K., Bhattacharyya, S., Browne, P., van den Broek, A., Castro, M., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., *et al.* (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res.* **34:**D 10–15.

[39] Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., Quackenbush, J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19:**651–652.

[40] Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N., *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34:**D 187–191.

[41] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34:**D 354–357.

[42] Krieger, C.J., Zhang, P., Mueller, L.A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S.Y., Karp, P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **32 Database issue:** D 438–442.

[43] Selkov, E., Maltsev, N., Olsen, G.J., Overbeek, R., Whitman, W.B. (1997) A reconstruction of the metabolism of Methanococcus jannaschii from sequence data. *Gene* **197:**GC 11–26.

[44] Galperin, M.Y., Brenner, S.E. (1998) Using metabolic pathway databases for functional annotation. *Trends Genet.* **14:**332–333.

[45] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. U S A* **96:**2896–2901.

[46] Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402:**86–90.

[47] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. U S A* **96:**4285–4288.

[48] Gelfand, M.S., Novichkov, P.S., Novichkova, E.S., Mironov, A.A. (2000) Comparative analysis of regulatory patterns in bacterial genomes. *Brief Bioinform.* **1:**357–371.

[49] Natale, D.A., Galperin, M.Y., Tatusov, R.L., Koonin, E.V. (2000) Using the COG database to improve gene recognition in complete genomes. *Genetica* **108:**9–17.

[50] Enault, F., Suhre, K., Poirot, O., Abergel, C., Claverie, J.M. (2004) Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Res.* **32:**W336–339.

[51] von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33:**D433–437.

[52] Haft, D.H., Selengut, J.D., Brinkac, L.M., Zafar, N., White, O. (2005) Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics* **21:**293–306.

[53] Date, S.V., Marcotte, E.M. (2005) Protein function prediction using the Protein Link EXplorer (PLEX). *Bioinformatics* **21:**2558–2559.

[54] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13:**2498–2504.

[55] Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., *et al*. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33:**5691–5702.

[56] Baliga, N.S., Bonneau, R., Facciotti, M.T., Pan, M., Glusman, G., Deutsch, E.W., Shannon, P., Chiu, Y., Weng, R.S., Gan, R.R., Hung, P., Date, S.V., Marcotte, E., Hood, L., Ng, W.V. (2004) Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea. *Genome Res.* **14:**2221–2234.

[57] Osterman, A., Overbeek, R. (2003) Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.* **7:**238–251.

[58] Smit, A., Mushegian, A. (2000) Biosynthesis of isoprenoids via mevalonate in Archaea: the lost pathway. *Genome Res.* **10:**1468–1484.

[59] Daugherty, M., Polanuyer, B., Farrell, M., Scholle, M., Lykidis, A., de Crécy-Lagard, V., Osterman, A. (2002) Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics. *J. Biol. Chem.* **277:**21431–21439.

[60] Klaus, S.M., Kunji, E.R., Bozzo, G.G., Noiriel, A., de la Garza, R.D., Basset, G.J., Ravanel, S., Rebeille, F., Gregory, J.F., 3rd, Hanson, A.D. (2005) Higher plant plastids and cyanobacteria have folate carriers related to those of trypanosomatids. *J. Biol. Chem.* **280:**38457–38463.

[61] Klaus, S.M., Wegkamp, A., Sybesma, W., Hugenholtz, J., Gregory, J.F., 3rd, Hanson, A.D. (2005) A nudix enzyme removes pyrophosphate from dihydroneopterin triphosphate in the folate synthesis pathway of bacteria and plants. *J. Biol. Chem.* **280:**5274–5280.

[62] Gerdes, S.Y., Kurnasov, O.V., Shatalin, K., Polanuyer, B., Sloutsky, R., Vonstein, V., Overbeek, R., Osterman, A.L. (2006) Comparative genomics of NAD biosynthesis in Cyanobacteria. *J. Bacteriol.* **188:**3012–3023.

[63] Rodionov, D.A., Vitreschak, A.G., Mironov, A.A., Gelfand, M.S. (2002) Comparative genomics of thiamin biosynthesis in procaryotes: new genes and regulatory mechanisms. *J. Biol. Chem.***277:**48949–48959.

[64] Rodionov, D.A., Hebbeln, P., Gelfand, M.S., Eitinger, T. (2006) Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: evidence for a novel group of ATP-binding cassette transporters. *J. Bacteriol.* **188:**317–327.

[65] Yang, C., Rodionov, D.A., Li, X., Laikova, O.N., Gelfand, M.S., Zagnitko, O.P., Romine, M.F., Obraztsova, A.Y., Nealson, K.H., Osterman, A.L. (2006) Comparative genomics and experimental characterization of N-acetylglucosamine utilization pathway of *Shewanella oneidensis*. *J. Biol. Chem.* M605052200.

[66] Ettema, T.J., de Vos, W.M., van der Oost, J. (2005) Discovering novel biology by in silico archaeology. *Nature Rev. Microbiol.* **3:**859–869.

[67] Daugherty, M., Polanuyer, B., Farrell, M., Scholle, M., Lykidis, A., de Crecy-Lagard, V., Osterman, A. (2002) Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics. *J. Biol. Chem.* **277:**21431–21439.

[68] Waas, W.F., de Crécy-Lagard, V. Schimmel, P. (2005) Discovery of a gene family critical to wyosine base formation in a subset of phenylalanine-specific transfer RNAs. *J. Biol. Chem.* **280:**37616–37622.

[69] Bishop, A.C., Xu, J., Johnson, R.C., Schimmel, P., de Crécy-Lagard, V. (2002) Identification of the tRNA-dihydrouridine synthase family. *J. Biol. Chem.* **277:**25090–25095.

[70] Bishop, A.C., Xu, J., Johnson, R.C., Schimmel, P., de Crécy-Lagard, V. (2002) Identification of the tRNA-dihydrouridine synthase family. *J. Biol. Chem.* **277:**25090–25095.

[71] Reader, J.S., Metzgar, D., Schimmel, P., de Crécy-Lagard, V. (2004) Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine. *J. Biol. Chem.* **279:**6280–6285.

[72] Van Lanen, S.G., Reader, J.S., Swairjo, M.A., de Crécy-Lagard, V., Lee, B., Iwata-Reuyl, D. (2005) From cyclohydrolase to oxidoreductase: discovery of nitrile reductase activity in a common fold. *Proc. Natl Acad. Sci. U S A* **102:**4264–4269.

[73] El Yacoubi, B., Bonnett, S., Anderson, J.N., Swairjo, M.A., Iwata-Reuyl, D., de .Crécy Lagard, V. (2006) Discovery of a new prokaryotic type I GTP cyclohydrolase family. *J. Biol. Chem.* **281**(49)**:** 37586–37593.

Molecular Simulations of Enzyme Catalysis

Martin J. Field

Modeling and Simulation Group, Institut de Biologie Structurale – Jean-Pierre Ebel,UMR 5075 CEA/CNRS/UJF,
41, Rue Jules Horowitz, 38027 Grenoble Cedex 1, France

E-Mail: Mjfield@Ibs.Fr

Received: 27th June 2006 / Published: 31st August 2007

Abstract

Molecular modelling and simulation techniques have proved **powerful** tools for helping to understand how proteins and other biomacromolecules function at an atomic level. The study of enzyme reactions is a particularly challenging application of these methods because of the variety of processes of differing length and time scales that can contribute to catalysis. Among these are the bond-breaking and forming chemical steps, the diffusion of ligands into and out of the active site and conformational changes in the enzyme's structure.

This contribution gives a synopsis of the range of molecular simulation techniques that are available for studying enzyme reactions with particular emphasis on methods designed for the investigation of the chemical catalytic steps. The capabilities and limitations of current approaches will be described and possible future developments discussed. Special attention is given to the interface between molecular simulation and systems biology modelling and to how the STRENDA guidelines would need to be adapted to allow the reporting of enzyme data determined from simulation.

Introduction

Numerical modelling and simulation are important tools for the study of biological systems and will undoubtedly become more so as the power and the sophistication of computers and their algorithms increases [1]. Enzyme reactions represent a particularly challenging area for simulation because of the wide range of length and time scales upon which processes that are important to catalysis occur.

The aim of this article is to provide an overview of the state-of-the-art in the simulation of enzyme reactions at an atomic level. It starts with a brief summary of the types of method that exist for simulating different aspects of an enzyme reaction and is followed by a more detailed presentation of the principles behind and an application of one particular approach – the hybrid potential method. The next sections highlight some of the advantages and limitations of current simulation techniques and the article terminates with a discussion of how atomic-level simulation could contribute to systems biology modeling and with various recommendations for how the STRENDA guidelines would need to be modified to report data derived from simulation [2].

Methods for Simulating Enzyme Catalysis

A number of processes, that span a wide-range of length- and time-scales and the relative importance of which varies with the enzyme, contribute to an enzyme-catalyzed reaction [3]. For an isolated enzyme, these processes include: (i) diffusion and binding of substrates in and out of the enzyme's active site; (ii) conformational structural changes, such as loop movements or domain closure, that may be necessary for substrate binding and release or for catalytic activity; and (iii) chemical catalytic steps that involve the breaking and forming of bonds and the transfer of electrons. In addition, there can be other indirect phenomena which influence a reaction. Thus, for example, the optimum catalytic activities of many enzymes can only be attained if they are in specific states, such as when they are bound to non-substrate ligands, when they are covalently modified in some fashion or when they are oxidized or reduced.

Because of the variety of processes entering into enzyme catalysis, no single theoretical technique suffices for modelling an enzyme reaction and so a diverse series of approaches have been developed [4]. Three of the most important categories of technique illustrated in Fig. 1 and are:

1. Quantum chemistry (QC). These methods are appropriate for studying the chemical steps in catalysis as they can be used to compute the wavefunction and, hence, the electron density of a molecular system. The most accurate methods are the *ab initio* and density functional theory methods but they are also the most expensive. They can be used to treat systems of a few tens of atoms on time-scales of the order of tens of picoseconds. By contrast, the less accurate but also

quicker semi-empirical QC methods can handle systems of a few hundred atoms on time-scales of a few nanoseconds.

2. Molecular mechanics (MM) and classical molecular dynamics (CMD). MM techniques are typically much faster than QC methods as they employ empirically-derived functions to calculate the potential energy of a system. They have the disadvantage, though, that they are less generally applicable than QC methods and are inappropriate for simulating chemical reactions. In conjunction with CMD, they can be used to simulate systems of up to several tens of thousand of atoms for time-scales of a few hundred nanoseconds. These methods are particularly well-adapted to studying processes, such as conformational change and ligand-binding, where atomic-level detail is needed but no chemical reactions occur.

3. Coarse-grained (CG) models and Brownian dynamics (BD). Unlike the QC and MM techniques, CG models of a molecular system do not attempt to represent its full atomic detail – instead, atoms are grouped and modelled as larger particles. As an example, common CG models of proteins use one or two particles to describe each amino acid rather than the ten to twenty that would normally be required [5]. CG models also often treat the solvent with some sort of continuum model and dispense with a particle-based representation altogether. CG models, because of their simplicity, can be used to study very large molecular and macromolecular systems and, in conjunction with BD, to simulate processes on time-scales of up to the order of milliseconds. CG/BD models are well-adapted for calculating the values of ligand–enzyme diffusional-encounter rate constants (see reference [6] and the chapter by Stein and co-workers in this volume) and for studying other mesoscopic dynamical processes where atomic-level detail is not needed.

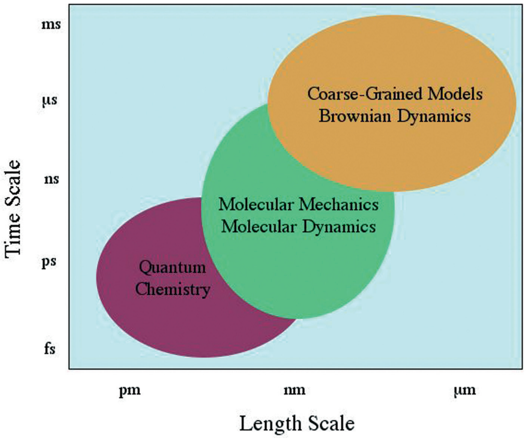


Figure 1. A schematic showing the appropriate length- and time-scales for three of the major classes of theoretical techniques that are employed for modelling enzyme reactions.

Hybrid Potential Approaches

This section and the next focuses on a method, called the hybrid or combined QC/MM potential method, that is one of the primary research interests of the author and which is designed for the study of the chemical catalytic steps in an enzyme reaction [7]. The technique is based upon the following rationale. The modelling of a chemical reaction necessitates the use of QC methods but these are impractical or too expensive to apply to systems of more than about a hundred atoms. This rules out the possibility of studying molecules the size of enzymes. On the other hand, MM methods are very good at being able to handle large systems but are not very good at simulating reactions. Therefore, why not combine the strengths of both methods and use a QC technique to treat the reacting portion of the system and an MM approach to represent the remaining atoms which, although non-reactive, could nevertheless play an important role?

The first hybrid potential was conceived in the 1970 s by Warshel and Levitt for the simulation of the reaction catalysed by lysozyme [8]. It was not until the early-1990 s, however, that they started to be widely used. This lag was due, in part, to a lack of computer power but also because a number of technical issues had to be resolved to have potentials that were sufficiently precise and robust [9]. Hybrid potentials are now employed in all areas of molecular computational science, not just for the simulation of enzymes.

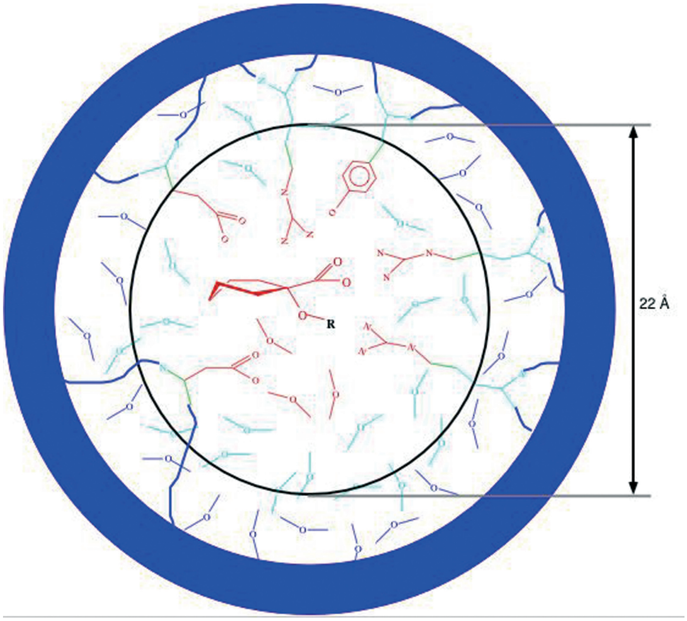


Figure 2. An illustration showing an example of how an enzyme, in this case influenza virus neuraminidase, is represented in a QC/MM hybrid potential simulation. The substrates and other enzyme and solvent groups that are implicated in the reaction are in the QC region (red atoms) whereas the remainder of the atoms in the system (light and dark blue atoms) are placed in the MM region.

Many flavours of hybrid potential have been developed but the simplest and also the most common involve the partitioning of a system between a single QC region and a single MM region. An example of such a partitioning is shown in Fig. 2. The proper formulation of a hybrid potential is quite intricate and depends upon the types of QC and MM potentials that are being combined. The crucial aspect of all formulations though is how the two potentials are coupled or, in other words, how the atoms of the QC and MM regions interact. In the hybrid potentials developed in the author's group, there are two classes of interaction [9,10]. The first class are the non-bonding interactions which occur in all hybrid potential studies. They comprise electrostatic terms between the electrons and nuclei of the atoms in the QC region and the charges of the atoms in the MM region and Lennard-Jones terms which account for the van der Waals and exchange–repulsion interactions between the two sets of atoms. The second class of interaction are the covalent QC/MM interactions which arise whenever a single molecule is split between different regions. Interactions of this type are nearly always present when studying enzyme systems as amino acid groups in the enzyme are almost always catalytically active. The treatment of these interactions is more complicated than that of the non-bonding ones but a number of competing algorithms have been developed that appear to be of roughly equivalent accuracy [7].

Although their formulation is distinct, hybrid potentials can be employed in much the same way as pure QC and pure MM potentials. Thus, for example, it is possible to perform geometry optimizations to locate the stable structures of an enzyme–ligand complex and to run molecular dynamics simulations to investigate the system's dynamics and to calculate its thermodynamics properties. Examples of the application of hybrid potentials in this way will be given in the next section and a full list of quantities accessible by simulation in the section after that.

An Example of a Hybrid-Potential Study

The author's group has studied approximately fifteen enzymes with hybrid potential methods. These include the nickel–iron hydrogenase from *Desulfovibrio gigas* [11], the influenza virus neuraminidase [12], spinach acetohydroxyacid isomeroreductase [13], chorismate mutase from *Bacillus subtilis* [14], rat aldehyde dehydrogenase [15], various class A beta-lactamases and penicillin-binding proteins (PBPs) [16] and cAMP-dependent protein kinase [17]. The goal in most of these studies was a better understanding of the reaction mechanism catalysed by the enzyme although not exclusively. Thus, for example, in the hydrogenase work the aim was to characterize the active-site structures of various different redox intermediates in the catalytic cycle [11], whereas the beta-lactamase and PBP work was undertaken to determine the binding modes of different classes of antibiotics in the active sites [16]. Almost all the hybrid-potential simulations performed in the author's group have been performed with the group's own simulation package, called Dynamo [4,10], which was conceived specifically for studies with hybrid potentials.

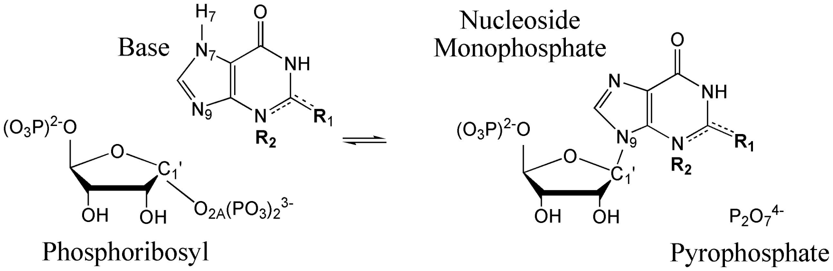


Figure 3. A scheme illustrating the reaction catalysed by the enzymes HGXPRTase and HGPRTase. (R1, R2) are (H, nothing), (NH2, nothing) and (O, H) for hypoxanthine, guanine and xanthine, respectively. The nucleoside monophosphates corresponding to these bases are inosine monophosphate, guanosine monophosphate and xanthosine monophosphate. Specially labelled atoms which will be mentioned in the text are C1', H7, N7, N9 and O2A.

This section describes briefly a study of the enzyme hypoxanthine-guanine-xanthine-phosphoribosyltransferase (HGXPRTase) from *Plasmodium falciparum* (Pf) and its human homologue, hypoxanthine-guanine-phosphoribosyltransferase (HGPRTase). This work is representative of the other applications of the Dynamo program [11 – 17] and illustrates nicely what can and cannot be achieved with hybrid-potential techniques. Only a brief discussion of the work will be given here as full technical details of the simulations and a discussion of the results may be found in references [18 – 20].

Pf is a protozoan and is one of the Plasmodium species responsible for malaria. The reaction catalysed by HGXPRTase is shown in Fig. 3 and involves transfer of a phosphoribosyl group between a base and a pyrophosphate group. The enzyme is active with hypoxanthine, guanine or xanthine as the base. One of the interests of HGXPRTase is that there exists a human equivalent, called HGPRTase, that catalyses the same reaction except that it has a much reduced activity with xanthine. Despite this, the proteins share an 80% sequence homology in the vicinity of the active site although this drops to about 40% overall. This raises two questions; first, what causes this difference in specificity and, second, could this difference be used to design inhibitors that selectively target the Pf enzyme and, hence, act as potential antimalarial drugs.

The hybrid potential simulations that were performed were designed to partially answer both of these questions. First, because the simulations were carried out to investigate the chemical steps of the reaction only and so could not differentiate specificity due to other processes, such as substrate binding, and, second, because detailed information about the structures along the pathway of the reaction resulting from the simulations could help in identifying features that are important for inhibitor design. A summary of the salient features of the simulations and their results are as follows.

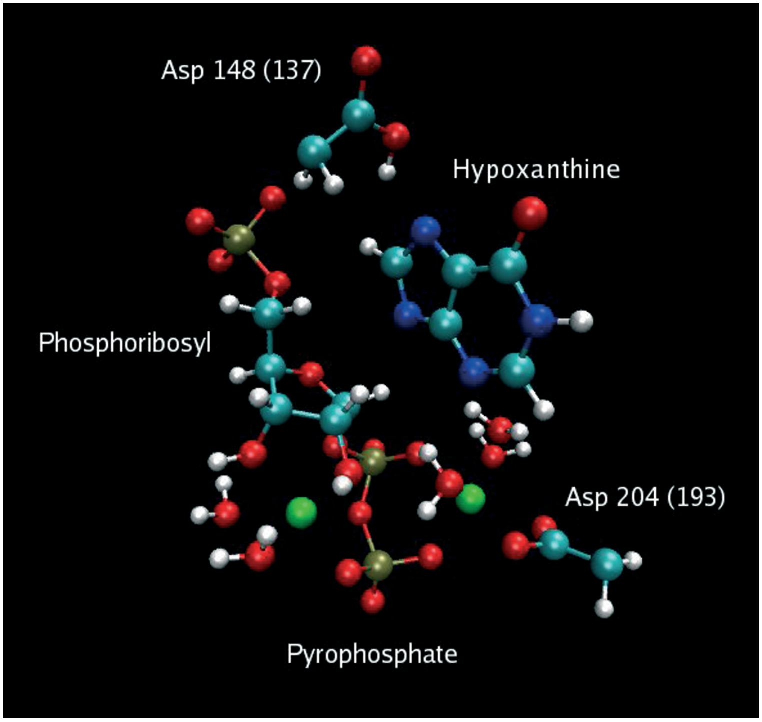


Figure 4. An image showing the atoms in the QC region of one of the HGXPRTase simulation models. Where appropriate the HGPRTase residue numbers are in brackets. The hypoxanthine is in its anionic form as its H7 proton has been transferred to Asp148. The magnesium cations are in green.

1. The reactions of HGXPRTase and HGPRTase were studied with both hypoxanthine and xanthine as substrates.

2. The techniques used to study reactions in systems as complicated as enzymes are currently such that it is not possible to expect the preferred mechanism to come out of the calculations directly. Instead, distinct hypotheses have to be made about how the reaction occurs and each hypothesis tested separately. The results of these tests may then be used to exclude particular mechanisms, if, for example, there is disagreement with experimental data, or, alternatively, to indicate that a mechanism is plausible. In the case of the HG(X)PRTases, the hypotheses tested included: (i) passage by a dissociative (SN1-like) mechanism with a stable dissociated intermediate; (ii) passage by various associative (SN2-like) mechanisms without stable intermediates; (iii) investigation of these mechanisms with different ribose sugar conformations; and (iv) investigation of these mechanisms with different protonation states for critical groups in the active site region.

3. Each hypothesis requires a separate simulation and simulation system. These typically comprised about 23,000 atoms including the enzyme, substrates, sur-

rounding water and counterions. All these atoms were treated in the MM region except for about 80 atoms which were put in the QC region (see Fig. 4). These included the substrates, the two magnesiums in the active site, their coordinating water and amino acid residues, and one or two catalytically active amino acid residue side chains.

4. The reaction mechanism for each simulation system was mapped out using a mixture of three different techniques: (i) saddle-point methods for locating the critical transition-state (TS) structures; (ii) reaction-path methods that generate intermediate structures for the mechanism by interpolating between the reactant and product structures; and (iii) free-energy calculations that determine the free energies for a mechanism as a function of specific reaction-coordinate variables. To give an idea of the computational expense required to test a single hypothesis, a complete study using all three techniques consumed approximately 20,000 hours (or 2.25 years) of computer time. Fortunately this was only necessary in a few instances as most hypotheses could be rejected with much less effort.
5. The preferred mechanism (of the hypotheses that were tested) was found to be identical with similar energetics for both enzymes and both substrates. The mechanism had two steps with an initial transfer of the proton H7 from the base to the adjacent aspartate side-chain followed by phosphoribosyl transfer. The second step was rate-limiting with a barrier of between 70 and 80 kJ per mole. This is in reasonable agreement with the experimental value of 65 kJ per mole and is the type of accuracy that can be expected from simulations of this type. These results indicated that the chemical steps are not responsible for the difference in specificity between HGXPRTase and HGPRTase but are due to other causes, such as binding.
6. Despite the identical mechanisms, an analysis of the TS structures for the rate-limiting phosphoribosyl transfer step in the two enzymes revealed some significant structural differences which could be important in inhibitor design. The most striking of these were the C1'-N9 distances which had average values of 2.56 and 1.82 Å in the Pf and human TS structures, respectively, and the C1'-O2A distances whose average values were 1.80 and 2.29 Å.

Quantities Accessible by Computation

Simulations are not a replacement for experiment but a complement, the results of which serve to test particular hypotheses about an enzyme reaction and to spur the design of new experiments. Quantities that can be determined more or less routinely by computation include:

1. Structures of stable enzyme–ligand complexes at various stages of the catalytic cycle.
2. Mechanisms. These can be elucidated by finding paths that link reactants and products and pass via known intermediate structures. The structures of unstable

species along the reaction path, such as saddle points or TSs, are inaccessible experimentally but they are of both fundamental and applied interest. First, they are necessary for calculation of the additional quantities occurring later on this list and, second, they are useful in processes, such as inhibitor design, for giving insight into the properties of a ligand that are necessary for tight-binding to an active site.

3. Free energies and associated thermodynamic quantities such as enthalpy and entropy. Free energies are determined as differences between stable or unstable structures. In the former case, the energies may be related to the equilibrium constants between species whereas, in the latter, they can provide an estimate of the rate of a process via transition state theory (TST).
4. Rate constants. For some processes, such as the diffusional encounter of an enzyme with a ligand, rates are calculated directly from simulation. In other cases, such as when investigating the chemical steps in the reaction, it is more usual to first obtain the TST estimate of the rate constant via free-energy calculations and then to correct the TST value for dynamical and quantum effects.
5. Kinetic isotope effects. These may be calculated if the TS structures for a mechanism have been calculated.

Simulation studies of enzyme reactions can require much effort, testing and trial and error. They have the advantage, though, that once a set of simulation systems has been obtained that encapsulate the process being studied, it is straightforward to rerun them under different physical conditions (ionic strength, pH value, pressure, temperature, etc.) or after (limited) mutations have been made in the enzyme or substrate structures.

Limitations of Current Approaches

Current modelling and simulation methods can provide much useful information about specific aspects of enzyme catalysis when properly applied. They do, however, have some fundamental limitations, among the most important of which are:

1. Almost all simulations require as input protein structures that are determined experimentally, usually by X-ray crystallography, but sometimes by other techniques, such as NMR. The requirements on the precision of these structures depends upon the type of simulation being performed. The results of hybrid potential calculations of the catalytic chemical steps in an enzyme reaction are often very sensitive to the arrangement and orientation of groups in the active site and so it is normal to start with high-resolution protein structures (of 2 Å resolution or better) that have been obtained in the presence of substrates, inhibitors or TS analogues. By contrast, calculations of the rates of diffusional encounter between an enzyme and its substrates are less stringent but, even so, structures need to be of sufficient resolution that accurate representations of the enzymes' charge distributions can be created.

2. Simulation studies are computationally intensive. The calculation of diffusional-encounter rates is one of the cheaper types of simulation and takes on the order of a day (or few) on a modern single processor computer for an average size protein. At the other extreme, the computation of free-energy profiles for the chemical steps in a reaction using hybrid potentials typically require several thousand hours of computer time and so special multiprocessor parallel computers must be employed.

3. There are still some aspects of enzyme catalysis that cannot reliably be studied with existing simulation techniques. Two specific examples are: (i) the investigation of reactions in which the enzyme undergoes a significant, but unknown, conformational change.
   Interpolative methods of reasonable precision are available for predicting how a conformational change occurs if structures of the two end states are available but extrapolative prediction is much more difficult; (ii) current QC techniques may not be accurate enough when studying systems with complicated electronic structures, such as radicals and transition metal complexes.

4. Probably the great majority of simulation studies published to date are irreproducible. There are two reasons for this: (i) there is a very large diversity of simulation methods and there has been little attempt at standardizing them. Even with techniques that are nominally the same, comparison can be impractical due to differences in the way that the techniques are programmed or in the parameter sets that are used for the simulations; (ii) much information is needed to reproduce a simulation with a particular program. This includes the simulation conditions, miscellaneous parameter sets and other data, such as atomic coordinates. Many journals do not require that this data is made available as a condition of publication, and, for those that do, there is little enforcement. It should be emphasized that the situation is generally better for calculations performed with purely QC techniques as these are easier to standardize [21]. Therefore, it is, in principle, possible to obtain ”identical” results for a particular system when using different QC programs as long as one has a detailed enough description of the methods employed and access to the starting data for the simulations.

Connections to Systems Biology

Modelling the behaviour of metabolic, signalling and other pathways has been an area of active research since at least the 1970 s but it has received increased attention in the last few years. This has been due to the emergence of systems biology as a discipline which seeks to provide a multiscale description of how particular aspects of a cell or organism function by integrating diverse experimental and theoretical approaches [22]. One of the principal limitations in systems biology modelling is the lack of experimental data that characterizes how parts of the system behave. This is especially true for the simulation of enzymatic pathways as only for a small number of pathways are the parameters and the laws that govern the kinetics of all the constituent enzymes known.

Modelling and simulation will not be able to resolve this impasse by themselves but they clearly have a role to play given the difficultly in obtaining much of this data experimentally. A two-fold strategy would be appropriate:

1. The use of fast, approximate methods to estimate parameters, such as binding and rate constants, for the enzymes in a pathway. The precision required of these estimates will depend upon the enzyme and the pathway but it seems possible that, for many enzymes, relatively crude estimates will suffice as the simulated behaviour of a pathway will be robust to parameter changes within reasonable bounds. Of the methods described in this chapter, only the CG/BD approach for computing diffusional-encounter rate constants can be considered fast in the sense envisaged here and so new methods will need to be developed.
2. The application of computationally intensive methods, such as hybrid potentials, only for the investigation of the ”critical” enzymes in a pathway.

Reporting Enzyme Data Obtained Via Simulation

The STRENDA guidelines for reporting enzymology data refer exclusively to data obtained experimentally and need adapting if they are to cover theoretically-derived values as well. A partial list of recommendations includes:

1. A way of distinguishing experimental and theoretical data. One model is employed by the Protein Data Bank (PDB) which separates data from the different sources [23].
2. Most modelling approaches require structures so the identity of the enzyme should be supplemented by the origin of the structures used in the simulation (e. g. the PDB codes).
3. An overview of the theoretical methodology employed. A literature reference would probably be enough for a standard method or a published result but in other cases a more detailed exposition would be necessary.
4. Details of the software and the machines used for the simulations.
5. A summary of the simulation protocol that includes, among other things, descriptions of the parameter sets, physical conditions and methods of data analysis.
6. A minimum set of data files that would allow other workers to reproduce the calculations given equivalent software and computing facilities. The type of data required would vary according to the simulation methodology but for the more complicated approaches it would mean providing starting coordinate and velocity sets for the enzyme system, parameter sets for the MM and QC potentials and sample program input files.

Acknowledgements

The author would like to thank all members of the Modeling and Simulation Group and, in particular, Aline Thomas for her work on the HG(X)PRTase enzymes and Ramon Crehuet for his work on reaction-path finding techniques. Acknowledgements are also due to the European Union, the French Atomic Energy Commission (CEA) and the French National Center for Scientific Research (CNRS) for financial support.

References

[1] Brent, R., Bruck, J. (2006) Can computers hope to explain biology? *Nature* **440:**416–417.

[2] Apweiler, R., Cornish-Bowden, A., Hofmeyr, J.-H.S., Kettner, C., Leyh, T.S., Schomburg, D., Tipton, K. (2005) The importance of uniformity in reporting protein-function data. *Trends Biochem. Sci.* **30:**11–12.

[3] Ferhst, A. (1999) *Structure and Mechanism in Protein Science*: *Guide to Enzyme Catalysis and Protein Folding*. W.H. Freeman, New York.

[4] Field, M.J. (1999) *A Practical Introduction to the Simulation of Molecular Systems*. Cambridge University Press, Cambridge.

[5] Tozzini, V. (2005) Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **15:**144–150.

[6] Gabdoulline, R.R., Wade, R.C. (2001) Protein-protein association: investigation of factors influencing association rates by Brownian dynamics simulations. *J. Mol. Biol.* **306:**1139–1155.

[7] Field, M.J. (2003) Simulating chemical reactions in complex systems. In: *Handbook of Numerical Analysis.(* Le Bris, C. Ed.).Vol. X, pp.667–697. Elsevier Science, Amsterdam.

[8] Warshel, A., Levitt, M. (1976) Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103:**227–249.

[9] Field, M.J., Bash, P.A., Karplus, M. (1990) A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.* **11:**700–733.

[10] Field, M.J., Albe, M., Bret, C., Proust-De Martin, F., Thomas, A. (2000) The dynamo library for molecular simmlations using hybrid quantum mechanical and molecular mechanical potentials. *J. Comput. Chem.* **21:**1088–1100.

[11] Amara, P., Volbeda, A., Fontecilla-Camps, J.-C., Field, M.J. (1999) A hybrid density functional theory/molecular mechanics study of nickel-iron hydrogenase: investigation of the active site redox states. *J. Am. Chem. Soc.* **121:**4468–4477.

[12] Thomas, A., Jourand, D., Bret, C., Amara, P., Field, M.J. (1999) Is there a covalent intermediate in the viral neuraminidase reaction? A hybrid-potential free-energy study. *J. Am. Chem. Soc.* **121:**9693–9702.

[13] Proust-De Martin, F., Dumas, R., Field, M.J. (2000) A hybrid-potential free-energy study of the isomerization step of the acetohydroxy acid isomeroreductase reaction. *J. Am. Chem. Soc.* **122:**7688–7697.

[14] Marti, S., Andres, J., Moliner, V., Silla, E., Tunon, I., Bertran, J., Field, M.J. (2001) Insights into enzyme catalysis: the chorismate mutase case. *J. Am. Chem. Soc.* **123:**1709–1712.

[15] Wymore, T., Deerfield II, D.W., Field, M.J., Hempel, J., Nicholas Jr, H.B. (2003) Initial catalytic events in class 3 aldehyde dehydrogenase: MM and QM/MM simulations. *Chem.-Biol. Interact.* **143/144:**75–84.

[16] Oliva, M., Dideberg, O., Field, M.J. (2003) Understanding the molecular mechanism of active-site serine penicillin-recognizing proteins: a molecular dynamics simulations study. *Proteins*: *Struct. Funct. Genet.* **53:**88–100.

[17] Diaz, N., Field, M.J. (2004) Insights into the phosphoryl-transfer mechanism of cAMP-dependent protein kinase from quantum chemical calculations and molecular dynamics simulations. J. *Am. Chem. Soc.* **126:**529–542.

[18] Thomas, A., Field, M.J. (2002) Reaction mechanism of the HGXPRTase from Plasmodium falciparum: a hybrid potential quantum mechanical/molecular mechanical study. *J. Am. Chem. Soc.* **124:**12432–12438.

[19] Crehuet, R., Thomas, A., Field, M.J. (2005) An implementation of the nudged-elastic-band algorithm and application to the reaction mechanism of HGXPRTase from Plasmodium falciparum. *J. Mol. Graph. Model.* **24:**102–110.

[20] Thomas, A., Field, M.J. (2006) A comparative QM/MM simulation study of the reaction mechanisms of human and Plasmodium falciparum HG(X)PRTases. *J. Am. Chem. Soc.*, in press.

[21] Boggs, J.E. (1999) Guidelines for presentation of methodological choices in the publication of computational results: ab initio electronic structure calculations. *J. Comput. Chem.* **20:**1587–1590.

[22] Werner, E. (2005) The future and limits of systems biology. *Science* STKE 278:pe16.

[23] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*. **28:**235–242.

Assaying Enzymes from Hyperthermophiles

Wilfred R. Hagen

Department of Biotechnology, Delft University of Technology,
Julianalaan 67, 2628 BC Delft, The Netherlands

E-Mail: w.r.hagen@tudelft.nl

Received: 4th July 2006 / Published: 31st August 2007

Abstract

The determination of kinetic and thermodynamic data from hyperthermophilic enzymes at physiological temperature (i.e. ≥ 80 °C) raises a number of technical and fundamental problems. Based on studies of purified enzymes from the model organism *Pyrococcus furiosus* several of these problems are identified and explored here. It is proposed that kinetic and thermodynamic data on hyperthermophilic enzymes be reported at the organism's growth temperature or, alternatively, at a lower temperature compatible with practical assay conditions with additional data obtained at yet lower temperatures to allow for extrapolation.

Introduction

The native structure of biomacromolecules is metastable with respect to a number of physico-chemical parameters: e.g., ionic strength, pH, radiation, pressure, temperature. To a limited extent the cell can develop specific biochemical capacities to protect itself from the detrimental effects of the extreme values of these environmental parameters. A well-known example is the capacity of the bacterium *Helicobacter pylori* to raise locally acidity of the human stomach from a nominal pH value of 1.5 to circa 5 – 6, by producing large quantities of the nickel enzyme urease for the production of ammonia from human-made urea [1]. A second example is the capacity of many halotolerant or halophilic micro-organisms to take up or to synthesize organic compatible solutes, such as the quaternary amine betaine, and thus to balance osmotic potential in an environment of high ionic strength [2]. However, for other environmental boundary conditions, notably extremes in temperature, micro-organisms have apparently been unable to develop machineries to stabilize their interior at mesophilic values. Thus, e. g., the archaeon *Pyrococcus furiosus* that grows optimally at an environmental temperature of 100 °C [3], also has an intracellular temperature of 100 °C, and, therefore, it must have its *entire* biochemistry adapted to this biologically extreme temperature. In this framework the enzymologist is not only faced with a fundamental problem (what is the biochemical nature of high-temperature adaptation) and with a practical problem (how does one measure biological activities at high temperatures), but also with a problem of normalization (under what conditions should 'hot' enzymes be assayed to maximize comparability with 'regular' enzymes).

Some two decades of biochemical research on hyperthermophiles has until now left the notion of unity in biochemistry unshaken. No fundamentally new concepts have been discovered related to the central pillars of life: the bioenergetics of oxidative phosphorylation, the transcription of DNA, translation of RNA, chaperone-assisted protein folding, and so on. Also, the use that hyperthermophiles make of building blocks (ATP), metabolites (glyceraldehyde-3-phosphate), and cofactors (NADPH) appears to be completely conventional, and this is remarkable in view of the limited lifetime versus thermal degradation of these compounds in dilute aqueous solution at 100 °C [4]. How hyperthermophiles succeed in stabilizing thermolabile intermediates is an unsolved problem. A partly solved problem is the thermostability of proteins from hyperthermophiles: comparisons with mesophilic counterparts at different levels ranging from pair wise comparison of 3D structures to predicted proteins from multiple genomes [5,6] suggest that the determinant is a multifaceted one, encompassing an increased number of salt bridges, hydrogen bridges, beta sheets, shortened loops, altered amino acid usage, etc. The implication is that protein thermostability in general is not predictable at this time, therefore, that mutagenesis towards increased stability is not yet possible in a rational way.

Our work on high-temperature enzymology has focused on the hyperthermophilic, anaerobic, marine euryarchaeoton *Pyrococcus furiosus* as a model system for several reasons. The organism is readily grown, e. g., on starch in 100 litre batch cultures at circa 93 °C with a doubling time of circa 40 minutes. Its biochemistry has been under study for nearly two decades. Its complete genome and also those of half a dozen closely related species (*Thermococcales* spp) are freely available. Several of its genes have been found to be readily (over)expressed in *Escherichia coli* as functional proteins. The biochemistry of *P. furiosus* and related species is mildly idiosyncratic, for example, in its strong preference for (and absolute dependence on) the 5 d transition element tungsten. In our ongoing studies we have identified a number of fundamental and practical problems, illustrated below, that may be of general relevance to quantitative 'hot' enzymology.

High Temperature Adapted Assay Instrumentation

The conceptually simple rise in temperature required to measure enzymes from hyperthermophiles at or near their physiological temperatures poses technical problems of variable complexity depending on the type of assay. Our work on *P. furiosus* has focused on redox catalysis and thus on the use of on-line assays based on the detection of redox dyes (spectrometry), electrons (direct electrochemistry) and also of gaseous substrates/products (amperometry).

UV-visible research spectrometers are commonly equipped with variable temperature accessories; however these usually have not been designed to operate in the 80 – 110 °C range. Furthermore, they are made to accept 1 cm square cuvettes, some of which, notably the fused quartz type, do not withstand high temperature for prolonged periods, and so high temperature colour-based assays can prove to be a costly exercise.

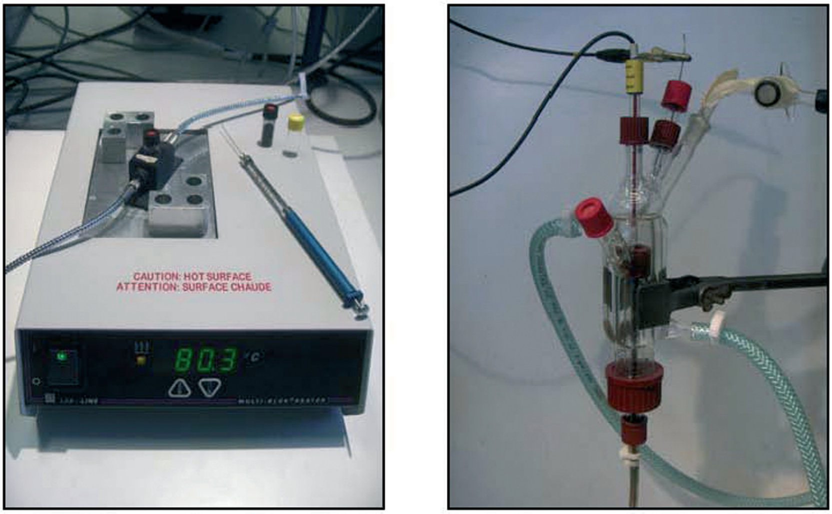


Figure 1. High temperature adaptations of routine assay equipment: (A) UV-vis spectrometry; (B) direct electrochemistry.

Figure 1a is a picture of a relatively simple solution suitable for routine assay of numerous samples, e. g., during protein purification. The cuvette house of a fibre optic spectrometer (Avantes – The Netherlands) has been built into an aluminum block that is part of a heating plate. The cuvette house has been adapted to take round glass bottles (HPLC type; 1 cm diameter). The heating block also contains additional holding positions for the pre-thermostatting of cuvettes. The HPLC bottles are closed with a septum which does not only allow for anaerobicity but also for the temperature to be raised up to a few degrees above 100 °C. The plastic cap allows for rapid cuvette transfer by hand at high temperature.

Direct electrochemistry on solid electrodes allows determining reduction potentials of electron transfer proteins and of some enzymes by cyclic voltammetry; it can also be used to assay redox enzymes in combination with natural or artificial electron transfer partners by measuring the extent of a catalytic wave in cyclic voltammetry. We have previously described a simple three electrode electrochemical cell for direct protein electrochemistry built around a small drop of solution (typically 10 – 50 µl) on top of a flat activated glassy carbon disc as the working electrode [7]. This design can be readily adapted for high temperature studies up to circa 90 °C by surrounding the cell with a thermostatted water jacket connected to a circulating water bath (Fig. 1b). The drop of solution is protected from evaporation by overlaying it with a small amount of immersion oil as used in microscopy or in PCR instruments for DNA multiplication. The reference electrode is of the Ag/AgCl type (saturated KCl) which, contrast to calomel eletrodes, can be operated up to at least 100 °C and provides a – temperature dependent – well defined reference potential [8].

A similar solution of isothermal junctions between electrodes by means of a temperature controlled water jacket is possible for amperometrically assaying gaseous substrates (O_2, H_2, NO, N_2O) with the familiar membrane covered electrochemical Clark cell, or oxygraph, the combined electrode of which is platinum versus Ag/AgCl. Here, solvent evaporation is not an issue in view of the larger cell volume (1 – 2 ml) and smaller evaporating surface (1 mm diameter port for injections/degassing). Temperature specs for commercial versions of the Clark cell will typically be limited to rather low values (40 ° C), but jacketed cells are readily home made of polycarbonate and can be run up to nearly boiling temperature.

The Problem of Choosing a Proper Assay Temperature

Figure 2 is a 'typical' plot of enzyme activity versus temperature [9]. The example is from one of the tungsto-enzymes of *P. furiosus*, Aldehyde OxidoReductase. AOR catalyses the two-electron oxidation of a range of aldehydes to their corresponding acids. The highest catalytic competence (k_{cat}/K_M) is for the substrate crotonaldehyde when assayed with benzyl viologen as electron acceptor (Fig. 2):

$$CH_3CHCHCHO + H_2O \leftrightarrow CH_3CHCHCOOH + 2\,H^+ + 2e^-$$

but the natural substrate(s) has not been unequivocally identified [10]. AOR is thought to function in catabolism of proteinaceous material (amino acid degradation). Figure 2 illustrates two key aspects of 'hot' enzymology.

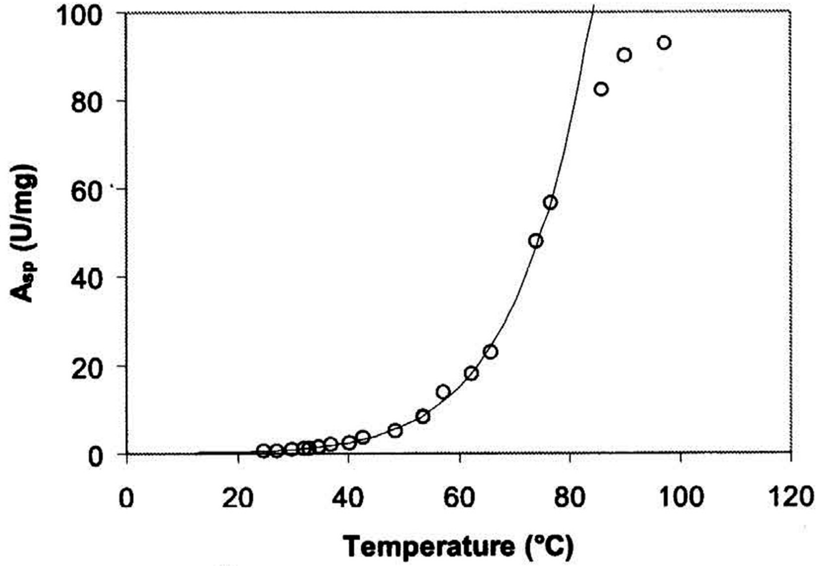


Figure 2. Apparent crotonaldehyde oxidation activity of *P. furiosus* aldehyde oxidoreductase as a function of temperature.

Firstly, as with *any* enzyme activity, the reaction rate is seen to approximately double with every 10 degrees increase in temperature. However, for hyperthermophilic enzymes the *dynamic range* of temperatures over which activity is practically measurable, is much greater than for mesophilic enzymes. This fact implies improved possibilities for, e. g., the study of temperature-dependent protein conformational changes, and the determination of activation energies associated with enzyme catalysis. Furthermore, a poor-man's version of pre-steady-state kinetics presents itself: at laboratory ambient temperatures hyperthermophilic enzymes are slowed down to the extent that trapping of enzyme kinetic intermediates would appear to no longer require advanced (stopped flow; rapid quench) equipment. This fascinating possibility of facilely creating kinetically relevant intermediates on a 'hand-mixing' time scale remains largely unexplored to date.

Secondly, and again as with *any* enzyme, increasing the assay temperature will eventually lead to the thermal degradation of activity. However, when employing the operational definition of physiological temperature as the temperature at which the hyperthermophilic organism exhibits maximal growth rate under optimized laboratory conditions, it is found more often than not that hyperthermophilic enzymes degrade thermally at a faster rate than would be compatible with the time scale of their activity assays. In Fig. 2 this is seen as a deviation of experimental points at high temperature from the fitted exponential. The dilemma for the enzymologist is obvious: the assay temperature has to be lowered to a value such that the enzyme will be stable at least for the time span that it takes to measure its activity. In the example of Fig. 2 this corresponds to circa 80 °C, which can still be considered a – suboptimal – physiological temperature as *P. furiosus* will still grow at 80 °C (be it at a reduced rate). However, different enzymes from the same species vary drastically in their thermal stability in dilute aqueous solution (*cf.* ferritin, below) and a single standard temperature for all enzymes could only be defined after all enzymes would have been purified and characterized. Clearly, such a normalized temperature would not be equal to the temperature of maximal growth. It is therefore suggested that kinetic data on (hyper)thermophilic enzymes be reported at the highest temperature at which their activity is stable over the time period required for a reliable assay (if lower than the temperature of maximal growth), and that this information be extended with data on activity and stability as a function of temperature.

Interference of Non-Catalytic Reaction at High Temperature

Ferritin is a small (circa 20 kDa) α-helical protein that spontaneously polymerizes into a cage-shaped homo 24-mer, and that ubiquitously occurs in all domains of life. Its main physiological function is thought to be the storage of iron and/or a protection against oxidative stress [11]. Ferritin takes up Fe(II) ions and converts these in the presence of an oxidant, e. g., O_2, into a core of ferrihydrite. Its activity can be assayed by measuring the increase in light scattering at, e. g., 315 nm from the growing Fe(III) core. Non-biological oxidation of Fe(II) by O_2 at ambient temperatures is usually very much slower than ferritin-catalysed oxidation. A structural ferritin gene in *P. furiosus* can be cloned and overexpressed in *E. coli* resulting in an extremely thermostable 24-mer the Fe(II) oxidation activity of which is resistant to 10 h boiling at 100 °C or 30 min autoclaving at 120 °C [12].

When assayed at 25 °C this ferritin exhibits cooperative kinetics (Hill coefficient $n \approx 2$) and a half-maximal activity for $K_{0.5}$= 5 mM Fe(II). At 25 °C the non-catalytic rate of Fe(II) oxidation for [Fe(II)]= 5 mM is negligible compared to the ferritin catalysed rate. When the temperature is raised to 85 °C the Fe(II) oxidation activity increases by circa two orders of magnitude consistent with an approximate two-fold increase in rate for every 10 °C increase in temperature. However, this activity can only be measured at relatively low [Fe(II)] $\leq$0.3 mM. At [Fe(II)]= 5 mM the rate of non-catalytic oxidation is comparable to that of ferritin-catalysed oxidation, and this interference precludes a complete kinetic analysis at high temperature. This example illustrates interference of a background reaction under *in vitro* assay conditions to the extent that kinetic analysis is limited to non-physiological temperature (*P. furiosus* does not grow at 25 °C).

High-temperature studies of *P. furiosus* ferritin have pointed to another technical problem of 'hot' biochemistry: the extreme thermostability of this protein is reflected in the fact that differential scanning calorimetric measurements fail to reveal a 'melting' temperature up to 120 °C [12]. It appears that calorimetry is not a practical option to study unfolding of these types of proteins.

Determination of Reduction Potentials at High Temperture

A key thermodynamic parameter to be determined in the study of redox proteins is the reduction potential, E_m, of the prosthetic group(s). An E_m value reflects the relative stability of the oxidized versus the reduced form of a compound and thus can be dependent on a number of environmental parameters, e. g., pH and ionic strength, but also: temperature. This raises the question at what temperature redox properties of (hyper)thermophilic proteins should be determined and reported. E_m values of proteins are not necessarily linear functions of temperature, e. g., due to temperature-dependent protein conformational changes. Consequently, E_m values should preferably be determined at physiological temperature, which is, however, not a trivial problem as can be illustrated on the example of *P. furiosus* rubredoxin, a small (6 kDa), thermostable electron transfer protein with a single Fe(II/III) redox prosthetic group. Its E_m can be determined as a function of temperature by direct voltammetry on activated glassy carbon as shown in Fig. 3. Apparently, no temperature-dependent conformational changes occur over the studied temperature range because the reduction potential is found to be linear in T with an approximate -1.5 mV change per degree increase [13].

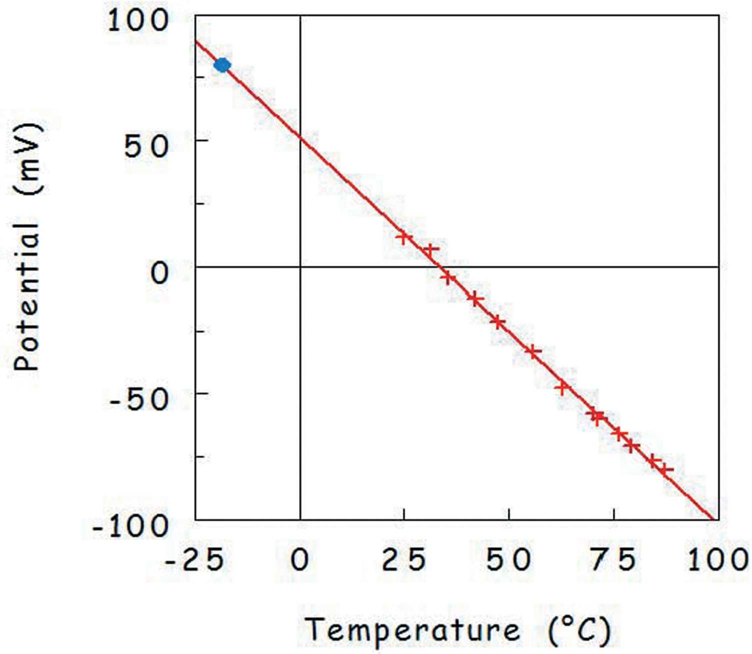


Figure 3. Reduction potential of *P. furiosus* rubredoxin as a function of temperature as determined with direct electrochemistry (((plus))) and with EPR monitored titration (●).

An alternative technique to determine E_m values is bulk titration monitored with EPR spectroscopy. Substoichiometric additions of reductant (dithionite) or oxidant (ferricyanide) are made to the protein solution in the presence of a cocktail of redox mediators to ensure redox equilibrium between the protein and the detection electrode (platinum), and when equilibrium is reached (i. e. a constant voltage reading) then a sample is drawn and rapidly frozen for cryogenic EPR analysis in order to determine the extent of reduction of the prosthetic group. The importance of this method lies in (1) its general applicability to metalloproteins that usually exhibit an EPR signal either in their oxidized or in their reduced state, and (2) the finding that many proteins, notably redox enzymes do not exhibit finite electron transfer rates with bare electrodes, which precludes general application of the direct voltammetry method. Remarkably, when the bulk titration method is applied to *P. furiosus* rubredoxin, the outcome is $E_m \approx +80$ mV *in*dependent of whether the titration is done at 20 °C or at 80 °C [13]. This observation suggests that, during the cooling down of the EPR sample towards its freezing point the rubredoxin protein sufficiently rapidly adapts its structure so that the determined E_m value always corresponds to the sample's freezing temperature whatever the initial sample temperature was (*cf.* Fig. 3). The general implication would be that reduction potentials of (hyper)thermophilic electron transfer proteins at physiological temperatures cannot be determined accurately by EPR monitored redox titrations. Whether this conclusion also holds for larger proteins, notably enzymes, is not yet clear at this time.

Enzyme Activation During Heat-Up

The enzyme hydrogenase catalyses the activation of molecular hydrogen in nature either for its oxidation to protons or for its formation from protons [14]:

$$2\,H^+ + 2e^- \leftrightarrow H_2$$

The most common form of the enzyme has an active centre consisting of a heterodinuclear cluster of nickel and iron bridged by cysteinato sulfurs and with structural CO and CN^- ligands coordinating the Fe ion. In purified NiFe-hydrogenase this unusual centre is frequently found to be in an inactivated form which can be activated by anaerobic incubation under hydrogen. The activation process may involve the removal of a bridging oxo or peroxo group and the concomitant or subsequent reduction of the dinuclear cluster [14]. The rate of activation depends on the source of the enzyme and on its history. *P. furiosus* makes a soluble NiFe-hydrogenase the metal centre of which is in an inactive, oxidized form after purification of the enzyme at ambient temperature. EPR spectroscopic studies have shown that this hydrogenase goes through an auto-activation cycle when heated up anaerobically to, e. g., 80 °C [15]. The mechanism of auto-activation is unknown, but it may involve one or more molecules of H_2 trapped in the protein's H_2 channelling system during isolation. The general implication of these observations is that (hyper)thermophilic enzymes may exhibit a broad range of extents of activation depending not only on the temperature of assay, but also on the time of pre-incubation at this temperature.

Change of Substrate Solvation with Temperature

In addition to the already mentioned aldehyde oxidoreductase, AOR, *P. furiosus* synthesizes at least four more tungsten-containing oxidoreductases, one of which is named formaldehyde oxidoreductase, FOR, because for all tested aldehyde substrates this enzyme has the highest k_{cat} for formaldehyde [16]. The physiological relevance of this observation has however been questioned because the apparent K_M for formaldehyde is unrealistically high, namely of the order of 10^{-2} M [16]. As with AOR, also FOR is thought to function in a degradation pathway of proteinaceous material.

Formaldehyde is a unique aldehyde in its very strong tendency to be hydrated in aqueous solution. The hydration equilibrium lies well towards the direction of methylene glycol formation such that only a very small fraction is in the free formaldehyde form:

$$\mathrm{HCHO} + \mathrm{H_2O} \leftrightarrow \mathrm{HOCH_2OH}$$

So what is the actual substrate of the FOR-catalysed oxidation reaction: is it methylene glycol or formaldehyde? A reasonable answer is suggested by temperature-dependent Michaelis–Menten analysis. At 20 °C the K_M= 40 mM and at 80 °C the K_M= 6 mM for total formaldehyde (free plus hydrated); however, when these values are re-calculated for free formaldehyde using the temperature dependence of the dissociation constant for methylene glycol [17] the $K_M \approx 0.03$ mM is *in*dependent of temperature [18]. It thus appears that *free* formaldehyde is an excellent substrate for the enzyme FOR. This example illustrates another general rule of 'hot' biochemistry: the non-biological hydration chemistry of the substrate may be very different at mesophilic versus (hyper)thermophilic temperatures.

Conclusions

In comparison to the study of their mesophilic counterparts exploration of the quantitative enzymology of thermophilic enzymes, and, *a fortiori*, of hyperthermophilic enzymes raises a number of additional problems of a practical and also of a fundamental nature.

Established methods of analysis have to be adapted to handle samples at high temperatures. For some methods this can be a relatively simple technical adjustment (colorimetric assays, *cf.* Fig. 1), but for other methods attempts at adaptation can reveal intrinsic limitation (EPR monitored redox titration, *cf.* Fig. 3).

Unfolding of hyperthermophilic proteins may be intrinsically difficult to study as these proteins may not exhibit melting in the temperature range in which commercial calorimetries operate.

The study of redox properties of hyperthermophilic redox enzymes may be intrinsically difficult because common methods to determine reduction potentials do not work at high temperatures.

Activation mechanisms for enzymes purified in an inactivated form can be a complex function of temperature.

Non-biological substrate chemistry (e. g., hydration; oxidation on air) may strongly vary with temperature and may thus complicate temperature-dependent enzymology.

When physiological temperature is defined as the optimal growth temperature of a hyperthermophilic micro-organism, it is frequently difficult to assay activities at this temperature because the enzymes may have limited stability in isolated form in dilute solutions.

In summary, with a view to the standardization of assay conditions for hyperthermophilic enzymes it is advised that data be reported either at optimal growth temperature or, if this is not feasible, at the highest possible temperature at which assays can be reliably run. Particularly in the latter case it is of relevance also to obtain data at lower temperatures to allow for cautious extrapolation.

References

[1] Stingl, K., Altendorf, K., Bakker, E.P. (2002) Acid survival of *Helicobacter pylori*: how does urease activity trigger cytoplasmic pH homeostasis? *TRENDS Microbiol.* **10**:70–74.

[2] Roberts, M.F. (2005) Organic compatible solutes of halotolerant and halophilic microorganisms. *Saline Systems* **1**:5 (doi: 10.1186/1746–1448–1-5).

[3] Fiala, G., Stetter, K.O. (1986) *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100 ^{0}C. *Archs Microbiol.* **145**:56–61.

[4] Daniel, R.M., Cowan, D.A. (2000) Biomolecular stability and life at high temperatures. *Cell. Mol. Life Sci.* **57**:250–264.

[5] Chakravarty, S., Varadarajan, R. (2000) Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Lett.* **470**:65–69.

[6] Suhre, K., Claverie, J.-M. (2003) Genomic correlates of hyperthermostability, an updata. *J. Biol. Chem.* **278**:17198–17202.

[7] Hagen, W.R. (1989) Direct electron transfer of redox proteins at the bare glassy carbon electrode. *Eur. J. Biochem.* **182**:523–530.

[8] Bard, A.J., Faulkner, L. (2001) *Electrochemical Methods; Fundamentals and Applications*, 2nd Edn. Wiley, New York.

[9] Hagedoorn, P.-L. (2002) *Metalloproteins Containing Iron and Tungsten: Biocatalytic Links between Organic and Inorganic Redox Chemistry*. PhD thesis, Delft University of Technology.

[10] Mukund, S., Adams, M.W.W. (1991) The novel tungsten-iron-sulfur protein of the hyperthermophilic archaebacterium *Pyrococcus furiosus* is an aldehyde ferredoxin oxidoreductase: evidence for its participation in a unique glycolytic pathway. *J. Biol. Chem.* **266**:14208–14216.

[11] Crichton, R.R. (2001) *Inorganic Biochemistry of Iron Metabolism: From Molecular Mechanisms to Clinical Consequences*, 2nd Edn. Wiley, Chichester.

[12] Tatur, J., Hagedoorn, P.-L., Overeijnder, M.L., Hagen, W.R. (2006) A highly thermostable ferritin from the hyperthermophilic archaeal anaerobe *Pyrococcus furiosus*. *Extremophiles* **10**:139–148.

[13] Hagedoorn, P.L., Driessen, M.C.P.F., van den Bosch, M., Landa, I., Hagen, W.R. (1998) Hyperthermophilic redox chemistry: a re-evaluation. *FEBS Lett.* **440**:311–314.

[14] Cammack, R., Robson, R., Frey, M. (2001) *Hydrogen as a Fuel – Learning from Nature*. Taylor & Francis, London.

[15] Silva, P.J., de Castro, B., Hagen, W.R. (1999) On the prosthetic groups of the NiFe sulfhydrogenase from *Pyrococcus furiosus*: topology, structure, and temperature-dependent redox chemistry. *J. Biol. Inorg. Chem.* **4**:284–291.

[16] Roy, R., Mukund, S., Schut, G., Dunn, D.M., Weiss, R., Adams, M.W.W. (1999) Purification and molecular characterization of the tungsten-containing formaldehyde ferredoxin odidoreductase from the hyperthermophilic archaeon *Pyrococcus furiosus*: the third of a putative five-membered tungstoenzyme family. *J. Bacteriol.* **181**:1171–1180.

[17] Winkelman, J.G.M., Voorwinde, O.K., Ottens, M., Beenackers, A.A.C.M., Janssen, L.P.B.M. (2002) Kinetics and chemical equilibrium of the hydration of formaldehyde. *Chem. Eng. Sci.* **57**:4067–4076.

[18] Bol, E., Bevers, L.E., Hagedoorn, P.-L., Hagen, W.R. (2006) Redox chemistry of tungsten and iron-sulfur prosthetic groups in *Pyrococcus furiosus* formaldehyde ferredoxin oxidoreductase. *J. Biol. Inorg. Chem.* **11:**999–1006.

Decomposition into Minimal Flux Modes: A Novel Flux-Balance Approach to Predict Flux Changes in Metabolic Networks from Changes of Enzyme Concentrations

Sabrina Hoffmann, Andreas Hoppe and Hermann-Georg Holzhutter

Institut für Biochemie, Universitätsmedizin Berlin (Charité), Humboldt-Universität zu Berlin, Monbijoustr. 2, 10117 Berlin, Germany

E-Mail: hergo@charite.de

Received: 25th September 2006 / Published: 31st August 2007

Abstract

The dynamic behaviour of metabolic networks is determined by the kinetic properties and the cellular levels of the enzymes and transporters involved. Changes in the concentrations of enzymes can be assessed by proteomics measurements or – more indirectly – by gene expression analyses. However, a straightforward interpretation of such data with respect to metabolic functions of the cell is difficult as a simple correlation between changes of enzyme levels and changes of fluxes in a metabolic network does not exist. Here we outline a theoretical concept to exploit information on changes of enzyme concentrations for predicting changes of stationary fluxes and this way to characterize changes in the functional status of cells or tissues. The basis of our concept is a novel variant of flux-balance analysis which we call *MinMode-decomposition*. The basic idea of this concept is to approximate flux distributions in metabolic networks as linear combinations of functionally motivated minimal flux modes (*MinModes*). They are defined as minimal flux modes supporting a unit flux through only one of the target reactions of the network. This theoretical concept will be applied to metabolic networks of bacteria (*Methylobacterium extorquens*) and human red blood. Based on simulated data we demonstrate that a good prediction of observed flux changes can be achieved if the decomposition of flux changes into MinModes is performed such that a maximal correlation with observed changes in enzyme activities is accomplished.

Introduction

All cellular functions are ultimately linked to the presence of metabolites (such as proteins, nucleotides, fatty acids, phospholipids etc.) produced by the so-called metabolic network comprising thousands of enzyme-catalysed chemical reactions and carrier-mediated transport processes. The rate (herein called flux) through a given process, i.e. the amount of material chemically converted or transported per time unit, is controlled by various regulatory mechanisms. The set of all fluxes in a metabolic network is called flux distribution. The flux distribution may dramatically change with changing functional status of the cell (e.g. turning on the glycolytic flux when switching from the resting to the working muscle). It is an important goal of quantitative biochemistry to determine the flux distribution that determines the functionality of the cell. Such studies may help to reveal the relative importance of a specific enzyme and to predict the impact on the flux distribution if the enzyme is not active, e.g. due to a mutation or due to the administration of an enzyme inhibitor. The latter aspect is of central importance for the development of novel drugs interfering with the cellular metabolism.

Experimental determination of metabolic flux rates by means of tracer studies is time-consuming and tedious. Therefore, various mathematical concepts have been developed to analyse the full spectrum of flux modes possible in a metabolic network (structural analysis) or to predict flux distributions (semi-quantitative analysis). The common basis for all these concepts is the stoichiometric matrix $S = (S_{ij})$ representing the number of molecules of metabolite (i) formed or utilized in reaction (j). The stoichiometric matrix **S** is a $m \times n$ matrix where m corresponds to the number of metabolites and n is the number of reactions for which at least one catalysing enzyme is available in a given cell type [1]. The presence of a particular reaction can be evidenced by biochemical studies or – with some precaution – deduced from proteomic or genomic data [2–5].

Most modelling approaches assume the spatial distribution of metabolites to be homogeneous so that the kinetic behaviour of the network can be described by a system of ordinary differential equation systems,

$$\frac{d[X_i]}{dt} = \sum_{j=1}^{m} S_{ij}\ v_j \qquad (1)$$

$[X_i]$ is the concentration of the i-th metabolite ($i = 1,2,...,m$) and v_j denotes the flux through the j-th reaction ($j = 1,2,...,n$). The fluxes v_j constitute the so-called flux vector $\mathbf{v} = (v_j)$, in this paper referred to as a flux distribution. For quasi-stationary metabolic states where changes of the external conditions of the cell (e.g. changes of substrate concentrations or hormonal effectors) are slow compared with the characteristic time-dependent response of the intra-cellular metabolism, a further simplification in the mathematical description of the network can be achieved by calculating the stationary solution of equation system (1),

$$\sum_{j=1}^{m} S_{ij}\, v_j = 0 \tag{2}$$

The most advanced and satisfactory modelling approach is to solve the equation systems (1) or (2) with explicit flux vector **v** composed of rate equations relating the fluxes to the concentrations of the metabolites and external signals. However, such a straightforward modelling approach requires detailed knowledge of the kinetic properties of each participating enzyme. Even for the relatively simple metabolic network of bacteria as, for example, *Escherichia coli* this information is currently only available for the minority of enzymes involved. Therefore, computational studies of whole-cell metabolic networks demand alternative mathematical concepts. Existing non-kinetic concepts can be subdivided into two categories: Structural methods of network analysis and flux-balance analysis (FBA). Structural network analysis aims at exploring the full set of flux modes that may exist in a network with known stoichiometry. Simply speaking, these methods provide an overview of the many routes along which a given metabolite can be converted into another metabolite. Various algebraic concepts have been developed to define a basic set of "fundamental" flux modes which linearly combine to all possible flux modes in the network [1, 6, 7]. The definitions of such basic flux modes differ in the way that the reversible reactions are partitioned in forward and backward rates [8]. The two most prominent "fundamental" sets of flux modes are the so-called elementary modes [7] and the extremal pathways [6]. They have been used to re-define metabolic pathways [9], to check the robustness of the metabolic network against enzyme knock-outs [10, 11], the identification of thermodynamically infeasible cycles [12, 13] and the determination of so-called minimal cut sets, i.e. minimal sets of enzymes that have to be knocked out in order to completely abolish the flux through a given set of reactions [14]. The main obstacle in the application of structural analyses to large networks is the enormous number of possible flux modes arising from the combinatorial multiplicity with which single reactions can be composed to a longer route. For example, for a simplified metabolic network of *Escherichia coli* consisting of 106 reactions and producing five different end products from one initial substrate, 27100 elementary modes exist. Addition of three further initial substrates increases the number of elementary modes to 507632 [15]. This combinatorial explosion [16] implies that basic mode sets for networks with several hundreds of reactions cannot be calculated on commonly available computers. A lot of effort has been put into the development of faster algorithms to reduce computation time [15]. However, even if better algorithms allow the computation of fundamental modes for larger networks, inspection of all these modes and evaluation of their physiological significance remains extremely difficult. Attempts have also been undertaken to decrease the size of fundamental mode sets by incorporating additional constraints, for example, transcriptional regulation and environmental conditions [17] or kinetic and physiological feasibility [18, 19] into the computing algorithm. However, the formulation of such constraints requires profound *a priori* knowledge of physiological and regulatory details.

Besides structural network analyses, the concept of flux-balance analysis (FBA) has become a widely used method to estimate unknown fluxes in metabolic networks [20]. FBA postulates an objective function relating the flux distribution to a specific physiological function of the cell and to determine a flux distribution that optimizes this objective function. The idea behind this approach is that cells are capable of setting up an optimal flux distribution to produce a functionally relevant metabolic output. Most applications of FBA have used an objective function that considers only a single cardinal function of the cell as, for example, the accumulation of cell material (biomass) during the S-phase of the cell cycle. However, even primitive cells have to generate a metabolic output that simultaneously meets several functional demands. To overcome the restriction of FBA to monofunctional objective functions we have recently proposed the principle of flux minimization [21 – 23]. According to this principle, functionally relevant target fluxes, i.e. fluxes generating metabolites that are either used as building blocks for the synthesis of complex biomolecules or exported, should be accomplished with a minimal sum of internal network fluxes.

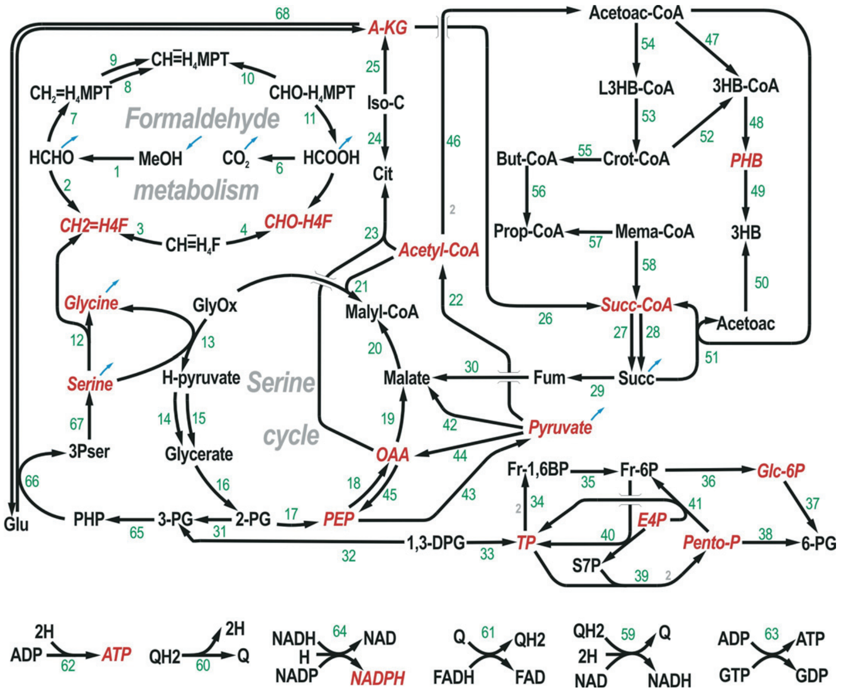
This work presents some ideas how the concept of structural network analyses can be unified with the concept of flux-minimization. As exemplary metabolic networks we will consider the central metabolism of *Methylobacterium extorquens* and the redox- and energy metabolism of human erythrocytes: Both networks have already been studied in previous work of our group [24, 25]. In the first part of this paper we show, only very few elementary modes are actually needed to decompose flux distributions calculated by flux-balance methods. However, the decomposition of the FBA solution into elementary modes is not unique and not all of the fundamental modes used in this decomposition allow for a clear physiological interpretation. Therefore, in the second part of the paper we propose a new type of fundamental modes which we call minimal flux modes (or short: *MinModes*). They are defined as minimal flux modes required to maintain a unit flux through a single target reaction of the network. According to this definition, there are only as many *MinModes* as there are functionally relevant target fluxes (17 in the network of Methylobacterium and 4 in the network of erythrocytes). Although MinModes do not form a basis in strict mathematical sense the examples considered in this article suggest that they can be linearly combined to provide a good approximation of a given flux distribution. The striking advantage of such a representation in terms of MinModes is that the coefficients used for the linear combination have a clear physiological meaning: they represent the metabolic output of the network in a given steady state and thus can be used as a measure for the functionality of the cell. Parts of these results have already been published [26].

In the third part of this article we study the relationship between changes of enzyme levels and changes in the flux distribution. As consistent data sets encompassing changes in the expression levels of metabolic enzymes and measured flux changes in the network are not available yet, we settle this study on simulated data. We take the change of the v_{max} value of an enzyme as a measure for the change of its concentration. Subsequently we use the validated kinetic model for the erythrocyte network to simulate changes of the steady state fluxes elicited by changes in the v_{max} values of the participating enzymes. As expected, there is no simple correlation between these two changes. However, using the decomposition of flux changes into MinModes as a side constraint we show that the prediction of flux changes from changes of enzyme levels can be greatly improved.

Exemplary Networks

The theoretical concepts considered in this paper will be applied to metabolic networks of two different cell types: The bacterium *Methylobacterium extorquens* AM1 (in the following referred to a B-network) and the human erythrocyte (in the following referred to as E-network). The metabolic scheme for the B-network was originally published by Van Dien [24, 25]. The scheme used in this paper contains some corrections which we have made in the light of recent findings [27, 28]. The B-network comprises 68 internal chemical reactions (43 of which are considered reversible), 8 exchange processes with the extracellular medium (for methanol, succinate, carbon dioxide, formate, formaldehyde, pyruvate, glycine, and serine) and 17 target reactions producing those metabolites required for biomass synthesis (see Table 1 for more details). While the model includes methanol, succinate, and pyruvate as alternative substrates, in all calculations methanol was considered the only available carbon source.

The reaction scheme for the E-model was originally published by Heinrich, Schuster and Holzhütter [29, 30]. It comprises 22 internal reactions, 4 exchange processes (for glucose, phosphate, pyruvate and lactate) and 4 target reactions delivering those metabolites which are essential for the integrity and functionality of the erythrocyte (2,3-bisphosphoglycerate, ATP, glutathione and phosphoribosyl pyrophosphate). For the E-network a detailed kinetic model is available [30] which takes into account all known kinetic properties of the participating enzymes. This kinetic model allows the computation of reliable stationary and time-dependent metabolic states which can be compared with results obtained by the FBA outlined in this paper. The reaction schemes for both networks are shown in Fig. 1. The involved reactions are explained in the legend of this figure.



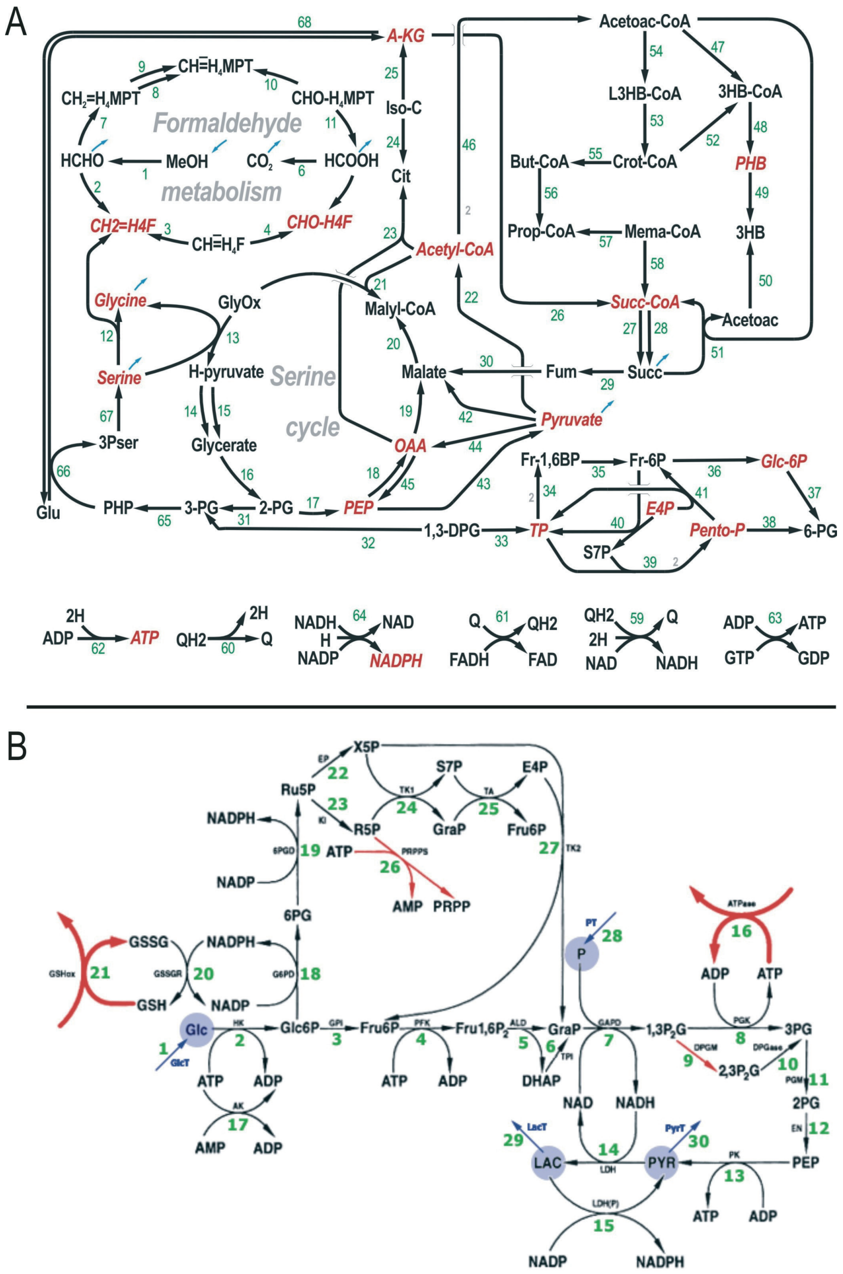
$$\mathrm{si}\left(MM,MM'\right)=\frac{1}{N_r}\sum_{i=1}^{N_r}\rho\left(MM_i,MM_i'\right),\quad \rho\left(MM_i,MM_i'\right)=\begin{cases}1\text{ if }\mathrm{sgn}(MM_i)=\mathrm{sgn}(MM_i')\\0\text{ else}\end{cases}$$

Figure 1. Metabolic networks considered in this article.

A. B-network: Reaction scheme of the central metabolism of *Methylobacterium extorquens*. The reaction arrows point in the direction of the net reaction under standard conditions. Compound names in red italic indicate utilization or generation of the corresponding metabolite during biomass production, blue arrows indicate exchange fluxes with the external environment. Cofactors have been dropped for better readability. The complete reaction scheme is shown in Table1A. The scheme is based on information outlined in [21] and derived from the KEGG data base (http://www.genome.ad.jp/kegg/).

Reactions/Enzymes:

1-methanol dehydrogenase (1.1.1.244), 2-not catalysed, 3-methylene H_4F dehydrogenase (MtdA)(1.5.1.5), 4-methenyl H_4F cyclohydrolase (3.5.4.9), 5-formyl H_4F synthetase (6.3.4.3), 6-formate dehydrogenase (1.2.1.2), 7-formaldehyde-activating enzyme, 8-methylene H_4MPT dehydrogenase (MtdB), 9-methylene H_4MPT dehydrogenase (MtdA) n/a, 10-methenyl H_4MPT cyclohydrolase (3.5.4.27), 11-formyl MFR:H_4MPT formyltransferase (1.2.99.5), 12-formyl MFR dehydrogenase (1.2.99.5), 13-serine hydroxymethyltransferase (2.1.2.1), 14- serine-glyoxylate aminotransferase (2.6.1.45), 15-hydroxypyruvate reductase (1.1.1.81), 16-glycerate kinase (2.7.1.31), 17-enolase (4.2.1.11), 18-PEP carboxylase (4.1.1.31), 19-malate dehydrogenase (1.1.1.37), 20-malate thiokinase (6.2.1.9), 21-malyl-CoA lyase (4.1.3.24), 22-pyruvate dehydrogenase (1.2.4.1), 23-citrate synthase (2.3.3.1), 24-aconitase (4.2.1.3), 25-isocitrate dehydrogenase (1.1.1.42), 26-a-KG dehydrogenase (1.2.1.52), 27-succinyl-CoA synthetase (6.2.1.4), 28-succinyl-CoA hydrolase (3.1.2.3), 29-succinate dehydrogenase (1.3.5.1), 30-fumarase (4.2.1.2), 31-phosphoglycerate mutase (5.4.2.1) ,32-phosphoglycerate kinase (2.7.2.3), 33-glyceraldehyde-3-P dehydrogenase (1.2.1.12), 34-aldolase (4.1.2.13), 35-fructose-1,6-bisphosphatase (3.1.3.11), 36-phosphoglucose isomerase (5.3.1.9), 37-glucose-6-phosphate dehydrogenase (1.1.1.49), 38 – 6-phosphogluconate dehydrogenase (1.1.1.44), 39 transketolase (2.2.1.1), 40-transaldolase (2.2.1.2), 41- transketolase (2.2.1.1), 42-malic enzyme (1.1.1.38), 43-pyruvate kinase (2.7.1.40), 44-pyruvate carboxylase (6.4.1.1), 45-PEP carboxykinase (4.1.1.32), 46- b-ketothiolase (2.3.1.16), 47- acetoacetyl-CoA reductase (NADPH) (1.1.1.36), 48- PHB synthase (2.3.1._), 49- PHB depolymerase (3.1.1.75), 50 b-hydroxybutyrate dehydrogenase (1.1.1.30), 51-acetoacetate-succinyl-CoA transferase (2.8.3.5), 52-d-crotonase (4.2.1.17), 53 l-crotonase (4.2.1.17), 54-acetoacetyl-CoA reductase (NADH) (1.1.1.35), 55-crotonyl-CoA reductase (1.3.1.8), 56-unknown pathway, 57-propionyl-CoA carboxylase (6.4.1.3), 58-methylmalonyl-CoA mutase (5.4.99.2), 59-NADH-quinone oxidoreductase (1.6.99.5), 60-cytochrome oxidase (1.10.2.2), 61- ubiquinone oxidoreductase (1.5.5.1), 62-ATPase, 63-NDP kinase (2.7.4.6), 64-transhydrogenase (1.6.1.2), 65 – 3-phosphoglycerate dehydrogenase (1.1.1.95), 66-phosphoserine transaminase (2.6.1.52), 67-phosphoserine phosphatase (3.1.3.3), 68-glutamate dehydrogenase (1.4.1.4)



B. E-network: Reaction scheme of the energy- and redox metabolism of human erythrocytes.

Reaction/Enzymes:

1-glucose transporter GlcT, 2-hexokinase HK (2.7.1.1), 3-phosphohexose isomerase GPI (5.3.1.9), 4-phosphofructokinase PFK (2.7.1.11), 5-aldolase ALD (4.1.2.13), 6-triosephosphate isomerase TPI (5.3.1.1), 7-triosephosphate dehydrogenase (NAD) GAPDH (1.2.1.12), 8-phosphoglycerate kinase PGK (2.7.2.3), 9-bisphosphoglycerate mutase DPGM (5.4.2.4), 10-bisphosphoglycerate phosphatase DPGase (3.1.3.13), 11-phosphoglycerate mutase PGM (5.4.2.1), 12-enolase EN (4.2.1.11), 13-pyruvate kinase PK (2.7.1.40), 14-lactate dehydrogenase LDH (NADH) (1.1.1.28), 15-lactate dehydrogenase LDH (NADPH) (1.1.1.28), 16-ATPase (total) ATPase, 17-myokinase (adenylate kinase) AK (2.7.4.3), 18-glucose-6-phosphate dehydrogenase G6PD (1.1.1.49), 19-phosphogluconate dehydrogenase 6PGD (1.1.1.44), 20-glutathione reductase GSSGR (1.8.1.7), 21-glutathione oxidation (total) GSHox, 22-phosphoribulose epimerase EP (5.1.3.1), 23-ribose phosphate isomerase KI (5.3.1.6), 24-transketolase (1) TK1 (2.2.1.1), 25-transaldolase TA (2.2.1.2), 26-phosphoribosylpyrophosphate synthetase PRPPS (2.7.6.1), 27-transketolase (2) TK2 (2.2.1.1), 28-phosphate transporter PT, 29- lactate exchange LacT, 30- pyruvate exchange PyrT

Computational Methods

Calculation of flux-minimized steady-state flux distributions

The computation of flux modes is based on flux balance analysis (FBA). The core of this method is the optimization of an objective function which relates the flux distribution to cellular functions. According to the principle of flux minimization [31] the objective function to be minimized is chosen as:

$$\Phi = \sum_{j} \mathrm{pos}\left(v_j\right) + K_j \ \mathrm{neg}\left(v_j\right) \tag{3}$$

where the sum runs over all fluxes in the network and K_j denotes the equilibrium constant for the j-th reaction. The real functions pos(x) and neg(x) return the absolute value of the argument x if $x \geq 0$ and $x \leq 0$, respectively, and otherwise 0. The functional state of the cell is defined by fixing non-zero fluxes through so-called 'target reactions' which together with the steady-state represent constraints of the minimization problem. The constrained flux minimization problem is solved using the software package CPLEX [32]. Details of the computational protocol have been described elsewhere [31].

The *in vivo* state of the B-network is determined by the following values of the fluxes through the 17 target reactions involved in the production of biomass:
$V_{77}= 13.4$ (glycine), $v_{82}= 1.96$ ($CH_2=H_4F$), $v_{86}= 11.1$ (pep), $v_{92}= 5.09$ (ery4P), $v_{89}= 1.92$ (tp), $v_{85}= 7.24$ (serine), $v_{83}= 11.0$ (CHO-H_4F), $v_{93}= 92.9$ (phb), $v_{90}= 16.4$ (glc6P), $v_{88}= 17.1$ (akg), $v_{84}= 41.8$ (pyruvate), $v_{81}= 53.5$ (acetylCoA), $v_{87}= 41.1$ (oaa), $v_{80}= 6.8$ (succCoA), $v_{78}= 585.3$ (atp), $v_{91}= 10.4$ (pentoP), $v_{79}= 235.13$ (nadph)
The unit of fluxes is moles precursor metabolite per 1000 g atoms C in biomass. Release of NADH during biomass production was included in the target flux for NADPH. The relations between the target flux, i. e. the stoichiometry with which the 17 precursor metabolites enter the biomass, have been determined by Van Dien [24].

The *in vivo* state of the E-network is determined by the following values of the fluxes through the 4 target reactions:
$v_9 = 0.49$ (2,3DPG), $v_{16} = 2.38$ (ATP), $v_{21} = 0.093$ (GSH), $v_{26} = 0.026$ (PRPP)
The unit of fluxes is mM/h.
In the following we will refer to the solution of the minimization problem (3) fulfilling the steady-state conditions (2) and providing the above values of target fluxes as the *global flux minimum.*

Decomposition of the global flux minimum into elementary modes

For the two exemplary networks, elementary modes and the convex basis of elementary modes were computed using a recent version of the software tool FluxAnalyzer [33]. Decomposition of *global flux minimum* into the convex basis was performed by solving the linear program

$$\mathbf{e}\,\mathbf{C} = \mathbf{v}_g \ , \ \|\mathbf{e}\| \text{ minimal} \tag{4}$$

where **C** is the convex basis of elementary modes written as a matrix, $\mathbf{v}_g$ is the global flux minimum and **e** is a vector of non-negative real numbers.

Definition of minimal flux modes (MinModes)

Besides the *global flux minimum* defined as the optimal flux distribution in the network with all target fluxes kept at pre-defined non-zero values, we computed special flux-minimized steady-states by putting the value of only one of the target fluxes to either unity (in the used flux units). The resulting minimized flux modes we call minimal flux modes (short: MinModes). They are defined as follows:

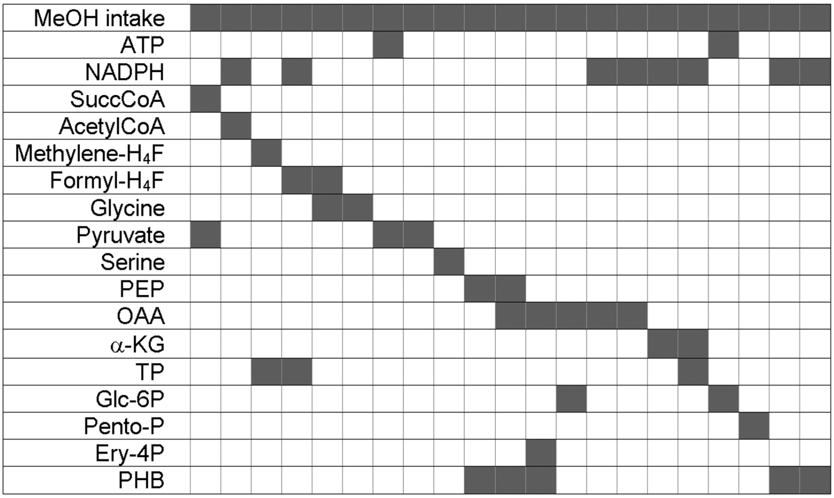
> A MinMode is a minimal (according to the flux minimization principle) steady state flux distribution that accomplishes a unit flux through one of the (independent) target reactions whilst the fluxes through the other target reactions are zero.

This definition is more rigorous than that given in [26], because it presumes the target fluxes to be independent from each other. Independency of target fluxes means the existence of a flux-minimized solution that accomplishes a non-zero flux through the chosen target flux without the necessity to have non-zero fluxes through other target reactions. This condition may be not always fulfilled. For example, if two metabolites (say A and B) are produced in one and the same reaction, then under steady state conditions utilization of A necessarily entails utilization of B. Hence, the first step towards the computation of MinModes requires to identify clusters of intrinsically coupled target reactions. In the following we will assume that the definition of 'target fluxes' in the network eventually includes clusters of coupled output reactions. For the two networks studied in this paper, the output fluxes are independent, i. e. there are 17 of such MinModes for the B-network and 4 MinModes for the E-network.

Results

Elementary modes of the network of Methylobacterium

The model of central metabolism of *Methylobacterium extorquens* shown in Fig. 1 was subjected to elementary mode analysis. The network model consists of 93 reactions whereby 43 reactions are considered irreversible. The complete set of elementary modes for this model is too large to be computable by means of the program *FluxAnalyzer* [33] using a PC equipped with a memory capacity of 768 MB. The enormous amount of elementary modes network can be envisaged by noting that there are 450251 elementary modes if the computation is simplified by setting 5 of the 17 target fluxes to zero. However, for the representation of an arbitrary flux distribution knowledge of the convex basis of elementary modes is sufficient [34]. A convex basis of elementary modes was computable for the complete B-network. It consists of 7033 elementary modes. Intriguingly, only 21 elementary flux modes (out of 7033) of the convex basis were actually needed (i. e. had non-zero coefficients in the linear representation) for the decomposition of the *global flux minimum*. Regarding the physiological interpretation of these 21 elementary flux modes further analysis showed that most of them contain more than one non-zero target flux (see Fig. 2). Moreover, the choice of basic elementary modes is not unambiguous. This makes it difficult to assign a specific output of the network to these elementary flux modes.



MeOH intake	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
ATP							■											■			
NADPH		■		■										■	■	■	■			■	■
SuccCoA	■																				
AcetylCoA		■																			
Methylene-H_4F			■																		
Formyl-H_4F				■	■																
Glycine					■	■															
Pyruvate	■						■	■													
Serine									■												
PEP										■	■										
OAA											■	■	■	■	■						
α-KG																■	■				
TP			■	■													■				
Glc-6P													■					■			
Pento-P																			■		
Ery-4P												■									
PHB										■	■	■								■	■

Figure 2. Occurrence of non-zero input and output fluxes in the 21 elementary modes required for the decomposition of the global flux minimum for the B-network. Each column represents an elementary mode, each row represents an input/output flux. Table fields shaded in grey indicate a non-zero value of the respective flux.

Elementary modes of the network of the Erythrocyte

The model of central metabolism of the human Erythrocyte has 20 elementary modes. Only 4 elementary modes were actually needed (i.e. had non-zero coefficients in the linear representation) for the decomposition of the global flux minimum where each of these modes exactly corresponds to a singular target flux. Thus, for this simple model the difficulties mentioned for the B-network do not occur.

Calculation of MinModes

According to the definition given above, MinModes are flux-minimized states of the network where only one of the (independent) target reactions carries a unit flux whereas the flux through all other target reactions is put to zero. As an important property, each of these modes is associated with a specific output of the network. For the B-network (*Methylobacterium extorquens)* each MinMode represents the minimal flux distribution required for the synthesis of a single biomass precursor metabolite. For the E-network (erythrocyte) the target metabolites ATP and GSH are essential for cellular integrity (maintenance of the intracellular ionic milieu by the ATP-driven Na-K-pump, protection against secondary reaction products of radicals via GSH), the two other target metabolites 2,3DPG and PRPP are indispensable for oxygen binding to haemoglobin and the salvage of adenine nucleotides. The MinModes for the two networks are listed in Table 1.

Inspecting the MinModes of the B-network, we note that the amount of methanol consumed in each of the MinModes can be split into a carbon providing and an energy (ATP and redox equivalents) providing part. As an example, the MinMode for the production of the biomass precursor glycine is plotted in Fig. 3.

Table 1. Model scheme and the corresponding minimal flux modes (MinModes) of the B-network. *cmm* and *gfm* denote the *MinMode composition* and the *global flux minimum*, respectively.

#	reaction	k_i	acetylCoA	aKG	ATP	CH_2=H4F	Ery4P	CHO-H4F	Glc6P	glycine	NADPH	OAA	PentoP	PEP	PHB	pyruvate	serine	Succ	TP	*cmm*	*gfm*
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**		
	Formaldehyde metabolism																				
1	MeOH + Q $\rightarrow$ HCHO + QH_2	100000	2	5	1	1	5	1	6.6	2.8	0.4	4	5.8	3.55	4	3.36	3.6	1	4.09	32.63	23.59
2	HCHO + $H_4F \rightarrow CH_2{=}H_4F$	100000	1	3	0	1	2	1	2	1	0	2	2	2	2	2	2	0	2	10.23	10.23
3	$CH{\equiv}H_4F$ + NADPH $\leftrightarrow CH_2{=}H_4F$ + NADP	7.14	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	-0.19	-0.19
4	$CH{\equiv}H_4F \leftrightarrow$ CHO-H_4F	2.3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0.19	0.19
5	HCOOH + H_4F + ATP $\leftrightarrow$ CHO-H_4F + ADP	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	HCOOH + NAD $\rightarrow CO_2$ + NADH	420	1	2	0	0	3	0	4.6	1.8	0.4	2	3.8	1.55	2	1.36	1.6	0	2.09	12.09	13.37
7	HCHO + $H_4MPT \rightarrow CH_2{=}H_4MPT$	100000	1	2	0	0	3	0	4.6	1.8	0.4	2	3.8	1.55	2	1.36	1.6	0	2.09	12.09	13.37
8	$CH_2{=}H_4MPT$ + NAD $\rightarrow CH{\equiv}H_4MPT$ + NADH	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1.62	0
9	$CH_2{=}H_4MPT$ + NADP $\rightarrow CH{\equiv}H_4MPT$ + NADPH	1	1	2	0	0	3	0	4.6	1.8	0.4	2	3.8	1.55	1	1.36	1.6	0	2.09	10.47	13.37
10	CHO-$H_4MPT \leftrightarrow CH{\equiv}H_4MPT$	1	-1	-2	0	0	-3	0	-4.6	-1.8	-0.4	-2	-3.8	-1.55	-2	-1.36	-1.6	0	-2.09	-12.09	-13.37
11	HCOOH + $H_4MPT \leftrightarrow$ CHO-H_4MPT	100000	-1	-2	0	0	-3	0	-4.6	-1.8	-0.4	-2	-3.8	-1.55	-2	-1.36	-1.6	0	-2.09	-12.09	-13.37
	Serine cycle																				
12	Serine + $H_4F \leftrightarrow$ Glycine + $CH_2{=}H_4F$	1022	-1	-3	0	0	-2	0	-2	-1	0	-2	-2	-2	-2	-2	-2	0	-2	-10	-10
13	Serine + GlyOx $\leftrightarrow$ Glycine + H-Pyruvate	1	1	3	0	0	2	0	2	2	0	2	2	2	2	2	2	0	2	10.23	10.23
14	H-Pyruvate + NADH $\rightarrow$ Glycerate + NAD	1	0	1	0	0	2	-1	4.4	2.2	0.6	1	3.2	2	2	2	2.4	0	2	11.26	10.23
15	H-Pyruvate + NADPH $\rightarrow$ Glycerate + NADP	1	1	2	0	0	0	1	-2.4	-0.2	-0.6	1	-1.2	0	0	0	-0.4	0	0	-1.02	0
16	Glycerate + ATP $\rightarrow$ 2-PG + ADP	100000	1	3	0	0	2	0	2	2	0	2	2	2	2	2	2	0	2	10.23	10.23
17	PEP $\leftrightarrow$ 2-PG	3	-1	-3	0	0	-1	0	-1	-1	0	-2	-1	-2	-2	-2	-1	0	-1	-9.28	-9.28
18	PEP + $CO_2 \rightarrow$ OAA	1	1	1	0	1	0	3	0	0	0	2	0	0	0	0	0	0	0	3.27	1.01
19	OAA + NADH $\rightarrow$ Malate + NAD	6260	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1.65	0
20	Malate + CoA + ATP $\rightarrow$ Malyl-CoA + ADP	100000	1	3	0	0	2	0	2	2	0	2	2	2	2	2	2	0	2	10.23	10.23
21	Acetyl-CoA + GlyOx $\leftrightarrow$ Malyl-CoA	345	-1	-3	0	0	-2	0	-2	-2	0	-2	-2	-2	-2	-2	-2	0	-2	-10.23	-10.23
	Citrate cycle																				
22	Pyruvate + CoA + NAD $\rightarrow$ Acetyl-CoA + NADH + CO_2	100000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	Acetyl-CoA + OAA $\rightarrow$ Cit + CoA	100000	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3	0.3
24	Iso-C $\leftrightarrow$ Cit	14.4	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.3	-0.3
25	α-KG + CO_2 + NADPH $\leftrightarrow$ Iso-C + NADP	1.3	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.3	-0.3
26	α-KG + CoA + NAD $\rightarrow$ Succ-CoA + CO_2 + NADH	100000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	Succ + CoA + GTP $\leftrightarrow$ Succ-CoA + GDP	1.68	0	-1	0	0	-1	0	-1	-1	0	0	-1	-1	0	-1	-1	1	-1	-2.05	-2.77
28	Succ-CoA $\rightarrow$ Succ + CoA	100000	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0.72	0
29	Succ + FAD $\leftrightarrow$ Fum + FADH	1	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	0	1	2.88	2.88
30	Fum $\leftrightarrow$ Malate	4.86	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	0	1	2.88	2.88
	Gluconeogenesis and pentose phosphate pathway																				
31	2-PG $\leftrightarrow$ 3-PG	7.3	0	0	0	0	1	0	1	1	0	0	1	0	0	0	1	0	1	0.95	0.95
32	1,3-DPG + ADP $\leftrightarrow$ 3-PG + ATP	3226	0	0	0	0	-1	0	-1	0	0	0	-1	0	0	0	0	0	-1	-0.59	-0.59
33	1,3-DPG + NADH $\leftrightarrow$ TP + NAD	2.77	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0.59	0.59
34	2 TP $\rightarrow$ Fru-1,6-BP	5555	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	Fru-1,6-BP $\leftrightarrow$ Fru-6-P	174	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	Fru-6-P $\leftrightarrow$ Glc-6-P	2.22	0	0	0	0	-1	0	-2	0	0	0	-2	0	0	0	0	0	0	-1.02	-1.02
37	Glc-6-P + NADP $\leftrightarrow$ 6-PG + NADPH	2.08	0	0	0	0	-1	0	-3	0	0	0	-2	0	0	0	0	0	0	-1.31	-1.31
38	Pento-P + CO_2 + NADPH $\leftrightarrow$ 6-PG + NADP	13.5	0	0	0	0	1	0	3	0	0	0	2	0	0	0	0	0	0	1.31	1.31
39	TP + S7P $\leftrightarrow$ 2 Pento-P	1.5	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0.47	0.47
40	TP + S7P $\leftrightarrow$ Ery-4-P + Fru-6-P	1	0	0	0	0	0	0	-1	0	0	0	-1	0	0	0	0	0	0	-0.47	-0.47
41	Ery-4-P + Pento-P $\leftrightarrow$ TP + Fru-6-P	10	0	0	0	0	-1	0	-1	0	0	0	-1	0	0	0	0	0	0	-0.56	-0.56
42	Malate + NAD $\leftrightarrow$ Pyruvate + CO_2 + NADH	1	0	-2	0	0	-1	0	-1	-1	0	0	-1	-1	-2	-1	-1	0	-1	-5.7	-7.35
43	PEP + ADP $\rightarrow$ Pyruvate + ATP	18000	0	2	0	0	1	0	1	1	0	0	1	1	2	2	1	0	1	6.43	8.08
44	Pyruvate + CO_2 + ATP $\leftrightarrow$ OAA + ADP	6.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	OAA + GTP $\leftrightarrow$ PEP + GDP + CO_2	12	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0.61	0
	PHB synthesis and glyoxylate regeneration cycle																				
46	2 Acetyl-CoA $\leftrightarrow$ Acetoac-CoA + CoA	1	0	1	0	0	1	0	1	1	0	1	1	1	1	1	1	0	1	4.5	4.5
47	Acetoac-CoA + NADPH $\rightarrow$ 3HB-CoA + NADP	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1.62	1.62
48	3HB-CoA $\rightarrow$ PHB + CoA	100000	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1.62	1.62
49	PHB $\rightarrow$ 3HB	100000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	Acetoac + NADH $\leftrightarrow$ 3HB + NAD	526	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
51	Acetoac-CoA + Succ $\leftrightarrow$ Acetoac + Succ-CoA	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(continuation on the next page)

Table 1: (Continued).

#	reaction	k_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	*cmm*	*gfm*
52	Crot-CoA ↔ 3HB-CoA	5.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
53	L3HB-CoA ↔ Crot-CoA	5.55	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	0	1	2.88	2.88
54	Acetoac-CoA + NADH → L3HB-CoA + NAD	1587	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	0	1	2.88	2.88
55	Crot-CoA + NADPH ↔ But-CoA + NADP	1	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	0	1	2.88	2.88
56	But-CoA + NAD → Prop-CoA + CO_2 + NADH	100000	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	0	1	2.88	2.88
57	Mema-CoA + ADP ↔ Prop-CoA + ATP + CO_2	123	0	-1	0	0	-1	0	-1	-1	0	-1	-1	-1	0	-1	-1	0	-1	-2.88	-2.88
58	Mema-CoA ↔ Succ-CoA	18.6	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	0	1	2.88	2.88
	Energy metabolism																				
59	NADH + Q ↔ NAD + QH_2 + 2 H	1	0	-1	0	0	-1	1	-1.8	-0.4	-0.2	0	-1.4	-0.91	-1	-1.27	-0.8	0	-0.82	-4.72	-2.16
60	QH_2 ↔ Q + 2 H	1	2	5	1	1	5	2	5.8	3.4	0.2	5	5.4	3.64	3	3.09	3.8	1	4.27	30.78	24.31
61	FADH-S + Q ↔ FAD-S + QH_2	1	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	0	1	2.88	2.88
62	ADP + 2 H ↔ ATP	1	2	4	1	1	4	3	4	3	0	5	4	3	2	2	3	1	4	26.26	23.29
63	GDP + ATP ↔ GTP + ADP	1.04	0	-1	0	1	-1	3	-1	-1	0	0	-1	-1	0	-1	-1	1	-1	-1.44	-2.77
64	NADH + NADP + H → NADPH + NAD	1	0	0	0	0	0	0	0	0	0	0	0	-0.55	0	-0.36	0	0	-1.09	-0.41	-2.28
	Serine biosynthesis																				
65	PHP + NADH ↔ 3-PG + NAD	10000	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	-1	0	0	-0.36	-0.36
66	α-KG + 3-Pser ↔ PHP + Glu	6.66	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	-1	0	0	-0.36	-0.36
67	3-Pser ↔ Serine	673	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0.36	0.36
68	α-KG + NADPH → Glu + NADP	100000	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0.36	0.36
	Exchange reactions																				
69	CO_2 → CO_2(out)	1	0	0	0	0	1	0	0.6	0.8	0.4	0	0.8	0.55	0	0.36	0.6	0	1.09	4.77	6.05
70	CHOOH → CHOOH(out)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
71	Glycine → Glycine(out)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
72	HCHO → HCHO(out)	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	10.31	0
73	Pyruvate → Pyruvate(out)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
74	Serine → Serine(out)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
75	MeOH(out) → MeOH	1	2	5	1	1	5	1	6.6	2.8	0.4	4	5.8	3.55	4	3.36	3.6	1	4.09	32.63	23.59
76	Succ → Succ(out)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Additional reactions for biomass synthesis																				
77	glycine → glycine_biomass	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0.23	0.23
78	ATP → ATP_biomass + ADP	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10.2	10.2
79	NADPH → NADPH_biomass + NAD	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	4.1	4.1
80	Succ-CoA → Succ-CoA_biomass + Succ + CoA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.12	0.12
81	Acetyl-CoA → Acetyl-CoA_biomass + CoA	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.93	0.93
82	CH_2=H_4F → CH_2=H_4F_biomass + H_4F	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03	0.03
83	CHO-H_4F → CHO-H_4F_biomass + H_4F	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0.19	0.19
84	pyruvate → pyruvate_biomass	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.73	0.73
85	serine → serine_biomass	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0.13	0.13
86	PEP → PEP_biomass	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0.19	0.19
87	OAA → OAA_biomass	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0.72	0.72
88	α-KG → α-KG_biomass	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3	0.3
89	TP → TP_biomass	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.03	0.03
90	Glc-6-P → Glc-6-P_biomass	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0.29	0.29
91	Pento-P → Pento-P_biomass	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0.18	0.18
92	Ery-4-P → Ery-4-P_biomass	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0.09	0.09
93	PHB → PHB_biomass	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1.62	1.62

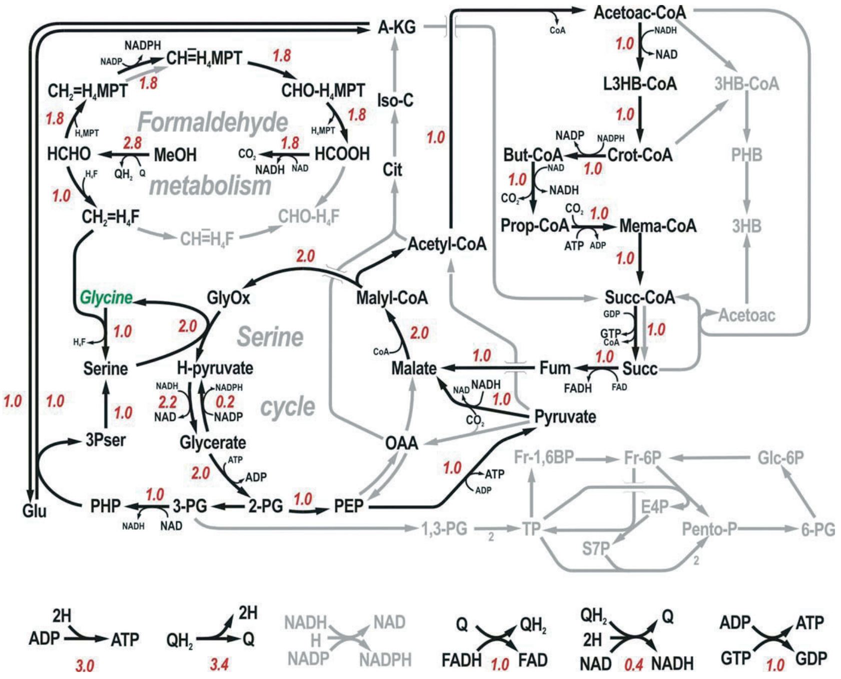


Figure 3. Minimal flux mode (*MinMode* $MM_{glycine}$) for glycine synthesis in the B-network. The MinMode was calculated by minimizing the objective function (3) where the flux v_{70} for the release of glycine to the biomass was put to unity and the other 17 target flux were put to zero. Flux values are depicted next to the reaction arrow, grey arrows indicate zero fluxes.

Here, the H_4MPT-cycle is used for the production of NADPH and NADH (utilized in other parts of the network) as well as for the complete oxidation of methanol to CO_2. The released carbon is fixed in another part of the network to provide the carbon needed for the actual synthesis of glycine. The second carbon atom is incorporated via reactions of the serine cycle. Only two of the reactions that traditionally form the citrate acid cycle, are used in this MinMode. They are catalysed by the succinate dehydrogenase (29) and the fumarase (30).

To illustrate the relative importance of the individual reactions for the functionality of the network we counted how often a non-zero flux through each of the reactions occurs in the 17 MinModes of the B-network. The corresponding statistics (Fig. 4) reveals the ubiquitous usage of reactions involved in methanol uptake and the energy metabolism. Based on the frequency with which a reaction has a non-zero flux in the various MinModes one may depict the 'backbone' of the network. For the B-network this backbone is constituted by the reactions of serine cycle and formaldehyde metabolism (H_4F and H_4MPT cycle). On the other hand, there are 15 apparently redundant reactions that have a zero-flux in all Min-Modes. This relatively high number of apparently abdicable reactions is certainly due to the fact that only those reactions are considered as targets of the network which deliver metabolites for the synthesis of cellular biomass. Because metabolites other than the biomass precursors may be relevant for cellular functionality, we have calculated an extended set of MinModes considering each of the 63 metabolites as a possible and relevant output of the network. In this set of 58 MinModes only 6 reactions turn out to be apparently dispensable (reactions: 5, 22, 26, 34, 49 and 52).

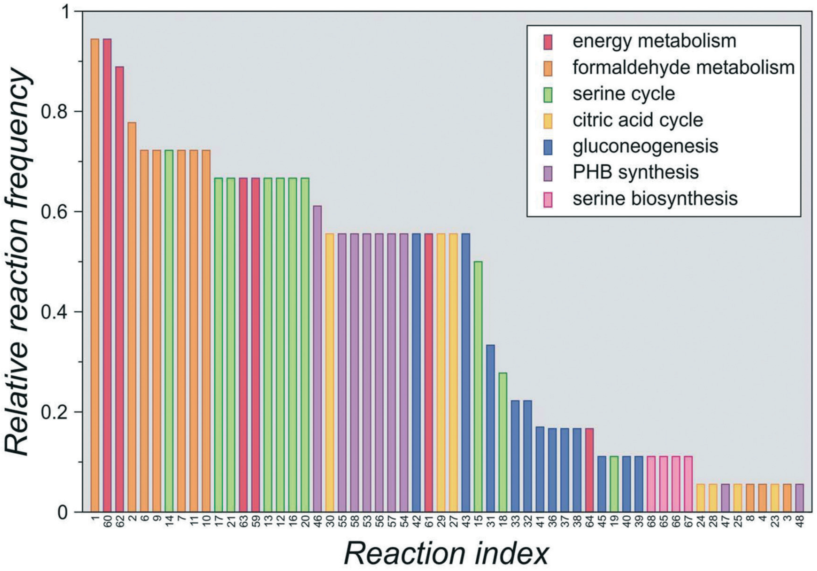


Figure 4. Frequency of non-zero fluxes in the 17 MinModes of the in the B-network

Inspecting the 4 MinModes of the E-network, which are equivalent to the 4 elementary modes required for the decomposition of the *global flux minimum*, the only difference between the two MinModes associated with the production of ATP and the production of 2,3DPG is the respective component referring to the particular target reaction.

Composition of flux distributions into MinModes

The MinModes represent canonical flux modes of the network supporting a single metabolic output (or a group of coupled outputs, see above). In real situations the flux distribution of a cell has to assure simultaneously a multitude of metabolic functions as, for example, the production of ATP, repair of DNA, elimination of reactive oxygen species or the synthesis of proteins. Because these metabolic functions must be controlled independently in the cell, it is straightforward to postulate that the total flux distribution is a linear combination of independent component fluxes rather than a globally optimized flux distribution. The concept presented here of minimal flux modes is an attempt to define those component fluxes by the following equation:

$$\mathbf{v_d} \approx \sum_i \alpha_i \, \mathbf{MM_i} \tag{5}$$

Here $\mathbf{MM_i}$ denotes the MinMode supporting the flux 1 (flux unit) through the i-th target reaction and the numerical value of the (dimensionless) coefficient α_i corresponds to the actual flux. For example, applying the linear combination (4) of MinModes to the B-network, the coefficient α_{gly} that is multiplied with $MM_{gly,}$ the MinMode for the production of the biomass precursor glycine (Gly), is put to 13.4 as the measured flux of glycine into biomass amounts to 13.4 flux units.

To check the feasibility of the MinMode decomposition we compared the resulting flux distribution $\mathbf{v_d}$ (=*MinMode decomposition*) with the *global flux minimum* and with observed flux values. For the B-network, Fig. 5 depicts the values of the individual fluxes predicted by the *MinMode decomposition* and those of the *global flux minimum*.

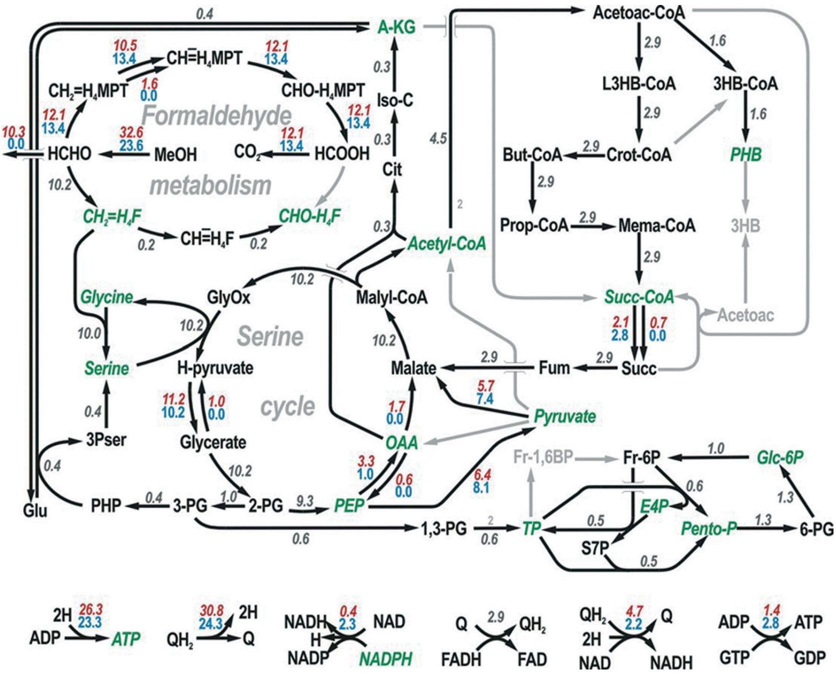
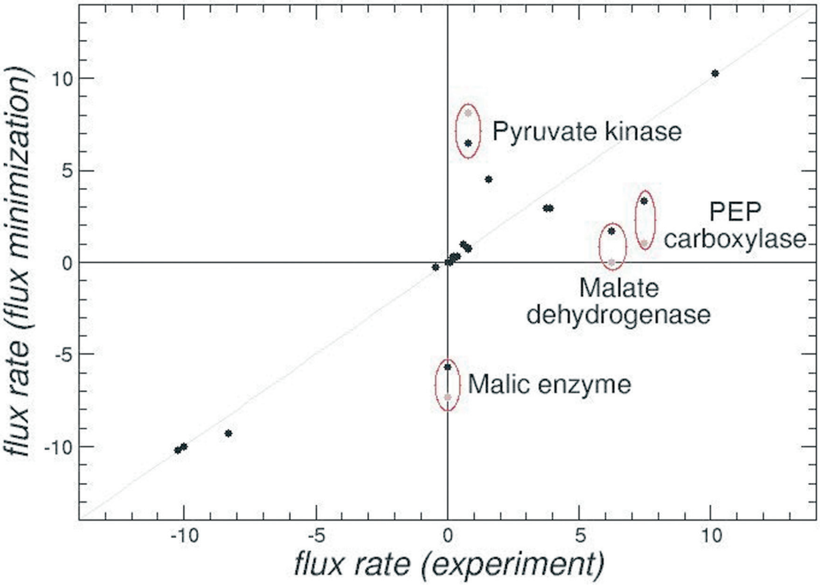


Figure 5. Comparison of flux values of the *global flux minimum* of the B-network (regular numbers) with flux values obtained by *MinMode composition* (italic numbers). Equal flux values in both approaches are displayed only once (grey/italic). All metabolites fed into biomass synthesis by target reactions are depicted with green/ italic letters.

Obviously, the larger the target flux with which a precursor metabolite enters the biomass the higher the influence of the corresponding MinMode in the combination. In the B-network, the ATP consumption for biomass formation has a large impact. In contrast to an "along the way" synthesis of ATP in the *global flux minimum*, the MinMode MM_{78} exclusively provides ATP. This makes it plausible, why the *MinMode composition* predicts somewhat higher fluxes for ATP synthesis. Minor differences in the flux pattern occurred for reactions of the H_4MPT-cycle, transhydrogenase, NDP kinase, and reactions of alternative pathways that work with different cofactors, as for example the conversion of PEP into malate (reactions 18, 19, 45 and 42, 43) or from methylene-H_4MPT into methenyl-H_4MPT (reactions 8 and 9). All flux differences can be accounted for by differences in the production, conversion or dissipation of energy. To further check the feasibility of the *MinMode composition* we compared it with measured fluxes [24]. For 18 out of 21 reactions the fluxes predicted by the *MinMode composition* are in good accordance with the experimental data (Fig. 6A). In the cases of malic enzyme, malate dehydrogenase, PEP carboxylase, and pyruvate kinase, significant differences between predictions and observations occurred. Noteworthy, the remaining flux discrepancies are even smaller than those with respect to the *global flux minimum.*



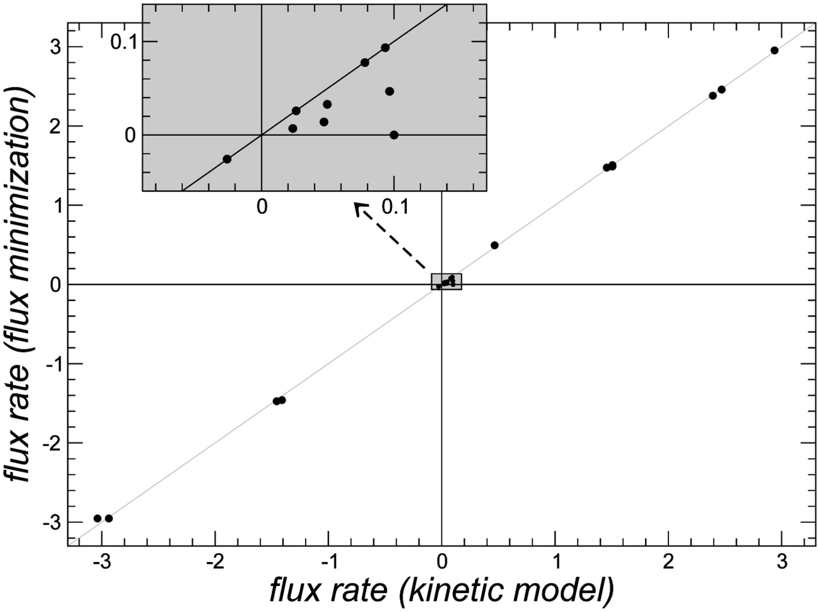


Figure 6. A.. B-network. Scattergram illustrating the correlation of experimental flux values [24] with flux values predicted by *MinMode composition.* Significant deviations between experimental flux values with flux values predicted by global optimization are displayed in grey.B. E-network. Scattergram illustrating the correlation of flux values calculated by means of a kinetic model [30] with flux values predicted by *MinMode composition*.

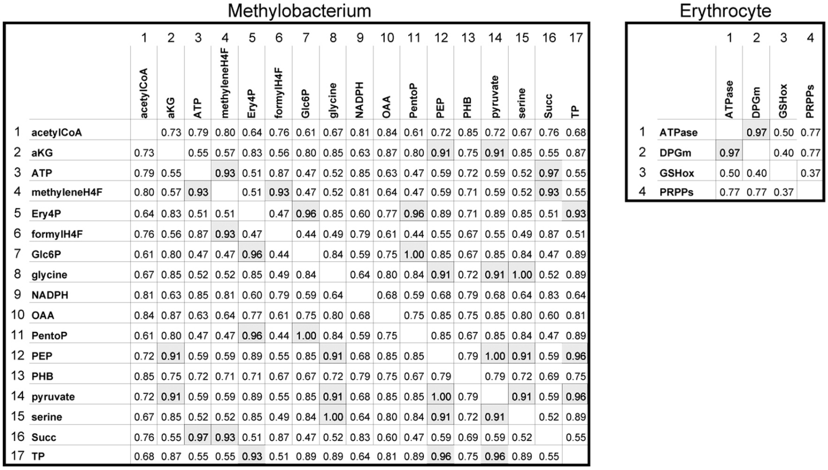
To check the feasibility of the *MinMode composition* for the E-network we compared the predicted fluxes with those calculated by means of a validated kinetic model [30]. Figure 6B reveals an almost perfect prediction for the larger fluxes and acceptable deviations for the small fluxes. Taken together, the quality of flux predictions based on the MinMode decomposition was equivalent with the quality achieved by global optimization [31].

Similarity analysis of *MinModes*

Vectors forming a basis in strict mathematical sense have to be orthogonal, i. e. independent from each other. This criterion does not hold for *MinMode*s. As demonstrated above with the *MinMode*s of the E-network, *MinMode*s belonging to different metabolic outputs, e. g. production of ATP and 2,3DPG, may be very similar. Thus it is practically impossible to conclude from observed changes of fluxes, which of the two target fluxes have changed. Therefore we represent those *MinMode*s exhibiting a strong similarity by a single *Principal MinMode* and we use the smaller set of such *Principal MinMode*s for the decomposition. To quantify the similarity of two *MinMode*s, Pearson's correlation coefficient turns out not to be a reliable measure because the components of flux vectors are not normally distributed (they contain many zero-fluxes and many tightly related flux values owing to the flux-balance conditions). We decided to quantify the similarity between two arbitrary flux modes MM and MM' by a similarity index (si) defined as the (relative) number of components having the same sign in both modes:

$$\mathrm{si}\left(\mathrm{MM},\mathrm{MM}'\right)=\frac{1}{\mathrm{N_r}}\sum_{i=1}^{\mathrm{N_r}}\rho\left(\mathrm{MM_i},\mathrm{MM}'_\mathrm{i}\right),\ \ \rho\left(\mathrm{MM_i},\mathrm{MM}'_\mathrm{i}\right)=\begin{cases}1 \text{ if } \mathrm{sgn}(\mathrm{MM_i})=\mathrm{sgn}(\mathrm{MM}'_\mathrm{i})\\ 0 \text{ else}\end{cases} \qquad (6)$$

In Equation 6, sgn(x) denotes the sign-function (sgn(x) = +1, -1 or 0 if $x > 0$, $x < 0$ or $x = 0$). The sum in Equation 6 runs over all components except those two referring to the target fluxes generating the two *MinMode*s MM and MM', respectively. The similarity indices form the (symmetric) *MinMode* similarity matrix of *MinModes*. Figure 7 shows the *MinMode* similarity matrices for the two exemplary networks. Values larger than the arbitrarily chosen threshold value of 0.9 (i. e. 90 % of the fluxes in the two *MinModes* under comparison are either both zero or point in the same direction) are marked in grey. We note that three pairs of *MinMode*s calculated for the B-network exhibit perfect similarity (si = 1). For example, the *MinMode*s associated with the production of the metabolites Glc6P and PentoseP do not differ either in the zero fluxes or in the directionality of the non-zero fluxes. Plotting the components of these two *MinMode*s against each other also reveals strong similarity (Fig. 8A). This is due to the fact that the pentose phosphates are formed in the pentose phosphate pathway which branches from glucose-6-phosphate, i. e. production of pentose phosphates necessarily involves the production of glucose-6-phosphate.



Methylobacterium

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
		acetylCoA	aKG	ATP	methyleneH4F	Ery4P	formylH4F	Glc6P	glycine	NADPH	OAA	PentoP	PEP	PHB	pyruvate	serine	Succ	TP
1	acetylCoA		0.73	0.79	0.80	0.64	0.76	0.61	0.67	0.81	0.84	0.61	0.72	0.85	0.72	0.67	0.76	0.68
2	aKG	0.73		0.55	0.57	0.83	0.56	0.80	0.85	0.63	0.87	0.80	0.91	0.75	0.91	0.85	0.55	0.87
3	ATP	0.79	0.55		0.93	0.51	0.87	0.47	0.52	0.85	0.63	0.47	0.59	0.72	0.59	0.52	0.97	0.55
4	methyleneH4F	0.80	0.57	0.93		0.51	0.93	0.47	0.52	0.81	0.64	0.47	0.59	0.71	0.59	0.52	0.93	0.55
5	Ery4P	0.64	0.83	0.51	0.51		0.47	0.96	0.85	0.60	0.77	0.96	0.89	0.71	0.89	0.85	0.51	0.93
6	formylH4F	0.76	0.56	0.87	0.93	0.47		0.44	0.49	0.79	0.61	0.44	0.55	0.67	0.55	0.49	0.87	0.51
7	Glc6P	0.61	0.80	0.47	0.47	0.96	0.44		0.84	0.59	0.75	1.00	0.85	0.67	0.85	0.84	0.47	0.89
8	glycine	0.67	0.85	0.52	0.52	0.85	0.49	0.84		0.64	0.80	0.84	0.91	0.72	0.91	1.00	0.52	0.89
9	NADPH	0.81	0.63	0.85	0.81	0.60	0.79	0.59	0.64		0.68	0.59	0.68	0.79	0.68	0.64	0.83	0.64
10	OAA	0.84	0.87	0.63	0.64	0.77	0.61	0.75	0.80	0.68		0.75	0.85	0.75	0.85	0.80	0.60	0.81
11	PentoP	0.61	0.80	0.47	0.47	0.96	0.44	1.00	0.84	0.59	0.75		0.85	0.67	0.85	0.84	0.47	0.89
12	PEP	0.72	0.91	0.59	0.59	0.89	0.55	0.85	0.91	0.68	0.85	0.85		0.79	1.00	0.91	0.59	0.96
13	PHB	0.85	0.75	0.72	0.71	0.71	0.67	0.67	0.72	0.79	0.75	0.67	0.79		0.79	0.72	0.69	0.75
14	pyruvate	0.72	0.91	0.59	0.59	0.89	0.55	0.85	0.91	0.68	0.85	0.85	1.00	0.79		0.91	0.59	0.96
15	serine	0.67	0.85	0.52	0.52	0.85	0.49	0.84	1.00	0.64	0.80	0.84	0.91	0.72	0.91		0.52	0.89
16	Succ	0.76	0.55	0.97	0.93	0.51	0.87	0.47	0.52	0.83	0.60	0.47	0.59	0.69	0.59	0.52		0.55
17	TP	0.68	0.87	0.55	0.55	0.93	0.51	0.89	0.89	0.64	0.81	0.89	0.96	0.75	0.96	0.89	0.55	

Erythrocyte

		1	2	3	4
		ATPase	DPGm	GSHox	PRPPs
1	ATPase		0.97	0.50	0.77
2	DPGm	0.97		0.40	0.77
3	GSHox	0.50	0.40		0.37
4	PRPPs	0.77	0.77	0.37	

Figure 7. Similarity matrices for the *MinModes* of the B- and E-network.
Similarity between two MinModes was assessed by the similarity index (si) defined in Equation 5. si-values larger than 0.9 are indicated in grey.

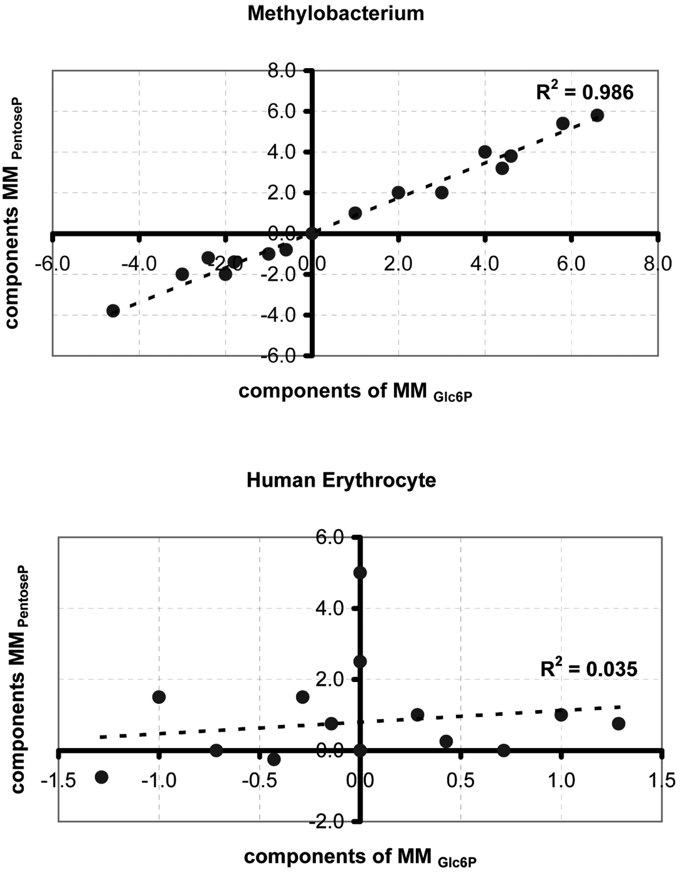


Figure 8. Scattergram illustrating the correlation between the *MinModes* for the formation of Glc6P and pentoseP.

The important point is, however, that with methanol as carbon source a large part of the network is used to produce glucose-6-phosphate. In contrast, in the E-network with glucose as substrate only two reactions are needed (Glct and HK) to form glucose-6-phosphate. As a consequence, the two *MinMode*s associated with the production of Glc6P and PentoseP are completely different (Fig. 8B). This shows that the similarity of two *MinMode*s associated with the production of a given pair of metabolites depends strongly upon the architecture of the network and the available extracellular substrates. From the two similarity matrices shown in Fig. 7 we can identify those target fluxes which are simultaneously affected by changes in the level of active enzymes. For example, in the case of ATP and 2,3DPG we will not be able to discriminate which cellular requirements have changed with respect to those two target functions by only inspecting the internal fluxes.

Based on the similarity matrix, *MinMode*s can be grouped into clusters encompassing all *MinMode*s with sufficiently high mutual similarity. Figure 9 shows the result of a cluster analysis of the *MinModes* for the B-network, performed by using the furthest neighbour method, i. e. measuring the overall similarity of all *MinMode*s assembled in a cluster by the smallest pair-wise similarity index.

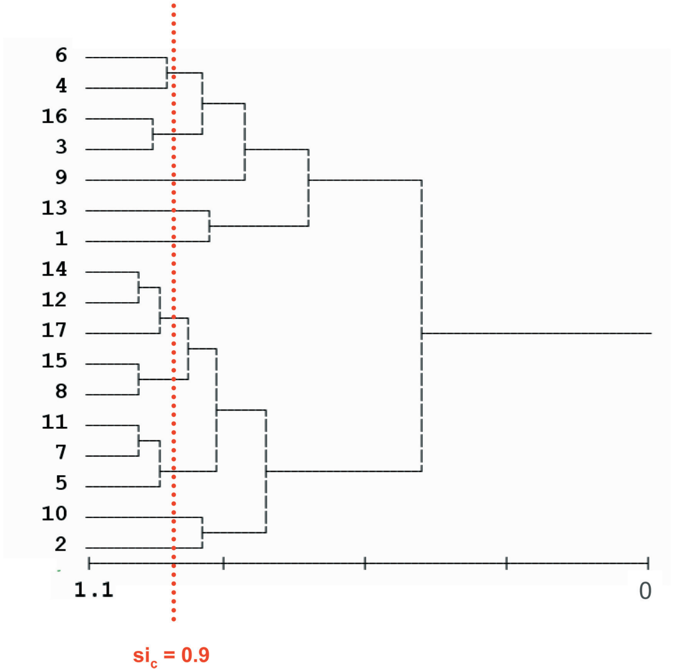


Figure 9. Dendogram illustrating the clustering of *MinModes* for the B- and E-network.
Cluster analysis was performed on the basis of the similarity matrices shown in Fig. 7 using the closest neighbour method, i. e. the smallest similarity index for all pairs of *MinModes* falling into one cluster is larger than the critical value indicated on the horizontal axis.

Extraction of *Principal MinMode*s

Using *MinModes* for the decomposition of flux distributions it seems feasible to represent those *MinModes* comprising a large degree of similarity by a single *Principal MinMode*. *Principal MinModes* exhibit a lesser degree of similarity and thus allow a more unambiguous decomposition. To this end, we have to define a cut-off value (si_c) for *MinMode* similarity. *MinModes* assembled in a cluster possessing a cluster similarity larger than this cut-off value are lumped together (as a linear combination of its elements) to a single *Principal MinModes* (PMMs). Note that *Principal MinModes* does not satisfy the MinMode definition. Further *Principal MinMode*s are given in terms of those *MinMode*s which do not fall into clusters with sufficiently high cluster similarity. The *Principal MinMode*s obtained by this procedure for the two exemplary networks at a cut-off value of $si_c = 0.9$ are depicted in Fig. 10.

Principal MinModes - Methylobacterium

PMM $_1$	$MM_5 + MM_7 + MM_{11}$	Ery4P, Glc6P, PentoseP
PMM $_2$	$MM_{12} + MM_{14} + MM_{17}$	PEP, Pyruvate, TP
PMM $_3$	$MM_8 + MM_{15}$	glycine, serine
PMM $_4$	$MM_4 + MM_6$	methyleneH4F, formylH4F
PMM $_5$	$MM_3 + MM_{16}$	Succ, ATP
PMM $_6$	MM_1	acetylCoA
PMM $_7$	MM_2	aKG
PMM $_8$	MM_9	NADPH
PMM $_9$	MM_{10}	OAA
PMM $_{10}$	MM_{13}	PHB

Principal MinModes - Erythrocyte

PMM 1	$MM_1 + MM_2$	DPGM, ATPase
PMM 2	MM_3	GSHox
PMM 3	MM_4	PRPPs

Figure 10. Definition of Principal *MinModes.*
MinModes falling into clusters with minimal similarity of 0.9 (see Fig. 9) have been lumped into a single Principal MinMode.

There are 10 PMMs (instead of 17 MMs) for the B-network and 3 PMMs (instead of 4) for the erythrocyte network. Obviously, the number of *Principal MinMode*s depends on the choice of the cut-off value si_c for mutual *MinMode* similarity. At $si_c = 0.8$ we would get only 3 PMMs for the B-network.

Flux changes induced by changes of enzyme levels: simulated gene expression

To study the changes of stationary fluxes accompanying changes of enzyme levels we used our comprehensive kinetic model of the erythrocyte metabolism to calculate stationary states at various enzyme levels. Variations in the amount of an enzyme were accomplished by varying its maximal velocity. We considered the following two extreme cases of gene expression. *Random gene expression* was simulated by multiplying the actual v_{max} values of all enzymes with factors randomly chosen within the interval [0.1, 10.0]. *'On demand' gene expression* (see Fig. 11 for a detailed explanation) was simulated by first changing the load

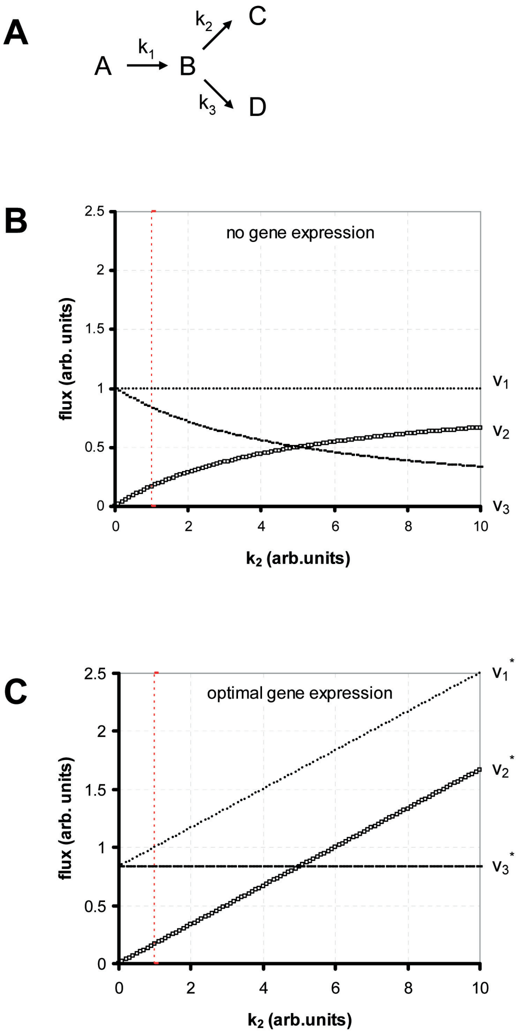


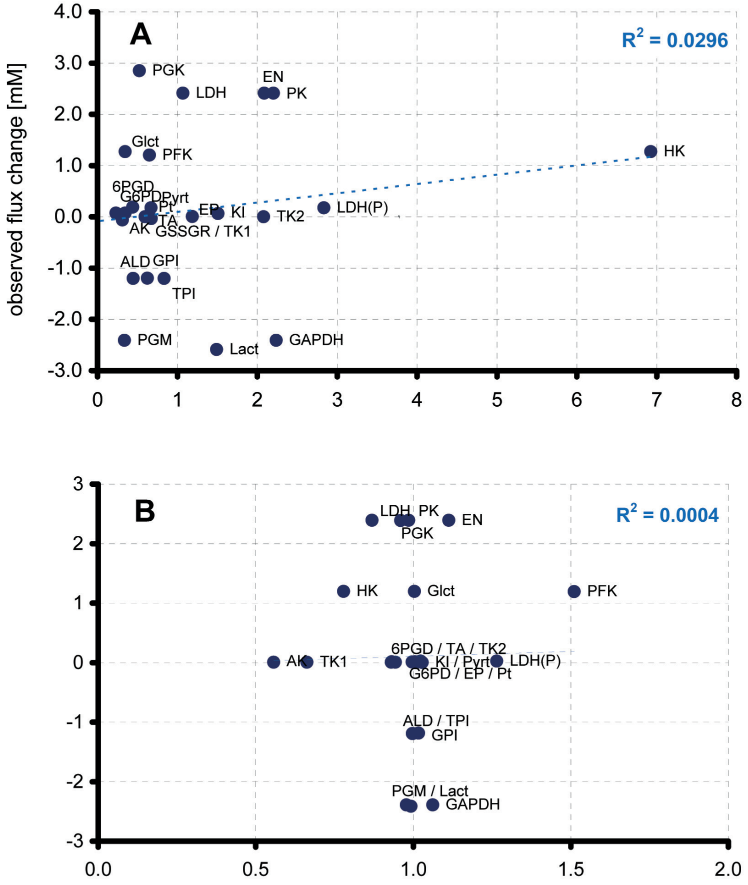
Figure 11. Illustration of the hypothesis on optimal gene expression.
The simplistic network in panel A is composed of three monomolecular reactions. The function of the network is to convert substrate A into two different end products C and D. Under non-saturating conditions the fluxes are given by $v_1 = k_1$ [A], $v_2 = k_2$ [B] and $v_3 = k_3$ [B] whereby the rate constants k_2 and k_3 represent the load parameters and the rate constant k_1 is proportional to the concentration of the enzyme catalysing the first reaction.. Metabolic steady states of the system are defined by the flux-balance condition $v_1 = v_2 + v_3$. Thus, at fixed concentration of the substrate [A]= 1, the output fluxes read $v_2 = \frac{k_1 k_2}{k_2 + k_3}$ and $v_3 = \frac{k_1 k_3}{k_2 + k_3}$.

Panel B illustrates the dependence of the three stationary fluxes on the load parameter k_2 at fixed values $k_3 = 5$ of the second load parameter and $k_1 = 1$ of the rate constant for the first reaction. Increasing values of the load parameter k_2 result in a decrease of flux v_3 whereby the increase of flux v_2 is sub-linear, i.e. the control coefficient $C_2 = \frac{\partial \ln(v_2)}{\partial \ln(k_2)}$ of the output flux v_2 with respect to the load parameter k_2 is smaller than unity. This sub-optimal behaviour becomes successively pronounced with increasing values of k_2. Optimal gene expression is hypothesized both to accomplish a maximal response of the output flux v_2 towards changes of the load parameter k_2 such that the flux-control coefficient becomes unity, $C_2 \rightarrow 1$, and to prevent a change of the other (independent) output flux v_3. This can be achieved by variable expression of enzyme catalysing the reaction A →B, i.e. adapting the rate constant k_1 to the load parameters according to $k_1 = \gamma\ (k_2 + k_3)$. The corresponding fluxes v_i^* $(i = 1,2,3)$ in the presence of optimal gene expression are shown in panel C whereby the proportionality constant was put $\gamma = {}^1/_6$ so that the load characteristics without and with gene expression match at $k_2 = 1$ (indicated by the dashed vertical line).

parameter k_i for target fluxes v_i by a given factor $\eta = k_i'/k_i$ where k_i' is the new value of the load parameter. This change of the load parameter implies a change of the target flux to the new value v_i' but, in general, this change is smaller than η, i.e. the flux-control coefficient $C_i = \frac{1}{\eta} \frac{v_i'}{v_i}$ of the chosen target flux with respect to the load parameter is smaller than unity.

New v_{max} values of all enzymes were than determined fulfilling two criteria: (i) the change of the chosen target flux was also η-fold, i.e. the flux-control coefficient of this target flux with respect to the load parameter was unity, $C_i = 1$, and (ii) the other target fluxes remained at their initial value. This simulated mode of gene expression assures high selectivity in the cellular response towards changes of the metabolic load.

For these two modes of simulated gene expression, changes in the steady-state fluxes were expressed as difference between new and initial flux values and these changes were plotted against the changes in the v_{max} values of the catalysing enzymes expressed as fold changes calculated by dividing the new v_{max} value by its initial value. The scattergrams in Fig. 12A-C and the associated measures of determination (R^2) reveal poor correlations not only for the random case but also for the two cases of 'on demand' gene expression optimizing the response of the metabolic system towards an increase in the load parameters k_{ATPase} for the energy consumption and k_{GSHox} for the consumption of GSH. This finding is in clear contrast to the well-known linear relationship between flux rate and enzyme activity holding for isolated reactions at fixed concentration of the reactants. In a reaction network, however, changes in the activity of a single enzyme remain not restricted to changes in the rate of the corresponding reaction but give rise to changes of all fluxes owing to the coupling of the reactions through shared reactants and allosteric effectors. This way a local perturbation of a single enzyme propagates through the whole network and the resulting new steady-state flux distribution depends on the specific kinetic properties of all enzymes in the network. Hence, from the kinetic point of view, a simple correlation between changes of enzyme levels and changes of the associated fluxes indeed cannot be expected. In the following paragraph we propose a method to exploit information on changes of enzyme concentrations to arrive at better predictions of flux changes in the network.



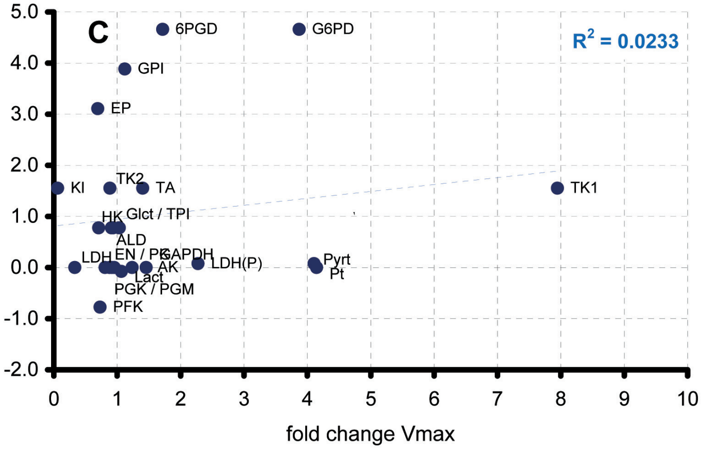


Figure 12. Scattergram illustrating the correlation between flux changes and changes of v_{max} values for the E-network. v_{max} values were changes either randomly (A) or applying an 'on demand' gene expression strategy (details given in the main text). 'Observed' flux changes were calculated by using the kinetic model [30] (B) Simulated 'on demand' gene expression accompanying an increase of the load parameter k_{ATPase} for ATP consumption by a factor of $\eta = 2$. (C) Simulated 'on demand' gene expression accompanying an increase of the load parameter k_{GSHox} for GSH consumption by a factor of $\eta = 100$.

Predicting flux changes from changes of enzyme levels by using the flux decomposition into *Principal MinModes*

Changes in the stationary fluxes are not independent – they are coupled through the balance conditions which hold even if there are changes in the amount of enzymes due to variable gene expression. For example, given that the postulated metabolic network of the erythrocyte is correct the fluxes through the glucose transporter (Glct) and the hexokinase (HK) have to be equal, and the same also holds for any flux changes through these two reactions. Looking at the data in Fig. 10 not in X → Y direction but in Y → X direction one would expect the data points for Glct and HK to coincide if there was a perfect correlation of flux changes with changes of enzyme activities. The scattergram in Fig. 12B ('on demand' gene expression at higher energetic load k_{ATPase}) shows that the v_{max} changes for these two proteins are different. Although there is no change in the activity of the glucose transporter and even a decrease (!) in the activity of the hexokinase (HK) the flux through both reactions has increased by a factor of about 1.2. This is a pure kinetic effect brought about by a lowered intracellular glucose concentration due to activation (higher expression) of the phosphofructokinase (PFK), one of the key regulatory glycolytic enzymes. This example illustrates that some of the flux changes in the network are due to kinetic effects induces by changes in the expression of those enzymes (as the PFK) exerting the dominant control over the desired changes in the metabolic output of the network. Regarding the problem of predicting flux changes from changes of enzyme levels we have to conclude that only some of the observed changes in enzyme activities are indicative for changes of the associated fluxes. For the considered example, the increase in the activity of the PFK actually reflects an increase in the flux through this enzyme whereas the unaltered activity of glucose transporter (Glct) does not.

Let us consider the pool of correlated fluxes, i.e. which are related to each other by fixed ratios in any conceivable flux distributions due to flux balance conditions. Given that within this pool there exists at least one flux for which the change is not kinetically determined but predominantly due to a change of the enzyme level. If so, it would make sense to represent all fluxes belonging to this pool by a single representative flux and to correlate its change with the average observed changes in the activities of the associated enzymes. Such a constrained correlation analysis can be accomplished by approximating the unknown vector $\Delta\mathbf{v}$ of flux changes by a linear decomposition into *Principal MinModes* $\mathbf{PMM_i}$ (see Equation 5) the components of which automatically obey the flux-balance conditions:

$$\Delta\mathbf{v} = \sum_{\mathrm{i}} \Delta\alpha_{\mathrm{i}}\ \mathbf{PMM}_{\mathrm{i}} \qquad (7)$$

The coefficients $\Delta\alpha_i$ (i = 1,2,..., number of PMMs) of this decomposition are than determined by maximizing the correlation between $\Delta\mathbf{v}$ and the observed changes of enzyme activities $\Delta\mathbf{A} = \left(E_1^* / E_1 , E_2^* / E_2 , \ldots , E_{ne}^* / E_{ne} \right)$, i.e.

$$\underset{(\alpha_{\mathrm{i}} \geq 0)}{\mathrm{MAX}}\ \mathrm{Corr}\left(\Delta\,\mathbf{v}, \Delta\mathbf{A}\right) \qquad (8)$$

Here, fold-changes in the v_{max} values are taken as measure for changes in the enzyme amounts. Solving the optimization problem (8) we obtain a prediction of the flux changes in the network. The value of the coefficient $\Delta\alpha_i$ indicates the changes of the target fluxes associated with the *Principal MinMode* $\mathrm{PMM_i}$. Hence, the set of coefficients $\Delta\alpha_i$ allow direct inferences to be made on those changes in the metabolic output the network which have provoked the observed changes in the enzyme levels.

This procedure was applied to the three “expression patterns” illustrated in Fig. 13A-C. The similarity index used to extract *Principal MinModes* for the erythrocyte network was put to 0.9 resulting in the following 3 *Principal MinModes* given in Fig. 10.

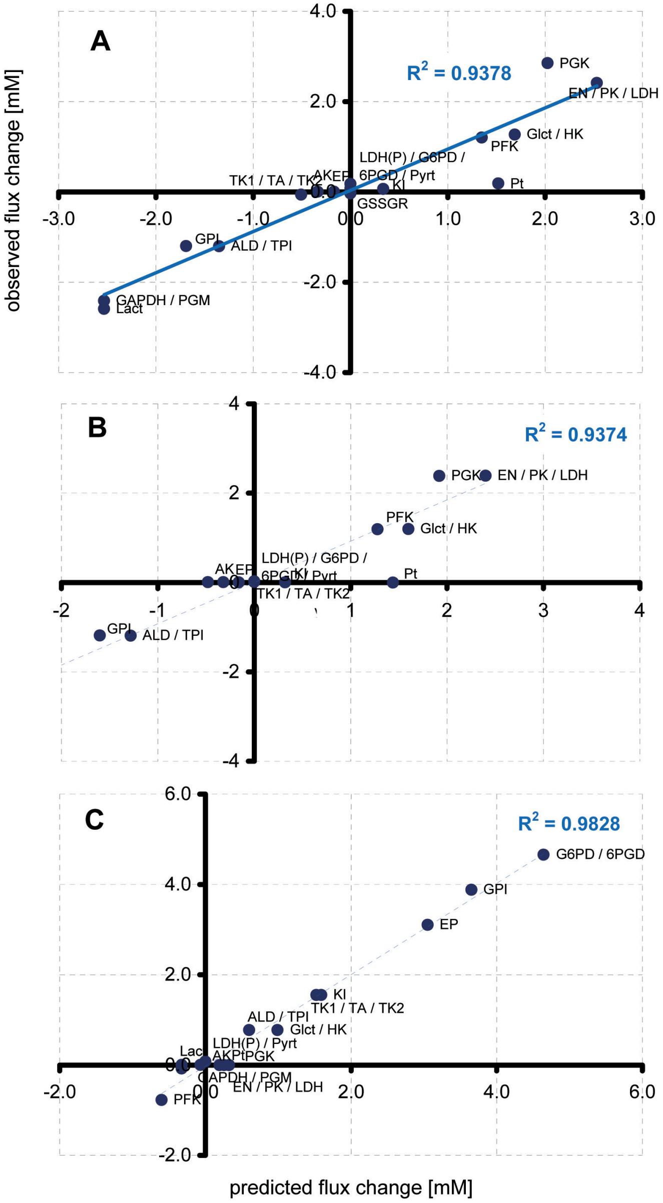


Figure 13. Scattergram illustrating the correlation between 'observed' flux changes and flux changes obtained by maximizing the correlation between changes of v_{max} values and the *MinMode* decomposition (6) based on the three Principal MinModes for the E-network given in Fig. 10. The three simulated cases A-C are explained in Fig. 11. (A) Random gene expression. Estimated values for the decomposition coefficients: $\Delta\alpha_1 = 0.08$, $\Delta\alpha_2 = 0.0$, $\Delta\alpha_3 = 0.06$. (B) 'On demand' gene expression at increased load parameter k_{ATPase}. $\Delta\alpha_1 = 0.07$, $\Delta\alpha_2 = 0.0$, $\Delta\alpha_3 = 0.0$, (C) 'On demand' gene expression at increased load parameter k_{GSHox}. $\Delta\alpha_1 = 0.0$, $\Delta\alpha_2 = 0.72$, $\Delta\alpha_3 = 0.0$.

Figure 13 shows the "observed" (= simulated) flux changes plotted against predicted flux changes obtained as solution of the optimization problem (7). Even for the simulated case of *random gene expression* (Fig. 13A) the concordance between predicted and observed flux changes is surprisingly good. The values of the coefficients α_1, α_2 and α_3 solving the optimization problem (7) for the three cases of simulated gene expression are given in the legend of Fig. 13. According to these values the changes in the maximal enzyme activities for case B and C were clearly identified as resulting from an increase in either the energetic or oxidative load.

Putting these findings together we may conclude that the proposed strategy of:

- decomposing the flux distribution into minimal flux modes
- lumping these *MinModes* together to redundant-free *Principal MinModes*,
- expressing the unknown flux changes as linear combination of *Principal MinModes*
- and determining the unknown coefficients of this linear combination by maximizing the correlation with observed changes of enzyme level (=v_{max} values)

provides a powerful means of predicting flux changes in the metabolic network as well as those changes in the output of the metabolic network having caused these flux changes.

Discussion

As demonstrated for the exemplary network considered herein, the projection of the global flux-minimized steady-state solution onto the convex basis of elementary modes resulted in a manageable set of elementary flux modes with non-vanishing coefficients. However, these "basic" elementary modes were difficult to assign to a specific metabolic output. Besides this, there are some further shortcomings rendering the convex basis of elementary modes unsuitable for the decomposition of flux distributions into functionally interpretable modes. First, the determination of the convex basis is not unique [8]. Thus, choosing another convex basis of elementary modes, their physiological interpretation can be vastly different [19]. Second, the coefficients for the non-negative linear decomposition are also not unique (in our computational protocol we have chosen the coefficients with minimal 1-norm) and thus their absolute values do not allow conclusions to be drawn with respect to the relative importance of the various elementary modes. Third, the (subjective) decision on reversible and irreversible reactions deeply affects the set of elementary modes and thus the convex basis. In the B-network, 43 chemical reactions are considered irreversible. One might argue that potentially every chemical reaction can be reversed by increasing the concentration of the products and/or reduction of the concentration of the substrates. Setting all reactions as reversible renders many elementary modes *a priori* physiologically irrelevant. Elementary modes appearing implausible to the biochemist are, for example, inner cycles that operate without exchange of matter with the environment or modes encompassing reactions with flux directions that are not in concordance with their (large) change of standard free energy.

In this work we introduced the concept of minimal flux modes (*MinMode*s), defined as flux minimized steady-state flux distributions enabling the production of a single metabolite. The production of a certain subset of metabolites defines the functionally relevant output of the network. The introduction of *MinMode*s means to decompose this output into separate contributions. The flux cone spanned by the set of *MinMode*s is a true subset of the mode space. For the Methylobacterium model the convex dimension of the *MinMode* space is 17, whereas the convex basis of elementary modes comprises 7033 vectors. The full vector space spanned by linear combinations of the convex basis with real number coefficients has the dimension 29. Hence the set of *MinMode*s is not complete, i.e. an arbitrary steady flux distribution cannot be exactly decomposed into a linear combination of *MinMode*s. On the other hand, *MinMode*s are attractive because of their clear assignment to specific output reactions. Because the *MinMode*s are biochemically feasible, the reduced flux cone spanned by the *MinMode*s is also feasible as a whole, in contrast to the complete flux cone. Elementary modes and minimal fluxes as introduced in this paper represent two different methodological concepts that ultimately have the same goal: To decompose the fluxes in the metabolic network into a set of more simple but physiologically relevant flux modes. Elementary mode analysis starts with a complete set of all possible and not further decomposable routes ("top down" approach). The resulting huge set of such elementary modes has to be reduced to a physiologically relevant and numerically tractable set by imposing additional constraints [17 – 19]. In contrast, minimal mode analysis starts with a very small set of modes each of them connected with one physiologically relevant output of the network ("bottom up" approach). Thus, the number of *MinMode*s cannot be larger than the number of metabolites occurring in the network. However, *MinMode*s allow only an approximate description of the true flux distribution as they do not form a complete basis in strict mathematical sense. Nevertheless, the striking advantage of the *MinMode* concept is its applicability to very large whole-cell networks (> 500 reactions) where the effective handling of elementary modes can be clearly ruled out for computational reasons.

The set of *MinMode*s can be investigated by similar methods as used for analyses based on elementary modes. For example, the frequency with which a reaction has a non-zero flux within the set of *MinMode*s can be taken to rank the functional relevance of reactions in the network. Such analyses may give insight into the evolution of metabolic networks: Reactions with many non-zero fluxes in the *MinMode*s may represent the ancient part of the network responsible for some basal functions. This part of the network was then successively complemented by reactions and pathways connected to more specific functions and thus being less represented in the *MinMode*s. It has to be noted that disabling reactions with a high number of non-zero fluxes in the set of *MinMode*s does not necessarily lead to lethality because besides the *MinMode*s (being special flux distributions) alternative routes may exist. For example, formate dehydrogenase (reaction 6) has a high participation frequency but was shown to be not essential for growth on methanol [35]. On the other hand, reactions which upon exclusion from the network (by setting the flux to zero) do not allow the determination of a *MinMode* for each output metabolite are to be considered essential. Omitting non-essential reactions merely leads to a different pattern of *MinModes*. If a mutant lacking a reaction predicted to be not essential is not able to survive or shows reduced growing capabilities, an important physiological function of this reaction has failed to be noticed. Therefore, relating the observed phenotype of knock-out mutants with network based classifications of essential and non-essential reactions may provide a valuable heuristics to unravel the physiological importance of metabolites.

Interestingly there are fifteen reactions of the B-network that do not contribute to any flux mode. Assuming the synthesis of biomass precursors to be the sole cellular function of the network, these 15 reactions should be abdicable for cellular growth using methanol as the only carbon source. Three of these reactions contribute to the degradation of the storage metabolite PHB. The enzymes PHB depolymerase, β-hydroxybutyrate dehydrogenase and acetoacetate-succinyl-CoA transferase, able to catalyse the conversion of PHB into acetoac-CoA are not required for cellular growth. However, this holds true only for environmental conditions where enough substrate is available. In a starvation phase, the apparently dispensable PHB degradation becomes important. Another example is the aldolase reaction converting 2 triose phosphates into Fru-1,6-BP (reaction 34). The flux through this reaction in every *MinMode* is zero. Thus, synthesis of Fru-1,6-BP as an intermediate seems to be not necessary for the fulfilment of the assumed metabolic tasks under the given environmental conditions. There are two possible explanations for this redundancy. Either, this reaction is required for enhanced stability and robustness of the network and is abdicable under conditions where the enzymes catalysing alternative routes are expressed or its metabolic task is not required under the given conditions. Or, Fru-1,6-BP is required for other biochemical processes not considered yet, either as a reactant in reactions not included in the network or as a regulatory metabolite. The latter explanation is very likely because knowledge of the full spectrum of metabolites necessary to ensure all cellular activities will be incomplete. For example, it has become known quite recently that presence of Fr-1,6-BP is an absolute requirement for lactate dehydrogenase activity in *Lactobacillus casei* [36] and a similar regulatory function of this metabolite cannot be excluded in *Methylobacterium extorquens*. Therefore, to take into account a potential role of all metabolites in cellular functionality a complete set of *MinModes* should be constructed enabling the production of all metabolites occurring in the network. If this more systematic approach is applied, only 6 of the 15 previously unused reactions remain. In case the model is correct this redundancy should be explained solely by network robustness to mutations and changed environmental conditions. Mutants not able to catalyse these reactions should grow normally. The example of α-KG dehydrogenase (reaction 34), one of the 6 abdicable reactions, supports this hypothesis. The enzyme catalyses the conversion of α-KG into Succ-CoA as part of the citric acid cycle. For growth on C1 resources this cycle is partly repressed and accomplishes an assimilatory role [24, 37]. It could be shown that a lack of this enzyme does not influence the growth behaviour of *Methylobacterium extorquens* (while growing on methanol only) [38]. Thus, the classification of reactions into those which are essential and non-essential has to be considered with precaution because such a classification depends strongly upon the specific external conditions as well as on the knowledge of the physiological functions that metabolites or reactions may have.

One may wonder whether the proposed computational approach to construct optimal flux distributions for each output variable of the network is more or less feasible than the calculation of an optimized flux distribution meeting all target flux simultaneously. The good concordance between experimentally determined fluxes and calculations based on *MinMode composition*s (see Fig. 6) may suggest that optimization of metabolic networks with respect to single target reactions has been an important goal of natural evolution. Considering that the relative importance of target fluxes may vary depending on the specific external conditions of the cell (at the extreme one target reaction as, for example, the production of glutathione in the presence of oxidative stress, might transiently overshoot all others) such a strategy appears to be not implausible because it allows independent regulation of different metabolic outputs. Moreover, adding new reactions (and thus functionality) to an already existing network during the course of natural evolution should not comprise already achieved optimality. Of course, it is too early to make a sound judgment, so that further applications of the proposed methods to other, more complex networks are needed. Based on the same assumption of a minimized total sum of fluxes, the flux distributions obtained by global and single target optimization show of course only little differences. Thus the new approach results in a flux distribution that is just as good as the one obtained by the previous approach. The Minimal Flux Modes 11 few differences can be illustrated by a simple network (Fig. 14A), consisting of 5 reactions v1,v2,v3,v4,v5 where v4 and v5 are considered as target reactions. Two minimal flux modes can be calculated for this network. For the *MinMode* that produces metabolite A by realizing the flux v4 = 1, we obtain the *MinMode* ~v =(10010), whereas for the realization of the flux v5 = 1, the *MinMode* is ~v =(01001). The sum of fluxes in each *MinMode* is 2. The combination of *MinMode*A and *MinMode*B results in the flux distribution ~v =(11011) with the sum of fluxes being 4. The basic flux minimization approach does not presume costs for enzyme synthesis or different activity levels for an enzyme. Therefore, demanding non-zero fluxes through both target reactions v4 and v5 gives the same flux distribution. Thus, single optimization and global optimization may lead to equal solutions as long as no currency metabolites (as ATP, NADH etc.) are involved. In the case, where v4 is an ATP consuming and v3 an ATP producing reaction (Fig. 14B), additional fluxes for ATP balancing become necessary (v6,v7). Here, the sum of fluxes for the single optimized solution would be 4 for *MinMode*A.

In combination with *MinMode*B the sum of fluxes is now 6. Global optimization yields a suboptimal path for the synthesis of B that contains a non-zero flux for the reaction v3, resulting in the vector ~v =(2011100) with a sum of fluxes equal to 5. Therefore, for a global optimization a suboptimal (carbon) route for one metabolite can be chosen if it is of advantage for the synthesis of another target metabolite.

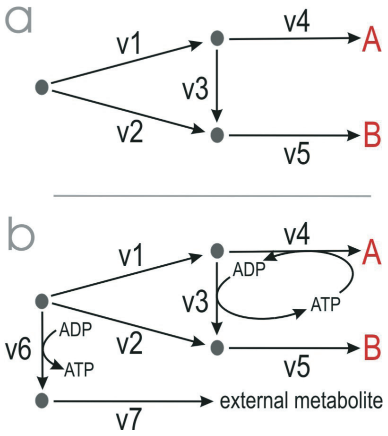


Figure 14. A simplistic network illustrating network effects which may cause differences between fluxes calculated by global minimization and MinMode decomposition.

The basic idea behind the concept of *MinMode*s presented in this paper is to interpret the flux distribution in a metabolic network as a superposition of various flux modes each related to one of the many functional requirements that the cells has to fulfil simultaneously but with different relative intensities. Hence, any metabolic status can be represented in terms of the coefficients entering the linear combination of *MinMode*s to the overall flux distribution. The use of those coefficients simplifies the flux-balance approach considerably and makes it possible to relate observed changes in metabolic fluxes directly to changes in the functionality of the cell.

As demonstrated for the *MinMode*s of the two exemplary networks considered, the *MinMode*s may exhibit a remarkable degree of similarity. Generally, *MinMode*s associated with target reactions located in close vicinity, i.e. belonging to the same pathway, should give rise to similar *MinMode*s. Thus, to employ *MinMode*s as a sort of 'basic vectors' it seems feasible to lump together similar *MinMode*s into a single *Principal MinMode*s. On one hand this leads to a further reduction of the set of relevant *MinMode*s, on the other it also reduces the clear-cut physiological interpretation of these *Principal MinMode*s as they are not associated with only one target flux but a certain group of target fluxes. In this article the definition of such *Principal MinMode*s was accomplished by clustering the *MinMode*s on the basis of a similarity index that counts the number of pair-wise fluxes pointing in the same direction or being zero. As with any statistical procedure, it is finally left to the user to define the minimal degree of similarity that has to be present among all *MinMode*s lumped together to a single *Principal MinMode*s.

In the last part of this article we have used the decomposition of flux changes into *Principal MinMode*s to predict changes in metabolic flux rates from observed changes of enzyme levels. It has to be noted that these results are based on simulated 'gene expression' experiments where we changed the maximal activities of erythrocyte enzymes and calculated the associated flux changes by means of a comprehensive kinetic model available for the E-network. These simulations required some *a priori* assumptions to be made about the regulatory principles underlying variable gene expression. These principles are still widely unknown. However, there is increasing theoretical and experimental evidence [39, 40] that temporal gene expression is an important means of cells to adapt their protein synthesizing capacity to changing external conditions such that the required metabolic output is achieved with high efficiency. This plausible strategy was the rational behind the simulations of 'on demand' gene expression which assures a high response of the network to changes in the demand of specific target reactions whereas the fluxes through other target reaction are kept at constant values. As expected, for the both extreme cases – random and 'on demand' gene expression – there was no significant correlation between changes in enzyme activities and changes in flux rates through the corresponding reactions. This theoretical finding, questions the naive interpretation of changes in gene expression profiles as to directly reflect changes in the activity of the underlying pathways.

Using the *MinMode* decomposition of the unknown flux changes as a side constraint and determining the coefficients of this decomposition to provide a maximal correlation with observed changes of enzyme activities, we obtained a significantly better prediction of flux changes. A further benefit of this strategy is that the values of the coefficients directly indicate the changes in the target fluxes that have elicited the changes in enzyme activities. This way it should be possible to make inferences on the functional strategy of cells just employing information of changes in enzyme levels.

#	enzyme	reaction	k_i	DPGM	ATPase	GSHox	PRPPS
1	Glct	Gluc(out) → Gluc	1	1	1	1	1
2	HK	Gluc + ATP → Glc6P + ADP	3900	1	1	1	1
3	GPI	Fru6P → Glc6P	2.55	-1	-1	1	-1
4	PFK	Fru6P + ATP → Fru1,6P + ADP	100000	1	1	0	1
5	ALD	DHAP + GraP → Fru1,6P	8.77	-1	-1	0	-1
6	TPI	GraP → DHAP	24.6	-1	-1	0	-1
7	GAPDH	1,3PG + NADH → GraP + Pi + NAD	5210	-1	-1	-1	-1
8	PGK	1,3PG + ADP → 3PG + ATP	1455	0	1	1	1
9	DPGM	1,3PG → 2,3PG	100000	1	0	0	0
10	DPGase	2,3PG → 3PG + Pi	100000	1	0	0	0
11	PGM	2PG → 3PG	6.9	-1	-1	-1	-1
12	EN	2PG → PEP	1.7	1	1	1	1
13	PK	PEP + ADP → Pyr + ATP	13790	1	1	1	1
14	LDH	Pyr + NADH → Lac + NAD	9090	1	1	1	1
15	LDH(P)	Pyr + NADPH → Lac + NADP	1420	0	0	0	0
16	ATPase	ATP → ADP + Pi	100000	0	1	0	0
17	AK	2 ADP –¿ ATP + AMP	0.64	0	0	-1	-1
18	G6PD	G6P + NADP → 6PG + NADPH	2000	0	0	1	0
19	6PGD	6PG + NADP → Ru5P + CO2 + NADPH	141.7	0	0	1	0
20	GSSGR	GSSG + NADPH → 2 GSH + NADP	3417.8	0	0	1	0
21	GSHox	GSH → GSSG	100000	0	0	1	0
22	EP	Ru5P → X5P	2.7	0	0	1	-1
23	KI	Ru5P → R5P	3	0	0	1	1
24	TK1	X5P + R5P → GraP + S7P	1.05	0	0	1	-1
25	TA	S7P + GraP → E4P + Fru6P	1.05	0	0	1	-1
26	PRPPS	R5P + ATP → AMP + PrPP	100000	0	0	1	1
27	TK2	X5P + E4P → GraP + Fru6P	1.2	0	0	1	-1
28	Pt	Pi(out) → Pi	1	0	0	1	1
29	Lact	Lac(out) → Lac	1	-1	-1	-1	-1
30	Pyrt	Pyr(out) → Pyr	1	0	0	0	0

Table 2. Model scheme and the corresponding minimal flux modes (MinModes) of the E-network.

References

[1] Clarke, B.L. (1988) Stoichiometric network analysis. *Cell Biophys.* **12:**237–253.

[2] Forster, J., Famili, I., Fu, P., Palsson, B.O., Nielsen, J. (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res.* **13:**244–253.

[3] Price, N.D., Papin, J.A., Schilling, C.H., Palsson, B.O. (2003) Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol.* **21:**162-169.

[4] Reed, J.L., Vo, T.D., Schilling, C.H., Palsson, B.O. (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol.* **4:**R54.

[5] Schilling, C.H., Covert, M.W., Famili, I., Church, G.M., Edwards, J.S., Palsson, B.O. (2002) Genome-scale metabolic model of Helicobacter pylori 26695. *J. Bacteriol.* **184:**4582–4593.

[6] Schilling, C.H., Letscher, D., Palsson, B.O. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theoret. Biol.* **203:**229**-**248.

[7] Schuster, S., Hilgetag, C., Woods, J.H., Fell, D.A. (2002) Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *J. Math. Biol.* **45:**153**-**181.

[8] Wagner, C., Urbanczik, R. (2005) The geometry of the flux cone of a metabolic network. *Biophys J.* **89:**3837–3845.

[9] Papin, J.A., Price, N.D., Wiback, S.J., Fell, D.A., Palsson, B.O. (2003) Metabolic pathways in the post-genome era. *Trends Biochem. Sci.* **28:**250–258.

[10] Papin, J.A., Price, N.D., Edwards, J.S., B, B.Ø. P. (2002) The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy. *J. Theoret. Biol.* **215:**67**-**82.

[11] Price, N.D., Papin, J.A., Palsson, B.Ø. (2002) Determination of redundancy and systems properties of the metabolic network of Helicobacter pylori using genome-scale extreme pathway analysis. *Genome Res.* **12:**760-769.

[12] Beard, D.A., Babson, E., Curtis, E.. Qian, H. (2004) Thermodynamic constraints for biochemical networks. *J. Theoret. Biol.* **228:**327–333.

[13] Beard, D.A., Liang, S.-d., Qian, H. (2002) Energy balance for analysis of complex metabolic networks. *Biophys. J.* **83:**79-86.

[14] Klamt, S., Gilles, E.D. (2004) Minimal cut sets in biochemical reaction networks. *Bioinformatics* **20:**226-234.

[15] Gagneur, J., Klamt, S. (2004) Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics* **5:**175.

[16] Klamt, S., Stelling, J. (2002) Combinatorial complexity of pathway analysis in metabolic networks. *Mol. Biol. Rep.* **29:**233-236.

[17] Covert, M.W., Palsson, B.O. (2003) Constraints-based models: regulation of gene expression reduces the steady-state solution space. *J. Theoret. Biol.* **221:**309-325.

[18] Schwartz, J.-M., Kanehisa, M. (2005) A quadratic programming approach for decomposing steady-state metabolic flux distributions onto elementary modes. *Bioinformatics* **21:**Suppl 2, ii204-ii205.

[19] Schwartz, J.-M., Kanehisa, M. (2006) Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis, *BMC Bioinformatics* **7:**186.

[20] Cornish-Bowden, A., Cárdenas, M.L. (2002) Metabolic balance sheets, *Nature* **420:**129–130.

[21] Holzhütter, H.-G. (2004) The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur. J. Biochem.* **271:**2905–2922.

[22] Holzhütter, H.-G. (2006) The generalized flux-minimization method and its application to metabolic networks affected by enzyme deficiencies, *Biosystems* **83:**98–107.

[23] Holzhütter, S., Holzhütter, H.G. (2004) Computational design of reduced metabolic networks. *Chembiochem* **5:**1401–1422.

[24] Dien, S.J. V., Lidstrom, M.E. (2002) Stoichiometric model for evaluating the metabolic capabilities of the facultative methylotroph Methylobacterium extorquens AM1, with application to reconstruction of C(3) and C(4) metabolism. *Biotechnol. Bioeng.* **78:**296–312.

[25] Dien, S.J. V., Strovas, T., Lidstrom, M.E. (2003) Quantification of central metabolic fluxes in the facultative methylotroph methylobacterium extorquens AM1 using 13C-label tracing and mass spectrometry. *Biotechnol. Bioeng.* **84:**45–55.

[26] Hoffmann, S., Hoppe, A., Holzhütter, H.G. (2006). Composition of metabolic flux distributions by functionally interpretable minimal flux modes (MinModes). Paper presented at the *Genome Informatics*.

[27] Dien, S.J. V. personal communication.

[28] Vorholt, J. personal communication.

[29] Heinrich, R., Holzhütter, H.G., Schuster, S. (1987) A theoretical approach to the evolution and structural design of enzymatic networks: linear enzymatic chains, branched pathways and glycolysis of erythrocytes. *Bull. Math. Biol.* **49:**539–595.

[30] Schuster, R., Holzhütter, H.G. (1995) Use of mathematical models for predicting the metabolic effect of large-scale enzyme activity alterations. Application to enzyme deficiencies of red blood cells. *Eur. J. Biochem.* **229:**403–418.

[31] Holzhütter, S., Holzhütter, H.-G. (2004) Computational design of reduced metabolic networks. *Chembiochem* **5:**1401-1422.

[32] CPLEX. http://www.ilog.com/products/cplex/.

[33] Klamt, S., Stelling, J., Ginkel, M., Gilles, E.D. (2003) FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics* **19:**261–269.

[34] Pfeiffer, T., Sánchez-Valdenebro, I., Nuño, J.C., Montero, F., Schuster, S. (1999) METATOOL: for studying metabolic networks. *Bioinformatics* **15:**251–257.

[35] Chistoserdova, L., Laukel, M., Portais, J.-C., Vorholt, J.A., Lidstrom, M.E. (2004) Multiple formate dehydrogenase enzymes in the facultative methylotroph Methylobacterium extorquens AM1 are dispensable for growth on methanol. *J. Bacteriol.* **186:**22–28.

[36] Arai, K., Hishida, A., Ishiyama, M., Kamata, T., Uchikoba, H., Fushinobu, S., Matsuzawa, H., Taguchi, H. (2002) An absolute requirement of fructose 1,6-bisphosphate for the Lactobacillus casei L-lactate dehydrogenase activity induced by a single amino acid substitution. *Protein Eng.* **15:**35–41.

[37] Dien, S.J. V., Marx, C.J., O'Brien, B.N., Lidstrom, M.E. (2003) Genetic characterization of the carotenoid biosynthetic pathway in Methylobacterium extorquens AM1 and isolation of a colorless mutant. *Appl. Environ. Microbiol.* **69:**7563–7566.

[38] Chistoserdova, L., Chen, S.-W., Lapidus, A., Lidstrom, M.E. (2003) Methylotrophy in Methylobacterium extorquens AM1 from a genomic point of view. *J. Bacteriol.* **185:**2980–2987.

[39] Klipp, E., Heinrich, R., Holzhutter, H.G. (2002) Prediction of temporal gene expression. Metabolic opimization by re-distribution of enzyme activities. *Eur. J. Biochem.* **269:**5406–5413.

[40] Zaslaver, A., Mayo, A.E., Rosenberg, R., Bashkin, P., Sberro, H., Tsalyuk, M., Surette, M.G., Alon, U. (2004) Just-in-time transcription program in metabolic pathways. *Nature Genet.* **36:**486- 491.

Problems of Currently Published Enzyme Kinetic Data for Usage in Modelling and Simulation

Ursula Kummer and Sven Sahle

Bioinformatics and Computational Biochemistry Group, EML Research,
Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany

E-Mail: ursula.kummer@eml-r.villa-bosch.de

Received: 14th August 2006 / Published: 31st August 2007

Abstract

Modelling, simulation and computational analysis have become important tools in modern biochemistry. Moreover, their tight integration with experimental approaches has become an integral part of systems biology which has attracted scientific and political interest all over the world. However, published enzymatic data often does not take a modeller's viewpoint into account, even though in many cases this would only demand minor adjustments and would serve the community a great deal. Supporting users by automating some of the steps in modelling and simulation adds even more requirements. In the following we would like to emphasize a few points that we feel should be further supported or that have been neglected in the discussion about the standardization of enzymatic data, but would be valuable for modellers.

Introduction

Even though computational biochemistry is a quite ancient part of life sciences, its impact and importance for experimental research has not been acknowledged until recently. The recent interest obviously stems from the fact that the sheer complexity of the biochemical network in a living cell (as opposed to simple isolated enzymatic reactions) calls for computational help. Thus, today systems biology is understood as the tight integration of computational and experimental research in order to understand biochemical systems in their entirety.

In order to set up decent biochemical models we rely heavily on data from experiments and from previous modelling, especially kinetic data. However, the way this data has been published in the past is often lacking information crucial for the set-up of models. This has been recognized recently and discussed at the previous ESCEC meeting. In addition, the more frequent use of modelling techniques and the increasing size and complexity of models has led to the development of software tools that support users in the process of modeling, e. g., Pedro Mendes' group (VBI) and our group have developed COPASI (http://www.copasi.org) which offers a user-friendly, platform independent facility to set-up models, and to simulate and analyse them. In the course of developing the software as well as when performing modelling studies ourselves, we have encountered many problems with the published enzyme kinetic information. Most of that has been thoroughly and extensively discussed during the previous meeting.

However, we feel that some problems still have been neglected or at least are underestimated and we would like to point these out in the following.

Specific Problems

Importance of the kinetic equation

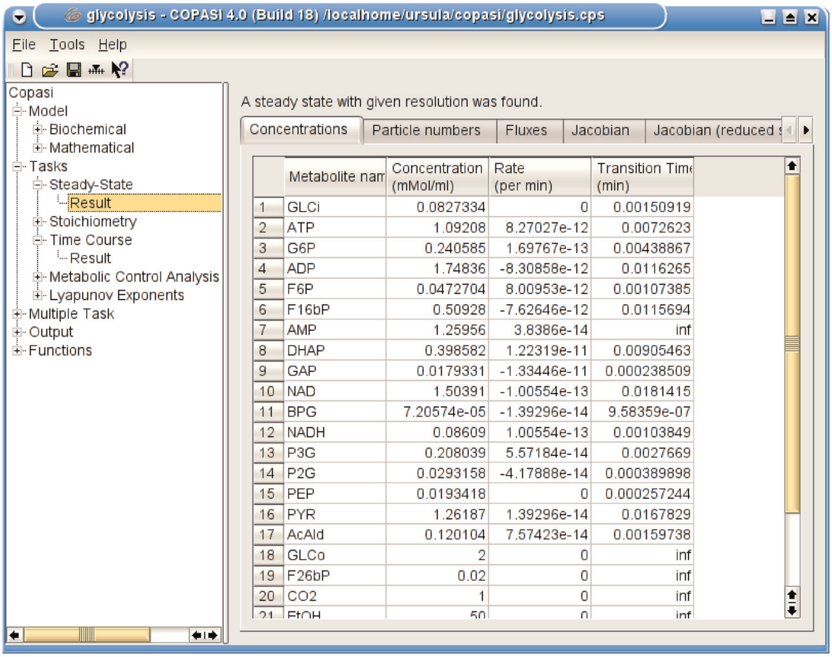
The vast majority of kinetic data published in the literature comprises V_{max} and K_{m} values or other individual rate constants. However, this is only part of the information necessary for modelling the respective system. In many cases the actual kinetic equation which is assumed or even was used to derive the published parameter (often by fitting to the equation) is missing. Without this crucial information, the value of publishing the actual parameter is greatly diminished. It also does not help too much if authors mention the name of the corresponding rate-law in the text, as e. g. Bi–Bi- Ping–Pong, etc.; since these terms are not used in an unambiguous way and therefore can be very misleading. What is actually needed is the explicit notation of the respective equation – nothing else. This would make sure that modellers do not have to guess which equation to use. In addition, wrong use, e. g. using a parameter with the wrong rate law would be avoided. Just to illustrate this obvious point a little bit further we use the following arbitrary example:
We exchange the kinetic term for the hexose transporter in a model for yeast glycolysis by Teusink *et al*. The original term

$$\frac{\frac{\mathrm{Vmax}\cdot(A-B)}{\mathrm{Kglc}}}{1+\frac{A+B}{\mathrm{Kglc}}+\frac{\mathrm{Ki}\cdot A\cdot B}{\mathrm{Kglc}^2}}$$

is exchanged against a somewhat simpler Uni–Uni term

$$\frac{\mathrm{Vf}\cdot\left(\mathrm{substrate}-\frac{\mathrm{product}}{\mathrm{Keq}}\right)}{\mathrm{substrate}+\mathrm{Kms}\cdot\left(1+\frac{\mathrm{product}}{\mathrm{Kmp}}\right)}$$

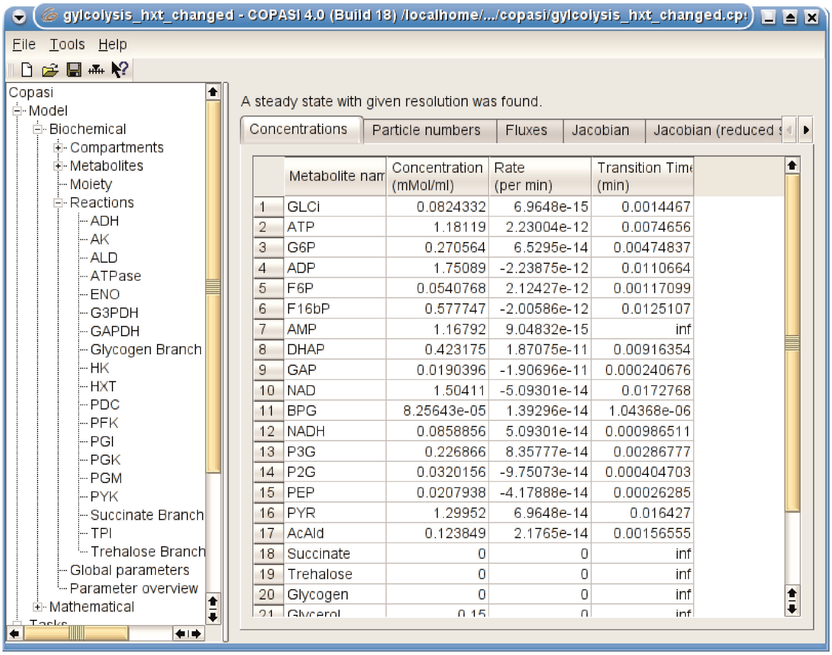
in which K_{eq} equals one. Since product and substrate are glucose and the respective K_m values are assumed to be the same, both terms are actually quite similar. We use the same parameters in both cases. The resulting models are analysed w.r.t. their steady state behaviour. This analysis is done using COPASI (http://www.copasi.org). The results are shown in Figs 1 and 2.



A steady state with given resolution was found.

	Metabolite nam	Concentration (mMol/ml)	Rate (per min)	Transition Time (min)
1	GLCi	0.0827334	0	0.00150919
2	ATP	1.09208	8.27027e-12	0.0072623
3	G6P	0.240585	1.69767e-13	0.00438867
4	ADP	1.74836	-8.30858e-12	0.0116265
5	F6P	0.0472704	8.00953e-12	0.00107385
6	F16bP	0.50928	-7.62646e-12	0.0115694
7	AMP	1.25956	3.8386e-14	inf
8	DHAP	0.398582	1.22319e-11	0.00905463
9	GAP	0.0179331	-1.33446e-11	0.000238509
10	NAD	1.50391	-1.00554e-13	0.0181415
11	BPG	7.20574e-05	-1.39296e-14	9.58359e-07
12	NADH	0.08609	1.00554e-13	0.00103849
13	P3G	0.208039	5.57184e-14	0.0027669
14	P2G	0.0293158	-4.17888e-14	0.000389898
15	PEP	0.0193418	0	0.000257244
16	PYR	1.26187	1.39296e-14	0.0167829
17	AcAld	0.120104	7.57423e-14	0.00159738
18	GLCo	2	0	inf
19	F26bP	0.02	0	inf
20	CO2	1	0	inf
21	EtOH	50	0	inf

Figure 1. Steady-state concentrations as computed by COPASI using the glycolysis model of Teusink *et al*.



A steady state with given resolution was found.

	Metabolite nam	Concentration (mMol/ml)	Rate (per min)	Transition Time (min)
1	GLCi	0.0824332	6.9648e-15	0.0014467
2	ATP	1.18119	2.23004e-12	0.0074656
3	G6P	0.270564	6.5295e-14	0.00474837
4	ADP	1.75089	-2.23875e-12	0.0110664
5	F6P	0.0540768	2.12427e-12	0.00117099
6	F16bP	0.577747	-2.00586e-12	0.0125107
7	AMP	1.16792	9.04832e-15	inf
8	DHAP	0.423175	1.87075e-11	0.00916354
9	GAP	0.0190396	-1.90696e-11	0.000240676
10	NAD	1.50411	-5.09301e-14	0.0172768
11	BPG	8.25643e-05	1.39296e-14	1.04368e-06
12	NADH	0.0858856	5.09301e-14	0.000986511
13	P3G	0.226866	8.35777e-14	0.00286777
14	P2G	0.0320156	-9.75073e-14	0.000404703
15	PEP	0.0207938	-4.17888e-14	0.00026285
16	PYR	1.29952	6.9648e-14	0.016427
17	AcAld	0.123849	2.1765e-14	0.00156555
18	Succinate	0	0	inf
19	Trehalose	0	0	inf
20	Glycogen	0	0	inf
21	Glycerol	0.15	0	inf

Figure 2. Steady-state concentrations of the same model as Fig. 1 with the term for the hexose transporter changed as explained in the text.

It is easy to see that the steady-state concentration of most variables differs by more that ten percent. Thus, the systems behaviour is significantly changed by this minor change in kinetics with the same parameters used. This trivial example illustrates the above said and calls for the inclusion of the notation of the kinetic equation in the standardization of published enzymatic data.

Finally, if a reaction involves participating species with different stoichiometries, it should be stated clearly to which participant (substrate or product) the rate law applies (as is often, but not always, done in literature). Preferably the rate law should be stated for a species with unity stiochiometry.

The V_{max} parameter

Another apparent (and recognized) problem is the publication of the V_{max} values. Since most studies are done *in vitro* the enzyme concentration contained in the V_{max} is the one in the test tube. However, modellers are usually interested in the enzyme concentration in the living cell instead. Even though the enzyme of interest has been isolated from cellular material in most cases, there is often not even an estimate of the amount present in the respective life material. An estimation of the original amount often is also not possible by calculating backwards since the results of the purification steps are not reported in sufficient detail.

In addition, instead of simply reporting the components of V_{max}, namely the enzyme concentration and the rate constant, many authors hamper the calculation of the individual rate constant by not explicitly writing down the respective enzyme concentration in the test tube, but rather giving the activity of the enzyme without giving amounts etc. (see unit notation below).

All in all, this effectively turns V_{max} into an unknown variable in most cases, introducing a lot of fuzziness into the system. Of course, in many if not most cases, there can be no exact quantificatation of the enzyme of interest in a specific cell type. This implies that parameter estimation techniques have to be used at some point in time. However, this procedure is obviously more reliable and much faster if the initial values are good guesses. These estimates could be very well provided in the primary literature.

Reversible rate laws

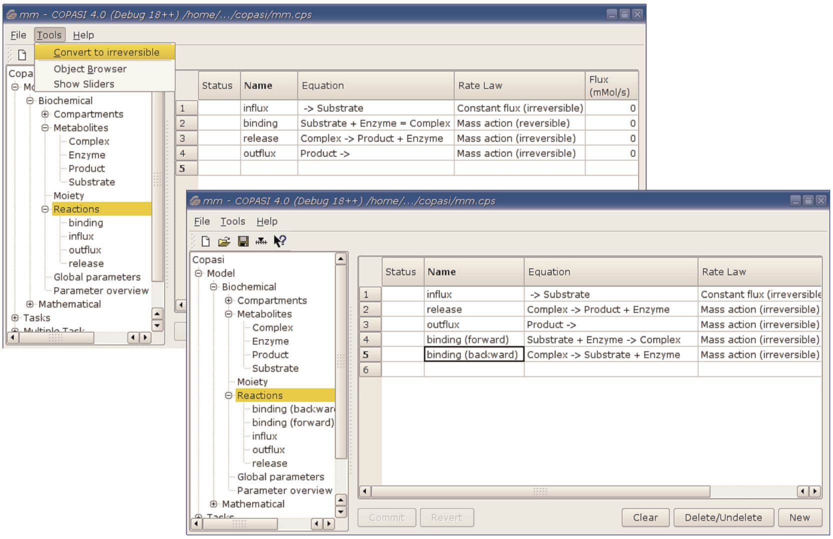
The notation of reversible rate laws is another, albeit less severe problem. Reversible rate laws do not pose any problem when models are written down using ordinary differential equations (ODEs). Forward and backward flows of a reversible reaction can cancel each other out so that the overall rate can be given as a single expression. Depending on the concentrations of the substrates and products the rate can be positive or negative, it is zero if the reaction is in equilibrium.

However, when modelling biochemical systems containing only relatively low numbers of the participating compounds, e.g. because of volume limitations (e.g. in vesicles) or because of functional necessity (e.g. signalling), we often have to refer to stochastic methods on discrete particle basis [1]. In the stochastic modelling and simulation framework each reaction is characterized by a reaction probability (instead of a reaction rate). A stochastical simulation works as follows: first the probabilities of all reactions are calculated. These depend on the concentrations of the species that take part in the reactions. Then, taking into account the probabilities of all the reactions, it is determined which reaction will take place next and at which point of time this will happen. This is done by drawing random numbers from a random number generator. The chosen reaction is then "executed" by increasing the particle numbers of the corresponding product species and decreasing the particle numbers of the substrates. So far one single reaction step was simulated. The whole process is repeated.

This stochastic simulation process ensures that the effects of discreteness (the fact that particle numbers are always integers) and the effects of stochasticity (the single reaction events happen at random points of time) are considered.

Concerning the relation between reaction rates and reaction probabilities it is clear that reaction rates can also be expressed as an average number of reaction events happening in a unit of time. This in turn can easily be translated into a reaction probability. Thus in many cases (and under certain conditions) the traditional rate laws and kinetic parameters can be utilized for stochastic simulations. A problem occurs, however, if the rate law describes a reversible reaction. Consider for example a reversible reaction in equilibrium. The net rate is zero, which means that substrate and product concentrations do not change due to this reaction. It does not matter that in reality many reaction events in both direction take place. In the stochastic simulation however every single (forward and backward) reaction event needs to be simulated. Since the reactions are random, this leads to fluctuations around the equilibrium. For some short time more forward reaction events may happen, after that more backward reaction events occur. Only as an average over some time the reaction rate is zero. Therefore, separate rate laws for the forward and backward part of the reactions need to be available.

Thus, if rate laws are given for reversible reactions these terms have to be dismantled which is of course possible to do manually. Due to the increasing size of biochemical systems modelled and the reuse of parts of a model in other models, an automatization of this process however would be useful. Thus, e.g. COPASI contains a preliminary tool which is able to dissect reversible reactions automatically into forward and backward reactions (Fig. 3), but right now (apart from the trivial mass action case) the respective kinetics have to be adjusted by the user.



	Status	Name	Equation	Rate Law	Flux (mMol/s)
1		influx	-> Substrate	Constant flux (irreversible)	0
2		binding	Substrate + Enzyme = Complex	Mass action (reversible)	0
3		release	Complex -> Product + Enzyme	Mass action (irreversible)	0
4		outflux	Product ->	Mass action (irreversible)	0
5					

	Status	Name	Equation	Rate Law
1		influx	-> Substrate	Constant flux (irreversible
2		release	Complex -> Product + Enzyme	Mass action (irreversible)
3		outflux	Product ->	Mass action (irreversible)
4		binding (forward)	Substrate + Enzyme -> Complex	Mass action (irreversible)
5		binding (backward)	Complex -> Substrate + Enzyme	Mass action (irreversible)
6				

Figure 3. Screenshot of COPASI demonstrating the tool that renders reversible reactions into two irreversible reactions. The two windows show the list of reactions before and after the conversion.

Simultaneous and combined use of different simulation methods would be facilitated if rate laws were either written down individually (for forward and backward reactions) or written down in such a way that they can easily be dismantled automatically by computer programs. Thus, if the forward reaction rate is simply the first term of the numerator divided by the denominator and the backward reaction rate is the second term of the numerator divided by the denominator as in the following example, an automatic dismantling is relatively simple, irrespective of e. g. brackets in this term.

An example taken from Holzhütter *et al.* [2] as stored in JWS online [3]:

$$\mathrm{v11[ENO]} = \frac{\mathrm{Vmaxv11}\left(\mathrm{Gri2P[t]} - \frac{\mathrm{PEP[t]}}{\mathrm{Keqv11}}\right)}{\mathrm{Gri2P[t]} + \mathrm{K2PGv11}\left(1 + \frac{\mathrm{PEP[t]}}{\mathrm{KPEPv11}}\right)}$$

However, another example from the same paper as stored in the database shows a case where this is not as simple:

$$\mathrm{v9[BPGP]} = \frac{\mathrm{Vmaxv9}\left(\mathrm{Gri23P2f[t]} - \frac{\mathrm{Gri3P[t]}}{\mathrm{Keqv9}} + \mathrm{MgGri23P2[t]}\right)}{\mathrm{K23P2Gv9} + \mathrm{Gri23P2f[t]} + \mathrm{MgGri23P2[t]}}$$

Coherent unit notation

Problems with unit notations are mostly associated with the notation of enzymatic activities and concentrations. It is still common to use units like e. g. “activity per mg freshweight”. As pointed out above, reuse of the respective kinetic data makes it necessary to compute the enzyme concentration in the assay. In order to do so, one has to gather all information from the text (if at all possible) about molecular weight, purity etc. This can be quite cumbersome and is probably done multiple times by different people in the community. Instead, it will be much easier if authors do this right away and provide the respective information in the original text.

Conclusions

Computational biochemistry relies more and more on tools that automate and facilitate individual steps in the setting up of models and their computational analysis. In addition to the general requirements of the modelling community, this development adds stronger and different requirements w.r.t. published enzymatic data. Some of these have been discussed above. We hope that enzymatic databases like SABIO-RK [4] and BRENDA [5] will also help by being a useful intermediate layer of information between the primary literature and the modeller being able to curate enzyme kinetic data in such a way that some of the above problems will be resolved.

Acknowledgements

We would like to thank the BMBF and the Klaus Tschira Foundation for funding. We are also grateful to Martin Golebiewski, Ralph Gauges, Stefan Hoops, Renate Kania, Pedro Mendes, Jürgen Pahle, Isabel Rojas, and Andreas Weidemann for fruitful discussions.

References

[1] Gillespie, D.T. (1976) A general method for numerically simulating the stochastic-time evolution of coupled chemical reactions. *J. Comp. Phys.* **22:**403–434.

[2] Holzhütter, H.-G. (2004) The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur. J. Biochem.* **271:**2905–2922.

[3] Olivier, B.G., Snoep, J.L.(2004) Web-based kinetic modelling using JWS Online. *Bioinformatics* **20:**2143–2144.

[4] Wittig, U., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann, A., Anstein, S., Saric, J., Rojas, I. (2006) SABIO-RK: Integration and curation of reaction kinetics data. *Lecture Notes in Bioinformatics* **4075:**94–103.

[5] Schomburg, I., Chang, A., Schomburg, D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.* **30:**47–49.

Adding Semantics in Kinetics Models of Biochemical Pathways

Nicolas Le Novere, Melanie Courtot, Camille Laibe

EMBL-EBI, Wellcome-Trust Genome Campus, Hinxton, CB10 1SD
United-Kingdom

E-Mail: lenov@ebi.ac.uk

Received: 24th August 2006 / Published: 31st August 2007

Abstract

The need to exchange and integrate models drove the community to design common data format such as SBML. However, as important as was the definition of a common syntax, we also need to tackle the semantics of the models. The community recently proposed MIRIAM, the Minimal Information Requested in the Annotation of Models, a set of rules for curating quantitative models of biological systems. This standard lists the condition an encoded model has to meet to fully correspond to its reference description, and describe how to annotate each of its components. The Systems Biology Ontology (SBO) aims to strictly index, define and relate terms used in quantitative modelling, and by extension quantitative biochemistry. SBO is currently made up of five different vocabularies: quantitative parameters, participant roles, modelling frameworks, mathematical expressions – that refers to the three previous branches – and events. SBO can be used not only to annotate quantitative models, but also biochemical experiements. It is expected that the adoption of those two semantic layers will favour the reusability of quantitative biochemical descriptions, whether parameters or models.

Introduction

Until very recent times, life science coped extremely well with fuzzy semantics, and that for very sound operational reasons. Except for hard-core biochemists and pharmacologists, concepts like *Kd*, *Kp*, EC_{50} etc. were globally known as inversely proportional to the "affinity", and it was sufficient to have a sensible discussion about biological processes. The Molecular Biology era reinforced even further that trend, by shifting the general interest from quantitative to qualitative description of physiology. Moreover, when quantitation nevertheless sneaked in, the results were quite often expressed as relative rather than absolute values. Typically it was the measurement in a mutant or upon a perturbation, normalised over the same measurement in a wild-type animal, or in unperturbed condition. With the notable exception of the sequence and structure of macromolecules, biologists were not really supposed to store and exchange experimental results, but rather to build on the interpretation of data provided by the authors themselves, who were judged the best suited to analyse it.

Then entered Functional Genomics. The large genome projects had demonstrated that it was possible to produce experimental data on a large scale, with a quality control that far surpassed the standards of isolated academic group [1]. New technologies now promised to allow a similar large-scale generation for more functional types of data. However, those technologies were complex and very expensive. As a consequence, the workforce and degree of expertise needed to put them in application meant that single research groups could not longer run their own data production. And the costs of experiments also meant that each dataset should be produced in a limited number of replicates. Biologists finally had to store and exchange raw data. Despite scale of the datasets, continuous improvements of computing power offered biologists efficient tools to perform the necessary archival/retrieval procedures.

The second shock came from the rise of Systems Biology, which increased the general awareness not only to modelling and simulation of biological processes, but also to quantitative biology in general. As a consequence, what was once the territory of a small population of specialists is now visited by various actors of biomedical research. In parallel, the formal models used in biology are growing, both in size and complexity. A given modeller is therefore less likely to be an expert of all the corners of a quantitative model, whether the biological knowledge or even the mathematical approaches.

This need for quantitative data of high quality calls for a shift of paradigm in the way experimental parameters are exchanged and re-used, but also theoretical concepts handled. There is no point to exchanging quantitative data or models if nobody can understand the meaning of the data and the content of the models beside their initial generators. The community has to define agreed-upon standards for kinetic data generation and curation, so that the experimental measurement can be safely reuse. Controlled vocabularies, where concepts are related one to the other, must be designed for annotating quantitative models with connections to biological data resource. Finally, one needs to integrate modelling work with the other sources of knowledge, and disseminate the large number of models produced.

To offer a possible answer to those issues, the BioModels.net initiative was launched in 2004 by Michael Hucka, Andrew Finney and the author. BioModels.net (http://biomodels.net) is an international effort to (1) define community standards for model curation, (2) design controlled vocabularies for annotating models with connections to biological data resources, and (3) provide a free, centralised, publicly-accessible database of annotated, computational models in SBML and other structured formats. In this paper we will describe two resources belonging to the first and second classes. The third objective has been tackled with BioModels Database [2] (http://www.ebi.ac.uk/biomodels/)

Minimal Information Requested in the Annotation of Biochemical Models

Searching for existing models relevant to a specific problem, a scientist comes across a model named *Model1*, describing the reactions *rA* and *rB* between the molecular components *X* and *Y*.
What can we make of this model? Where does this model come from? What are exactly the components *X* and *Y* in molecular or cellular terms? It could help a lot to know what biological process is modelled by *rA* and *rB*. Providing one finally elucidates the origin of the model, and the identity of its components, how can we know that when instantiated, this model will provide the correct numerical results?

The aim of MIRIAM [3] is to define processes and schemes that will increase the confidence in model collections and enable the assembly of model collections of high quality. The first part of the guidelines is a standard for reference correspondence, dealing with the syntax and semantics of the model. A second part is an annotation scheme, that specifies the documentation of the model by external knowledge. The scheme for annotation can itself be further subdivided into two sections. The *attribution* covers the minimum information that is required to associate the model with a reference description and an actual encoding process. The *external data resources* covers information required to relate the components of quantitative models to established data resources or controlled vocabularies.

The aim of standard for reference correspondence is to ensure that the model is properly associated with a reference description and is consistent with that reference description. The reference description can be a scientific article, but also any other unique publication, on print or online, that describes precisely the structure of the models, list the quantitative parameters, and described the expected output. In order to be declared MIRIAM-compliant, a quantitative model must fulfil the following rules:

1. The model must be encoded in a public, standardised, machine-readable format such as (but not restricted to) SBML[4] or CellML[5].
2. The model must be clearly related to a single reference description. If a model is derived from several initial reference descriptions, there must still be a unique reference description that references a set of results that one can expect to reproduce when simulating the derived/combined model.
3. The encoded model structure must reflect the biological processes listed in the reference description (this correspondence is not necessarily one-to-one).
4. Quantitative attributes of the model, such as initial conditions and parameters, as well as kinetic expressions for all reactions, have to be defined, in order to allow to instantiate simulations.
5. The model, when instantiated within a suitable simulation environment, must be able to reproduce all results given in the reference description that can readily be simulated.

In order to be confident in re-using an encoded model, one should be able to trace its origin, and the people who were involved in its inception. The following information should always be joined with an encoded model:

- A citation of the reference description with which the model is associated, either as a complete bibliographic record, or as a unique identifier, Digital Object Identifier (http://www.doi.org), PubMed identifier (http://www.pubmed.gov), unambiguous URL pointing to the description itself etc.
- Name and contact information for the creators, that is the persons who actually contributed to the encoding of the model in its present form.
- The date and time of creation, and the date and time of last modification.
- A precise statement about the terms of distribution. The statement can be anything from "freely distributable" to "confidential". MIRIAM being intended to allow models to be communicated better, terms of distribution are essential for that purpose.

The aim of the external data resources annotation scheme is to link the components of a model to corresponding structures in existing and future open access bioinformatics resources. Such data resources can be, for instance, database or ontologies. This will permit not only the identification of model components and the comparison of components between different models, but also the search for models containing specific components.

This annotation must permit to unambiguously relate a piece of knowledge to a model component. The referenced information should be described using a triplet {"data-type", "identifier", "qualifier"}.

- The "data-type" is a unique, controlled, description of the type of data, written as a Unique Resource Identifier. The URIs can be expressed as a Uniform Resource Locator, e.g. http://www.uniprot.org/ or a Uniform Resource Name, e.g. urn:lsid:uniprot.org. They have no physical meaning, and if expressed as an URL, does not have to correspond to an existing website. For instance the URI representing the enzyme classification is http://www.ec-code.org/.
- The "identifier", within the context of the "data-type", points to a specific piece of knowledge. For instance 2.7.11.17 for a calcium/calmodulin regulated protein kinase.
- The optional "qualifier" is a string that serves to refine the relation between the referenced piece of knowledge and the described constituent. Although MIRIAM standard does not impose any restriction on the use of qualifiers, biomodels.net nevertheless provides predefined qualifiers, described below.

Such a triplet can easely be exported later using the Resource Description Framework (http://www.w3.org/RDF/), to ease further automatic treatment. RDF is at the core of what is called "Semantic Web" (http://www.w3.org/2001/sw/), and one of the basic technologies that enables modern data interoperability in life science [9]
The following qualifiers are examples that can be used to characterize model components:

[is] The modelling object represented by the model component is the subject of the referenced resource. For instance, this qualifier might be used to link the encoded model to a database of models.

[isDescribedBy] The modelling object represented by the component of the encoded model is described by the referenced resource. This relation might be used to link a model or a kinetic law to the literature that describes this model or this kinetic law.

The following qualifiers are examples that can be used to characterize the biological entity represented by model components.

[is] The biological entity represented by the model component is the subject of the referenced resource. This relation might be used to link a reaction to its exact counterpart in KEGG or Reactome for instance.

[hasPart] The biological entity represented by the model component includes the subject of the referenced resource, either physically or logically. This relation might be used to link a complex to the description of its components.

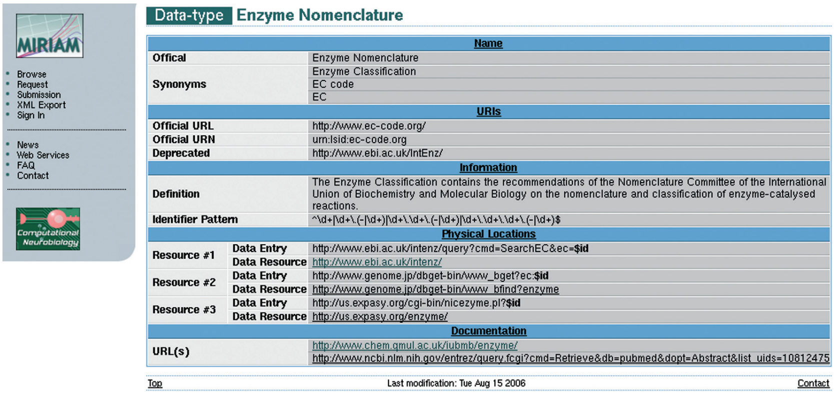
[isPartOf] The biological entity represented by the model component is a physical or logical part of the subject of the referenced resource. This relation might be used to link a component to the description of the complex is belongs to.

[isVersionOf] The biological entity represented by the model component is a version or an instance of the subject of the referenced resource.

[hasVersion] The subject of the referenced resource is a version or an instance of the biological entity represented by the model component.

[isHomologTo] The biological entity represented by the model component is homolog, to the subject of the referenced resource, i.e. they share a common ancestor.

[isDescribedBy] The biological entity represented by the model component is described by the referenced resource. This relation should be used for instance to link a species or a parameter to the literature that describes the concentration of the species or the value of the parameter.



Data-type Enzyme Nomenclature

<table>
<tr><th colspan="3">Name</th></tr>
<tr><td colspan="2">Offical</td><td>Enzyme Nomenclature</td></tr>
<tr><td colspan="2">Synonyms</td><td>Enzyme Classification
EC code
EC</td></tr>
<tr><th colspan="3">URIs</th></tr>
<tr><td colspan="2">Official URL</td><td>http://www.ec-code.org/</td></tr>
<tr><td colspan="2">Official URN</td><td>urn:lsid:ec-code.org</td></tr>
<tr><td colspan="2">Deprecated</td><td>http://www.ebi.ac.uk/IntEnz/</td></tr>
<tr><th colspan="3">Information</th></tr>
<tr><td colspan="2">Definition</td><td>The Enzyme Classification contains the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzyme-catalysed reactions.</td></tr>
<tr><td colspan="2">Identifier Pattern</td><td>^\d+|\d+\.(-|\d+)|\d+\.\d+\.(-|\d+)|\d+\.\d+\.\d+\.(-|\d+)$</td></tr>
<tr><th colspan="3">Physical Locations</th></tr>
<tr><td rowspan="2">Resource #1</td><td>Data Entry</td><td>http://www.ebi.ac.uk/intenz/query?cmd=SearchEC&ec=<b>$id</b></td></tr>
<tr><td>Data Resource</td><td>http://www.ebi.ac.uk/intenz/</td></tr>
<tr><td rowspan="2">Resource #2</td><td>Data Entry</td><td>http://www.genome.jp/dbget-bin/www_bget?ec:<b>$id</b></td></tr>
<tr><td>Data Resource</td><td>http://www.genome.jp/dbget-bin/www_bfind?enzyme</td></tr>
<tr><td rowspan="2">Resource #3</td><td>Data Entry</td><td>http://us.expasy.org/cgi-bin/nicezyme.pl?<b>$id</b></td></tr>
<tr><td>Data Resource</td><td>http://us.expasy.org/enzyme/</td></tr>
<tr><th colspan="3">Documentation</th></tr>
<tr><td colspan="2">URL(s)</td><td>http://www.chem.qmul.ac.uk/iubmb/enzyme/
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=10812475</td></tr>
</table>

Top Last modification: Tue Aug 15 2006 Contact

Figure 1: Example of the entry in MIRIAM database describing the Enzyme Classification.

To enable interoperability, a set of standard valid URIs has to be maintained, and tool provided to automatically retrieve valid URL(s) corresponding to a given URI. This is the purpose of MIRIAM Database and the associated Web Services (http://www.ebi.ac.uk/compneur-srv/miriam/).

MIRIAM is maintained at the EBI using an open relational database management system (MySQL) and a web application using a free implementation of Java Server Page (servlet container Apache Tomcat http://tomcat.apache.org/). Each entry of the database contains a diverse set of details about a given data-type: official name and synonyms, the URIs (URL and/or URN forms), patterns of identifiers, links to documentation documentation etc. In addition, each data-type can be associated with several physical locations. An example is shown on figure.

Users are able to perform queries such as retrieving valid physical locations (URLs) corresponding to a given URI (physical location of a generic data-type or of a precise piece of knowledge), retrieving all the information stored about a data-type (such as its name, its synonyms, links to some documentation etc.). Moreover, a programmatic access through Web Services http://www.w3.org/2002/ws/ (based on Apache Axis (http://ws.apache.org/axis/) and SOAP messages [10]) allows a software not only to re-solved model annotations, but also to generate the correct URIs based on resource name and accession numbers.

For instance, a software generating content needs to annotate an enzymatic activity: The query getURI(”enzyme nomenclature”, ”2.7.11.17”) returns the result http://www.ec-code.org/#2.7.11.17. Conversely, an interface to a database of enzymatic activity needs to generate an hyperlink: The query `getDataEntries(\)http://www.ec-code.org/#2.7.11.17\))` returns the result (at the time of redaction of this chapter) `http://www.ebi.ac.uk/intenz/query? cmd=SearchE-C&ec=2.7.11.17, http://www.genome.jp/dbget-bin/www_bget? ec:2.7.11.17,http://us.expasy.org/cgi-bin/nicezyme.pl? 2.7.11.17`.

Dozens of other methods are available to interact with the database and make the most of annotations.

Systems Biology Ontology

Whilst many controlled vocabularies exist that can directly be used to relate quantitative models to biological knowledge, there were no classification of the concepts themselves used in quantitative modelling. BioModels.net partners recognized that several additional small controlled vocabularies were required to enable the systematic capture of information in those models started to develop their own ontology.

The word ontology is defined here in its information science meaning, as a hierarchical structuring of knowledge. In our case, it is a set of relational vocabularies, that is a set of terms linked together. Each term has a definition and a unique identifier. Those ontologies have seen their role in structuring our knowledge growing steadily in life science over the last few years. The most famous ontology in life-science is Gene Ontology (GO) [11]. They have actually been used by life scientists for a while, also not recognised as such. The most obvious examples are the various taxonomies, of organisms, of sequences families or of protein domains. Less apparent is the fact that other biochemical knowledge management frameworks, such as the Enzyme Classification, also fulfill many of the criteria necessary to qualify as an ontology.

One of the goals of the Systems Biology Ontology (SBO, http://www.ebi.ac.uk/sbo/) is to facilitate the immediate identification of the relation between a model component and the model structure. SBO is currently made up of five different vocabularies (Figure 2). Within a vocabulary, the terms are related by “is a” inheritances, which represent sub-classing.

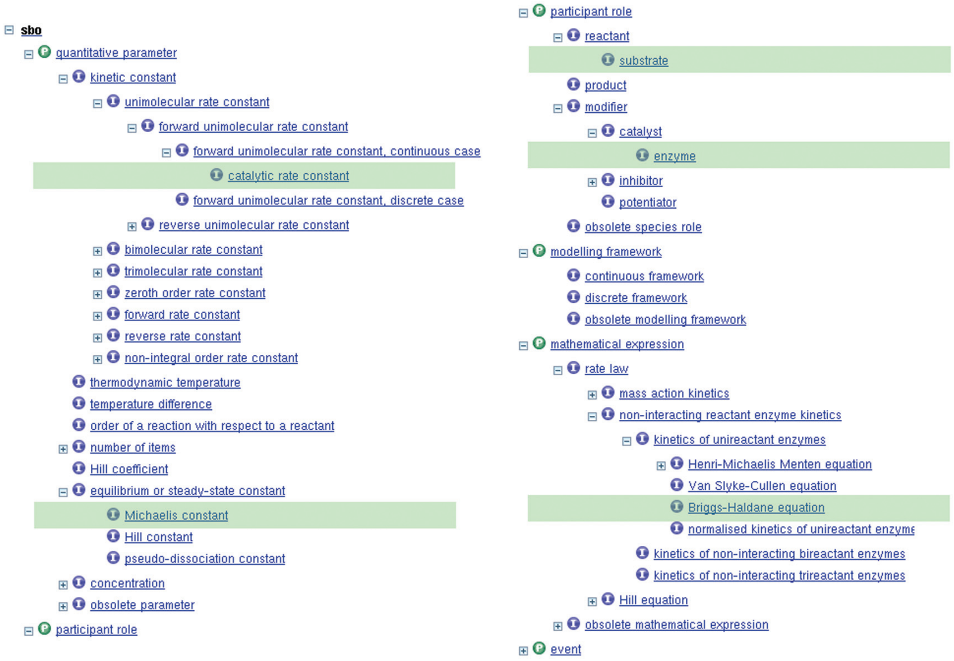

sbo
quantitative parameter
kinetic constant
unimolecular rate constant
forward unimolecular rate constant
forward unimolecular rate constant, continuous case
catalytic rate constant
forward unimolecular rate constant, discrete case
reverse unimolecular rate constant
bimolecular rate constant
trimolecular rate constant
zeroth order rate constant
forward rate constant
reverse rate constant
non-integral order rate constant
thermodynamic temperature
temperature difference
order of a reaction with respect to a reactant
number of items
Hill coefficient
equilibrium or steady-state constant
Michaelis constant
Hill constant
pseudo-dissociation constant
concentration
obsolete parameter
participant role
participant role
reactant
substrate
product
modifier
catalyst
enzyme
inhibitor
potentiator
obsolete species role
modelling framework
continuous framework
discrete framework
obsolete modelling framework
mathematical expression
rate law
mass action kinetics
non-interacting reactant enzyme kinetics
kinetics of unireactant enzymes
Henri-Michaelis Menten equation
Van Slyke-Cullen equation
Briggs-Haldane equation
normalised kinetics of unireactant enzyme
kinetics of non-interacting bireactant enzymes
kinetics of non-interacting trireactant enzymes
Hill equation
obsolete mathematical expression
event

Figure 2: Partially unfolded view of SBO tree. Highlighted are the terms we use in the example described in the text.

1. A controlled vocabulary for parameter roles in quantitative models. This CV includes terms such as “forward unimolecular rate constant”, “Hill coefficient”, “Michaelis constant” etc.
2. A taxonomy of the roles of reaction participants, including the following terms: “catalyst”,
3. “substrate”, “competitive inhibitor” etc.
4. A list of modelling framework, that precises how to interpret a mathematical expression, such as “deterministic”, “stochastic”, “boolean” etc.
5. A classification of mathematical expression used in biochemical modelling. In particular this controlled vocabulary contains a taxonomy of kinetic rate equations. Examples of terms are “mass action kinetic”, “Henri-Michaelis-Menten kinetics”, “Hill equation” etc.

6. A branch containing the classification of events represented by biochemical models, such as “binding”, “transport” or “degradation”.

Each SBO term is made up of a stable identifier, a name, a definition, synonyms, a list of parentages, comments, and, for the mathematical expression branch, an equation. The identifier is a unique string that is never deleted once it is created. If a term needs to be suppressed, it is made child of the “obsolete” branch of the corresponding vocabulary. The name is unique in the ontology, but can change over time. The parentages are of two types, a subclassing (or subsumption or hyponymy) “Is A”, and a dissection (or meronymy) “Part Of”. Contrarily to other ontologies such as Gene Ontology, the latter is used only to link direct children of the root (the five vocabularies).

As an example, the term describing Briggs-Haldane kinetics is described on figure 3.

The annotation of quantitative model components with SBO terms will be an essential step to reach MIRIAM-compliance. Such an annotation will add the layer of semantics necessary to link mathematical representations of biochemical models encoded in SBML or CellML with graphical notations such as the Systems Biology Graphical Notation (http://www.sbgn.org/), or semantically enriched computing formats to represent biochemical knowledge such as BioPAX [12] (http://www.biopax.org). SBO will enhance our capacity to understand and to programmatically analyse models. Finally, SBO will also power the search strategies used by the databases of models and kinetics. In the following we present some examples of SBO use.

SBO to Discriminate between Implicit Hypothesis

The conversion between a continuous and a discrete modelling framework sometimes requires the transformation of a unique complex rate-equation into the description of several elementary reactions. The complex rate-equation has been generally derived using hypothesis that most often are not explicit from the equation itself. As an example, let's consider the case of a simple irreversible unireactant enzyme catalysis. The transformation of a substrate S into a product P by an enzyme E as been formalised by Victor Henri in 1903 [13] and later by Leonor Michaelis and Maud Menten in 1913 [14] as following the kinetic law:

$$v = [E] \times \frac{k_{cat} \times [S]}{K_S + [S]}$$

[Term]
id: SBO:0000031
name: Briggs-Haldane equation

def: ”Rate-law presented in ”G.E. Briggs and J.B.S. Haldane (1925) A note on the kinetics of enzyme action, *Biochem. J.*, **19**: 339 – 339”. It is a general rate equation that does not require the restriction of equilibrium of Henri-Michaelis-Menten or irreversible reactions of Van Slyke, but instead make the hypothesis that the complex enzyme-substrate is in quasi-steady-state. Although of the same form than the Henri-Michaelis-Menten equation, it is semantically different since Km now represents a psudoequilibrium constant, and is equal to the ratio between the rate of consumption of the complex (sum of dissociation of substrate and generation of product) and the association rate of the enzyme and the substrate.
mathml:

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <semantics definitionURL="http://biomodels.net/SBO/\#SBO:0000062">
    <lambda>
      <bvar><ci definitionURL="http://biomodels.net/SBO/\#SBO:0000025">kcat</ci></bvar>
      <bvar><ci definitionURL="http://biomodels.net/SBO/\#SBO:0000014">Et</ci></bvar>
      <bvar><ci definitionURL="http://biomodels.net/SBO/\#SBO:0000015">S</ci></bvar>
      <bvar><ci definitionURL="http://biomodels.net/SBO/\#SBO:0000027">Km</ci></bvar>
      <apply>
        <divide/>
        <apply>
          <times/>
          <ci>kcat</ci>
          <ci>Et</ci>
          <ci>S</ci>
        </apply>
        <apply>
          <plus/>
          <ci>Km</ci>
          <ci>S</ci>
        </apply>
      </apply>
    </lambda>
  </semantics>
</math>
```

is_a: SBO:0000028 ! kinetics of unireactant enzymes

Figure 3: SBO term describing Briggs-Haldane kinetics using the OBO flat format

k_{cat}x[E] is equal to the experimental maximal velocity, and K_S corresponds to the experimental substrate concentration required to reach half-maximal velocity. Henri-Michaelis-Menten mechanism assumed an underlying set of three elementary reactions:

$$E + S \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} ES \xrightarrow{k_{\text{cat}}} E + P$$

In addition, those authors supposed a fast equilibrium between the enzyme/substrate complex and the free enzyme and substrate. As a consequence, K_S=k_{off}/k_{on}
One year later, Donald Van Slyke and Glenn Cullen [15] proposed another explanation based on two successive and irreversible reactions:

$$E + S \xrightarrow{k_{\text{on}}} ES \xrightarrow{k_{\text{cat}}} E + P$$

Although the microscopic mechanism is different, the general form of the velocity is equivalent. However, the constant, equivalent to K_S, is now $K=k_{cat}/k_{on}$.
Finally, in 1924, George Edward Briggs and John Burdon Sanderson Haldane generalised the mechanism than Michaelis and Menten described, releasing the hypothesis of fast equilibrium. Instead they replaced it with the famous quasi-steady-state approximation for the enzyme/substrate complex. The velocity follows yet again the same rate-law.

However, $K_m = \frac{k_{off} + k_{cat}}{k_{on}}$.

Now let's say we come accross a model describing a reaction using the Henri-Michaelis-Menten equation. Here is the SBML description of the reaction:

[SBML code]

```
<reaction>
  <listOfReactants>
    <speciesReference species="S" />
  </listOfReactants>
  <listOfProducts>
    <speciesReference species="P" />
  </listOfProducts>
  <listOfModifiers>
    <speciesReference species="E" />
  </listOfModifiers>
  <kineticLaw>
    <listOfParameters>
      <parameter id="Km"/>
      <parameter id="kp"/>
    </listOfParameters>
    <math xmlns="http://www.w3.org/1998/Math/MathML">
      <apply>
        <times/>
        <ci>compartment</ci>
        <apply>
          <divide/>
          <apply><times/><ci>E</ci><ci>kp</ci><ci>S</ci></apply>
          <apply><plus/><ci>Km</ci><ci>S</ci></apply>
        </apply>
      </apply>
    </math>
  </kineticLaw>
</reaction>
```

There are several situations where we have to develop an elementary step equivalent, instead of using directly the combined non-linear version. For instance, we cannot use such a rate-law when the condition of substrate excess is not met or the total concentration of enzyme varies significantly. Another situation is the use of stochastic simulation tools. First, we have to create first an extra species *ES*. Then we have three possibilities for the reaction scheme.

$$\mathrm{E} + \mathrm{S} \underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} \mathrm{ES}$$

$$\mathrm{ES} \xrightarrow{k_{\mathrm{cat}}} \mathrm{E} + \mathrm{P}$$

With $k_{off}{=}K_m \times k_{on}$.

$$\mathrm{E} + \mathrm{S} \xrightarrow{k_{\mathrm{on}}} \mathrm{ES}$$

$$\mathrm{ES} \xrightarrow{k_{\mathrm{cat}}} \mathrm{E} + \mathrm{P}$$

With $k_{on} = \frac{K_m}{k_{cat}}$.

$$\mathrm{E} + \mathrm{S} \underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} \mathrm{ES}$$

$$\mathrm{ES} \xrightarrow{k_{\mathrm{cat}}} \mathrm{E} + \mathrm{P}$$

With $k_{off} = K_m \times k_{on} - k_{cat}$

The second case is determined. In the first and third cases, one of the parameters has to be estimated, either from external knowledge or using a parameter estimation procedure[1].

If the model description is provided without additional information, there is no way to choose between the three alternatives. On the contrary, the annotation of the model is sufficient, not only to help us to decide between the three alternatives, but also to automatically convert the parameters and the rate-laws. Note the absence of MathML description of the rate-law in the kineticLaw element, unecessary in this case. Since all the parameters are local, a software can reconstruct the the rate-law by matching the SBO term reference on species, parameters and kineticLaw with the ones on the bvar of the SBO term MathML:

[1] k_{on} was choosen as the unknown, because it does not truly depend on the characteristics of the enzymatic reaction. Instead, it depends only on the environment (molecular crowding, viscosity) and scales with the square-root of the mass.

[MathMl code]

```
<reaction>
  <listOfReactants>
    <speciesReference species="A" definitionURL="http://www.biomodels.net/SBO/#SBO:0000015" />
  </listOfReactants>
  <listOfProducts>
    <speciesReference species="B" definitionURL="http://www.biomodels.net/SBO/#SBO:0000011" />
  </listOfProducts>
  <listOfModifiers>
    <speciesReference species="C" definitionURL="http://www.biomodels.net/SBO/#SBO:0000014" />
  </listOfModifiers>
  <kineticLaw definitionURL="http://www.biomodels.net/SBO/#SBO:0000031" >
    <listOfParameters>
      <parameter id="U" definitionURL="http://www.biomodels.net/SBO/#SBO:0000027" />
      <parameter id="V" definitionURL="http://www.biomodels.net/SBO/#SBO:0000025" />
    </listOfParameters>
  </kineticLaw>
</reaction>
```

An SBO-aware software will have access to the vocabularies of SBO, either as a local copy, or using a programmatic access to the master copy. It will recognize that the kineticLaw represents a Briggs-Haldane kinetics and transform the description of the enzymatic reaction into the following elementary steps:

```
<listOfParameters>
  <parameter id="kon" definitionURL="http://www.biomodels.net/SBO/#SBO:0000036" constant="false" />
  <parameter id="koff" definitionURL="http://www.biomodels.net/SBO/#SBO:0000038" />
  <parameter id="U" definitionURL="http://www.biomodels.net/SBO/#SBO:0000027" />
  <parameter id="V" definitionURL="http://www.biomodels.net/SBO/#SBO:0000025"/>
</listOfParameters>
<listOfRules>
  <assignmentRule variable="kon">

    <math xmlns="http://www.w3.org/1998/Math/MathML">
      <apply>
        <divide/>
        <apply><plus/><ci>koff</ci><ci>V</ci></apply>
        <ci>U</ci>
      </apply>
    </math>
  </assignmentRule>
</listOfRules>
<reaction id="v1" reversible="true">
  <listOfReactants>
    <speciesReference species="A" definitionURL="http://www.biomodels.net/SBO/#SBO:0000015" />
    <speciesReference species="B" definitionURL="http://www.biomodels.net/SBO/#SBO:0000014" />
  </listOfReactants>
  <listOfProducts>
    <speciesReference species="AB" definitionURL="http://www.biomodels.net/SBO/#SBO:0000011" />
  </listOfProducts>
  <kineticLaw definitionURL="http://www.biomodels.net/SBO/#SBO:0000101" />
</reaction>
<reaction id="v2" reversible="false">
  <listOfReactants>
    <speciesReference species="AB" definitionURL="http://www.biomodels.net/SBO/#SBO:0000010" />
  </listOfReactants>
  <listOfProducts>
    <speciesReference species="C" definitionURL="http://www.biomodels.net/SBO/#SBO:0000011" />
  </listOfProducts>
  <kineticLaw  definitionURL="http://www.biomodels.net/SBO/#SBO:0000049" />
</reaction>
```

The kineticLaw of the reaction v1 is annotated with the SBO term “second order forward with two reactants, first order reverse, reversible mass action kinetics, continuous scheme”, while the kineticLaw of the reaction v2 is annotated with the SBO term “first order irreversible mass action kinetics, continuous scheme”. Note that the species AB is associated with different SBO terms according to the reaction. The procedure would be exactly the same if instead of continuous descriptions for the elementary reactions, one wanted to use discrete rate-laws. The only changes would be the SBO terms on the kineticLaw elements.

Use of SBO to annotate experimental measurements

Precise annotation is not only necessary for theoreticians, but also for experimentalists. It is unfortunately all too frequent to come across confusions between V_{max} and k_{cat}, or K_p, K_d and IC_{50}. Similarly, the rate-law used to fit the experimental data-point and extract parameters is sometimes omitted. This potentially results in incorrect interpretations. This confusion reduces much the reuse of quantitative information in biochemistry, or even worse, lead to false interpretations.

A careful annotation of both rate-laws and parameters with the relevant SBO terms would increase the amount of information transferred from the data generation step to the data analysis one, minimising the risk of confusion, maximising the value for money of biochemical experimentation, and finally avoiding the continuous reiteration of the same data generation for different purposes.

Such an annotation could be directly reused by databases of quantitative biochemistry such as BRENDA [16] or SABIO-RK [17]. SBO terms could serve as a glue between various part of those knowledge management systems, but could also be used to query the resources, searching for a given parameter or a type of kinetics.

In addition, SBO annotation could help automatically generating part of the resources. For instance, using a mathematical expression term, one can directly create the adequate forms to enter the concentrations and parameters, as well as the corresponding structures in RDBMS tables.

SBO development and export

The Systems Biology Ontology is now listed as part of the Open Biomedical Ontologies (OBO). OBO is an umbrella for well-structured controlled vocabularies for shared use across different biological and medical domains. OBO seek to enforce some criteria of quality, orthogonality and stability among its ontologies. In addition, OBO ontologies share common formats and processing tools. As other OBO ontologies, SBO is an open-resource, developed and maintained by the scientific community, and reusable under the terms of the artistic license (http://www.opensource.org/licenses/artistic-license.php).

Everybody can submit request for new terms or suggestions to modify the structure of the ontology, or of the associated services through the Sourceforge project (http://sourceforge.net/projects/sbo/)

To curate and maintain SBO, we developed a dedicated resource (http://www.ebi.ac.uk/sbo/). A relational database management system (MySQL) at the back-end is accessed through a web interface based on JSP and JavaBeans. Its content is encoded in UTF8, therefore supporting a large set of characters in the definitions of terms. Distributed curation is made possible by using a tailored locking system allowing concurrent access. This system allows a continuous update of the ontology with immediate availability and suppress merging problems.

At the time we are writing this chapter, SBO is exported in the OBO flat format (http://www.godatabase.org/dev/doc/obo_format_spec.html). This format is rather unstructured, easely human-readable and shared by the majority of OBO ontologies. However, these qualities make the format a rather poor substrate for automated treatments, particularly in our case, where a portion of the content in in a highly structured form (MathML). At the time the chapter will be published however, it is likely that SBO will alsop be exported in OBO-XML and OBO-OWL. OBO-XML contains the same information that OBO-flat, but is expressed in eXtensible Markup Language [18], that permits extensive computing treatment. In addition it will make the incorporation of the MathML component of the mathematical expression branch trivial.
Finally, we seek to export SBO also using the Web Ontology Language (http://www.w3.org/2004/OWL/). OWL builds on RDF http://www.w3.org/RDF/ and URIs [7], and adds more vocabulary for describing properties and classes, thus improving the semantics of the format and facilitating automated interpretation.

We are also developing WebServices that will allow software to process SBO, or use it either to annotate dataset, or to interpret their annotation.

Conclusion and Perspectives

The need to exchange and integrate models drove the community to design common data format such as SBML. However, as important as was the definition of a common syntax, we also need to tackle the semantics of the models. It is expected that the adoption of MIRIAM and the Systems Biology Ontology will enhance the semantic content of quantitative biochemical description and favour their reusability.

Acknowledgements

BioModels.net and the associated projects has been initiated by Michael Hucka, Andrew Finney and NLN. We are grateful to all the people from the community who gave us their feedback. MC and CL are supported by a grant from the National Institute of General Medical Sciences (USA).

References

[1] A Felsenfeld, J Peterson, J Schloss, and M Guyer. Assessing the quality of the dna sequence from the human genome project. *Genome Res,* **9**:1–4, 1999.

[2] Nicolas Le Nov'ere, Benjamin Bornstein, Alexander Broicher, M'elanie Courtot, Marco Donizelli, Harish Dharuri, Lu Li, Herbert Sauro, Maria Schilstra, Bruce Shapiro, Jacky L Snoep, and Michael Hucka. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res*, **34** (Database issue): D 689 –D 691, 2006.

[3] Nicolas Le Nov'ere, Andrew Finney, Michael Hucka, Upinder S. Bhalla, Fabien Campagne, Julio Collado-Vides, Edmund J. Crampin, Matt Halstead, Edda Klipp, Pedro Mendes, Poul Nielsen, Herbert Sauro, Bruce Shapiro, Jacky L. Snope, Hugh D. Spence, and Barry L. Wanner. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology*, **23(12)**:1509–1515, 2005.

[4] M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, J.C. Doyle, A.P. Arkin, B.J. Bornstein, D. Bray, A. Cornish-Bowden, A.A. Cuellar, S. Dronov, E.D. Gilles, M. Ginkel, V. Gor, I.I. Goryanin and W.J. Hedley, T.C. Hodgman, J.-H. Hofmeyr, P.J. Hunter, N.S. Juty, J.L. Kasberger, A. Kremling, U. Kummer, N. Le Nov'ere, L.M. Loew, D. Lucio and P. Mendes, E. Minch, E.D. Mjolsness, Y. Nakayama, M.R. Nelson, P.F. Nielsen, T. Sakurada, J.C. Schaff, B.E. Shapiro, T.S. Shimizu, H.D.Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics,* **19(4)**:524–531, 2003.

[5] CM Lloyd, MD Halstead, and PF Nielsen. CellML: its future, present and past. *Prog. Biophys. Mol. Biol.*, **85**:433–450, 2004.

[6] T Berners-Lee. Uniform resource locators (URL). a syntax for the expression of access information of objects on the network. Available via the World Wide Web at http://www.w3.org/Addressing/URL/url-spec.txt.

[7] T Berners-Lee, R Fielding, and L Masinter. Uniform resource identifier (URI): Generic syntax. Available via the World Wide Web at http://www.ietf.org/rfc/rfc3986.txt.

[8] R Moats. URN syntax. Available via the World Wide Web at http://www.ietf.org/rfc/rfc2141.txt.

[9] Xiaoshu Wang, Robert Gorlitsky, and Jonas S Almeida. From XML to RDF: how semantic web technologies will change the design of 'omic' standards. *Nat Biotechnol*, **23(9)**:1099–1103, Sep 2005.

[10] N Mitra. Soap version 1.2 part 0: Primer. Available via the World Wide Web at http://www.w3.org/TR/soap12-part0/.

[11] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, **25**:25–29, 2000.

[12] JS Luciano. Pax of mind for pathway researchers. *Drug Discov Today*, **10**:937–942, 2005.

[13] V Henri. Lois G'en'erales de l'Action des Diastases, pages 85–93. Herman, 1903.

[14] L Michaelis and ML Menten. Die kinetik der invertinwirkung. *Biochem. Z.*, **49**:333–369, 1913.

[15] Donald D Van Slyke and Glenn E Cullen. The mode of action of urease and of enzymes in general. *J. Biol. Chem.*, **19**:141–180, 1914.

[16] Ida Schomburg, Antje Chang, and Dietmar Schomburg. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res*, **30(1)**:47–49, Jan 2002.

[17] U Wittig, M Golebiewski, R Kania, O Krebs, S Mir, A Weidemann, S Anstein, J Saric, and I Rojas. SABIO-RK: integration and curation of reaction kinetics data. *Lecture Notes in Bioinformatics*, **4075**:94–103, 2006.

[18] T Bray, J Paoli, CM Sperberg-McQueen, E Maler, F Yergeau, and J Cowan. Extensible markup language (xml) 1.1 (second edition), 2006.

Beilstein-Institut

ESCEC, March 19th – 23rd, 2006, Rüdesheim/Rhein, Germany

Beyond Flat Files: Data Modelling, Editing, Archival and Interchange

Steffen Neumann

Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany

E-Mail: sneumann@IPB-Halle.de

Received: 12th June 2006 / Published: 31st August 2007

Abstract

Software engeneering today provides tools which minimize the need for manual coding of the typical components of an application, such as database, frontend and web application. Visual modelling brings together users and developers, and allows quick and direct communication about the topic. In the metabolomics community data models and XML formats for data interchange such as mzData are currently emerging. Using these standards as a show case, we present an infrastructure to support the use of these data standards and the process of getting there.

Introduction

Most communities in the Life Sciences are facing the problem of how to represent their data in a suitable way. The perfect data model should be flexible, to represent both standard and customized experimental set ups, stringent, to allow for validation and error-detection, machine readable, for storage and retrieval, open to ensure long-term archiving and accessibility, readable by the human eye, for debugging purposes – and of course easy to use. This contribution gives some experience of implementing software and the infrastructure for some emerging community data models.

In recent years metabolomics has become an important technology in solving functional genomics challenges [1] and mass spectrometry (both GC–MS and LC–MS methods) have been adapted to provide high throughput and broad coverage of metabolites [2, 3]. Large-scale metabolomics experiments can produce huge amounts (up to 1 TB per machine per year) of raw data. Structured storage is the key to efficient access to the data for further processing and analysis. In addition to raw mass spectrometry data experimental meta-information is needed to match and compare results from different experiments. A standardized data exchange format allows community-wide collaboration and provides the basis for the large training sets needed in machine learning approaches.

Flat Files have been a commonly used storage model for biological data in the past years. For MS data exchange and as a vendor neutral format, both plain text peaklists or the (binary) netCDF format are being used. Both provide very little metadata – if at all – about the measurement set up, such as machine parameters, software used or by whom the experiment had been conducted. All of this information becomes important if the data is going to be archived for later (re-)processing. However, this requires parsers and converters for each client application processing the data.

Community-wide accepted data standards for interchange are currently emerging, such as mzXML[4] or mzData[5] in the context of the *Proteome Standards Initiative* (PSI). Converters from proprietary vendor file formats to mzData exist for e. g. Applied Biosystems, Bruker, Thermo Finnigan etc. For details see the web site of the Sashimi project[1] and the PSI[2]. The *Architecture for Metabolomics* (ArMet) describes both metadata and results of metabolomics experiments [6], and is compliant with the recommended *Minimum Information about a Metabolomics experiment* (MIAMET)[7]. ArMet has been used in the Setup-X database [8]. All these emerging standards and data models can be used with current software engineering technologies.

The formalism of choice to describe these data models is a UML (class) diagram, which shows the “things” or more formally objects that are to be modelled. Examples in Fig. 3 are the User or a Peaklist. Each object has a set of named attributes of a given data type, such as Name of type String or Creation_Date of type Time Stamp in the example.

An *instance* of this data model consists of the set of objects with values assigned to the attributes. The purpose of a data exchange format is to allow transfer, without loss of information, to other pieces of software or even remote sites. The conversion process is also called serialization.

The *Extensible Markup Language* (XML) is a well-structured markup language. Content encoded in XML can easily be read by XML parsers, which exist for virtually any programming environment. An example of XML is shown in Fig. 1.

[1] http://sashimi.sf.net/
[2] http://psidev.sf.net/ms/

```
<admin>
  <sampleName>tt4_Batch1_1</sampleName>
  <sampleDescription comment="sampleNumber 1">
    <cvParam cvLabel="psi" accession="PSI:1000001"
             name="SampleNumber" value="Arabidopsis Seeds 1" />
    <cvParam cvLabel="psi" accession="PSI:1000006"
             name="SampleConcentration" value="2.57" />
    <cvParam cvLabel="psi" accession="PSI:1000002"
             name="SampleName" value="test sample" />
  </sampleDescription>
  <sourceFile>
    <nameOfFile>tt4_Batch1_1.wiff</nameOfFile>
    <pathToFile>file:/home/sneumann/data/</pathToFile>
    <fileType>Analyst QS 1.1</fileType>
  </sourceFile>
  <contact>
    <name>Christoph Boettcher</name>
    <institution>IPB Halle</institution>
    <contactInfo>+49 (0) 345 5582 0</contactInfo>
  </contact>
</admin>
```

Figure 1. XML excerpt of an mzData entry. Information is given either as an attribute (like cvParam) or in the body (like IPB Halle) of an attribute.

Modelling

Regardless of which software development process (e.g. Waterfall or Extreme Programming) is adopted, during the early phase the purpose of the system needs to be defined. This can be done by describing typical *use cases* or requirements that the software has to fulfil. An example of such a use case for a repository system is given in Fig. 2.

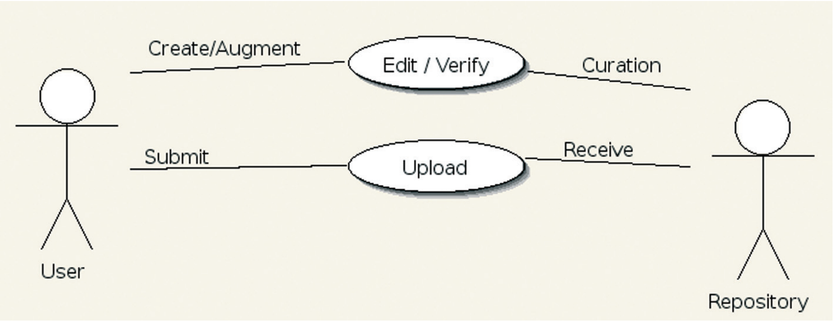


Figure 2. Use Cases for a repository system, with the user and the repository-institution as actors, and the two tasks edit/verify and upload as use cases. The connecting lines are annotated with the role an actor plays.

The actual data model should be created in close dialogue with the customers or users. A suitable model representation also for discussion are UML Class Diagrams. Initially, all "things" of interest should be collected, which will then usually end up as objects or entities in the final model. Next, their relationships have to be defined, such as "Experiment contains measurements" or "A Paper has one corresponding author". The cardinality specifies that an author is *needed* for a paper, otherwise it will be rejected.
For data interchange, files need to be exported and imported on different current and future hardware (Intel, PowerPC) and operating systems (Windows, Unix and others) or sent over networks such as the internet, so the file representation has to ensure that any differences in encoding are either recorded for special treatment or that only the minimal consensus is used. XML is such a file format. The structure of the content can be described using either a *Document Type Definition* (DTD) or – more powerful in its expressiveness – an *XML Schema Definition* (XSD).

Collaborative Development

For the success (or community-wide acceptance) of a data standard a large body of initial contributors and supporters is essential. The development should adopt the release-often-release-early approach also taken in many open source software projects, mentioned as one of the key points in Eric Raymonds essay "The cathedral and the bazaar" [9]. This will invite a broad range of comments and possibly fixes to the development version of a project.

This process of developing open standards and related open source software differs from commercial software development. Without the personal contact and meetings held in a company, there is a need for an efficient collaboration platform, wich supports at least a code repository for sharing the current development, keeping the associated change-logs and allows release management. The other important task is to foster communication between the developers, hosting mailing lists (or equivalent functionality) with archives and search facilities.

The actual choice of platform depends on availability and personal opinion, and can vary between a general-purpose platform such as SourceForge[3] or more targeted environments such as ProteomeCommons Project[4]

[3] http://www.sourceforge.net/

[4] http://www.proteomecommons.org

MS Data Model:Mzdata

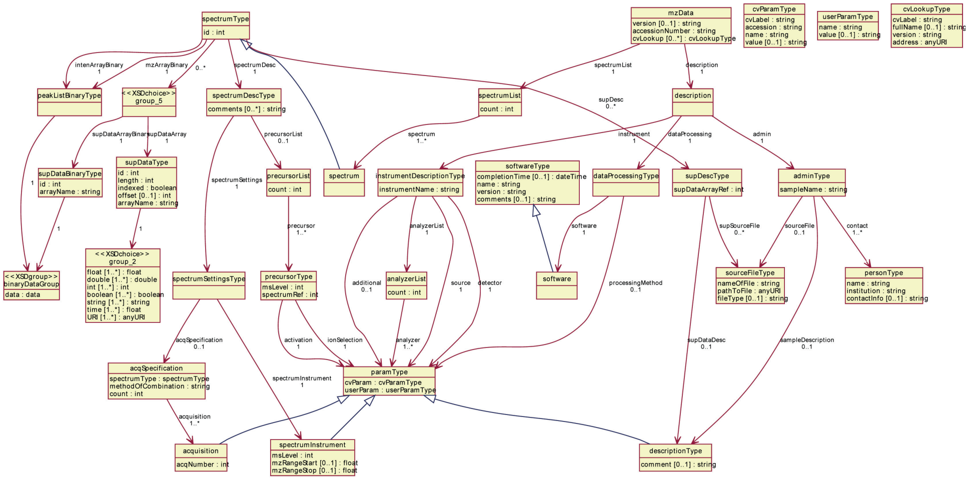


Figure 3. MzData data model. Root object is mzData class, which has descriptive elements and MS data with their own meta-data.

The mzData standard [10] has been designed over the past two years by the MS working group of the *Proteome Standards Initiative* (PSI), with contributions from both academic and industrial members. It is intended mainly as a file exchange format and shares some features with the mzXML format which has been developed initially at the Institute for Systems Biology in Seattle (ISB).

A UML view of the mzData schema is shown in Fig. 3, which has reached version 1.05 since January 2005[5]. The schema covers an administrative description (such as a contact person or sample ID), a set of mass spectrometry relevant parameters (instrument description, ion source, resolution etc.) and a description of the software (-pipeline) and relevant arguments that have been involved in creating the file at hand. Finally, the model contains (a set of) base64 encoded fields with the binary representation of the peaklists (mass, intensity and optionally further supplementary data, e. g. peak quality etc.).

The files are created either from the instrument software directly, or through converters for the instrument specific file formats to mzData. Sometimes, where only support for the mzXML format is available, mzXML can be used as an intermediate with a subsequent conversion through an mzXML to mzData converter. Up to date information is available on the respective project web sites.

[5] A revision of mzData is under review, and expected later in 2006.

Sample Implementation

The implementation is focused around the mzData model, since mzData has been created and described through a model in the *Unified Modeling Language* (UML) (see Fig. 3) and is available as XML schema. This description includes data types, classes, inheritance and constraints. First we describe the use cases for a simple mass spectrometry repository, then details on the third-party libraries and components are shown.

Use cases

The following use cases briefly define which actions should be supported by the infrastructure and applications for a MS repository. The six use cases underlying the implemented applications are:

Use Case1: Preparation for submission

A step which is necessary after an experiment has been performed, and the raw data has been converted to mzData. Depending on the converter, some fields might be filled with default or dummy values, such as `<institution\s) Not set </institution\s)` or `<cvParam cvLabel=\)psi\) accession=\)PSI:1000002\) name=\)SampleName\) value=\)test sample\)/\s)` Such values need to be edited before uploading to a repository: the biologist loads the generated mzData file into an editor and checks the metadata. Once these have been corrected and fields added, the file needs to be verified against the schema and the defined constraints. If necessary, the file has to be edited until it passes validation.

Use Case 2.1: Submission of data

This involves both the biologist and the repository system. The biologist selects a file for upload to the submission form of the repository via a normal web browser. The repository validates the data against the schema and accepts or rejects the file. Finally, the data is persisted in the RDBMS.

Use Case 2.2: Batch import

Batch import is needed by the administrators if a large collection of data files need to be added to the database. A command line tool reads the files and persists them in the database.

Use Case 3: Curation of data

This is performed by the repository's curators, and is necessary if data needs to be changed after submission upon request, or to ensure the data quality. The curator connects to the database, selects an entry and reviews the corresponding values. Changes are persisted in the database, and a validation step guarantees consistency with the mzData schema.

Use Case 4: Browsing the repository

Allows members of the community to list and search the data in the public repository, and to download the corresponding XML file.

Use Case 5: Processing of stored MS data
A user can request this through the XCMS tool at the repository web site. After browsing the repository as described in Use Case 4, the (set of) mzData entries for processing is selected, and XCMS-specific parameters are adjusted. The raw MS signals are processed, retention times are aligned onto a common basis and the results are presented both in a tabular and graphical form.

Software Choices

Creating such a large system would be impossible without (re-)using a range of third-party libraries and tools. In this section I describe those that we have chosen for our MetWare system.

The type of databases most commonly used today are *Relational Database Management Systems* (RDBMS). The data is stored in tables, where each row is an entry, having its attribute values in the columns. Data manipulation and queries are formulated in the *Structured Query Language* (SQL), which declares 1) which tables are used in a query, 2) how they are to be joined, and 3) which attributes are extracted.

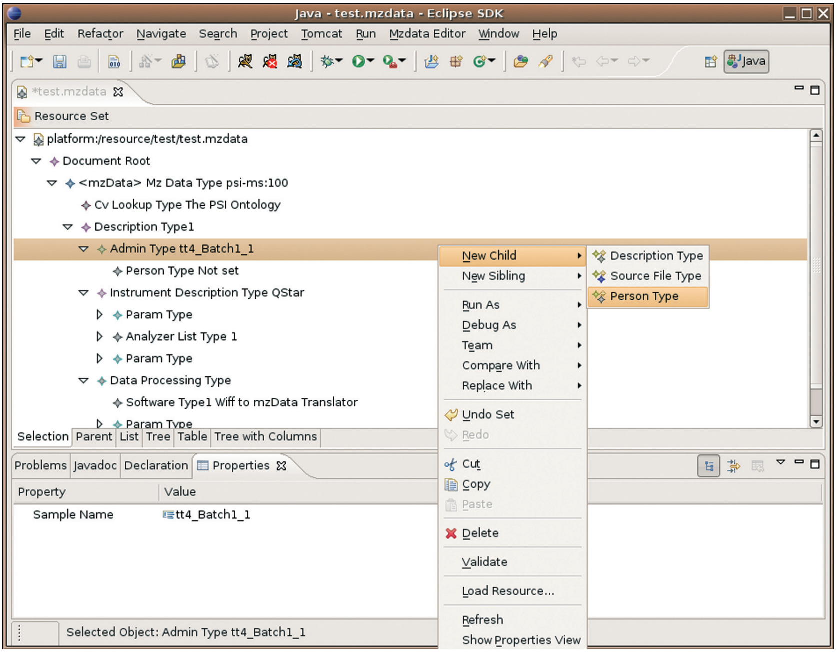
The connection between the database and the clients is done though the *Java Database Connectivity* (JDBC) for Java based standalone clients or the Web frontend, or via *Open Database Connectivity* (ODBC) libraries for non-Java clients. These layers eliminate the need to use proprietary client APIs and wrap them if the database is exchanged for a different brand. Though the connection is vendor-independent, the SQL dialects are not, and common pitfalls exist when e. g. porting a MySQL query to Oracle. Another layer of indirection introducing database adapters can convert generic query statements into the vendor-specific dialect.

Model Driven Architecture

In a Model Driven Architecture the data model is defined as a *Platform Independent Model* (PIM) and afterwards transformed into a *Platform Specific Model* (PSM) for a specific architecture and language.

The Eclipse Modelling Framework[6] provides code generation facilities for Java classes implementing the model, adapters for viewing, change notification and undo capabilities and a basic editor with validation against the model schema. EMF has been used to import the mzData model and create the model implementation and editor. However, the EMF itself does not provide database persistence.

[6] http://www.eclipse.org/emf/


Java - test.mzdata - Eclipse SDK
File Edit Refactor Navigate Search Project Tomcat Run Mzdata Editor Window Help
Java
*test.mzdata
Resource Set
platform:/resource/test/test.mzdata
Document Root
<mzData> Mz Data Type psi-ms:100
Cv Lookup Type The PSI Ontology
Description Type1
Admin Type tt4_Batch1_1
Person Type Not set
Instrument Description Type QStar
Param Type
Analyzer List Type 1
Param Type
Data Processing Type
Software Type1 Wiff to mzData Translator
Param Type
Selection Parent List Tree Table Tree with Columns
Problems Javadoc Declaration Properties
New Child
New Sibling
Run As
Debug As
Team
Compare With
Replace With
Undo Set
Redo
Cut
Copy
Paste
Delete
Validate
Load Resource...
Refresh
Show Properties View
Description Type
Source File Type
Person Type

Property	Value
Sample Name	tt4_Batch1_1

Selected Object: Admin Type tt4_Batch1_1

Figure 4. Generated Editor for mzData. The entry is shown as a tree-view, with properties (values) of the tags shown at the bottom of the window. Context sensitive menus provide schema-compatible insertion of children or siblings and the verification of a (sub-)tree.

The persistence of the EMF objects is handled through an object relational mapping. *Java Data Objects* (JDO) from Sun[7] offer access to different data stores and manage transactions. Persisted data can be queried and transformed into native Java programming language objects. JPOX[8] is the reference implementation of the JDO2.0 specification and can attach to most available relational databases. The eclipse plugin from Springsite[9] generates the metadata for JDO and integrates code that readily allows the editor frontend to be used on data stored in the database.

The presentation layer of the web application is implemented using *Java Server Faces* (JSF) from Sun[10], which provide the framework for handling user sessions, lifecycle of backing objects and navigation between the pages. JSF Tag libraries provide additional widgets which can be used to present tree views, show popup help or integrate a layout templating engine.

[7] http://java.sun.com/products/jdo/
[8] http://www.jpox.org/
[9] http://elver.org/
[10] http://java.sun.com/javaee/javaserverfaces/

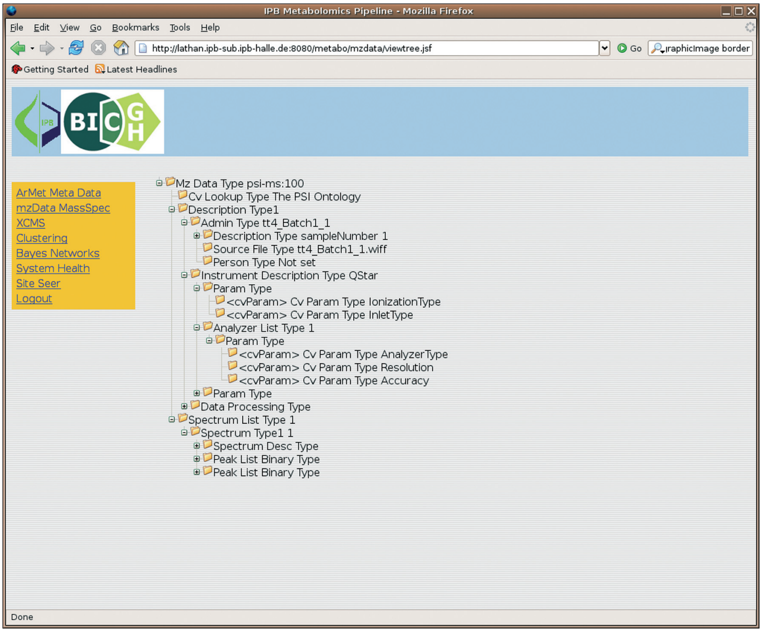
Data Processing

For analysis software written in Java that can readily incorporate and use the JDO libraries, data access can be done in the *JDO Query Language* (JDOQL), similar to the SQL query language.

For signal processing tasks mentioned in Use Case 5 (alignment of retention time shifts and higher level analysis) we integrate a backend service using the XCMS package from the statistics software R and Bioconductor [11] project. XCMS performs peak picking, retention time alignment of multiple LC–MS or GC–MS runs and generates a list of differential mass signals. Communication between R and the application server is done via the Rserve protocol[11]. To connect XCMS to the database backend, we created SQL queries which retrieve the binary data from the RDBMS and feed it into the modified mzData parser. For a detailed description of XCMS see [12].

Results

We have focused on the creation of the backend storage and applications for the use cases Use Case 1 to Use Case 3. In the following paragraphs we describe first experience with the implementation.



```
Mz Data Type psi-ms:100
  Cv Lookup Type The PSI Ontology
  Description Type1
    Admin Type tt4_Batch1_1
      Description Type sampleNumber 1
      Source File Type tt4_Batch1_1.wiff
      Person Type Not set
    Instrument Description Type QStar
      Param Type
        <cvParam> Cv Param Type IonizationType
        <cvParam> Cv Param Type InletType
      Analyzer List Type 1
        Param Type
          <cvParam> Cv Param Type AnalyzerType
          <cvParam> Cv Param Type Resolution
          <cvParam> Cv Param Type Accuracy
      Param Type
    Data Processing Type
  Spectrum List Type 1
    Spectrum Type1 1
      Spectrum Desc Type
      Peak List Binary Type
      Peak List Binary Type
```

[11] http://stats.math.uni-augsburg.de/Rserve/

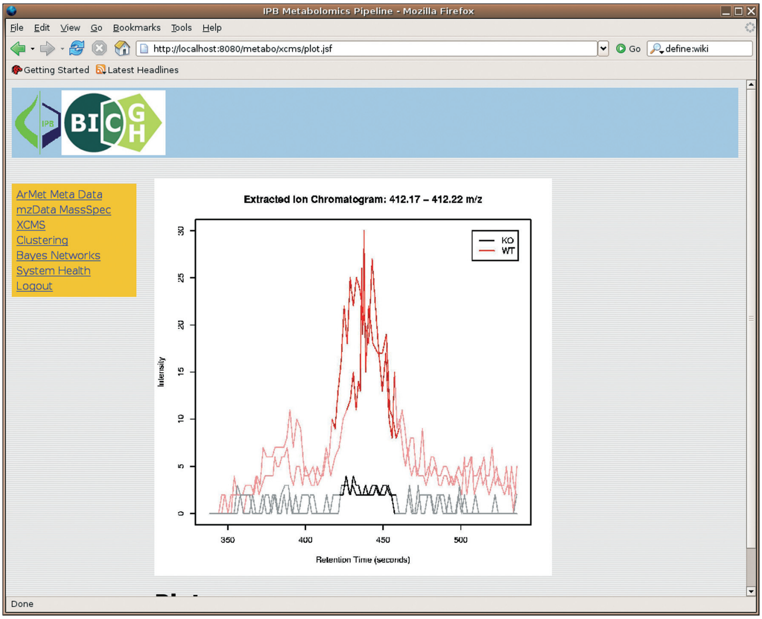


Figure 5. The web interface for the mzData model, showing a part of the tree view. Example of XCMS output: aligned raw data for a differential mass signal.

The editor for Use Case 1 (preparation of mzData XML files for submission) was the first to be finished using EMF. A screenshot is shown in Fig. 4. It can easily handle data files of around 100 MB, which had been acquired on our LC–MS set up using an Applied Biosystems QStar mass spectrometer and were transformed from the instrument-specific wiff format to mzData with a vendor supplied converter. The validation of said 100 MB file is completed in less than a second and is no additional burden to the biologist. The editor can be downloaded at http://msbi.ipb-halle.de/.

The persistence enabled Editor used for Use Case 3 (Curation) that connects to the RDBMS via JDO offers the same functionality as the standalone version. Since lazy loading is implemented, only the relevant parts of the data are requested from the database. Even large collections can be accessed this way.

The web-system is currently being evaluated and improved to provide a biologist-friendly user interface design for the outlined use cases Use Case 4 and Use Case 5, with modules (see Fig. 5) existing for both of them. The architecture of the system (application-, R-statistics- and database server) allows for an easy integration of high-level analyses. Prototypes for these modules are included in the web application.

Conclusion

The chosen data standards are currently gaining a wider acceptance in the metabolomics community. A flexible software development process is necessary to accommodate frequent changes without the need for manual adaption of the resulting software. The overall system consists of the database, R server and web application server, all of which can run on different machines. To scale to a large number of concurrent users, all three services can be run on a cluster of machines, sharing the load. A common filesystem layout is not needed.

We provide the service to biologists working in our institute and close collaborators. A demo database is available at http://msbi.ipb-halle.de/. In the future we plan to implement a similar system for the ArMet metadata, and tight integration of externally controlled vocabulary and ontologies.

Projects starting a standardization effort should consider modelling their data on a public platform and invite other parties to comment or even participate. Getting the actual model "right" (flexible, stringent, machine-/human-readable and easy to use) can be expected to be the hardest task. The standard should be closely followed by software implementing data capture and handling, with the database access coming last. The MDA approach makes it possible to recreate the necessary code basis and backend database with minimal manual coding, since the data standard is hopefully going to evolve.

Acknowledgements

Nigel Hardy, Helen Jenkins, Chris Taylor, Kai Runte and many others for the ArMet and MzData models. Dierk Scheel, Jürgen Schmidt for their valuable discussions on mass spectrometry and biology.

The work is supported under BMBF grant 0312706G.

References

[1] Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G., Kell, D.B.(2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* **22**(5): 246–252..

[2] Roepenack-Lahaye, E.v., Degenkolb, T., Zerjeski, M., Franz, M., Roth, U., Wessjohann, L., Schmidt, J., Scheel, D., Clemens, S. (2004) Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole Time-of-Flight mass spectrometry. *Plant Physiol.* **134**:548–559.

[3] Roessner, U., Wagner, C., Kopka, J., Trethewey, R., Willmitzer, L. (2000) Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-**mass** spectrometry**.** *Plant J.* **23**:131–142.

[4] Pedrioli, P.G.A., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R., Cheung, K., Costello, C.E., Hermjakob, H., Huang, S., Julian, R.K., Kapp, E., McComb, M.E., Oliver, S.G., Omenn, G., Paton, N.W., Simpson, R., Smith, R., Taylor, C.F., Zhu, W., Aebersold, R.(2004) A common open representation of mass spectrometry data and its application to proteomics research**.** *Nature Biotechnol.* **22**(11):1459–1466.

[5] Orchard, S., Hermjakob, H., Binz, P., Hoogland, C., Taylor, C., Zhu, W., Julian, R.J., Apweiler, R.(2005) Further steps towards data standardisation. *Proteomics* **5**(2):337–339.

[6] Jenkins, H., Hardy, N., Beckmann, M., Draper, J., Smith, A.R., Taylor, J., Fiehn, O., Goodacre, R., Bino, R.J., Hall, R., Kopka, J., Lane, G.A., Lange, B.M., Liu, J.R., Mendes, P., Nikolau, B.J., Oliver, S.G., Paton, N.W., Rhee, S., Roessner-Tunali, U., Saito, K., Smedsgaard, J., Sumner, L.W., Wang, T., Walsh, S., Wurtele, E.S., Kell, D.B. (2004) **A** proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnol.* **22**(12):1601–1606.

[7] Bino, R., Hall, R., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B., Mendes, P., Roessner-Tunali, U., Beale, M., Trethewey, R., Lange, B., Wurtele, E., Sumner, L. (2004) Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* **9**(9):418–425.

[8] Fiehn, O.S.M., Wohlgemuth, G. (2005) Automatic annotation of metabolomic mass spectra by integrating experimental metadata**.** In**:** *Proceedings of DILS 2005*, no. 3615 in Proc. Lect. Notes Bioinformatics, pp.224–239. Springer.

[9] Raymond, E.S. (1999) *The Cathedral and the Bazaar*. O'Reilly & Associates, Inc., Sebastapol, CA, USA.

[10] Orchard, S., Taylor, C., Hermjakob, H., Zhu, W., Julian, R., Apweiler, R. (2004) Current status of proteomic standards development. *Expert Rev. Proteomics* **1**(2):179–183.

[11] Gentleman, R.C., Carey, V.J., B DM, Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y. H., Zhang, J. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5**:R80, [[http://genomebiology.com/2004/5/10/R80]].

[12] Smith, C., Want, E., O'Maille, G., Abagyan, R., Siuzdak, G**.** (2006) XCMS**:** Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Analyt. Chem.* **78**(3):779–787.

Beilstein-Institut

ESCEC, March 19th – 23rd, 2006, Rüdesheim/Rhein, Germany

Representing Enzyme Function in Mechanistically Diverse Enzyme Superfamilies

Scott C.-H. Pegg[1], Shoshana Brown[1], Patricia C. Babbitt[1,2]

[1]Dept. of Biopharmaceutical Sciences and [2]Dept. of Pharmaceutical Chemistry
University of California, San Francisco, 94143, USA

E-Mail: spegg@mako.ucsf.edu

Received: 24th February 2006 / Published: 31st August 2007

Abstract

Computational representation of enzyme function should include the structural elements of enzymes which deliver catalytic ability. This is especially important in mechanistically diverse enzyme superfamilies, whose members catalyze different overall reactions. In such superfamilies, evolutionarily conserved elements of structure can be correlated with only conserved aspects of function. The representation of enzyme function in the Structure-Function Linkage Database, in particular the specific structure-function relationships, at multiple levels of evolutionary conservation, aids in the annotation of enzyme function and in designing enzyme engineering experiments.

Introduction

Computational representations of enzyme function, especially the specific ways in which enzyme structure delivers catalytic function, aids our ability to predict the function of newly sequenced enzymes [1, 2] and in efforts to engineer new functions into existing enzymes. [3] Any such computational representation should have at three main properties. First, it should be rapidly searchable. Second, there should be valid similarity metrics defined between any two reactions, allowing users to identify reactions (or substrates or products) that are similar to other reactions (substrates, products). This ability is especially important in enzyme engineering, where a user desires as a starting structural scaffold an enzyme with a functionality similar to the function being engineered. Third, the specific contributions to function by structural elements of the enzyme (e.g. active site residues) should be represented. This allows users to search for specific mechanistic abilities in potential engineering scaffolds and aids in the annotation of newly sequenced or structurally characterized enzymes.

Several representations of enzyme function are currently available, but they fail to make the explicit connection between enzyme structure and function, especially with regard to how conserved structural elements deliver catalytic abilities. The Enzyme Classification (E.C.) system [4], developed before the wide availability and diversity of crystal structures or enzyme sequences, classifies enzyme function according to the overall reaction catalyzed by an enzyme. While the E.C. representation, a series of four hierarchical numbers, allows for rapid computation and simple similarity functions, it doesn't include the contributions of the enzyme structure. Reactions in the E.C. system are considered independent of enzyme structure, leading to cases where enzymes with very different structure-function relationships are classified as similar and vice versa [5]. Recently developed databases of enzyme reactions such as EzCatDB [6] and MACiE [7] have cataloged a large number of enzyme reactions and, where available, the individual mechanistic steps they're comprised of, including the specific amino acids involved in the reactions. These resources, however, do not provide a representation that has similarity metrics defined upon it, nor do they represent some of the more subtle ways in which enzyme structure can contribute to function (e.g. stabilization of a charged intermediate via backbone dipoles). A computational framework for representing enzyme function in a platform independent manner using XML has recently been proposed [8]. While the motivation behind CMLReact is reasonable, it's unclear how well such a scheme, which remains largely undeveloped, will be able to provide similarity metrics and capture the contributions of enzyme structure. As an extensible scheme, however, there remains the potential for other computational representations of function to be absorbed into the CMLReact format.

We focus here on the computational representation of enzyme function within mechanistically diverse enzyme superfamilies [9]. These superfamilies are sets of homologous enzymes which, while often sharing very little sequence similarity to each other, and often catalyzing different overall reactions with a variety of substrates and products, share the same fold and conserve a specific partial reaction (of some other aspect of mechanism) enabled by a conserved set of residues. Study of these superfamilies, especially their conserved structure-function relationships, provides insights into enzyme evolution and significantly aids enzyme engineering efforts.

Methods

We have created an online resource for the study of mechanistically diverse enzyme superfamilies, the Structure-Function Linkage Database (SFLD) [10, 11]. This database structures enzymes into a three level hierarchy using both structural and functional criteria. At the top (superfamily) level are enzymes that share a common partial reaction step, mediated by conserved elements of structure, while at the bottom (family) level are enzymes that catalyze identical overall reactions, via identical mechanisms, using the same conserved aspects of enzyme structure. The middle (subgroup) level contains sets of enzymes where particular structure-function relationships are shared, and are specific to each superfamily. An example of this hierarchy is shown in Figure 1. The SFLD is a rich resource, containing curated alignments, mechanisms, structures, and sequences of widely divergent enzymes that share conserved structure-function relationships. Most fields are also annotated with evidence codes similar to the Gene Ontology (GO) [12] evidence codes and links to relevant literature references. Due to the time and effort involved in the curation of mechanistically diverse enzyme superfamilies, the SFLD remains a deep resource, containing a wealth of structure-function information about particular superfamilies, as opposed to a broad resource that covers all of enzyme space, although more superfamilies are in the process of being added. The SFLD is freely accessible at http://sfld.rbvi.ucsf.edu.

Computational representation of enzyme function in the SFLD is accomplished primarily through the SMILES/SMARTS [13] representation of small molecules and reactions. Overall reactions are stored as well as their constituent partial reactions. These reactions can be searched rapidly using SMARTS queries, allowing users to search for substructures in substrates and/or products. Figure 2 shows some examples of this type of query. Individual residues involved in delivering function, as well as their specific participation, where known, are stored for every structure, and across all sets of proteins at each level of the SFLD hierarchy. This allows users to quickly align a sequence to a curated alignment and determine from the annotated residue positions if the query sequence is likely to have a similar structure-function relationship.

The value of these capabilities is illustrated by our experiments in annotating structures solved by the Structural Genomics Initiatives (SGI) [14]. We scanned 1,605 structures solved by the SGI using hidden Markov Models (HMMs) [15] built on the curated sequence alignments in the SFLD, and compared their Protein Data Bank (PDB) [16] annotations to our own predictions of function. Our predictions were made according to the level(s) of the SFLD hierarchy for which a HMM matched an SGI sequence and the fraction of annotated conserved active site residues that were matched in the alignment of the sequence to the curated multiple alignment upon which the HMM was built. In some cases, we were able to make very specific predictions of enzyme function which have been validated experimentally by our collaborators. (Gerlt, JA, unpublished)

Results

Table 1 shows the SGI structures which matched at least one HMM in the SFLD. Targets for which our predictions of function agree with the current PDB annotations have a white background. In green are cases where we were able to increase the knowledge about the target protein, adding some information about the reaction the enzyme is likely to perform. In the case of 1WUE and 1WUF, targets annotated as being of unknown function, we accurately predicted their ability to catalyze the synthesis of o-succinylbenzoate, a function that was subsequently confirmed experimentally [11]. Our analysis was also able to identify target 1UIY as having been misannotated (orange background in Table 1). This target, while aligning well to the enoyl-CoA hydratase family of the SFLD, is missing a critical glutamic acid residue required for catalysis [17].

A key aspect of the organization of enzyme structure-function relationships within the SFLD is that it allows annotation at multiple levels of granularity. In some cases we can make predictions of overall function with some certainty (such as with 1WUF), but in others we can only state that the enzyme performs a partial reaction conserved throughout the subgroup or superfamily (such as with 1RVK). A more comprehensive discussion of our annotation of several of the SGI targets listed in Table 1 has recently been published [11].

Conclusion

The hierarchy of conserved structure-function relationships within an enzyme superfamily helps us not only avoid the overprediction of enzyme function, but also to make guided decisions when performing enzyme engineering. Our representation of enzyme function in the SFLD allows users to rapidly search for similar substrates and products, and through the annotation of functional residues at each level of the SFLD hierarchy to obtain information about how particular aspects of enzyme structure deliver catalytic function. This information can then be used to identify appropriate starting scaffolds [3, 18].

Our current representation of function is somewhat incomplete, however. While rapidly searchable and with adequately defined similarity metrics based upon small molecule chemical similarity, it lacks a formal representation of some aspects of enzyme participation. For example, the terpene synthase superfamily displays a variety of methods of stabilizing the positive charge on the carbocation intermediates of its reactions, including dipole-charge interactions from sidechains and backbone carbonyls, and cation-pi interactions with aromatic sidechains [19]. These aspects are currently stored as text descriptions in a table of conserved residues, a representation that is not amenable to the sort of similarity queries we'd like to make. Ultimately, we desire a representation of enzyme function in which we can quickly answer such queries as, “what are the enzymes that use a backbone carbonyl to stabilize a positive charge?” and “what are the partial reactions in which an lysine acts as a Schiff base?” While such queries can be answered through string matching of the text descriptions of conserved residue function, the results are inconsistent due to the freeform nature of the text field-different curators will describe identical functions in different ways. A more structured representation of the contributions of a particular aspect of an enzyme structure to a given catalytic step is required to accurately answer the sorts of questions posed above.

Our current attempts at such a representation involve development of extensions to the SMILES representation. This allows us to retain some of the major benefits of SMILES, such as its wide acceptance in third party software, which allows us to implement rapid substructure searching and well developed similarity metrics between the chemical structures represented. Work on this extension of SMILES remains an ongoing research project in our laboratory.

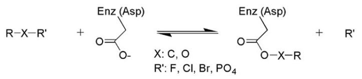
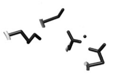
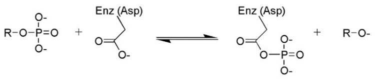
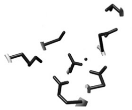
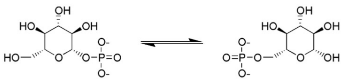
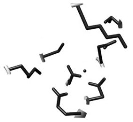
SFLD hierarchy	conserved reaction	conserved functional residues
superfamily Haloacid dehalogenase		
subgroup Phosphatase-like I		
family β-phosphoglucomutase		

Figure 1: An example of the SFLD hierarchy. This example shows the β-phosphoglucomutase family, which belongs to the "phosphatase-like I" subgroup, which in turn belongs to the haloacid dehalogenase superfamily. The middle column shows the conserved reaction across all members of the hierarchical level (row) and the rightmost column shows the active site residues conserved at each level.

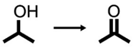
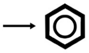
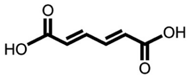
SMARTS Query	Meaning	Schematic
`C([OD1])>>C(=O)`	Reaction contains the conversion of an alcohol to a ketone	
`>>c1ccccc1`	Product of the reaction contains a benzene ring	
`C(=O)([OD1])C=CC=CC(=O)[OD1]`	Either the substrate or the product of the reaction contains mandelate	

Figure 2: Examples of SMARTS queries and their chemical meanings.

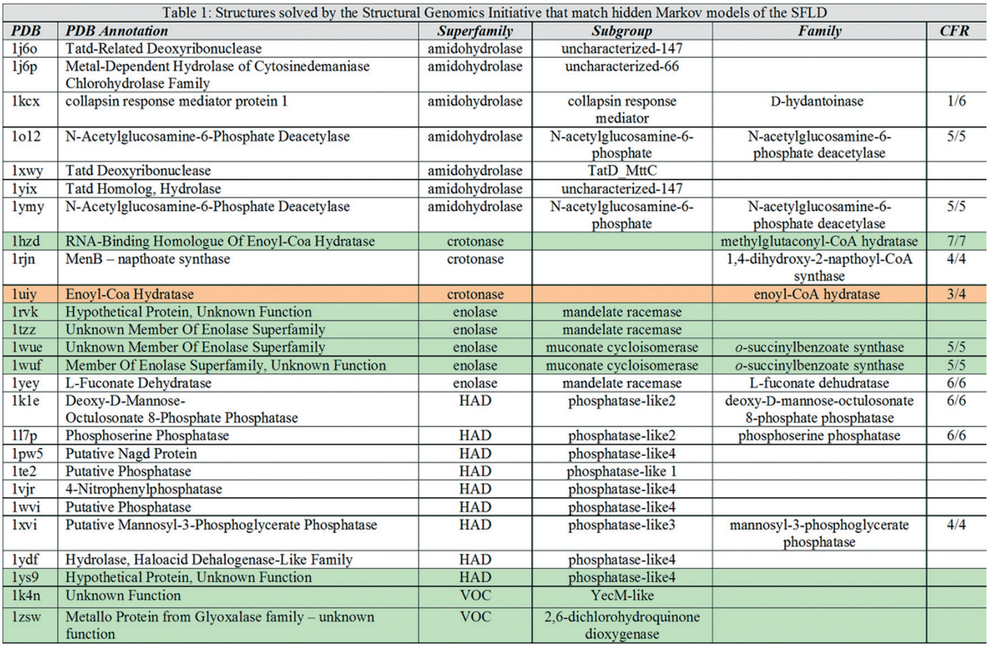


Table 1: Structures solved by the Structural Genomics Initiative that match hidden Markov models of the SFLD

PDB	*PDB Annotation*	*Superfamily*	*Subgroup*	*Family*	*CFR*
1j6o	Tatd-Related Deoxyribonuclease	amidohydrolase	uncharacterized-147		
1j6p	Metal-Dependent Hydrolase of Cytosinedemaniase Chlorohydrolase Family	amidohydrolase	uncharacterized-66		
1kcx	collapsin response mediator protein 1	amidohydrolase	collapsin response mediator	D-hydantoinase	1/6
1o12	N-Acetylglucosamine-6-Phosphate Deacetylase	amidohydrolase	N-acetylglucosamine-6-phosphate	N-acetylglucosamine-6-phosphate deacetylase	5/5
1xwy	Tatd Deoxyribonuclease	amidohydrolase	TatD_MttC		
1yix	Tatd Homolog, Hydrolase	amidohydrolase	uncharacterized-147		
1ymy	N-Acetylglucosamine-6-Phosphate Deacetylase	amidohydrolase	N-acetylglucosamine-6-phosphate	N-acetylglucosamine-6-phosphate deacetylase	5/5
1hzd	RNA-Binding Homologue Of Enoyl-Coa Hydratase	crotonase		methylglutaconyl-CoA hydratase	7/7
1rjn	MenB – napthoate synthase	crotonase		1,4-dihydroxy-2-napthoyl-CoA synthase	4/4
1uiy	Enoyl-Coa Hydratase	crotonase		enoyl-CoA hydratase	3/4
1rvk	Hypothetical Protein, Unknown Function	enolase	mandelate racemase		
1tzz	Unknown Member Of Enolase Superfamily	enolase	mandelate racemase		
1wue	Unknown Member Of Enolase Superfamily	enolase	muconate cycloisomerase	*o*-succinylbenzoate synthase	5/5
1wuf	Member Of Enolase Superfamily, Unknown Function	enolase	muconate cycloisomerase	*o*-succinylbenzoate synthase	5/5
1yey	L-Fuconate Dehydratase	enolase	mandelate racemase	L-fuconate dehudratase	6/6
1k1e	Deoxy-D-Mannose-Octulosonate 8-Phosphate Phosphatase	HAD	phosphatase-like2	deoxy-D-mannose-octulosonate 8-phosphate phosphatase	6/6
1l7p	Phosphoserine Phosphatase	HAD	phosphatase-like2	phosphoserine phosphatase	6/6
1pw5	Putative Nagd Protein	HAD	phosphatase-like4		
1te2	Putative Phosphatase	HAD	phosphatase-like 1		
1vjr	4-Nitrophenylphosphatase	HAD	phosphatase-like4		
1wvi	Putative Phosphatase	HAD	phosphatase-like4		
1xvi	Putative Mannosyl-3-Phosphoglycerate Phosphatase	HAD	phosphatase-like3	mannosyl-3-phosphoglycerate phosphatase	4/4
1ydf	Hydrolase, Haloacid Dehalogenase-Like Family	HAD	phosphatase-like4		
1ys9	Hypothetical Protein, Unknown Function	HAD	phosphatase-like4		
1k4n	Unknown Function	VOC	YecM-like		
1zsw	Metallo Protein from Glyoxalase family – unknown function	VOC	2,6-dichlorohydroquinone dioxygenase		

Table 1: Structures solved by the Structural Genomics Initiative that match hidden Markov models of the SFLD. Targets with a white background have PDB annotations that agree with our annotations using the SFLD. Targets with green backgrounds represent cases in which the SFLD annotations add useful information to the current PDB annotations. Targets with an orange background represent misannotations in the PDB that are corrected by the SFLD annotations. Although targets 1kcx and 1uiy match a family HMM in the SFLD, the fact that they are missing at least one functionally important residue suggests that they do not perform the designated family reaction. (CFR: Conserved Functional Residue)

Acknowledgements

This work was performed with support from NIH grant GM60595 and NSF grant 0234768 (to PCB).

References

[1] Roberts, R.J. (2004). Identifying protein function–a call for community action. *PLoS Biol* **2**, E42.

[2] Saghatelian, A. & Cravatt, B. (2005). Assignment of protein function in the post-genomic era. *Nat Chem Biol* **1**, 130–142.

[3] Schmidt, D.M., Mundorff, E.C., Dojka, M., Bermudez, E., Ness, J.E., Govindarajan, S., Babbitt, P.C., Minshull, J. & Gerlt, J.A. (2003). Evolutionary potential of (beta/alpha)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. *Biochemistry* **42**, 8387–93.

[4] Webb, E.C. (1993). Enzyme nomenclature: a personal retrospective. *Faseb J* **7**, 1192–4.

[5] Babbitt, P.C. (2003). Definitions of enzyme function for the structural genomics era. *Curr Opin Chem Biol* **7**, 230–7.

[6] Nagano, N. (2005). EzCatDB: the Enzyme Catalytic-mechanism Database. *Nucleic Acids Res* **33**, D407–12.

[7] Holliday, G.L., Bartlett, G.J., Almonacid, D.E., O'Boyle N, M., Murray-Rust, P., Thornton, J.M. & Mitchell, J.B. (2005). MACiE: a database of enzyme reaction mechanisms. *Bioinformatics*.

[8] Holliday, G.L., Murray-Rust, P. & Rzepa, H.S. (2006). Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions. *J Chem Inf Model* **46**, 145–57.

[9] Gerlt, J.A. & Babbitt, P.C. (2001). Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem* **70**, 209–246.

[10] Pegg, S.C., Brown, S., Ojha, S., Huang, C.C., Ferrin, T.E. & Babbitt, P.C. (2005). Representing structure-function relationships in mechanistically diverse enzyme superfamilies. *Pac Symp Biocomput*, 358–69.

[11] Pegg, S.C., Brown, S., Ojha, S., Seffernick, J., Meng, E.C., Morris, J.H., Chang, P.J., Huang, C.C., Ferrin, T.E., Babbitt, P.C. (2006). Leveraging Enzyme Structure-Function Relationships for Functional Inference and Experimental Design: The Structure-Function Linkage Database. *Biochemistry Epub* ahead of print.

[12] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–9.

[13] Weininger, D.J. (1988). SMILES. 1. Introduction and encoding rules. *Jour. Chem. Inf. Comput. Sci.* **28**, 31–46.

[14] Chance, M.R., Bresnick, A.R., Burley, S.K., Jiang, J.S., Lima, C.D., Sali, A., Almo, S.C., Bonanno, J.B., Buglino, J.A., Boulton, S., Chen, H., Eswar, N., He, G., Huang, R., Ilyin, V., McMahan, L., Pieper, U., Ray, S., Vidal, M. & Wang, L.K. (2002). Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci* **11**, 723–38.

[15] Eddy, S. (1996). Hidden Markov models. *Curr. Op. Struct. Biol.* **6**, 361–365.

[16] Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D. & Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* **58**, 899–907.

[17] Bahnson, B.J., Anderson, V.E. & Petsko, G.A. (2002). Structural mechanism of enoyl-CoA hydratase: three atoms from a single water are added in either an E1cb stepwise or concerted fashion. *Biochemistry* **41**, 2621–9.

[18] Glasner, M.E., Gerlt, J.A., Babbitt, P.C. (2006). Mechanisms of Protein Evolution and Their Application to Protein Engineering. In Advances in Enzymology and Related Areas of Molecular Biology. Wiley & Sons.

[19] Lesburg, C.A., Zhai, G., Cane, D.E. & Christianson, D.W. (1997). Crystal structure of pentalenene synthase: mechanistic insights on terpenoid cyclization reactions in biology. *Science* **277**, 1820–4.

A Universal Rate Equation for Systems Biology

Johann M. Rohwer, Arno J. Hanekom and Jan-Hendrik S. Hofmeyr

Triple-J Group for Molecular Cell Physiology, Department of Biochemistry, Stellenbosch University, Private Bag X1, ZA-7602 Matieland, South Africa

E-Mail: jr@sun.ac.za

Received: 1st October 2006 / Published: 31st August 2007

Abstract

Classical enzyme kinetics, as developed in the 20th century, had as a primary objective the elucidation of the mechanism of enzyme catalysis. In systems biology, however, the precise mechanism of an enzyme is less important; what is required is a description of the kinetics of enzymes that takes into account the systemic context in which each enzyme is found. In this paper we present the generalized reversible Hill equation as a universal rate equation for systems biology, in that it takes into account (i) the kinetic and regulatory properties of enzyme-catalysed reactions, (ii) the reversibility and thermodynamic consistency of all reactions, and (iii) the modification of enzyme activity by allosteric effectors. Setting the Hill coefficient to one yields a universal equation that can successfully mimic the behaviour of various detailed non-cooperative mechanistic models. Subsequently, it is shown that the bisubstrate Hill equation can account for substrate-modifier saturation, in agreement with experimental data from *Bacillus stearothermophilus* pyruvate kinase. In contrast, the classical Monod–Wyman–Changeux (MWC) equation cannot account for this effect. The proposed reversible Hill equations are all independent of underlying enzyme mechanism, are of great use in computational models and should lay the groundwork for a "new" enzyme kinetics for systems biology.

Introduction

One central aim of classical enzyme kinetics has been the determination of an enzyme's mechanism from initial rate studies with varying substrate and product concentrations [1, 2]. Kinetic equations have been derived for almost every conceivable mechanism, using (partial) equilibrium binding or steady-state kinetics, and there has been a continuing focus on experimental analyses that will be able to discriminate between these mechanisms [2]. The focus in enzyme kinetic analyses has thus been on the characterization of individual enzyme mechanisms, and many of the resulting kinetic equations (especially those for cooperative or multi-substrate reactions) are complex and contain numerous parameters.

In the post-genomic era the field of computational systems biology has received increasing prominence. Its aim is to build kinetic models of cellular pathways, with the individual pathway components (e. g. enzymes) quantitatively described by mathematical rate laws. As such, the overall behaviour of the pathway can then be calculated by the models, needing only the properties of the individual enzymes as input. As a consequence, the focus of enzyme kinetics has shifted. For a kinetic model, we require a kinetic rate law that that will describe the response of an enzyme to changes in substrate, product and modifier concentrations; however, for the overall pathway behaviour, the exact enzyme catalytic mechanism is unimportant as long as the enzyme rate as a function of substrate, product and modifier concentrations is adequately described. When building kinetic models from literature data, one is often faced with the problem that enzymes have not been characterized fully. For example, most often only K_m values for substrates are available and the exact mechanism or some of the other kinetic parameters such as K_i values have not been determined. This forces the modeller to make additional assumptions.

The model construction process would thus be greatly facilitated by a generic equation that contains fewer parameters and yet describes the kinetic behaviour of the enzyme adequately. In this paper, we present the reversible Hill equation as a candidate for fulfilling this task. Firstly, the uni-substrate Hill equation is generalized to an arbitrary number of substrates and products. Secondly, the non-cooperative version of the bi-substrate rate equation is shown to successfully describe the behaviour of two more complex detailed mechanistic models, i. e. ordered and ping-pong kinetics. Thirdly, the Hill and Monod–Wyman–Changeux (MWC) models are compared in terms of their description of allosteric modifier behaviour, with specific emphasis on whether the modifier effect saturates. Fourthly, experimental data for pyruvate kinase show modifier saturation, in agreement with the Hill model but not with the MWC model (Section 4). Finally, the implications of this work for computational systems biology are summarized.

A Generalized Reversible Hill Equation

The development of a universal rate equation for systems biology relies strongly on the foundations of the work of Hofmeyr and Cornish-Bowden [3], who generalized the Hill equation for cooperativity [4] to its reversible form. For a reaction $A \Leftrightarrow P$ this reads:

$$v = \frac{V_{\mathrm{f}}\alpha\left(1-\frac{\Gamma}{K_{eq}}\right)(\alpha+\pi)^{h-1}}{\frac{1+\mu^h}{1+\sigma^{2h}\mu^h}+(\alpha+\pi)^h} \tag{1}$$

where α is the concentration of substrate A scaled by its half-saturation constant $A_{0.5}$ ($\alpha = a / A_{0.5}$), π is the concentration of P scaled by $P_{0.5}$, Γ is the mass-action ratio, K_{eq} the equilibrium constant, h the Hill coefficient, μ the concentration of allosteric modifier M scaled by its half-saturation constant $M_{0.5}$, and σ is an interaction factor quantifying the extent to which binding of a modifier molecule affects substrate and product binding to the enzyme, thus leading to allosteric inhibition or activation. Apart from its ability to describe reversible reactions, this equation is significant in that it takes into account-and separates-the thermodynamic, kinetic and regulatory properties of the reaction [3]. One particularly useful aspect of this equation is the operational definition of its half-saturation constants. For example, at zero product and in the absence of modifier ($\pi = \mu = 0$), setting the concentration of A equal to its half-saturation constant ($\alpha = 1$), yields $v/V_f = 0.5$. The value of $A_{0.5}$ can thus easily be determined in an experiment as that concentration of substrate which yields half of the limiting (maximal) rate.

By following a similar approach as in [3], we have derived the reversible Hill equation for the two-substrate two-product (bi-bi) reaction $A_1 + A_2 \Leftrightarrow P_1 + P_2$ ([5]; Rohwer *et al.*, in preparation). For the case without allosteric modification this equation reads:

$$v = \frac{V_{\mathrm{f}}\alpha_1\alpha_2\left(1-\frac{\Gamma}{K_{eq}}\right)(\alpha_1+\pi_1)^{h-1}(\alpha_2+\pi_2)^{h-1}}{\left(1+(\alpha_1+\pi_1)^h\right)\left(1+(\alpha_2+\pi_2)^h\right)} \tag{2}$$

with the equation parameters and half-saturation constants defined as in Equation 1. By deriving the equation for the three-substrate three-product (ter–ter) case and extending the general pattern, it is possible to obtain a reversible Hill equation describing a reaction comprising an arbitrary number of substrate–product pairs [5]. For the reaction $A_1 + A_2 + ... + A_{\mathrm{n}} \Leftrightarrow P_1 + P_2 + ... + P_{\mathrm{n}}$ this equation reads:

$$v = V_{\mathrm{f}}\prod_{i=1}^{n_s}\alpha_i\left(1-\frac{\Gamma}{K_{eq}}\right)\prod_{i=1}^{n_s}\left(\frac{(\alpha_i+\pi_i)^{h-1}}{1+(\alpha_i+\pi_i)^h}\right) \tag{3}$$

where n_s is the number of substrate-product pairs, and other parameters defined as in Equations 1 and 2. Moreover, it has been possible to obtain Hill equations for the one-substrate two-product (uni–bi) and bi–uni, as well as the bi–ter and ter–bi cases, which broadens the level of applicability of the reversible Hill equation. Equations 1 – 3 can all be transformed to their non-cooperative counterparts by setting the Hill coefficient h equal to one. The derivations will not be shown here, but will be published in detail elsewhere (Rohwer *et al.*, in preparation); the reader is also referred to the Master's thesis of Hanekom, [5] which can be obtained from the authors of this paper on request.

Equations 1 – 3 thus constitute a set of reversible Hill equations that should be able to describe most, if not all, enzyme-catalysed reactions occurring in cellular pathways. All the relevant combinations of different numbers of substrates and products are covered, and by varying the Hill coefficient, the equation can describe reactions exhibiting positive, negative or no cooperativity. From this perspective, the equation indeed appears "universal" in terms of its applicability to cellular reactions. Yet, to be truly worthy of the label "universal", it is insufficient merely to be able to apply the equation to all reactions; the equation will also have to exhibit *realistic* kinetic properties. This will be addressed in the following sections in two different ways: first, we investigate the behaviour of the non-cooperative generalized equation and compare it to more detailed mechanistic models; and next, we compare the behaviour of the cooperative, allosterically inhibited equation to the MWC model and validate the results with experimental data from pyruvate kinase.

Non-Cooperative Bi-Substrate Kinetic Models

The kinetics of enzymes with two or more substrates have been studied in great detail. The field was pioneered by Cleland, who developed kinetic formulations for most conceivable mechanisms in the early 1960 s [6]. An important focus of this original work was to be able to derive the mechanism of an enzyme-catalysed reaction from kinetic studies, and any differences in kinetic behaviour between the various mechanistic models were thus exploited. The fact that reactions can proceed by (partial) equilibrium binding or steady-state kinetics, and that the mechanism can proceed via a ternary complex or substituted enzyme, leads to a multitude of possible formulations, which have been summarized comprehensively by Segel in the definitive textbook on the topic [2].

Ordered vs. ping-pong kinetics

A kinetic mechanism for a bi-substrate reaction that proceeds via a ternary complex with compulsory order binding differs substantially from a mechanism that proceeds via a substituted enzyme. In the former case (termed "ordered" mechanism from here on), both substrates are bound to the enzyme before catalysis occurs and the products are released. Moreover, the substrates cannot bind randomly but rather have to bind to the enzyme in a fixed sequence; likewise, the products are released in a fixed order. For the reaction $A + B \Leftrightarrow P + Q$ this can be symbolized as follows (*cf.* [1]):

$$E \Leftrightarrow EA \Leftrightarrow [EAB \cdot EPQ] \Leftrightarrow EQ \Leftrightarrow E \qquad (4)$$

and leads to the following formulation of a rate equation [1, 2, 6]:

$$\frac{v}{V_{\mathrm{f}}} = \frac{\frac{ab}{K_{\mathrm{iA}}K_{\mathrm{mB}}}\left(1-\frac{pq/(ab)}{K_{\mathrm{eq}}}\right)}{\begin{array}{l}1+\frac{a}{KiA}+\frac{K_{\mathrm{mA}}b}{K_{\mathrm{iA}}K_{\mathrm{mB}}}+\frac{K_{\mathrm{mQ}}p}{K_{\mathrm{mP}}K_{\mathrm{iQ}}}+\frac{q}{K_{\mathrm{iQ}}}+\frac{ab}{K_{\mathrm{iA}}K_{\mathrm{mB}}}+\frac{K_{\mathrm{mQ}}ap}{K_{\mathrm{iA}}K_{\mathrm{mP}}K_{\mathrm{iQ}}}+\frac{K_{\mathrm{mA}}bq}{K_{\mathrm{iA}}K_{\mathrm{mB}}K_{\mathrm{iQ}}}\\ +\frac{pq}{K_{\mathrm{mP}}K_{\mathrm{iQ}}}+\frac{abp}{K_{\mathrm{iA}}K_{\mathrm{mB}}K_{\mathrm{iP}}}+\frac{bpq}{K_{\mathrm{iB}}K_{\mathrm{mP}}K_{\mathrm{iQ}}}\end{array}} \tag{5}$$

By contrast, in the substituted enzyme mechanism (also termed “ping-pong” mechanism) the enzyme is modified after the first substrate molecule has bound and product molecule has been released. The modified enzyme then binds the second substrate, and upon catalysis and release of the second product, the original enzyme is returned. Mechanistically this is symbolized as follows, with *E*((prime)) denoting the modified enzyme:

$$E \Leftrightarrow [EA \cdot E^{'}P] \Leftrightarrow E^{'} \Leftrightarrow [E^{'}B \cdot EQ] \Leftrightarrow E \tag{6}$$

and leads to the following rate equation formulation for the ping-pong mechanism [1,2,6]:

$$\frac{v}{V_{\mathrm{f}}} = \frac{\frac{ab}{K_{\mathrm{iA}}K_{\mathrm{mB}}}\left(1-\frac{pq/(ab)}{K_{\mathrm{eq}}}\right)}{\frac{a}{K_{\mathrm{iA}}}+\frac{K_{\mathrm{mA}}b}{K_{\mathrm{iA}}K_{\mathrm{mB}}}+\frac{p}{K_{\mathrm{iP}}}+\frac{K_{\mathrm{mP}}q}{K_{\mathrm{iP}}K_{\mathrm{mQ}}}+\frac{ab}{K_{\mathrm{iA}}K_{\mathrm{mB}}}+\frac{ap}{K_{\mathrm{iA}}K_{\mathrm{iP}}}+\frac{K_{\mathrm{mA}}bq}{K_{\mathrm{iA}}K_{\mathrm{mB}}K_{\mathrm{iQ}}}+\frac{pq}{K_{\mathrm{iP}}K_{\mathrm{mQ}}}} \tag{7}$$

Is the universal generic equation a good enough approximation?

The fact that ordered and ping-pong mechanisms and their associated rate equations are quite dissimilar, prompted us to investigate whether these differences result in markedly altered kinetic behaviour and, hence, are important from a systems biological perspective. Moreover, both Equations 5 and 7 contain numerous parameters (a K_i value, in addition to a K_{m} value, for each substrate and product), which have seldom all been determined in kinetic characterizations reported in the literature. When constructing kinetic models of pathways, this frequently leads to a lack of data, forcing the modeller to assume parameter values, particularly for K_i (for an example from our own work see [7]. In contrast to both the ordered and ping-pong models, the corresponding bi-substrate generic equation, which is derived from the reversible Hill equation (Equation 2) with $h = 1$, has fewer parameters, i.e. only a K_{m} value for each substrate and product and no K_i values, and reads as follows:

$$\frac{v}{V_{\mathrm{f}}} = \frac{\frac{ab}{K_{\mathrm{A}}K_{\mathrm{B}}}\left(1-\frac{pq/(ab)}{K_{\mathrm{eq}}}\right)}{\left(1+\frac{a}{K_{\mathrm{A}}}+\frac{q}{K_{\mathrm{Q}}}\right)\left(1+\frac{b}{K_{\mathrm{B}}}+\frac{p}{K_{\mathrm{P}}}\right)} \tag{8}$$

We thus set out to investigate whether the kinetic behaviour of both the ordered and ping-pong mechanisms could be described equally well by the generic equation. To do this, data sets were generated with both the ordered and ping-pong mechanistic equations (Equations

5 and 7) by varying both substrates and both products independently over two orders of magnitude for three parameter sets (i.e. combinations of K_m and K_i values). The generic equation (Equation 8) was then fitted to both the ordered and the ping-pong data and goodness of fit assessed by the r^2 value [8]. As can be seen from the representative example in Fig. 1, the quality of fit was near perfect in some cases, and in all cases the r^2 value was grater than 0.94, indicating that the generic equation is capable of successfully mimicking the kinetic behaviour of both the ordered and the ping-pong mechanistic models.

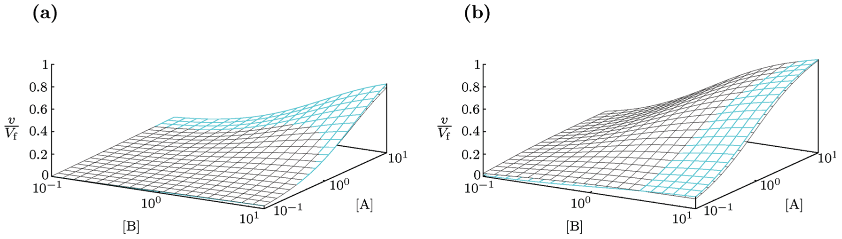


Figure 1. Examples of good fits of the generic rate equation to ordered and ping-pong model data. The generic bi–bi rate equation was fitted to data (a) from the ordered rate equation (Equation 5) with the following parameters: $K_{mA} = 3.3$, $K_{mB} = 3.3$, $K_{mP} = 0.83$, $K_{mQ} = 0.83$, $K_{iA} = 1.0$, $K_{iB} = 3.0$, $K_{iP} = 7.5$, $K_{iQ} = 10.0$; and (b) from the ping-pong rate equation (Equation 7) with the following parameters: $K_{mA} = 1.0$, $K_{mB} = 1.0$, $K_{mP} = 1.0$, $K_{mQ} = 10.0$, $K_{iA} = 1.0$, $K_{iB} = 1.0$, $K_{iP} = 1.0$, $K_{iQ} = 10.0$. K_{eq} was fixed at 10 in all cases. The original ordered and ping-pong data are indicated in black, the generic bi–bi fitted model in cyan. The generic equation was fitted on the complete data set where both substrates and both products were varied independently over two orders of magnitude. The plots show the fits at $p = q = 0.1$. Reproduced from [8] with permission from the Institution of Engineering and Technology.

While the fit was not always as good as in Fig. 1, it could be improved considerably by reducing the range of product concentrations included in the analysis. Since we varied both substrates and products over two orders of magnitude, this analysis really presents a "worst-case" scenario and in many cases of real-life kinetic models, the variation in substrates and products will be less.

This section has evaluated the performance of the universal rate equation in non-cooperative cases. In the next section, we investigate how the equation fares when dealing with cooperative and allosteric kinetics.

Cooperativity and Allosteric Modifier Saturation

Enzymes following normal Michaelian kinetics require an 81-fold increase in substrate concentration to "switch on" (increase their rate from 0.1 Vf to 0.9 Vf). By comparison, enzymes that obey cooperative Hill kinetics only need a 9-fold increase in substrate concentration for the same effect (for a Hill coefficient of 2). Cooperative enzymes are thus sensitive to small changes in substrate concentration and it is important that such cooperative enzymes be regulated with a high degree of precision [1]. Inhibition or activation by allosteric effectors is one mechanism that accomplishes such regulation, and it is a ubiquitous motif in metabolic pathways. It therefore becomes important to be able to describe the kinetics of such allosteric enzymes accurately in kinetic models.

Allosterically regulated cooperative enzyme reactions are usually modelled with irreversible MWC kinetics [9]. However, when Hofmeyr and Cornish-Bowden [3] derived the uni-substrate reversible Hill equation (Equation 1 above), they also demonstrated that this equation predicts substantially different allosteric inhibition kinetics, compared to the MWC equation. In particular, the Hill model shows modifier saturation in that at high substrate concentration the allosteric inhibitor ceases to have an effect, whereas the MWC equation does not show this saturation and the inhibitor always has an effect irrespective of the substrate concentration. Allosteric inhibitors in the MWC equation thus behave analogously to competitive inhibitors. The effect was, however, only demonstrated for the uni–uni case.

Since most enzyme-catalysed reactions in biochemical pathways have two or more substrates and products, we first set out to demonstrate that the difference between the Hill and MWC models with respect to allosteric modifier saturation also exists for the bi-substrate case. As the inhibitor saturation effect has to our knowledge not been demonstrated experimentally, we subsequently present data for the allosteric enzyme pyruvate kinase, which also show saturation of the allosteric modifier effect, thus lending support to the Hill equation and contrasting with the MWC model.

Modifier saturation in Hill vs. MWC

The bi-substrate Hill equation for the irreversible reaction $A + B \rightarrow P + Q$ with allosteric modifier M reads as follows [5]:

$$v = \frac{V_{\mathrm{f}}\alpha^h\beta^h}{\left(\frac{1+\mu^h}{1+\sigma^{4h}\mu^h}\right) + \left(\frac{1+\sigma^{2h}\mu^h}{1+\sigma^{4h}\mu^h}\right)[\alpha^h + \beta^h] + \alpha^h\beta^h} \tag{9}$$

with $\beta = b/B_{0.5}$ and the other parameters defined as in Equations 1 and 2. The MWC equation for the same reaction is given by:

$$v = \frac{V_{\mathrm{f}}\left(\frac{[\mathrm{A}][\mathrm{B}]}{K_{mA}K_{mB}}\right)\left(1+\frac{[\mathrm{A}]}{K_{mA}}\right)^{n-1}\left(1+\frac{[\mathrm{B}]}{K_{mB}}\right)^{n-1}}{\left(1+\frac{[\mathrm{A}]}{K_{mA}}\right)^{n}\left(1+\frac{[\mathrm{B}]}{K_{mB}}\right)^{n} + L_0\left(1+\frac{[\mathrm{I}]}{K_i}\right)^{n}} \tag{10}$$

where Vf is the limiting enzyme rate, K_{mA} and K_{mB} are the intrinsic dissociation constants for substrates A and B from the R-form of the enzyme, [I] is the inhibitor concentration, K_i is the intrinsic dissociation constant for inhibitor I from the T-form of the enzyme, n is the number of enzyme subunits and L_0 is the equilibrium ratio of L_0/R_0 in the absence of substrates and products. This equation was derived by simplifying the generalized MWC model of Popova and Sel'kov [10] along the assumptions of the original paper of Monod *et al.* [9]: (i) the reaction is irreversible, (ii) the T-form of the enzyme does not participate in catalysis, (iii) the inhibitor only binds to the T-form, and (iv) the substrates bind only to the R-form of the enzyme, which is catalytically active.

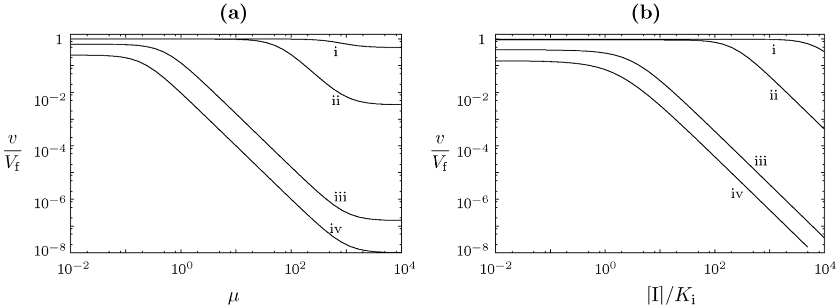


Figure 2. Enzyme activities of the Hill and MWC models as a function of inhibitor concentration at different substrate conditions. (a) Bi-substrate Hill equation (Equation 9) with $h = 2$ and $\sigma = 0.1$. (b) Bi-substrate MWC equation (Equation 10) with $n = 2$ and $L = 10$. Data are plotted in double logarithmic space. Substrates A and B were varied simultaneously; their scaled concentrations are i: 150, ii: 25, iii: 2 and iv: 1. Reproduced from [19] with permission from the Institution of Engineering and Technology.

These two models (Equations 9 and 10) were then compared by plotting the reaction rate as a function of the concentration of allosteric inhibitor for different values of the substrate concentrations, which were increased together (Fig. 2). The results clearly demonstrate that the bi-substrate Hill model shows substate-modifier saturation in that increasing the modifier concentration above a certain threshold (here, $\mu \approx 10^3$) ceases to have an effect on the reaction rate. Moreover, the inhibitory effect is nullified at high substrate concentrations. The bi-substrate MWC model does not show this saturating effect, analogous to the uni-substrate case [3].

Experimental verification of modifier saturation in pyruvate kinase

Since the Hill and MWC models can be clearly distinguished using the effect of modifier saturation, we investigated experimentally whether this effect would be present in a bi-substrate cooperative enzyme. *Bacillus stearothermophilus* pyruvate kinase is a microbial cooperative enzyme that exhibits cooperativity towards its substrate phosphoenolpyruvate (PEP) [11]. The cooperative kinetics, structure and thermal stability of this enzyme have been studied in detail [11 – 13], and it has both allosteric activators and inhibitors. Moreover, it can be conveniently assayed with a simple spectrophotometric protocol [14]. Here, the kinetics of the allosteric inhibitor inorganic phosphate (P_i) were investigated (Fig. 3).

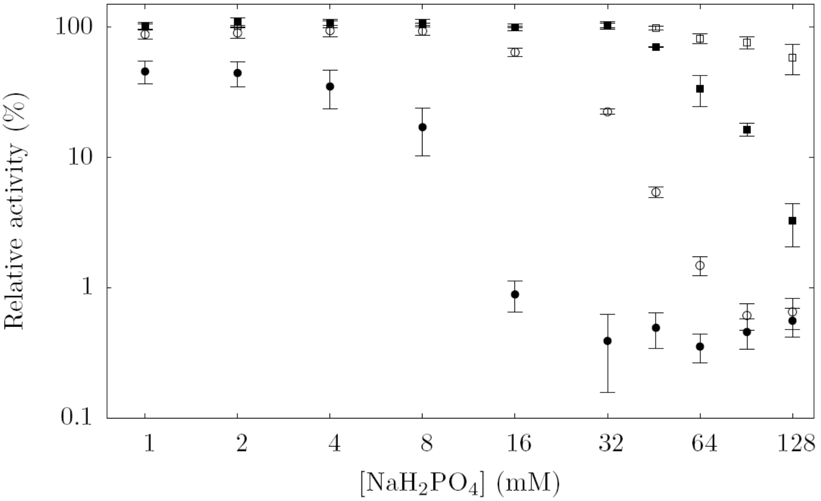


Figure 3. Relative pyruvate kinase activity as a function of inhibitor (NaH_2PO_4) concentration at increasing substrate concentrations. Data were normalized to the limiting rate measured at 20 mM PEP and 20 mM ADP in the absence of inhibitor. Note that data are presented in double-logarithmic space. The assay mixture contained equimolar concentrations of PEP and ADP: 1 mM (●), 4 mM (((odot))), 10 mM (((square))) and 20 mM (((osquare))). All data points are the average of 3 – 5 independent determinations ± SE. Experiments demonstrating the saturation of the inhibitory effect (4 mM substrates at ≥ 91 mM inhibitor; 1 mM substrates at ≥ 32 mM inhibitor) were all performed in five-fold. Reproduced from [19] with permission from the Institution of Engineering and Technology.

At high substrate concentrations, P_i could no longer inhibit the enzyme, even at high levels. In addition, modifier saturation is clearly visible when the substrates were present at 1 mM (P_i concentrations above 32 mM did not inhibit the enzyme further). At 4 mM substrate concentrations, saturation of the inhibitory effect was also visible for $[P_I]^3$91 mM. When comparing these results with the kinetic plots in Fig. 2, it is clear that the data are consistent with the Hill model but not with the classical MWC model.

Discussion and Conclusion

This paper has described a new universal rate equation for systems biology, which is based on the reversible Hill equation. The equation can be written for an arbitrary number of substrate–product pairs, as well as for uni–bi, bi–uni, bi–ter and ter–bi reactions. In addition, an arbitrary number of either independent or competing allosteric modifiers can be treated [5]. By varying the Hill coefficient through values ranging from less than one to greater than one, the equation can exhibit negative cooperativity, no cooperativity (i.e.

Michaelian kinetics) or positive cooperativity. These features make the universal rate equation so generic and versatile that it should be possible, in principle, to use it for describing the kinetics of any enzyme-catalysed reaction.

The derivation of the generalised reversible Hill equation was based on the same assumptions as the uni-substrate case [3], i.e. (i) the limiting case of cooperativity (active sites are either empty or fully occupied, partially liganded enzyme species are not considered), (ii) random equilibrium binding of substrates, products and modifiers to the enzyme, also in the form of dead-end complexes, (iii) independently acting binding sites that do not influence each other, and finally (iv) generalization from the number of subunits (*n*) to the Hill coefficient (*h*), which can take on non-integer values (also less than one).

Although allosteric effects in the generalized reversible Hill equation presented in this paper only affect the binding strength of substrates or products through changing their apparent half-saturation constants (i.e. so-called "K-enzymes"), it should be pointed out that the reversible Hill equation has also been rewritten to include effects of allosteric modifiers on the catalytic properties of an enzyme (i.e. so-called "V-enzymes") [15, 16]. The details are not included here for lack of space; however, they contribute to the universality of the reversible Hill equation in its application to computational systems biology.

The non-cooperative formulation of the universal rate equation is capable of succesfully mimicking the kinetic behaviour of both the ordered and the ping-pong mechanistic models (Fig. 1). The equation for random bi–bi kinetics was not included in the analysis, since its derivation is based on equilibrium binding of substrates and products [2, 6] (the derivations of the ordered and ping-pong models are based on steady-state kinetics). The common ordered bi–bi mechanism is thus identical to our generic bi-substrate model barring the existence of the dead-end complexes, which should therefore lead to an even better correspondence than for the ordered and ping-pong mechanistic equations.

The cooperative version of the equation shows substrate–allosteric modifier saturation, in contrast to the irreversible MWC model, which does not (Fig. 2), and the validity of the reversible Hill model is corroborated by experimental data for the enzyme pyruvate kinase (Fig. 3), which also show modifier saturation. Together, these data provide *in silico* and *in vitro* evidence for validation of the universal equation.

The results are significant for two reasons: First, in general, generic equations based on the reversible Hill equation contain fewer parameters than mechanistic equations. As a result, fewer parameters need to be measured experimentally, which lessens the burden for experimental kinetic characterization. Moreover, the parameters can be determined directly because of the clear operational definition of the half-saturation constants (see Section 2). In contrast, MWC equations, for example, are mechanistic models that contain intrinsic metabolite dissociation constants, which cannot be determined directly in such an operational way, but only through fitting. Secondly, it is unnecessary to know the detailed mechanism of an enzyme in order to simulate its kinetics for modelling. For computational systems biology, enzyme mechanism as such is less important but an accurate kinetic description in terms of quantification of the reaction rate as a function of substrates, products and effectors is crucial.

It should be emphasized that not all MWC equations are unable to account for modifier saturation; it is only the commonly used uni-substrate irreversible formulation [9] and its bi-substrate form (Equation 10) that have this limitation. In fact, we have shown that the generalized MWC model of Popova and Sel'kov [10, 17] gives near indistinguishable behaviour from the generalized reversible Hill equation, including allosteric inhibitor saturation [18]. The reason for this is that in the generalized MWC model [17], all species interacting with the enzyme (be it substrates, products or allosteric effectors) can in principle bind to both the T- and R-forms and both these enzyme forms are catalytically active (albeit to different extents), whereas in the original formulation of Monod, Wyman and Changeux [9], the restrictions outlined below Equation 10 were imposed. However, in experimental applications the original model has been used almost without exception, and the generalized form of Popova and Sel'kov has rarely been applied, which makes the distinction between Hill and MWC equations important.

Although mechanistic equations were derived for initial-rate kinetics (see e.g. [6]), the universal rate equation presented here is not limited to the analysis of initial rates. We have developed a new experimental method to obtain kinetic parameters through fitting of progress curve data obtained from time-course NMR spectroscopy (Hanekom *et al.*, in preparation). This is especially relevant for systems biology, since many of the high-throughput techniques of modern biology (transcriptomics, proteomics, metabolomics) generate such time-series data.

In conclusion, we propose that the universal rate equation presented in this paper should form the basis of a "new" enzyme kinetics for systems biology. It is simpler than mechanistic rate equations, can account for positive, negative or no cooperativity, is thermodynamically consistent and contains fewer parameters than mechanistic equations.

Acknowledgement

This work was supported by the National Bioinformatics Network (South Africa).

References

[1] Cornish-Bowden, A. (1995) *Fundamentals of Enzyme Kinetics*. Portland Press, London.

[2] Segel, I. H. (1975) *Enzyme Kinetics. Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems*. John Wiley and Sons, New York.

[3] Hofmeyr, J.-H. S., Cornish-Bowden, A. (1997) The reversible Hill equation: how to incorporate cooperative enzymes into metabolic models. *Comp. Appl. Biosci.* **13:**377–385.

[4] Hill, A. V. (1910) The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J. Physiol. (Lond.)* **40:**iv–vii.

[5] Hanekom, A. J. (2006) Generic kinetic equations for modelling multisubstrate reactions in computational systems biology. Master's thesis, Stellenbosch University.

[6] Cleland, W. W. (1963) The kinetics of enzyme-catalyzed reactions with two or more substrates or products. I. Nomenclature and rate equations. *Biochim. Biophys. Acta* **67:**104–137.

[7] Rohwer, J. M., Botha, F. C. (2001) Analysis of sucrose accumulation in the sugar cane culm on the basis of *in vitro* kinetic data. *Biochem. J.* **358:**437–445.

[8] Rohwer, J. M., Hanekom, A. J., Crous, C., Snoep, J. L., Hofmeyr, J.-H. S. (2006) Evaluation of a simplified generic bi-substrate rate equation for computational systems biology. *IEE Proc.-Syst. Biol.* **153:**338–341.

[9] Monod, J., Wyman, J., Changeux, J.-P. (1965) On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12:**88–118.

[10] Popova, S. V., Sel'kov, E. E. (1978) Description of the kinetics of two-substrate reactions of the type $S1 + S2 \Leftrightarrow S3 + S4$ by a generalized Monod-Wyman-Changeux model. *Mol. Biol. (Mosk.)* **13:**129–139.

[11] Lovell, S. C., Mullick, A. H., Muirhead, H. (1998) Cooperativity in *Bacillus stearothermophilus* pyruvate kinase. *J. Mol. Biol.* **276:**839–851.

[12] Sakai, H., Suzuki, K., Imahori, K. (1986) Purification and properties of pyruvate kinase from *Bacillus stearothermophilus*. *J. Biochem.* **99:**1157–1167.

[13] Sakai, H., Ohta, T. (1987) Evidence for two activated forms of pyruvate kinase from *Bacillus stearothermophilus* in the presence of Ribose 5-phosphate. *J. Biochem.* **101:**633–642.

[14] Bücher, T., Pfleiderer, G. (1955) Pyruvate kinase from muscle. *Methods Enzymol.* **1:**435–440.

[15] Westermark, P. O., Hellgren-Kotaleski, J., Lansner, A. (2004) Derivation of a reversible Hill equation with modifiers affecting catalytic properties. *WSEAS Trans. Biol. Med.* **1:**91–98.

[16] Hofmeyr, J.-H. S., Rohwer, J. M., Snoep, J. L. (2006) Conditions for effective allosteric feedforward and feedback in metabolic pathways. *IEE Proc.-Syst. Biol.* **153:**327–331.

[17] Popova, S. V., Sel'kov, E. E. (1975) Generalization of the model by Monod, Wyman and Changeux for the case of a reversible monosubstrate reaction *SP. FEBS Lett.* **53:**269–273.

[18] Olivier, B. G., Rohwer, J. M., Snoep, J. L., Hofmeyr, J.-H. S. (2006) Comparing the regulatory behaviour of two cooperative, reversible enzyme mechanisms. *IEE Proc.-Syst. Biol.* **153:**335–337.

[19] Hanekom, A. J., Hofmeyr, J.-H. S., Snoep, J. L., Rohwer, J. M. (2006) Experimental evidence for allosteric modifier saturation as predicted by the bi-substrate Hill equation. *IEE Proc.-Syst. Biol.* **153:**342–345

SABIO-RK (System for the Analysis of Biochemical Pathways-Reaction Kinetics)

Isabel Rojas, Martin Golebiewski, Renate Kania, Olga Krebs, Saqib Mir, Andreas Weidemann and Ulrike Wittig

Scientific Databases and Visualization Group, EML Research GmbH, Heidelberg, Germany

E-Mail: isabel.rojas@eml-r.villa-bosch.de

Received: 14th June 2006 / Published: 31st August 2007

Abstract

SABIO-RK is a database designed to store and offer access to information about biochemical reactions and their kinetics in a comprehensive and standardized manner. It integrates information from several sources to form a backbone of information necessary to include information about the kinetics of biochemical reactions. The kinetic data itself is primarily extracted from literature along with descriptions of the experimental conditions under which they were determined. This process is supported by the use of a web-based user interface which complies with most of the recommendations of the STRENDA committee for reporting on the results of enzyme/reaction kinetics. In this paper we describe the main characteristics of the SABIO-RK and its search and input interfaces.
Availability: http://sabio.villa-bosch.de/sabiork

Introduction

The simulation of biochemical reaction networks depends on the combination of experimental data with modelling methods. A simulation requires information about the kinetics of the biochemical reactions participating in the network, such as the kinetic laws describing the dynamics of the reactions with their respective parameters determined under certain experimental conditions. These data are widely scattered through various publications and described in many different formats. Moreover, each special field uses its own vocabulary and concepts. Thus, the process of integrating the kinetic data to simulate a biochemical network would be enormously facilitated by the definition and use of standards for reporting and exchanging the data obtained, both from experimentalist to modellers and for the feedback from modellers to experimentalists.

In order to compare kinetic data and integrate them into models of biochemical networks, kinetic parameters need to be consistently described and related to the kinetic mechanisms, the equations representing the kinetic laws and the environmental conditions. The known mechanisms of biochemical reactions should be reflected in mathematical formulas, which have to be linked to the corresponding parameters, such as kinetic constants and concentrations of each reaction participant. As kinetic constants highly depend on environmental conditions, they can only be specified completely by describing these conditions used for determination. Data sets based on experiments assayed under similar experimental conditions should be associated to each other to facilitate the comparison.

There is currently a small number of databases containing kinetic data of biochemical reactions. BRENDA [1] is a comprehensive database on information about enzymes. The enzyme entries also contain information about the reactions catalysed by the enzyme including data describing their reaction kinetics and in some cases information about the mechanism associated with the reaction's kinetics. Swiss-Prot [2] started to include experimental data like pH- and temperature dependence and kinetic parameters as comments related to biophysicochemical properties. The BioModels database [3] rather stores published mathematical simulation models of biological interest that are annotated and linked to relevant data resources (e.g. publications or databases), than experimental kinetic data of single reactions. The models include kinetic law equations and their parameters represented in SBML (Systems Biology Mark-up Language) format [4] and can be used for simulations of biochemical reactions or networks.

SABIO-RK, extends and supplements the information content of these databases by storing highly interrelated information about biochemical reactions and their kinetics, this last mainly experimentally obtained. It includes reactants and modifying compounds (i.e. inhibitors or activators) of reactions, information about the catalysing enzymes, and the kinetic laws governing the reactions, the latter with their parameters and information about experimental conditions under which they were determined. Data about biochemical reactions and their rate equations and parameters can be exported in SBML file format.

Most of the above mentioned databases manually obtain their information from publications. Data is typically loaded using in-house software, which has been designed on the basis of the structure of the underlying database. However, the ideal case would be that experimentalists or modellers could use a standard format to report their findings and that this format could be used by the databases to import kinetics data. Systems biologists use SBML format [4] to exchange models of biochemical reactions. However, it does not offer support to describe much of the information that documents the conditions and constraints of a given model or single experiment, unless this information is included in an unstructured open format as a comment or a description of the model. Information such as: under which experimental conditions does the model hold, or for which organism the data is reported, are not supported by SBML. It is planned however that this will change in the near future. In order to facilitate the integration of information the SBML community has incorporated and recommends the annotation of SBML files with references to controlled vocabularies and ontologies (see [5]). The STRENDA [6] (**St**andards for **R**eporting **En**zymology **Da**ta) commission is working on the definition of a standard for reporting on enzyme activity. The standard should contain the minimum amount of information that should accompany any published enzyme activity data. The use of references to controlled vocabularies and ontologies is also of great importance for the implementation of the STRENDA guidelines.

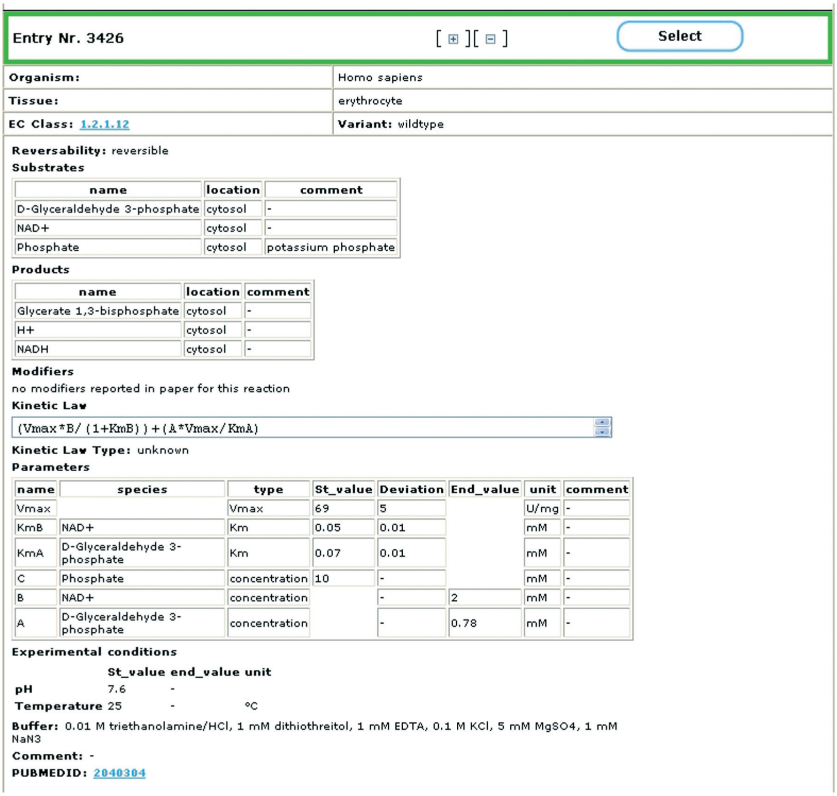
In this paper, we will report on SABIO-RK and the input interface used to load and store information about reactions and enzyme kinetics, and how this interface matches in most points with the current definition of the STRENDA standards especially with respect to the kinetics of enzymes and reactions. This interface would enable scientists to enter the results and conditions of their experiments into the database and to export these using a (to be defined) STRENDA format that can then be used to exchange the data.

SABIO-RK (System for the Analysis of Biochemical Pathways-Reaction Kinetics)

SABIO-RK is an extension of the SABIO (System for the Analysis of Biochemical Pathways) biochemical pathway database, also developed at EML Research [7]. SABIO stores the fundamental information about biochemical pathways, like reactions and their participants (enzymes, compounds, etc.). It also offers support for the storage of information about proteins, protein complexes and genes, all this linked to organism (including strains) and to biochemical reactions (in the case of enzymes). SABIO integrates data from different sources, to establish a broad information basis. Most of the reactions, their associations with biochemical pathways, and their enzymatic classifications (enzyme classifications of the International Union of Biochemistry and Molecular Biology [8]) are extracted from the KEGG database (Kyoto Encyclopaedia of Genes and Genomes) [9].

SABIO-RK combines the data about biochemical reactions stored in SABIO with information about their kinetic properties. The kinetic data is mainly manually extracted from published scientific articles and then verified by curators. A kinetic law – if available in the article – is associated with a biochemical reaction (defined in terms of its substrates, products and modifiers) and its catalysing enzyme (typically defined by an Enzyme Classification number and a description of the enzyme variant, e.g. isoenzyme or mutant). A reaction can have multiple kinetic laws defined within one or multiple publications. This may depend on environmental and experimental conditions, enzyme variants, and the absence or presence of modifiers. As we will see in the next section, a kinetic law entry will contain data about the organism, tissue, and cellular location where the reaction takes place, as well as the type of the kinetic law and the reaction's rate equation. The latter is shown with its parameters and the experimental conditions (e.g. pH, temperature, buffer) under which the parameters were determined or for which the parameters hold.

The SABIO-RK database has been conceived to serve the Systems Biology community as its main user. However it also contains useful information for experimentalist or researchers interested in information about biochemical reactions and their kinetics. It aims to support modellers with high quality data in setting up *in-silico* models describing biochemical reaction networks.



Entry Nr. 3426

Organism:	Homo sapiens
Tissue:	erythrocyte
EC Class: 1.2.1.12	**Variant:** wildtype

Reversability: reversible

Substrates

name	location	comment
D-Glyceraldehyde 3-phosphate	cytosol	-
NAD+	cytosol	-
Phosphate	cytosol	potassium phosphate

Products

name	location	comment
Glycerate 1,3-bisphosphate	cytosol	-
H+	cytosol	-
NADH	cytosol	-

Modifiers

no modifiers reported in paper for this reaction

Kinetic Law

(Vmax*B/(1+KmB))+(A*Vmax/KmA)

Kinetic Law Type: unknown

Parameters

name	species	type	St_value	Deviation	End_value	unit	comment
Vmax		Vmax	69	5		U/mg	-
KmB	NAD+	Km	0.05	0.01		mM	-
KmA	D-Glyceraldehyde 3-phosphate	Km	0.07	0.01		mM	-
C	Phosphate	concentration	10	-		mM	-
B	NAD+	concentration		-	2	mM	-
A	D-Glyceraldehyde 3-phosphate	concentration		-	0.78	mM	-

Experimental conditions

	St_value	end_value	unit
pH	7.6	-	
Temperature	25	-	°C

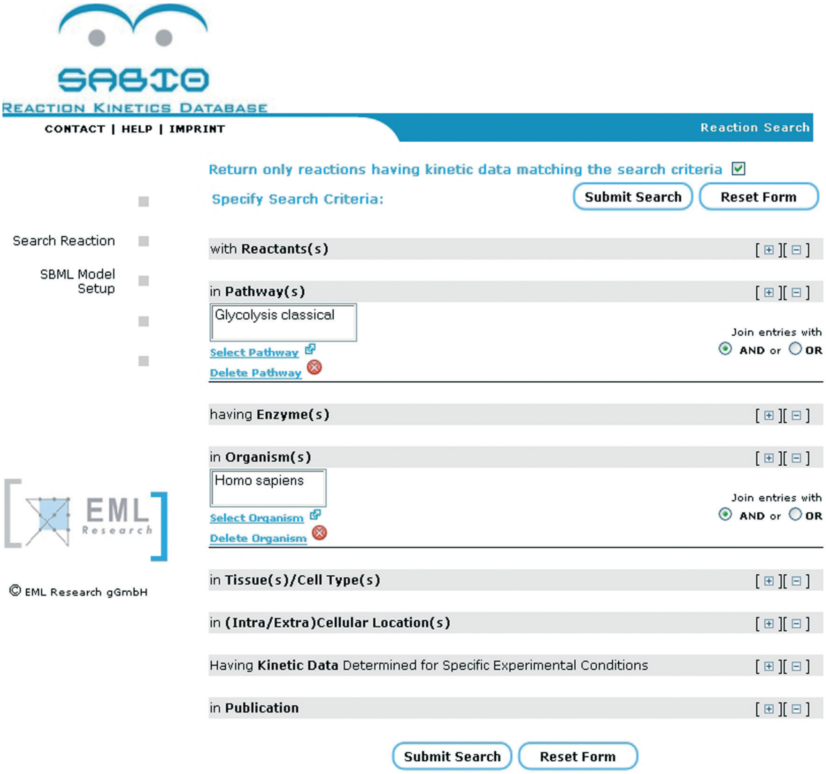
Buffer: 0.01 M triethanolamine/HCl, 1 mM dithiothreitol, 1 mM EDTA, 0.1 M KCl, 5 mM MgSO4, 1 mM NaN3

Comment: -

PUBMEDID: 2040304

Figure 1: General concepts contained in the SABIO-RK database. (We have included plural definitions to facilitate reading.)

Figure 1 shows a general concepts contained in the database (not corresponding to tables in the database) and their relations. The current version the SABIO-RK web interface allows users to perform searches for reactions by specifying characteristics (one or many) of the reactions of interest (Fig. 2). For example the user can specify the pathway to which the reactions searched should belong to, e.g. Glycolysis; or he or she can specify one or more reaction participants (reactants or enzymes), organisms, tissues, or cell types in which the reaction is reported to occur. Additional search terms include cellular locations, environmental conditions (pH and temperature), or publications in which kinetic data are reported. The next version of the interface will also enable the user to search for networks or paths of reactions between two compounds or enzymes.



SABIO

REACTION KINETICS DATABASE

CONTACT | HELP | IMPRINT

Reaction Search

Search Reaction

SBML Model Setup

EML Research

© EML Research gGmbH

Return only reactions having kinetic data matching the search criteria ☑

Specify Search Criteria:

Submit Search | Reset Form

with Reactants(s) [⊞][⊟]

in Pathway(s) [⊞][⊟]

Glycolysis classical

Select Pathway

Delete Pathway

Join entries with ● AND or ○ OR

having Enzyme(s) [⊞][⊟]

in Organism(s) [⊞][⊟]

Homo sapiens

Select Organism

Delete Organism

Join entries with ● AND or ○ OR

in Tissue(s)/Cell Type(s) [⊞][⊟]

in (Intra/Extra)Cellular Location(s) [⊞][⊟]

Having Kinetic Data Determined for Specific Experimental Conditions [⊞][⊟]

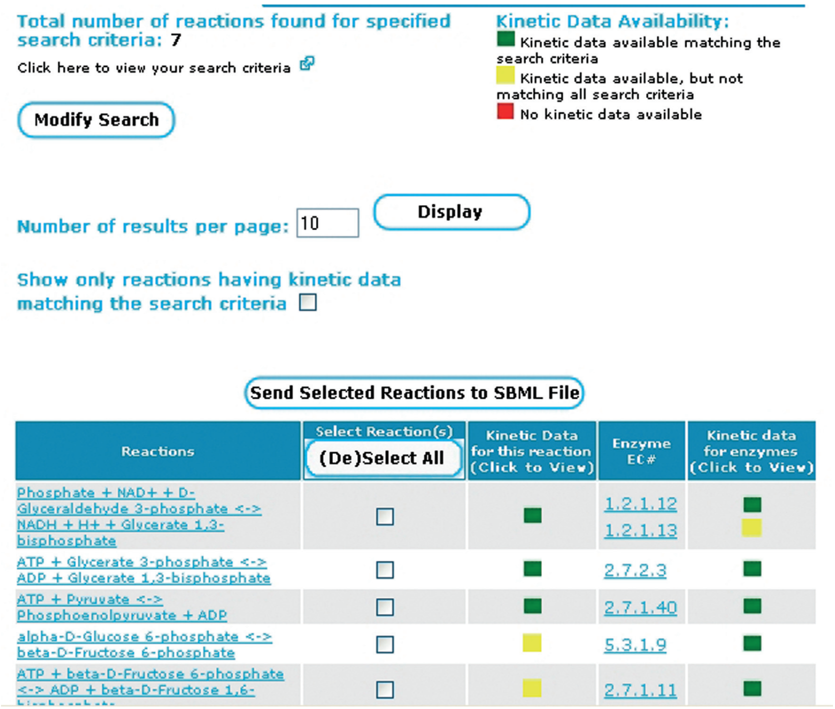
in Publication [⊞][⊟]

Submit Search | Reset Form

Figure 2: Search facilities in SABIO-RK. Currently the system only offers the possibility of searching for reactions and their kinetics, but we plan to expand the search facilities to search for enzymes, specific parameters, and for compounds.

The system retrieves all entries satisfying the given criteria and indicates whether there is kinetic information available. A three colour-code is used to indicate this. Green means that for the associated reaction there are kinetic data available matching all search criteria. For a search like "find all reactions within the Glycolysis pathway for *Homo sapiens* which take place in liver", this would mean that there is kinetic data reported on the respective reaction in human liver. Yellow means there are kinetic data available, but not matching all search criteria. For example, the kinetic data were not determined for *Homo sapiens* but for *Rattus sp,* or not in liver but in heart. Red indicates that there are no kinetic data stored for the reaction reported.

Apart from showing the availability of kinetic data for the specified reactions, the system will also indicate whether there is kinetic data available for the enzymes catalysing each of these reactions (see Fig. 3). We took this approach to offer complementary or alternative information about kinetic data for related reactions catalysed by the same enzyme. The availability of kinetic data for the enzyme is shown using the same three- colour code as used for the reactions. By clicking on a reaction, further information about it is displayed: Reactants, pathways in which it participates and enzymes catalysing this reaction that are reported with kinetic data in the database for a specific organism. Additional information about the enzyme (name, synonyms, classification and reactions it catalyses) can similarly be obtained by clicking on the EC number.



Total number of reactions found for specified search criteria: 7

Click here to view your search criteria

Kinetic Data Availability:

🟩 Kinetic data available matching the search criteria

🟨 Kinetic data available, but not matching all search criteria

🟥 No kinetic data available

Number of results per page: 10

Show only reactions having kinetic data matching the search criteria □

Reactions	Select Reaction(s) (De)Select All	Kinetic Data for this reaction (Click to View)	Enzyme EC#	Kinetic data for enzymes (Click to View)
Phosphate + NAD+ + D-Glyceraldehyde 3-phosphate <-> NADH + H+ + Glycerate 1,3-bisphosphate	□	🟩	1.2.1.12 1.2.1.13	🟩 🟨
ATP + Glycerate 3-phosphate <-> ADP + Glycerate 1,3-bisphosphate	□	🟩	2.7.2.3	🟩
ATP + Pyruvate <-> Phosphoenolpyruvate + ADP	□	🟩	2.7.1.40	🟩
alpha-D-Glucose 6-phosphate <-> beta-D-Fructose 6-phosphate	□	🟨	5.3.1.9	🟩
ATP + beta-D-Fructose 6-phosphate <-> ADP + beta-D-Fructose 1,6-bisphosphate	□	🟨	2.7.1.11	🟩

Figure 3: Results screen, showing the entries found for the given criteria (Glycolysis in *Homo sapiens*) and for each of these the availability of kinetic data.

From the result screen listing the specified reactions, the user can view the kinetic data belonging to each reaction, or all kinetic data available for the enzymes catalysing this reaction. In a new window the entries containing kinetic data for one reaction or one enzyme are listed. The user is presented with an overview showing for each entry the data on organism, tissue, enzyme classification and the variant of the enzyme. The expanded version of an entry shows all the kinetic data and additional information extracted from a publication, like environmental conditions. The information source of each database entry is indicated and linked to the PubMed database [10] in order to allow the user to refer to the original publication (Fig. 4).

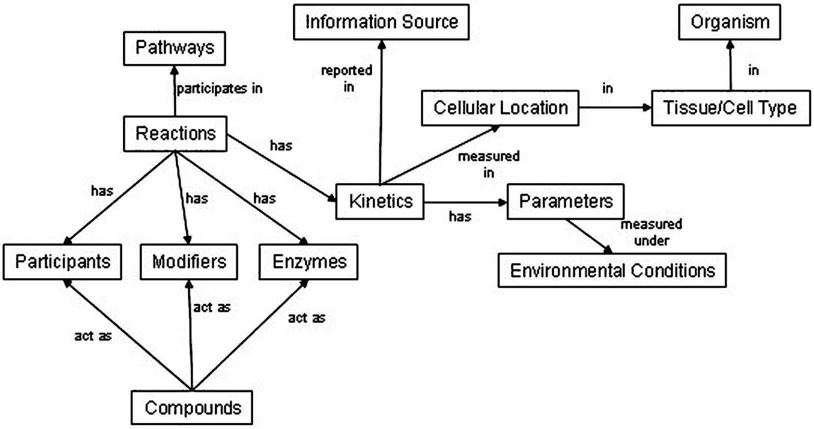


Figure 4: Kinetic data entry.

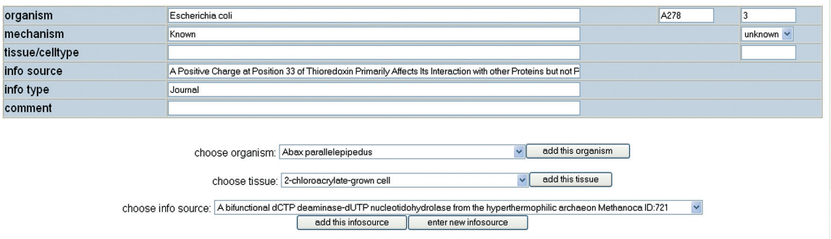
The results on reactions and their corresponding kinetic laws and parameters can be stored and exported in a SBML (Systems Biology Mark-Up Language) formatted file. This format has been established as a standard exchange format between different tools including modelling and simulation software. The export is facilitated by using the libSBML API [4]. Not only single reactions, but also reaction clusters can be exported. The SBML file lists all the compounds (named species in SBML) belonging to the reactions as participants or modifiers. If a compound is present in more than one reaction, it will only be defined once in the file and will be referred to in the corresponding reactions. Thus, the reactions are coupled by the overlapping compounds.

Due to the limitations of the SBML file format, the data exported requires some simplifications. For example no information about the experimental conditions, under which the parameters were determined, can be exported yet, although we plan to incorporate this information as annotations in the SBML file. Because parameter values can only be single values but no ranges, we include as parameter value the mean of the parameter range (if given). Also, the standard deviations of the parameters stored in the database, cannot be exported. Another restriction of the SBML format is the limitation to one kinetic law for each reaction. Thus, multiple kinetic laws (e. g. pseudo-first order kinetics) for one and the same reaction cannot be exported in one file.

As of June 2006, data of over 520 publications were inserted into the SABIO-RK database, corresponding to over 5100 database entries for 1100 different biochemical reactions and 325 distinct EC classes in 194 organisms. The stored parameters mainly describe steady-state kinetics for metabolic reactions. Around 40% of all entries have a rate equation. The database entries describe around 4600 enzyme activities (like rate constants, k_{cat} or V_{max}), 4600 K_{m} and 1000 K_i (inhibitor constant) values.

Data Input

The information about the kinetics of biochemical reactions is mainly extracted from text in a manual process carried out by student helpers. They use a web-based interface (Fig. 5) to enter the data into a temporary database. The main objective of this user interface is to supply a uniform format that the students and curators can employ to include the data found in the publications. The interface supports the students by pre-processing the data introduced and by offering the possibility to choose terms from predefined thesauri (here of course also allowing the introduction of new terms); this helps to avoid redundancies just because of aberrant notations or typing errors. The system will verify amongst other things if the parameters defined in a kinetic law are all defined as parameters, even if they do not have values associated with them.



organism	Escherichia coli	A278	3
mechanism	Known		unknown
tissue/celltype			
info source	A Positive Charge at Position 33 of Thioredoxin Primarily Affects Its Interaction with other Proteins but not P		
info type	Journal		
comment			

choose organism: Abax parallelepipedus

choose tissue: 2-chloroacrylate-grown cell

choose info source: A bifunctional dCTP deaminase-dUTP nucleotidohydrolase from the hyperthermophilic archaeon Methanoca ID:721

Figure 5: Input of the reaction data (substrates, products and modifiers) together with information about the pathway (optional) and about the enzyme.

Ideally students extract the following information for each reaction reported within a publication:

- Reaction defined by substrates and products
- Modifiers of the reaction (activators, inhibitors, catalysts, cofactors)
- Cellular location of compounds
- Enzyme classification number
- Swiss-Prot accession number(s) (of the enzyme)
- Variant of the enzyme (wild type or a certain isoenzyme or mutant)
- Kinetic law type (e. g. Michaelis–Menten, Ping–Pong Bi–Bi)
- Kinetic law formula

- Kinetic parameters (e. g. K_m, k_{cat}, V_{max})
- Concentrations used for reactants, enzymes and modifiers
- Experimental conditions (e. g. temperature, pH, buffer composition)
- Biological source (e. g. cell type, tissue, organism, strain)
- Information source (reference)

For most of this information, comment lines are available to add information, for example about synthetic, labelled derivatives of physiological compounds or host organism for a recombinant enzyme.

In order to provide a better understanding of the interface, let us now go over the support offered by the system in the introduction of the fields mentioned above.

Input of reactions' data

To begin with, the student may enter some of the names of the reactants (substrates and products; we will also refer to these as species), followed by a database search which in turn displays all reactions stored in the database in which the reactants are involved. By choosing the appropriate reaction, all relevant information is automatically extracted from the database and displayed in the corresponding fields such as: species name, species stoichiometry and species role (substrate, product). However, it might be that the reaction is not found; in this case the user may enter all information manually. After the introduction of the substrates and products (in which ever way), the species can be associated to a location; this is also supported by offering a list of locations. Determining whether a reaction or a compound is already included in SABIO, is not a trivial issue, given that the search by name may not suffice to determine synonymic expressions. If a new reaction is given curators have to verify (as much as possible) whether this reaction is really new or if it is already in the database with a different notation. To support the curators, we are working on the development of linguistic methods to obtain compound structures from names and compare compounds at the level of their chemical structure [11].

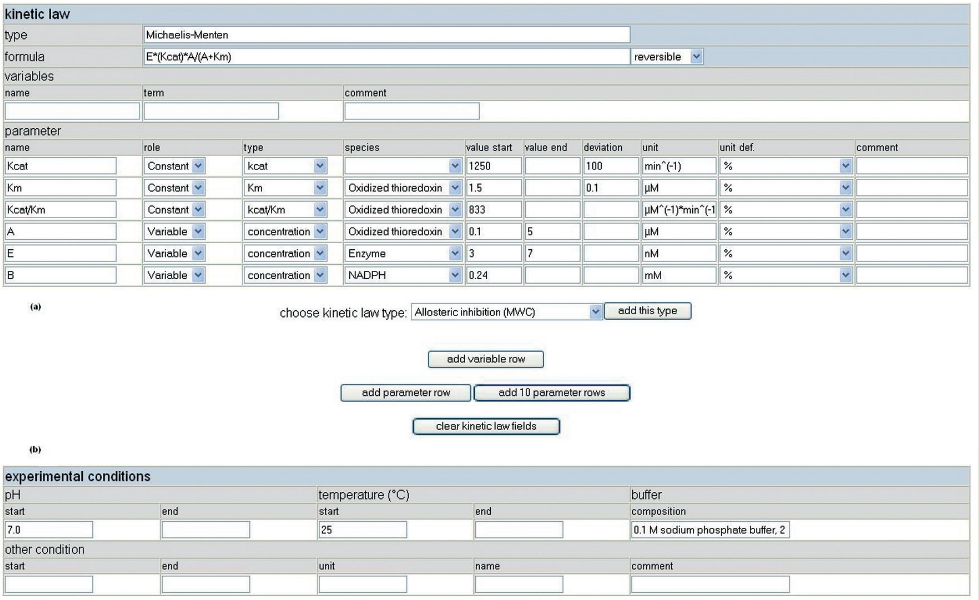
Apart from the reactants the user should specify information about the enzyme (if applicable) like enzyme classification of the reaction plus Swiss-Prot identifier(s) of the protein or protein complex. The information of the pathway in which the reaction participates is optional; for reactions in the database this information is already present.

Addition of kinetic laws **(Figure 6a)**

By the addition of the kinetic law information, the user is supported by providing a list of possible kinetic law types. Originally the system automatically offered a default formula for each type, which could be used by users as a basis; however this feature has been taken out by petition of the users, who manifested the preference in directly introducing the mathematical formula as specified in the paper. The user can define parameters and variables for the kinetic law. In order to avoid the proliferation of unit definitions, the user is supplied with a list of units. New units can be added, but this is not encouraged unless completely necessary (no equivalent found in the list). Additionally, the user is supported by some verification procedures: all parameters and species referred to in the kinetic law formula must be defined in the parameter and species lists, respectively; the brackets in the kinetic law formula must be mathematically correct; naming of parameters must be consistent with SBML rules (e. g. no special characters allowed); parameter types can only be chosen from a given list of predefined terms (e. g. V_{max}, K_m, K_i); in case of a parameter–species relationship (e. g. for K_m or concentration value)only predefined species from the reaction list can be entered. In addition to this, a browser plug-in has been implemented to allow the visualization of the kinetic law formula as a mathematical formula and not just as text, helping to verify its correctness.

Experimental conditions (Figure 6b)

In this section the user should introduce the experimental conditions under which the kinetics were determined. Currently we consider the pH, temperature and the specification of the buffer, but the system allows the introduction of other conditions.



kinetic law

type	Michaelis-Menten	
formula	E*(Kcat)*A/(A+Km)	reversible

variables

name	term	comment

parameter

name	role	type	species	value start	value end	deviation	unit	unit def.	comment
Kcat	Constant	kcat		1250		100	min^(-1)	%	
Km	Constant	Km	Oxidized thioredoxin	1.5		0.1	µM	%	
Kcat/Km	Constant	kcat/Km	Oxidized thioredoxin	833			µM^(-1)*min^(-1	%	
A	Variable	concentration	Oxidized thioredoxin	0.1	5		µM	%	
E	Variable	concentration	Enzyme	3	7		nM	%	
B	Variable	concentration	NADPH	0.24			mM	%	

choose kinetic law type: Allosteric inhibition (MWC)

experimental conditions

pH		temperature (°C)		buffer
start	end	start	end	composition
7.0		25		0.1 M sodium phosphate buffer, 2

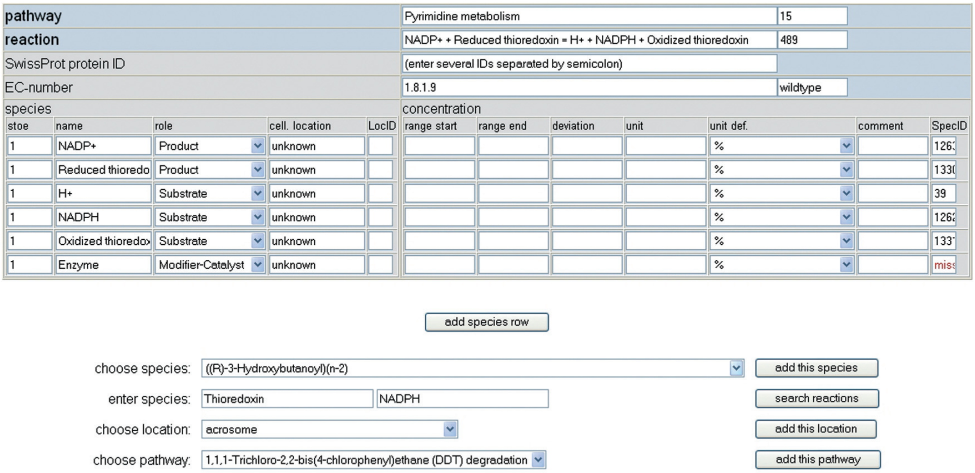
other condition

start	end	unit	name	comment

Figure 6: Input of the kinetic data (a) along with the information about the environmental conditions under which these were determined (b).

General Information (Figure 7)

In this section of the entry form the user is asked to give information about the organism and tissue (if known) for which the kinetics were determined. Here again the user is supported by lists of names. Although these fields should belong to the experimental description, they have been put here due to the fact that typically a publication will report on the kinetics for multiple reactions under multiple experimental conditions, however the organism and tissue are commonly constant within a publication. All data in the general section can be kept for its use for several kinetic data entries. The information source should also be given in this section, this is a compulsory field (there cannot be any entry without information source). The user can select from the list of publications in the database (using a search function) or introduce a new source. Also included in here is the possibility to add comments (general to the entry) and currently we have a field to indicate whether or not the paper provides detail information about the reactions mechanism; this information will be used when the system supports a detailed description of the reactions' mechanisms (see future plans).



pathway	Pyrimidine metabolism	15
reaction	NADP+ + Reduced thioredoxin = H+ + NADPH + Oxidized thioredoxin	489
SwissProt protein ID	(enter several IDs separated by semicolon)	
EC-number	1.8.1.9	wildtype

species					concentration						
stoe	name	role	cell. location	LocID	range start	range end	deviation	unit	unit def.	comment	SpecID
1	NADP+	Product	unknown						%		126:
1	Reduced thioredo	Product	unknown						%		133(
1	H+	Substrate	unknown						%		39
1	NADPH	Substrate	unknown						%		126:
1	Oxidized thioredo›	Substrate	unknown						%		133'
1	Enzyme	Modifier-Catalyst	unknown						%		miss

choose species: ((R)-3-Hydroxybutanoyl)(n-2)

enter species: Thioredoxin NADPH

choose location: acrosome

choose pathway: 1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation

Figure 7: Input of complementary information to the entry, which very often is shared amongst many entries within the same publication.

Before the data is finally transferred to SABIO-RK, it is approved, complemented, and verified by a team of biological experts so as to detect possible errors and inconsistencies. The curators are faced with problems like synonymic or aberrant notations of compounds and enzymes, multiplicity of parameter units and missing information about assay procedures and experimental conditions. Frequently, the methods used are described fragmentarily or by a simple reference to another publication, which in turn refers to a third publication. Hence, it is sometimes almost impossible to get the complete description. Moreover, the description of a buffer can be very complex, containing for example information about coupled enzyme reactions or synthetic derivatives of physiological compounds. Chemical compounds and enzymes often have various alternative names, organisms can be described by their common or systematic name, and units of kinetic parameters and concentrations can be written in different ways or can be based on different unit systems. Furthermore, we are often faced by the problem of missing or partial information in the literature. For example, a reaction definition can be incomplete, which means that only substrates of reactions are named without a definition of the reaction products. If the chemical mechanism of the enzymatic reaction is known, the reaction equation can be completed, but in most cases this work is very time-consuming, and the result may also be imprecise.

During the curation process, the data is unified and structured consistently in order to facilitate the comparison of the kinetic data extracted from different sources, since it was usually obtained under different experimental conditions or from different organisms, tissues etc. Furthermore, structured data enable the user to conclude general rules concerning the dependence of a biochemical reaction or an enzyme on environmental changes like for example increase of temperature or pH variations.

The interface is also used by the curators to check and complete the entries, and supports them in the administrational work (assignment of papers, statistics etc.). The publications to be revised have been obtained from PubMed [9], by using several queries leading to papers, which very likely contain information about biochemical reaction kinetics.

The information supported by this input interface covers most of the fields present in the STRENDA commissions' recommendations for the reports about reaction kinetics. Currently the input interface is being used only internally by the development team on SABIO-RK, however we hope that in the future experimental partners can directly introduce their data into the database and make it thus available through the SABIO-RK database interface.

Future Directions

The SABIO-RK project started at the beginning of 2005. Currently the database contains mainly data about metabolic reactions. However, since cellular signal transduction is a fast-growing emerging field, one of our main objectives is to incorporate more kinetic information about signalling reactions. This includes the representation of molecules in different activation states, for instance modifications of signalling molecules like phosphorylation or acetylation of proteins. Another very important objective is the incorporation of detailed information about reactions' mechanisms. This will allow the user to obtain information about the kinetic properties of sub-reactions or binding mechanisms of enzymes and substrates. As mentioned in the data input section, we are keeping track of the publications having this information to facilitate the process of returning to the adequate literature. An extension of the data model to store reactions' mechanisms and the corresponding kinetic data has already been developed and will soon be implemented together with adaptations of the user interface.

In order to allow the users to refer to additional information about reactions, pathways, chemical compounds and enzymes, we are working on the cross-linking and annotation of the database content to other database resources. In addition, we will apply and annotate controlled vocabularies and ontologies such as those specified in the Open Biomedical Ontologies (OBO) [12] to enhance the standardization and comparability of the data stored in SABIO-RK. With these goals, we will adopt the proposed set of rules for the annotation of biochemical models described in MIRIAM (Minimum information requested in the annotation of biochemical models) [5]. The annotations will not only be used for cross-linking our database to other resources, but also can be exported in the SBML files.

The information describing environmental conditions under which the parameters were determined, as well as the literature source from which the data was extracted, cannot be completely exported in a structured and defined format in SBML. For this reason, we plan to define new XML based export schemas, in order to facilitate the exchange of detailed kinetic information together with their constraints.

On the side of the user interface of SABIO-RK we are working on the extension of search facilities, less reaction oriented, to permit searches for parameters and kinetic laws, e.g. search for all reactions that follow a certain kinetic law type or for all enzymes of the pathway glycolysis for which K_m values are known. Also planned in the near future is to enable the user to search for networks or paths of reactions between two compounds or enzymes. Visual display of the reactions found as well as of the kinetic parameter values is also scheduled.

One of our biggest aims is to convince scientists to use the input interface to enter data directly into the database. As a result, all the needed information can be given by the experimenters and no information is lost. In doing so, users would be able to directly compare their own experimental results in SABIO-RK with similar kinetic data extracted from literature or entered by other users.

Summary

SABIO-RK is a database storing highly interrelated information about biochemical reactions and their kinetics, within the context of cellular locations, tissues and organisms. The database has a web-based user interface that enables the user to search for biochemical reactions and their kinetics, based on the characteristics of the reactions and on the environmental conditions under which its kinetics were obtained. Although the main motivation of SABIO-RK was to act as a resource for modellers of biochemical networks to assemble information about reactions and their kinetics, the database is also aimed at experimenters wanting to obtain information about reactions kinetics and compare their own results with similar published data. The kinetics data is mainly extracted from literature sources by students and then revised and supplemented by a group of curators. The students employ a web-based interface to introduce the data in a standardized format. We hope that in the future both, experimentalists and modellers will be able to use this interface to directly introduce kinetic study results of their respective experiments or simulations into the database.

Acknowledgements

First of all we would like to thank the Klaus Tschira Foundation as well as the German Research Council (BMBF) for their funding. We would also like to thank the members of the Bioinformatics and Computational Biochemistry and the Molecular and Cellular Modelling Groups of EML Research for their helpful discussions and comments. Last but not least, we thank all the student helpers, who have contributed to the population of the database.

References

[1] Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., Schomburg, D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* **32:**D 431–433.

[2] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider M. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. **31:**365–370.

[3] Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J.L., Hucka, M. (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res*. **34:**D 689–691.

[4] Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novere, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19:**524–531

[5] Le Novere, N., Finney, A., Hucka, M., Bhalla, U.S., Campagne, F., Collado-Vides, J., Crampin, E.J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J.L., Spence, H.D., Wanner, B.L. (2005): Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnol.* **23:**1509–1515.

[6] STRENDA: http://www.strenda.org

[7] Rojas, I., Bernardi, L., Ratsch, E., Kania, R., Wittig, U., Saric, J. (2002) A database system for the analysis of biochemical pathways. *In Silico Biol*. **2:**0007.

[8] IUBMB: http://www.chem.qmul.ac.uk/iubmb/enzyme/

[9] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M. (2006): From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34:**D 354–357

[10] PubMed: http://www.pubmed.gov

[11] Wittig, U., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann, A., Anstein, S., Saric, J., Rojas, I. (2006) SABIO-RK: integration and curation of reaction kinetics data. *Lecture Notes in Bioinformatics* **4075:**94–103.

[12] Open Biomedical Ontologies (OBO): http://obo.sourceforge.net/

Investigation of Proteases – Suggestions

Hartmut Schlüter

Core Facility Protein Analytik, Charité,
Tucholskystr. 2, 10117 Berlin, Germany

E-Mail: hartmut.schlueter@charite.de

Received: 6th November 2006 / Published: 31st August 2007

Proteases are enzymes catalysing the hydrolysis of peptides or proteins. They are key players in a wide range of biological processes such as the release of peptide hormones, nutrient acquisition, cell growth, differentiation, antigen processing and protein turnover, in all living organisms. Furthermore it is becoming more and more obvious that the abnormal functioning of some proteases may lie behind several types of diseases, including inflammation, cancer and Alzheimer's disease. Therefore proteases are attracting an increasing interest.

The MEROPS database, which is specialized in proteases, lists 555 known and putative genes encoding proteases in *Homo sapiens* (31st of August 2006). The number of proteins acting as proteases in the human organism may even be much higher, since proteins can develop proteolytic activities although they are not assigned as proteases. The protein disulfide isomerase A3 (PDIA3, primary accession number: P30101), which main function is protein folding, is an example for the latter case [1,2], since Kito, Urade and coworkers published convincing data about a protease activity of PDIA3 [1].

From many of these protease-encoding genes the endogenous substrates and as a result the physiological roles are as yet unknown. One strategy for deciphering the physiological roles of proteases is to start with known reaction products of the proteolytic action of unknown proteases (Fig. 1). For example, the peptide urotensin-II, a potent vasoconstrictor, is cleaved from its inactive urotensin-precursor by the proteolytic action of an unknown protease. The knowledge of the sequence of both, the peptide urotensin and its precursor, allows a reaction specific probe (substrate) to be developed, which can be used in an assay like the MES (mass spectrometry assisted enzyme screening system) [1] for detecting urotensin-II-generating activity. After having developed the enzyme assay it can be used for screening for the presence of the target enzyme in protein fractions and guiding the purification of the target enzyme to near homogeneity (Fig. 1). For the identification of the target enzyme the purified active fraction can be subjected to enzymatic cleavage and mass spectrometric analysis of the enzymatic peptide fragments followed by database research. After protein identification the results must be validated by either expressing recombinantly the identified protein candidate or by simply purchasing it, if possible and by demonstrating that the enzymatic activity and properties are identical to that of the purified enzyme and its properties.

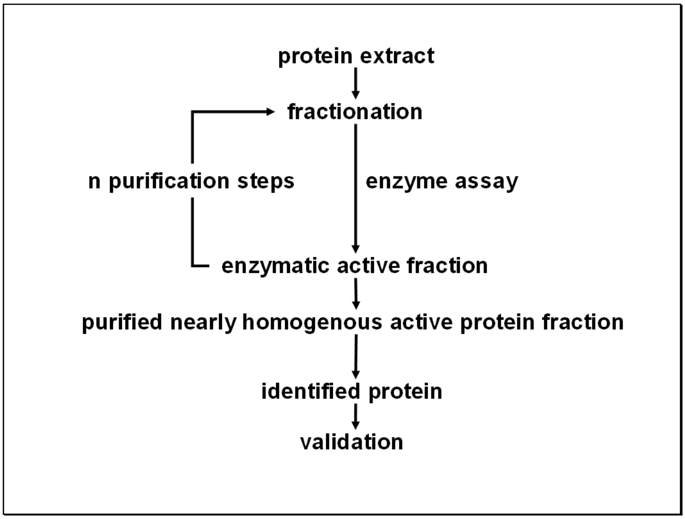


Figure 1 Scheme showing the strategy for detection, purification, identification and verification of proteases from protein extracts.

The MES system is one of the core instruments within the protease-deciphering strategy. Figure 2 demonstrates a typical read out of the MES system. In this case 2 different protein fractions were monitored for angiotensin-II-generating activity by incubating the proteins with the reaction specific probe angiotensin-I. Since the signal intensity of the reaction product angiotensin-II in the incubate of fraction A increases faster with increasing incubation time than in fraction B, the angiotensin-II-generating activity of fraction A is higher.

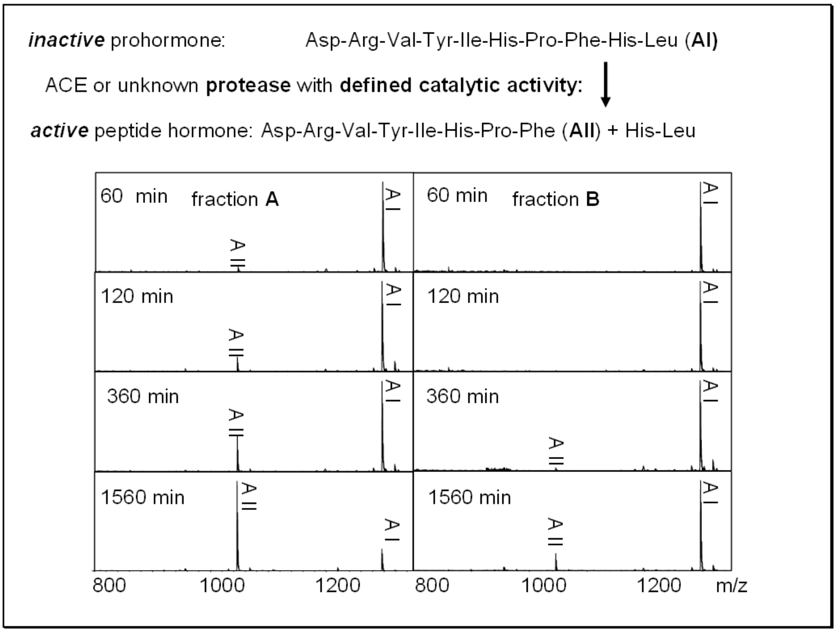


Figure 2. MES results of a MES assay monitoring angiotensin-II-generating activity of 2 chromatographic fractions A and B. MALDI-MS spectra of the reaction products of the incubation of immobilized proteins of fraction A and fraction B derived from porcine renal tissue. AI: Angiotensin I; AII: Angiotensin II; A(1 – 7): Angiotensin (1 – 7).

Mass spectrometry based enzyme assays are advantageous compared to UV- or fluorescence based enzyme assays because they give information about the identity of the reaction products and about the fate of the substrate. The control of the identity of the reaction products reduces the risk of false positive results. Being able to monitor the fate of the substrate gives the opportunity to notice the presence of additional proteolytic activities accompanying the target protease. An example for this latter case is given in Fig. 3A. The MES mass spectrum was obtained after incubating a crude protein fraction obtained from porcine renal tissue with angiotensin-I. In the spectrum in Fig. 3A, beside the signal of angiotensin-II, additional signals point to the presence of several other peptidases. The peptide des-Asp-A-I may be generated by the enzyme ACE-II, which is known to be present in renal tissue and A(1 – 7) by the renal peptidase neprilysin. With increasing purity of the angiotensin-II-generating activity the number of the additional peptides decreases. Nearly homogenous fractions yield the signal of the reaction product of the target protease only (Fig. 3B).

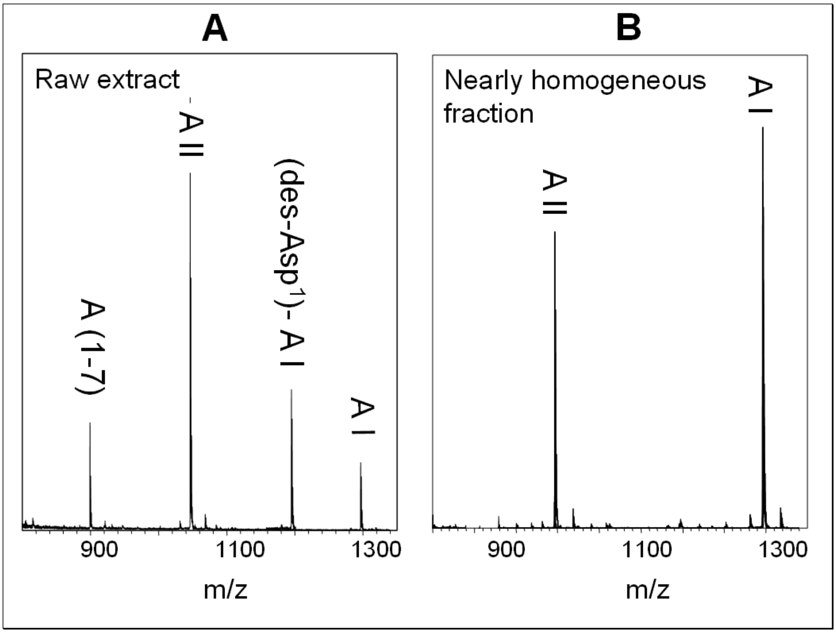


Figure 3. Typical MES results of a MES assay monitoring angiotensin-II-generating activity: MALDI-MS spectra of the reaction products of the incubation of immobilized proteins of a raw extract of porcine renal tissue (A) and of proteins from a fraction purified to near homogeneity. AI: Angiotensin I; AII: Angiotensin II; A(1 – 7): Angiotensin (1 – 7); (des-Asp^1)-A-I: des-asparaginc acid angiotensin I.

The experience with the mass-spectrometry based assay system MES results in the 1st suggestion:
Control experiments should include a mass spectrometric analysis of the enzymatic reaction products thus minimizing false positive results, verifying the chemistry of the catalytic conversion and being able to detect other accompanying enzymatic activities, which may interfere with the target enzyme.

After purifying the active fraction to near homogeneity the protein will be identified (Fig. 1). Usually within the purified fractions not only one but several proteins are identified. Therefore the question arises as to which of the identified proteases may have proteolytical activities. Comparison of the own experimental data describing the properties of the protease with those described in the literature helps to verify the identification data. The verification procedure may be easy, if the identified protein is known as a protease, the proteolytic activity is its main function and the enzymatic properties are well described. However proteins may be identified, which have several different functions. An example for the latter case is the protein disulfide isomerase A3 (PDIA3_human, P30101), which major catalytic property comprises the disulfide isomerase activity. In such cases it has to be proven, if the proteolytic activity is physiologically relevant. Therefore a comprehensive database analysis and analysis of the original papers is necessary. If this work will give more confidence about the proteolytic activity of the candidate, the protein should be recombinantly expressed to prove its proteolytic activity experimentally.

Performing the database analysis about the candidate usually is accompanied with some trouble, arising from the many synonyms often used for a protein encoded by one single gene and the still missing standardization of the nomenclature of proteins. Here an example will be given. The database Swiss-Prot summarizes the following molecular functions of PDIA3: Cysteine-type endopeptidase activity, phospholipase C activity, protein disulfide isomerase activity, protein retention in ER, protein import into nucleus and signal transduction. In the next step the original papers have to be searched for. Using the synonyms shown in Swiss-Prot for PubMed database searches yielded the results given in Table 1. Because of the confusion concerning nomenclature, some authors used several synonyms within the title of their papers: “Association of the chaperone **glucose-regulated protein 58** (**GRP58/ER-60/ERp57**) with Stat3 in cytosol and plasma membrane complexes” [1, 2].

Table 1. The numbers indicate the hits in the database search of PubMed performed with the synonyms of PDIA3 without and with additional keywords.

Keywords **Synonyms**	Σ	**Isomerase** **+ human**	**Isomerase** **+ human + protease**
ERp57	119	73	**5**
p58	**480**	**8**	1
58 kDa microsomal protein	0		
58 kDa glucose regulated protein	0		
ER60	23	10	**4**
ER-60	32	10	**5**
ERp60	**833**	73	**8**
PDIA3	98	97	**10**

Table 1 lists only a few of many synonyms known for proteins encoded by the gene PDIA3. A PubMed data base search was performed with each of the synonyms, with the combination of the synonyms with the key words “isomerase” and “human” and with the keywords “isomerase”, “human” and “protease”. The synonyms “p58”, “58 kDa microsomal protein”, “58 kDa glucose regulated protein” yielded the worst results, especially “p58”, but also “ERp60” yielded a huge number of false positive results. More helpful are the synonyms “ER-60”, “ER60” and “PDIA3”. However, only in those papers, where the gene sequence of the amino acid sequence or the complete amino acid sequence is given, one can be sure, that the protein described is identical with that described under PDIA3 in Swiss-Prot. Therefore in many cases there remains some doubt, whether the properties described in the publication really belong to PDIA3.

Suggestion 2:

A standardized nomenclature for enzymes (and all other proteins) is needed, e. g. the accession number, which gives an unambiguous hint towards its origin and which should be used by all databases and all journals!

As soon as the database research has yielded positive results towards the proteolytic function of the candidate protein a validation by an appropriate experiment is necessary. In some cases a recombinant expression of the candidate protein may be circumvented because it can be purchased. Independently from the source of the candidate protein it is strongly recommended that both the identity and the purity of the candidate protein preparation are checked. The following example demonstrates the importance of this recommendation: A protease preparation was purchased, here named protease 1. A size exclusion chromatography (SEC) of the protease 1 containing fraction was performed (Fig. 4). The UV-absorption profile of the SEC chromatogram (Fig 4A) monitored at 280 nm already shows that the protease 1 containing fraction is not pure. The protease 1 activity of the fractions of the SEC was measured with an appropriate substrate. The protease 1 activity co-eluted with the fraction with the highest UV absorption. An LC–MS analysis of the tryptic peptides of the active fraction confirmed the identity of protease 1. Furthermore the fractions were tested for the target enzyme activity with the MES assay described above. Surprisingly the target enzyme activity eluted in front of the protease 1 fraction. The purchased protease 1 fraction was chromatographed with an affinity chromatography and an aliquot of the eluate was applied to the size exclusion chromatography again (Fig. 4B). In the resulting fractions no protease 1 activity was detected any more beside the target protease activity. An LC–MS analysis of the fraction with the target protease activity confirmed that the target protease is not identical to protease 1.

The example of the impurified protease 1 demonstrates the need for controlling the identity of the protease responsible for the activity as well as the control of the purity. Without controlling the purity the false protease would have been assigned with the defined proteolytic reaction.

The problem of impurities is also given with recombinantly expressed proteins, even after passing affinity chromatography. Therefore suggestion 3 is recommended.

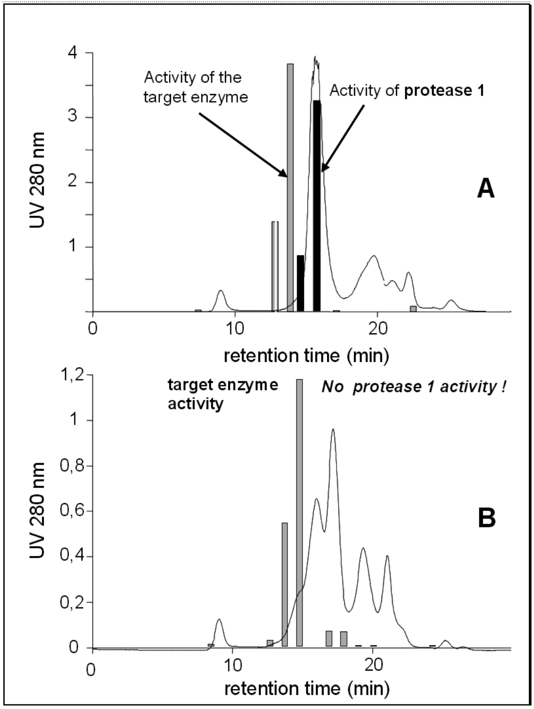


Figure 4. Chromatograms of size exclusion chromatographies of a purchased protein fraction before (A) and after purification (B) with an affinity chromatography. Black bars: Activity of protease 1. Grey bars: Activity of the target protease.

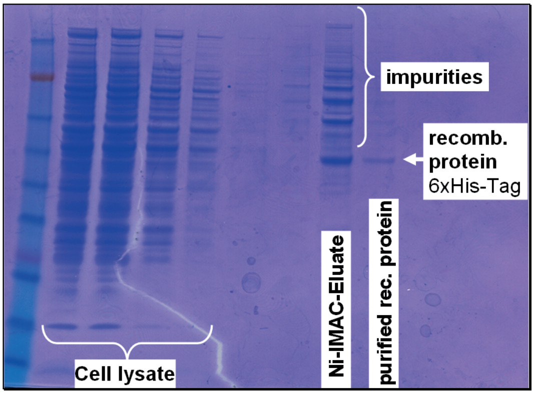


Figure 5. SDS-PAGE analysis of the lysate of the host cells expressing a recombinant protein with a His-tag and the eluate of a immobilized-Ni affinity chromatography (Ni-IMAC)

Suggestion 3:

Publishing data about the properties of an enzyme should also include 1, proof of the purity of the enzyme fraction and 2, proof of the identity of the enzyme. The proof of the purity should not be performed by *SDS–PAGE* but 2DE including identification of all proteins visible in the 2DE-gel *or* tryptic digest followed by LC–MS/MS analysis.

References

[1] Spiess, C., Ehrmann, M. (1999) A temperature-dependent switch from chaperone to protease in a widely conserved heat shock protein. *Cell* **97:**339–347.

[2] Krojer, T., Garrido-Franco, M., Huber, R., Ehrmann, M., Clausen, T. (2002) Crystal structure of DegP (HtrA) reveals a new protease chaperone machine. *Nature* **416:**455–459**.**

[3] Urade, R**.,** Oda**,** T**.,** Ito**,** H**.,** Moriyama, T., Utsumi, S., Kito, M., (1997) Functions of characteristic Cys-Gly-His-Cys (CGHC) and Gln-Glu-Asp-Leu (QEDL) motifs of microsomal ER-60 protease. *J. Biochem.* **122:**834–842.

[4] Schlüter, H., Jankowski, J., Rykl, J., Thiemann, J., Belgardt, S., Zidek, W., Wittmann, B., Pohl, T. (2003) Detection of protease activities with the mass-spectrometry-assisted enzyme-screening (MES) system. *Analyt. Bioanalyt. Chem.* **377:**1102–1107.

[5] Guo, G.G., Patel, K., Kumar, V., Shah, M., Fried, V.A., Etlinger, J.D., Sehgal, P.B. (2002) Association of the chaperone glucose-regulated protein 58 (GRP58/ER-60/ERp57) withStat3 in cytosol and plasma membrane complexes. *J. Interferon Cytokine Res.* **22:**555–563**.**

[6] Kita, K., Okumura, N., Takao, T., Watanabe, M., Matsubara, T., Nishimura, O., Nagai, K.(2006) Evidence for phosphorylation of rat liver glucose-regulated protein 58,GRP58/ERp57/ER-60, induced by fasting and leptin. *FEBS Lett.* **580:**199–205.

New Developments at the BRENDA Enzyme Information System

Jens Barthelmes, Christian Ebeling, Antje Chang, Ida Schomburg and Dietmar Schomburg

Technical University Braunschweig, Bioinformatics and Systems Biology,
Langer Kamp 19b, 38106 Braunschweig, Germany

E-Mail: d.schomburg@tu-bs.de

Received: 31st October 2006 / Published: 31st August 2007

Abstract

The BRENDA enzyme information system (http://www.brenda.uni-koeln.de) is the largest publicly available enzyme information system worldwide. The major part of its content is manually extracted from primary literature. It is not restricted to specific groups of enzymes, but includes information on all identified enzymes irrespective of the source of the enzyme. The range of data encompasses functional, structural, sequence, localization, disease-related, isolation, stability information on enzyme and ligand-related data. Each single entry is linked to the enzyme source and to a literature reference. Recently the data repository was complemented by text mining data which is stored in AMENDA and FRENDA. A genome browser, membrane protein prediction and full text search capacities were added. The newly implemented web service provides instant access to the data for programmers via a SOAP interface. The BRENDA data can be downloaded in the form of a text file from the beginning of 2007.

Introduction

The BRENDA (BRaunschweig ENzyme DAtabase) enzyme information system [1, 2] is a manually annotated repository for enzyme data. Originally published as a series of books [3] in 1987, it was integrated into a publicly available database in 1998 and has been curated and continuously improved and updated at the University of Cologne since then. Its contents are not restricted to specific groups of enzymes, but include information on all enzymes that have been classified in the EC scheme of the IUBMB (International Union of Biochemistry and Molecular Biology) irrespective of the enzyme's source. The range of data includes the catalysed reaction, detailed description of the substrate, cofactor and inhibitor specificity, kinetic data, structure properties, information on purification and crystallization, properties of mutant enzymes, participation in diseases, and amino acid sequences. Each single entry is linked to the enzyme source (organism and, if applicable, the tissue, and/or the protein sequence) and to the literature reference. Data queries can be performed in a number of different ways, including an EC-tree browser, a taxonomy-tree browser, an ontology browser, and a combination query of up to 20 parameters. However the huge amount of literature on enzymes does not allow the manual annotation of the complete literature for all enzymes. The capacity for manual annotation has been restricted to ~8,000 references per year. To be able to include more literature, text-mining programs have been developed. Recently, two additional databases (AMENDA and FRENDA) which contain the results of these procedures, have been added to the BRENDA host. They complement the existing database with respect to organisms, tissues and references.

Contents of BRENDA

At present, BRENDA contains ~1.9 million manually annotated data for more than 4,000 EC-numbers, on average 500 single entries per EC-number. These data are stored in ~120 tables in a relational database system enabling extensive search modes, i.e. quick search, full text search, advanced search, substructure search, sequence search, TaxTree search, ECTree browser, searches in the Genome browser, and searches in more than 20 different ontologies.

Functional parameters

In total BRENDA holds data for > 140,000 kinetic parameters (Table 1). In addition to the numeric values, the experimental conditions are given in a commentary as a text in order to account for the different procedures for enzyme characterization in the laboratory. A web portal for the deposition of enzyme kinetic parameters has been developed in cooperation with the STRENDA commission (http://www.strenda.org/) [4]. This will increase the availability of well-defined kinetic parameters that are essential for systems biology approaches. Each entry in BRENDA is linked to a literature reference. This makes it possible to retrieve detailed information from the original literature (provided the literature is accessible as online version).

The pI-value has recently been included into the section of functional parameters. The isoelectric point provides information about the pH at which the protein carries no net electrical charge. This value is of significance for the purification procedure allowing conclusions about the solubility of the enzyme and its motility in electrophoretic procedures. Presently BRENDA contains more than 1,700 pI-values.

The reactivity of mutant enzymes can reveal detailed insights into the catalytic process and may give valuable clues about the active sites, the mechanism of the reaction, or the regulation. Meanwhile ~19,000 engineered enzymes are described in the database. For each single modification of the protein sequence, the properties of the resulting enzyme are described. Kinetic data for these enzymes are included in the respective database sections.

Table 1. Data statistics for the various sections of the database.

Enzyme Information	Single Data*
Nomenclature	70972
Isolation & preparation	53364
Stability	34532
Reaction & specificity	396760
Enzyme structure	232824
Functional & kinetic parameters	191134
Km Value	76894
Ki Value	14014
pI Value	1745
Turnover Number	20493
Specific Activity	30070
pH Optimum	26220
pH Range	6344
Temperature Optimum	13354
Temperature Range	2000
Organism-related information	80964
Source Tissue	56557
Localization	24407
References	91403
Enzyme application	3854
Enzyme-related diseases	52558
Mutant enzymes	18194

* These numbers refer to the combination of enzyme–organism–(protein-)value.

Organism-related information

Because enzymes and their properties vary greatly depending on the organism (e.g. eukaryotic or prokaryotic) it is highly important to link enzyme data to their source organism. Presently BRENDA covers information on enzymes of more than 7,500 different organisms (Fig. 1). With ~170,000 single data human enzymes are the most thoroughly described in the literature, followed by enzymes of the rat (~132,000 entries) and *Escherichia coli* (~93,000 entries).

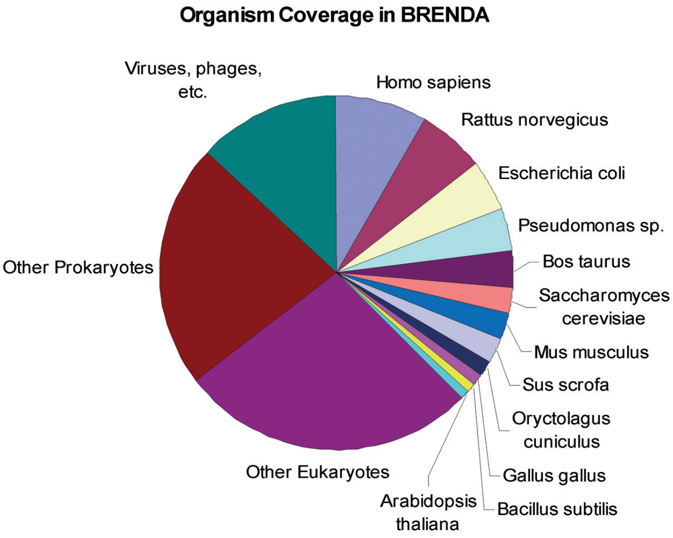
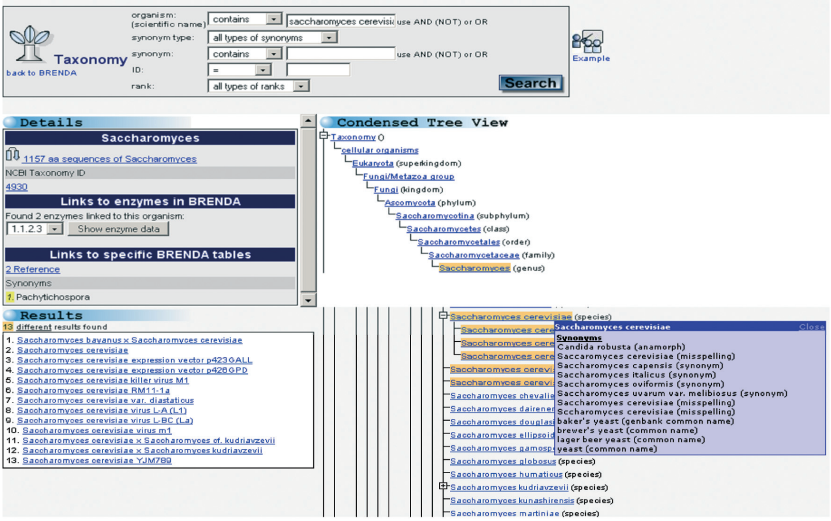


Figure 1. Organism coverage in BRENDA data.

All organisms are integrated into the BRENDA TaxTree (Fig. 2). The researcher may search along the TaxTree or switch to higher or lower branches to get an overview in e.g. a class or family or may focus the search on a specific species. Most of the TaxTree entries are linked to the NCBI taxonomy database. A small number of organisms cannot be linked to this tree because they do not appear in the NCBI taxonomy tree.



organism: (scientific name) contains | saccharomyces cerevisi: use AND (NOT) or OR
synonym type: all types of synonyms
synonym: contains | use AND (NOT) or OR
ID: =
rank: all types of ranks

Details

Saccharomyces

1157 aa sequences of Saccharomyces

NCBI Taxonomy ID

4930

Links to enzymes in BRENDA

Found 2 enzymes linked to this organism:

1.1.2.3 | Show enzyme data

Links to specific BRENDA tables

2 Reference

Synonyms

1. Pachytichospora

Condensed Tree View

```
Taxonomy ()
 └cellular organisms
   └Eukaryota (superkingdom)
     └Fungi/Metazoa group
       └Fungi (kingdom)
         └Ascomycota (phylum)
           └Saccharomycotina (subphylum)
             └Saccharomycetes (class)
               └Saccharomycetales (order)
                 └Saccharomycetaceae (family)
                   └Saccharomyces (genus)
```

Results

13 different results found

1. Saccharomyces bayanus x Saccharomyces cerevisiae
2. Saccharomyces cerevisiae
3. Saccharomyces cerevisiae expression vector p423GALL
4. Saccharomyces cerevisiae expression vector p426GPD
5. Saccharomyces cerevisiae killer virus M1
6. Saccharomyces cerevisiae RM11-1a
7. Saccharomyces cerevisiae var. diastaticus
8. Saccharomyces cerevisiae virus L-A (L1)
9. Saccharomyces cerevisiae virus L-BC (La)
10. Saccharomyces cerevisiae virus m1
11. Saccharomyces cerevisiae x Saccharomyces cf. kudriavzevii
12. Saccharomyces cerevisiae x Saccharomyces kudriavzevii
13. Saccharomyces cerevisiae YJM789

```
⊟Saccharomyces cerevisiae (species)
 ├Saccharomyces cere
 ├Saccharomyces cere
 └Saccharomyces cere
├Saccharomyces cerevi
├Saccharomyces cerevi
├Saccharomyces chevalie
├Saccharomyces dairener
├Saccharomyces douglasi
├Saccharomyces ellipsoid
├Saccharomyces gamosp
├Saccharomyces globosus (species)
├Saccharomyces humaticus (species)
⊞Saccharomyces kudriavzevii (species)
├Saccharomyces kunashirensis (species)
├Saccharomyces martiniae (species)
```

Saccharomyces cerevisiae Close

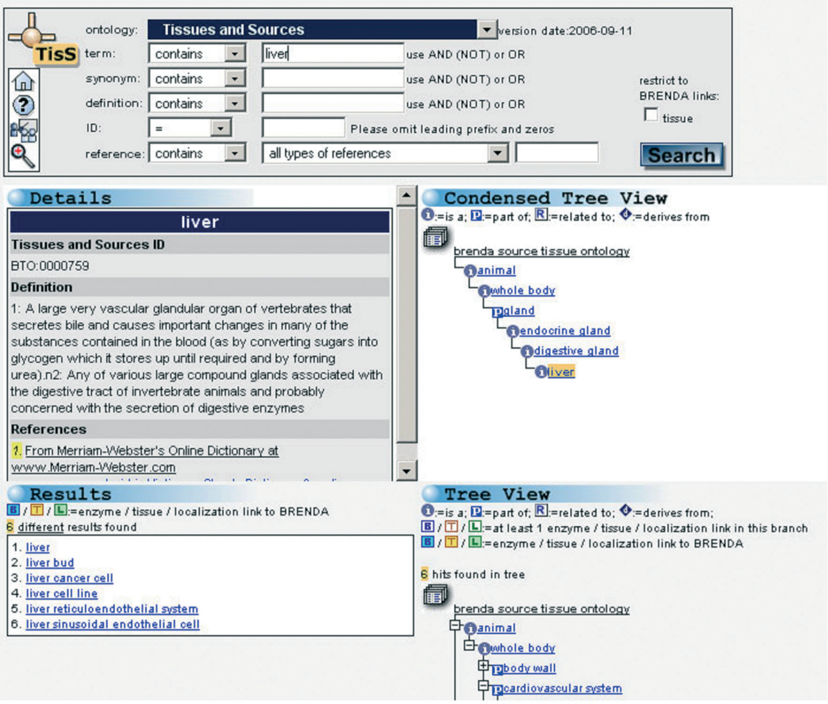
Synonyms
Candida robusta (anamorph)
Saccaromyces cerevisiae (misspelling)
Saccharomyces capensis (synonym)
Saccharomyces italicus (synonym)
Saccharomyces oviformis (synonym)
Saccharomyces uvarum var. melibiosus (synonym)
Saccharomyes cerevisiae (misspelling)
Sccharomyces cerevisiae (misspelling)
baker's yeast (genbank common name)
brewer's yeast (common name)
lager beer yeast (common name)
yeast (common name)

Figure 2. Sample of the search in the TaxTree.

BRENDA tissue ontology

For multicellular organisms it is not sufficient to relate enzyme data to the organism alone. The biochemical and molecular properties of one enzyme in different tissues or cell types can vary enormously. The information about the source of an enzyme, i.e. the tissue or cell-lineage therefore is vitally important. The occurrence of enzymes can be restricted to a specific cell type, cell line, or tissue from uni- and multicellular organisms, or can occur ubiquitously. BRENDA has developed its own ontology [1] in which the tissues are sorted hierarchically, corresponding to the format and rules of the Gene Ontology Consortium [5]. The tissue tree in BRENDA is divided into four areas, i.e. animal, plant, fungi and other sources, separated into subtrees. Most of the terms have definitions and synonyms which all can be displayed in the hierarchical tree.

In addition to the occurrence of the tissue, the localization of the enzyme within the cell is given. BRENDA provides a controlled vocabulary in cooperation with the GO consortium. A common shared vocabulary of the cellular components terms has been developed.
Both, the tissue and localization terms are classified in a concise ontology and the localization vocabulary is consistent with the GO terms.



ontology: Tissues and Sources — version date:2006-09-11
term: contains — liver — use AND (NOT) or OR
synonym: contains — use AND (NOT) or OR
definition: contains — use AND (NOT) or OR
ID: = — Please omit leading prefix and zeros
reference: contains — all types of references
restrict to BRENDA links: □ tissue

Details

liver

Tissues and Sources ID

BTO:0000759

Definition

1: A large very vascular glandular organ of vertebrates that secretes bile and causes important changes in many of the substances contained in the blood (as by converting sugars into glycogen which it stores up until required and by forming urea).n2: Any of various large compound glands associated with the digestive tract of invertebrate animals and probably concerned with the secretion of digestive enzymes

References

1. From Merriam-Webster's Online Dictionary at www.Merriam-Webster.com

Condensed Tree View

i:=is a; P:=part of; R:=related to; ◊:=derives from

brenda source tissue ontology
- i animal
  - i whole body
    - P gland
      - i endocrine gland
        - i digestive gland
          - i liver

Results

B / T / L:=enzyme / tissue / localization link to BRENDA

6 different results found

1. liver
2. liver bud
3. liver cancer cell
4. liver cell line
5. liver reticuloendothelial system
6. liver sinusoidal endothelial cell

Tree View

i:=is a; P:=part of; R:=related to; ◊:=derives from;
B / T / L:=at least 1 enzyme / tissue / localization link in this branch
B / T / L:=enzyme / tissue / localization link to BRENDA

6 hits found in tree

brenda source tissue ontology
- ⊟ i animal
  - ⊟ i whole body
    - ⊞ P body wall
    - ⊟ P cardiovascular system

Figure 3. Sample of the search in the BRENDA Ontology (BTO).

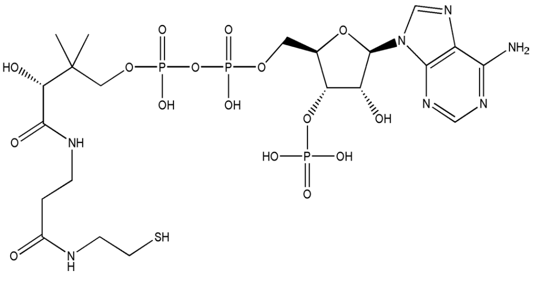
Ligands and metabolites

Enzymes interact with ligands in manifold ways. These can be substrates, products, prosthetic groups, cofactors, but also activating, stabilizing or inhibiting compounds. The present version contains ~88,500 different ligand names. Of these 52,250 molecules are stored as 2D structures in MOL-format. Generic compound names (e. g. "dextrans" or "carboxylic acid") amount to ~10,000 entries. Applying the organism-specific search option ligands occur in:

- 737,240 enzyme/ligand relationships
- 424,186 enzyme/substrate relationships
- 396,270 enzyme/product relationships
- 16,010 enzyme/cofactor relationships
- 107,331 enzyme/inhibitor relationships
- 17,563 enzyme/activating compound relationships
- 26,303 enzyme/metal or ion relationships

When searching for enzyme ligands or response modifiers two different query procedures are possible:

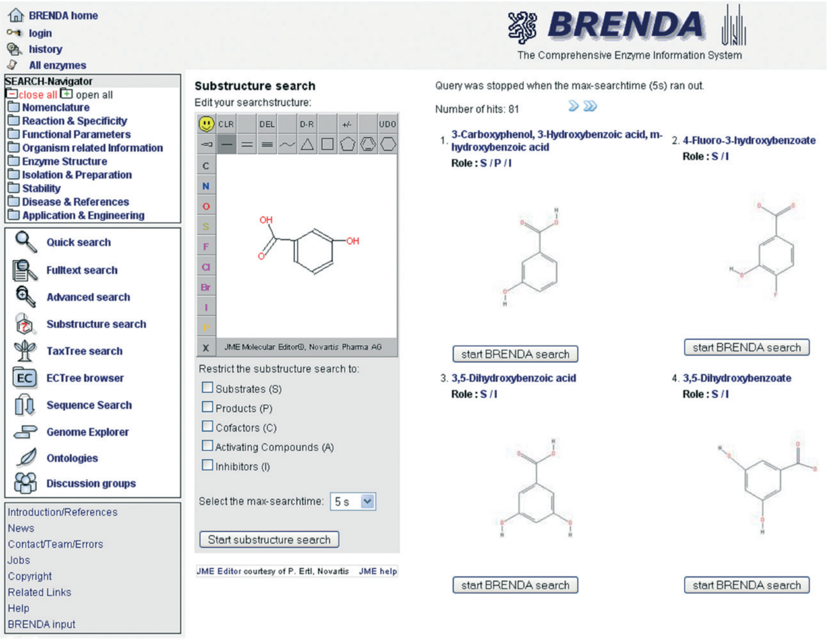
- Using the name of the compound: This option returns not only the data stored for the ligand under the given name but applies the integral molecular thesaurus. The newly generated thesaurus is based on the InChI (IUPAC International Chemical Identifier) [4] codes of the molecular structures stored as molfiles. An InChI is a non-proprietary identifier for chemical substances that can be used in printed and electronic data sources thus enabling easier linking of diverse data compilations. In earlier versions of the database unique isomeric SMILES [7, 8] were used for the calculation of the thesaurus. This procedure has been abandoned since it sometimes caused problems with complex structures.



InChI = 1/C 21H36N7O16P3S/c1 – 21(2,16(31)19(32)24 – 4-3 – 12(29)23 – 5-6 –
48)8 – 41 – 47(38,39)44 – 46(36,37)40 – 7-11 – 15(43 – 45(33,34)35)14(30)20(42 –
11)28 – 10 – 27 – 13 – 17(22)25 – 9-26 – 18(13)28 /h9 – 11,14 – 16,20,30 – 31,48 H,3 –
8H2,1 – 2H3,(H,23,29)(H,24,32)(H,36,37)(H,38,39)(H2,22,25,26)(H2,33,34,35)/t11-
,14-,15-,16+,20-/m1 /s1 /f/h23 – 24,33 – 34,36,38 H,22H2

Figure 4. Structure and InChI code for coenzyme A.

- Performing a substructure search (Fig. 5) with the integrated JME Editor [9]. This is an easy to use Java application for drawing molecules. The search can be restricted to a specific function (e. g. substrates). The results page displays the images, names, and synonyms of the found compounds, their function when interacting with the enzyme and also provides a button for an immediate BRENDA search.



Restrict the substructure search to:

□ Substrates (S)
□ Products (P)
□ Cofactors (C)
□ Activating Compounds (A)
□ Inhibitors (I)

Select the max-searchtime: 5 s

Query was stopped when the max-searchtime (5s) ran out.

Number of hits: 81

1. 3-Carboxyphenol, 3-Hydroxybenzoic acid, m-hydroxybenzoic acid
Role : S / P / I

2. 4-Fluoro-3-hydroxybenzoate
Role : S / I

3. 3,5-Dihydroxybenzoic acid
Role : S / I

4. 3,5-Dihydroxybenzoate
Role : S / I

Figure 5. Substructure search.

New Databases at the BRENDA Host

For the BRENDA enzyme database the references for manual annotation are chosen from the results of database searches in literature databases such as PubMed [10] and Chemical Abstracts (SciFinder) [11]. For some enzyme classes it is possible to include the complete literature that has been published for a specific enzyme. For the vast majority of enzymes, however, this is impossible for several reasons.

- the number of annually published references is too large to keep up with in the manual annotation capacities
- The literature on enzymes also covers aspects which are not in the focus of the BRENDA database. These may be reports on the genome-annotation, global expression of proteins, and literature in which the enzyme is used in a standard

> assay as a tool without any information on the enzyme's properties. References of this kind are not taken into consideration for BRENDA since they would only increase the statistic number of references per enzyme without providing more information and may even reduce the conciseness

For specific projects the user however might wish to retrieve a complete list of references for an enzyme. This would require a PubMed [10] search not only with the recommended name of the enzyme, but also with all the synonyms which are used. Conducting a single search for each synonym might be very time-consuming because most enzymes are used with different names, some even with hundreds of names as can be seen from Table 2

Table 2. Multiple synonyms for enzymes.

EC-number	Recommended Name	No. of Synonyms
2.7.10.1	receptor protein-tyrosine kinase	416
3.1.21.4	type II site-specific deoxyribonuclease	368
1.6.5.3	NADH dehydrogenase (ubiquinone)	169
3.1.3.48	protein-tyrosine-phosphatase	176
5.2.1.8	Peptidylprolyl isomerase	161

Similarly, searching for the complete literature on an enzyme in a specific organism or tissue would require searching with all known synonyms, common and scientific names. Since especially organism names have changed frequently because of taxonomic requirements a search in PubMed [10] with organism synonyms would require much time and a good knowledge in taxonomy. Similarly a search for an enzyme in a specific tissue would require a detailed knowledge of animal or plant anatomy.

In order to provide complete sets of references for all enzymes two databases were added at the BRENDA host.

FRENDA

FRENDA (**F**ull **RE**ference **EN**zyme **DA**ta) is an additional database to BRENDA available to the academic community with BRENDA release 6.2 (June 2006). FRENDA aims at providing an exhaustive collection of indexed literature references containing organism-specific enzyme information. Compared to a standard PubMed [10] query, FRENDA also returns all references on the enzyme published under one of its synonyms.

FRENDA currently covers 1.4 million enzyme/organism combinations from 550,000 distinct references, automatically extracted from more than 16 million PubMed abstracts (June 2006) [10]. The scientific articles are pre-filtered using MeSH terms [12] – only references declared as ”enzyme” hits are used (1.6 million remaining abstracts). FRENDA uses a dictionary-based approach for recognizing named entities (enzymes, organisms) in titles and abstracts. The dictionaries are compiled from BRENDA and NCBI Taxonomy [10]. The text-mining proceeds in a two-step approach:

1. Identification of the enzyme names (recommended names and synonyms) in title, abstract or MeSH terms,
2. Searching for co-occurring organism names (scientific names and synonyms) in title, abstract or MeSH terms.

The results of this indexing process were classified into 4 reliability categories depending on the occurrence of search terms in title and/or abstract and/or MeSH terms.

- Enzyme name and organism occur in the title or abstract but not in the same sentence. These hits are discarded.
- + Enzyme name and organism occur in the same sentence in the abstract or they both occur in the title
- ++ EC-number occurs in the MeSH-Terms or in the abstract, the organism occurs in the title or in the Abstract
- +++ Enzyme name and organism occur in the same sentence in the abstract and they both occur in the title
- ++++ Enzyme name and organism occur in the same sentence in the abstract, they both occur in the title and the EC-number is found in the abstract or in the MeSH terms

This classification is provided with the commentaries in the FRENDA database.
The manual evaluation of the quality of the FRENDA approach using 250 randomly chosen results indicates a precision of 64.8% with a recall of 72% from a set of 250 manually annotated enzyme-related literature references.

AMENDA
As a subset of FRENDA, AMENDA (Automatic Mining of ENzyme DAta) currently covers organism-specific information on enzyme localization (more than 30,000 records, compared to 17,000 records in BRENDA) and source tissues (roughly 150,000 records, compared to 38,000 records in BRENDA) from a text-mining procedure (to be published).

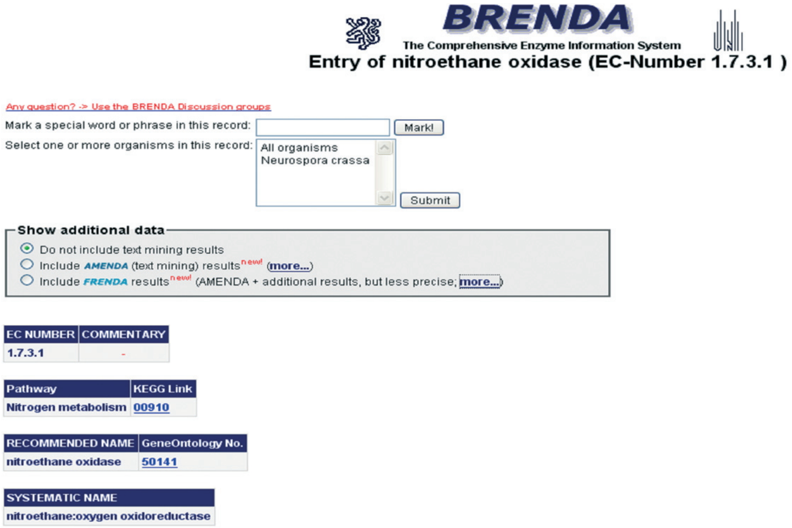
Search terms for enzyme names, organism names, localization, and sources and tissues are compiled from BRENDA enzyme synonyms, the BRENDA tissue-tree (http://obo.sourceforge.net/cgi-bin/detail.cgi?_brenda) and the NCBI Taxonomy [10]. AMENDA is based on the FRENDA co-occurrence approach. Protozoa, viruses, and bacteria are excluded for tissue search. References with enzyme/organism hits are searched for occurrences of tissue terms (singular and plural) and localization terms in title, abstract, and MeSH terms and further evaluated based on text-mining criteria.

- + Enzyme name, localization (or tissue), and organism (or the corresponding synonyms) occur in the title or in the same sentence in the Abstract

- ++ Enzyme name, localization (or tissue) and organism (or the corresponding synonyms) occur in the title. EC-number is contained in the MESH terms assigned to this article or EC-number occurs in the Abstract
- +++ Enzyme name, localization (or tissue), and organism (or the corresponding synonyms) occur in the title and in the same sentence in the Abstract
- ++++ Enzyme name, localization (or tissue) and organism (or the corresponding synonyms) occur in the title and in the same sentence in the abstract. EC-number is contained in the MESH terms assigned to this article or EC-number occurs in the Abstract

The text mining approach described above was tested on 200 randomly selected results. A precision of approximately 76.0% for the combined search terms enzyme–organism–tissue/localization was achieved. In a way similar to FRENDA, the commentaries indicate the individual reliability level for each data set.

When searching for enzyme data the user can choose which data should be displayed. In the default selection only the manually annotated BRENDA data are displayed. With each data set an additional box is displayed which gives the choice to display FRENDA resp. AMENDA results. Entries from these databases are specifically flagged in order to distinguish them from the BRENDA data.



Entry of nitroethane oxidase (EC-Number 1.7.3.1)

Any question? -> Use the BRENDA Discussion groups

Mark a special word or phrase in this record: Mark!

Select one or more organisms in this record: All organisms, Neurospora crassa

Submit

Show additional data

● Do not include text mining results
○ Include AMENDA (text mining) results new! (more...)
○ Include FRENDA results new! (AMENDA + additional results, but less precise; more...)

EC NUMBER	COMMENTARY
1.7.3.1	-

Pathway	KEGG Link
Nitrogen metabolism	00910

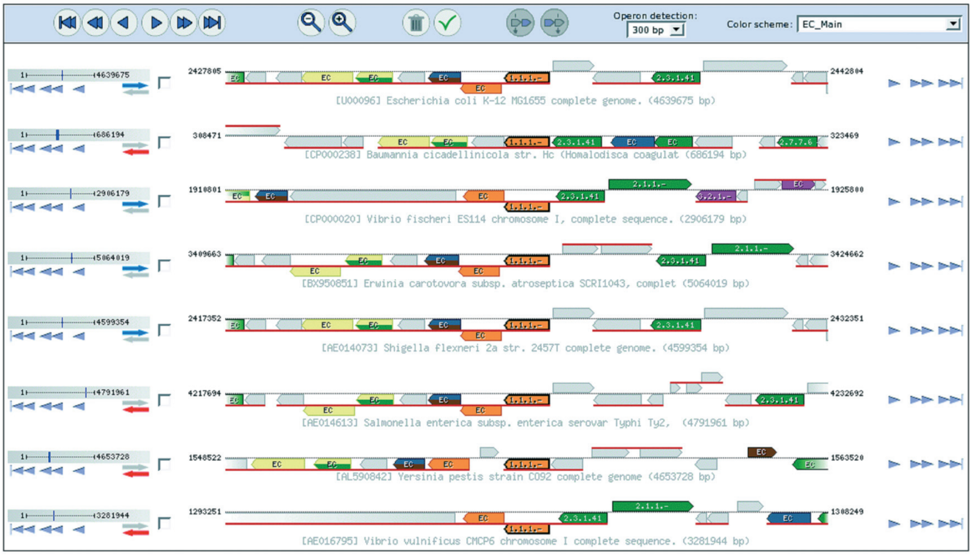
RECOMMENDED NAME	GeneOntology No.
nitroethane oxidase	50141

SYSTEMATIC NAME
nitroethane:oxygen oxidoreductase

Figure 6. The databases AMENDA and FRENDA can be displayed simultaneously with the BRENDA data.

BRENDA Genome Explorer

The BRENDA Genome Explorer is an enzyme-centred genome visualization tool for browsing and comparing enzyme annotations in full genomes. It closes the gap between genomic and enzymatic data and allows the alignment of genomes at a given enzyme-coding gene and its orthologs, thus allowing visual comparison of the genomic environment of the gene in different organisms (Fig. 2). The underlying genome database is compiled from EBI Genomes [13] and ENSEMBL [14] and supplemented by UniProt [15] annotations. It can be searched for specific proteins via names, EC-numbers, or UniProt accessions, allowing for a highly target-oriented search.



Operon detection: 300 bp

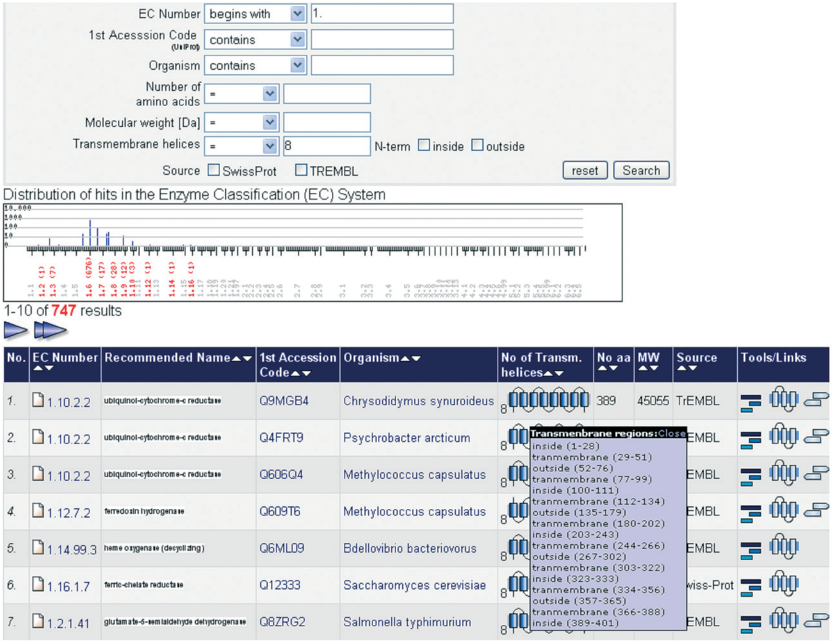
Color scheme: EC_Main

2427805
EC
EC
EC
EC
EC
1.1.1.-
2.3.1.41
2442804
[U00096] Escherichia coli K-12 MG1655 complete genome. (4639675 bp)
1 4639675
308471
EC
EC
1.1.1.-
2.3.1.41
EC
EC
2.7.7.6
323469
[CP000238] Baumannia cicadellinicola str. Hc (Homalodisca coagulat (686194 bp)
1 686194
1910801
EC
EC
EC
2.1.1.-
2.3.1.41
1.1.1.-
3.2.1.-
EC
1925800
[CP000020] Vibrio fischeri ES114 chromosome I, complete sequence. (2906179 bp)
1 2906179
3409663
EC
EC
EC
EC
1.1.1.-
2.3.1.41
2.1.1.-
3424662
[BX950851] Erwinia carotovora subsp. atroseptica SCRI1043, complet (5064019 bp)
1 5064019
2417352
EC
EC
EC
EC
EC
1.1.1.-
2.3.1.41
2432351
[AE014073] Shigella flexneri 2a str. 2457T complete genome. (4599354 bp)
1 4599354
4217694
EC
EC
EC
EC
EC
1.1.1.-
2.3.1.41
4232692
[AE014613] Salmonella enterica subsp. enterica serovar Typhi Ty2, (4791961 bp)
1 4791961
1548522
EC
EC
EC
EC
1.1.1.-
EC
EC
1563520
[AL590842] Yersinia pestis strain CO92 complete genome (4653728 bp)
1 4653728
1293251
EC
2.1.1.-
2.3.1.41
1.1.1.-
EC
1308249
[AE016795] Vibrio vulnificus CMCP6 chromosome I complete sequence. (3281944 bp)
1 3281944

Figure 7. BRENDA Genome Explorer showing a part of a genome alignment for *Escherichia coli* erythronate-4-phosphate dehydrogenase, EC 1.1.1.290.

Transmembrane Protein Prediction

Transmembrane helices for enzymes are predicted with TMHMM (TransMembrane Hidden Markov Model) developed by Sonnhammer *et al.* [16]. With the aid of this tool it is possible to predict the number, the size and the location of trans-membrane helices, thereby discriminating soluble and membrane-bound enzymes.



EC Number	begins with	1.	
1st Acesssion Code (UniProt)	contains		
Organism	contains		
Number of amino acids	=		
Molecular weight [Da]	=		
Transmembrane helices	=	8	N-term □ inside □ outside
Source	□ SwissProt	□ TREMBL	

Distribution of hits in the Enzyme Classification (EC) System

1-10 of 747 results

No.	EC Number	Recommended Name	1st Accession Code	Organism	No of Transm. helices	No aa	MW	Source
1.	1.10.2.2	ubiquinol-cytochrome-c reductase	Q9MGB4	Chrysodidymus synuroideus	8	389	45055	TrEMBL
2.	1.10.2.2	ubiquinol-cytochrome-c reductase	Q4FRT9	Psychrobacter arcticum	8			EMBL
3.	1.10.2.2	ubiquinol-cytochrome-c reductase	Q606Q4	Methylococcus capsulatus	8			EMBL
4.	1.12.7.2	ferredoxin hydrogenase	Q609T6	Methylococcus capsulatus	8			EMBL
5.	1.14.99.3	heme oxygenase (decyclizing)	Q6ML09	Bdellovibrio bacteriovorus	8			EMBL
6.	1.16.1.7	ferric-chelate reductase	Q12333	Saccharomyces cerevisiae	8			wiss-Prot
7.	1.2.1.41	glutamate-5-semialdehyde dehydrogenase	Q8ZRG2	Salmonella typhimurium	8			EMBL

Transmenbrane regions:Close
inside (1-28)
tranmembrane (29-51)
outside (52-76)
tranmembrane (77-99)
inside (100-111)
tranmembrane (112-134)
outside (135-179)
tranmembrane (180-202)
inside (203-243)
tranmembrane (244-266)
outside (267-302)
tranmembrane (303-322)
inside (323-333)
tranmembrane (334-356)
outside (357-365)
tranmembrane (366-388)
inside (389-401)

Figure 8. Characteristic output of the trans-membrane prediction tool.

Accessibility

BRENDA is accessible via the various search options (quick search, advanced search, ontologies, sequence search, Genome Explorer etc.). The database will be downloadable as a text file from January 2007 on. Access to AMENDA and FRENDA requires a registration.

SOAP-Based Web Service

Web services provide a simple way to access the data collection without the need for downloading, parsing, and preparing an entire database for local queries. Web services are independent of the internal organization of the database and avoid parsing problems caused by changes in the text file structure.

BRENDA now provides a SOAP (Simple Object Access Protocol, http://www.w3.org/TR/soap) based web service comprised of 148 methods covering 52 data fields. Flexible queries can be performed directly from programs written in different programming languages (Perl, Java, C++, Python, PHP) on data fields such as substrate, K_m-value and pH-optimum. For any given record returned, a set of complete literature references can be retrieved using unique reference identifiers. Every data field may be queried by providing at least one of the three parameters EC-number, organism, or – if applicable – ligand structure identifier. The ligand structure identifier, which can be queried with the name of a chemical compound, is used to ensure that all synonyms for a given molecular structure are also retrieved.

The BRENDA web service also gives access to the data using identifiers from other databases like UniProt [14] or NCBI Taxonomy [10], as well as ontologies like Gene Ontology [5] or BRENDA Tissue Ontology [1]. The ontology-based search allows for queries based on entire branches of the hierarchy, avoiding a complex search for all leaves in the given branch. For example, an ontology-based search for the term 'brain' or the respective Gene Ontology identifier will return all tissues and cell types under the umbrella term 'brain'. The same method can also be applied to search for whole groups of organisms. The documentation of the BRENDA web service including examples in different programming languages is available at http://www.brenda.uni-koeln.de/soap.

Conclusions

The BRENDA enzyme information system has made a big step forward not only by a formidable increase in the annotation speed but also by inclusion of data based on text-mining approaches and by the development of different new methods for data access. The new funding by an EU grant allows the annotation speed to be increased even further to bring the backlog down to less than one year and will also allow a substantial increase in the percentage of ligands with full structural information.

Acknowledgements

This work was supported by the European Union (FELICS, Free European Life-Science Information and Computational Services: 0121902 (RII3)) and the German BMBF, Federal Ministry of Education, Science, Research and Technology (NGFN, Nationales Genomforschungsnetz: 0313398B)

References

[1] Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G., Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*. **32:**D 431 –D 433.

[2] Schomburg,I., Chang,A., Schomburg,D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res*. **30:**47–49.

[3] Schomburg,D., Schomburg,I. (2001–2006) *Springer Handbook of Enzymes*. 2nd Edn. Springer, Heidelberg, Germany.

[4] Kettner,C., Hicks,M.G. (2005) The dilemma of modern functional enzymology. *Curr. Enzyme Inhib*. **1:**171–181.

[5] Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*. **34:**D 322 –D 326.

[6] Stein, S.E., Heller,S.R., Tchekhovski, D. (2003) An Open Standard for chemical structure representation – The IUPAC Chemical Identifier. *Nimes International Chemical Information Conference Proceedings*, pp. 131–143.

[7] Weininger,D. (1988) SMILES, a chemical language and information system.1. Introduction to methodology and encoding rules. *J. Chem. Inform. Comput. Sci*. **28:**31–36.

[8] Weininger, D., Weininger, A., Weininger, J. (1989) SMILES. 2. algorithm for generation of unique SMILES notation. *J. Chem. Inform. Comput. Sci*. **29:**97–101.

[9] Csizmadia, P. (2000) MarvinSketch and MarvinView: molecule applets for the World Wide Web. *Proceedings of ECSOC-3 and Proceedings of ECSO-4,* 367–369.

[10] Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S. *et al*. (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **34:**D 173–180.

[11] Ridley,D.D. (2002) *SciFinder and SciFinder Scholar*. J. Wiley & Sons, New York.

[12] National Library of Medicine. (1960) Medical subject headings: main headings, subheadings, and cross references used in the Index Medicus and the National Library of Medicine Catalog. 1st Edn. Washington, DC: U.S. Department of Health, Education, and Welfare.

[13] Cochrane, G., Aldebert, P., Althorpe, N., Andersson, M., Baker, W., Baldwin, A., Bates, K., Bhattacharyya, S., Browne, P., v. d. Broek, A. *et al*. (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res.* **34:**D 10 –D 15.

[14] Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*. **34:**D556–561.

[15] Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*. **34:**D187–191.

[16] Sonnhammer, E. l. L., von Heijne, G., Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. In: *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*. (Glasgow, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff, D., Sensen, C. Eds), pp.175–182. AAAI Press, Menlo Park, CA.

JWS Online: A Web-Based Tool for Curation, Review, Storage and Analysis of Kinetic Models

Jacky L. Snoep[1,2,3], **Carel van Gend**[1],**Cor Stoof**[3], **Brett G. Olivier**[1], **Riaan Conradie**[1], **Franco B. Du Preez**[1], **Du Toit W.P. Schabort**[1], **Gerald Penkler**[1] **and Kora Holm**[1]

[1]Triple-J group for Molecular Cell Physiology, Department of Biochemistry, Stellenbosch University, Matieland 7602, South Africa

[2]Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester M1 7ND, U. K.

[3]Molecular Cell Physiology, Vrije Universiteit, De Boelelaan 1087, 1081 HV Amsterdam, The Netherlands

E-Mail: jls@sun.ac.za

Received: 9th November 2006 / Published: 31st August 2007

Abstract

In this contribution we report on the JWS Online project and the progress that has been made since the first ESCEC meeting. Whilst maintaining the same user interface, we have completely redesigned the server part of JWS Online, now a) using webMathematica as the interface between the HTML pages and the Mathematica [1] Kernel and b) storing all models as Mathematica packages, and c) using a PostgresQL [2] database to store a full description of each model.
In the last few years a number of new initiatives have started, of which some fulfil comparable roles to JWS Online and with some of which we collaborate. Here we compare JWS Online to these initiatives focusing on the three aims of JWS Online: 1) to be a repository for curated kinetic models of biological systems, 2) to be an easy to use simulator that can be accessed over the internet, 3) to help in the reviewing of manuscripts containing kinetic models.

Introduction

Mathematical Biology has a long history, especially in the field of population dynamics, with famous examples such as the description of Fibonacci for the growth of an idealized rabbit population and the Lotka–Voltera equations for predator–prey interactions. In these earlier models equations were often selected more on the basis of ease of use in mathematical analysis and less on knowledge of the biological system. In the neurosciences the work of Hodgkin and Huxley was a major breakthrough, not only for the understanding of the generation of the action potential but also in their approach to build a kinetic model of the neuron using kinetic parameters that were experimentally determined. More recently there has been a tremendous increase in the interest of applying kinetic models in the field of molecular and cellular biology. Whereas pioneering work in this field was done in the 1960 s by Chance, Garfinkel, Higgins and Hess e. g. [3], an enormous increase in the last decade in the construction of detailed kinetic models can probably be related to 1) the development in experimental fields (e. g. genomics) leading to detailed and information-rich data sets, 2) the increase in computing power and strength of simulation tools and 3) further development of strong analysis frameworks, (e. g. dynamical systems analysis, metabolic and hierarchical control analysis). The combined use of such theoretical, computational and experimental approaches has been characteristic for the field of **Systems Biology**, developed over the last five years, aiming at an understanding at a systemic level via the integration of our knowledge of the system's components (including their interactions).

Detailed kinetic models form a core component of **Systems Biology** studies. These models contain the experimental information on the components of the systems and their integration should result in the systemic **behaviour** observed for the complete system. These models are different from the traditional models that are made as simple as possible; such so-called core models are used to test a hypothesis or to illustrate a theoretical concept. Their simplicity makes core models amenable to robust mathematical analysis, without being sidetracked by unnecessary detail. Examples are the two and three variable models often used in bifurcation analysis of dynamic systems. Core models are important to get an understanding of the general behaviour of a system or a set of equations but it is often not possible to directly relate the model to experimental data and model validation is often made in more qualitative terms. In addition to these core models, systems biology has a need for a different kind of model, with a high level of detail and a direct, mechanistic interpretation of the model components. In **the Silicon Cell initiative** we have advocated the use of models with a high level of detail, containing experimentally determined parameter values (e. g. [4, 5]). We suggest measuring the model parameters of the isolated components (either *in vitro* or *in vivo*) and validating the model against the behaviour of the complete system, thus clearly separating model construction from model validation. In addition we suggested a modular approach, i. e. building detailed models of parts of the system, subsequently validating these models and combining them, followed by an additional round of validation. Ultimately such an approach would lead to a kinetic description of a complete system, for instance making a detailed kinetic model for the yeast *Saccharomyces cerevisiae*.

The Silicon Cell initiative would result in a significant number of models to be constructed. Even a simple unicellular system contains several thousand reactions, and a sensible split over modules would need to be made. Adding to the complexity is the capacity of living organisms to adapt themselves via regulation at the level of gene expression potentially any of the reaction steps can be modulated. This variable gene expression is one of the reasons one should model the cell in detail at the level of the enzyme catalysed reaction step. With time, a large collection of models, including metabolic, signal transduction, cell cycle and gene expression regulation models will be constructed which upon grouping will ultimately cover the complete cell.

To be able to link kinetic models together they must obey certain standards in terms of **annotation** (e.g. variable names should be identical) and the models should be described in a standard format (e.g. SBML [6]). In addition to the standardization of formats the models should also be available in **curated** form in repositories. In this contribution we highlight one initiative, JWS Online, which in addition to being such a repository for curated models is also a simulator with a web interface, making it possible to run the models in a browser. A third important aspect of JWS Online is its collaboration with scientific journals to assist in reviewing manuscripts that contain models. We start by describing JWS Online and its current set-up, focusing on the server side (the user interface was described in the last ESCEC contribution and has largely remained the same). Subsequently we will compare JWS Online with other initiatives that have comparable functionalities, i.e. the Virtual Cell, BioModels, DOQCS, Sigpath, JSim, ModelDB, Web-Cell and the SBML and CellMl repositories. We limit ourselves to these web-based initiatives, (and apologize for potential omissions), and have not included stand-alone simulators.

JWS Online

JWS Online [7] is hosted at the National Bioinformatics Node of the University of Stellenbosch. The first version of the web site went online in 2000 and since then a number of important updates have been made but the three main aims have remained the same. JWS Online is: 1) a repository of curated models, 2) a web-driven simulator and 3) a review facility for scientific journals. Models have been added steadily and currently 70 models are available, in three categories, Silicon Cell models, Core models and Demonstration models. JWS Online is mirrored at the Vrije Universiteit in Amsterdam (http://jjj.bio.vu.nl), at the Virginia Bioinformatics Institute (http://jjj.vbi.vt.edu) and at Manchester University (http://jjj.mib.man.ac.uk).

JWS Online works together with four journals to facilitate reviewing of manuscripts that contain kinetic models: *FEBS J.*, *Microbiology*, *IEE Proceedings Systems Biology* and *Metabolomics*. Authors who submit a manuscript containing a kinetic model are requested to submit their model to JWS Online (jls@sun.ac.za) in electronic format (i.e. SBML or JWS Online input form). Subsequently the model is converted into a Mathematica package that is stored in the JWS Online database. Using the JWS Online facility the simulations of the authors are repeated and if the results cannot be reproduced the authors are contacted to resolve the problem. Once the model is curated in this way, a letter is sent to the reviewers stating how they can access the model on a secure site and reproduce the results of the authors and otherwise interrogate the model. Once the reviewers have come to a decision regarding the manuscript, the model is either moved to the public database or deleted.

JWS Online collaborates with a number of other initiatives: **the Silicon Cell** initiative, **Biomodels** (see below), **YSBN**, the Yeast Systems Biology Network (http://www.gmm.gu.se/YSBN/), and **HepatoSys**, the BMBF funded German systems biology competence network of hepatocytes (http://www.systembiologie.de/en/index.html), and the **COPASI** team (http://www.copasi.org/).

The functionality of JWS Online was discussed in the first ESCEC proceedings [8] and here we will only briefly summarize the functionality of the simulator and describe the way the server side of JWS Online works.

JWS Online set-up

The JWS simulation system is based on a client–server architecture, where commands issued by the client (a Java applet in a web browser) are fulfilled by an instance of Mathematica running on the server, see Fig. 1 for a flow diagram. This is facilitated by a webserver (Apache Tomcat [9]) running webMathematica, which is responsible for allocation of Mathematica kernels from a pool, accepting client commands and sending these to the Kernel for evaluation, and returning the results to the client.

The JWS models are stored as Mathematica packages. These include values for the model input parameters, and also define the functionality available for the model. In particular, functions may be defined which calculate and plot a time course of the model, display the steady state of the model, or display the results of a metabolic control analysis. The details of each of these calculations are specific to a particular model, and are described in the package. In addition, the package may define only a subset of these functions, depending on what is appropriate for the model. An SQL database contains a full text description of each model, as well as links to the Mathematica package for that model.

On visiting the JWS site, the user is presented with a welcome screen, displaying basic site information. The user may then opt to choose a model from the database. An initial selection page is displayed, in which the user may select any or all of model organism, model category and subcategory and model author.

The selection request is then sent to the web server, where a Python [10] script extracts from the database those models that satisfy the selection criteria. These are displayed in the user's browser as a list, from which the user may choose to display detailed information about a particular model, or opt to run a model.

The request to run a model is returned to the server, where the webMathematica kernel manager allocates a Mathematica kernel from the kernel pool. This kernel then loads the model from the Mathematica model description, and passes the model parameters to the

JWS Java applet, which is downloaded to the browser. The applet is configured according to the functionality available for the specific model chosen; certain models, for example, allow a time-course simulation, steady-state analysis and the determination of metabolic control analysis information, while others allow only a subset of these.

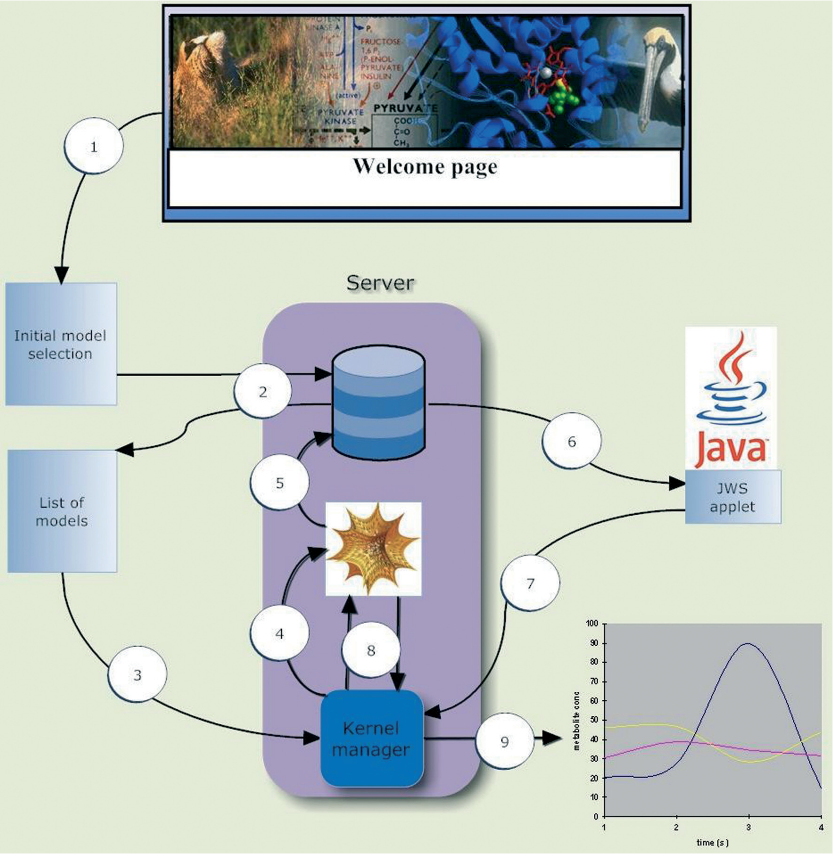


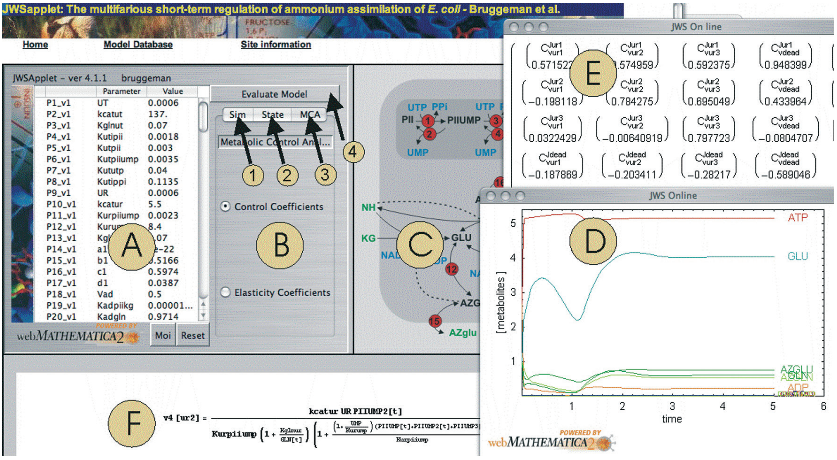
Figure 1. Flow diagram of the JWS Online set-up.

1) The user selects the link to the database of models, and a page is presented which allows the user to restrict the models to be displayed by organism type, model category and sub-category, and model author.
2) A request is sent to a Python script on the web server, which selects those models from the database, which satisfy the chosen criteria. These are then returned to the browser, which displays the list of possible models.
3) The user selects the model to run, and a request for the model details is sent to the server.
4) The webMathematica kernel manager allocates a Mathematica kernel.
5) The Mathematica kernel looks up the model details in a package file.
6) These are then passed to a Java applet, which is downloaded to the browser.

7) The applet sends a request to the webMathematica kernel manager to evaluate the model.
8) The evaluation request is sent to the Mathematica kernel, and on completion the results are returned to the kernel manager.
9) Finally, the results of the computation are sent to and displayed on the client machine.

JWS Online functionality

Interaction with JWS Online is done through a graphical user interface (GUI). A screen shot and two result windows are shown in Fig. 2. The interface consists of a number of panels where the user can make changes to the default parameter values of the model (Fig. 2, A), control the type of analysis that is required (Fig. 2, B), view a scheme of the model (Fig. 2, C) and its rate equations (Fig. 2, F) by moving the mouse over the red ovals in the scheme. Results are shown in separate windows (Fig. 2, D and E) depending on the type of simulation that is selected. In panel B the user can select for 1) a time simulation (Fig. 2, arrow 1), giving the options to plot either metabolite concentrations or flux values, 2) a steady-state analysis (Fig. 2, arrow 2), giving the options to do different types of structural analyses or to analyse for the steady-state solution, or 3) to do a Metabolic Control Analysis (Fig. 2, arrow 3), giving the options to either calculate the control coefficients or the elasticity coefficients. After selecting an analysis type the user can evaluate the model by clicking the Evaluate button (Fig. 2, arrow 4) and the results will be shown in a separate window. Examples of results windows are shown for a time simulation (Fig. 2, D) and a MCA analysis for control coefficients (Fig. 2, E).



JWSapplet: The multifarious short-term regulation of ammonium assimilation of *E. coli* - Bruggeman et al.

JWSApplet – ver 4.1.1 bruggeman

	Parameter	Value
P1_v1	UT	0.0006
P2_v1	kcatut	137.
P3_v1	Kglnut	0.07
P4_v1	Kutipii	0.0018
P5_v1	Kutpii	0.003
P6_v1	Kutpiiump	0.0035
P7_v1	Kututp	0.04
P8_v1	Kutippi	0.1135
P9_v1	UR	0.0006
P10_v1	kcatur	5.5
P11_v1	Kurpiiump	0.0023
P12_v1	Kuru	8.4
P13_v1	Kg	.07
P14_v1	a1	e-22
P15_v1	b1	.5166
P16_v1	c1	0.5974
P17_v1	d1	0.0387
P18_v1	Vad	0.5
P19_v1	Kadpiikg	0.00001...
P20_v1	Kadgln	0.9714

Metabolic Control Anal...

● Control Coefficients

○ Elasticity Coefficients

JWS On line

C^{Jur1}_{vur1} 0.57152	C^{Jur1}_{vur2} .574959	C^{Jur1}_{vur3} 0.592375	C^{Jur1}_{vdead} 0.948399
C^{Jur2}_{vur1} −0.198118	C^{Jur2}_{vur2} 0.784275	C^{Jur2}_{vur3} 0.695049	C^{Jur2}_{vdead} 0.433964
C^{Jur3}_{vur1} 0.0322429	C^{Jur3}_{vur2} −0.00640919	C^{Jur3}_{vur3} 0.797723	C^{Jur3}_{vdead} −0.0804707
C^{Jdead}_{vur1} −0.187869	C^{Jdead}_{vur2} −0.203411	C^{Jdead}_{vur3} −0.28217	C^{Jdead}_{vdead} −0.589046

JWS Online

Figure 2. Screen shot of the JWS Online user interface and result windows. A screen shot is made of the JWS Online implementation of the model for the regulation of ammonia assimilation in *Escherichia coli* [23]. The Interface consists of different panels, A,B,C,F that allow control over the simulation and give information on the model (see text for details). In addition two result windows are shown (panel D and E) displaying a time simulation and an MCA result respectively.

The JWS Online team

Initially JWS Online was started in 2000 by Jacky Snoep and Brett Olivier as a challenge to see whether we could run Mathematica simulations over the internet. At that time web-Mathematica had not been developed and we used JLink to connect Java and Mathematica. JWS Online was launched in 2000 and has subsequently been significantly improved in a number of steps, the last one being the conversion to webMathematica on which we report here and for which Cor Stoof did the necessary Java programming.

At present Jacky Snoep is the PI of the JWS Online project with Carel van Gend as full time programmer. On a part-time basis Brett Olivier maintains the web site and Riaan Conradie, Franco Du Preez and Du Toit Schabort assist in coding models for the repository, Gerald Penkler and Kora Holm draw the metabolic schemes and make literature searches for manuscripts containing models.

Other Initiatives

Here we give a brief description of some other initiatives that have overlapping functionality with JWS Online. We have only listed initiatives that provide a repository of kinetic models for biological systems that are accessible via the internet (Table 1).

Table 1. A comparison between several web-based initiatives that store kinetic models and/or make models available for simulation. The initiatives are compared on their functionality with respect to whether they allow simulations to be run on the site (simulation), whether they store a collection of models (repository), whether the stored models are curated (i.e. do the models show the same behaviour as the published model, curation), whether the models are annotated (annotation) and whether the initiative is actively busy to add more models (here copying from other initiatives is not considered active, addition).

Initiative	URL	Simulator	Curation	Annotation	Addition
JWS Online	http://jjj.biochem.sun.ac.za	Yes	Yes	No	Yes
Virtual Cell	http://www.nrcam.uchc.edu	Yes	No	No	No
Biomodels	http://www.ebi.ac.uk/biomodels	No	Yes	Yes	Yes
WebCell	http://webcell.kaist.ac.kr	Yes	No	No	No
CellML	http://www.cellml.org	No	No	No	Yes
SBML	http://sbml.org	No	Yes	No	Yes
DOQCS	http://doqcs.ncbs.res.in/	No	Yes	Yes	Yes
ModelDB	http://senselab.med.yale.edu/senselab/ModelDB/	No	Yes	No	Yes
JSim	http://nsr.bioeng.washington.edu/	Yes	No	No	Yes
SigPath	http://www.sigpath.org/	No	No	Yes	No

The Virtual Cell [11,12] is hosted at the National Resource for Cell Analysis and Modeling (NRCAM) at the University of Connecticut Health Center and is a computational environment that helps in the construction and simulation of models that are cast in terms of ODEs or PDEs. The Virtual Cell follows a client–server set-up running Java applets;

clients can store models in a repository and import/export facilities for SBML, CellML and Matlab exist. The models are not curated or annotated (the client is responsible) and the Virtual Cell team does not actively add models to the repository.

The **Biomodels** [13] database is hosted at EMBL-EBI (UK) and is a collaborative effort between this institute, the SBML team (U.S.A.), the Systems Biology Group of the Keck Graduate Institute (U.S.A.), the Systems Biology Institute (Japan) and JWS Online. Biomodels focuses on model curation, annotation and import/export formats of published models. Models are curated to ensure that the published results can be reproduced. In the annotation process model components are linked to controlled vocabularies and other data resources. Models to be included in the database must be compliant with MIRIAM standards [14]. Both the Biomodels and JWS Online project are actively involved in adding models to their databases and these models are exchanged in SBML format between the two initiatives.

DOQCS [15] is hosted at The National Centre for Biological Sciences (NCBS) and is part of the Tata Institute of Fundamental Research in Bangalore. The Database of Quantitative Cellular Signaling is a repository of models of signalling pathways. It includes reaction schemes, concentrations, rate constants, as well as annotations on the models. The database provides a range of search, navigation and comparison functions. Export of models is available in GENESIS [16] and MATLAB (http://www.mathworks.com) format.

The **CellML** [17] and former **SBML** repositories hosted at the University of Auckland and CalTech respectively are repositories of kinetic models in XML format. The two modelling languages have significant overlap, CellML is aiming more at describing systems at the cellular level while SBML is better geared for reaction pathway models. CellML appears to have more freedom to define entities as components but is not as widely accepted as a format in simulation software. The models of the SBML repository have been improved and incorporated into BioModels Database.

SigPath [18] is hosted at the Weill Medical College of Cornell University, and at the Mount Sinai School of Medicine, it is an information management system designed to support quantitative studies on the signalling pathways and networks of the cell. SigPath focuses on storing, curating and annotating of quantitative information concerning signalling pathways. This information can be manipulated and reactions can be linked to form kinetic models. Some of these models (which are not curated as such) are available as a repository and can be exported in a number of formats amongst which SBML.

ModelDB [19] is a repository for published models from the neurosciences, it is part of the SenseLab project and hosted at Yale University. The models are available in the format in which it was submitted to the database (e. g. Fortran, NEURON).

JSim [20] is a simulation environment that can be used for the construction of models, a selected number of models is also available as Java applets and can be run over the web. JSim is closely linked to the NSR Physiome project, which provides comprehensive and downloadable physiological models [21].

WebCell [22] is hosted at the Korea Advanced Institute of Science and Technology (KAIST) and uses a client–server set-up with Java Servlet Pages and applets. New models can be added by the clients and stored in the database. The current models in the database are taken from the JWS Online, Biomodels and SBML repositories. The simulation functionality is similar to JWS Online.

Acknowledgements

The National Bioinformatics Network (NBN; South Africa) has provided funding for the infrastructure of the Stellenbosch node, including the JWS Server. In addition, C.v.G., B.G.O., R.C., and F.D.P. receive generous support from the NBN.

References

[1] Wolfram Research, Inc., Mathematica, Version 5.2, Champaign, IL (2005).

[2] http://www.postgresql.org/, PostgreSQL, Version 8.1.

[3] Chance, B., Garfinkel, D., Higgins, J., Hess, B. (1960) Metabolic control mechanisms V. A solution for the equations representing interaction between glycolysis and respiration in ascites tumor cells. *J. Biol. Chem.* **235:**2426–2439.

[4] Snoep, J.L., Westerhoff, H.V. (2004) The silicon cell initiative. *Current Genomics* **5:**687–697.

[5] Snoep, J.L. (2005) The SiliconCell initiative: working towards a detailed kinetic description at the cellular level. *Current Opin. Biotechnol.* **16:**336–343.

[6] Hucka, M., Finney, A., Sauro, H.M., Bolouri, H, Doyle, J.C., Kitano, H. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19:**524–531.

[7] Olivier, B.G., Snoep, J.L. (2004) Web-based kinetic modelling using JWS Online. *Bioinformatics* **20:**2143–2144.

[8] Snoep, J.L., Olivier, B.G., Westerhoff, H.V. (2004) JWS Online cellular systems modeling and the silicon cell. In: *Experimental Standard Conditions of Enzyme Characterizations.* (Hicks, M.G., Kettner, C. Eds), pp. 129–143. Logos Verlag, Berlin, Germany.

[9] http://tomcat.apache.org, Apache Tomcat, Version 5.5

[10] http://www.python.org, Python, Version 2.4

[11] Loew, L.M., Schaff, J.C. (2001) The Virtual Cell: a software environment for computational cell biology. *Trends Biotechnol.* **19:**401–406.

[12] Slepchenko, B.M., Schaff, J.C., Macara, I., Loew, L.M. (2003) Quantitative cell biology with the Virtual Cell. *Trends Cell Biol.* **13:**570–576.

[13] Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J.L., Hucka, M. (2006) BioModels Database: A free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.* **34:**D689 – D691.

[14] Le Novère, N., Finney, A., Hucka, M., Bhalla, U., Campagne, F., Collado-Vides, J., Crampin, E., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J.L., Spence, H.D., Wanner, B.L. (2005) Minimal information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnol.* **23:**1509–1515.

[15] Sivakumaran, S., Hariharaputran, S., Mishra, J., Bhalla, U.S. (2003) The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks. *Bioinformatics* **19:**408–415.

[16] Bhalla, U.S. (2002) Use of Kinetikit and GENESIS for modeling signaling pathways. *Method. Enzymol.* **345:**3**–**23.

[17] Lloyd, C.M., Halstead, M.D.B., Nielsen, P.F. (2004) CellML: its future, present and past. *Progr. Biophys. Molec. Biol.* **85:**433–450.

[18] Campagne, F., Neves, S., Chang, C., Skrabanek, L., Ram, P.T., Iyengar, R., Weinstein, H. (2004) Quantitative Information Management for the Biochemical Computation of Cellular Networks, Sci. STKE 248, PL11.

[19] Migliore, M., Morse, T.M., Davison, A.P., Marenco, L., Shepherd, G.M., Hines, M.L. (2003) ModelDB Making models publicly accessible to support computational neuroscience. *Neuroinformatics* **1:**135–140.

[20] http://nsr.bioeng.washington.edu, Jsim

[21] http://www.physiome.org/Models/

[22] Lee, D.-Y., Yun, C., Cho, A., Hou, B.K., Park, S., Lee, S.U. (2006) WebCell: a web-based environment for kinetic modeling and dynamic simulation of cellular networks. *Bioinformatics* **22:**1150–1151.

[23] Bruggeman, F.J., Boogerd, F.C., Westerhoff, H.V. (2005) The multifarious short-term regulation of ammonium assimilation of Escherichia coli: dissection using an in silico replica. *FEBS J.* **272:**1965–1985.

The Estimation of Kinetic Parameters in Systems Biology by Comparing Molecular Interaction Fields of Enzymes

Matthias Stein, Razif R. Gabdoulline, Bruno Besson, Rebecca C. Wade

EML Research gGmbH, Molecular and Cellular Modeling Group,
Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany

E-Mail: matthias.stein@eml-r.villa-bosch.de

Received: 16th June 2006 / Published: 31st August 2007

Abstract

The kinetic modelling of biochemical pathways requires a consistent set of enzymatic kinetic parameters. We report results from software development to assist the user in systems biology, allowing the retrieval of heterogeneous protein sequence, structural and kinetic data. For the simulation of biological networks, missing enzymatic kinetic parameters can be calculated using a similarity analysis of the enzymes' molecular interaction fields. The quantitative PIPSA (qPIPSA) methodology relates changes in the molecular interaction fields of the enzymes with variations in the enzymatic rate constants or binding affinities. As an illustrative example, this approach is used to predict kinetic parameters for glucokinases from *Escherichia coli* based on experimental values for a test set of enzymes. The best correlation of the electrostatic potentials with kinetic parameters is found for the open form of the glucokinases. The similarity analysis was extended to a large set of glucokinases from various organisms.

Introduction

One of the aims of systems biology is to provide a mathematical description of metabolic or signalling protein networks. This can be achieved by constructing a set of differential equations describing changes in concentrations of compounds with time [1]. Enzyme-specific parameters, such as ligand binding affinity and catalytic turnover, are needed for solving these equations. These parameters need to be valid under the desired experimental conditions. Despite recent developments in enzymatic high-throughput assays, experimental values of many of the required parameters often are not available for the chosen organism or enzyme, or have not been determined at the desired temperature or pH [2].

For the construction of a kinetic model, it is essential to have a consistent and reliable set of enzymatic kinetic parameters. The importance of the uniformity of the measurement and reporting of enzymatic functional data has been emphasized in [3].

Molecular systems biology deals with the intrinsic molecular interactions and enzymatic reaction mechanisms of each enzyme involved in the systems biology network [4]. The generation of quantitative structure–function relationships which relate the enzyme's activity to molecular interactions between the substrate molecules and critical components of the enzyme represents one of the challenges of modern enzymology [5].

The SYCAMORE (**SY**stems biology’s **C**omputational **A**nalysis and **MO**deling **R**esearch **E**nvironment) is being developed as part of the German systems biology initiative “HepatoSys” [6] (Platform Bioinformatics and Modelling, Groups of Dr Ursula Kummer and Dr Rebecca Wade, EML Research) and aims at providing guidance to the user in setting up a biochemical kinetic model, running and analysing the results (see legend of Fig. 1 for details). When kinetic parameters are absent or inconsistent, structure-based modelling of the missing kinetic parameters is started.

PIPSA (Protein Interaction Property Similarity Analysis) is used as a means of comparing the molecular interaction fields of a test set of proteins and relating differences in enzymatic rate constants to variations in the electrostatic potentials exerted by the protein. The PIPSA methodology has been used previously to cluster different proteins according to the similarity of their electrostatic potentials. Applications include PH domains [7], E2 domains [8], triose phosphate isomerases [9], and Cu,Zn-superoxide dismutases [10]. We have extended the use of PIPSA to a more quantitative approach (qPIPSA) to relate the variations of the protein electrostatic potential within a family of enzymes to kinetic parameters.

The aim of this paper is to present an example of the application of the structure-based modelling module of the SYCAMORE project. We demonstrate the retrieval of heterogeneous protein structural and sequence information from distributed sources. The information on a protein from related organisms is then used to estimate the kinetic parameters for a corresponding protein from a different organism using the PIPSA methodology. This approach enables the user to detect inconsistent experimental values of kinetic parameters.

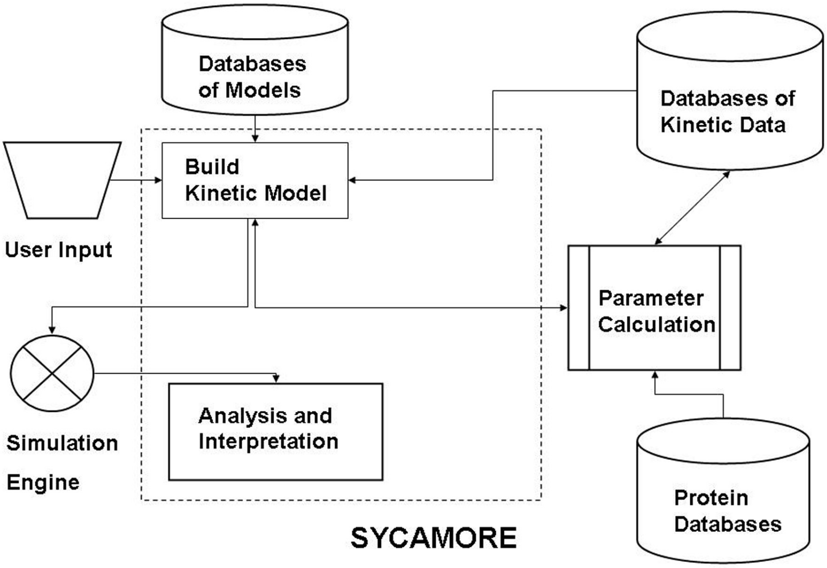


Figure 1. The SYCAMORE (SYstems biology's Computational Analysis and MOdeling Research Environment) assists the user in setting up and performing simulations in systems biology. The user can create a mathematical model by hand or use models from a depository such as Biomodels [43] or JWS online [44]. During the setup of the model, experimental kinetic parameters can be retrieved from BRENDA [14] or SABIO-RK[15]. When experimental parameters are not available for the desired organism but for a related organism or obtained under different environmental conditions, the modelling of these parameters from protein sequence and structural information can be initiated. The generated data then flow back into the kinetic model before the complete model is given to an external simulation engine (such as COPASI [45]). The final step is the analysis and interpretation of the results of the network modelling.

As a test case, we apply the method to the discrimination between mammalian and non-mammalian glucokinases and in particular to the assignment of a K_m value to the enzyme from *Escherichia coli*. The biochemistry and evolution of glucokinases has been reviewed in [11 – 13].

Methods

Retrieval of enzymatic structural and kinetic information

The structure-based modelling module within SYCAMORE is a link between the databases of experimental kinetic data, protein sequence and structure databases and the mathematical kinetic model (see Fig. 1). It is coded in Java as a server–client architecture and browser-based to allow for maximum portability and ease of accessibility.

This module is still under development. Currently the user can query the BRENDA [14] and SABIO-RK [15] databases for existing experimental kinetic parameters. Protein structural models can be retrieved from the Protein Data Bank (PDB) [16], theoretical models from ModBase [17] and from the Swiss-Model Repository [18]. Protein sequences are taken from the Swiss-Prot/UniProt database [19].

The module uses servlets and core classes. The results pages are generated using Java Server Pages (JSP) which allow static HTML to be mixed with dynamically-generated HTML pages so that the generated web pages have a dynamic content. The result pages display in any web browser compliant with XHTML and ECMAscript (Javascript). The Systems Biology Standard Markup Language (SMBL) [20] was chosen as the file format standard to communicate between the various applications and modules.

The user has the opportunity to choose retrieved sequence, structural and kinetic data from the various sources and in the end to review his choice, modify parameters or insert user-generated alternative values.

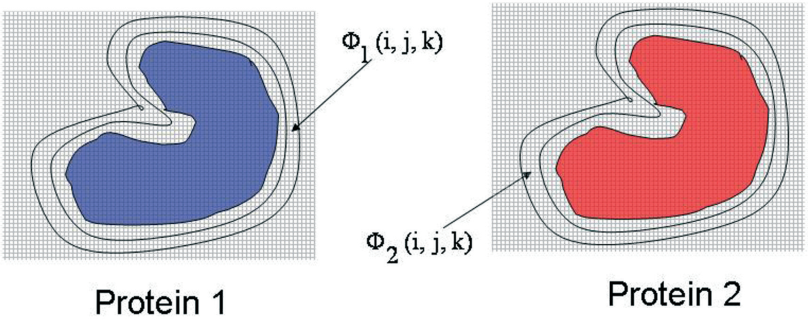
Protein interaction property similarity analysis

The structure-based systems biology calculations are performed by comparing molecular interaction fields such as the electrostatic potential or a hydrophobic field. The PIPSA method has been described elsewhere [7, 21].

The molecular interaction fields of proteins are compared on a three-dimensional grid over the superimposed proteins. The difference in the molecular interaction fields can be quantified by the calculation of similarity indices which were originally developed for the comparison of small molecules. The Hodgkin similarity index detects differences in sign, magnitude and spatial behaviour in the potential [22, 23].

Generation of protein models

Protein amino acid sequences were taken from the Swiss-Prot database [19]. Multiple sequence alignment of amino acid sequences was performed using the program ClustalW [24]. Comparative protein structural modelling was done using Modeller 8v1 [25]. Polar hydrogens were added using the program WHATIF [26]. The OPLS non-bonded parameter set was used to assign partial atomic charges and radii. The electrostatic potentials were calculated with the program UHBD [27]. The linearized form of the Poisson–Boltzmann equation (LPBE) was solved using the Choleski preconditioned conjugate gradient method. An ionic strength of 50 mM, a grid dimension of $150 \times 150 \times 150$ Å^3 and a grid spacing of 1.0 Å was employed. The relative dielectric constant of the solvent was 78.0 and that of the solute was set to 4.0.



$$(\mathbf{p}_1, \mathbf{p}_2) = \sum_{i,j,k} \Phi_1(\mathbf{i}, \mathbf{j}, \mathbf{k}) \Phi_2(\mathbf{i}, \mathbf{j}, \mathbf{k})$$

Hodgkin Similarity Index

$$\mathrm{SI}_{12} = \frac{2(\mathbf{p}_1, \mathbf{p}_2)}{(\mathbf{p}_1, \mathbf{p}_1) + (\mathbf{p}_2, \mathbf{p}_2)}$$

Figure 2. Calculation of molecular fields Φ_1 and Φ_2 on three dimensional cubic grids for two proteins and definition of the scalar product of the molecular interaction fields by summing over every grid point on a skin**.** The Hodgkin similarity index [22,23,46] is a measure of the pair-wise similarity of the molecular fields**.**

Results and Discussion

Here we give an illustrative example of the application of structure-based systems biology for the detection of inconsistent kinetic parameters and the generation of missing parameters for use in mathematical modelling of biochemical protein networks.

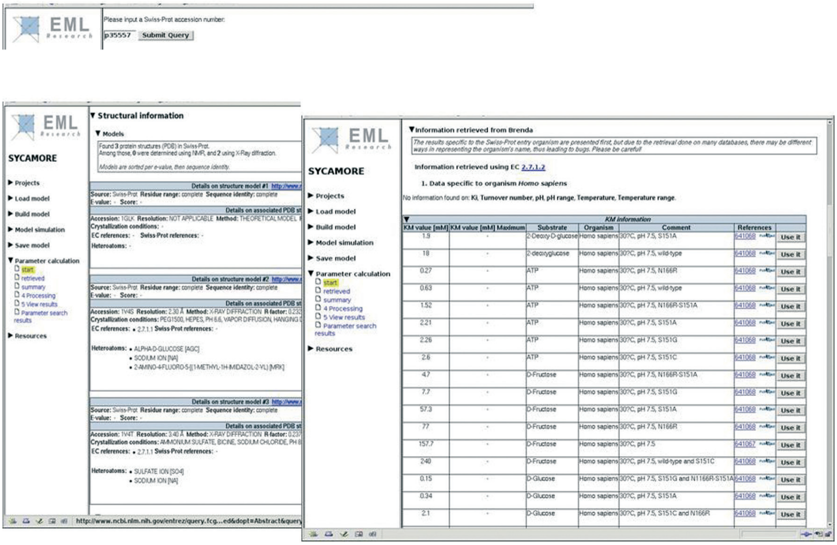
The conversion of chemical energy in the glycolytic (Emden–Meyerhof) pathway is one of the best investigated and understood metabolic pathways. The glucokinases (EC 2.7.1.2) catalyse the first chemical reaction in glycolysis. They phosphorylate glucose at the 6 position by abstracting a phosphate group from ATP. This yields glucose-6-phosphate and ADP. The virtually irreversible reaction is one of the control sites in glycolysis since the mammalian glucokinase is not product inhibited.

$$\text{Glucose} + \text{ATP} \rightarrow \text{Glucose-6-phosphate} + \text{ADP} \qquad (1)$$

We create here the scenario of a user wanting to model the glucokinase from *E. coli* by starting from knowledge about the enzyme in *Homo sapiens*.

Retrieval of protein information from distributed resources

In the structure-based estimation of kinetic parameters, the user is faced with the distribution of necessary data over various resources. The protein information retrieval module within SYCAMORE simplifies the accession to distributed protein sequence, structural and kinetic information.



Please input a Swiss-Prot accession number:
p35557

▼ Structural information

▼ Models

Found 3 protein structures (PDB) in Swiss-Prot.
Among those, 0 were determined using NMR, and 2 using X-Ray diffraction.

Models are sorted per e-value, then sequence identity.

Details on structure model #1 http://www.n
Source: Swiss-Prot **Residue range:** complete **Sequence identity:** complete
E-value: - **Score:** -
Details on associated PDB st
Accession: 1GLK **Resolution:** NOT APPLICABLE **Method:** THEORETICAL MODEL
Crystallization conditions: -
EC references: - **Swiss-Prot references:** -
Heteroatoms: -

Details on structure model #2 http://www.n
Source: Swiss-Prot **Residue range:** complete **Sequence identity:** complete
E-value: - **Score:** -
Details on associated PDB st
Accession: 1V4S **Resolution:** 2.30 Å **Method:** X-RAY DIFFRACTION **R-factor:** 0.232
Crystallization conditions: PEG1500, HEPES, PH 6.6, VAPOR DIFFUSION, HANGING D
EC references: • 2.7.1.1 **Swiss-Prot references:** -
Heteroatoms:
- ALPHA-D-GLUCOSE [AGC]
- SODIUM ION [NA]
- 2-AMINO-4-FLUORO-5-[(1-METHYL-1H-IMIDAZOL-2-YL) [MRK]

Details on structure model #3 http://www.n
Source: Swiss-Prot **Residue range:** complete **Sequence identity:** complete
E-value: - **Score:** -
Details on associated PDB st
Accession: 1V4T **Resolution:** 3.40 Å **Method:** X-RAY DIFFRACTION **R-factor:** 0.237
Crystallization conditions: AMMONIUM SULFATE, BICINE, SODIUM CHLORIDE, PH 8
EC references: • 2.7.1.1 **Swiss-Prot references:** -
Heteroatoms:
- SULFATE ION [SO4]
- SODIUM ION [NA]

http://www.ncbi.nlm.nih.gov/entrez/query.fcg...ed&dopt=Abstract&query

▼Information retrieved from Brenda

The results specific to the Swiss-Prot entry organism are presented first, but due to the retrieval done on many databases, there may be different ways in representing the organism's name, thus leading to bugs. Please be careful!

Information retrieved using EC 2.7.1.2

1. Data specific to organism *Homo sapiens*

No information found on: **Ki, Turnover number, pH, pH range, Temperature, Temperature range.**

▼ *KM information*

KM value [mM]	KM value [mM] Maximum	Substrate	Organism	Comment	References	
1.9	-	2-Deoxy-D-glucose	Homo sapiens	30?C, pH 7.5, S151A	641068	Use it
18	-	2-deoxyglucose	Homo sapiens	30?C, pH 7.5, wild-type	641068	Use it
0.27	-	ATP	Homo sapiens	30?C, pH 7.5, N166R	641068	Use it
0.63	-	ATP	Homo sapiens	30?C, pH 7.5, wild-type	641068	Use it
1.52	-	ATP	Homo sapiens	30?C, pH 7.5, N166R-S151A	641068	Use it
2.21	-	ATP	Homo sapiens	30?C, pH 7.5, S151A	641068	Use it
2.26	-	ATP	Homo sapiens	30?C, pH 7.5, S151G	641068	Use it
2.6	-	ATP	Homo sapiens	30?C, pH 7.5, S151C	641068	Use it
4.7	-	D-Fructose	Homo sapiens	30?C, pH 7.5, N166R-S151A	641068	Use it
7.7	-	D-Fructose	Homo sapiens	30?C, pH 7.5, S151G	641068	Use it
57.3	-	D-Fructose	Homo sapiens	30?C, pH 7.5, S151A	641068	Use it
77	-	D-Fructose	Homo sapiens	30?C, pH 7.5, N166R	641068	Use it
157.7	-	D-Fructose	Homo sapiens	30?C, pH 7.5	641067	Use it
240	-	D-Fructose	Homo sapiens	30?C, pH 7.5, wild-type and S151C	641068	Use it
0.15	-	D-Glucose	Homo sapiens	30?C, pH 7.5, S151G and N1166R-S151A	641068	Use it
0.34	-	D-Glucose	Homo sapiens	30?C, pH 7.5, S151A	641068	Use it
2.1	-	D-Glucose	Homo sapiens	30?C, pH 7.5, S151C and N166R	641068	Use it

Figure 3. Snapshot of protein information retrieval module within SYCAMORE**.** It retrieves heterogeneous protein information such as protein structure, existing experimental kinetic data and sequence information (see text for details).

Figure 3 shows a snapshot of the protein information retrieval module within SYCAMORE. When querying for the glucokinase from *Homo sapiens* (Swiss-Prot ID P35557) in Swiss-Prot, three related protein structures are found: these are the X-ray crystal structures of the enzyme from *Homo sapiens* in its closed form (PDB entry 1V4S) and its open form (PDB entry 1V4T) [28] plus a theoretical model for the human glucokinase (PDB code 1GLK) based on its homology to the enzyme from yeast. The user may select one of the three models for subsequent structural modelling.

Below, relevant additional structural information for kinetic modelling from the IntAct [29] database at EBI are given, such as the interaction of human glucokinase with the glucokinase regulatory protein (GCKR) and the 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase I.

The next screen displays relevant kinetic information that was found in BRENDA [14] when searching for enzymes with the same EC number. First data for the glucokinase from *Homo sapiens* such as K_{m} values and specific activities for a range of substrates and the influence of single point mutations on K_{m} are reported. Then data specific to other organisms are also reported.

The user may choose any of the reported parameters for subsequent mathematical modelling by clicking on the "use it" button. The user then has the option to review his choice of parameters, correct or modify them or insert his own parameters manually for the mathematical modelling of the enzyme glucokinase.

PIPSA of the electrostatic potential of glucokinases

Here we present an illustrative case of the structure-based generation of kinetic parameters from a PIPSA of the electrostatic potential of glucokinases. We analyse the similarity of the electrostatic potentials of a test set of 8 different glucokinases for which experimental K_{m} constants for the substrate glucose could be found in the BRENDA database. We set our focus on the glucokinase from *E. coli* and demonstrate a procedure to assist the user in the choice of an appropriate K_{m} value when constructing a kinetic model.

Kinetic constants and comparative protein structural modelling

For the glucokinases from *Homo sapiens*, *Rattus norvegicus*, *Escherichia coli*, *Aspergillus niger*, *Hansenula polymorpha*, *Saccharomyces cerevisiae*, *Streptococcus mutans* and *Zymomonas mobilis* K_{m} values for the substrate glucose could be found in the BRENDA database. They all catalyse an identical chemical reaction. However, they do so with very different substrate binding affinity, represented by the K_{m} value.

The experimental values found in BRENDA are 0.028 mM (*S. cerevisiae*) [30], 0.05 mM (*H. polymorpha*) [30], 0.063 mM (*Asp. niger*) [30], 0.095 mM (*Z. mobilis*) [31], 0.61 mM (*S. mutans*) [32] to 6 mM (*H. sapiens*) [33] and 7.7 mM *(R. norvegicus*) [34] and thus cover a range of more than 2 orders of magnitude.

For the glucokinase from *E. coli*, the available experimental K_{m} values range from 0.78 mM [35] to 0.15 mM [36]. Since no experimental error bars are given, we would like to check the completeness and consistency of these values. The user has to make a choice when setting up a kinetic model of glycolysis in *E. coli*. We apply the PIPSA method to compare the electrostatic potentials around the active site and correlate with experimental K_{m} values from other organisms to suggest a value for *E. coli*.

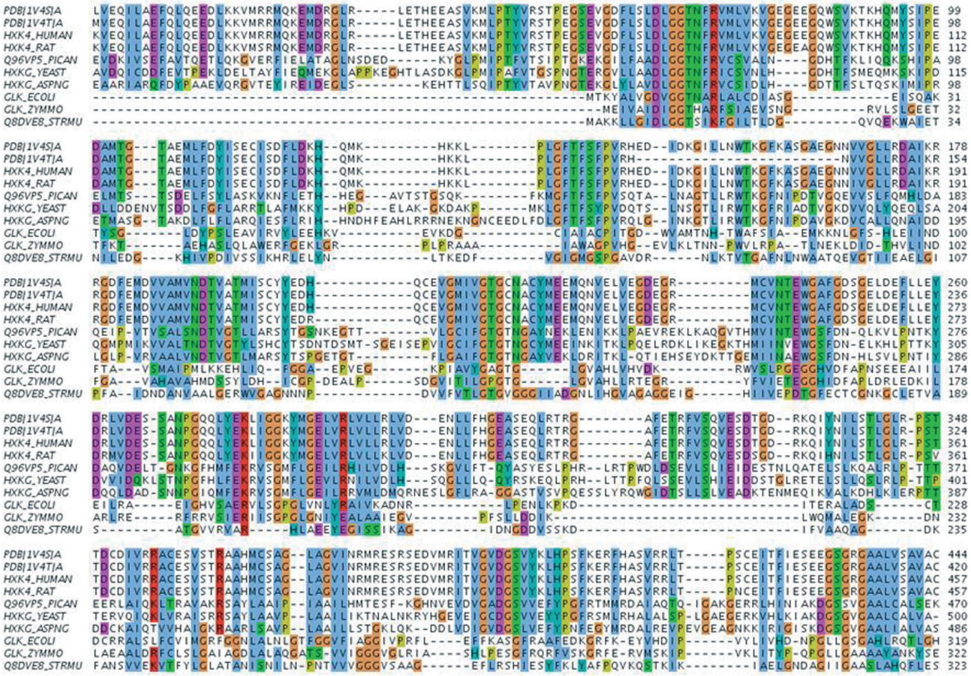


Figure 4. Multiple sequence alignment of glucokinases from *Homo sapiens, Rattus norvegicus, Hansenula polymorpha*, *Saccharomyces cerevisiae*, *Aspergillus niger*, *Escherichia coli*, *Zymmomonas mobilis* and *Streptococcus mutans*. The amino acid sequences of the template structures of the open (PDB code 1V4T) and closed (PDB code 1V4S) [28] forms of the human glucokinases are also given.

Figure 4 shows the ClustalW multiple sequence alignment of glucokinases with the sequences from *Homo sapiens* of the closed (1V4S) and open forms (1V4T) of the enzyme. The multiple sequence alignment was used to generate protein structural models by mapping the target sequences from *Homo sapiens*, *Rattus norvegicus*, *Escherichia coli*, *Aspergillus niger*, *Hansenula polymorpha*, *Saccharomyces cerevisiae*, *Streptococcus mutans* and *Zymomonas mobilis* to the template protein structure of the open (PDB code 1V4T) and closed forms (PDB code 1V4S) of human glucokinase. For each of the generated protein models, the electrostatic potential was calculated.

Calculation and comparison of the electrostatic potentials for glucokinases

The mammalian glucokinase undergoes a large conformational change upon substrate binding [28]. Two of the three layers of the small domain of glucokinase rotate at an angle of 99 ° around a hinge region [28]. The substrate glucose binds to the bottom of the deep cleft between the large domain and the small domain. In the closed form, glucose is coordinated by residues from the large and the small domains.

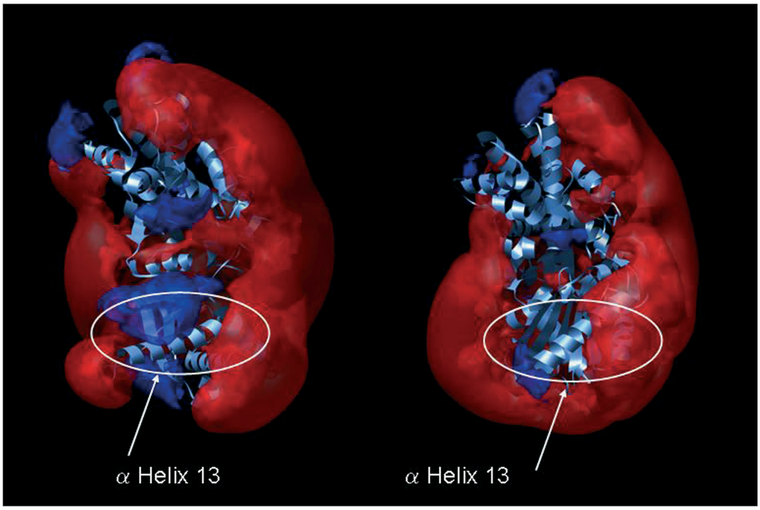


Figure 5. Calculated electrostatic isopotential isosurface at (0.6 kcal mol^{-1} e^{-1} of the open (left) and closed (right) form of the Hexokinase IV from *Homo sapiens* [28].

Figure 5 shows the calculated electrostatic potential for the open form (1V4T; left in Fig. 5) and the closed form (1V4S, right in Fig. 5). The two forms differ in electrostatic potential in particular around the α13 helix which moves in a different direction to the small domain upon conformational change [28].

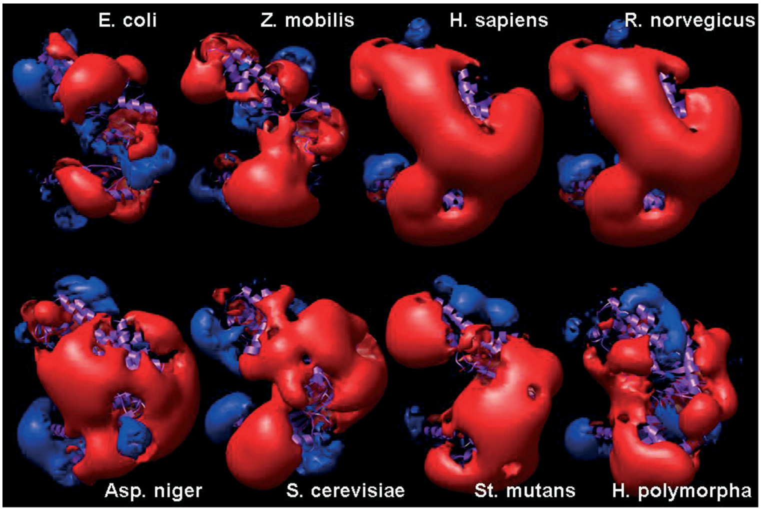


Figure 6. Calculated electrostatic potentials of glucokinases from eight organisms for which substrate K_m values were found in the BRENDA database**.** The isosurfaces are shown at 0.6 kcal mol$^{\mathbf{-1}}$ **e**$^{\mathbf{-1}}$ for the open form of the enzyme**.**

Figure 6 shows the computed electrostatic potential of the glucokinases in the other organisms. All have a large negative patch near the ATP binding region (right side) and a more positive patch on the left. Visual inspection shows that the electrostatic potential of the glucokinases from *Homo sapiens* and *Rattus norvegicus* appear indistinguishable. There is, however, a large variation in the distribution of the electrostatic potential across the organisms.

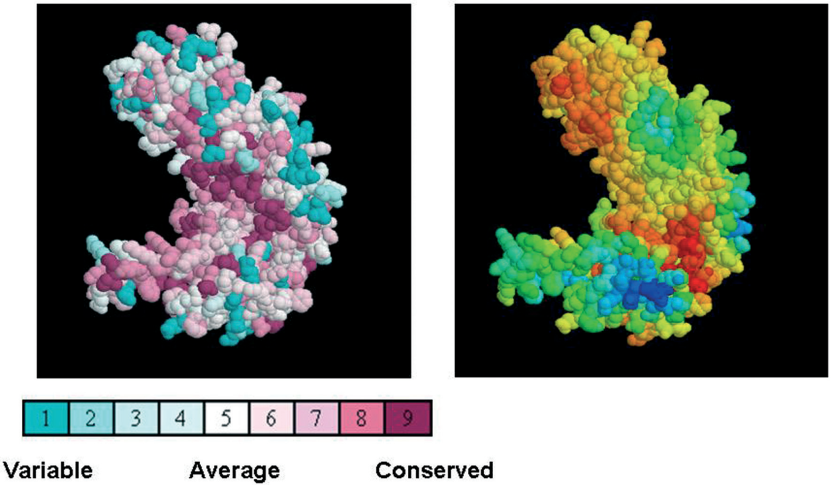


Figure 7:

Left: Conservation of the amino acid residues in the multiple sequence alignment displayed on the open form of human hexokinase IV (1V4T) using the Consurf algorithm (47).

Right: Conservation of the calculated electrostatic potential. Pairwise comparison of the calculated electrostatic potentials.

Figure 7 shows the conservation of the positions of amino acid residues of the eight glucokinases mapped onto the crystal structure of the human enzyme in its open form (left). The most conserved amino acid residues are found in the cleft between the large and small domains: this is the site where the ligand co-crystallizes in the closed form; and a patch of conserved amino acid residues in proximity to the ligand binding site, potentially the entry channel of the substrate. Figure 7 (right) shows the conservation of the electrostatic potential. The most conserved patches of the electrostatic potential of the set of glucokinases, ranging from blue (no conservation), yellow (intermediate) to patches of high conservation (coloured in red). The most conserved electrostatic region approximately overlaps with the region of most conserved amino acid sequences between the two protein domain and may refer to the entry channel of the substrate. The electrostatic potential near the ligand binding site, however, is not strictly conserved. The variations in the electrostatic potentials at this spot may explain the large range of K_m values between mammalian and non-mammalian enzymes.

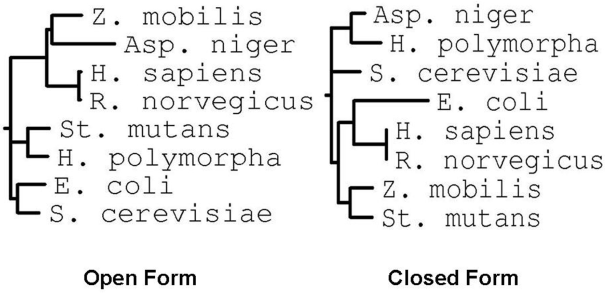


Figure 8: Tree diagram of the similarities of the electrostatic potentials of glucokinases in a region of radius 15 Å around the ligand binding site.

A more quantitative comparison of the electrostatic potentials is possible with the Hodgkin similarity indices. The pairwise similarities can be easily visualized in phylogenetic trees [8]. Figure 8 displays tree diagrams of the similarities of the electrostatic potentials in the test set of eight glucokinases of the open (left) and closed (right) forms. We used a radius of 15 Å around the ligand binding site for the comparison of the electrostatic potentials since the conservation of the active site was also observed in a phylogenetic analysis of the primary sequences of hexokinases [11].

For the closed form, the nearest neighbours of the glucokinase from *E. coli* are the mammalian glucokinases from *Homo sapiens* and *Rattus norvegicus*. This would suggest a K_m value of the *E. coli* glucokinase in the mM range. This assignment seems improbable since the sequence identity is very low between glucokinases from *E. coli* and *Homo sapiens* (14% overall sequence identity).

The mammalian glucokinases in liver (hexokinases IV) possess a high K_m value (6 – 7 mM) and act as a sensor of high glucose levels in the blood since the physiological role of glucokinases in vertebrates is significantly different from that of invertebrates. In mammalians, the glucokinase (hexokinase IV) is the liver-specific isozyme with a glucose sensor function in hepatocytes [11] and represents 95% of the total hexokinase activity of hexokinases. The liver enzymes phosphorylate glucose only when it has reached a high concentration in the blood. Thus, isozymes in brain and muscle, which have 50-fold lower K_m values, are activated first. Only when glucose is abundant, is the liver isozyme active and ensures that glucose is not wasted.

When the electrostatic potentials are computed for protein structural models of the open form (Fig. 8, left), the closest glucokinase to *E. coli* is from *S. cerevisiae* and suggests a K_m value around 0.03 mM for *E. coli*. This predicted K_m value is clearly outside the range of K_m values retrieved from BRENDA: 0.15 mM [35] to 0.78 mM [36]. This discrepancy was analysed further. The glucokinase from *E. coli* displays only weak similarity to the other glucokinases. This had been noticed already by Cardenas *et al.* [37]. The absence of homology with other hexokinases suggested an early divergent evolution of hexokinases in plants, vertebrates, yeast and bacterial hexokinases. The current investigation suggests that the glucokinase from *E. coli* is a very specific hexokinase with a predicted very low K_m value of the same order of magnitude as yeast.

The recently solved X-ray structure of the ATP-dependent glucokinase from *E. coli* displayed a RNase H-like fold [38] which is also found for *Homo sapiens* [28] and yeast [39] glucokinases and justifies *a posteriori* the use of the template protein structure from *Homo sapiens* despite the low sequence identity.

When searching for additional investigations of the kinetics of the glucokinase from *E. coli* that are not yet included in BRENDA, we found a recent report by Millar and Raines of a K_m value of the glucokinase from *E. coli* of 0.076 mM [40]. This is significantly lower than the K_m values reported previously ranging from 0.15 mM to 0.78 mM.

This supports our assignment of the glucokinase from *E. coli* to the family of very specific bacterial glucokinases with a very low K_m value: 0.028 mM (*S. cerevisiae*) and 0.063 mM (*Asp. niger*).

In general, we found a better correlation of the kinetic parameters for the open form of the enzyme. This was also noticed by Xu *et al*. who correlated calculated interaction energies of various sugars with measured k_{cat}/K_m values [41]. They came to the conclusion that the substrate sugar molecules are recognized by binding to the open form of glucokinase.

PIPSA of a large set of glucokinases

The previous application of the PIPSA classification of glucokinases was limited to a small set of eight experimentally characterized organisms. In systems biology one aims at an understanding of enzymes in context and also across a larger number of organisms.

The investigation of the similarity of the electrostatic potentials of glucokinases was extended to a larger set of proteins. All protein sequences that were annotated as either glucokinases or classified with the EC number 2.7.1.2 were aligned according to their amino acid sequence identity. Sequences which were annotated as polyphosphate glucokinases, ROK (repressor, open reading frame, and kinase) or for which only fragments were available, were removed. This led to a set of 164 aligned protein sequences. Protein structural models were generated based on the template structure of the human hexokinase IV (HXK4_HUMAN) in its open form. Electrostatic potentials were calculated by solving the linearized Poisson–Boltzmann equation (as described above in detail).

The 164 proteins were classified according to their Hodgkin similarity indices of the electrostatic potential in a region of 15 Å radius around the ligand binding site (see Fig. 9). The inserts show magnifications of selected glucokinases from *E. coli*, Yeast and *Homo sapiens*.

The nearest neighbours to *E. coli* are the glucokinases from *E. coli O6*, *Shigella flexnen*, *Salmonella typhi* and *Salmonella typhimarium*. The enzyme from yeast is closest to various glucokinases from *Xylella fastidiosa* and *Yersinia pestis*. From PIPSA of the electrostatic potentials, one may expect glucokinases from *Sparus aurata* (Gilthead sea bream), *Cyprinus carpio* (Common carp), hexokinase IV from *mouse*, *Oncorhynchus mykiss* (Rainbow trout) to exhibit similar kinetic parameters to the enzymes from *Homo sapiens* and *R. norvegicus*. Also the glucokinase EMI2_Yeast (Early Meiotic Induction Protein 2 [42] is predicted to possess similar kinetic parameters. This glucokinase is involved in sporulation and is required for the full activation of the early meiotic inducer EMI1 [41]. This glucokinase performs a different physiological role from bacterial glucokinases and thus a high K_m value may be expected.

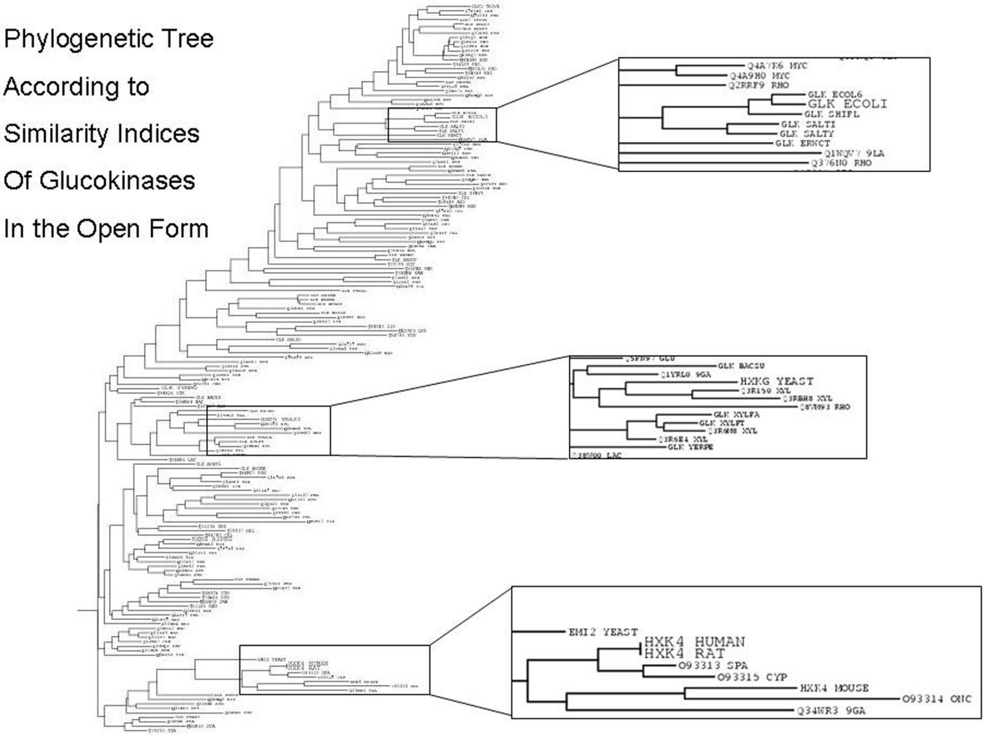


Figure 9: Tree diagram of 164 glucokinases EC 2.1.7.2 classified by their similarity in electrostatic potential of the open form in a region of radius 15 Å around the ligand binding site.

Conclusion and Outlook

Structure-based systems biology provides detailed insight into cellular processes at a molecular level. It is thus complementary to the abstract mathematical modelling of protein signaling or metabolic networks. The PIPSA method provides a quantitative structure to function relationship for enzymes. It quantifies the similarity of molecular interactions between the substrate molecule and the protein active site for the same enzyme from a large number of organisms. The large-scale application of PIPSA allows the classification of enzymes previously uncharacterized and the detection of relationships with other enzymes.

Furthermore, the PIPSA method can be used to detect outliers from a series of well-characterized enzymes. For this use it is critical to have:

i) an extensive annotation of experimental conditions
ii) a detailed and consistent set of experimental data.

Further application and extension of the qPIPSA method to predicting enzymatic K_{m} and k_{cat}/K_{m} values and the comparative modelling of the glycolytic pathway across multiple organisms is in progress.

Acknowledgements

We thank the Klaus-Tschira Foundation and the BMBF Systems Biology Initiative HepatoSys for financial support. We are grateful to Dr A. Weidemann for the design and layout of the SYCAMORE user interface and Dr U. Kummer and Dr. I. Rojas for the scientific collaboration.

References

[1] Kitano, H. (2002) *Nature* 420, 206–210.

[2] Gabdoulline, R.R., Kummer, U., Olsen, L.F. & Wade, R.C. (2003) *Biophys. J.* 85, 1421–1428.

[3] Apweiler, R., Cornish-Bowden, A., Hofmeyr, J.-H.S., Kettner, C., Leyh, T.S., Schomburg, D. & Tipton, K. (2005) *Trends in Biochemical Sciences* 30, 11–12.

[4] Aloy, P. & Russell, R.B. (2005) *FEBS Letters* 579, 1854–1858.

[5] Kettner, C. & Hicks, M.G. (2005) *Curr. Enz. Inhib.* 1, 171–181.

[6] http://www.systembiologie.de.

[7] Blomberg, N., Gabdoulline, R.R., Nilges, M. & Wade, R.C. (1999) *Proteins* 37, 379–387.

[8] Winn, P.J., Religa, T.L., Battey, J.N., Banerjee, A. & Wade, R.C. (2004) *Structure* 12, 1563–1574.

[9] Wade, R.C., Gabdoulline, R.R. & Luty, B. (1998) *Proteins* 31, 406–416.

[10] Wade, R.C., Gabdoulline, R.R., Luedemann, S. & Lounnas, V. (1998) *Proc. Natl. Acad. Sci. USA* 95, 5942–5949.

[11] Cardenas, M.L., Cornish-Bowden, A. & Ureta, T. (1998) *Biochim. Biophys. Acta* 1401, 242–264.

[12] Cardenas, M.L. (1997) *Biochemical Society Transactions* 25, 131–135.

[13] Cardenas, M.L. (2003) in *Glucokinase and Glycemic Diseases*, ed. Magnuson, M.A. (Karger, Basel).

[14] Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G. & Schomburg, D. (2004) *Nucleic Acids Res.* 32, D431–3.

[15] Rojas, I., Kania, R., Wittig, U., Weidemann, A., Goblewski, M. & Krebs, O. (2005) in *Proceedings of the 4th Workshop on Computation of Biochemical Pathways and Genetic Networks*, eds. Kummer, U., Pahle, J., Surovtsova, I. & Zobeley, J. (Logos Verlag, Berlin), pp. 63–67.

[16] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000) *Nucl. Acids Res.* 28, 235–242.

[17] Pieper, U., Eswar, N., Davis, F.P., Braberg, H., Madhusudhan, M.S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B.M., Eramian, D., Shen, M.Y., Kelly, L., Melo, F. & Sali, A. (2006) *Nucleic Acids Research.* 34, D291–295.

[18] Kopp, J. & Schwede, T. (2004) *Nucl. Acids Res.* 32, D230-D234.

[19] Bairoch, A., Boeckmann, B., Ferro, S. & Gasteiger, E. (2004) *Brief. Bioinform.* 5, 39–55.

[20] Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.-H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novere, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J. & Wang, J. (2003) *Bioinformatics* 19, 524–531.

[21] Wade, R.C., Gabdoulline, R.R. & Rienzo, F.D. (2001) *Intl. J. Quant. Chem.* 83, 122–127.

[22] Good, A.C., Hodgkin, E.E. & Richards, W.G. (1992) *Journal of Computer-Aided Molecular Design* 6, 513–520.

[23] C. Burt, W.G. Richards & Huxley, P. (1990) *Journal of Computational Chemistry* 11, 1139–1146.

[24] Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994) *Nucl. Acids Res.* 22, 4673–4680.

[25] Sali, A. & Blundell, T.L. (1993) *J. Mol. Biol.* 234, 779–815.

[26] Vriend, G. (1990) *J. Mol. Graph.* 8, 52–56.

[27] Davis, M.E., Madura, J.D., Luty, B.A. & McCammon, J.A. (1991) *Computer Physics Communications* 62, 187–197.

[28] Kamata, K., Mitsuya, M., Nishimura, T., Eiki, J. & Nagata, Y. (2004) *Structure* 12, 429–438.

[29] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. & Apweiler, R. (2004) *Nucl. Acids Res.* 32, D452–455.

[30] Laht, S., Karp, H., Kotka, P., Jarviste, A. & Alamae, T. (2002) *Gene* 296, 195–203.

[31] Scopes, R.K. & Bannon, D.R. (1995) *Biochim. Biophys. Acta* 1249, 173–9.

[32] Porter, V. & Chassy, B.M. (1982) *Methods Enzymol.* 90, 25–30.

[33] Xu, L.Z., Harrison, R.W., Weber, I.T. & Pilkis, S.J. (1995) *J. Biol. Chem.* 270, 9939–46.

[34] Tu, J. & Tuch, B.E. (1996) *Diabetes* 45, 1068–75.

[35] Meyer, D., Schneider-Fresenius, C., Horlacher, R., Peist, R. & Boos, W. (1997) *J. Bacteriol.* 179, 1298–306.

[36] Arora, K.K. & Pedersen, P.L. (1995) *Arch. Biochem. Biophys.* 319, 574–579.

[37] Cardenas, M.L. (2004) in *Glucokinase and Glycemic Diseases*, eds. Matschinsky, F.M. & Magnuson, M.A. (Karger, Basel), Vol. 16, pp. 31–41.

[38] Lunin, V.V., Li, Y., Schrag, J.D., Iannuzzi, P., Cygler, M. & Matte, A. (2004) *Journal of Bacteriology* 186, 6915–6927.

[39] Kuser, P.R., Krauchenco, S., Antunes, O.A. & Polikarpov, I. (2000) *J. Biol. Chem.* 275, 20814–20821.

[40] Miller, B.G. & Raines, R.T. (2004) *Biochemistry* 43, 6387–6392.

[41] Xu, L.Z., Weber, I.T., Harrison, R.W., Gidh-Jain, M. & Pilkis, S.J. (1995) *Biochemistry* 34, 6083–6092.

[42] Enyenihi, A.H. & Saunders, W.S. (2003) *Genetics* 163, 47–54.

[43] Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J.L. & Hucka, M. (2006) *Nucl. Acids Res.* 34, D689–691.

[44] Olivier, B.G. & Snoep, J.L. (2004) *Bioinformatics* 20, 2143–2144.

[45] http://www.copasi.org.

[46] Hodgkin, E.E. & Richards, W.G. (1987) *Int. J. Quant. Chem. Quant. Biol. Symp.*, 105–110.

[47] Glaser, F., Pupko, T., Paz, I., Bell, R.F., Bechor-Shental, D., Martz, E. & Ben-Tal, N. (2003) *Bioinformatics* 19, 163–164.

Kinetic Characterization of Alcohol Dehydrogenases and Matrix Metalloproteinases: A Reflection on Standardization of Assay Conditions

Jan-Olof Winberg

Department of Medical Biochemistry, Institute of Medical Biology,
Faculty of Medicine, University of Tromsø, 9037 Tromsø, Norway

E-Mail: janow@fagmed.uit.no

Received: 3rd July 2006 / Published: 31st August 2007

Abstract

The present paper will focus on the characterization of enzymes from two different types of family, Short Chain Dehydrogenases/Reductases and Matrixins. The former family includes over 3000 enzymes, and I have worked mainly with different allelic variants of alcohol dehydrogenase (ADH) from the fruit fly *Drosophila melanogaster* and the ADH in *Drosophila lebanonensis*. To date, approximately 25 matrix metalloproteinases are known in humans. I will focus here on both similarities and differences in problems regarding the standardization of assay conditions and parameters that I have experienced during my work with these two different enzyme systems.

Introduction

The biochemical characterization of enzymes requires careful and well planned experimental set-ups. Among parameters that need to be considered are the type of buffer to be used, what pH value is relevant to use, the ionic strength of the assay, are additives necessary, relevant temperature and what type of assay can be used in kinetic characterizations. Enzymes vary in their *in vivo* localization, their interactions with other proteins and cellular components that may affect their stability as well as their biological activity. By purification an enzyme is removed from its environment, which results in that some enzymes need additives to compensate for the loss of interaction partners. This of course creates a problem with respect to standardization of enzyme assay conditions, which were nicely described by Tipton and co-workers [1] in the 2003 meeting on Experimental Standard Conditions of Enzyme Characterizations. In the present paper, I will focus therefore on two problems that frequently occur in the literature with respect to standardization. The first problem concerns the determination of enzyme concentration used to calculate kinetic coefficients. The second problem concerns the use of additives that have an effect on the biochemical parameter studied and to what extent the description of experimental conditions is sufficient to reproduce reported results. I will elucidate these problems mainly from my own work with two different enzyme systems, alcohol dehydrogenase from *drosophila* (DADH) and matrix metalloproteinases (MMPs). First, the two enzyme systems will be briefly described, and thereafter I will continue with the standardization problems.

Drosophila Alcohol Dehydrogenase

The ADH (EC 1.1.1.1) from insects is involved in the metabolism of short and medium sized primary and secondary alcohols, which is converted to their corresponding aldehydes and ketones (Equation 1), using the coenzyme NAD$^+$ [2]. The ADH is also involved in the oxidation of the formed aldehydes to their corresponding carboxylic acids (Equation 2) [3,4].

$$\text{alcohol} + \text{NAD}^+ \leftrightarrow \text{aldehyde/ketone} + \text{NADH} + \text{H}^+ \qquad (1)$$

$$\text{aldehyde} + \text{H}_2\text{O} + \text{NAD}^+ \leftrightarrow \text{carboxylate}^- + \text{NADH} + 2\,\text{H}^+ \qquad (2)$$

ADH has been found in most of the *drosophila* species investigated, and some of these species are polymorphic with respect to the Adh gene such as *D. melanogaster*, while other species such as *D. lebanonensis* are monomorphic [5].

The insect ADHs differ from the well known ADHs from other species such as vertebrates and plants in that it lacks metal ions and has a much shorter polypeptide chain [6, 7]. At the beginning of the 1980 s Jörnvall and colleagues used these differences to divide the dehydrogenases into families [7] and today, over 3000 open reading frames has been detected for the family of Short-Chain Dehydrogenases/Reductases (SDR), the family to which DADH belongs [8]. Enzymes belonging to the SDR family have been found in all species from humans to viruses [8] and they involve various enzyme classes such as oxidoreductases, lyases and isomerases. Structurally the SDR enzymes differ from the other families of dehydrogenases and reductases in that they are one domain enzymes where the N-terminal part of the polypeptide chain builds up the coenzyme binding region and the C-terminal part the catalytic region [6, 8]. It was first in 1998 that the first 3D structure of a DADH was reported, and later on followed by binary DADH–coenzyme and ternary DADH–coenzyme–substrate/product complexes [9–11]. Many studies on DADH have been performed in order to understand the evolution and the metabolic function of this enzyme [5, 12]. DADHs have also been characterized with respect to substrate specificity, coenzymes and substrate stereospecificity, inhibitory kinetics, reaction mechanism, pH and temperature dependence, and interconversion of electrophoretic variants [2, 13–19].

Matrix Metalloproteinases

Matrix metalloproteinases (MMPs) is the name of a group of enzymes either secreted into the extracellular matrix (ECM) or bound to the cell membrane that together are able to degrade almost all the structural ECM proteins as well as several non-ECM proteins [20]. MMPs belong to the Clan MA, subclan MAM, family M10, subfamily A (Merops database) [21]. Typical for MMPs is that they are zinc and calcium dependent. They contain two zinc ions, one catalytic and one structural. Calcium is necessary both for the stability and the activity of these enzymes [20]. Based on the substrate specificity, similarities in the primary structure and organization of the protein domains, the MMPs can be divided into six classes, matrilysins, collagenases, gelatinases, stromelysins, membrane-type MMPs and others/new MMPs [22, 23]. The general domain structure of MMPs is shown in Fig. 1 along with the structure of the different classes of MMPs. Most MMPs contain an N-terminal signal and pro-domain, a catalytic domain containing the catalytic zinc ion, a hinge domain and a C-terminal hemopexin like domain. In four of the six membrane-type MMPs (MT-MMPs), the C-terminal domain ends in a type I transmembrane domain, while two binds to the cell membrane through a glycosyl-phosphatidyl-inosityl (GPI) anchor. Two other potential MT-MMPs (MMP-23A and B, which have the same primary structure, but are coded by two different genes) contain a type II transmembrane domain (signal anchor) N-terminal to the pro-domain, and instead of a hemopexin domain they contain a unique “cysteinarray” and an immunoglobulin-like (Ig) domain. The two gelatinases (MMP-2 and MMP-9) also contain a fibronectin II-like insert in their catalytic domains, while the hinge-region of MMP-9 also contains a collagen V-like domain.

What these enzymes have in common is that most are synthesized and secreted into the extracellular tissues as inactive proenzymes that need to be activated. ProMMPs can be activated by other proteinases including active MMPs in the tissues or on the cell membrane, by chaotropic agents, organomercurials, reactive oxygen species or oxidized glutathione [24]. Due to the unique sequence (RX[K/R]R) in the end of the pro-domain of the MT-MMPs, MMP-11, -21 and -28 these enzymes can be activated intracellularily by furin, a serine proteinase that belongs to the convertase family [22–24]. The activity of MMPs is also regulated by endogeneous inhibitors such as α2-macroglobulin and the specific tissue inhibitors of MMPs (TIMPs) [22, 23, 25].

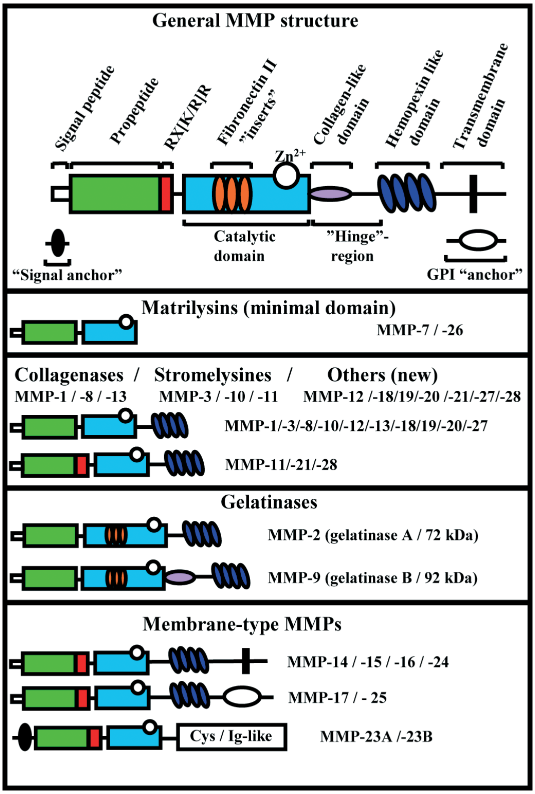


Figure 1. Schematic representation of the domain structure of MMPs. The general domain structure of MMPs is shown (top) along with the individual human MMPs that are classified according to their substrate specificity, similarities in the primary structure and organization of the protein domains.

Detection of Kinetic Coefficients Requires that the Amount of Functional Enzyme Active Sites is Determined

To get a full description of an enzyme and its ability to act on various substrates, it is necessary to determine the kinetic coefficients with substrates, coenzymes and other factors that are involved in the reaction. Equations 3 and 4 are examples of nomenclature for a two substrate reaction where *S* and *C* represent substrate and coenzyme, respectively.

$$\frac{e}{v} = \phi_0 + \frac{\phi_1}{[C]} + \frac{\phi_2}{[S]} + \frac{\phi_{12}}{[C][S]} \qquad (3)$$

$$\frac{1}{v}=\frac{1}{V_m}+\frac{K_{m1}}{V_m[C]}+\frac{K_{m2}}{V_m[S]}+\frac{K_{m2}K_{ia}}{V_m[C][S]} \tag{4}$$

$$\frac{e}{v}=\frac{1}{k_{cat}}+\frac{K_{m1}}{k_{cat}[C]}+\frac{K_{m2}}{k_{cat}[S]}+\frac{K_{m2}K_{ia}}{k_{cat}[C][S]} \tag{5}$$

Independent of nomenclature, to obtain a full description of the kinetic coefficients the concentration of functional enzyme active sites is required. As can be seen from the above described examples of nomenclature, in Equation 3 [26] the enzyme concentration is incorporated in the rate equation while this is not the case in equation 4 [27]. In the latter case it is necessary to convert V_m to k_{cat}, i.e. the coefficient for the catalytic centre activity of the enzyme. As $k_{cat} = V_m/[e]$, Equation 4 can be rewritten to Equation 5. With knowledge of k_{cat} it is possible to get a description of the enzymes capability to act on a substrate, and also to compare the activity with other similar enzymes.

A large problem is to find a good and reliable method to determine the amount of functional enzyme active sites in order to calculate k_{cat} ($1/N_0$). In the literature, it can often be seen that the amount of enzyme used in the calculations is not based on a reliable method that determines the concentration of functional active sites. Instead, the amount of protein is determined by a protein detection method such as Bradford, or A_{280nm} and a well defined extinction coefficient for the enzyme in question. Even if the enzyme preparation can be regarded to be homogeneous based on SDS-PAGE and isoelectric focusing, none of these methods are acceptable to determine the amount of functional enzyme. The reason is that these methods are based on the assumption that the protein concentration is identical with the concentration of functional active sites in the enzyme, which is not always the case. Therefore, a reliable value for k_{cat} ($1/N_0$) can be obtained only if the amount of functional enzyme is determined by a method that is based on active-site titration. How the titration is performed depends on the enzyme, and several methods have been described [28]. A good example of active site titration of ADHs was first shown by Theorell [29], which was based on the formation of a dead end ternary complex using the alcohol competitive inhibitor pyrazole. This method has been used in several studies of ADHs [30–32]. Unfortunately not all ADHs form a strong ternary complex with pyrazole, which is a necessity for its use as a titrating agent [33]. Under such conditions, it is necessary to find alternative methods. The method of Theorell [29] has been used on sorbitol dehydrogenase (SDH) from sheep liver, where DTT (a substrate competitive inhibitor) was used instead of pyrazole [34]. Several titration methods have been used on proteinases, including the classical titration method of chymotrypsin [28]. With MMPs, various methods have been used. All are based on a strong interaction between a synthetic inhibitor or one of the TIMPs and the enzyme active site [35–37]. Here of course it is important not to use a TIMP that is known to bind to a proMMPs C-terminal hemopexin-like region such as TIMP-2 to proMMP-2 and TIMP-1 to proMMP-9 [25].

The question is whether it is correct or not to report a kinetic coefficient such as k_{cat} and k_{cat}/K_{m} for a substrate when a homogeneous enzyme preparations has been used and where it is not possible to obtain the amount of functional enzyme by an active site titration method. Personally I think this is wrong, even if it is a good reason to assume that the concentration of functional active-sites is identical with the amount of enzyme detected with for example $A_{280\mathrm{nm}}$ and a well defined extinction coefficient for the enzyme in question. In such cases it would have been much better to introduce new coefficients that for example could be denoted $k_{cat(\text{-t})}$ and $k_{cat(\text{-t})}/K_{\mathrm{m}}$ where (-t) shows that the coefficient is not based on active site titration.

Detection of the Substrate Specificity of an Enzyme with or without Knowledge of The Absolute Concentration of the Functional Active Sites

Under some conditions it is not possible to determine the concentration of functional enzyme active sites and hence, the absolute value of the kinetic coefficients. This will of course limit our ability to compare the absolute activity of enzymes, but it is still possible to obtain various kinetic characteristics such as the substrate specificity for an enzyme and compare this with the substrate specificity of another enzyme, detection of inhibitory compounds and reaction mechanism. A typical example is our early studies of DADHs [38, 39]. We intended to determine the topology of the enzyme active site long before a 3D-structure of DADH was available. As the topology of the active-site determines the substrate specificity of the enzyme, we decided to investigate the substrate specificity of the enzyme by using approximately 100 different structurally well defined alcohols (primary, secondary, linear, cyclic and bi-cyclic). We faced several problems during these early studies of DADH, one was the small amounts of enzyme available which were not enough to perform active site titration, and hence it was not possible to determine the absolute values of the various kinetic coefficients. The second was the large amount of alcohols that we planned to use, and how to determine the substrate specificity without obtaining all the kinetic coefficients and their absolute values. This of course required optimal reaction conditions in order to determine the specificity of the various DADHs. I will try do describe some of the problems and how we solved them.

Quantitative estimation of functional enzyme without active site titration

How did we ensure that the same amount of functional enzyme was used in each experiment? This problem was solved simply by using a high saturating concentration of ethanol as a standard at optimal conditions as described below. We also showed that the enzyme activity under the condition used with a fixed ethanol concentration was linear with the variation in enzyme concentration ($v = k$ x [e]), although the absolute [e] was not known. We therefore presented all data as V_m, V_m/K_{m} and activity at fixed alcohol concentrations (see below) relative to the activity of ethanol [32, 38, 39]. Later on, when we had enough

DADH to perform active-site titrations, we used the standard conditions above in the development of a rate assay that was calibrated against the titration [31, 32]. This of course allowed us to convert all old relative data to absolute data.

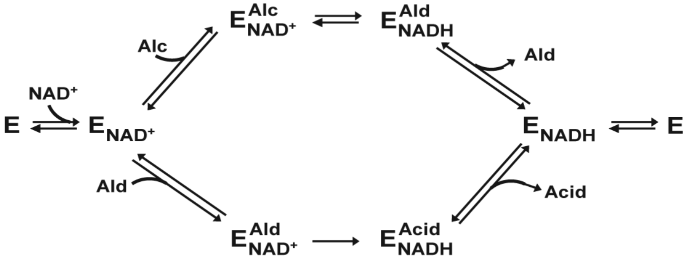
Detection of substrate specificity using a single coenzyme concentration

Which of the kinetic coefficients reflect an enzyme's substrate specificity? As DADH is a two substrate enzyme as described above, the substrate specificity is reflected in the kinetic coefficient N_2 in Equation 3 or K_{m2}/k_{cat} in Equation 5 for various alcohols. Our aim was to get a picture of the substrate specificity by using a fixed NAD^+ concentration, vary the concentration of some selected alcohols and thereafter determine the activity for all alcohols at a fixed concentration. In order to use a fixed NAD^+ concentration, it is important that this is high enough so the obtained (app)k_{cat}/K_{m2} and (app)k_{cat} values are as close as possible to the values for an infinite coenzyme concentration. This requires that $K_{m1}/(k_{cat}$ $[NAD^+])$ is much less than $1/k_{cat}$ and that K_{m2} $K_{ia}/(k_{cat}$ $[NAD^+])$ is much less than K_{m2}/k_{cat}. The problem was to find the experimental conditions that were optimal in order to obtain reliable results. We decided to use a temperature that was used in the classical experiments on horse liver ADH by Theorell and McKinley-McKee [40] and by Dalziel [41]. Our initial experiments revealed that optimal conditions were obtained using 0.1 M glycine–NaOH buffer pH 9.5 and a fixed concentration of 0.5 mM of NAD^+. The reason to choose such a high pH compared to physiological pH is of course the equilibrium of the reaction and the amount of NAD^+ needed to obtain acceptable values of (app)k_{cat} and (app)k_{cat}/K_{m2}. As an example, at neutral pH with a NAD^+ concentration of 1 mM, $N_1/[NAD^+]$ and $N_{12}/[NAD^+]$ are approximately the same as N_0 and N_2, respectively [42, 43]. Using 10 mM of NAD^+ would have reduced the ratios to be 5 – 10% of the corresponding N coefficient. However at basic pH (9.5 – 10) these two relations are approximately 2% of corresponding N_0 and N_2 coefficient, using 0.5 mM of NAD^+ [42, 43]. These calculations are based on the two substrates ethanol and propan-2-ol using the *D. melanogaster* alleloenzyme ADH^S and the *D. lebanonensis* ADH [42, 43]. This can be compared with results for sheep liver Sorbitol dehydrogenase (SDH), where the corresponding relations $(1+ N_1/(N_0\ [NAD^+])$ and $1 + N_{12}/(N_2\ [NAD^+]))$ using 1 mM NAD^+ are close to 2 at both pH 7.4 and 9.5 using sorbitol as varied substrate [34]. Studies of this SDH revealed that the substrate specificity was the same at neutral and at basic pH [44].

DADH is also able to oxidize aldehydes in the presence of NAD^+ to their corresponding acids (Equation 2) [3, 4, 45, 46]. At pH 7.0, it is not possible to follow this reaction by determining the production of NADH. This is due to the dismutation reaction (Equation 6; which is the sum of Equations 1 and 2), i. e. as fast as NADH is produced, it reacts with the aldehyde and produces alcohol.

$$2 \text{ aldehyde} + H_2O \leftrightarrow \text{carboxylate}^- + \text{alcohol} + H^+ \qquad (6)$$

However, above pH 9.0 it has been possible to detect NADH production with DADH as the reduction reaction of aldehyde to alcohol is slower than at neutral pH [3, 4, 45, 46], and hence an unequal amount of alcohol and acid is produced in the dismutation reaction. It has been argued that the increase in A_{340nm}, i.e. the release of NADH, is not a direct measure of the aldehyde oxidation reaction and acid production, and that the resulting kinetic values cannot be compared with those for alcohol dehydrogenation. This indicates that aldehyde oxidation can only be studied with methods such as ^{1}H-NMR, gas chromatography or pH-stat titrations. Due to the amount of enzyme needed, as well as initial-rate measurements cannot be performed with the two former methods, one would expect that this would limit the possibility of doing kinetic studies on the aldehyde oxidation reaction. Even if this is correct to a certain extent, we have shown that it is possible to do kinetic studies by following the initial-rate production of NADH at pH 9.5, by using a very sensitive filter fluorimeter specially built to study dehydrogenase reactions [4]. With this instrument we could detect the continuous production of NADH, with a detection limit as low as 10 nM. We performed substrate specificity studies, as well as detecting kinetic coefficients for the aldehyde oxidation reaction and compared this with both the alcohol oxidation reaction and aldehyde reduction reaction [4]. The combination of dead-end and product inhibitors was used to determine the reaction mechanism for the aldehyde oxidation pathway, which like the interconversion between alcohols and aldehydes was consistent with a compulsory ordered mechanism as shown in Scheme 1. It is important to emphasize that it is necessary to avoid buffers containing primary or secondary amine groups, as these formed Schiff bases with the aldehydes. This shows the possibilities to do studies of enzymes if optimal reaction conditions and optimal instrumentation is used, and that some type of studies is not possible to perform at neutral pH.



Scheme 1. Reaction mechanism for DADH. The upper pathway shows the interconversion between an alcohol (Alc) and an aldehyde (Ald), and the lower pathway the oxidation of an aldehyde (Ald) to a carboxylic acid (Acid). The mechanism for these reactions was consistent with a compulsory ordered pathway, where the coenzymes form binary enzyme complexes.

Determination of substrate specificity using a single fixed alcohol and NAD$^+$ concentration

In order to use only one alcohol concentration, which is the optimal concentration to use? As N_2 (K_{m2}/k_{cat}) reflects the activity at low alcohol concentrations, one should use a concentration that is below K_m. We used 1 mM of the different alcohols, which was assumed to be an acceptable concentration. This also appeared to be the case for the primary alcohols and a lot of the secondary alcohols. Although this concentration proved to be a little too high with respect to some of the secondary alcohols, it reflected in an acceptable way the (app)k_{cat}/K_{m2} values in those cases where these were obtained [32, 38, 39]. The substrate specificity obtained at pH 9.5 has been shown to reflect the substrate specificity at neutral pH for DADH [42, 43].

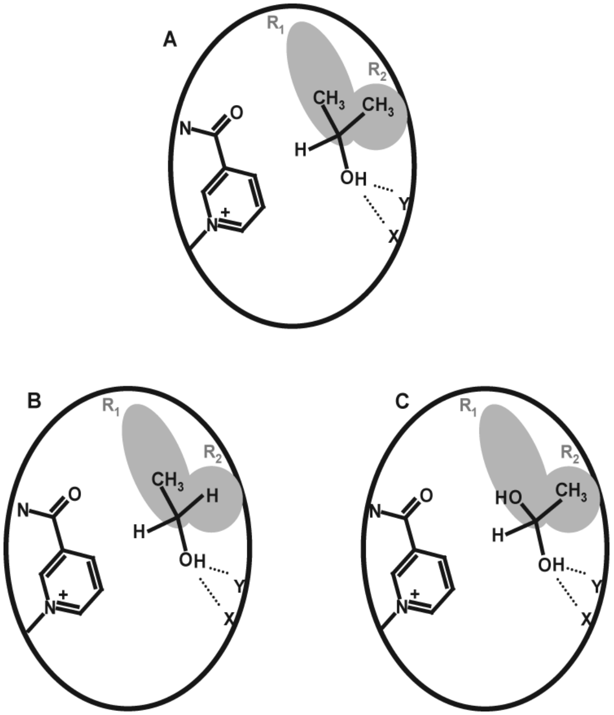


Figure 2. Schematic representation of the alcohol binding site in ternary DADH-NAD$^+$-substrate complexes. The binding of (**A**) propan-2-ol, (**B**) ethanol and (**C**) acetaldehyde (diol) is shown. The hydrophobic and bifurcated part of the enzyme active site that interacts with the alkyl groups in alcohols and aldehydes is shown in grey and labelled as R_1 and R_2. Also shown is the nicotinamide part of the oxidized coenzyme NAD$^+$ and the OH-group of the substrates that interacts with the OH-group in the two conserved residues tyrosine-151 (Y) and serine-138 (X) using *D. lebanonensis* numbering.

Professor Ladenstein and his group at Karolinska Institutet in Sweden obtained the 3D-structure of several ternary DADH-NAD$^+$-ketone complexes through X-ray crystallography [10], and their description of the topology of the active site was exactly as we depicted from our substrate specificity studies more then 15 years earlier [32, 38, 39]. The kinetic and X-ray crystallography data showed a hydrophobic, bifurcated substrate-binding site in DADH, which results in optimal binding and activity with secondary alcohols (Fig. 2a). Kinetic and X-ray crystallographic data has also shown that the alkyl chain in ethanol and other primary alcohols as well as aldehydes during reduction with NADH to alcohols binds to the R_1 part of this bifurcated alcohol binding part of the active site (Fig. 2b) [10, 47]. However, in the oxidation of aldehydes to acids, the alkyl chains in the aldehyde binds to the R_2 binding part of the active site (Fig. 2c) [11].

Additives in a Purified Enzyme Preparation May Alter the Biochemical Properties of the Enzyme

In this part I will take up the importance of a careful description of an enzyme assay, i.e. the conditions used including the concentrations of all the constituents in the assay. The example used shows that the amount of additives present in a preparation of proMMP-2 determines whether or not trypsin will act as an activator of this MMP.

The literature states that serine proteinases like trypsin cannot activate proMMP-2. This has been based on very careful studies by Okada *et al.*, [48] in which they studied the activation of proMMP-2 by the organic mercury compound 4-aminophenyl mercury acetate (APMA). In this study, several proteinases including trypsin were also tested as proMMP-2 activators, and none of these activated the enzyme. The conditions used for the activation with APMA are very well documented, while the conditions used when trypsin and the other proteinases were tested, are less well documented. They used various amounts of trypsin (0.1 – 100 µg/ml) at 22 °C from 5 minutes to 30 hours. What was not explicitly cited was the concentration used of $CaCl_2$, and if they used Brij-35 in the assay and if so, what was the concentration. In another article, it was shown that trypsin-2 is an activator of proMMP-9, but could only partly activate proMMP-2 [49]. However, nothing was mentioned with respect to reaction conditions such as added Brij-35 or $CaCl_2$.

These results fitted badly with my own studies on the expression of MMP-2 from cultured fibroblasts [50, 51]. After harvesting the cell-conditioned serum-free medium, we used to add $CaCl_2$, BSA and Hepes (pH 7.5) to a final concentration of 10 mM, 0.2 % and 0.1 M respectively. This was done in order to protect the enzyme in the freezing (–20 °C) and thawing processes. The proMMP-2 in these serum-free media was always activated by trypsin, and we showed that this was not due to the activation of another MMP (collagenase 1/MMP-1) in the media, that then could activate proMMP-2 [51]. In a recent study, we did check whether trypsin could activate recombinant proMMP-2 [35]. In these studies we decided to test whether the discrepancies between our results, using cell conditioned media and those that used purified proMMP-2, could be ascribed to differences in experimental reaction conditions, or, that the activation of proMMP-2 in the cell conditioned media actually was due to trypsin-induced activation of a latent MMP-2 activator in the media and not through a direct activation of proMMP-2.

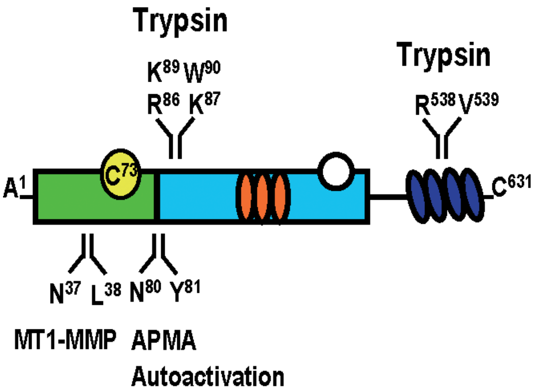


Figure 3. Schematic drawing showing the cleavage sites in proMMP-2 produced by MT1-MMP, APMA, autoactivation and trypsin. MT1-MMP cleaves N-terminal for the invariant C^{73} that is linked to the active site zinc in the proenzyme. The intermediate formed is further processed by autoactivation that generates the fully active 62 kDa form of MMP-2. Treatment of proMMP-2 with AMPA results in autoactivation. Trypsin cleaves C-terminal for the autoactivation site, and at several sites in the C-terminal region, ending up with a cleavage between R^{538} and V^{539} generating an active 50 kDa form.

The commercial recombinant proMMP-2 used in our studies was delivered from Chemicon, and contained 100 µg/ml proMMP-2 in 5 mM Tris–HCl (pH 7.5), 0.1 mM $CaCl_2$ and 0.005% Brij-35. In our activation experiments with trypsin, this proMMP-2 stock solution always ended up 30 – 40 times diluted in 0.1 M Hepes, pH 7.5 prior to the addition of the different amounts of trypsin, $CaCl_2$ and Brij-35. Other conditions varied were temperature (4, 16, 22 and 37 °C) and incubation time with trypsin (2 minutes to 24 hours). Our results showed that trypsin is actually an activator of proMMP-2 that first removes the pro-domain from the 72 kDa proMMP-2 and generates an active 62 kDa form. This is followed by a trypsin-induced successive removal of the most C-terminal parts of the hemopexin-like domain that ends up in a 50 kDa active form of the enzyme as shown in Fig. 3. Without exogenous added $CaCl_2$ and Brij-35, trypsin induced activation at the low temperatures, while at 37 °C, the proMMP-2 was only degraded. Both $CaCl_2$ and Brij-35 stabilized the MMP, which could be activated at 37 °C if only one of these compounds were present. However in the presence of 0.05% Brij-35, trypsin-induced activation decreased with increasing concentrations of $CaCl_2$. At 5 and 10 mM $CaCl_2$ (approximately 5 – 10 times the physiological concentration in tissues) only a small fraction was activated and almost all the enzyme remained in the proform. Thus the discrepancy in the literature cannot be ascribed experimental faults or the activation of an unknown proMMP-2 activator in cell conditioned media, but was due to various reaction conditions. The trypsin-induced activation of proMMP-2 generates an active MMP-2 with a slightly shorter N-terminal than the enzyme activated by the assumed most important biological activator, MT1-MMP, or the organic mercurial compound APMA (Fig. 3). This difference in structure also resulted in an altered capacity of the enzyme to degrade the biological substrate, gelatin, and a chromogenic substrate, as well as an altered binding strength (K_i) to the biological inhibitor TIMP-1 [35].

These results clearly demonstrate the importance of various additives and to report their concentration, as they may affect the parameters studied. By reporting all the additives and their concentrations, authors allow others to extend their investigation as well as to test the substance in the published results. As shown above, due to the presence of various additives in the reaction assay, an erroneous statement about a biological parameter of an enzyme has been introduced in the literature which is hard to erase.

Conclusions

In order to obtain a full description of the kinetic coefficients, the concentration of functional enzyme active sites is required. This should be obtained by a method based on active-site titration.

If it is not possible to obtain the amount of functional enzyme by active site titration methods, my view is that it is wrong to present kinetic coefficients like k_{cat} and k_{cat}/K_m using units such as s^{-1} and $\mathrm{mM}^{-1}\ \mathrm{s}^{-1}$, respectively. In a lot of cases it is much better to present the kinetic coefficients as specific activities or relative activities using V_m and V_m/K_m. In other cases it may be better to introduce new kinetic coefficients, which for example could be denoted $\mathrm{k}_{\mathrm{cat(-t)}}$ and $k_{cat(\mathrm{-t})}/K_\mathrm{m}$ (using units such s^{-1} and $\mathrm{mM}^{-1}\mathrm{s}^{-1}$), where (–t) shows that the catalytic activity is not based on active site titration.

It should not be necessary to stress that a clear description of conditions used, including all additives, should be reported.

Standardization of parameters such as pH and temperature can be done to a certain extent where it is appropriate.

References

[1] Boyce, S., Tipton, K., McDonald, A.G. (2003) Extending enzyme classification with metabolic and kinetic data: Some difficulties to be resolved. In: *1st International Beilstein Workshop on Experimental Standard Conditions of Enzyme Characterizations* (Hicks, M., Kettner, C., Eds), pp. 17–43, Beilstein-Institut, Rüdesheim/Rhein, Germany.

[2] Winberg, J.O., McKinley-McKee, J.S. (1992) Kinetic interpretations of active site topologies and residue exchanges in Drosophila alcohol dehydrogenases. *Int. J. Biochem.* **24**:169–181.

[3] Henehan, G.T., Chang, S.H., Oppenheimer, N.J. (1995) Aldehyde dehydrogenase activity of Drosophila melanogaster alcohol dehydrogenase: burst kinetics at high pH and aldehyde dismutase activity at physiological pH. *Biochemistry* **34**:12294–12301.

[4] Winberg, J.O., McKinley-McKee, J.S. (1998) Drosophila melanogaster alcohol dehydrogenase: mechanism of aldehyde oxidation and dismutation. *Biochem. J.* **329**:561–570.

[5] Chambers, G.K. (1991) Gene expression, adaptation and evolution in higher organisms. Evidence from studies of Drosophila alcohol dehydrogenases. *Comp. Biochem. Physiol. [B]* **99**:723–730.

[6] Jörnvall, H., Persson, B., Krook, M., Atrian, S., Gonzalez-Duarte, R., Jeffery, J., Ghosh, D. (1995) Short-chain dehydrogenases/reductases (SDR). *Biochemistry* **34**:6003–6013.

[7] Jörnvall, H., Persson, M., Jeffery, J. (1981) Alcohol and polyol dehydrogenases are both divided into two protein types, and structural properties cross-relate the different enzyme activities within each type. *Proc. Natl Acad. Sci. U S A* **78**:4226–4230.

[8] Kallberg, Y., Oppermann, U., Jörnvall, H., Persson, B. (2002) Short-chain dehydrogenases/reductases (SDRs): Coenzyme-based functional assignments in completed genomes. *Eur. J. Biochem.* **269**:4409–4417.

[9] Benach, J., Atrian, S., Gonzalez-Duarte, R., Ladenstein, R. (1998) The refined crystal structure of Drosophila lebanonensis alcohol dehydrogenase at 1.9 A resolution. *J. Mol. Biol.* **282**:383–399.

[10] Benach, J., Atrian, S., Gonzalez-Duarte, R., Ladenstein, R. (1999) The catalytic reaction and inhibition mechanism of Drosophila alcohol dehydrogenase: observation of an enzyme-bound NAD-ketone adduct at 1.4 A resolution by X-ray crystallography. *J. Mol. Biol.* **289**:335–355.

[11] Benach, J., Winberg, J.O., Svendsen, J.S., Atrian, S., Gonzalez-Duarte, R., Ladenstein, R. (2005) Drosophila alcohol dehydrogenase: acetate-enzyme interactions and novel insights into the effects of electrostatics on catalysis. *J. Mol. Biol.* **345**:579–598.

[12] Geer, B.W., Heinstra, P.W. H., Kapoun, A.M., Van der Zel, A. (1990) Alcohol dehydrogenase and alcohol tolerance in drosophila melanogaster. In: *Ecological and Evolutionary Genetics of Drosophila* (Barker, J.S. F. e. a., Ed.), Plenum Press, New York.

[13] Allemann, R.K., Hung, R., Benner, S.A. (1988) Stereochemical profile of the dehydrogenases of drosophila melanogaster. *J. Am. Chem. Soc.* **110**:5555–5560.

[14] Benner, S.A., Nambiar, K.P., Chambers, G.K. (1985) A stereochemical imperative in dehydrogenases: New data and criteria for evaluating function-based theories in bioorganic chemistry. *J. Am. Chem. Soc.* **107**:5513–5517.

[15] Winberg, J.O., Brendskag, M.K., Sylte, I., Lindstad, R.I., McKinley-McKee, J.S. (1999) The catalytic triad in drosophila alcohol dehydrogenase: pH, temperature and molecular modelling studies. *J. Mol. Biol.* **294**:601–605.

[16] Winberg, J.O., McKinley-McKee, J.S. (1988) Drosophila melanogaster alcohol dehydrogenase. Biochemical properties of the NAD+-plus-acetone-induced isoenzyme conversion. *Biochem. J.* **251**:223–227.

[17] Winberg, J.O., McKinley-McKee, J.S. (1994) Drosophila melanogaster alcohol dehydrogenase: product-inhibition studies. *Biochem. J.* **301**:901–909.

[18] Winberg, J.O., Thatcher, D.R., McKinley-McKee, J.S. (1982) Alcohol dehydrogenase from the fruitfly Drosophila melanogaster. Inhibition studies of the alleloenzymes AdhS and AdhUF. *Biochim. Biophys. Acta* **704**:17–25.

[19] Winberg, J.O., Thatcher, D.R., McKinley-McKee, J.S. (1983) Drosophila melanogaster alcohol dehydrogenase: an electrophoretic study of the AdhS, AdhF, and AdhUF alleloenzymes. *Biochem. Genet.* **21**:63–80.

[20] Nagase, H., Woessner, J.F., Jr. (1999) Matrix metalloproteinases. *J. Biol. Chem.* **274**:21491–21494.

[21] Rawlings, N.D., Tolle, D.P., Barrett, A.J. (2004) MEROPS: the peptidase database. *Nucleic Acids Res.* **32**:D 160 –D 164.

[22] Overall, C.M., López-Otin, C. (2002) Strategies for MMP inhibition in cancer: Innovations for the post-trial era. *Nature Rev. Cancer* **2**:657–672.

[23] Vihinen, P., Kähäri, V.M. (2002) Matrix metalloproteinases in cancer: Prognostic markers and therapeutic targets. *Int. J. Cancer* **99**:157–166.

[24] Nagase, H. (1997) Activation mechanisms of matrix metalloproteinases. *Biol. Chem.* **378**:151–160.

[25] Brew, K., Dinakarpandian, D., Nagase, H. (2000) Tissue inhibitors of metalloproteinases: evolution, structure and function. *Biochim. Biophys. Acta* **1477**:267–283.

[26] Dalziel, K. (1957) Initial steady state velocities in the evaluation on enzyme-coenzyme-substrate reactions. *Acta Chem. Scand.* **11**:1706–1723.

[27] Cleland, W.W. (1977) Determining the chemical mechanisms of enzyme-catalyzed reactions by kinetic studies. *Adv. Enzymol. Relat. Areas Mol. Biol.* **45**:273–387.

[28] Brocklehurst, K. (1996) Active site titration. In: *Enzymology: LabFax* (Engel, P.C., Ed.), pp. 59–66, Academic Press, San Diego.

[29] Theorell, H., Yonetani, T. (1963) Liver alcohol dehydrogenase-DPN-Pyrazole complex: A model of a ternary intermediate in the enzyme reaction. *Biochem. Z.* **338**:537–553.

[30] Ganzhorn, A.J., Green, D.W., Hershey, A.D., Gould, R.M., Plapp, B.V. (1987) Kinetic characterization of yeast alcohol dehydrogenases: Amino acid residue 294 and substrate specificity. *J. Biol. Chem.* **262**:3754–3761.

[31] Winberg, J.O., Hovik, R., McKinley-McKee, J.S. (1985) The alcohol dehydrogenase alleloenzymes AdhS and AdhF from the fruitfly Drosophila melanogaster: an enzymatic rate assay to determine the active-site concentration. *Biochem. Genet.* **23**:205–216.

[32] Winberg, J.O., Hovik, R., McKinley-McKee, J.S., Juan, E., Gonzalez-Duarte, R. (1986) Biochemical properties of alcohol dehydrogenase from Drosophila lebanonensis. *Biochem. J.* **235**:481–490.

[33] Ganzhorn, A.J., Plapp, B.V. (1988) Carboxyl groups near the active site zinc contribute to the catalysis in yeast alcohol dehydrogenase. *J. Biol. Chem.* **263**:5446–5454.

[34] Lindstad, R.I., Hermansen, L.F., McKinley-McKee, J.S. (1992) The kinetic mechanism of sheep liver sorbitol dehydrogenase. *Eur. J. Biochem.* **210**:641–647.

[35] Lindstad, R.I., Sylte, I., Mikalsen, S.O., Seglen, P.O., Berg, E., Winberg, J.O. (2005) Pancreatic trypsin activates human promatrix metalloproteinase-2. *J. Mol. Biol.* **350**:682–698.

[36] Minond, D., Lauer-Fields, J.J., Nagase, H., Fields, G.B. (2004) Matrix metalloproteinase triple-helical peptidase activities are differentially regulated by substrate stability. *Biochemistry* **43**:11474–11481.

[37] Park, H.I., Turk, B.E., Gerkema, F.E., Cantley, L.C. (2002) Peptide substrate specificities and protein cleavage sites of human endometastase/matrilysin-2/matrix metalloproteinase-26. *J. Biol. Chem.* **38**:35168–35175.

[38] Hovik, R., Winberg, J.O., McKinley-McKee, J.S. (1984) Drosophila melanogaster alcohol dehydrogenase: substrate specificity of the ADHF alleloenzyme. *Insect Biochem.* **14**:345–351.

[39] Winberg, J.O., Thatcher, D.R., McKinley-McKee, J.S. (1982) Alcohol dehydrogenase from the fruitfly Drosophila melanogaster. Substrate specificity of the alleloenzymes AdhS and AdhUF. *Biochim. Biophys. Acta* **704**:7–16.

[40] Theorell, H., McKinley-McKee, J.S. (1961) Liver alcohol dehydrogenase: I. Kinetics and equilibria without inhibitors. *Acta Chem. Scand.* **15**:1797–1810.

[41] Dalziel, K. (1962) Kinetic studies of liver alcohol dehydrogenases. *Biochem. J.* **84**:244–254.

[42] Brendskag, M.K., McKinley-McKee, J.S., Winberg, J.O. (1999) Drosophila lebanonensis alcohol dehydrogenase: pH dependence of the kinetic coefficients. *Biochim. Biophys. Acta* **1431**:74–86.

[43] Winberg, J.O., McKinley-McKee, J.S. (1988) The AdhS alleloenzyme of alcohol dehydrogenase from Drosophila melanogaster. Variation of kinetic parameters with pH. *Biochem. J.* **255**:589–599.

[44] Lindstad, R.I., Köll, P., McKinley-McKee, J.S. (1998) Substrate specificity of sheep liver sorbitol dehydrogenase. *Biochem. J.* **330**:479–487.

[45] Heinstra, P.W., Geer, B.W., Seykens, D., Langevin, M. (1989) The metabolism of ethanol-derived acetaldehyde by alcohol dehydrogenase (EC 1.1.1.1) and aldehyde dehydrogenase (EC 1.2.1.3) in Drosophila melanogaster larvae. *Biochem. J.* **259**:791–797.

[46] Moxon, L.N., Holmes, R.S., Parsons, P.A., Irving, M.G., Doddrell, D.M. (1985) Purification and molecular properties of alcohol dehydrogenase from drosophila melanogaster: Evidence from NMR and kinetic studies for function as an aldehyde dehydrogenase. *Comp. Biochem. Physiol. [B]* **80B**:525–535.

[47] Winberg, J.O., Martinoni, B., Roten, C., McKinley-McKee, J.S. (1993) Drosophila alcohol dehydrogenase: stereoselective hydrogen transfer from ethanol. *Biochem. Mol. Biol. Int.* **31**:651–658.

[48] Okada, Y., Morodomi, T., Enghild, J.J., Suzuki, K., Yasui, A., Nakanishi, I., Salvesen, G., Nagase, H. (1990) Matrix metalloproteinase 2 from human rheumatoid synovial fibroblasts. Purification and activation of the precursor and enzymic properties. *Eur. J. Biochem.* **194**:721–730.

[49] Sorsa, T., Salo, T., Koivunen, E., Tyynela, J., Konttinen, Y.T., Bergmann, U., Tuuttila, A., Niemi, E., Teronen, O., Heikkila, P., Tschesche, H., Leinonen, J., Osman, S., Stenman, U.H. (1997) Activation of type IV procollagenases by human tumor-associated trypsin- 2. *J. Biol. Chem.* **272**:21067–21074.

[50] Winberg, J.O., Gedde-Dahl, T. (1986) Gelatinase expression in generalized epidermolysis bullosa simplex fibroblasts. *J. Invest. Dermatol.* **87**:326–329.

[51] Winberg, J.O., Gedde-Dahl, T., Jr. (1992) Epidermolysis bullosa simplex: expression of gelatinase activity in cultured human skin fibroblasts. *Biochem. Genet.* **30**:401–420.

Biographies

Robert A. Alberty

I graduated from the University of Nebraska in 1943, did research on blood plasma at the University of Wisconsin, received my PhD in 1947, and became an Instructor at the University of Wisconsin. I became interested in enzyme kinetics and was a postdoc with Linus Pauling at CalTech in 1950 – 51. Back at the University of Wisconsin we isolated fumarase, determined the rate equations for both the forward and reverse reactions, and confirmed the Haldane equation, among other things. In 1963 I became Dean of the Graduate School, and in 1967 I became Dean of the School of Science at MIT. I was so deeply involved in administration that I had to stop research. When I left the Deanship in 1982, I decided to use computers to study petroleum processing, and that lead me to the use of Legendre transforms to define new thermodynamic properties. In 1991 I had my "eureka'' moment when I realized that when the pH is used as an independent variable in biochemistry, you should not use the Gibbs energy G, but you need to use a Legendre transform to define a transformed Gibbs energy G'. This lead to a IUPAC-IUBMB report in 1994. I have been building a database (BasicBiochemData3) on the thermodynamics of biochemical reactions, and have written two books on the subject, the most recent one in Mathematica.

Rolf Apweiler

is a Team Leader and Senior Scientist at the European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. He studied Biology with a focus on Biochemistry and Molecular Biology in Heidelberg, Germany and Bath, UK, and worked in drug discovery in the pharmaceutical industry. He became involved in Bioinformatics through the Swiss-Prot project in 1987. He received his PhD in 1994 from the Center for Molecular Biology, University of Heidelberg, Germany and joined the European Bioinformatics Institute the same year. Dr Apweiler has coordinated the Swiss-Prot work at the European Bioinformatics Institute since 1994. He also started, among other projects, the TrEMBL protein database, the Integrated resource of protein families, domains and functional sites (InterPro), Gene Ontoloy Annotation (GOA), the Integr8 web portal, the Genome Reviews, and the UniProt resource (the sucessor of the Swiss-Prot, TrEMBL and PIR projects). These projects have organised large amounts of protein information, provided comparisons between proteomes and aim to produce dynamic, controlled vocabularies that can be applied to all organisms. In addition, Dr Apweiler has been in charge of the EMBL nucleotide sequence database since 2001. Dr Apweiler served on many review and editorial boards and published more than hundred peer-reviewed articles and numerous book chapters. Rolf Apweiler has also a long-standing interest in data standards and nomenclature as exemplified in his engagement in the IUBMB Nomenclature Committee, the HUGO gene nomenclature committee, and in the HUPO Proteomics Standards Initiative.

URLs:
http://www.ebi.ac.uk/seqdb/
http://www.uniprot.org
http://www.ebi.ac.uk/interpro/
http://www.ebi.ac.uk/integr8
http://www.ebi.ac.uk/GOA/
http://www.ebi.ac.uk/embl/
http://www.ebi.ac.uk/GenomeReviews/

Jildau Bouwman

Jildau Bouwman was educated in neuroscience at the vrije Universiteit Amsterdam. She did a molecular biology internship in synapse development and an electrophysiological internship on receptor subunit switches during development. In 1999 she started with a PhD studentship at the Rudolf Magnus Institute for neurosciences in Utrecht. In 2004 she started her recent position as post-doc in the department of Molecular Cell Physiology at the Vrije Universiteit Amsterdam. She is involved in the “Vertical Genomics” project. The eventual goal of the project is to be able to predict how a change in gene expression can influence metabolic fluxes.

Richard Cammack

is Professor of Biochemistry at King's College, University of London. He graduated from the University of Cambridge with a BA in Natural Sciences in 1965 and PhD in Enzymology, under Malcolm Dixon in 1968. He has over 200 publications on mechanisms of electron transfer and enzyme catalysis, particularly in iron-sulfur proteins such as hydrogenases and aromatic dioxygenases. He is currently using EPR spectroscopy to study the role of iron in health and disease. He is past Chairman (2000 – 2005) of the Nomenclature committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and Joint commission on Biochemical Nomenclature (JCBN), and Editor-in-Chief of the second edition of the Oxford Dictionary of Biochemistry and Molecular Biology.

Athel Cornish-Bowden

carried out his undergraduate studies at Oxford, obtaining his doctorate with Jeremy R. Knowles in 1967. After three post-doctoral years in the laboratory of Daniel E. Koshland, Jr., at the University of California, Berkeley, he spent 16 years as Lecturer, and later Senior Lecturer, in the Department of Biochemistry at the University of Birmingham. Since 1987 he has been Directeur de Recherche in three different laboratories of the CNRS at Marseilles. Although he started his career in a department of organic chemistry virtually all of his research has been in biochemistry,with particular reference to enzymes, including pepsin, mammalian hexokinases and enzymes involved in electron transfer in bacteria. He has written several books relating to enzyme kinetics, including *Analysis of Enzyme Kinetic Data* (Oxford University Press, 1995) and *Fundamentals of Enzyme Kinetics* (3 rd edition, Portland Press, 2004). Since moving to Marseilles he has been particularly interested in multi-enzyme systems, including the regulation of metabolic pathways. More generally, he has long had an interest in biochemical aspects of evolution, and his semi-popular book in this field, *The Pursuit of Perfection*, will be published by Oxford University Press in 2004.

Kirill N. Degtyarenko

Born in Moscow region, Russia in 1967.
In 1989, graduated from the Russian State Medical University, Medico-Biological Faculty (M.D.; M.Sc. in Biochemistry). Since 1986, he worked under guidance of Prof. Valentin Uvarov, first at the Department of Biochemistry, MBF and later at the Institute of Biomedical Chemistry, Moscow.
In 1992, he defended his Ph.D. thesis on Molecular Evolution of the P450 Superfamily at the Institute of Biomedical Chemistry (supervisors: Prof. Alexander Archakov and Valentin Uvarov).
He spent one year at the International Centre for Genetic Engineering and Biotechnology, Trieste, Italy, before joining the Department of Biochemistry and Molecular Biology, the University of Leeds, UK in 1995. Since 1998, Kirill has been working at the European Bioinformatics Institute, Hinxton (near Cambridge).

Martin Field

1982	Undergraduate degree (BA) from St.Catharine's College, Cambridge in Natural Sciences.
1982 – 1985	PhD at the University of Manchester in quantum chemistry.
1985 – 1989	Postdoctorate at the University of Harvard – theoretical studies of enzymatic reaction mechanisms and protein dynamics.
1989 – 1992	Posts at the University of Geneva and at the NIH, Bethesda, Maryland.
1992+	Group leader of the modeling and simulation laboratory at the Institut de Biologie Structurale in Grenoble.

Martin's general research involves using molecular modeling and simulation approaches for studying problems of biological interest. Specific interests include the development and application of hybrid potential techniques for studying enzymatic reaction mechanisms and other condensed phase processes.

Wilfred R. Hagen

is a Professor of Enzymology in the Department of Biotechnology at Delft University of Technology in Delft, The Netherlands. The central research theme in his group is the role of metal ions in redox biocatalysis. Fred Hagen completed his PhD on EPR of metalloproteins at the University of Amsterdam in 1982 with SPJ Albracht and EC Slater. He then took up an EMBO fellowship, and subsequently an NIH fellowship, at the Biophysics Research Division of The University of Michigan in Ann Arbor, to work on g-strain (the theory of EPR spectra from biomacromolecules) with WR Dunham and RH Sands.

In 1984 he returned to The Netherlands to join the Biochemistry Department of C Veeger at Wageningen University to set up a group on metalloproteins. In 1995 he was appointed to a chair of Physical Chemistry at the University of Nijmegen, where he headed the high-frequency EPR spectroscopy group. In 1998 he was also appointed professor of Bioinorganic Chemistry in Wageningen. In 2000 he resigned from both positions and moved to Delft University of Technology to take up the chair of Enzymology in the Department of Biotechnology.

http://www.bt.tudelft.nl/enz.

Jan-Hendrik Hofmeyr

is Professor in the Department of Biochemistry at the University of Stellenbosch, South Africa. He obtained his Ph.D. in 1986 at the University of Stellenbosch after collaborating with Henrik Kacser (one of the founders of metabolic control analysis) and the enzymologist Athel Cornish-Bowden. Jannie and his colleagues Jacky Snoep and Johann Rohwer form the Triple-J Group for Molecular Cell Physiology, a research group that studies the control and regulation of cellular processes using theoretical, computer modelling and experimental approaches. He has made numerous fundamental contributions to the development of metabolic control analysis and computational cell biology, and with Athel Cornish-Bowden developed both co-response analysis and supply-demand analysis as a basis for understanding metabolic regulation. He is a Fellow of the Academy of Science of South Africa and, with the other Triple-Js, chairs the International Study Group for BioThermoKinetics. He recently won the Harry Oppenheimer Fellowship Award, South Africa’s most prestigious science award.

Hermann-Georg Holzhütter

became professor in 1998 and head of the research group “Theoretical Systemsbiology” at the Institute of Biochemistry of the Medical School (Charité) of the Humboldt-University in Berlin. His academic roots extend back to the late 60 s / early 70 s when he studied Physics at the Humboldt-University. In 1976, he was awarded his Ph. D. for his research on the theory of transport in small gap semiconductors and in 1986 he received his habilitation (Dr. rer. nat. habil.) for theoretical studies on the dynamics and evolution of enzymatic networks. Today, his research topics are enzyme kinetics and the modelling of complex enzymatic networks with an immunological focus.

Carsten Kettner

studied biology at the University of Bonn and obtained his diploma at the University of Göttingen in the group of Prof. Gradmann which had the pioneering and futuristic name – "Molecular Electrobiology". This group consisted of people carrying out research in electrophysiology and molecular biology in fruitful cooperation. In this mixed environment, he studied transport characteristics of the yeast plasma membrane using patch clamp techniques. In 1996 he joined the group of Dr. Adam Bertl at the University of Karlsruhe and undertook research on another yeast membrane type. During this period, he successfully narrowed the gap between the biochemical and genetic properties, and the biophysical comprehension of the vacuolar proton-translocating ATP-hydrolase. He was awarded his Ph.D for this work in 1999. As a post-doctoral student he continued both the studies on the biophysical properties of the pump and investigated the kinetics and regulation of the dominant plasma membrane potassium channel (TOK1). In 2000 he moved to the Beilstein-Institut to represent the biological section of the funding department. Here, he is responsible for the organization of symposia (sic!), research (proposals) and development of new products considering the ideas of the Beilstein-Instiut, such as a medical plant database, considering the ideas of the Beilstein-Institut. He also co-ordinates the work of the STRENDA commission which is concerned with the standardization of enzyme data (see also www.strenda.org).

Ursula Kummer

After finishing her Abitur in Baden-Baden, Ursula Kummer studied Biochemistry, Physics and Chemistry in Tübingen, Germany and Eugene, Or, USA. She received a MSc in Chemistry at the University of Oregon, Eugene, Or, USA, a Vordiplom in Physics, a Diplom in Biochemistry and a PhD in Biochemistry at the University of Tübingen.

Her PhD thesis which combined experimental and computational studies was finished in 1996 and dealt with the Nonlinear Dynamics of Enzymatic Systems.

After postdoctoral time in Tübingen she joined the EML in Heidelberg where she became group leader in 2000. Since then her group, the Bioinformatics and Computational Biochemistry Group has been working on development of methodologies for the simulation, modeling and analysis of biochemical networks and on their application. Ursula Kummer is one of the coordinators of the BIOMS center in Heidelberg.

Nicolas Le Novere

started his career in the team of Jean-Pierre Changeux at the Pasteur Institute in 1992. He investigated, using both experimental and bioinformatics methods, the structure and function of cerebral nicotinic acetylcholine receptors until 1999. After a post-doc in the team of Dennis Bray at the University of Cambridge, where he worked on the modelling of bacterial chemotaxis, he came back to France as a CNRS research fellow. He is now Group Leader at the European Bioinformatics Institute, the british outstation of the EMBL. He shares his efforts between the modelling of neuronal signalling and the development of tools and services for Computational Systems Biology. Nicolas Le Novere is co-author of 50 scientific publications. He received in 2004 the Jean-Marie Le Goff award, of the French Academy of Science, for his work concerning the bioinformatics analysis of Ligand-Gated Ion channels.

Thomas S. Leyh

is a Professor of Biochemistry at the Albert Einstein College of Medicine (USA). He is deeply interest in all levels of protein function: structure, dynamics, ground- and transition-state structure and energetics, ligand-binding, allostery, the conformational coupling of energetics, and the higher-order organization of catalysis in the cell. His current projects, many of which are structurally grounded, include numerous enzymes that are loosely centered around biomedically relevant issues in sulfur metabolism, isoprenoid biosynthesis and antibiotic development. Dr. Leyh reviews manuscripts for numerous journals, and has been a Member of the Editorial Board of the *Journal of Biological Chemistry*. He has served as a Member of the *Molecular Biochemistry* Study Section at the NSF, and the NIH *Biochemistry* Study Section where he served as Chairman. He is currently a Member of the *Molecular Structure Function A* Study Section at the NIH. He recently spearheaded an NIH workshop on Functional Genomics, which lead to a new NIH-sponsored program. Dr. Leyh a member of the *Strenda Commission*.

Steffen Neumann

studied “Computing in the Natural Sciences” at Bielefeld University, where he focused on Pattern Recognition, Distributed Systems and Bioinformatics, combined with Neurobiology, -psychology and Cybernetics. In 1994/95 he took part in the Erasmus exchange program at Dublin City University (DCU). From 1999 to 2003 he was assistant researcher in the group of Prof. Gerhard Sagerer, where he completed his Ph. D. on Protein Docking. In 2004 Steffen Neumann held a Post Doc Position in the Plant Data Warehouse Group at the Institute of Plant Genetics and Crop Plant Research (IPK) in Gatersleben, before he became head of the Bioinformatics and Mass Spectrometry Group at the Leibniz Institute of Plant Biochemistry (IPB) in Halle.

The Group is developing a Platform for Metabolomics Research, standardisation and exchange formats and integrating data from other -omics fields.

Scott Pegg

Education

University of California, San Francisco, 1996 – 2001
Ph. D. in Pharmaceutical Chemistry, 2001
Area of Specialization: Bioinformatics and Computer-aided Molecular Design

University of California, Berkeley, 1990 – 1995
B. A. in Molecular and Cell Biology & Computer Science, with honors

Research

Current Research: Dept. of Biopharmaceutical Sciences, UCSF

- Computational methods of describing enzyme function, particularly the explicit role of enzyme structure-function relationships.
- Computational methods for the development of biosynthetic routes to small molecules.

Postdoctoral Research: Dept. of Biopharmaceutical Sciences, UCSF 2001 – 2003 (research advisor: Dr. Patricia C. Babbitt).

- Construction of the Structure-Function Linkage Database, providing links between protein sequence, structure, and specific chemical function.

Doctoral Research: Dept. of Pharmaceutical Chemistry, UCSF, 1996 – 2001 (research advisors: Dr. Irwin D. Kuntz and Dr. Patricia C. Babbitt).

- Development of a genetic algorithm for the de-novo design of small molecule ligands.
- Development and analysis of a methodology for detection of remote protein homologies in sequence databases.
- Analysis of docking simulations using protein homology models.

Teaching

Instructor: Bioinformatics Algorithms, U.C. San Francisco, 2001 – present.
Lecturer: Introduction of Bioinformatics, U.C. San Francisco, 2001 – present.

Awards and Honors

Eino Nelson Prize for Graduate Research Achievement, U.C.S.F., 1999
NIH Biotechnology Training Grant, 1997
U.C. Regents Graduate Fellowship, 1996
Computer Science Departmental Achievement Award, 1995
NCAA Student Athlete Award (waterpolo), 1994

Johann Rohwer

is Associate Professor in the Department of Biochemistry at Stellenbosch University, South Africa. He obtained his Ph.D. in 1997 from the University of Amsterdam, working on the control and regulation of the bacterial phosphotransferase system under the supervision of Hans Westerhoff. He then joined Stellenbosch University, where he and his colleagues Jannie Hofmeyr and Jacky Snoep constitute the "Triple-J Group for Molecular Cell Physiology“, a research group that studies the control and regulation of cellular processes using theoretical, numerical and experimental approaches.

Johann has contributed to the theoretical development of metabolic control analysis, to its experimental application, and to the development of software tools for computational systems biology. His main research interests are the construction of kinetic models of cellular function, and the application of NMR spectroscopy to the non-invasive study of metabolism in vivo. He has received the President's Award from the South African National Research Foundation and the Silver Medal of the South African Society of Biochemistry and Molecular Biology.

Together with the other Triple-Js, he chairs the BTK: International Study Group for Systems Biology, and he represents his university on the South African National Bioinformatics Network.

Isabel Rojas

Born in Caracas, Venezuela. She graduated as Licentiate in Computer Science at the Universidad Central de Venezuela (UCV) in 1990 and obtained a Master of Science and a Diploma in Computer Science from the Imperial College, UK in 1993. She did her PhD in computer science at the University of Edinburgh, UK, from 1993 to 1997.

Before joining the EML she worked as a database development consultant for a period of 4 years, managing a group of developers as well as training personnel in multiple companies. She has worked as a lecturer in several computer science disciplines in several Universities and High-Education Institutions.

She leads the Scientific Databases and visualisation group at the EML Research since June 1999. The group mainly works on the development and databases to support the study and analysis of biochemical pathways. The development of user interfaces and visualisation methods for better understanding of the data form also part of the group’s work. Besides these topics the group works on the development of biological ontologies and methods for the extraction of biological information from text and biochemical databases.

Since end of 2004 the group has been working on the development of the SABIO (System for the Analysis of Biochemical Pathways) – Reaction Kinetics (SABIO-RK) database, a web-accessible system setup to support researchers interested in information about biochemical reactions and their kinetics.

Hartmut Schlüter

1981 – 1988: Westfälische-Wilhelms-University, Münster (Chemistry)

1988: Diploma (= M. Sc.) in Biochemistry, Faculty of Chemistry, University of Münster

1991: Ph. D. (Dr. rer. nat.) in Biochemistry, University of Münster, Faculty of Chemistry,
Thesis supervisor: Prof. Dr. H. Witzel

1994: Heinz Maier-Leibnitz prize

1995: Gerhard Hess award (DFG)

1995: Bennigsen-Foerder prize

1991 – 1996: Postdoctoral fellowship at the Medical Faculty of the University of Münster

1996: Habilitation (Dr. rer. nat. habil.) in Pathobiochemistry at the Medical Faculty of the University of Münster

1996 – 2000: Group leader at the Medical Faculty of the Ruhr-University of Bochum

2000-current: Senior Scientist and Head of the Bioanalytical Laboratory of Nephrology, University hospital Benjamin-Franklin, Free University of Berlin,
now: Charité – University Medicine Berlin, Campus Benjamin-Franklin, Joint Facility of the Free University of Berlin and the Humboldt-University of Berlin

2003-current: (apl.) Professor at the Campus Benjamin-Franklin, Free University of Berlin

Dietmar Schomburg

1974: Diplom in Chemistry at the Technical University ”Carolo-Wilhelmina” in Braunschweig

1976: Dr. rer.nat. in Chemistry (Structural Chemistry of Organo-phosphorus compounds)

1985: Habilitation (Dr. rer.nat.habil.) for Structural Chemistry

Scientific Career:

1976 – 1978: Post-Doc in the Chemistry Department at Technical University Braunschweig.

1978 – 1979: Research Fellow at Harvard University in Cambridge, Mass., U. S. A. in Professor W.N. Lipscomb's and Professor F.H. Westheimer's groups.

1979 – 1981: Post-Doctoral Fellow in the Chemistry Department at Braunschweig Technical University

1981 – 1983: Assistant Professor (Hochschulassistent), Braunschweig Technical University

1983 – 1986:	Head of the x-ray lab at the German Centre for Biotechnology – GBF (Gesellschaft für Biotechnologische Forschung), Braunschweig
1987 – 1996:	Head of the GBF Department of "Molecular Structure Research."
1989 – 1995:	Head of CAPE (Center of Applied Protein Engineering)
1990 – 1996:	(apl.) Professor at the Technical University Braunschweig
1996 – 2007:	Full Professor of Biochemistry, University of Cologne
since 2007:	Full Professor of Biochemistry, Technical University of Braunschweig

Jacky Snoep

received his Ph D in 1992 in the fields of microbial physiology and enzymology working on the control of pyruvate catabolism in bacterial systems. He subsequently worked as a postdoctoral fellow, first specializing in molecular techniques to apply control analysis together with Prof. Ingram at the University of Florida and second together with Prof. Westerhoff at the Netherlands Cancer Institute working on theoretical and modelling aspects of biological systems.

Currently Snoep is appointed in Cellular BioInformatics at the Free University of Amsterdam and in Biochemistry at the University of Stellenbosch. He has successfully applied the multidisciplinary approach of combining theory, computer modelling and experiment to understand biological systems to topics as diverse as DNA supercoiling and metabolic engineering of lactic acid bacteria. Since 2001 Snoep has been active in setting up a database for kinetic models that can be interactively run and interrogated over the internet at http://jjj.biochem.sun.ac.za.

Matthias Stein

obtained a degree in chemistry and a Ph D in biophysical chemistry from the Technische Universität Berlin. He investigated the enzymatic mechanism of biological hydrogen conversion by means of magnetic resonance spectroscopy and advanced electronic structure calculations. He also obtained a Master of Science degree in theoretical chemistry from the University of Manchester, UK. After his Ph D he worked as an Administrative Manager of the Collaborative Research Centre (SFB) 498 in Berlin. He did a postdoc at the Royal Institute of Technology in Stockholm. He then spend three years in industry and worked for a biotech company in the area of scientific computing and computer-aided drug design. He is currently a Research Associate at the EML Research gGmbH in Heidelberg and is working on the derivation of kinetic parameters from protein structures for simulations in systems biology. The work presented here is part of the German systems biology initiative within the HepatoSys network.

Keith Tipton

Degrees etc.
B. Sc. (Biochemistry), St Andrews University (1962); M. A. (1965), Ph. D. (1966); Cambridge University; M.R.I.A. (1984)

Main Posts:
University of Cambridge: Demonstrator & Lecturer (1965 – 1977). Fellow of King's College Cambridge (1965 – 1977).
University of Dublin: Professor of Biochemistry (1997 – present).
Fellow of Trinity College, Dublin (1979 – present).
Visiting Professor: Universities of Florence (1976, 1993 & 2003) & Siena (1987 & 1999); Autonomous University of Barcelona (1988 – 89).

Publications:
Over 250 papers in refereed journals; 35 papers as chapters in books; editor of 19 books, > 150 abstracts; 1 patent, co-author of three books.

Research Interests:
Enzymology: regulation, kinetics, inhibition, isolation, applications and classification. Metabolic analysis and simulation. Neurochemistry: depression, degenerative diseases and 'neuroprotection'. Biochemical Pharmacology: drug design, ethanol.

Jan-Olof Winberg

Education:
1982, Cand. real., University of Oslo
1990, Dr. philos., University of Oslo.

Positions/graduate employments:
1983 – 86, Research scholar, Genetic Department, The Norwegian Radium Hospital, Oslo.
1986 – 1993, Senior research officer and head of the Biochemical section, Genetic Department, University Hospital of Northern Norway, 9038 Tromsø.
1993-, Professor, Department of Biochemistry, IMB, MF, University of Tromsø.

Research activities:
Biochemical and kinetic characterization of alcohol dehydrogenases and matrix metalloproteinases.
In vivo, *ex vivo* and *in vitro* expression of matrix metalloproteinases and their tissue inhibitors in diseases such as epidermolysis bullosa and various types of cancer.
Characterization of factors that are involved in the regulation of matrix metalloproteinases in various types of cancer.
Detection of mutations of collagen type VII in patients with the recessive dystrophic form of epidermolysis bullosa.

Author's Index

Index