

Proceedings
of the
1st International Beilstein Workshop
on
EXPERIMENTAL STANDARD CONDITIONS
OF
ENZYME CHARACTERIZATIONS

October, 5th - 8th 2003

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, October 5th - 8th 2003, Rüdesheim, Germany.

Copyright © 2004 Beilstein-Institut zur Förderung der Chemischen Wissenschaften.
Copyright of this compilation by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften. The copyright of specific articles exists with the author(s).

Permission to make digital or hard copies of portions of this work for personal or teaching purposes is granted provided that the copies are not made or distributed for profit or commercial advantage and that copies bear the full citation and copyright notice. To copy otherwise requires prior permission of the publisher.

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: Beilstein-Institut

Cover Illustration: Joelle Heyer
Beilstein GmbH

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 Beilstein Institute organises and sponsors scientific meetings, workshops and seminars, with the aim to catalyse advances in chemical and biological science by facilitating the interdisciplinary exchange and communication of ideas amongst the attendees.

Functional characterization of enzymes and the subsequent computational analysis and modelling of the cellular metabolism and the interaction of cells within tissues and organs led to the foundation of a new branch within the life sciences called systems biology. At the present time, data from metabolic simulations show broad value ranges with high uncertainty because the accessible experimental data have been obviously generated under non-standardized experimental conditions. Successful biological analysis requires, however, comparable and reliable data from both enzyme and physiological interactions collected under standardized experimental conditions. The standardization or recommendation of experimental conditions firstly needs broad discussions within the scientific community, which hopefully will lead to enough common acceptance so that each researcher will carry out his/her experiments in accord with these recommendations.

Participants, as well as speakers were confronted with the following complex questions from experimental and theoretical enzymology:

- are any standards used in the field of functional characterization of enzymes?
- are there any standard procedures or instructions for experimental conditions?
- is it possible to define laboratory procedures for common use?
- do current repositories for enzyme characterization data meet the demands of users?
- which data types for metabolic simulations are necessary?
- are there any demands for the transfer of standardized experimental data to journals or databases?

Over the three days of the workshop, the participants not only heard excellent talks, took part in lively discussions, but in the time between the official sessions of the scientific program, exchanged ideas and thoughts and generally made a valuable and personal contribution to find a way out of the dilemma mentioned above. Whilst this meeting did not find answers to all questions, it succeeded in initiating a dialog between scientists from the different areas of enzymology. One notable outcome is the foundation of the STRENDa commission (see also <http://www.strenda.org>) under the auspices of the Beilstein-Institut.

Preface

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

Frankfurt/Main, October 2004

Martin G. Hicks
Carsten Kettner

CONTENTS

	Page
 Kettner, C. and Hicks, M.G.	
Chaos in the World of Enzymes - How Valid is Functional Characterization without Methodological Experimental Data?	1
 Boyce, S., Tipton, K. and McDonald, A.G.	
Extending Enzyme Classification with Metabolic and Kinetic Data: Some Difficulties to be Resolved	17
 Bock, H.G., Körkel, S., Kostina, E. and Schlöder, J.P.	
Methods of Design of Optimal Experiments with Application to Parameter Estimation in Enzyme Catalytic Processes	45
 Fernie, A.R. and Sweetlove, L.J.	
Broad-Range Metabolite Analysis: Integration into Genomic Programs	71
 Schlüter, H., Jankowski, J., Thieman, A., Rykl, J., Kurzawski, S. and Runge D.	
Determination of Enzyme Activities by Mass Spectrometry - Benefits and Limitations	87
 Holzhütter, H.-G.	
Studying Enzyme Kinetics by Means of Progress-Curve Analysis	99
 Schuster, S. and Zevedei-Oancea, I.	
Multifunctional Enzymes and Pathway Modelling	115
 Snoep, J.L. Olivier, B.G. and Westerhoff, H.V.	
JES Online Cellular Systems Modelling and the Silicon Cell	129
 Degtyarenko, K.	
Controlled Vocabularies and Ontologies in Enzymology	143
 Andreassi, J.L. and Leyh, T.	
Profiles of Molecular Function - Genomic Enzymology	175

**Schomburg, I., Chang, A., Ebeling, C., Huhn, G., Hofmann, O.
and Schomburg, D.**

Experimental Enzyme Data as Presented in BRENDA - a Database for Metabolic
Research, Enzyme Technology and Systems Biology 185

Cammack, R.

Systematic Names for Systems Biology 203

Biographies 215

Index 225

CHAOS IN THE WORLD OF ENZYMES - HOW VALID IS FUNCTIONAL CHARACTERIZATION WITHOUT METHODOLOGICAL EXPERIMENTAL DATA?

CARSTEN KETTNER AND MARTIN G. HICKS

Beilstein-Institut, D-60487 Frankfurt/Main, Germany

E-Mail: ckettner@beilstein-institut.de

Received: 15th April 2004 / Published: 1st October 2004

ABSTRACT

Functional characterization of enzymes plays an essential role in one of the major aims of proteomics research: the modelling of sections of the cellular metabolism with a view to being able to model the whole cellular metabolism and the interaction of cells within tissues and organs. With these purposes in mind, the scientific community established a new branch within the life sciences, called systems biology. However, meaningful modelling by necessity requires comparable and reliable data from standardized enzyme characterizations. From a short, but detailed, investigation of the BRENDA database, it is shown here that the quality of experimental data of enzymes is insufficient for the needs of theoretical biology.

The first step to remedy the situation is to ensure that measurements carried out on enzymes are done so under standard conditions and that all the important information is recorded. With the aim of arriving at an acceptable set of recommendations for experimental conditions, the Beilstein Institut has initiated broad discussions within the scientific community and is further willing to organize and present them as long as appropriate and there is sufficient support.

QUO VADIS, SYSTEMS BIOLOGY?

Continuous advances and improvements have enabled proteome analyses to proceed with increased depth and efficiency. 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 not available. The problem is twofold; on the one hand, deriving data from experimental work is expensive and very time consuming, being a non-trivial process. On the other hand, 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 investigation of metabolic networks, the regulation of developing and developed cells, cell and tissue specification and further highly complex cellular processes are the central aims of systems biology. Researchers in this discipline use the rapidly grown biological databases to create biological models and simulations as well to develop further powerful algorithms which have made it possible to explore many cellular physiological functions. Systems biology intends to reconstruct the cellular metabolism as a series of overlapping mathematical models. By using all the theoretical and experimental achievements of the various genome and proteomes projects, the data can then be analysed by computational, mathematical and engineering methods in order that predictive models of single biochemical processes, cells, tissues or even entire organs can be generated [1]. Furthermore, a model that works well will be useful for designing initial or further experiments that will verify or refute both working hypothesis of physiological functions of certain modelled systems and the predictions previously carried out. The integration of experimental and mathematical methods in biology also requires that enzymologists make a conceptual shift; enzyme systems do not reach equilibrium even when they seem stationary - the system is in constant flux. Traditionally, functional characterization of enzymes includes the determination of V_{\max} (maximum velocity of an enzyme reaction) and K_m to help define the kinetic behaviour of a given enzyme. However, these parameters are often either unknown or extrapolated from different organisms or species.

Furthermore, traditional enzymology uses enzymatic assays under steady state conditions, so that researchers record the turnover of large amounts of substrate molecules by multiple enzyme molecules. Technological advances now allow enzymologists to capture the kinetics of single enzyme molecules by using a variety and combination of methods and tools such as classical biochemical and biophysical techniques, i.e. fluorescence microscopy, FRET, electrophysiology etc.. However, using these disjunct sets of information, often measured under different conditions, it is not possible to reach the goal of understanding complete pathways. Thus, enzymology should move from the relatively simplistic determination of single enzyme kinetics towards non-equilibrium thermodynamics of populations of enzymes in metabolic contexts. In addition, systems biology can only work successfully if the theoretical part is continuously fed with a large quantity of highly qualitative experimental data.

INTEGRATION OF EXPERIMENTAL AND COMPUTATIONAL BIOLOGY

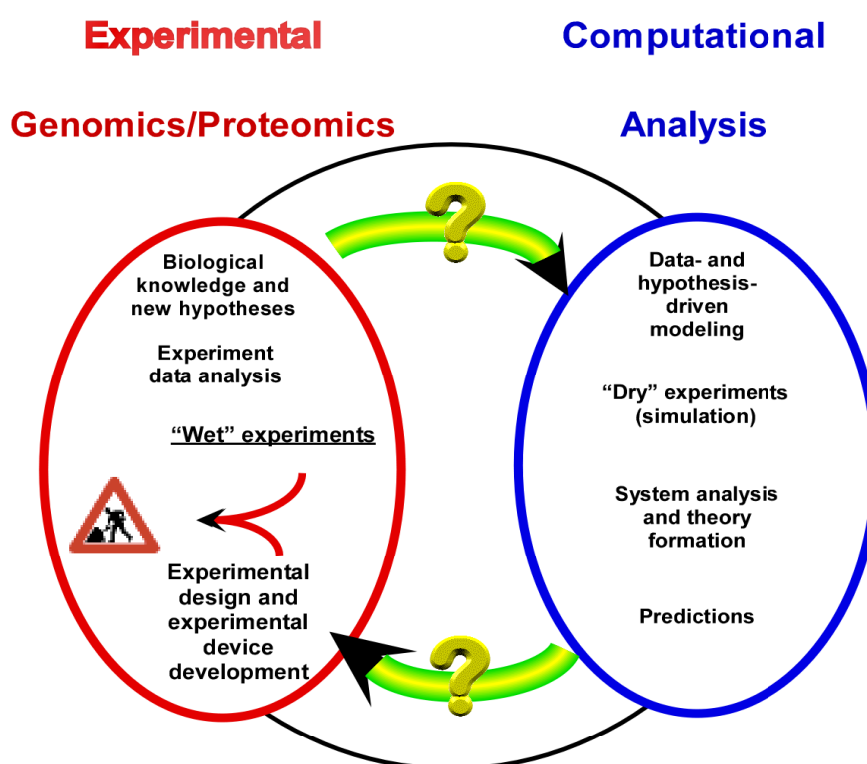


Figure 1. Hypothesis-driven research in modern biology whose one representative discipline is systems biology (adapted from Kitano, H. (2002). *Science* **295**:1662-1664).

The research cycle in modern biology is illustrated in Fig. 1. The cycle can commence either with data- or hypothesis-driven modelling, or with experimental design related to a specific scientific problem.

The models are determined by a computable set of assumptions and hypotheses, which must be tested or confirmed experimentally. The validity of assumptions and hypotheses embedded in each model is proved by "dry" experiments, such as simulations. Inadequate models that show inconsistencies with established experimental facts will be rejected or modified. Those models that pass this test undergo system analysis resulting in a number of predictions which themselves undergo test procedures in "wet" experiments. Data from successful experiments either verify or show the inadequacies of the computational models and enter into the pool of biological knowledge that is source for new hypotheses. By contrast, analysed data from "wet" experiments pass also into this pool of biological knowledge and can be subjected to modelling and simulation. Successful modelling and computational "dry" experiments require strictly comparable data of high quality.

THE COMPREHENSIVE VIEW ON ENZYMES

There is no doubt that proteins and, especially, enzymes, are the most important elements of biochemical processes. Thus, comprehensive characterization of enzymes provides a series of requirements for unambiguous description:

- i) proper identification of the protein as a product of a single or several different genes,
- ii) proper determination of the structure,
- iii) proper description of the enzymatic reaction and classification of the enzyme according to the Enzyme Nomenclature (see also the articles by Richard Cammack, by Kirill Degtyarenko, and by Keith Tipton et al. in this book), and
- iv) identification of the catalysing mechanisms from a consideration of the physicochemical properties of the biochemically active sites in the enzyme.

These active sites include both those catalysing binding pocket(s) for the metabolic reaction and binding pockets for inhibitors and/or activators. The results of this research lead to a comprehensive picture of the function-structure relationships of an enzyme. Numerical data about these relationships for a whole series of catalysing enzymes within a physiological pathway would provide systems biology with the opportunity to model, simulate or calculate this pathway under different known physiological and non-physiological conditions.

Functional data of enzymes include measurements of the catalysing behaviour which is dependent upon pH, temperature, ionic strength, inhibiting and activating compounds, substrate specificity, etc.. These are usually numerical data. They are required to describe, for example, the kinetics of a given enzyme and, subsequently, of entire pathways.

However, if a number of enzymes in a given pathway is investigated under (at least) comparable experimental conditions, this data will also be comparable and will be suitable to feed further steps of analysis. In general, further analysis means *in silico*, i.e. computer-based, investigation of the kinetics of branches or entire pathways. Results of these analyses reflect the metabolism of biological compounds under "normal" and stressing physiological conditions. These analyses are certainly core topics of systems biology. However, researchers encounter increasing numbers of collections of data on enzyme characterizations, all of which should be used cautiously. If functional data lacks proper comparability, sound numerical analysis of pathways, cells, tissues or even entire organs, will not be possible.

Apart from metabolic networks mentioned above, systems biology also attempts to understand cellular networks of molecular interactions. When such a network has been established together with gene expression profiles, it is possible to explain and predict both interactions (and also generate hypotheses which have to be proven) that regulate the observable expression dynamics and why and when a gene is turned on and off in response to the state of this network [2]. It should then be possible for systems biology to not only depict the cellular metabolic pathways, such as those in the well-known Boehringer poster, but to do this in three dimensions with a higher level of information than for example the KEGG pathway map. The application of these digitized maps may be found in the understanding and simulation of the treatment of diseases such as diabetes. The results can be used for the development of new "intelligent" drugs [3].

FIRST STEPS TOWARDS MODERN BIOLOGY

It may be of interest to note here that a series of projects to simulate entire cells, such as the e-cell project led by Masaru Tomita [4], are currently in their initial phases. The *Escherichia coli* cell simulation used by this group represents a hybrid model and includes both quantitative kinetic and qualitative stoichiometric data. The last data type had to be used due to the lack of available kinetic data. Another group led by Jacky Snoep and Hans Westerhof [5] has established The Silicon Cell project to model parts of metabolic pathways.

Part of which allows the simulation of some of the common metabolic pathways via the Internet with a Java-based program (see also Jacky Snoep's article in this book). In this project, the researchers were hindered through the lack of kinetic data. The group is now having to re-determine the required enzymatic reaction data under internal standardized experimental conditions. Some other functional data are retrievable from KEGG [6] or EcoCyc [7].

The intention here is to show the necessity for the standardization of experimental conditions on the basis of one example only. We present some enzymological and methodological data for the key enzyme involved in the glucose degradation pathway, called glycolysis, also known as the Fructose-1,6-bisphosphate pathway, or Embden-Meyerhof-Parnas pathway.

THE WAY INTO THE LABYRINTH OF SYSTEMATIC DISORDER

The Cologne BRENDA database was chosen as the preferred data repository for the investigation of enzymological data and associated experimental conditions. BRENDA is the acronym for "Braunschweig Enzyme Database". This database was developed at, and is maintained by, Dietmar Schomburg's group at the University of Cologne and offers an exceptional collection of functional enzyme data [8] (see also Dietmar Schomburg's contribution in this book). In contrast to the databases SwissProt [9], PIR [10] or PDB [11] which provide predominantly information on protein sequences and structures, this database covers a wide range of functional enzyme characteristics. It contains approximately half a million entries in more than 50 fields that can be searched in various combinations for information on about 22,000 enzymes with some 3500 "different" enzymes. The data on enzyme function for this database are extracted directly from the primary literature by graduate biologists and chemists.

This enzyme database was used to evaluate the quality and quantity of the functional enzyme data of glycolysis (Embden-Meyerhof-Pathway) (Fig. 2). This pathway was selected for our evaluation because it is almost certainly one of the best-understood metabolic pathways within the cell where it plays a key role in the degradation of sugar and the subsequent provision of chemical energy for the cellular catabolism. Furthermore, it was expected that for glycolysis the best information about the enzymes involved with respect to their functional characteristics in most organisms would be available. To avoid biochemically less interesting enzymes and obtaining poor data, three key enzymes from the glycolysis were selected: glucokinase (GK, EC 2.7.1.1), 6-phosphofructokinase (6-PF, EC 2.7.1.11) and pyruvate kinase (PK, EC 2.7.1.40),

from baker's yeast *Saccharomyces cerevisiae* and *Escherichia coli* as well as from *Bacillus stearothermophilus* and the slime fungus *Dictyostelium discoideum*.

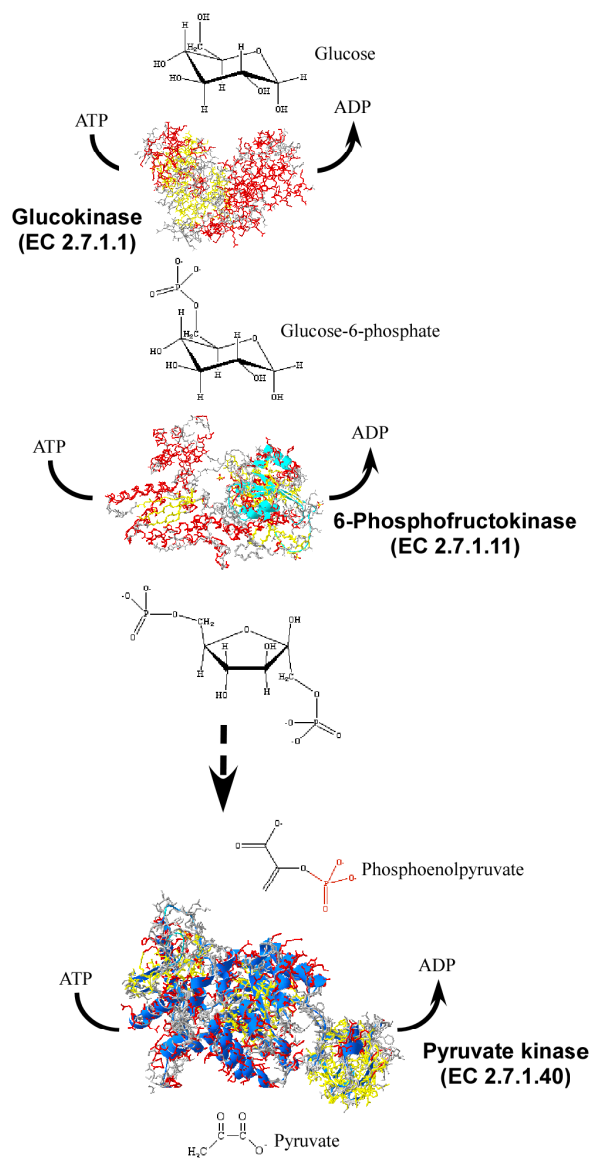


Figure 2. Glycolysis: key enzymes and their substrates [19, 20].

In our case, the main criteria for the functional description of these enzymes were data on the turnover kinetics such as activity and Michaelis constant (K_M), information about activating and/or inhibiting compounds and molecules such as cofactors, allosterically acting compounds as well as ions. Additionally, we were interested in temperature and pH profiles which give information on the maximum activity of the enzyme at a given temperature or pH.

After this information had been collated for all three enzymes and all four organisms, the experimental conditions for some of this information were examined using the original literature, firstly to study the reasons for the differences in the data set within one descriptive criterion and secondly, to see if the data was comparable and suitable for modelling and simulation.

It should be noted that, since this investigation was carried out in mid 2003, BRENDA has continued to increase in both content and in data structure, it is therefore possible that there will be some inconsistencies between our results and the current information available from BRENDA.

IN THE MIDDLE OF THE LABYRINTH

The functional data of the three key enzymes of the glycolysis of the four different organisms, two prokaryotes and two eukaryotes, were reviewed. Examples of the experimental methods and conditions which led to the functional characteristics of the enzymes, as described in the literature cited in BRENDA, have also been collected.

The investigation of the functional characteristics of the three key enzymes of the glycolysis gave the results shown in Table 1. The study on the functional enzyme data was commenced with queries for the yeast *Saccharomyces cerevisiae* and the coliform bacterium *E. coli*. But since the retrieval results from BRENDA were zero for glucokinase and low for pyruvate kinase, it was necessary to expand the investigations to two further biochemically well-known organisms: *Bacillus stearothermophilus* and the slime fungus *Dictyostelium discoideum*.

As can be seen from Table 1, even the results for yeast and *E. coli* were surprising since these two organisms appear to have been well studied by both biochemical and molecular biological methods. Yeast plays an important economic role, its genome was completely sequenced eight years ago and since then functional and structural proteomics has made great strides [12-15]. On the other hand, *E. coli* is the main organism in expression studies and acts as the main transformation vector [16-18]. Thus, comprehensive data on their fundamental metabolic pathway was expected. The best investigated enzyme of both organisms seems to be 6-phosphofructokinase (6-PF).

Functional data for 6-PF can be obtained from several publications, it being the most important enzyme in glycolysis. However, there are many missing data for pyruvate kinase for both organisms: for example, data on the inhibitors, temperature and pH range are missing for *E. coli* as well as for *S. cerevisiae*. Kinetic data, such as K_M value and specific activity, for *E. coli* were completely absent. The best data for all three enzymes were found for *B. stearothermophilus* which is a prominent member of extremophile organisms. All criteria for the functional description of the glycolytic key enzymes seem to be fulfilled. Finally, the slime fungus *D. discoideum* provided the least amount of data; there are only a few functional data available for the first step of the glucose degradation carried out by the enzyme glucokinase.

Chaos in the World of Enzymes

Table 1. Comparison of the functional characteristics of the glycolytic key enzymes from four different organisms.

Enzyme / EC-Nr.		Glucokinases			6-Phosphofructokinase 2.7.1.11			Pyruvatekinase 2.7.1.40			
Parameter	Bacillus spec.	S. c.	E. c.	Dictyostelium d.	Bacillus spec.	S. cerevisiae	E. coli	Bacillus spec.	S. cerevisiae	E. coli	D. d.
K _m Value [mM]	0.15; 0.06 (ATP), 0.52 (gluc)			0.12 (gluc), 1.1 (ATP)	0.0392 (F6P), 0.055 (ATP)			0.2 (PEP), 0.7 (ADP)	0.049 (IDP), 0.16 (ADP)		
Substrates	ATP + D-glucose			ATP + D-glucose	ATP + D-Fruc 6-P	ATP + D-Fruc 6-P (Glu-1-P, SedHep-7-P)	ATP + D-Fruc 6-P	ADP + PEP	ADP + PEP	ADP + PEP	
Products	ADP + D-gluc 6-P			ADP + D-gluc 6-P	ADP + D-Fruc 1,6 BisP	ADP + D-Fruc 1,6 BisP	ADP + D-Fruc 1,6 BisP	ATP + Pyruvat	ATP + Pyruvat	ATP + Pyruvat	
Cofactors					ADP, AMP		GDP	AMP (allosteric), GMP (activation), CMP (activation), GDP (activation)		AMP (PKII)	
Activating Compound					Fruc 2,6-BisP			Gluc-6-P, Rib-5-P, 3-P-Glyc		Fruc 1,6 DiP (PKI), Gluc 6-P, Rib 5-P (PKII)	
Inhibitors	N-Acetyl-alpha-D-glucosamine, D-Maltose, AgNO ₃ , PCMB, HgCl ₂ , Pb(NO ₃) ₂ , 1,10-Phenanthroline, Iodoacetamide, N-Ethylmaleimide, p-Hydroxymercuribenzoate, Glucose 6-phosphate			Glucose 6-phosphate, ADP	Citrate, Diphosphate, Phosphoenolpyruvate (most potent inhibitor), F6P, MgATP ²⁻ , F1,6BisP		Citrate, ATP, MgATP ²⁻	Mg ²⁺ , ADP, PO ₄ ³⁻ , Ca ²⁺ , Cu ²⁺ , Ni ²⁺ , Sr ²⁺		ATP, Succinyl-CoA (PKI)	
MW [Da]	87,000/67,000				135,000	835,000	142,000	242,000	209,400	190,000 (PKII), 225,000 (PKI)	
Metals/Ions	Mg2+ (required); Mn2+ (45% of the activity with Mg2+); Co2+(68% of the activity with Mg2+)			Mg ²⁺ (completely dependent on presence of Mg ²⁺)	Mg ²⁺ , K ⁺ , NH ₄ ⁺ , Li ⁺	Mg ²⁺ , K ⁺ , NH ₄ ⁺ , PO ₄ ³⁻	Mg ²⁺ , MgATP, Mn ²⁺ , MnATP ²⁻ , K ⁺ , NH ₄ ⁺	Mn ²⁺ , Mg ²⁺ , Co ²⁺ , K ⁺ , NH ₄ ⁺ , Na ⁺ , Cu ²⁺ , Ni ²⁺ , Sr ²⁺	Mg ²⁺ , K ⁺		
Specific Activity [μM/min/mg]	334/304			0.51	160	148	215, 263, PFK1: 190; PFK2: 205	210, 333	340, 250, 219		
Temperature Optimum [°C]	30/37 (assay at), 60			20 (assay at)	30 (assay at)	25 (assay at)	27	30 (assay at)	30 (assay at)	25 (assay at)	
Temperature Stability [°C]	60/70 (10% loss)				30 (reverses inhibition by Mg ²⁺ , ATP and PEP), 30 (assay at)		(PFK1): 6.5 with two optima at 6.5 + 8.5; (PFK2): 8.5 with two	25 (in 25 mM Tris pH 7.5); 65 (rapid inactivation);			
pH Optimum	9			7	8, 2 (8, 7)	7		6.8 (60°C), 7.2 (30°C)			
pH Range	6, 10.5 (max)			6, 9 (max)				5.5, 7.4 (max) at 60°C; 6.4, 8.2 (max) at 30°C			
Literature	Goward et al., 1968, Biochem. J.	Baumann, 1969, Biochemistry			Marschke et al., 1982, Methods Enzymol.	Hofmann et al., 1982, Methods Enzymol.	Kotlarz et al., 1982, Methods Enzymol.	Sakai et al., 1986, J. Biochem.	Aust et al., 1975, Methods Enzymol.	Malcovati & Valentini, 1982, Methods Enzymol.	
	Hengartner et al., 1973, FEBS Lett.				Shirakihara et al., 1988, J. Mol. Biol.	Stellwagen, 1975, Methods Enzymol.	Kernerer et al., 1975, Methods Enzymol.	Tuominen & Bernlöhr, 1975, Methods Enzymol.	Hunsley & Suelter, 1969, J. biol. Chem.		
					Byrnes et al., 1994, Biochemistry	Kruger et al., 1988, Arch. Biochem. Biophys.					

What might the reasons be for this lack of data?

Data available from the SwissProt, KEGG, PDB and PIR databases indicate that the listed enzymes of the four organisms mentioned above have been well investigated with respect to their structures and sequences. Queries within these databases give a comprehensive collection of data on protein identification, subunit composition and stoichiometry as well as crystallographic data, information about isolation and storage of purified proteins. But comments on the function of these proteins are extremely short. Furthermore, the functional enzyme data found in special enzyme databases such as BRENDA are fragmentary and some enzymes lack any functional information. This was a surprising and alarming result of our short study. On the basis of such poor data availability, it is hard to imagine that metabolic simulation and modelling could be carried out successfully.

The next step in our evaluation was to obtain information from the original research literature on the experimental conditions. An investigation of the material & methods sections of the appropriate publications which describe among others, the functional characterization of a given enzyme led to the analysis shown in Table 2.

At first glance, the functional data of each enzyme of all the organisms appear to have been obtained by comparable methods: the coupled optical test and/or the pH stat assay. The first noted method was carried out by means of different sets of auxiliary enzymes depending on the glycolytic enzyme studied to record the rate of NADH oxidation. The second method records the flow of protons by pH-sensitive electrodes. However, the decisive differences within the applied methods are the basic experimental conditions. Measurements were performed under different temperatures (ranging from 25°C to 37°C), also apparently different wavelengths to record NADH oxidation (which might be less critical) were used, and, finally, the assay buffers had different compositions. In particular, for the assay buffers, there is a wide range of simple compositions (e.g., for pyruvate kinase of *E. coli*) and rather "complicated" compositions (e.g., for pyruvate kinase of *S. cerevisiae*) with respect to the number and types of compounds. Investigation of enzyme function is supposed to be carried out under the simplest conditions to avoid side effects that would lead to changes in enzyme reactions which are hard to interpret. In conclusion, from the standpoint of single methodological aspects, such as temperature or pH, it is generally conspicuous that the laboratory experimental conditions chosen to determine kinetic key parameters such as activity or K_M lead to values that are hardly comparable for all three enzymes.

Chaos in the World of Enzymes

Michaelis-Menten kinetics and specific activity both strongly depend on these basic parameters mentioned above.

The above noted fundamental differences in the application of obviously commonly used methods confirmed the necessity to standardize experimental conditions for enzyme characterizations.

Table 2. Analysis of methods and experimental conditions, obtained from literature.

A) for yeast, i.e. *Saccharomyces cerevisiae*. Differences in the methods of yeast fractionation, protein isolation and purification.

Enzyme	Phosphofructokinase (EC 2.7.1.11)	Pyruvatekinase (EC 2.7.1.40)
Reaction	Fruc-6-P + ATP --> Fruc-1,6-bis-P + ADP + H ⁺	PEP + H ⁺ + ADP --> ATP + Pyruvate
Principle	coupled optical test with aldolase, triose-phosphate isomerase and glycerol-3-phosphate dehydrogenase (DH)	uptake of protons with a pH stat, decrease in absorption at 230 nm due to loss of PEP or coupled optic test with lactate DH
Product formation	Fruc-1,6-bis-P formation indicated by NADH dependent reduction of dihydroxy-acetone-P to glycerol-3-P	reduction of pyruvate, oxidation of NADH
Optimization	addition of AMP as strong allosteric activator	
Reagents	<u>Assay mixture:</u> Imidazole/HCl, 100 mM, pH 7.3, Fruc-6-P, 3 mM, ATP, 0.6 mM, pH 7.2, MgSO ₄ , 5 mM, (NH ₄) ₂ SO ₄ , 5 mM, AMP, 1 mM, NADH, 0.2 mM, <u>Auxiliary enzymes:</u> Imidazole/HCl, 100 mM, pH 7.2, Aldolase, 14 u/ml, Triosephosphate isomerase, 136 u/ml, Glycerol-3-P-DH, 12 u/ml <u>Buffer mixture:</u> Imidazole/HCl, 100 ml, pH 7.2 Fruc-6-P, 1 mM 2-Mercaptoethanol, 5 mM PMSF, 0.5 mM	<u>Assay mixture:</u> (CH ₃) ₄ N cacodylate, 100 μM, pH 6.2, MgCl ₂ , 24 μM, KCl, 100 μM, Tricyclohexylammonium FDP, 1 μM NADH, 0.16 (0.15) μM, <u>Auxiliary enzymes:</u> Lactate DH, 33 μg (3 mg/ml in 200 mM Tris-HCl, pH 7.5)
Measurement	NADH oxidation at 340 nm and 25°C against water	NADH oxidation at 230 (?) (340) nm and 30°C against water
Literature	- Hofmann & Kopperschlaeger (1982). <i>Methods Enzymol.</i>	- Aus <i>et al.</i> (1975). <i>Methods Enzymol.</i> - Hunsley & Suelter (1969). <i>J. biol. Chem.</i>

Kettner, C. & Hicks, M.G.

B) for *Escherichia coli*

Enzyme	Phosphofructokinase (EC 2.7.1.11)	Pyruvatekinase (EC 2.7.1.40)
Reaction	Fruc-6-P + ATP --> Fruc-1,6-bis-P + ADP + H ⁺	PEP + H ⁺ + ADP --> ATP + Pyruvate
Principle	coupled optical test using aldolase, triose-phosphate isomerase, glycerol-3-P-DH; uptake of protons with pH stat	direct: proton uptake with pH stat; coupled optical test using lactate DH
Product formation	conversion of Fruc-1,6-bis-P to glycerol-3-P; Oxidation of 2 µM NADH per µM Fruc-1,6-bis-P formed	reduction of pyruvate, oxidation of NADH
Optimization		
Reagents	<u>Assay mixture:</u> Tris-Cl, 80 mM, pH 8.2, MgCl ₂ , 1 mM, Fruc-6-P, 6.7 mM, <u>Auxiliary enzyme mixture:</u> Tris-Cl, 10 mM, EDTA, 2mM, Aldolase, 1.5 mg/ml, Glycerol-3-P-DH, 0.5 mg/ml, NADH, 0.15 mM, ATP, 1 mM, pH 8, <u>pH stat assay:</u> KCl, 80 mM, MgCl ₂ , 10 x [ATP], ATP, variable, Fruc-6-P, variable, pH 8.5	<u>Assay mixture:</u> Tris or HEPES, 10 mM, pH 7.5, MgCl ₂ , 10 mM, KCl, 50 mM, ADP, 2 mM, PEP, 10 mM, <u>Auxiliary enzyme mixture:</u> Lactate-DH, 22 u/ml, NADH, 0.12 mM, PK, 0.03 - 0.1 u/ml
Measurement	oxidation of NADH at 340 nm and 29°C against water; change of [H ⁺] recorded by calomel and glass electrode at pH 8.5	oxidation of NADH at 340 nm at 25°C against water
Literature	- Kemerer et al. (1975). <i>Methods Enzymol.</i>	- Malcovati & Valentini (1982). <i>Methods Enzymol.</i>

Chaos in the World of Enzymes

C) for *Bacillus stearothermophilus*.

Enzyme	Glucokinase (EC 2.7.1.1)	Phosphofructokinase (EC 2.7.1.11)	Pyruvatekinase (EC 2.7.1.40)
Reaction	Glucose + ATP --> Gluc-6-P + ADP	Fruc-6-P + ATP --> Fruc-1,6-bis-P + ADP + H ⁺	PEP + H ⁺ + ADP --> ATP + Pyruvate
Principle	coupled optical test using Gluc-6-P-DH	coupled optical test with aldolase, triosephosphate isomerase and glycerol-3-phosphate-DH	coupled optical test using PEP and Lactic-DH
Product formation	Gluc-6-P formation and oxidation of NADPH	Fruc-1,6-bis-P formation indicated by NADH dependent reduction of dihydroxyacetone-P to glycerol-3-P; 1 mol Fruc-1,6-bis-P --> Oxidation of 2 mol NADH	Lactate formation and oxidation of NADH
Reagents	<u>Reaction mixture:</u> Tris/HCl, 50 mM, pH 9, Glucose, 10 mM, ATP, 2 mM, MgCl ₂ , 3 mM, NADP ⁺ , 1 mM, Gluc-6-P-DH, 5 µg/ml	<u>Assay mixture (am):</u> Imidazole-HCl, 20 mM, pH 6,82, KCl, 100 mM, MgCl ₂ , 3 mM, β-mercapto-EtOH, 1 mM <u>Auxiliary enzymes (ae):</u> BSA, 1.5 mg, Aldolase, 4 mg, Triosephosphate isomerase, 32 u/ml, Glycerol-3-P-DH, 8 u/ml --> dilution of enzyme solution in 2.5 ml aq.dest.; 0.1 ml of ae was used per 1 ml am	<u>Assay mixture:</u> Imidazole-HCl, 25 µM, pH 7.2 (50 µM, pH 7.1), KCl, 25 µM (50 µM), MgCl ₂ , 3.5 µM (7 µM), ADP, 1 µM (NaADP, 4 µM) NADH, 0.06 µM (0.12 µM) <u>Auxiliary enzymes:</u> PEP, 2 µM, Lactic-DH, 10 µM (40µM)
Measurement	NADPH oxidation at 340 nm and 37°C against water	NADH oxidation at 340 nm and 30°C against water	NADH oxidation at 230 (340) nm and 30°C against water
Literature	- Hengartner & Zuber (1973). <i>FEBS Lett.</i>	- Marschke & Bernlohr (1973). <i>Methods Enzymol.</i>	- Sakai et al. (1986). <i>J. Biochem.</i> - Tuominen & Bernlohr (1975). <i>Methods Enzymol.</i> - Bücher & Pfeleiderer (1955). <i>Methods Enzymol.</i>

ESCAPING FROM ENZYMOLOGICAL CONFUSION?

Much of the functional data of enzymes have been generated under non-standardized experimental conditions. Moreover, these data are usually determined by individual laboratory-specific application and implementation of the experimental design. Consequently, the chances of success for systems biology to escape from the verbally overused *-omics*-sciences are poor, as long as the quality of the input as well as the out-going modelling data cannot be improved. Furthermore, from the point of view of the applications of systems biology, the importance of reliable experimental data is without question.

In conclusion, there is a definite requirement for standardization of the experimental conditions mentioned above. The current situation is approaching chaos in that functional characterization is being carried out without methodological experimental data. Of course, such standardization, firstly needs broad discussions within the scientific community which hopefully will lead to some common acceptance so that each researcher will carry out his/her experiments accordingly. Moreover, it is necessary to improve the integration of mathematical biology with experimental biology to offer experimentalists the opportunity to learn the language of mathematics and dynamical modelling and theorists to learn the language of biology. Both practitioners and theorists are invited to participate in the discussions of the requirements needed for the successful development of systems biology.

The Beilstein-Institut will support this process of discussion and standardization process by hosting:

- a) the 1st ESCEC symposium,
- b) the discussion of suggestions for standardizing laboratory conditions, and
- c) the ESCEC commission which will consist of representatives of several diverse directions of enzymology (see also www.strenda.org).

ACKNOWLEDGMENTS

Many thanks to Dr. Allan D. Dunn, Beilstein-GmbH, Frankfurt, Germany, for carefully proofreading, as well as for helpful advice and discussions.

REFERENCES

- [1] Noble, D. (2002) Modeling the heart-from genes to cells to the whole organ. *Science* **295**:1678-1682.
 - [2] Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Caestani, C., Yuh, C., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Lee, P.Y., Revilla, R., Rust, A.G., Pan, Z., jun., Schilstra, M.J., Clarke, P.J.C., Arnone, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L., Bolouri, H. (2002) A genomic regulatory network for development. *Science* **295**:1669-1678.
 - [3] Werner, E. (2002) Systems biology: the new darling of drug discovery? *Drug Discov. Today* **7**(18): 947 - 949.
 - [4] www.e-cell.org , Keio University, Japan.
 - [5] www.siliconcell.net, Universities of Stellenbosch, South Africa, and Amsterdam, The Netherlands.
 - [6] Kyoto Encyclopaedia of Genes and Genomes, www.genome.ad.jp/kegg.
 - [7] Encyclopaedia of Escherichia coli K12 Genes and Metabolism", www.ecocyc.org.
 - [8] BRENDA, www.brenda.uni-koeln.de
 - [9] SwissProt Protein knowledgebase TrEMBL Computer-annotated supplement to SwissProt, <http://au.expasy.org/sprot/>.
 - [10] PIR, Protein Information Resource, <http://pir.georgetown.edu/>.
 - [11] PDB, Protein Data Bank, <http://www.rcsb.org/pdb/>.
 - [12] Guilliermond, A. (1920) The Yeasts. J. Wiley, New York.
 - [13] Beech, M.J., Davenport, R.R. (1971) Isolation, purification and maintenance of yeasts. In: *Methods in Microbiology*. Vol. 4 (Booth, C. Ed.), pp. 153-182. Academic Press, London.
 - [14] Bassett, D.E., Basarai, M.A., Connelly, C., Hyland, K.M., Kitagawa, K., Mayer, M.L., Morrow, D.M., Page, A.M., Resto, V.A., Skibbens, R.V., Hieter, P. (1996) Exploiting the complete yeast genome sequence. *Curr. Opin. Genet. Devel.* **6**:763-766.
 - [15] Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G. (1996) Life with 6000 genes. *Science* **274**: 546, 563-567.
 - [16] Blattner, F.R., Plunkett, G. 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**: 1453-1462.
-

- [17] Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M., Shinagawa, H. (2001) Complete genome sequence of Enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with laboratory strain K-12. *DNA Res.* **8**: 11-22, 47-52.
 - [18] Lee, P.S., Lee, K.H. (2003) *Escherichia coli* - a model system that benefits from and contributes to the evolution of proteomics. *Biotechnol. Bioenerg.* **84**(7): 801-814.
 - [19] Schwede, T., Kopp, J., Guex, N., Peitsch, M.C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucl. Acids Res.* **31**: 3381-3385.
 - [20] Guex, N., Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. *Electrophoresis* **18**:2714-2723.
-

EXTENDING ENZYME CLASSIFICATION WITH METABOLIC AND KINETIC DATA: SOME DIFFICULTIES TO BE RESOLVED

SINEAD BOYCE, KEITH TIPTON* AND ANDREW G. McDONALD

Department of Biochemistry, Trinity College, Dublin 2, Ireland

E-Mail: *ktipton@tcd.ie

Received: 13th January 2004 / Published: 1st October 2004

ABSTRACT

Classification of enzymes according to the reaction(s) catalysed is a relatively straightforward procedure, as it deals with more-or-less factual data. However, attempting to add meaning to those data by adding metabolic or kinetic information takes one into the field of parameters rather than absolutes. Thermodynamic data have been assembled for a number of reactions, but the direction in which a reaction is favoured in isolation does not necessarily mean that that will be the direction of the reaction in cellular metabolism; there are many metabolic examples of enzyme reactions proceeding in the thermodynamically less-favoured direction. Attempts to predict "missing enzymes" from metabolic pathways should also be treated with caution, since there are several cases where such guesses have proven to be wide of the mark. Incorporation of kinetic data requires the definition of standard conditions, which should ideally bear some relevance to the physiological situation in which the enzyme operates. However, not all enzymes operate under the same physiological conditions and there are, as yet, no universally accepted standard conditions, or sets of conditions, of temperature, pH, ionic strength etc. for the collection of such data.

INTRODUCTION

The International Union of Biochemistry and Molecular Biology (IUBMB) Enzyme List classifies enzymes in terms of the reactions they catalyse (see [1, 2] for definitive versions). It is restricted to classification and recommendations on nomenclature. As such, the data contained within it are, as far as possible, strictly factual and should provide a system for the unambiguous identification of the enzyme(s) being studied. Thus, it should provide a solid basis for the incorporation of data on enzyme behaviour, compartmentation etc. to facilitate studies on the behaviour of biological systems. However, the value of such reconstructive approaches will depend on the quality of the data incorporated.

This brief account will consider the structure of the Enzyme List (see [3,4] for fuller details) and its limitations. Then some of the problems and pitfalls in attempts to assign metabolic and kinetic data will be considered with particular reference to the wide variations in assay conditions that have been used in the literature. The examples used are far from comprehensive and are largely drawn from systems with which we are familiar from our own work.

PRINCIPLES OF ENZYME CLASSIFICATION

Unlike many other classification systems, where classification is based on structure or function, enzymes are classified according to the reactions they catalyse, with each enzyme being assigned a four-digit number, called the EC (Enzyme Commission) number. An EC number takes the form w.x.y.z, where w, x, y and z represent the class, subclass, sub-subclass and serial number, respectively. At present, there are six enzyme classes, each of which covers a different type of reaction, as summarized in Table 1.

The subclass normally provides information about the type of compound or group involved.

For example, in EC 1.x.-., the subclass number, x, indicates the group oxidized, with 1 indicating a CH-OH group, 2 an aldehyde or oxo group, 3 a CH-CH group, 4 a CH-NH₂ group etc.

The sub-subclass further specifies the type of reaction involved, often the "other" substrate.

For example, in EC 1.-.y.-, the sub-subclass (y) provides information on the group reduced, with 1 indicating NAD(P)⁺, 2, a cytochrome, 3, O₂, 4, S-S...99, others. The forth digit, z, is a serial number that identifies individual enzymes within a sub-subclass.

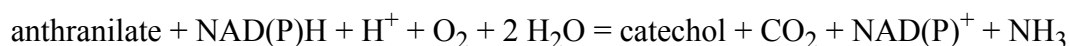
 Extending Enzyme Classification with Metabolic and Kinetic Data

Table 1. Enzyme classes

Class	Name	Reaction catalysed
1	Oxidoreductases	$AH_2 + B = A + BH_2$ or $AH_2 + 2 B^+ = A + 2 B + 2 H^+$
2	Transferases	$AX + B = A + BX$
3	Hydrolases	$AB + H_2O = AH + BOH$
4	Lyases	$A=B + X-Y = A-B$ X Y
5	Isomerases	$A = B$
6	Ligases	$A + B + NTP^* = A-B + NDP + P$ or $A + B + NTP = A-B + NMP + PP$

*NTP = nucleoside triphosphate

Sometimes an enzyme might fit into more than one class, e.g. EC 1.14.12.1, anthranilate 1,2-dioxygenase (deaminating, decarboxylating), which catalyses the reaction:



might also be classed among the deaminases (EC 3.5.-) or the decarboxylases (EC 4.1.1.-). In such cases, the general rule is that the lower EC class number takes precedence.

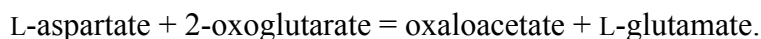
Entries within the Enzyme List have a standardized format, although not all enzymes will contain each of the fields described below.

Common name

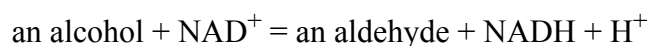
This is often the most commonly used name for the enzyme, provided that it is neither misleading nor ambiguous. Some generic words indicating reaction type are used in common names, but not in the systematic names, e.g. dehydrogenase, reductase, oxidase, peroxidase, kinase, tautomerase, deaminase and dehydratase. The common name may indicate the direction in which the reaction is perceived to operate (either thermodynamically or in vivo), e.g., use of the term 'reductase' in the common name indicates that the reaction occurs in the opposite direction to that written (see Reaction).

Reaction

This shows the actual reaction catalysed, written, where possible, in the form of a 'biochemical' equation, for example:

EC 2.6.1.1 (aspartate transaminase)

Sometimes, when an enzyme has wide specificity, the reaction is written in terms of the general type of reactant, for example:

EC 1.1.1.1 (alcohol dehydrogenase)

and sometimes as a description, for example:

EC 3.2.1.1 (α -amylase)

Endohydrolysis of 1,4- α -D-glucosidic linkages in polysaccharides containing three or more 1,4- α -linked D-glucose units.

It must be stressed that the reaction as written is not meant to indicate the preferred equilibrium of the reaction or the direction in which some may believe the enzyme to operate *in vivo*. In any given sub-subclass, the direction chosen for the reaction is the same for all enzymes. Systematic names are based on this written reaction. Frequently, such biochemical equations are not charge-balanced.

Other name(s)

This field contains other names that have been used for the same enzyme, and is as comprehensive as possible to facilitate searching. It should be noted that the inclusion of a name in this list does not mean that its use is encouraged. In some cases where the same name has been given to more than one enzyme, or when the common name is misleading, this may be indicated.

Systematic name

This attempts to describe in unambiguous terms the reaction that the enzyme actually catalyses, and consists of two parts. The first contains the name of the substrate or, in the case of a bimolecular reaction, of the two substrates, separated by a colon. The second part, ending in *-ase*, indicates the nature of the reaction.

A number of generic words indicating a type of reaction are used: *oxidoreductase*, *oxygenase*, *transferase* (with a prefix indicating the nature of the group transferred), *hydrolase*, *lyase*, *racemase*, *epimerase*, *isomerase*, *mutase* and *ligase*. Where additional information is needed to make the reaction clear, a word or phrase indicating the reaction or a product is added in parentheses after the second part of the name, e.g. (*ADP-forming*), (*dimerizing*) or (*CoA-acylating*).

Comments

This field may contain information on the nature of the reaction catalysed, possible relationships to other enzymes, species differences, metal-ion and cofactor requirements, etc.

References

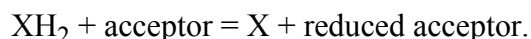
Key references on the identification, nature, properties and function of the enzyme are listed. It is important to note that the characterization of an enzyme is a prerequisite to its inclusion in the Enzyme List.

CONVENTIONS

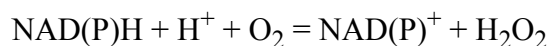
The way in which the reaction and systematic name are presented follow certain conventions. These are designed for consistency and in order to make it easier for proposers of new enzymes to suggest valid entries. Each class has its own specific conventions.

(a) The oxidoreductases (class 1)

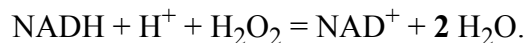
For most enzymes in this class, the reaction is written in the general form:



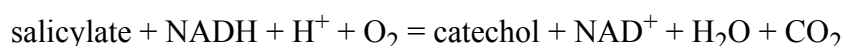
Where the acceptor may be NAD(P)^+ , a cytochrome, oxygen, a disulfide etc. The exceptions are those enzymes where two acceptors are involved. These include the subclass EC 1.6 where NAD(P)H is regarded as the oxidized substrate, as for example in the case of NAD(P)H oxidase (EC 1.6.3.1), where the reaction is written as:



Those enzymes that use peroxide as an acceptor (EC 1.11; the peroxidases) also depart from the normal formulation when the other substrate is NAD(P)H. For example the reaction catalysed by NADH peroxidase (EC 1.11.1.1) is presented as:



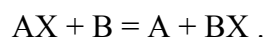
The subclass EC 1.14, which includes enzymes acting on paired donors, with incorporation or reduction of molecular oxygen, is also treated differently. For example, the reaction for salicylate 1-monooxygenase (EC 1.14.13.1) is written as:



The systematic names are written in terms of these directions and substrate orders.

(b) Enzymes catalysing transfer reactions (class 2)

For these reactions the donor is written first, in the general form:



Examples are:



Note that the order is preserved in the products (e.g. AdoHcy before methyl-X and ADP before X-P).

Two enzymes involved in the synthesis of sucrose illustrate this convention.

(1) UTP-glucose-1-phosphate uridylyltransferase (EC 2.7.7.9)



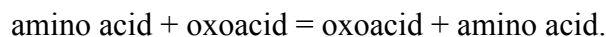
and has the systematic name UTP: α -D-glucose-1-phosphate uridylyltransferase.

(2) The reaction catalysed by sucrose synthase (EC 2.4.1.13) has the reaction:

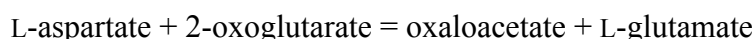


and the systematic name NDP-glucose:D-fructose 2- α -D-glucosyltransferase.

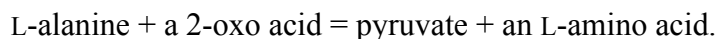
In all of the above examples there is little difficulty in deciding which is the donor substrate and which is the acceptor substrate. The aminotransferases (subclass 2.6) catalyse the general reaction:



Since the reaction could be written in either direction, by convention, when 2-oxoglutarate or an unspecified oxoacid is involved, this is usually treated as the acceptor. For example, the reaction catalysed by aspartate transaminase (EC 2.6.1.1; L-aspartate:2-oxoglutarate aminotransferase) is written as:

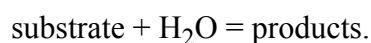


and that for alanine-oxo-acid transaminase (EC 2.6.1.12; L-alanine:2-oxo-acid aminotransferase) is:

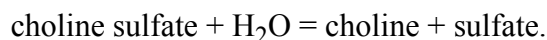


(c) The hydrolases (class 3)

These are generally straightforward, with the reaction written as:



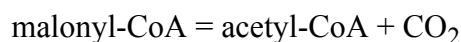
For example, the choline-sulfatase (EC 3.1.6.6) reaction is written as:



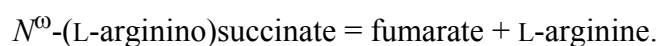
The systematic name, choline-sulfate sulfohydrolase, indicates both the substrate and the group removed by hydrolysis.

(d) The lyases (class 4)

These differ from other enzymes in that two substrates are involved in one direction and only one in the other. The reaction is written in the direction of less to more. Thus, malonyl-CoA decarboxylase (EC 4.1.1.9; malonyl-CoA carboxy-lyase) catalyses the reaction:

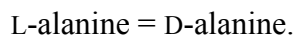


and the reaction catalysed by argininosuccinate lyase (EC 4.3.2.1; *N*-(L-argininosuccinate) arginine-lyase) is written as:

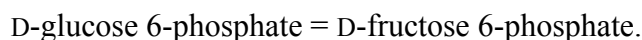


(e) The isomerases (class 5)

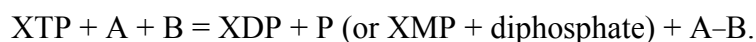
These are straightforward, one substrate, one product reactions. The reaction catalysed by alanine racemase (EC 5.1.1.1) is, for example,



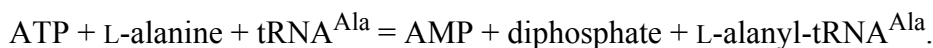
Similarly the reaction for glucose-6-phosphate isomerase (EC 5.3.1.9; D-glucose-6-phosphate ketol-isomerase) is written as:

**(f) The ligases (class 6)**

These enzymes catalyse the joining of two molecules with the concomitant hydrolysis of a diphosphate bond in ATP or a similar triphosphate. The reactions are normally written in the order:



Thus, the reaction catalysed by alanine-tRNA ligase (EC 6.1.1.7; L-alanine:tRNA^{Ala} ligase (AMP-forming)) is written as:



The pyruvate carboxylase reaction (EC 6.4.1.1; pyruvate:carbon-dioxide ligase (ADP-forming)) is, likewise, written as:

**PEPTIDASES AND RESTRICTION ENZYMES ARE TREATED DIFFERENTLY**

The group of enzymes that catalyse the hydrolysis of peptide bonds in proteins and peptides are grouped together under the hydrolases as EC 3.4.11.z – EC 3.4.25.z according to the type of reaction catalysed or the catalytic type of the enzyme involved. For example, EC 3.4.11 contains the aminopeptidases and EC 3.4.13 contains the dipeptidases, whereas EC 3.4.21 and EC 3.4.22 contain those endopeptidases where serine and cysteine residues, respectively, are involved in the catalytic process. Similarly, EC 3.4.24 contains the metalloendopeptidases. A further sub-subclass EC 3.4.99 is reserved for endopeptidases for which the catalytic mechanism is not yet known.

Extending Enzyme Classification with Metabolic and Kinetic Data

Where necessary, the amino acids in the peptide substrate are represented by P1...P_n, numbered towards the N-terminus and P1'...P_n', numbered towards the C-terminus. The peptide bond cleaved (the scissile bond) is indicated by the symbol ∇ .



This departure from the normal logic of the Enzyme List is a result of the demand for classification of a large number of peptidases with similar substrate specificities. If the reaction catalysed were used as the sole basis of classification, there would be rather few peptidases classified. Another difference concerns the naming of the peptidases. There are no systematic names, since the overlapping specificities would make it impossible to assign unique systematic names to each of them. Whereas in the remainder of the Enzyme List, it was decided to adopt the most commonly used name as the common name of an enzyme, even though such a name would not be one that might have been chosen as an adequate description of the reaction catalysed (e.g., catalase, pyruvate kinase), the absence of a systematic name necessitates the recommendation of a unique name for each peptidase. Thus, in this case, names like trypsin, pepsin A, pepsin B and renin are recommended names. Some examples are given in Table 2.

Table 2. The peptidases - some examples

EC 3.4.16.2

Recommended name: lysosomal Pro-X carboxypeptidase

Reaction: Cleavage of a -Pro ∇ Xaa bond to release a C-terminal amino acid

EC 3.4.21.1

Recommended name: chymotrypsin

Reaction: Preferential cleavage: Tyr ∇ , Trp ∇ , Phe ∇ , Leu ∇

EC 3.4.22.2

Recommended name: papain

Reaction: Hydrolysis of proteins with broad specificity for peptide bonds, but preference for an amino acid bearing a large hydrophobic side chain at the P2 position. Does not accept Val in P1'.

Table 2. continued**EC 3.4.23.1****Recommended name:** pepsin A**Reaction:** Preferential cleavage: hydrophobic, preferably aromatic, residues in P1and P1' positions. Cleaves Phe¹↯Val, Gln⁴↯His, Glu¹³↯Ala, Ala¹⁴↯Leu,Leu¹⁵↯Tyr, Tyr¹⁶↯Leu, Gly²³↯Phe, Phe²⁴↯Phe and Phe²⁵↯Tyr bonds in the B chain of insulin**EC 3.4.23.38****Recommended name:** plasmepsin I**Reaction:** Hydrolysis of the -Phe³³↯Leu- bond in the α -chain of hemoglobin, leading to denaturation of the molecule**EC 3.4.21.10****Recommended name:** acrosin**Reaction:** Preferential cleavage: Arg↯Lys↯

The restriction deoxyribonucleases also constitute a large family of enzymes with specificities that sometimes overlap. However, in this case, they can be divided into three families, as shown in Table 3. The Enzyme List directs users to the Restriction Enzyme Database (REBASE) [5] for further information about what is currently known about individual enzymes in these subclasses.

Table 3. The restriction deoxyribonuclease types**EC 3.1.21.3****Common name:** type I site-specific deoxyribonuclease**Reaction:** Endonucleolytic cleavage of DNA to give random double-stranded fragments with terminal 5'-phosphates; ATP is simultaneously hydrolysed**Other name(s):** type I restriction enzyme; deoxyribonuclease (ATP- and *S*-adenosyl-L-methionine-dependent); restriction-modification system; deoxyribonuclease (adenosine triphosphate-hydrolyzing); adenosine triphosphate-dependent deoxyribonuclease; ATP-dependent DNase

Extending Enzyme Classification with Metabolic and Kinetic Data

Table 3. continued. Comments on EC 3.1.21.3

Comments: This is a large group of enzymes which, together with those now listed as EC 3.1.21.4 (type II site-specific deoxyribonuclease) and EC 3.1.21.5 (type III site-specific deoxyribonuclease), were previously listed separately in sub-subclasses EC 3.1.23 and EC 3.1.24. They have an absolute requirement for ATP (or dATP) and *S*-adenosyl-L-methionine. They recognize specific short DNA sequences and cleave at sites remote from the recognition sequence.

They are multifunctional proteins that also catalyse the reactions of EC 2.1.1.72 [site-specific DNA-methyltransferase (adenine-specific)] and EC 2.1.1.73 [site-specific DNA-methyltransferase (cytosine-specific)], with similar site specificity. A complete listing of all of these enzymes has been produced by R.J. Roberts and is available at <http://rebase.neb.com/rebase/rebase.html>.

EC 3.1.21.4

Common name: type II site-specific deoxyribonuclease

Reaction: Endonucleolytic cleavage of DNA to give specific double-stranded fragments with terminal 5'-phosphates

Other name(s): type II restriction enzyme

Comments: This is a large group of enzymes which, together with those now listed as EC 3.1.21.3 (type I site-specific deoxyribonuclease) and EC 3.1.21.5 (type III site-specific deoxyribonuclease), were previously listed separately in sub-subclasses 3.1.23 and 3.1.24. They require only Mg^{2+} . They recognize specific short DNA sequences and cleave either within, or at a short specific distance from, the recognition site. A complete listing of all of these enzymes has been produced by R.J. Roberts and is available at <http://rebase.neb.com/rebase/rebase.html>.

EC 3.1.21.5

Common name: type III site-specific deoxyribonuclease

Reaction: Endonucleolytic cleavage of DNA to give specific double-stranded fragments with terminal 5'-phosphates

Other name(s): type III restriction enzyme; restriction-modification system

Comments: This is a large group of enzymes which, together with those now listed as EC 3.1.21.3 (type I site-specific deoxyribonuclease) and EC 3.1.21.4 (type II site-specific deoxyribonuclease), were previously listed separately in sub-subclasses EC 3.1.23 and EC 3.1.24. They have an absolute requirement for ATP but do not hydrolyse it; *S*-adenosyl-L-methionine stimulates the reaction, but is not absolutely required. They recognize specific, short DNA sequences and cleave a short distance away from the recognition sequence.

Table 3. continued. Comments on EC 3.1.21.5

These enzymes exist as complexes with enzymes of similar specificity listed under EC 2.1.1.72 [site-specific DNA-methyltransferase (adenine-specific)] or EC 2.1.1.73 [site-specific DNA-methyltransferase (cytosine-specific)]. A complete listing of all of these enzymes has been produced by R.J. Roberts and is available at <http://rebase.neb.com/rebase/rebase.html>.

RESTRICTIONS/LIMITATIONS OF THE PRESENT CLASSIFICATION SYSTEM

There are a number of limitations associated with the classification system used for enzymes. These are listed below.

(a) The same EC number may be assigned to many different proteins

For example, any isoenzymes and species differences come under the umbrella of the single enzyme alcohol dehydrogenase (EC 1.1.1.1) as they all perform the same reaction, i.e., convert an alcohol to an aldehyde and in the process reduce NAD^+ . Only when there is a clear difference in specificity is an enzyme assigned a different EC number and name. Thus there are other enzymes classified that catalyse the dehydrogenation of alcohols, as shown in Table 4.

Table 4. Some alcohol dehydrogenase enzymes

EC number	Common Name	Reaction
EC 1.1.1.1	alcohol dehydrogenase	an alcohol + NAD^+ = an aldehyde or ketone + $\text{NADH} + \text{H}^+$
EC 1.1.1.2	alcohol dehydrogenase (NADP^+)	an alcohol + NAD(P)^+ = an aldehyde + $\text{NAD(P)H} + \text{H}^+$
EC 1.1.1.71	alcohol dehydrogenase [NAD(P)^+]	an alcohol + NAD(P)^+ = an aldehyde + $\text{NAD(P)H} + \text{H}^+$
EC 1.1.99.8	alcohol dehydrogenase (acceptor)	a primary alcohol + acceptor = an aldehyde + reduced acceptor
EC 1.1.1.192	long-chain-alcohol dehydrogenase	a long-chain alcohol + $2 \text{NAD}^+ + \text{H}_2\text{O}$ = a long-chain carboxylate + $2 \text{NADH} + 2 \text{H}^+$
EC 1.1.1.194	coniferyl-alcohol dehydrogenase	coniferyl alcohol + NADP^+ = coniferyl aldehyde + $\text{NADPH} + \text{H}^+$
EC 1.1.1.21	aldehyde reductase	alditol + NAD(P)^+ = aldose + $\text{NAD(P)H} + \text{H}^+$
EC 1.1.1.184	carbonyl reductase (NADPH)	$\text{R-CHOH-R}' + \text{NADP}^+ = \text{R-CO-R}' + \text{NADPH} + \text{H}^+$

(b) Different EC numbers may be assigned to the same protein

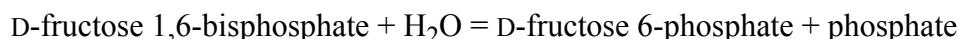
For example, the tryptophan synthase complex may be a single protein containing some of the following [EC 2.4.2.18 (anthranilate phosphoribosyltransferase), EC 4.1.1.48 (indole-3-glycerol-phosphate synthase), EC 4.1.3.27 (anthranilate synthase), EC 4.2.1.20 (tryptophan synthase) and EC 5.3.1.24 (phosphoribosylanthranilate isomerase)] (see [1,2]). The Enzyme List also indicates that 6-phosphofructo-2-kinase (EC 2.7.1.105) may be part of the same protein as fructose-2,6-bisphosphate 2-phosphatase (EC 3.1.3.46). Other examples include: two distinct domains on the NadR protein *Haemophilus influenzae* allow it to function both as a nicotinamide-nucleotide adenylyltransferase (EC 2.7.7.1) and a ribosylnicotinamide kinase (EC 2.7.1.22) [6], and human maleylacetoacetate isomerase (EC 5.2.1.2) is also a glutathione transferase (EC 2.5.1.18) zeta isoenzyme [7].

(c) The reaction equations and systematic names do not necessarily indicate the direction in which the reaction may be perceived to operate *in vivo*

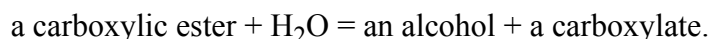
In many cases, such as many of the enzymes of glycolysis, the reaction direction depends upon the cellular metabolic conditions, whereas in some other cases the direction is not known.

(d) Whereas the reaction equations are generally mass-balanced, they are not necessarily charge-balanced

This is because they are written as pH-independent equations. For example, the reaction catalysed by fructose-bisphosphatase (EC 3.1.3.11) is written as:

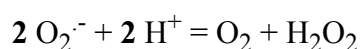


with no attempt to indicate the charges on the phosphates or whether metal-ion complexes are involved. In fact it would only be possible to indicate these if the reaction conditions, e.g., pH and metal cation composition, were specified. Similarly hydrogen ions are generally omitted from reactions which might be expected to produce them at physiological pH values, e.g. carboxylesterase (EC 3.1.1.1)



An exception to this is NAD(P) in oxidoreductase reactions where these oxidized forms are, by convention, written as NAD(P)^+ and the reduced forms as $\text{NAD(P)H} + \text{H}^+$.

It has been recognized that this is a somewhat arbitrary departure from the rule that charges are generally omitted from the biochemical equations used (see [8, 9]) and, indeed, NADP, for example does not have a net positive charge at physiological pH values. However, the convention is too well embedded amongst the biochemical community and the suggestion [10] that it would be more appropriate to represent the redox pairs as NAD(P) and NAD(P)H₂, or NAD(P) and reduced NAD(P), did not meet with favour. Furthermore, these alternative formulations would not make it easy to represent the NAD radicals that occur in some reactions. Where it is convenient, charges may also be used to represent radical species, e.g., the reaction catalysed by superoxide dismutase (EC 1.15.1.1) is written as:



(e) The Enzyme List does not provide information on the mechanism of a reaction

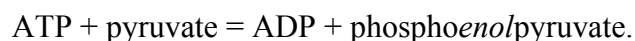
Only the overall reaction catalysed is considered, although some information on the mechanism may be provided in the comments. For example, fructose-bisphosphate aldolase (EC 4.1.2.13), which catalyses the reaction:

D-fructose 1,6-bisphosphate = glyceraldehyde 3-phosphate + D-glyceraldehyde 3-phosphate,

contains enzymes that operate by very different chemical mechanisms. This is indicated in the Comments' section, where it is stated that "The yeast and bacterial enzymes are zinc proteins. The enzymes increase electron-attraction by the carbonyl group, some (Class I) forming a protonated imine with it, others (Class II), mainly of microbial origin, polarizing it with a metal ion, e.g. zinc".

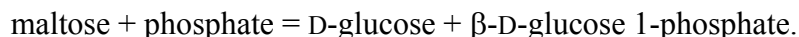
(f) Since the Enzyme List is based on the overall reaction catalysed, immediate reaction products that rapidly and spontaneously convert to a more stable form may not be indicated

Thus the pyruvate kinase (EC 2.7.1.40) reaction is written as:



However, the product of this reaction, which operates in the direction of ADP phosphorylation in mammalian cells, is enolpyruvate which immediately isomerizes to the *keto* form.

On the other hand, if the product isomerizes rather slowly, such that the immediate product can be detected without recourse to rapid-reaction techniques, it is the immediate product that is shown, as in the case of the reaction catalysed by maltose phosphorylase (EC 2.4.1.8)

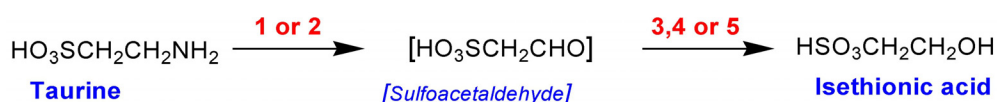


(g) The Enzyme List does not provide any information on any nonenzymic functions

An increasing number of enzymes are recognized to have additional nonenzymic functions (see [11] for discussion). Such functions are not included in the Enzyme List.

ATTEMPTS TO PREDICT MISSING ENZYMES

While it is tempting to predict the existence of an enzyme based on a 'gap' in a metabolic pathway, where it could be argued that an enzyme must exist to convert the product of one reaction into the substrate of another, this can lead to incorrect assumptions. For example, in the case of taurine metabolism, it would be reasonable to presume that the reaction from the substrate taurine to the product isethionic acid should proceed as follows:



with 1 being an amine oxidase, 2 an aminotransferase, 3 an aldehyde oxidase, 4 an aldehyde dehydrogenase and 5 an alcohol dehydrogenase. Indeed, such pathways do occur in some bacteria and fungi. However, this does not appear to be what happens in mammalian systems, where taurine is first converted spontaneously to taurine chloramine in the presence of hypochlorous acid and it is taurine chloramine and not taurine that takes part in the enzyme-catalysed reaction, shown in Fig. 1 [12].

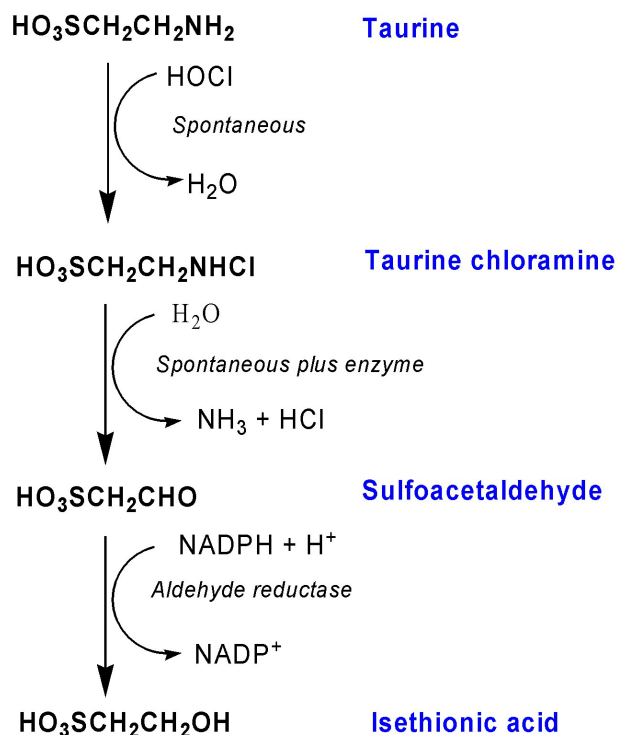


Figure 1. Breakdown of taurine in rat. Taurine reacts non-enzymically with hypochlorous acid (HOCl) to form *N*-chlorotaurine (taurine chloramine) and this is then converted to sulfoacetaldehyde and isethionic acid (see [12] for further details).

NEW ENZYMES

The purpose of the Enzyme List is, as far as possible, to provide unambiguous data on enzymes and the reactions they catalyse. Such data are then used by a variety of other databases (see Table 5 for examples). The criteria for the addition of a new enzyme are simple but strict. The proposed new enzyme must be shown actually to catalyse a reaction that is significantly different from those catalysed by enzymes already listed. Forms for submitting new enzymes or corrections/updates to existing entries are available on-line [1, 13].

Extending Enzyme Classification with Metabolic and Kinetic Data

Table 5. Some databases that use the EC classification system

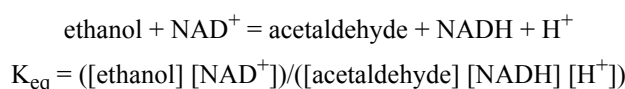
Database	URL
BRENDA	http://www.brenda.uni-koeln.de/
CarBank	http://bssv01.lancs.ac.uk/gig/pages/gag/carbbank.htm
Database Enzyme (UK HGMP Resource Centre)	http://www.hgmp.mrc.ac.uk/Bioinformatics/Databases/enzyme-help.html
Directory of p450-containing systems	http://www.icgeb.trieste.it/~p450srv/
EcoCyc	http://ecocyc.org
EMP Database of enzymes and metabolic pathways	http://wit.mcs.anl.gov/WIT2/EMP/
Enzyme information and structure database	http://restools.sdsc.edu/biotools/biotools12.html
Enzyme Nomenclature	http://www.chem.qmul.ac.uk/iubmb/enzyme/
Enzyme Structures Database	http://www.biochem.ucl.ac.uk/bsm/enzymes/
ExPasy	http://ca.expasy.org/enzyme/
GTD (thermodynamics of enzyme catalysed reactions)	http://www.biotech.nist.gov:8030/enzyme/
HUGO	http://www.gene.ucl.ac.uk/nomenclature/
KEGG (Kyoto Database of Genes and Genomes)	http://www.genome.ad.jp/kegg/
Klotho	http://www.biocheminfo.org/klotho/
LIGAND	http://www.genome.ad.jp/dbget/ligand.html
MaizeDB	http://www.maizegdb.org/
MEROPS	http://merops.sanger.ac.uk/
PDB	http://www.rcsb.org/pdb/
Phosphoprotein database	http://www.lecb.ncifcrf.gov/phosphoDB/
PROMISE	http://metallo.scripps.edu/PROMISE/
REBASE	http://rebase.neb.com/rebase/rebase.html
UMBBD (Biocatalysis/Biodegradation)	http://umbbd.ahc.umn.edu/
WIT	wit.mcs.anl.gov/WIT
Worthington Enzyme Manual	http://www.worthington-biochem.com/index/manual.html

APPLICATION OF THERMODYNAMIC DATA

Because the Enzyme List is restricted to providing factual data, it is the function of other databases to provide activity and thermodynamic data that may be used for simulating metabolic processes, reconstructing systems, determining control properties etc. Thermodynamic data for many enzymes can be found in the GTD Thermodynamics of Enzyme-catalysed Reactions database (see Table 5) and kinetic data are included in the BRENDA and WIT databases (see Table 5). However, for these to be meaningful, it is necessary that such data refer to 'physiologically relevant conditions' (see [9]).

Attempts to use thermodynamic data for isolated enzyme-catalysed reactions to predict the direction of flux *in vivo* should be treated with extreme caution, since enzymes usually form parts of metabolic systems. As a simple example, the equilibrium constants for the reaction catalysed by alcohol dehydrogenase are shown in Table 6. From the value at neutral pH, it is quite clear that we should not be very good at metabolizing ethanol. However, we have no great problem with this in the tissues because the acetaldehyde produced is rapidly converted to acetate by aldehyde dehydrogenase (NAD⁺)(EC1.2.1.3), which has a very low K_m value for that substrate and catalyses a reaction that is essentially irreversible.

Table 6. Equilibrium constants for the alcohol dehydrogenase (EC 1.1.1.1) reaction



pH	(K_{eq})
7.0	1.1×10^{-4}
8.0	7.1×10^{-4}
9.0	1.05×10^{-2}
10.0	9.0×10^{-2}

Temperature 293,15°K

Source GTD - see Table 5

ASSAY CONDITIONS

Comparison of enzyme activities and kinetic parameters and the reconstruction of metabolic systems require data to be obtained under comparable conditions. However, even a brief survey of the literature will indicate that this is far from the case. Even with what is apparently the same enzyme, different laboratories often assay under different conditions and the assay conditions used for different enzymes in the same metabolic pathway can differ markedly. Some recommendations have been formulated (see [14]), but these are somewhat imprecise.

This section will not deal with the assay procedures, importance of determining initial rates or the essential controls for progress-curve analysis, since these have been dealt with in detail elsewhere [15, 16]. Despite this, there have been attempts to define standard assay conditions to facilitate comparison of data.

a) Temperature

Originally many studies were conducted at 'room temperature', which could, of course, vary widely between laboratories. It had been recommended that enzymes should be assayed at 25°C, which was regarded as a standard 'room temperature'. However, not all laboratories were able to meet this requirement and the standard assay temperature was raised to 30°C. Even this gradual thermal inflation does not satisfy those studying human enzymes, who would regard a temperature of 37°C as being more appropriate for most tissues. However, this definition of physiological temperature for a mammalian system would not be appropriate, for example, for a thermophilic bacterium or a poikilotherm.

Perhaps the only logical way of dealing with this issue would be to specify recommended assay temperatures for specific organisms, or groups of organisms, remembering that to stick to a single temperature for the study of organisms that might be subjected to quite large temperature fluctuations could lead to the loss of important information.

b) pH value

The recommended pH value for enzyme assays is also not very helpful. Although it has been suggested that the assay pH should "where practicable, be optimal", this is not a great deal of help. Optimal pH is to some extent part of a circular argument since this may depend on the choice of substrate, the substrate concentrations, buffer, temperature and ionic strength and there are no strict recommendations for any of these. Furthermore, the optimum pH may be far removed from the pH at which an enzyme is perceived to operate *in vivo*. For example, the optimum pH for arginase (EC 3.5.3.1) is reported to be about pH 10 in horse, pH 9.8 in rat and pH 11 in *Bacillus brevis*. Those working with mammalian systems might favour an assay pH of about 7.2, which is believed to be around the physiological pH within the cell, but clearly this would be unphysiological for gastrointestinal enzymes, such as pepsin and trypsin, or for lysosomal enzymes. Thus it would not be helpful to recommend a standard pH value that would be appropriate for all systems. An alternative might be to devise individual standards for each organism, organ and organelle to be studied.

A further complication that must be borne in mind is that it is sometimes not easy to assay an enzyme at the desirable pH. For example, as discussed above, the equilibrium constant for the reaction catalysed by alcohol dehydrogenase is so far towards acetaldehyde at neutral pH values that it is difficult to assay the enzyme in the direction of ethanol oxidation.

Because of this, many studies of that reaction have utilized a high assay pH value and also, sometimes, included an aldehyde-trapping reagent, such as semicarbazide.

c) Substrates and substrate concentrations

Naturally it would be appropriate to use physiological substrates for enzyme assays and, in the case of enzymes that use more than one substrate that compete with one-another for the enzyme, it would be necessary to study each in turn. In reality, however, many studies have used non-physiological substrates for ease of manipulation and assay. For example, acetylthiocholine is frequently used to assay acetylcholinesterase (EC 3.1.1.7) because the thiocholine produced can be readily detected by reaction with sulfhydryl reagent 5,5'-dithiobis-2-nitrobenzoate (Nbs₂), releasing a yellow-coloured compound the formation of which can be followed at 412 nm [17]. Other examples include the use of pyroglutamyl-histidyl-prolylamido-4-methyl coumarin instead of the physiological substrate thyroid-stimulating hormone (pyroglutamyl-histidyl-prolylamide) to assay pyroglutamyl-peptidase II (EC 3.4.19.6) [18], the use of esters rather than peptides to assay a number of peptidases, the use of dyes as electron acceptors in oxidoreductase assays and the use of *p*-nitrophenyl phosphate to assay alkaline phosphatase (EC 3.1.3.1). The demand for higher assay sensitivity and high-throughput procedures has resulted in the development of an increasing number of chromogenic and fluorogenic substrates and, in some cases, it is difficult to find the structures of these compounds. Clearly, in such cases, a considerable amount of work would be necessary to show that the enzyme behaves identically towards such substrates as it does towards its physiological substrates.

Unless the K_m and V_{max} values are to be determined, the substrate concentration used in an assay will, of course, affect the activity values obtained. It is often recommended that saturating substrate concentrations should be used (i.e. $> 10 K_m$) for all substrates. This, of course, assumes that the K_m value has already been determined. Furthermore, this might not always be practicable because of factors such as solubility, the occurrence of high-substrate inhibition or a high absorbance of the assay mixture affecting the behaviour of optical assays (see [15, 16]). Furthermore, it should be remembered that any change in the assay conditions (e.g., pH, temperature, ionic strength) may affect the K_m values.

Although the above considerations indicate the desirability of performing a thorough kinetic study to determine the K_m and V_{max} values for all physiological substrates, there are many situations where simple activity level comparisons are sufficient, as, for example, in comparing the activities of plasma enzymes in the diagnosis of different pathological conditions. In such cases, one may be able to use non-physiological substrates and assay conditions, provided these are fully specified to allow the results to be replicated by others.

d) Buffers and ionic strength

The buffers and ionic strengths used in enzyme assays vary widely and are often far from physiological. It might be helpful if it were possible to recommend a simple standard buffer for use in enzyme assays. Unfortunately, this goal appears to be unobtainable, because at least some enzymes are unhappy in one or other of the common buffers. Furthermore, buffers that contain physiologically occurring compounds can be problematical in that they are likely to be substrates for some enzymes. For example, inorganic phosphate is a substrate for several enzymes and citrate is a substrate for enzymes such as aconitate hydratase (EC 4.2.1.3). Assay of such enzymes in a buffer based on these anions would of course preclude studies at varying substrate concentrations.

Phosphate buffer is widely used for enzyme studies but it can act as a product inhibitor for some enzymes that release phosphate. It is also, for example, an inhibitor of arylsulfatase (EC 3.1.6.1) (see [14]). Tris buffer (commonly known as Tris-HCl) is hardly physiological and also interferes with the assay of aldehyde dehydrogenase, and glutamate dehydrogenase (EC 1.4.1.2) from some sources is unstable in this buffer (see [19]). Pyrophosphate buffer can inhibit some enzymes because it is a metal-ion chelator and is also a product inhibitor of several other enzymes. It is, however, the buffer in which some isoenzymes of aldehyde dehydrogenase are most active and has commonly been used in assays for those enzymes (see e.g., [20]). Citrate is a chelator and therefore inhibits several enzymes that require metal ions for activity. The range of so-called Good buffers fare no better, for example, some of these inhibit the amine oxidases (EC 1.4.3.4 & EC 1.4.3.6) and HEPES and Tris inhibit carbamoyl-phosphate synthase (ammonia) (EC 6.3.4.16) [21].

From the above examples, it should be clear that it is unlikely that a universal buffer medium, suitable for all enzymes, will easily be found.

Perhaps the answer will lie in more complex mixtures, including proteins, as buffers, that more closely mimic the *in vivo* environments of groups of enzymes. At present, it appears that specifying the buffer and its components might be the only alternative. Even then, present usage is often too imprecise. It is common to read statements such as 0.1 M phosphate buffer pH 7.2, with no information as to its precise composition or whether sodium or potassium phosphate was used.

The ionic strength of assay media is seldom stated, although this can be calculated if the full composition of the assay mixture is given, which is not always the case. Several enzymes are sensitive to inhibition by high ionic strengths and altering the concentrations of charged substrates and the pH of the buffer may also affect the ionic strength. It would be helpful if all authors were required to state the ionic strength of their assay mixtures.

e) Other additives

Assay mixtures often contain additional components to stabilize, protect or activate the enzyme. Each of these has to be assay-specific, since compounds that facilitate some enzyme assays may act as substrates for some other enzymes or inhibit them.

Some enzymes that contain reactive sulfhydryl groups that are essential for activity (e.g. papain (EC 3.4.22.2), aldehyde dehydrogenase and glycoprotein *N*-palmitoyltransferase (EC 2.3.1.96) are assayed in the presence of reducing agents, such as cysteine, glutathione, ethanethiol (2-mercaptoethanol) or dithiothreitol (DTT). These compounds inhibit some other enzymes. Since glutathione and cysteine are physiological compounds, they are substrates for several other enzymes. Dithiothreitol is also a substrate for enzymes, the vitamin-K-epoxide reductases (EC 1.1.4.1 & EC 1.1.4.2) and can replace reduced thioredoxin and glutathione in the reactions catalysed by methionine-*S*-oxide reductase (EC 1.8.4.5) and adenylyl-sulfate reductase (glutathione) (EC 1.8.4.9), respectively. Ethanethiol can act as a substrate for thioether *S*-methyltransferase (EC 2.1.1.96) [22]. Such compounds may also interfere with some enzyme-assay procedures, for example, assays based on sulfhydryl-group detection, such as that for acetylcholinesterase, referred to above.

In some cases, a reducing agent is added to protect the substrate from oxidation, such as the addition of ascorbate, which is a substrate for some other enzymes, to solutions of adrenaline for monoamine oxidase (EC 1.4.3.4) assays.

Clearly the alternative of preventing non-enzymic adrenaline oxidation by working anaerobically is not possible for that oxygen-requiring enzyme.

The addition of a chelating agent such as EDTA is common for the assay of some enzymes that are sensitive to inhibition by traces of heavy-metal ions, such as papain and fructose biphosphatase (EC 3.1.3.11). Chelating agents are also sometimes used to buffer the free concentrations of divalent cations in solution. However, they can also inhibit a number of metal-ion-requiring enzymes, such as arylalkylphosphatase (EC 3.1.8.1), diisopropyl-fluorophosphatase (EC 3.1.8.2) and some metallopeptidases (e.g. EC 3.4.11.6, EC 3.4.17.10, EC 3.4.24.29).

Many enzymes require the addition of a metal cation for activity, either because the substrate for the enzyme is a metal chelate, as is, for example, the case with many reactions involving ATP, or because a metal ion is an essential activator of the enzyme. In some cases both of these factors may operate as, for example, in the case of pyruvate carboxylase (EC 6.4.1.1), where the reaction involves the binding of Mg-ATP to an enzyme-Mg complex [23]. In many cases, the metal ion may be tightly bound to the enzymes, but in other cases it dissociates readily. For example, it is necessary to add Fe^{2+} to cytoplasmic aconitate hydrolase (EC 4.2.1.3), to replace that lost in extraction and purification, in order to detect activity [24].

The specificity for metal ions can be high, for example, Mg^{2+} is required for adenylate cyclase (EC 4.6.1.1) activity but the enzyme is inhibited by Ca^{2+} [25]. Useful listings of metal and buffer ions as inhibitors of specific enzymes have been compiled by Zollner [26]. In addition, several enzymes where the true substrate is the metal-substrate complex are inhibited by high concentrations of uncomplexed metal ion and/or substrate.

Other factors are necessary for specific enzymes. The activator *N*-acetyl-L-glutamate is included in assays for carbamoyl-phosphate synthase (ammonia) (see [21]), pyruvate carboxylase has very little activity in the absence of acetyl-CoA (see [23]), ADP is frequently added, as an activator, for the assay of glutamate dehydrogenase [NAD(P)^+] (EC 1.4.1.3) (see [10]) and one form of glutaminase (EC 3.5.1.2) has little activity in the absence of phosphate (see e.g., [27]). All such compounds are inhibitors and/or substrates of other enzymes.

f) The enzyme

The question of whether studies of isolated enzymes can provide data that are relevant to cellular metabolism is almost as old as enzymology. It is often posed and just as often ignored because, despite many wishful claims to the contrary, the available technology does not offer a viable alternative. Sometimes even the simplest steps to ensure that the enzyme preparation is adequate are not taken. It is common to find that proteolysed preparations are used, either by design or accident, with the assumption that if the enzyme preparation has activity, it must be satisfactory. However, there is a considerable amount of evidence that this may not be a valid assumption. Removal of 4 or 5 amino-acid residues from the N-terminus of glutamate dehydrogenase, which can readily occur during extraction and purification, has been shown to affect its regulatory properties [28]. Similarly, proteolytic cleavage of fructose biphosphatase affects its pH optimum and allosteric regulation (see [29]). Despite such 'cautionary tales' an increasing number of studies have been conducted with preparations that are truncated, contain tags such as poly-His, lack glycosylation or are suspended in some odd detergent. The relevance of such studies is not clear.

Since the enzyme is a catalyst and usually present at concentrations very much lower than those of the substrates, one would expect the initial velocity of the reaction to be proportional to the enzyme concentration and this is true in the majority of cases. Situations in which this proportionality does not apply, which may be a result of assay artefacts, impurities in the enzyme preparation or the assay mixture, or as a result of dissociating systems that may be of physiological significance, have been discussed in detail elsewhere [15,16]. However, they indicate that it is essential that such proportionality is checked experimentally and that the causes of any departures are investigated. The often-quoted assertion that Michaelis-Menten kinetics do not apply if the concentration of the substrate is much less than that of the enzyme, which arises from the failure to distinguish between total substrate concentration (used in test-tube experiments) and free substrate concentration, which is the only valid parameter within the cell (see [22]), can easily be shown to be misleading. If the latter is maintained constant, owing to steady-state or equilibrium conditions in a metabolic system, there is no departure from the Michaelis-Menten equation when $[S] \ll [\text{Enzyme}]$.

CONCLUSIONS

The necessity of fully describing the assay mixtures used should not need stressing. Neither should the necessity for more care in ensuring that the enzyme preparation used corresponds to that existing *in vivo*. Certainly, temperature and pH might be more standardized, where appropriate. However, the above discussion indicates that it would be counter-productive to attempt to develop a universal assay mixture for the assay of all enzymes, since not all enzymes share the same environments. Even if satisfactory buffer mixtures were developed for the study of groups of enzymes in discrete systems, the necessity to have other additives in some assays that may be inimical to others will prevent the development of universal assay cocktails.

REFERENCES

- [1] <http://www.chem.qmul.ac.uk/iubmb/enzyme/>
 - [2] <http://www.ebi.ac.uk/intenz/index.html>
 - [3] Boyce, S., Tipton, K.F. (2000) Enzyme classification and nomenclature. In: *Nature Encyclopedia of Life Sciences*, Nature Publishing Group, London. <http://www.els.net/> [doi:10.1038/npg.els.0000710]
 - [4] Boyce, S., Tipton, K.F. (2000) History of the enzyme nomenclature system. *Bioinformatics* **16**: 34-40.
 - [5] <http://rebase.neb.com/rebase/rebase.html>
 - [6] Singh, S.K., Kurnasov, O.V., Chen, B., Robinson, H., Grishin N.V., Osterman, A.L., Zhang, H. (2002) Crystal structure of *Haemophilus influenzae* NadR protein: a bifunctional enzyme endowed with NMN adenylyltransferase and ribosylnicotinamide kinase activities. *J. biol. Chem.* **277**: 33291-33299.
 - [7] Blackburn, A.C., Woollatt, E., Sutherland G.R., Board, P.G. (1998) Characterization and chromosome location of the gene GSTZ1 encoding the human Zeta class glutathione transferase and maleylacetoacetate isomerase. *Cytogenet. Cell Genet.* **83**: 109-114.
 - [8] Alberty, R.A., Cornish-Bowden, A., Gibson, Q.H., Goldberg, R.N., Hammes, G., Jencks, W., Tipton, K.F., Veech, R., Westerhoff, H.V., Webb, E.C. (1996) Recommendations for nomenclature and tables in biochemical thermodynamics. *Eur. J. Biochem.* **240**: 1-14.
 - [9] Alberty R.A. (1991) Equilibrium compositions of solutions of biochemical species and heats of biochemical reactions. *Proc. Natl. Acad. Sci. USA* **88**: 3268-3271.
 - [10] <http://www.chem.qmul.ac.uk/iubmb/newsletter/1996/news7.html>
 - [11] Tipton, K.F., O'Sullivan, M.P., Davey, G.P., O'Sullivan, J. (2003) It can be a complicated life being an enzyme. *Biochem. Soc. Trans.* **31**: 711-715.
-

-
- [12] Cunningham, C., Tipton, K.F., Dixon, H.B.F. (1998) Conversion of taurine into N-chlorotaurine (taurine chloramine) and sulphoacetaldehyde in response to oxidative stress. *Biochem. J.* **330**: 933-937.
 - [13] <http://us.expasy.org/enzyme>
 - [14] Dixon, M., Webb, E.C., Thorne, C.J.R., Tipton, K.F (1979) *Dixon and Webb: Enzymes*, Longman, London.
 - [15] Tipton, K.F. (2002) Principles of enzyme assay and kinetic studies. In: *Enzyme Assays: A Practical Approach* (Eisenthal, R., Danson, M.J. Eds) pp. 1-47, Oxford University Press, Oxford.
 - [16] McDonald, A.G., Tipton, K.F. (2002) Kinetics of catalyzed reactions - biological. In: *Encyclopedia of Catalysis* (Horváth, I.T. Ed.), John Wiley & Sons, Inc., New York. <http://www.mrw.interscience.wiley.com/enccat/> (DOI:10.1002/0471227617.eoc127).
 - [17] Ellman, G.L., Courtney, K.D., Andres, V., Feather-Stone R.M. (1961) A new and rapid colorimetric determination of acetylcholinesterase activity. *Biochem. Pharmacol.* **7**: 88-95.
 - [18] Kelly, J.A., Slator, G.R., Tipton, K.F., Williams, C.H., Bauer, K. (1999) Development of a continuous, fluorometric coupled enzyme assay for thyrotropin-releasing hormone-degrading ectoenzyme. *Analyt. Biochem.* **274**: 195-202.
 - [19] Tipton, K.F., Couée, I. (1988) Glutamate dehydrogenase. In: *Glutamine and Glutamate in Mammals*. (Kvamme, E. Ed.) pp. 81-100, C.R.C. Press, Boca Raton.
 - [20] Hill, J.P., Buckley, P.D., Blackwell, L.F., Motion, R.L. (1991) Effect of pyrophosphate ions and alkaline pH on the kinetics of propionaldehyde oxidation by sheep liver cytosolic aldehyde dehydrogenase. *Biochem. J.* **273**: 691-693.
 - [21] Lund, P., Wiggins, D. (1987) Inhibition of carbamoyl-phosphate synthase (ammonia) by Tris and Hepes. Effect on K_a for N-acetylglutamate. *Biochem. J.* **243**: 273-276.
 - [22] Carrithers, S.L., Hoffman, J.L. (1994) Sequential methylation of 2-mercaptoethanol to the dimethyl sulfonium ion, 2-(dimethylthio)ethanol, *in vivo* and *in vitro*. *Biochem. Pharmacol.* **48**: 1017-1024.
 - [23] Warren, G.B., Tipton, K.F. (1974) Pig liver pyruvate carboxylase. The reaction pathway for the carboxylation of pyruvate. *Biochem. J.* **139**: 311-320.
 - [24] Kennedy, M.C., Emptage, M.H., Dreyer, J.L., Beinert, M. (1983) The role of iron in the activation-inactivation of aconitase. *J. biol. Chem.* **258**: 11098-11105.
 - [25] Steer, M.L., Levitzki, A. (1975) The control of adenylate cyclase by calcium in turkey erythrocyte ghosts. *J. biol. Chem.* **250**: 2080-2084.
 - [26] Zollner, H. (1999) *Handbook of Enzyme Inhibitors*, Wiley-VCH, Weinheim.
 - [27] Kvamme, E., Torgner, I.A., Roberg, B. (2001) Kinetics and localization of brain phosphate activated glutaminase. *J. Neurosci. Res.* **66**: 951-958.
-

- [28] McCarthy, A.D., Tipton, K.F. (1985) Ox liver glutamate dehydrogenase. Comparison of the kinetic properties of native and proteolysed preparations. *Biochem. J.* **230**: 95-99.
 - [29] Nimmo, H.G., Tipton, K.F. (1982) Fructose-biphosphate from ox liver. *Methods Enzymol.* **90**: 330-334.
-

METHODS OF DESIGN OF OPTIMAL EXPERIMENTS WITH APPLICATION TO PARAMETER ESTIMATION IN ENZYME CATALYTIC PROCESSES

HANS GEORG BOCK, STEFAN KÖRKEL, EKATERINA KOSTINA*
AND JOHANNES P. SCHLÖDER

Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg,
Im Neuenheimer Feld 368, 69120 Heidelberg, Germany
E-Mail: *ekaterina.kostina@iwr.uni-heidelberg.de

Received: 20th April 2004 / Published: 1st October 2004

ABSTRACT

This paper deals with the identification of kinetic parameters in enzyme catalytic processes. Experience shows that the experiments performed do not deliver measurement data sufficient for the identification of parameters. New optimal experiments are needed. We suggest effective algorithms and software for parameter estimation and design of optimal experiments, based on multiple shooting and specially tailored, structure-exploiting reduced Gauss-Newton and SQP methods. The methods are applied to optimal experimental design and parameter estimation in enzyme catalytic processes.

INTRODUCTION

We consider the problem of identifying enzyme operating stability which is very important e.g. in industrial production of enzymes or in understanding enzymatic pathways in a living system. Traditionally enzyme stability is determined empirically by time-consuming experiments: under constant reaction conditions (temperatures) the half-life of the catalyst is measured [1]. This procedure is repeated under different temperatures. Another approach for faster identification of enzyme operating stability is suggested in [2]. In an instationary operated reactor the reaction temperature is increased linearly and parameters of a kinetic model based on Arrhenius equations are estimated. Then the stability properties of enzymes are estimated as functions of the kinetic parameters. Experience shows however that such experiments do not deliver enough data to identify the kinetic parameters reliably.

Optimal experiments (temperature profiles) are necessary that allow the estimation of the kinetic parameters of enzyme catalytic processes and thus stability properties. This paper focuses on numerical methods for parameter estimation and design of optimal experiments.

The paper is organized as follows. Firstly, we describe parameter estimation problems in ordinary differential equations as well as specifying the model of enzyme reaction under consideration. Secondly, the method used for parameter estimation is described. It follows a very successful solution approach, namely the boundary value problem (BVP) approach, which combines infeasible point optimization algorithms like the generalized Gauss-Newton method with BVP techniques like multiple shooting to discretize the dynamic system. The discretized equations are then treated as nonlinear equality constraints. We solve the resulting finite-dimensional, possibly large-scale, nonlinear constrained least squares problem with the generalized Gauss-Newton method. Under certain regularity assumptions the method shows the final linear rate of local convergence. To ensure global convergence, a newly developed efficient step-size strategy is used. Special attention is paid to the discussion of why the generalized Gauss-Newton method is particularly appropriate for parameter estimation and does not converge to solutions with large residuals. The quality of parameter estimates is defined by statistical analysis, and the results of parameter estimation for an immobilized enzyme system finish this part of the paper. Thirdly, theoretical justification methods for the design of optimal experiments as well as numerical and experimental results are presented. Optimal experiments are designed for the nominal values of parameters which are known only to lie in a confidence region. Fourthly, the design of experiments that are less sensitive to parameter uncertainty is described. The paper is finished by conclusions.

PARAMETER ESTIMATION PROBLEM FOR ENZYME CATALYTIC PROCESSES

Parameter estimation problems in dynamic processes

The problem of identification of unknown parameters in dynamic models is among the most important tasks in mathematical modelling of dynamic processes. It can be described as follows. Let the dynamics of the model be described by a system of ordinary differential equations

$$\dot{y} = f(t, y(t), p, q, u(t)), t \in [t_0, t_{end}]$$

where the right-hand side f depends on an vector of unknown parameters $p \in R^{n_p}$, given control functions $u : [t_0; t_{end}] \rightarrow R^{n_u}$ and given control variables $q \in R^{n_q}$. It is assumed, that at the given times $t_j, j = 1, \dots, M$, measurements $\eta_{ij}, i = 1, \dots, M_j, j = 1, \dots, M$, of the observation functions b_{ij} are available

$$\eta_{ij} = b_{ij}(y(t_j), p^{true}, q) + \varepsilon_{ij}$$

which are subject to measurement errors ε_{ij} . Here p^{true} are the "true" values of the parameters. Note, that several model quantities can be measured at a time t_j . According to the common approach, in order to determine the unknown parameters an optimization problem is solved. The possible constraints of this problem describe the specifics of the model (constraints on the initial and terminal states, constraints on parameters, etc.) and can be formally written as constraints at times $\theta_1, \dots, \theta_\kappa \in [t_0, t_{end}]$

$$r_{con}(y(\theta_1), \dots, y(\theta_\kappa), p, q) = 0$$

As an objective functional in the optimization problem, typically a norm of the measurement errors is used. The type of the norm is motivated by the statistical distribution of the measurement errors. If the errors are independent, normally distributed with zero mean and known variances $(N(0, \sigma_{ij}^2))$, minimizing a weighted least squares function

$$\min \sum_{i,j} (\eta_{ij} - b_{ij}(y(t_j), p, q))^2 / \sigma_{ij}^2$$

yields a maximum likelihood estimate.

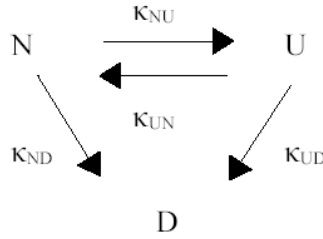
Summing up, mathematically the problems of parameter estimation can be written as follows:

Minimize the deviation of model response to measurement values such that the dynamics and the initial conditions of the dynamic process are fulfilled and possible further constraints are satisfied:

$$\begin{aligned}
& \min_{y, p} \sum_{j=1}^M \sum_{i=1}^{M_j} (\eta_{ij} - b_{ij}(y(t_j), p, q))^2 / \sigma_{ij}^2 \quad (1) \\
& \text{s.t. } \dot{y}(t) = f(t, y(t), p, q, u(t)), t \in [t_0, t_{end}] \\
& r_{con}(y(\theta_1), \dots, y(\theta_\kappa), p, q) = 0.
\end{aligned}$$

Model for enzyme catalytic processes

We consider parameter estimation for the catalytic reaction of enzymes [3]. The native (N), unfolded (U), and deactivated (D) enzymes are involved in the reaction which takes place in a reactor with a continuous inflow and a corresponding outflow. The reaction scheme is the following



where k_{NU} , k_{UN} , k_{UD} , k_{ND} denote rates of the corresponding reactions. Additionally a substrate S is involved, which is contained in the inflow and is degraded depending on the concentration of native enzymes.

Our mathematical model of this reaction scheme is based on the following assumptions:

- The concentrations C_N , C_U , C_D of the native, the unfolded and the deactivated enzymes respectively satisfy a conservation law:

$$C_N + C_U + C_D = C_{E_0}$$

where C_{E_0} denotes an initial amount of active enzymes.

-
- Native and unfolded enzymes are in a quasi-steady-state:

$$\frac{C_U}{C_N} = K_U$$

with some constant K_U .

- A biochemical reaction involving the substrate S takes place. This reaction is described by the Michaelis-Menten kinetics:

$$r = r_{max} \frac{C_S}{k_m + C_S}, r_{max} = k_r C_N$$

where C_S is the concentration of the substrate, k_m is the Michaelis-Menten constant, r_{max} is the maximal rate of the biochemical reaction, k_r is a proportionality coefficient depending on the temperature T .

- The reactions rates k_{NU} , k_{UN} , k_{UD} , k_{ND} , as well as the steady state constant K_U and the proportionality factor k_r depend on the temperature T according to the Arrhenius law:

$$k = k_1 \exp\left(\frac{\Delta k_2}{RT}\right) \quad (2)$$

where R is the gas constant. In particular:

$$k_{ND} = k_N^0 \exp\left(\frac{(-\Delta h_N^*)}{RT}\right), k_{UD} = k_d^0 \exp\left(\frac{(-\Delta h_u^*)}{RT}\right),$$

$$K_U = \exp\left(-\Delta \frac{h_u^0}{RT}\right) \exp\left(\frac{(\Delta S_u^0)}{RT}\right), k_r = A \exp\left(\frac{(-\Delta h_E^*)}{RT}\right),$$

where

- A denotes the activation constant,
 - Δh_E^* denotes the activation enthalpy,
 - ΔS_u^0 denotes the deactivation entropy,
 - Δh_u^0 denotes the deactivation enthalpy,
 - k_d^0 denotes the decay constant,
 - Δh_u^* denotes the activation enthalpy,
 - k_N^0 denotes the native enzyme decay constant,
 - Δh_N^* denotes the native enzyme activation enthalpy.
- The change of the substrate concentration C_S due to the inflow and the outflow of the substrate is defined by

$$\frac{\dot{V}}{V}(C_S^0 - C_S)$$

where V is the volume of the reactor, \dot{V} is the rate of the substrate inflow, C_S^0 is the inflow concentration of the substrate.

Taking into account these assumptions one can derive the mathematical model for the enzyme catalytic reaction

$$\frac{dC_D}{dt} = (k_d^0 \exp(\frac{(-\Delta h_u^*)}{RT}) K_U + k_N^0 \exp(\frac{(-\Delta h_N^*)}{RT})) (C_{E0} - C_D) \frac{1}{(1 + K_U)}, C_D(0) = 0, \quad (3)$$

$$\frac{dC_S}{dt} = \frac{\dot{V}}{V} (C_S^0 - C_S) - r_{max} \frac{C_S}{(k_m + C_S)}, C_S(0) = C_S^0.$$

with

$$K_U = \exp\left(-\Delta \frac{h_u^0}{RT}\right) \exp\left(\Delta \frac{S_u^0}{R}\right), r_{max} = A \exp\left(-\Delta \frac{h_E}{RT}\right) \frac{(C_{E_0} - C_D)}{(1 + K_U)}.$$

Summing up, the catalytic reaction is modelled by a system of two differential equations (3) where

- the state variables are the concentration of the deactivated enzyme C_D and the concentration of the substrate C_S ,
- the parameters to be estimated are

$$\begin{aligned} p_1 &= \ln A, & p_2 &= \Delta h_E^* & p_3 &= \Delta S_u^0 & p_4 &= \Delta h_u^0 \\ p_5 &= k_d^0 & p_6 &= \Delta h_u^* & p_7 &= k_N^0 & p_8 &= \Delta h_N^* \end{aligned}$$

The process may be controlled by the temperature, this means that the control function u is a temperature profile T .

For numerical stability and better scaling we reformulate the Arrhenius kinetic terms (2) as follows:

$$k_1 \exp\left(\frac{-k_2}{RT}\right) = \exp\left(k_{1new} \frac{(T^{-1} - T_0^{-1})}{(T_1^{-1} - T_0^{-1})} + k_{2new} \frac{(T^{-1} - T_1^{-1})}{(T_1^{-1} - T_0^{-1})}\right),$$

where T_1 and T_0 are some temperature values to be chosen by the use, e.g. maximal and minimal temperature used in the experiments. Physically the coefficients k_{1new} and k_{2new} describe the rate of the corresponding reaction at T_1 and T_0 respectively.

We use this transformation for parameters describing reactions $N \rightarrow D$ and $U \rightarrow D$:

- instead of the parameters p_5 and p_6 we introduce the parameters p_{5new} and p_{6new} according to the formulae:

$$k_d^0 \exp\left(\frac{(-\Delta h_u^*)}{RT_0}\right) = \exp(p_{6new}), k_d^0 \exp\left(\frac{(-\Delta h_u^*)}{RT_1}\right) = \exp(p_{5new}),$$

- and analogously instead of the parameters p_7 and p_8 , we introduce the parameters p_{7new} and p_{8new} according to the formulae:

$$k_N^0 \exp\left(\frac{(-\Delta h_N^*)}{RT_0}\right) = \exp(p_{8new}), k_N^0 \exp\left(\frac{(-\Delta h_N^*)}{RT_1}\right) = \exp(p_{7new}).$$

For the reader's convenience, we give the explicit relation between "old" and "new" parameters:

„new“ parameters

$$p_{5new} = \ln(k_d^0) - \frac{(\Delta h_u^*)}{RT_1}$$

$$p_{6new} = \ln(k_d^0) - \frac{(\Delta h_u^*)}{RT_0}$$

$$p_{7new} = \ln(k_N^0) - \frac{(\Delta h_N^*)}{RT_1}$$

$$p_{8new} = \ln(k_N^0) - \frac{(\Delta h_N^*)}{RT_0}$$

„old“ parameters

$$k_d^0 = \exp\left(\frac{(T_1 p_{5new} - T_0 p_{6new})}{(T_1 - T_0)}\right)$$

$$\Delta h_u = (RT_1 T_0 \frac{(p_{5new} - p_{6new})}{(T_1 - T_0)})$$

$$k_N^0 = \exp\left(\frac{(T_1 p_{7new} - T_0 p_{8new})}{(T_1 - T_0)}\right)$$

$$\Delta h_N = (RT_1 T_0 \frac{(p_{7new} - p_{8new})}{(T_1 - T_0)})$$

Half-life

The important stability feature of enzymes is characterized by half-life [2].

Half-life (HL) is a time required to reduce the amount of a native enzyme to a half of the initial amount at a constant temperature.

Mathematically half-life is given by

$$HL(T) = \ln 2 (1 + K_U) / (K_U k_d^0 \exp\left(\frac{(-\Delta h_u^*)}{RT}\right) + k_N^0 \exp\left(\frac{(-\Delta h_N^*)}{RT}\right)).$$

Measurement function

The observation function - velocity of consumption of base necessary to neutralize the acidic reaction product - is given by the dosage of base

$$b: = \frac{\dot{V}(C_S^0 \angle C_S(t))}{B_0}$$

where B_0 is the given concentration of base. Note that in the model under consideration we measure one quantity at a time.

BOUNDARY VALUE PROBLEM METHODS FOR PARAMETER ESTIMATION

A typical solution approach to parameter estimation which is found very often in practice is the initial value or single shooting approach: the ODE system is repeatedly solved as an initial value problem, and unknown parameters including possibly initial values are iteratively improved by some optimization procedure.

In contrast to that, our numerical solution of the parameter estimation problem is based on the Boundary Value Problem (BVP) approach going back to [4]. The basic idea consists in parameterizing the dynamic equations (initial or boundary value problem) like a boundary value problem (e.g., by multiple shooting) and then performing simultaneously (in one iteration loop) the minimization of the cost function and the fulfilment of the constraints given by the discretized boundary value problem. It has been shown [4,5], that BVP methods (based on multiple shooting or collocation) are much more stable and efficient than the single shooting approach when solving parameter estimation problems.

Multiple shooting

The scheme of multiple shooting consists of the following. First one chooses a suitable grid of multiple shooting nodes τ_j

$$t_0 = \tau_0 < \tau_1 < \dots < \tau_m = t_{end}$$

covering the interval where measurements are given.

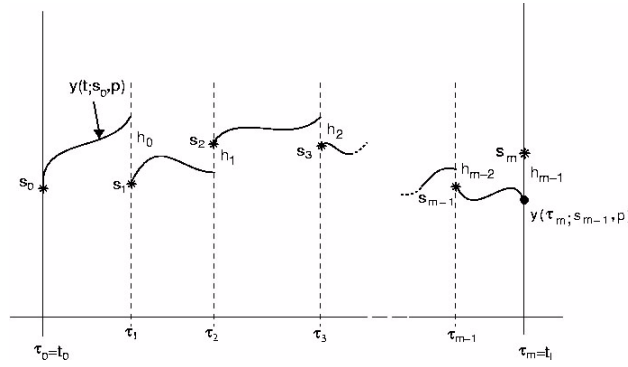


Figure 1. Multiple shooting approach.

At each grid point the values of the state variables s_j are chosen as additional unknowns and m ODE initial value problems

$$\dot{y} = f(t, y, p, q, u), \quad y(\tau_j) = s_j$$

are solved on each subinterval $I_j := [\tau_j, \tau_{j+1}]$ to yield a solution $y(t; s_j, p, q, u)$ for $t \in I_j$. Solutions of dynamic systems, generated by this procedure, are usually not continuous at τ_j . This has to be enforced by additional matching conditions. Inserting the computed values $y(t, s_j, p, q, u)$, $\tau_j \leq t \leq \tau_{j+1}$ into problem (1) one obtains a constrained optimization problem in the variables $(s, p) := (s_0, \dots, s_m, p)$

$$\min_{(s, p)} \sum_{j=1}^M \sum_{i=1}^{M_j} (\eta_{ij} - b_{ij}(y(t_j), p, q))^2 / \sigma_{ij}^2, \quad (4)$$

$$h_j(s_j, s_{j+1}, p) := y(\tau_{j+1}; s_j, p, q, u) - s_{j+1} = 0, \quad j = 1, \dots, m-1,$$

$$r_{con}(y(\theta_1), \dots, y(\theta_K), p, q) = 0.$$

Multiple shooting possesses several advantages:

1. It is possible to include a priori information about the state variables, e.g. from the measurements or from expert knowledge, by a proper choice of initial guesses for the additional variables s_j . Thus, it can be ensured that the initial solutions $y(t; s_j, p, u)$ remain close to the observed data. It can be shown that this damps the influence of poor parameter guesses.

2. The adequate choice of initial guesses for the state variables (and the application of a Gauss-Newton method for the solution of the constrained least squares problem) typically avoids convergence to local minima with large residuals.
3. The scheme is numerically stable. The splitting of the integration interval limits error propagation and makes it possible to solve parameter estimation problems even for unstable or chaotic systems.
4. The BVP discretization induces a very specific structure in the problem equations, which can be exploited in particular for parallelization.

Gauss-Newton method

Parametrization of the dynamics yields a finite dimensional, possibly large-scale, nonlinear constrained least squares problem which can be formally written as

$$\begin{aligned} \min_{\chi \in R^n} \quad & \|F_1(\chi)\|_2^2, \\ \text{s.t.} \quad & F_2(\chi) = 0 \end{aligned}$$

Note, that the equalities $F_2(\chi) = 0$ include the matching conditions induced by multiple shooting. We assume that the functions $F_i : D \subset R^n \rightarrow R^{m_i}$, $i = 1, 2$, are twice continuously differentiable. The number of variables in problem (5) is equal to number of differential equations multiplied by the number of multiple shooting nodes plus the number of parameters. To solve problem (5) we use a generalized Gauss-Newton method according to which a new iterate is (basically) generated by

$$\chi^{k+1} = \chi^k + t^k \Delta \chi^k, \quad 0 < t_k \leq 1, \quad (6)$$

where the increment $\Delta \chi$ is the solution of the following linear constraint l_2 problem at $\chi = \chi^k$

$$\begin{aligned} \min_{\Delta \chi \in R^n} \quad & \|F_1(\chi) + J_1(\chi) \Delta \chi\|_2^2, \quad (7) \\ \text{s.t.} \quad & F_2(\chi) + J_2(\chi) \Delta \chi = 0. \end{aligned}$$

Here, $J_1(\chi)$ and $J_2(\chi)$ denote the Jacobians of $F_1(\chi)$ and $F_2(\chi)$ respectively.

$$\text{rank } J_2(X) = m_2, \text{rank } J = n, J = J(X) = \begin{pmatrix} J_1(X) \\ J_2(X) \end{pmatrix}$$

then a linearized problem (7) has a unique solution $\Delta \chi^k$ and a unique Lagrange vector λ^k satisfying the following optimality conditions

$$J_1^T(X) J_1(X) \Delta \chi^k - J_2^T(X) \lambda^k = -J_1^T(X) F_1(X), \quad (8)$$

$$J_2(X) \Delta \chi^k = -F_2(X).$$

Using (8) one can easily show that $\Delta \chi^k$ can be formally written with the help of a solution operator J^+

$$\Delta \chi^k = -J^+(X^k) F(X^k), \quad F(X) = \begin{pmatrix} F_1(X) \\ F_2(X) \end{pmatrix}.$$

The solution operator J^+ is a generalized inverse, that is it satisfies $J^+ J J^+ = I$, and is explicitly given by

$$J^+(X) = \begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} J_1^T(X) J_1(X) & J_2(X)^T \\ J_2(X) & 0 \end{pmatrix}^{-1} \begin{pmatrix} J_1(X)^T & 0 \\ 0 & I \end{pmatrix}.$$

Choosing the step length t^k by means of classical line search methods based on the exact penalty function

$$T_1(X) := \|F_1(X)\|_2^2 + \sum_{i=1}^{m_2} \alpha_i |F_{2_i}(X)|$$

with sufficiently large weights $\alpha_i < 0$, $i = 1, \dots, m_2$, ensures global convergence. However, it is well known that already in mildly ill-conditioned problems such a step-size strategy may be very inefficient since it may produce very small step-sizes. Therefore we use the "restrictive monotonicity test", see [5,6], that has proved to be very effective in practical applications.

Note, that the generalized Gauss-Newton method (6) - (7) has several advantages. First it does not use second order derivative information and the local linearized problems are linear constrained least squares problems. Under certain regularity assumptions at the solution, the method shows a good linear rate of local convergence. There are problems, however, for which the Gauss-Newton method may have a rather slow local convergence rate or may even fail. The reason is that the linearized model (7), which forms the basis of the Gauss-Newton method, is an inadequate representation of such nonlinear problems, since the second-order information cannot be ignored. These problems are called problems with large residuals. Using SQP-type methods for the nonlinear constrained l_2 parameter estimation problem one could force convergence to a solution even in such a case. However, such solutions are undesirable in a certain sense. We can show that a solution of this type, even if it is a strict minimum of the nonlinear constrained l_2 problem (5), cannot be expected to be a continuous deformation of the "true" parameter values under perturbations caused by the measurement errors. Thus, slow local convergence of the full-step ($t^k \equiv 1$) Gauss-Newton indicates deficiencies in the model or lack of data and can be considered as an advantage of the method. For a detailed analysis see [5].

Solving the linear l_2 problem

At each iteration of a Gauss-Newton method a linear least squares problem (7) has to be solved. First, we make use of the special block structure of the Jacobian $J(\chi)$ which is induced by multiple shooting:

$$J = \left[\begin{array}{cccc|c} D_0 & D_1 & \dots & D_m & D^p \\ \hline G_0^l & G_0^r & & & G_0^p \\ & \ddots & \ddots & & \vdots \\ & & \ddots & 0 & \vdots \\ & & & \ddots & \vdots \\ 0 & & & \ddots & \vdots \\ & & & G_{m-1}^l & G_{m-1}^r & G_{m-1}^p \end{array} \right].$$

Every block column corresponds to the derivatives with respect to the discretization variables and parameters in one subinterval. The block rows with G-matrices are the derivatives of the continuity conditions

$$G_j^l := \partial h_j / \partial s_j \quad G_j^r := \partial h_j / \partial s_{j+1} = -I, \quad G_j^p := \partial h_j / \partial p.$$

The block rows with D-matrices correspond to the derivatives of the functions F_i of the cost functional and the constraints of the nonlinear problem (5) excluding the continuity conditions. We use a fast, stable and efficient structure exploiting decomposition (see [4, 5]) to reduce the large linear least squares problem to a linear least squares problem with smaller dimension. The number of variables in the resulting problem is equal to the number of parameters plus the number of differential equations. This so-called condensed problem may be solved using the methods described in [7].

Statistical sensitivity analysis for the estimates

An important question in parameter estimation is how good the computed estimates are. The answer is provided by sensitivity analysis. If the experimental data is normally distributed then the estimated solution χ^* of the parameter estimation problem is also a random variable which is normally distributed in the first order $\chi^* \sim N(\chi^{true}, C)$ with the (unknown) true value χ^{true} as expected value and the variance-covariance matrix C given by

$$C = C(\chi, q, u) = J^+ \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} J^{+T}. \quad (9)$$

The variance-covariance matrix describes the confidence ellipsoid which is an approximation of the nonlinear confidence region of the estimated variables. The $100(1 - \alpha)\%$ linearized confidence ellipsoid ($0 \leq \alpha \leq 1$) can be described by (see [5])

$$G_L(\alpha, \chi^{true}, q, u) = \{ \chi : \chi = \chi^{true} + J^+ \begin{pmatrix} \eta \\ 0 \end{pmatrix}, \|\eta\|_2^2 \leq v^2(\alpha) \}$$

Here, the probability factor $\gamma^2(\alpha) = \chi_{n \angle m_2}^2$ ($1 \angle \alpha$) where $\chi_{n \angle m_2}^2$ ($1 \angle \alpha$) is the quantile of the χ^2 distribution. The linearized confidence ellipsoid $G_L(\alpha, \chi^{true}, q, u)$ is contained exactly in a box determined by the confidence intervals

$$G_L(\alpha, \chi^{true}, q, u) \subset X \left[\chi_i^{true} - \theta_i, \chi_i^{true} + \theta_i \right],$$

where $\theta_i = \sqrt{C_{ii}}\gamma(\alpha)$. Here, C_{ii} denotes the diagonal elements of the covariance matrix C . The values $\sqrt{C_{ii}}$ are known as standard deviations of the variables χ_i .

Results of parameter estimation for *Candida antarctica* lipase on ionic resin ("Novozym")

Applying the methods of parameter estimation described earlier, "Novozym" shows that the data derived from the initial experiment does not deliver enough information to identify all parameters, see Table 1. The settings of the initial experiment are defined by the temperature shown in Fig. 2. The parameter estimation results show that the information received in the experiment is not at all enough to estimate parameters reliably. Thus we need to design additional experiments in order to provide sufficiently good data. The theoretical justification and methods for design of optimal experiments are briefly described in the next section.

Table 1. Estimated values of parameters \pm standard deviation after parameter estimation.

	Initial temperature profile
p1	27.86 ± 4.42
p2	48.98 ± 10.92
p3	$1.73 \pm 2.39 \times 10^5$
p4	$634.20 \pm 806.00 \times 10^6$
p5	$-1.43 \pm 1.50 \times 10^7$
p6	$-7.50 \pm 4.16 \times 10^7$
p7	-4.15 ± 0.091
p8	-8.63 ± 2.00

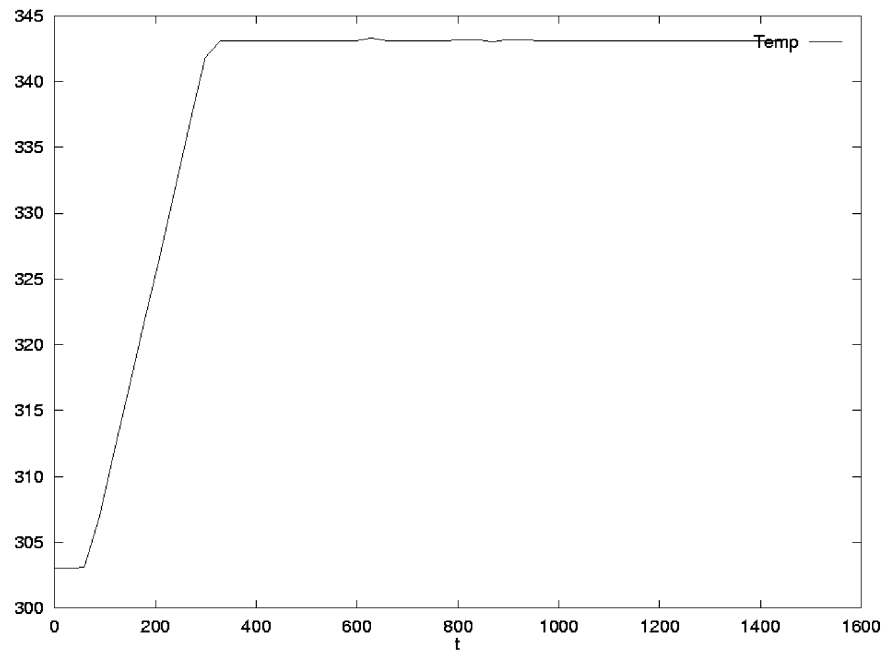


Figure 2. Initial temperature profile

OPTIMUM EXPERIMENTAL DESIGN: PROBLEM STATEMENT AND NUMERICAL METHODS

In this section we consider the problem of nonlinear optimum experimental design and discuss a solution approach to this problem.

Experimental design optimization problem

Since the experimental data is randomly distributed, the estimated parameters are also random variables. Data evaluated under different experimental conditions leads to estimations of parameters with good or poor confidence regions depending on the experimental conditions. We would like to find those experiments that result in the best statistical quality for the estimated parameters and at the same time satisfy additional constraints e.g. experimental costs, safety, feasibility of experiments, validity of the model etc.

The aim of optimum experimental design is to construct N_{ex} experiments by choosing appropriate experimental variables

$q = (q_1, \dots, q_{N_{ex}})$, and experimental controls $u = (u_1, \dots, u_{N_{ex}})$ in order to maximize the statistical reliability of the unknown variables under estimation. For this purpose we optimize a design criterion depending on the variance-covariance matrix C (9).

The optimum experimental design approach leads to an optimal control problem in ODE systems of the following form

$$\begin{aligned} \min_{q, u} \quad & \Phi(C(y, q, u)) \quad (10) \\ \text{s.t.} \quad & c_i(t, y_i(t), p, q_i, u_i(t)) \geq 0, i = 1, 2, \dots, N_{ex}, \quad (11) \\ & \dot{y}_i = f_i(t, y_i, p, q_i, u_i), i = 1, 2, \dots, N_{ex}, \\ & r_{con, i}(y_i(\theta_1^i), \dots, y_i(\theta_{\kappa^i}^i), p, q_i) = 0, i = 1, 2, \dots, N_{ex}. \quad (12) \end{aligned}$$

The constraint (11) describes control and path constraints for each experiment of N_{ex} experiments. Free variables of the optimization problem are the control profiles $u_i(t)$ (e.g. temperature profiles of cooling/heating) and the time-independent control variables q_i (e.g. initial concentrations, properties of the experimental device) for all experiments. Since the "size" of a confidence region is described by variance-covariance matrix C any suitable function of matrix C has to be taken as a cost functional. Typical choices are

$$\Phi(C) = \text{trace}(C)$$

$$\Phi(C) = \lambda_{\max}(C), \text{ where } \lambda_{\max} \text{ denotes the largest eigenvalue of } C$$

$$\Phi(C) = \max_i c_{ii}.$$

Numerical methods

The experimental design optimization problem is a nonlinear constrained optimal control problem. The main difficulty lies in the non-standard objective function which is nonseparable and implicitly defined on the sensitivities of the underlying parameter estimation problem, i.e. on the derivatives of the solution of the ODE system with respect to the parameters and initial values.

The numerical methods are based on the direct approach, according to which the control functions are parametrized on an appropriate grid by support functions, the solution of the ODE systems and the state constraints are discretized. As a result we obtain a finite-dimensional constrained nonlinear optimization problem which is solved by an SQP method. The main effort for the solution of the optimization problem by the SQP method is spent on the calculation of the values of the objective function and the constraints as well as its gradients.

Efficient methods for derivative computations combining internal numerical differentiation [5] of the integration method and automatic differentiation of the model functions [8] have been developed, see [9,10,11]. For more detailed discussion of the numerical methods for nonlinear optimum experimental design see [11,12].

Results of parameter estimation for *Candida antarctica* lipase on ionic resin ("Novozym") with data from additional experiments

Using the methods of optimum experimental design we have optimized 5 additional temperature profiles. The following set-up has been chosen for the experiments: the duration of each experiment is 20 h; measurements are taken every half-hour; the temperature profile is parametrized as follows

$$T_i(t) = T_i + \text{slope}_i(t - t_i), t \in [t_i, t_{i+1}], T_i(t_{i+1}) = T_{i+1}, t_i = 60i, i = 0, \dots, 19, t_{20} = 1200.$$

Additionally, there are the following restrictions on the design parameters T_i and slope.

$$293 \leq T_i \leq 343, i = 0, \dots, 20, -0.12 \leq \text{slope}_i \leq 0.25, i = 0, \dots, 19,$$

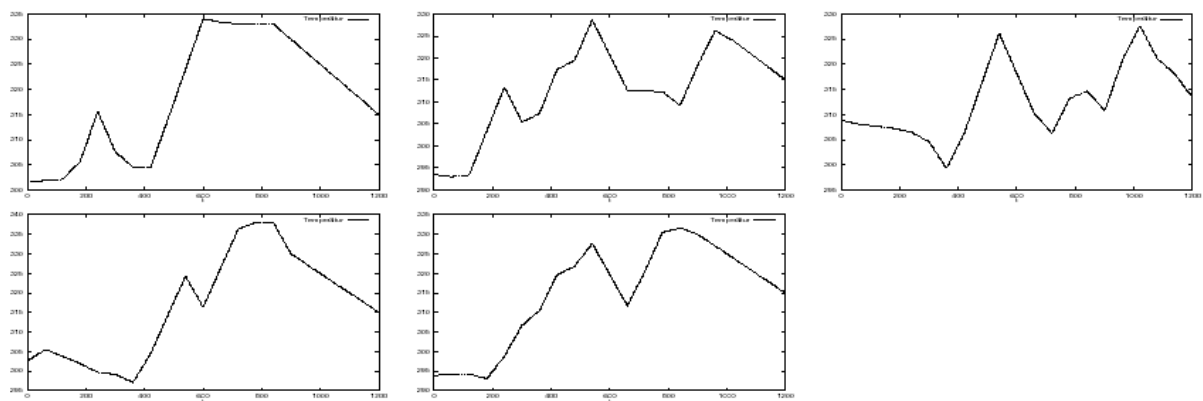
Results of parameter estimation for "Novozym" with data from the initial and additional experiments are presented in Tables 2 and 3. Table 3 shows the estimated values of half-life over an interesting temperature range. The optimal temperature profiles are presented in Fig. 3. Figure 4 presents the fits before and after parameter estimation. The results of parameter estimation show that now we can estimate all half-lives with standard deviation below 30 %.

Table 2. Estimated values of parameters \pm standard deviation after parameter estimation.

	Initial profile	5 designed + 1 initial profile
p ₁	27.86 ± 4.42	27.69 ± 0.81
p ₂	48.98 ± 10.92	49.96 ± 1.97
p ₃	$1.73 \pm 2.39 \times 10^5$	0.548 ± 0.075
p ₄	$634.20 \pm 806.00 \times 10^6$	184.88 ± 26.10
p ₅	$-1.43 \pm 1.50 \times 10^7$	-4.32 ± 0.26
p ₆	$-7.50 \pm 4.16 \times 10^7$	-6.20 ± 2.20
p ₇	-4.15 ± 0.091	-8.88 ± 1.59
p ₈	-8.63 ± 2.00	-11.34 ± 7.54

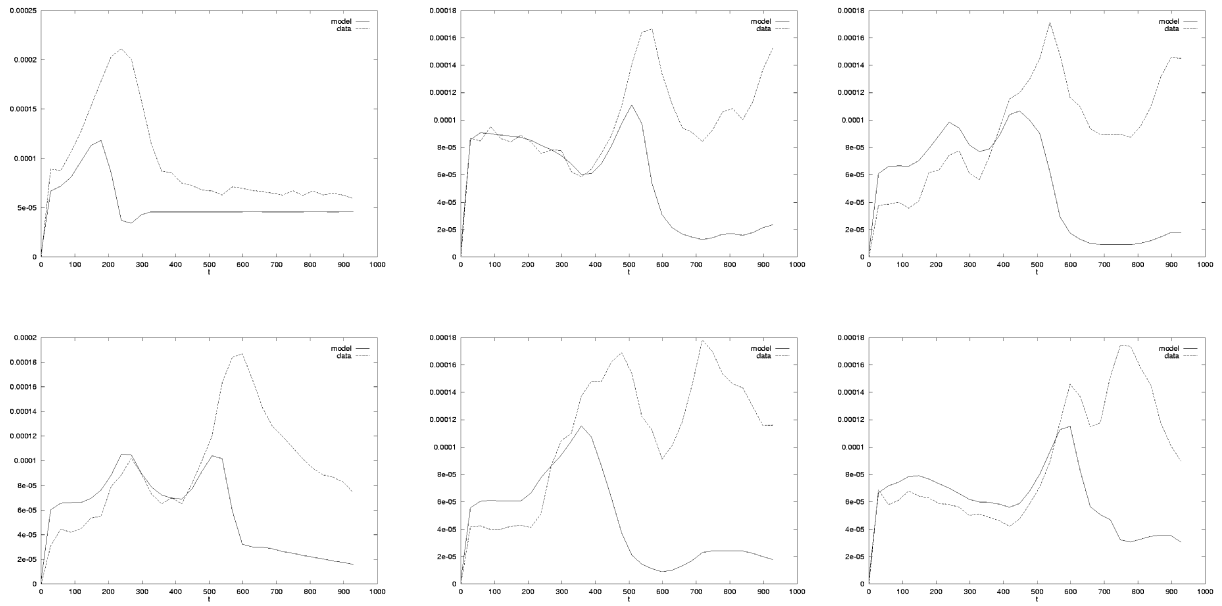
Table 3. Estimated values of half-life \pm standard deviation after parameter estimation.

Temperature [°C]	5 optimal + 1 initial profiles
48.0	1912.01 ± 574.2
50.0	1239.43 ± 266.8
52.0	800.68 ± 121.8
54.0	520.17 ± 57.42
56.0	342.81 ± 30.2
58.0	231.09 ± 17.4
60.0	160.53 ± 10.4

**Figure 3.** Optimal additional temperature profiles.

A)

Before parameter estimation

**B)**

After parameter estimation

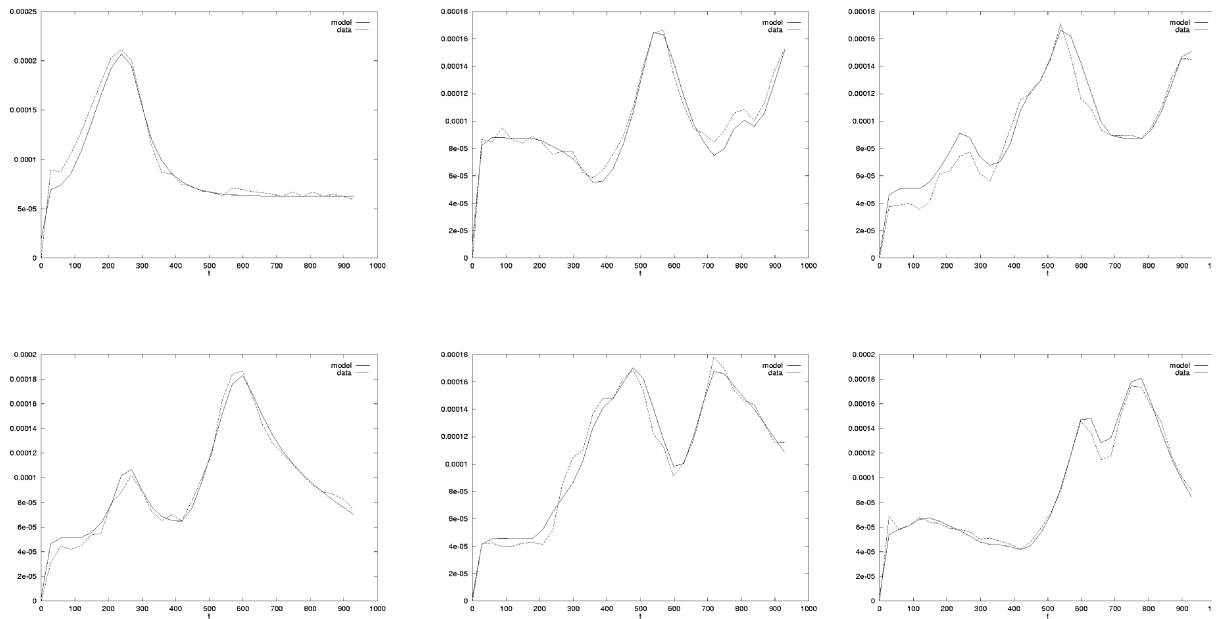


Figure 4. Model (straigth line) vs. measurements (dashed line) before (A) and after (B) parameter estimation.

DESIGN OF ROBUST OPTIMAL EXPERIMENTS OF NONLINEAR MODELS

As we can see the experimental design optimization problem (10), (11), (12) is formulated for the assumed parameter values which are, however, only known to lie in a possibly large confidence region. In this section we discuss how to construct robust experiments, that is experiments that are less sensitive to parameter uncertainty. We assume that the parameters are lying in the following ellipsoid around the nominal values of parameters p^0

$$E := E(v, p^0) = \{p : (p - p^0)^T \Sigma^{-1} (p - p^0) \leq v^2\}.$$

Here Σ is a positive definite matrix. To get a robust experimental design we formulate a worst-case problem, minimizing over the design variables ξ , the maximal value of $\phi(C)$ over the ellipsoid E

$$\min_{\xi \in \Omega} \max_{p \in E} \phi(C(\xi, p)). \quad (13)$$

In this problem, we first consider constraints which do not depend on the parameters. The remaining (control) constraints are summarized by $\xi \in \Omega$. The optimization problem (13) is a semi-infinite programming problem. The solution methods for such problems require the determination of global optima of nonlinear subproblems [13] which may be computationally too expensive. In order to compute robust designs we suggest a modified approach. We apply Taylor expansion w.r.t. p to the min-max objective function:

$$\min_{\xi \in \Omega} \max_{p \in E} \left(\phi(C(\xi, p^0)) + \frac{\partial}{\partial p} \phi(C(\xi, p^0)) (p - p^0) \right).$$

The inner problem is the maximization of a linear function subject to a convex quadratic constraint which can be solved explicitly:

$$\max_{p \in E} \left(\phi(C(\xi, p^0)) + \frac{\partial}{\partial p} \phi(C(\xi, p^0)) (p - p^0) \right) = \phi(C(\xi, p^0)) + v \left\| \frac{\partial}{\partial p} \phi(C(\xi, p^0)) \Sigma^{1/2} \right\|_2$$

This leads to the robust experimental design optimization problem

$$\min_{\xi \in \Omega} (\phi(C(\xi, p^0)) + \nu \left\| \frac{\partial}{\partial p} \phi(C(\xi, p^0)) \sum_{i=1}^2 \right\|).$$

The second term in the cost function can be interpreted as a penalty for uncertainty in the parameters.

Parameter dependent constraints are treated in an analogous way. We substitute the nonrobust constraints

$$\psi(\xi, p^0) \leq 0,$$

with the following constraints for the robust experimental design optimization problem.

$$\psi_R(\xi, p^0) := \psi(\xi, p^0) + \nu \left\| \frac{\partial}{\partial p} \psi(\xi, p^0) \sum_{i=1}^2 \right\| \leq 0.$$

Numerical results

In this section we present numerical results on comparing robust and nonrobust designs for enzyme reaction kinetics. To illustrate the performance of the method, experiments for identification of 4 out of 8 parameters were designed in the sequential mode. The idea of the sequential approach is schematized in Fig. 5.

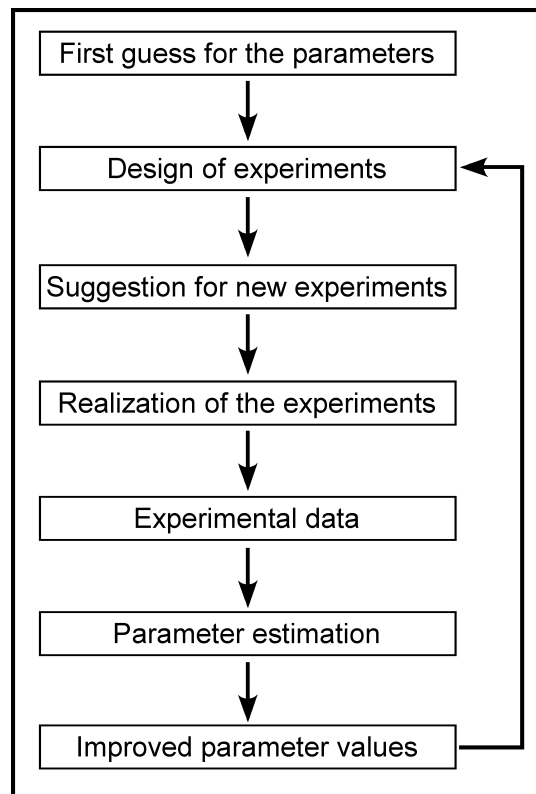


Figure 5. Sequential approach to experiment design.

For the "true" values of the parameter $p_1^{true} = 27.77$, $p_2^{true} = 50.15$, $p_3^{true} = 0.55$, $p_4^{true} = 185.25$ the normally distributed data were simulated using initial temperature profile as shown in Fig. 2.

The results of parameter estimation based on these data

$$p_1 = 48.61 \pm 20.59$$

$$p_2 = 103.58 \pm 52.77$$

$$p_3 = 0.58 \pm 0.10$$

$$p_4 = 189.69 \pm 32.50$$

were used for the computation of a nonrobust and a robust designs. The diagonal matrix with diagonal elements of the covariance matrix corresponding to the solution of parameter estimation problem was chosen as the matrix Σ in the robust design. Using the new optimized temperature profiles the data were simulated, parameter estimation was repeated with the new data and further new designs were constructed using the new estimates for the parameters. This procedure was repeated until we obtained trustworthy parameter estimates, see Table 4.

Table 4. Results of parameter estimation for initial design and 5 non-robust experiments (second column) and for initial design and 2 robust experiments (third column).

"true" values	$p_i \pm \sqrt{c_{ii}}$	$p_i \pm \sqrt{c_{ii}}$
$p_1 = 27.77$	27.44 ± 0.743	27.63 ± 1.266
$p_2 = 50.15$	49.33 ± 1.980	49.71 ± 3.330
$p_3 = 0.55$	0.620 ± 0.068	0.512 ± 0.072
$p_4 = 185.25$	209.85 ± 24.35	172.26 ± 24.41

The sequential design of robust experiments after 2 designs provide us with reliable parameter estimates, while we need 5 nonrobust experiments to get similar quality of the estimates. Although a computational time for computing a robust experiment is significantly greater than the time to compute a nonrobust experiment, clearly application of the robust sequential design makes it possible to reduce the number of real experiments and thus to reduce drastically experimental costs and the time necessary to identify the parameters.

CONCLUSIONS

Application of advanced methods for parameter estimation in dynamic processes governed by ODE and for design of robust optimal experiments allows reliable estimation of enzyme operating stability and the reduction of the number of experiments as well as experimental costs.

ACKNOWLEDGEMENTS

We thank the Deutsche Forschungsgemeinschaft (DFG) for financial support through SFB 359. The part of the research was made in cooperation with Degussa AG.

REFERENCES

-
- [1] Gupta, M.N. (1993) *Thermostability of Enzymes*. Springer, New York.
 - [2] Boy, M., Dominik, A., Vos, H. (1998) A method for fast determination of biocatalyst process stability. *Chem. Engng Technol.* **21**: 570-575.
 - [3] Bommarius, A., Estler, M., Kluge, A., Werner, H., Vollmer, H., Bock, H.G., Schlöder, J.P., Kostina, E. (2001) Method to determine the process stability of enzymes. Patent Application EP 1 067 198 A1, European Patent Office, *Patentblatt* 2.
 - [4] Bock, H.G. (1981) Numerical treatment of inverse problems in chemical reaction kinetics. In: *Modelling of Chemical Reaction Systems*. (Ebert, K.H. Deuflhard, P., Jäger, W. Eds) Springer Series in Chemical Physics **18**, Heidelberg.
 - [5] Bock, H.G. (1987) Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen. *Bonner Mathematische Schriften* **183**: 264.
 - [6] Bock, H.G., Kostina, E., Schlöder, J.P. (2000) On the role of natural level functions to achieve global convergence for damped Newton methods. In: *System Modelling and Optimization: Methods, Theory and Applications*. (Powell, M.J.D., Scholtes, S. Eds) pp. 51-74. Kluwer, Boston.
 - [7] Stoer, J. (1971) On the numerical solution of constrained least squares problems. *Siam J. Numer. Anal.* **8**(2): 382-411.
 - [8] Griewank, A. (2000) *Evaluation Derivatives. Principles and Techniques of Algorithmic Differentiation*. Frontiers in Applied Mathematics. SIAM
 - [9] Bauer, I., Bock, H.G., Körkel, S., Schlöder, J.P. (1999) Numerical methods for initial value problems and derivative generation for DAE models with application to optimum experimental design of chemical processes. In: *Scientific Computing in Chemical Engineering II*. (Keil, F., Mackens, W., Voss, H., Werther, J. Eds) **2**, pp. 282-289, Springer, Berlin.
 - [10] Bauer, I., Bock, H.G., Körkel, S., Schlöder, J.P. (2000) Numerical methods for optimum experimental design in DAE systems. *J. Comput. Appl. Math.* **120**:1-25.
 - [11] Körkel, S. (2002) *Numerische Methoden für Optimale Versuchsplanungsprobleme bei nichtlinearen DAE-Modellen*. PhD thesis, Universität Heidelberg.
 - [12] Körkel, S., Bauer, I., Bock, H.G., Schlöder, J.P. (1999) A sequential approach for nonlinear optimum experimental design in DAE systems. In: *Scientific Computing in Chemical Engineering II*. (Keil, F., Mackens, W., Voss, H., Werther, J. Eds) **2**, pp. 338-345, Springer, Berlin.
 - [13] R. Reemtsen, R., Rückmann, J.-J. (Eds.) (1998) *Semi-Infinite Programming. Nonconvex Optimization and its Applications*. Kluwer, Boston.
-

BROAD-RANGE METABOLITE ANALYSIS: INTEGRATION INTO GENOMIC PROGRAMS

ALISDAIR R. FERNIE^{1*} AND LEE J. SWEETLOVE²

¹Max-Planck-Institut für Molekulare Pflanzenphysiologie,
Am Mühlenberg 1, 14476 Golm, Germany

²Department of Plant Science, University of Oxford,
South Parks Rd, Oxford, OX1 4RB, U.K.

E-Mail: * fernie@mpimp-golm.mpg.de

Received: 13th April 2004 / Published: 1st October 2004

ABSTRACT

In recent years the focus of experimental biology has shifted from reductionist towards more holistic approaches. This shift has been driven by the development of genetic tools that have allowed the creation of an unprecedented base of genetic diversity and by the development of technologies allowing the rapid determination of the genetic, transcript, protein and metabolite complements of biological systems. Here we will describe experiences with broad-range metabolite analysis of potato and tomato development over the last few years: we will furthermore describe what information can be garnered from these experiments as well as describing recent attempts to analyse systems at the level of more than one molecular entity. Finally, the need for interdisciplinary collaboration and a perspective for this research field will be discussed.

METHODOLOGY

The landmark *Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and human genome sequencings facilitated the emergence of systems biology - a science that is currently progressing on both experimental and theoretical fronts. Experimentalists carry out comprehensive and high-throughput analyses of the various molecule entities of the cell, namely, mRNAs, proteins and metabolites (for a review see [1]), whereas theoreticians

concentrate on the analysis of the regulatory interactions of these molecules and predicting the effects of changing the state of the system (e.g. [2, 3]). Although complementary, these approaches have normally been taken in isolation and for this reason we will discuss them separately. Transcript analysis by hybridization is a relatively mature technology and has been the focus of high quality studies in many biological systems [4-7]. These studies have allowed the determination of differential gene expression under a range of environmental or developmental conditions as well as the identification of the high level of co-ordinated, correlated changes across genes and ultimately to the establishment of gene regulatory networks. Furthermore, of the various levels of analysis it is the only one which can be said to be truly comprehensive: with the current proportional coverage of profiling technologies decreasing following the order mRNA > protein > metabolite. For this reason transcript profiling is currently the experimental approach of choice for systems biology [8, 9]. That said, the emergent technologies of proteomics [10-12] and metabolomics [13, 14] have made rapid progress in recent years and methods such as ICAT (proteins) and mass spectrometry coupled chromatography (metabolites) now allow rapid profiling of these molecules at high analytical precision. Technical aspects of the most commonly used platforms have been reviewed elsewhere (e.g. [15, 16]) so we will not detail them here but rather discuss their importance within integrated approaches.

SYSTEMS APPROACHES

Although a wealth of data can be obtained when using any of the genomic technologies described above, it is clear that transcription, translation, post-translational modifications and the turnover of mRNA, proteins and metabolites do not occur in isolation but are heavily interconnected with one another (for an example of the level of complexity involved see [17]). For this reason it makes sense to move toward integrated approaches wherein transcripts, proteins and metabolites are measured from the same sample. Such approaches have been carried out recently in the microbial and medicinal fields. In one recent study the transcriptome, proteome and protein interactions of 20 systematic perturbations of galactose utilization in the yeast *Saccharomyces cerevisiae* were monitored providing evidence that approximately 15 of 289 detected proteins are regulated post-transcriptionally, and identify explicit physical interactions governing the cellular response to each perturbation [18].

A similar approach combining large-scale perturbation analyses, in combination with computational methodologies, genomic data, cis-regulatory analysis, and molecular embryology was used to define a regulatory network underpinning sea urchin development and revealing how given cells generate their ordained fates in the embryo [8]. These experiments are by nature multidisciplinary and multi-laboratory in that they combine molecular and cellular biologies with protein biochemistry and bioinformatics. While this does not pose a problem in itself, it is vital to note that in order to gain a meaningful insight into interactions between the various molecular entities it is imperative that experiments are performed on samples that are spatially, temporally and micro-environmentally identical. In the following section we intend to describe similar approaches we have taken in plants, albeit on a smaller scale, and to propose a framework by which large-scale systematic approaches can be achieved in plants.

Many recent plant studies have provided data sets comprising transcripts, enzyme activities and metabolites including those focused on various aspects of carbon metabolism interactions including the responses to nitrate [19, 20] and to diurnal changes [21], as well as studies of individual branches of secondary metabolism such as the recent study of the early stages of triterpene saponin biosynthesis [22]. Even though these studies are restricted in their coverage, the potential of this approach is still visible. Whilst in some cases changes in transcripts are accompanied by changes in enzyme activities and shifts in metabolism [20, 22], in other cases they are not (for example the impact of sugars and light and carbon on nitrate reductase: [19]), indicating a major role for post-transcriptional modification in the regulation of the pathways involved. Expansion of such approaches using the tools at hand has vast potential in aiding the understanding of the complex change underlying diverse patterns, such as development and circadian rhythms. We have recently studied potato tuber development and plants exhibiting altered sucrose metabolism using both metabolic and transcript profiling techniques [23]. A comparison of the discriminatory power of metabolite and RNA profiling to distinguish between different potato tuber systems suggests that metabolite profiling has a higher resolution than RNA profiling. Furthermore, when performing comparisons across the molecular entities we established a correlation between 571 of the 26, 616 possible metabolite: transcript pairs. Most of these observations were novel and notably included several strong correlations to nutritionally important compounds, such as vitamins and essential amino acids [23]. We therefore believe that this combinatorial approach is of high potential value in the identification of candidate genes for modifying the metabolite content of biological systems.

INTERPRETATION OF RESULTS HARVESTED FROM SYSTEMS BIOLOGICAL APPROACHES

Although it is probably too early to comment on the effectiveness of reverse genetic versus environmental system perturbation it is important to note that there is a fundamental difference in the data obtained via these methods. Although genetic changes give insight into mechanisms of system robustness and into gene/protein functionality, unless you have a range of gene inhibitions, they do not allow an interrogation of control (since removal of a gene-product merely tells you how well the system copes without it). In contrast, environmental perturbations will result in effects at many genetic loci and it will therefore be very hard to attribute function to any particular gene/protein. However, the ability to perturb the system to a variety of extents may allow the identification of regulatory control points. The interpretation of data sets resulting from the latter example is however clearly far more complicated. Ideally, as many different (types of) system perturbation should be applied (see [18]) and in many instance the use of transgenic lines expressing a range of activities of an enzyme would be preferable to single knock-out mutants. A further advantage of using both environmental and genetic perturbations, is that the identification of common patterns of changes following the different experimental approaches, allows greater surety that the recorded response is a direct result of the desired perturbation, rather than a pleiotropic artefact of the method used to elicit the perturbation.

Inverting this argument, the use of systems perturbations can be utilized to infer common mechanisms by which plant cells respond to different treatments, for example in studying the metabolic complements of variously modified potato tuber systems, Roessner et al. [24] revealed that those expressing a yeast invertase at an apoplastic location could be faithfully phenocopied by feeding glucose (and to a lesser extent fructose) to potato tuber discs. Whilst this example is somewhat trivial (since the resultant conclusions are what would be expected) it highlights the possibilities open for large-scale systematic approaches.

WHAT ELSE DO WE NEED TO ASSAY?

As mentioned above proteomic and metabolomic coverage is far from complete. However, with few exceptions the majority of pathways of plant primary metabolism can be studied in detail with the tools presently available at the protein [25, 26] and intermediary metabolite level [14, 24] in addition to the RNA level.

These studies have allowed description of the mitochondrial proteome and have uncovered new molecular mechanisms involved in mitochondrial defence, as well as cataloging degradation of sensitive protein components including TCA cycle enzymes in response to oxidative stress. Such studies have also highlighted the degree of systemic change following a relatively simple perturbation such as altered hexose supply. Furthermore, metabolic profiling of potato tubers expressing more efficient pathways of sucrose degradation revealed that the tuber contained all the necessary biosynthetic machinery for the *de novo* biosynthesis of amino acids - knowledge that is a prerequisite for any rational attempt to modify free amino acid content in this tissue. In the majority of instances approaches such as these are adequate to (or even preferable to sifting through data from esoteric pathways) answer the biological question raised. That said, it is clear that for certain approaches, such as unravelling circadian or developmental patterns, truly holistic approaches are needed. For such questions further developments are required in the areas of proteomics and metabolomics.

Whilst new proteomic methodologies (e.g. MudPIT in conjunction with ICAT) dispense with 2D gels and allow high-throughput analysis of thousands of proteins [27, 28] and protein CHIP technology may ultimately allow a simultaneous analysis of the entire proteome [29, 30], increasing coverage of the metabolome presents a greater challenge. Indeed metabolite analysis within systems approaches faces something of a dichotomy, since on the one hand it is important to increase the scope of the metabolites measured (without compromising the accuracy of the measurements) and on the other it is necessary to gain greater understanding of the subcellular levels of metabolites.

Although LC-MS based methodologies have recently been developed allowing the measurement of several important classes of secondary metabolites including alkaloids, flavonoids, glucosinolates, isoprenes, oxylipins, phenylpropanoids, pigments and saponins, and high-throughput spectrophotometric assays have been developed for GC-MS unfriendly primary metabolites such as phosphorylated intermediates and acetyl CoA (reviewed in [16]), it is clear that expansion of the coverage of metabolite profiling methods remains a daunting task. The problem of obtaining information on subcellular information is particularly acute for metabolites which lack the targeting signatures of proteins and turnover too rapidly to allow measurement following aqueous fractionation procedures regularly used in protein analysis (see [25]).

Despite these problems, methods have been developed to obtain the subcellular information on metabolite levels in intact plants - which is ultimately essential to allow accurate modelling of metabolism. The first of these methods, non-aqueous fractionation of lyophilized material involves the separation of small cell portions on an organic density gradient and the use of simultaneous equations to estimate metabolite concentrations with respect to marker enzymes of various organelles in the cell [31]. A second method involves the production of chimeric proteins that differentially fluoresce upon the binding of a certain metabolite. Such proteins can now be created by the fusion of periplasmic binding proteins to green fluorescent proteins and subsequent monitoring of fluorescence energy resonance transfer (FRET), allowing imaging of changes in the concentration of selected metabolites in real time [32]. Although both methods have the potential to provide subcellular spatial information, they both have severe drawbacks - current methods of non-aqueous fractionation only allow the discrimination of three compartments the vacuole, plastid and cytosol, whereas multiple independent chimeric proteins are required for each metabolite measured using the FRET approach - suggesting further research effort will be needed to refine such procedures. However, it is clearly preferable to use the estimated plastid metabolite concentrations than the average cellular concentrations when modelling plastidial metabolism, and the coupling of the non-aqueous fractionation method to GC-MS-based metabolite profiling methods, gave insight into the subcellular distribution of a wider range of metabolites than had been determined to date [33].

In addition to understanding steady-state metabolite levels, it is imperative that high-throughput methods of determining cellular flux between these metabolites are developed for plants. Although frameworks for such experiments have existed for many years in medicinal and microbial sciences (see [34], flux studies in plants are normally carried out using low resolution, highly time consuming protocols based on following the redistribution of radiolabelled substrates. Whilst these are useful in gaining information on the bulk flow through the major pathways of primary metabolism (for an example see [35]) they offer little information about other pathways and focus largely on pathway endpoints. More comprehensive methods that are commonly used in microbial sciences, utilize a combination of stable carbon isotope labelling and NMR or MS-based detection systems, in order to determine positional information of the fed label throughout metabolism.

This method offers the advantage that the labelling pattern of metabolic intermediates can be studied and the position of labelling within the carbon skeleton of end products can allow retrospective evaluation of the metabolic route by which they were formed.

Although used to a limited extent in studies of substrate cycles in primary metabolism [36], perhaps the best example of this technology to date is its use in understanding of the metabolism of storage lipids and proteins in developing *Brassica napus* [37]. In this paper the authors demonstrated the different bioenergetic contributions that amino acids make to the cytosolic and plastidial acetyl CoA pools, highlighting the potential of flux analysis in the understanding of both pathway importance and location. A further example of the importance of both flux measurements and understanding of subcellular location is provided by our recent finding that enzymes of glycolysis are functionally associated with the mitochondria in *Arabidopsis* [38]. We established this using a combination of proteomic analyses of a highly purified mitochondrial fraction which we then confirmed by enzyme activity assays. The sensitivity of these activities to protease treatments indicated that the glycolytic enzymes are present on the outside of the mitochondrion. Furthermore, when supplied with appropriate cofactors, isolated, intact mitochondria were capable of the metabolism of C-13-glucose to C-13-labelled intermediates of the trichloroacetic acid cycle, suggesting that the complete glycolytic sequence is present and active in this subcellular fraction. Whilst highly novel in plants such associations have been previously reported in *Tetrahymena pyriformis* [39] and more recently in human heart [40]. It is clear that the understanding of such localized pathways is of critical importance for the design of reliable metabolic models.

In addition to the development of efficient flux phenotyping platforms is the continued development of the analysis of higher order modules. Whilst protein-protein interactions which inform assignation of gene function and represent another method of pathway definition have been the focus of much recent research effort [41-43], others such as protein-DNA [44] and protein-lipid [4] interactions have only just begun to be studied at this level. If the analysis of these higher order modules can be carried out in parallel with analysis at the various independent levels under a range of environmental and developmental conditions, then a far greater understanding of factors governing metabolic regulation would be achieved. The hope being that once regulatory properties of the system are known at this level we can build accurate models of metabolism that could be interrogated *in silico* to facilitate the design of rational engineering strategies to generate desired properties in plants.

INTERPRETATION OF DATA SETS

The analysis of molecular entities in a high-throughput manner and on a global scale places a particular premium on the extraction of meaning from extremely large data sets. Currently, the largest data sets tend to be those that contain mRNA transcript abundances and many of the techniques of data analysis that we will describe have been applied principally to transcriptomic data. Nevertheless, it is worth pointing out that the problems and solutions are the same, whether one is dealing with mRNA-transcripts, proteins or metabolites. Microarrays are now available that can be plausibly described as genomic (the Affymetrix ATH1 genechip reports on some 24,000 *Arabidopsis thaliana* genes) and their use generates data sets that consist of tens of thousands of data-points. With such a high data density, it is difficult to display the entire data set in a way that allows meaningful interpretation. Instead, methods must be used to filter the data to retrieve only that data which satisfies criteria relevant to the experimental query. These criteria can be manifold, but generally the most informative are those based on either the degree of change of the data value between two conditions or on similarity of change of data value.

Many microarray-based experiments are a simple pairwise comparison between two conditions, generally a control and a genetic variant or a different treatment (environmental condition, pathogen attack etc.) and the aim is to identify genes that are differentially expressed between the two conditions. These genes can be identified by filtering the data according to a simple heuristic rule such as an expression cut off. Typically, this only considers genes to be differentially expressed if there is a change in relative mRNA transcript abundance above a defined threshold; typically 1.5-2.0 fold (see for example [6, 24]). While this method allows a rapid filtering of the data, it will introduce a high number of false negatives since many biologically-relevant changes may occur at a level that is below this arbitrary threshold. The need to identify such changes has led to the application of more sophisticated statistical methods to the analysis of array data [46]. These include: the nonparametric t-test, the Wilcoxon rank sum test and the ideal discriminator method [47]. Each of these methods performs differently, with some being more conservative than others. Such statistical approaches greatly reduce the number of false negatives and impart some much needed rigour to the analysis of differential gene expression on microarrays.

While differential expression of genes (or indeed changes in protein and metabolite abundances) provides the primary level of information as to the changes that accompany a biological event, there is much more to be gained from such data sets than just establishing what goes up and what goes down. Often, we are not so much interested in which genes are altered in expression as we are in the *relationships* between those changes. In particular, statistical techniques can be used to group data together on the basis of common patterns of changes. Such groupings are particularly useful in the context of deciphering gene function, since it is likely that elements that share the same function, or participate in the same process, are coordinately regulated [48]. Co-responses of elements of unknown- and known-function allow novel functional associations to be proposed [49]. This type of analysis has been used to propose novel functions of genes [6], their protein products [49] and also to discriminate coordinately regulated groups of metabolites [14] and is one of the principal ways in which system network structure can be established. Two methods have been used to group data points: hierarchical clustering analysis (HCA) [50] and principal components analysis (PCA) [51].

The former considers data objects as points in n -dimensional space or as n -dimensional vectors (where n is the number of samples for comparison) and measures the distance (or similarity) between these objects in n -dimensional space. This distance matrix can then be clustered using standard clustering algorithms and the results presented as a dendrogram [52]. Recently, more sophisticated clustering algorithms have been developed that automatically calculate the optimal distribution of data objects over clusters and overcome problems related to robustness and the establishment of optimal linear ordering of the cluster [53]. The complementary method of PCA also establishes n -dimensional vectors and focuses on the vector that gives the greatest separation between samples, the so-called principal component. The results are usually displayed as a two-dimensional plot with the first principal component on the x-axis and the second principal component on the y-axis. A refinement of this approach is to use a supervised projection method such as discriminant function analysis (DFA) [54] which exploits user-defined information (such as which samples are replicates of one another) to determine within- and between-group variation. This information is then used in combination with principal components to define discriminant functions that separate the groups.

COMPUTATIONAL BIOLOGY

The ultimate aim of systems biology is the construction of a complete mathematical description of the system. The idea is to exploit the increasing experimental knowledge of the behaviour of the system in terms of different molecular entities, to devise a set of rules that accurately model the behaviour of the system. There are clear advantages to such a mathematical approach. First, such a model represents a way of storing and describing our current understanding of a system. Second, mathematical models allow *in silico* interrogations of the behaviour of the system. Such *in silico* experiments have the advantage of both speed and power over conventional "wet" experimentation. Speed, in that any number of parameters can be varied and the effect of these variations on the system behaviour or output can be calculated virtually instantaneously. The power of the approach is derived from the fact that the effects of such changes are calculated in the context of the entire system, the behaviour of which is too complex to be understood intuitively. The realization of this aim of constructing a single model that describes an entire system is still some way off. However, models are being devised that describe discrete parts of the system and can be thought of as modules. It is likely that these modules can be joined together to form a higher order model that represents a more complete coverage of the system.

Most of the mathematical models that have been devised, concentrate on describing the control of flux through a metabolic pathway in terms of the amounts of the enzymes present and tend to ignore other levels of the regulatory hierarchy (e.g. regulatory-gene expression). The reason for this concentration on enzymes is that there are well established mathematical frameworks, such as metabolic control analysis (MCA), that describe the control of metabolic pathways in terms of enzymes and metabolites. Such frameworks are an essential starting point for mathematical models. Indeed, most of the models of plant metabolism to date have used MCA as the basic conceptual tool (e.g. [55, 56]). These models generally use kinetic parameters of enzymes to derive the control structure. An alternative approach is that of metabolic flux analysis (MFA) which can be used to interpret *in vivo* metabolic flux data and derive the metabolic network and its control structure [57]. Both approaches suffer from the fact that the experimental inputs required are generally not known in their entirety for a given system. In the case of models based on kinetic parameters of enzymes, a 'mix and match' approach is often taken where kinetic constants of enzymes from different organisms are combined.

Although core pathways such as glycolysis are highly conserved, there are nevertheless key differences in their regulation between different organisms and such an approach is unlikely to adequately describe the subtleties of different control structures in different organisms. In the case of metabolic flux analyses, the situation is slightly better as there are a number of theoretical approaches such as metabolic-flux balancing [58] and optimization approaches [59] that can be used to derive unknown fluxes. Recently, a new approach to the modelling of metabolic pathways has emerged that is based upon the principle of stoichiometric analysis [60].

The basis of this approach is to define elementary flux modes - non-decomposable sub-networks that account for every possible flux within the network. Elementary flux modes are non-decomposable in the sense that each mode contains a minimal set of enzymes such that if only enzymes belonging to this set are operating, then complete inhibition of one of these enzymes would lead to a complete cessation of pathway flux [2]. This approach allows one to mathematically define and describe all metabolic routes that are both stoichiometrically and thermodynamically feasible and is an extremely useful tool for the definition of network structure [61]. When applied on a sufficiently large scale the approach allows cellular behaviour to be reconstructed from network topology and thus represents a genuine systems analysis [62]. Although stoichiometric analysis concentrates on enzymes and metabolic pathways, the related approach of gene circuit analysis [63] deals with regulatory modules that operate at the genetic level. An integration of these two approaches would potentially describe the majority of regulatory features that are known to occur in a metabolic network and bring us ever closer to a truly holistic description of a biological system. However, to achieve such an integration will require much concerted effort between experimentalists and theoreticians.

CONCLUSION

The availability and continued improvement of high-throughput analytical techniques has brought about a distinct shift in the way biologists are approaching the solution of metabolic control networks. Instead of the reductionist enzyme-by-enzyme approach, we are instead attempting to take a more system-wide approach, in which a comprehensive analysis of a broad range of molecular responses across the system are made and are used as the basis of a holistic understanding of the system driven by computational methods.

While this new approach has the potential to provide a quantum leap in our ability to understand the control of metabolic networks, fulfilment of that potential will ultimately depend on a number of key developments. First, several analytical challenges need to be met to ensure that analysis of all types of molecular entity is as comprehensive as possible. These challenges include a broadening of the coverage afforded by protein and metabolite profiling technologies with the latter representing a considerable obstacle. In addition, the analysis of these molecules needs to be refined to include single cells and different sub-cellular compartments. Second, the field of computational biology needs to continue its progress in developing increasingly sophisticated tools and approaches to the extraction of biological meaning from genomic data sets. It is clear that the true success of the systems biology approach will be determined by the extent to which theoretical and computational biologists can work together with experimental biologists. The latter need to adopt new experimental strategies that are specifically tailored to the systems approaches, while the former need to ensure that computational tools are made available to the wider community such that the systems approach becomes a standard part of the experimental arsenal and its use extends beyond a handful of test cases.

REFERENCES

- [1] Oliver, S.G. (2002) Functional genomics: lessons from yeast. *Phil. Trans. R. Soc. Lond. B* **357**: 17-23.
- [2] Schuster, S., Dandekar, T., Fell, D.A. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.* **17**: 53-60.
- [3] Brazhnik, P., de la Fuente, A., Mendes, P. (2002) Gene networks: how to put the function in genomics. *Trends Biotechnol.* **20**: 467-472.
- [4] Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D., Brown, P.O. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* **283**: 83-87.
- [5] Furlong, E.E.M., Andersen, E.C., Null, B., White, K.P., Scott, M.P. (2001) Patterns of gene expression during *Drosophila* mesoderm development. *Science* **293**: 1629-1633.
- [6] Ramonell, K.M., Somerville, S. (2002) The genomics parade of defense responses: to infinity and beyond. *Curr. Opin. Plant Biol.* **5**: 291-294.
- [7] Ruuska, S.A., Girke, T., Benning, C., Ohlrogge, J.B. (2002) Contrapuntal networks of gene expression during Arabidopsis seed filling. *Plant Cell* **14**: 1191-1206.
- [8] Davidson, E.H. et al. (2002) A genomic regulatory network for development. *Science* **295**: 1669-1678.

Broad-Range Metabolite Analysis

-
- [9] Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**: 929-934.
 - [10] Shevchenko, A., Jensen, O.N., Podtelejnikov, A.V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H., Mann, M. (1996) Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* **93**: 14440-14445.
 - [11] Mann, M., Hendrickson, R., Pandey, A. (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70**: 437-473.
 - [12] Lilley, K.S., Razzaq, A., Dupree, P. (2002) Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. *Curr. Opin. Chem. Biol.* **6**: 46-50.
 - [13] Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N., Willmitzer, L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**: 1157-1161.
 - [14] Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L., Fernie, A.R. (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13**: 11-29.
 - [15] Celis, J.E., Kruhoffer, M., Gromova, I., Frederiksen, C., Ostergaard, M., Thykjaer, T., Gromova, P., Yu, J., Palsdottir, H., Magnusson, N., Ornoft, T.F. (2000) Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett.* **480**: 2-16.
 - [16] Fernie, A.R. (2003) Metabolome characterization in plant system analysis. *Funct. Plant Biol.* **30**: 1-10.
 - [17] Zhao, J., Williams, C.C., Last, R.L. (1998) Induction of Arabidopsis tryptophan pathway enzymes and camalexin by amino acid starvation, oxidative stress and an abiotic elicitor. *Plant Cell* **10**: 359-370.
 - [18] Ideker, T., Galitski, T., Hood, L. (2001) A new approach to decoding life: Systems biology. *Annu. Rev. Genom. Hum. Genet.* **2**: 343-372.
 - [19] Muller, C., Scheible, W.R., Stitt, M., Krapp, A. (2001) Influence of malate and 2-oxoglutarate on the NIA transcript level and nitrate reductase activity in tobacco leaves. *Plant Cell Environ.* **24**: 191-203.
 - [20] Scheible, W.R., Krapp, A., Stitt, M. (2000) Reciprocal diurnal changes of phosphoenolpyruvate carboxylase expression and cytosolic pyruvate kinase, citrate synthase and NADP-isocitrate dehydrogenase expression regulate organic acid metabolism during nitrate assimilation in tobacco leaves. *Plant Cell Environ.* **23**: 1155-1167.
 - [21] Masclaux-Daubresse, C., Valadier, M.H., Carayol, E., Reisdorf-Cren, M., Hirel, B. (2002) Diurnal changes in the expression of glutamate dehydrogenase and nitrate reductase are involved in the C/N balance of tobacco source leaves. *Plant Cell Environ.* **25**: 1451-1462.
-

-
- [22] Suzuki, H., Achnine, L., Xu, R., Matsuda, S.P.T., Dixon, R.A. (2002) A genomics approach to the early stages of triterpene biosynthesis in *Medicago trunculata*. *Plant J.* **32**: 1033-1048.
 - [23] Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L., Fernie, A.R. (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Reports* **4**: 989-993.
 - [24] Roessner, U., Willmitzer, L., Fernie, A.R. (2001) High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.* **127**: 749-764.
 - [25] Sweetlove, L.J., Heazlewood, J.L., Herald, V., Holtzapffel, R., Day, D.A., Leaver, C.J., Millar, A.H. (2002) The impact of oxidative stress on Arabidopsis mitochondria. *Plant J.* **32**: 891-904.
 - [26] Millar, A.H., Sweetlove, L.J., Giege, P., Leaver, C.J. (2001) Analysis of the Arabidopsis mitochondrial proteome. *Plant Physiol.* **127**: 1711-1727.
 - [27] Whitelegge, J.P. (2002) Plant proteomics: BLASTing out of a MudPIT. *Proc. Natl. Acad. Sci. USA* **18**: 11564-11566.
 - [28] McDonald, W.H., Ohi, R., Miyamoto, D.T., Mitchison, T.J., Yates, J.R. (2002) Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *Intl. J. Mass Spec.* **219**: 245-251.
 - [29] Zhu, H., Klemic, J.F., Chang, S., Bertone, P., Casamayor, A., Klemic, K.G., Smith, D., Gerstein, M., Reed, M.A., Snyder, M. (2000) Analysis of yeast protein kinases using protein chips. *Nat. Genetics* **26**: 283-289.
 - [30] Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R.A., Gerstein, M., Snyder, M. (2001) Global analysis of protein activities using proteome chips. *Science* **293**: 2101-2105.
 - [31] Gerhardt, R., Stitt, M., Heldt, H.W. (1983) Subcellular metabolite determination in spinach leaves through non-aqueous fractionation. *Physiol. Chem.* **364**: 1130-1131.
 - [32] Fehr, M., Frommer, W.B., Lalonde, S. (2002) Visualisation of maltose uptake in living yeast cells by fluorescent nanosensors. *Proc. Natl. Acad. Sci. USA* **99**: 9846-9851.
 - [33] Farre, E.M., Tiessen, A., Roessner, U., Geigenberger, P., Trethewey, R.N., Willmitzer, L. (2001) Analysis of the compartmentation of glycolytic intermediates, nucleotides, sugars, organic acids, amino acids, and sugar alcohols in potato tubers using a non-aqueous fractionation method. *Plant Physiol.* **127**: 685-700.
 - [34] Szyperski, T. (1998) ¹³C-NMR, MS and metabolic flux balancing in biotechnology research. *Quarterly Rev. Biophys.* **31**: 41-106.
 - [35] Lytovchenko, A., Sweetlove, L., Pauly, M., Fernie, A.R. (2002) The influence of cytosolic phosphoglucomutase on photosynthetic carbohydrate metabolism. *Planta* **215**: 1013-1021.
-

Broad-Range Metabolite Analysis

-
- [36] Dieuaide-Noubhani, M., Raffard, G., Canioni, P., Pradet, A., Raymond, P. (1995) Quantification of compartmented metabolic fluxes in maize root tips using isotope distribution from ^{13}C - or ^{14}C - labeled glucose. *J. biol. Chem.* **270**: 13147-13159.
 - [37] Schwender, J., Ohlrogge, J.B. (2002) Probing in vivo metabolism by stable isotope labeling of storage lipids and proteins in developing *Brassica napus* embryos. *Plant Physiol.* **130**: 347-361.
 - [38] Giege, P., Heazlewood, J.L., Roessner-Tunali, U., Millar, A.H., Fernie, A.R., Leaver, C.J., Sweetlove, L.J. (2003) Enzymes of glycolysis are functionally associated with the mitochondrion in Arabidopsis cells. *Plant Cell* **15**: 2140-2151.
 - [39] Srere, P.A. (1987) Complexes of sequential metabolic enzymes. *Annu. Rev. Biochem.* **56**: 89-104.
 - [40] Taylor, S.W., Fahy, E., Zhang, B., Glenn, G.M., Warnock, D.E., Wiley, S., Murphy, A.N., Gaucher, S.P., Capaldi, R.A., Gibson, B.W., Ghosh, S.S. (2003) Characterisation of the human heart mitochondrial proteome. *Nat. Biotechnol.* **21**: 281-286.
 - [41] Panicot, M., Minguet, E.G., Ferrando, A., Alcazar, R., Blazquez, M.A., Carbonell, J., Altabella, T., Koncz, C., Tiburcio, A.F. (2002) A polyamine metabolon involving aminopropyl transferase complexes in Arabidopsis. *Plant Cell* **14**: 2539-2551.
 - [42] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**: 399-403.
 - [43] Gavin, A.C. et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**: 141-147.
 - [44] Smith, J., Khin, E.K., Zaitseva, E.M., Freebern, W., Dzekunova, I., Gardner, K. (2002) Genome wide analysis of protein/DNA interactions. *FASEB J.* **16**: A1104.
 - [45] Howell, N.K., Herman, H., Li-Chan, E.C.Y. (2001) Elucidation of protein-lipid interactions in a lysozyme-corn oil system by Fourier transform Raman spectroscopy. *J. Agric. Food Chem.* **49**: 1529-1533.
 - [46] Pan, W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18**: 546-554.
 - [47] Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D., Altman, R.B. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* **18**: 1454-1461.
 - [48] Walhout, A.J., Reboul, J., Shtanko, O., Bertin, N., Vaglio, P., Ge, H., Lee, H., Doucette-Stamm, L., Gunsalus, K.C., Schetter, A.J., Morton, D.G., Kempthues, K.J., Reinke, V., Kim, S.K., Piano, F., Vidal, M. (2002) Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.* **12**: 1952-1958.
 - [49] Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M.C., Berden, J.A., Brindle, K.M., Kell, D.B., Rowland, J.J., Westerhoff, H.V., van Dam, K., Oliver, S.G. (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* **19**: 45-50.
-

-
- [50] Brazma, A., Vilo, J. (2000) Gene expression data analysis. *FEBS Lett.* **480**: 17-24.
 - [51] Jolliffe, I.T. (1986) *Principal Components Analysis*. Springer-Verlag, New York.
 - [52] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**: 14863-14868.
 - [53] Lukashin, A.V., Fuchs, R. (2001) Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics* **17**: 405-414.
 - [54] Windig, W., Haverkamp, J., Kistemaker, P.G. (1983) Interpretation of sets of pyrolysis mass spectra by discriminant-analysis and graphical rotation. *Analyt. Chem.* **55**: 81-88.
 - [55] Pettersson, G. Ryde-Pettersson, U. (1988) A mathematical model of the Calvin photosynthesis cycle. *Eur. J. Biochem.* **175**: 661-672.
 - [56] Thomas, S., Mooney, P.J., Burrell, M.M., Fell, D.A. (1997) Metabolic Control Analysis of glycolysis in tuber tissue of potato (*Solanum tuberosum*): explanation for the low control coefficient of phosphofructokinase over respiratory flux. *Biochem. J.* **322**: 119-127.
 - [57] Wiechert, W., Mollney, M., Isermann, N., Wurzel, M., de Graaf, A.A. (1999) Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. *Biotechnol. Bioengng.* **66**: 69-85.
 - [58] Bonarius, H., Schmid, G., Tramper, J. (1997) Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends Biotechnol.* **15**: 308-314.
 - [59] Torres, N., Voit, E., Gonzalez-Alcon, C. (1996) Optimization of nonlinear biotechnological processes with linear programming: application to citric acid production by *Aspergillus niger*. *Biotechnol Bioengng.* **49**: 247-258.
 - [60] Cornish-Bowden, A., Cardenas, L. (2002) Systems biology: Metabolic balance sheets. *Nature* **420**: 129-130.
 - [61] Schuster, S., Fell, D.A., Dandekar, T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* **18**: 326-332.
 - [62] Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., Gilles, E.D. (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**: 190-193.
 - [63] de la Fuente, A., Brazhnik, P., Mendes, P. (2002) Linking the genes: inferring quantitative gene networks from microarray data. *Trends Genet.* **18**: 395-398.
-

DETERMINATION OF ENZYME ACTIVITIES BY MASS SPECTROMETRY - BENEFITS AND LIMITATIONS

HARTMUT SCHLÜTER^{1,*}, JOACHIM JANKOWSKI¹, ACHIM THIEMAN¹,
JANA RYKL¹, SANDRA KURZAWSKI¹, DIETER RUNGE²

¹Nephrology, Charité - University Medicine Berlin; Campus Benjamin Franklin,
Garystr. 5, 14195 Berlin, Germany

²Immunology, University Rostock, Faculty of Medicine, Schillingallee 70,
18055 Rostock, Germany

E-Mail: *Hartmut.Schlueter@charitie.de

Received: 3rd March 2004 / Published: 1st October 2004

ABSTRACT

Enzymatic activities in complex protein fractions are often detected with spectroscopic methods, necessitating substrates, which are modified by chromogenic or fluorogenic agents or with radioactive isotopes. However, both approaches lack the control of the identity of the reaction products risking incorrect positive results. Mass spectrometry-assisted enzyme assays allow the direct and sensitive analysis of the reaction products of the enzymatic conversion of authentic natural substrates and give confidence about the identity of the reaction products. The newly developed mass-spectrometry-assisted enzyme screening (MES) method enables the determination of enzyme activities by mass spectrometry even in raw extracts and cell lysates without sample pre-treatment prior to MS.

INTRODUCTION

In the field of enzyme kinetics spectrophotometric methods are used extensively to monitor product formation during the enzymatic reaction. This most often requires artificial substrates that undergo a change in absorbency at a given wavelength upon turnover. Although this is a simple and effective means for kinetic analysis, the scope of substrates that can be studied is highly restricted.

Furthermore, natural substrates for the enzyme generally cannot be assayed in this manner. Therefore radioactive substrates are often preferred because of their identical chemical nature to the natural substrates as well as their sensitivity of detection.

However, radiometric assays require the separation of the radioactive products by thin layer chromatography or other chromatographic methods, and subsequent liquid scintillation counting. Optical as well as radiometric methods share the problem, that there is an ambiguity about the fate of the chemical structure of the substrate after the enzymatic conversion. Therefore incorrect positive results cannot be excluded. A different type of detector, which responds universally to all substrates and reaction products, would have definite advantages over spectrophotometric systems. Because enzymatic reactions change the chemical structure of the reactants this change is generally accompanied by a change in the molecular weight. Therefore for the detection of enzymatic activities mass spectrometric techniques are rapid, sensitive and reproducible alternatives. The applicability of mass spectrometry (MS) for the determination of enzyme kinetics has been demonstrated by a number of investigators in the past eight years. The first application of MS in conjunction with liquid chromatography for the real-time analysis of enzyme kinetics was reported in 1995 [1]. Bothner et al. demonstrated that in cases where introduction of a chromophore drastically changes the fate of the reaction as a result of the structural features of the substrate, electrospray MS (ESI-MS) has been found to be especially valuable [2]. Wu et al. reported that ESI-MS could be used as a rapid, sensitive and accurate quantitative assaying tool for inhibitor libraries [3].

Matrix-assisted laser desorption/ionization (MALDI)-MS is generally more robust than ESI-MS towards buffer solutions and is more suitable for complex mixture analysis. It is therefore generally better suited for direct screening of enzyme activities requiring only minimal sample pretreatment. As a typical application MS-based enzyme activity determination was used to elucidate key fluxes in the central metabolism of lysine producing *Corynebacterium glutamicum* during batch culture [4]. Nevertheless a restriction for MALDI-MS for quantitative analysis of small enzymatic reaction products is the interference of matrix signals with analyte signals in the mass range between 22 and 500 Da. An appropriate selection of the MALDI matrix may help to solve this problem.

Our group developed a method, named "mass spectrometry assisted enzyme screening (MES)", by which enzyme activities in complex protein fractions can be measured with a mass spectrometer [5, 6]. In this study experiments are shown, which demonstrate the advantage of the MES method.

MATERIALS AND METHODS

All chemicals and enzymes were purchased from Sigma (Deisenhofen) if not stated otherwise. CNBr-activated-Sepharose 6MB beads were bought from Amersham Biotech (Freiburg). Porcine kidneys were obtained from the local slaughterhouse.

A. Covalent immobilization of proteins

Human renin, porcine renin, proteins extracted from porcine renal tissue and a protein lysate of hepatocytes were immobilized to BrCN-activated Sepharose Beads (Amersham Bioscience, Freiburg) according to the instruction manual. Briefly, for each experiment, 50 μ L from the fractions (protein concentration 10 μ g/ μ L) and 50 μ L water as a control were each mixed with 150 μ L 0.1 M NaHCO₃, pH 8.3 and 50 μ L CNBr-activated-Sepharose beads. The mixtures were incubated for 2 h at room temperature. After immobilization the beads were blocked with 20 μ L 0.2 M glycine in 0.1 M NaHCO₃, pH 8.3 for 2 h at room temperature. After blocking, the beads were washed three times with double-distilled water and stored at 4°C.

B. Incubation of proteins with the reaction specific probes

For the detection of renin activity human renin substrate or porcine renin substrate (Bachem, Weil) was dissolved as reaction specific probes in HPLC-grade water (final concentration 10⁻⁵ mol/l). ACE activity was measured with angiotensin-I (Bachem, Weil) as reaction specific probe (final concentration 10⁻⁵ mol/l). To the immobilized proteins (10 μ L beads) 20 μ L of the probe-containing aqueous liquid was added. After defined incubation times 1 μ L-aliquots (in triplicate) were removed and passed to the MALDI-MS analysis (C.).

C. MALDI-MS analysis of the reaction mixtures

For pipetting the reaction solutions to the MALDI-target a pipetting robot (Multiprobe II, Perkin Elmer, Rodgau) was used. All mass spectra were acquired on a Reflex III-MALDI mass spectrometer (Bruker-Daltronics, Bremen). The software package XMASS 5.1 and a 384-microtiter-well format MALDI target and an AnchorChipTM technology target were obtained from Bruker-Daltronics (Bremen).

One microlitre of the reaction mixture was applied on a 384-format AnchorChipTM target in triplicate. Next, 1 μ l of matrix solution (1:10 dilution of a saturated solution of α -cyano-4-hydroxy-cinnamic acid in a 1:1 mixture of acetonitrile and water containing 0.1% TFA. The mixture was dried on the target before introduction into the mass spectrometer. Positively charged ions were analysed in the reflector mode of the MALDI mass spectrometer, using delayed ion extraction. Spectra were recorded with a 2-GHz data-sampling rate. Instrument high voltages were left on between analyses to ensure stable instrument performance. Unless otherwise stated, the extraction delay time was 150 ns and deflection was used to suppress ions up to m/z 800. In this study, a nitrogen laser with an emission wavelength of 337 nm and 3 ns pulse duration was used. Typically, the laser beam was focused to 50 mm diameter at an angle of 45° to the surface of the target. Microscopic sample observation was possible. For each sample, 100 single-shot spectra were accumulated, which result from 5 different spots per sample (20 spectra per spot). The complete MALDI-MS analysis was performed automatically using the Reflex III software. All further processing was performed in batch mode using the software package XMASS 5.1. Automated peak picking was performed using the SNAP algorithm provided by XMASS 5.1. This algorithm uses the data points for all recorded monoisotopic mass signals of a peptide to assign an m/z value to the first monoisotopic peak.

D. Measurement of enzymatic activity with a fluorospectro-photometer

As a fluorescent substrate the following renin substrate was used: Arg-Glu(EDANS)-Ile-His-Pro-Phe-His-Leu-Val-Ile-His-Thr-Lys(DABCYL)-Arg (Molecular probes). The substrate was dissolved in DMSO (final concentration 500 μ M). For the determination of renin activity 4 μ l of the substrate containing DMSO was mixed with 80 μ l assay buffer (100 mM NaCl, 50 M Tris, pH 8, 1 mM EDTA) and 10 μ l of the protein fraction.

Fluorescence was measured with a spectrofluoro-photometer (Fluoroscan, Thermo, Dreieich) at 355 nm (excitation) and 460 nm (emission).

E. Preparation of the protein extract from porcine renal tissue

Porcine kidneys were placed in ice-cooled physiological saline solution immediately after excision and processed within 30 min. The tissue was cut into small pieces (about 1 cm³), frozen in liquid nitrogen, and stored at -80°C for 24 h.

The frozen tissue was lyophilized and powdered. The freeze-dried powder (1 g dry weight) was suspended for 2 min in 10 ml 20 mM potassium phosphate buffer pH 7 at 4°C and homogenized. The homogenate was centrifuged at 30.000 g for 30 min at 4°C. The pellet was discarded. One aliquot of the supernatant was used for the immobilization of the proteins.

F. Preparation and incubation of cultured human hepatocytes

Human liver tissue was obtained from partial hepatectomy due to liver metastases of a colorectal carcinoma. The isolated hepatocytes were seeded at a final density of 1.5×10^6 cells/well into multi-well plates in Williams-E medium with insulin, dexamethason and 10% FCS. Prior to cell seeding, each culture plate was coated with collagen type I (Biochrom-Seromed, Berlin). When hepatocytes had attached firmly to the collagen matrix, culture medium was removed from the culture plates and hepatocytes were overlaid and cultured with Williams-E medium with insulin, dexamethason, but not with 10% FCS. Then, this medium was exchanged against HHMM (Human Hepatocyte Maintenance Medium) with 40 ng/ml HGF, 20 ng/ml dexamethason and 2,75 µg/ml insulin. The hepatocytes were cultured in the latter medium for 10 days [7]. The medium was exchanged every 48 h. After 10 days the hepatocytes were incubated for 5, 10, 20, 30 and 60 min with 10 nM glucagon dissolved in a glucose-free DMEM medium. For the control experiment, the hepatocytes were incubated with the glucose-free DMEM medium in the absence of glucagon. The incubation was stopped by washing the hepatocytes 3 times with an 0.9 % NaCl solution, cooled with ice. After washing, the cells were immediately frozen by adding liquid nitrogen. The cells were stored at -80°C. For the MES assay the hepatocytes were lysed by thawing the cells with the coupling buffer necessary for the immobilization of the proteins.

G. Measurement of kinase activity

For the detection of the cAMP-dependent kinase activity kemptide (Bachem, Weil) and ATP were dissolved in HPLC-grade water at final concentrations of 10^{-5} mol/l, 20 μ l of the substrate mixture was added to 20 μ l of the immobilized proteins. After defined incubation times 1 μ l-aliquots (in triplicate) were removed and transferred to the MALDI-MS analysis. The signal in the resulting mass spectra indicated that kinase activity is 80 Da larger than the signal of non-phosphorylated kemptide.

RESULTS

The analytical procedure of the MES method is based on covalent immobilization of proteins to beads. By immobilizing proteins, proteolytic degradation is prevented and the removal of those molecules from the protein fraction is achieved, which otherwise would interfere with the mass spectrometric detection of the enzymatic reaction products. The enzymatic activity is determined by incubating the immobilized proteins with a reaction specific probe, followed by the analysis of the reaction mixture with the MALDI-MS after defined incubation times. Locating a signal in the mass spectrum, which fits the molecular mass of the expected reaction product, validates the type of the enzymatic reaction.

Figure 1 gives an example of the detection of enzymatic activity with the MES method. In Fig. 1 the mass spectra of the reaction mixture of porcine renin substrate (Asp-Arg-Val-Tyr-Ile-His-Pro-Phe-His-Leu-Leu-Val-Tyr-Ser) incubated for 1 h and 2 h with immobilized porcine renin and with immobilized human renin are shown. Porcine renin yielded the reaction product angiotensin-I. The intensity of the signal significantly increased with increasing incubation times. After two hours the signal intensity of porcine renin substrate had significantly decreased. In contrast human renin did not hydrolyze the porcine renin substrate. The mass spectra show no signals from angiotensin-I. The incubation of human renin substrate with immobilized human renin yielded signals from angiotensin-I in the mass spectra (data not shown). This experiment demonstrates that the high species-dependent reaction specificity of renin is maintained when using the MES method. The immobilization of the enzymes did not affect their enzymatic properties.

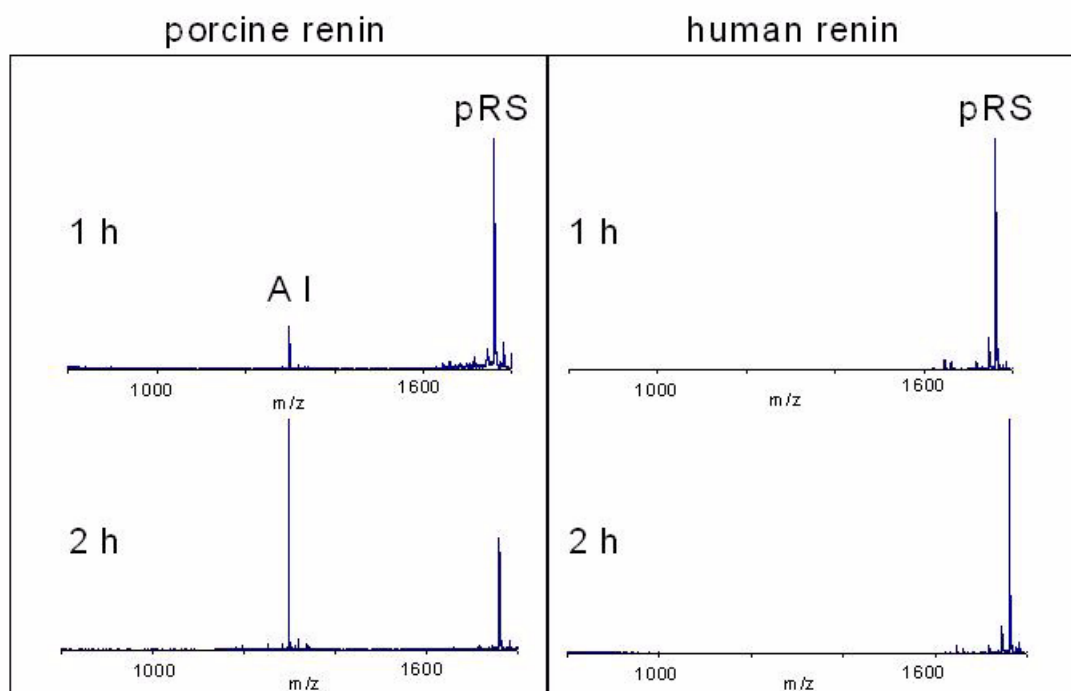


Figure 1. Detection of renin activity with the MES system. MALDI mass spectra of the reaction mixtures after 1 h and 2 h incubation of immobilized porcine renin and human renin with porcine renin substrate (pRS). AI: Angiotensin-I.

The second experiment demonstrates the risk of obtaining incorrect positive results with substrates, which are labelled with fluorescent moieties. In this case a fluorescent analogue of renin substrate was used, representing the sequence of human renin substrate (Asp-Arg-Val-Tyr-Ile-His-Pro-Phe-His-Leu-Val-Ile-His-Asn). This substrate is converted to angiotensin-I (Asp-Arg-Val-Tyr-Ile-His-Pro-Phe-His-Leu) by human renin but not by porcine renin (Table 1). However, incubating the fluorescent human renin substrate with a protein fraction from porcine renal tissue extract yielded a time dependent increase in fluorescence, representing the hydrolysis of the fluorescent human renin substrate. Therefore it must be assumed that the fluorescent substrate in the presence of the renal protein extract was hydrolyzed by another enzyme, which is not identical to renin. Because any cleavage of the peptide bonds of the fluorescent substrate yields fluorescence the resulting reaction products of the incubation of the fluorescent renin substrate with the renal protein extract must not be identical with the reaction products of human renin.

Table 1. Detection of renin activity in different protein fractions with a fluorophor-labelled renin-substrate Arg-Glu(EDANS)-Ile-His-Pro-Phe-His-Leu-Val-Ile-His-Thr-Lys(DABCYL)-Arg. Fluorescence was measured with a spectrofluoro-photometer.

Incubation of human renin substrate with	increase in fluorescence
human renin	+
porcine renin	-
porcine renal protein extract	+

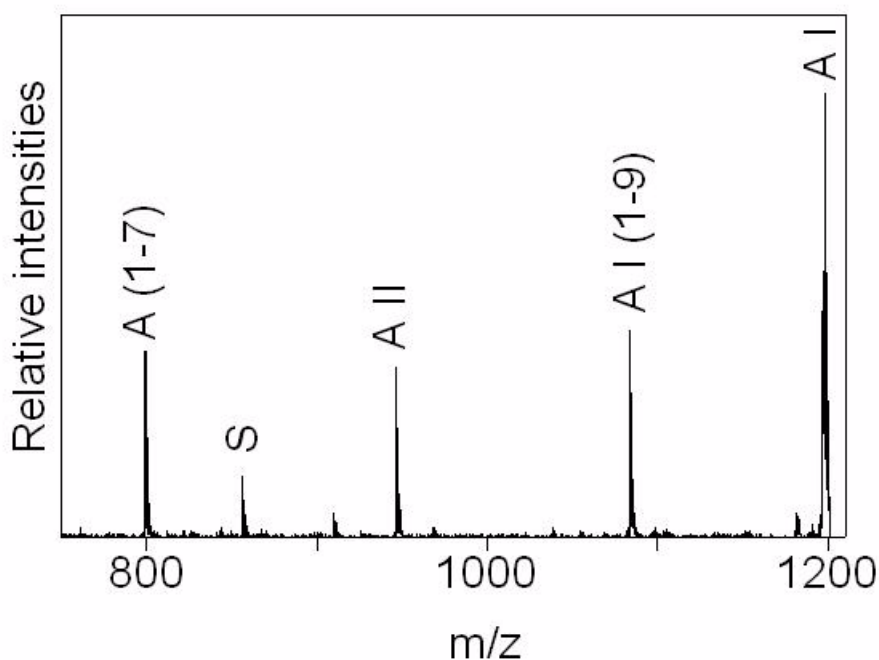


Figure 2. Detection of multiple proteolytic activities in a protein extract from porcine renal tissue. MALDI mass spectrum of the reaction mixture of the incubation of porcine renin substrate with the immobilized renal tissue proteins. A (1-7): Angiotensin (1-7); AII: Angiotensin-II; AI (1-9): Angiotensin-I (1-9); s: Internal standard.

In contrast with the MES method it is possible to detect the presence of several proteolytic enzymes in complex mixtures such as the renal tissue protein extract, as shown in Fig. 2. Here, the mass spectrum of the reaction products of the incubation of porcine renin substrate with immobilized proteins from a porcine renal tissue extract is given.

Beside angiotensin-I, the reaction product of renin, the signal of angiotensin-II represents the presence of an angiotensin-converting enzyme (ACE) activity, the signal of angiotensin-I (1-9) presumably demonstrates the presence of an ACE-2 activity and angiotensin (1-7) points to the presence of a neprilysin activity. This result clearly shows the advantage of the MES method. The mass spectrometric detection allows verification of the identity of the expected reaction products of an enzymatic conversion of the reaction specific probe. Furthermore unexpected enzyme activities can be observed via the identification of their reaction products.

A further advantage of MES is given by its high sensitivity. For the detection of ACE-activity a detection-limit for the MES system was determined, which was a factor of 1000 more sensitive than a comparable fluorescence-based ACE assay [6]. Because of its high sensitivity, the MES method was used for monitoring a cAMP-dependent protein kinase activity during the incubation of cultured human hepatocytes with glucagon (Fig. 3). For the MES kinase assay the peptide kemptide together with ATP was incubated with the immobilized proteins from the lysate of the hepatocytes. By the cAMP-dependent kinase activity kemptide becomes phosphorylated, which results in a second signal in the MALDI mass spectrum, which is 80 Da larger than the signal of kemptide.

Five minutes after starting the incubation with glucagon a significant increase in kinase activity was measured, reaching a maximum after 10 minutes. This experiment demonstrates that the MES method is suitable for the comparison of activities of low abundant enzymes, even in complex protein mixtures such as cell lysates.

Furthermore, the MES method closes the gap between transcriptome-, metabolome-, and proteome-analysis, because MES yields additional information about the modulation of activities of enzymes, which are already expressed in the cell. Therefore MES is an additional tool for systems biology.

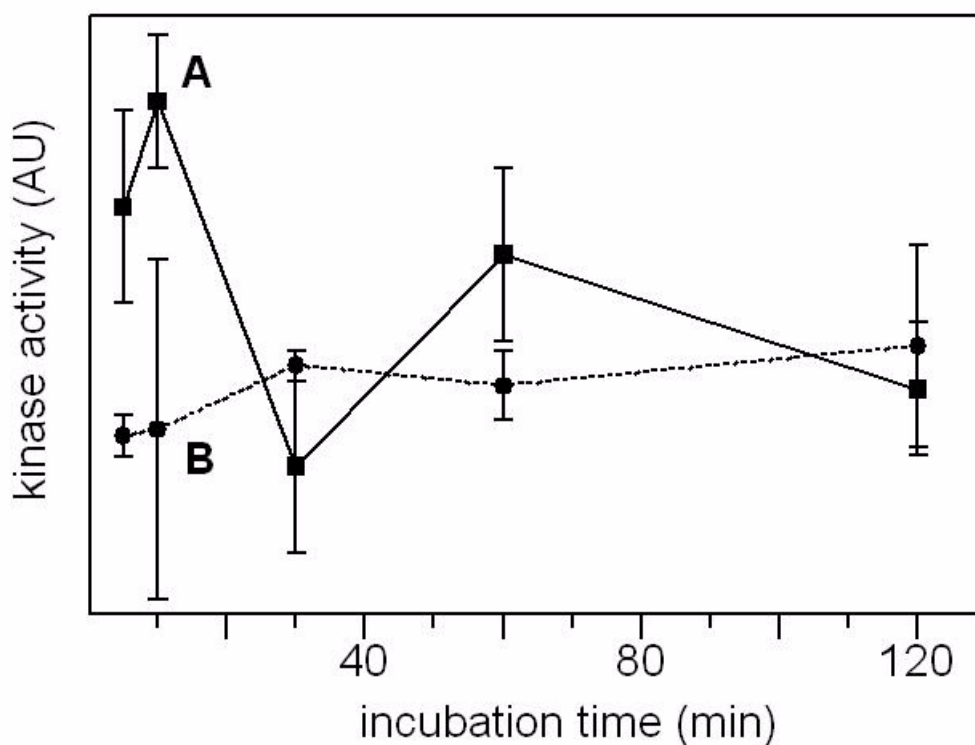


Figure 3. Time course of a cAMP-dependent kinase activity of hepatocytes stimulated with glucagon (A, solid line) and of unstimulated hepatocytes (B, dotted line). The relative enzyme activities were measured by the MES method using kemptide and ATP as substrates.

CONCLUSIONS

In conclusion the combined use of enzymology with MS provides much more detailed insight into the qualitative aspects of enzyme-catalysed reactions. The method is suitable for the analysis of the enzymatic conversion of low molecular weight substances as well as large biopolymers. Unlike the use of chromophore-labelled or radioactive substrates, there is little ambiguity as to the identity of the signal being measured. MS/MS analysis can offer additional structural information of the analyte being monitored. MS-based determination of enzymatic activities offers excellent accuracy, reproducibility and is especially well suited for assaying reactions that cannot be followed photometrically. The small sample size, minimal handling requirements, along with the potential for high-throughput, are further benefits of the combined use of enzymology with MS.

The MES approach enables the highly sensitive and reliable detection of enzymatic activities even in complex protein mixtures and therefore is a suitable tool for the determination of enzymatic activities in body fluids, cells or tissues.

REFERENCES

- [1] Hsieh, F.Y., Tong, X., Wachs, T., Ganem, B., Henion, J. (1995) Kinetic monitoring of enzymatic reactions in real time by quantitative high-performance liquid chromatography-mass spectrometry. *Analyt. Biochem.* **229**: 20-25.
 - [2] Bothner, B., Chavez, R., Wei, J., Strupp, C., Phung, Q., Schneemann, A., Siuzdak, G. Monitoring enzyme catalysis with mass spectrometry. (2000) *J. biol. Chem.* **275**: 13455-13459.
 - [3] Wu, J., Takayama, S., Wong, C.H., Siuzdak, G. (1997) Quantitative electrospray mass spectrometry for the rapid assay of enzyme inhibitors. *Chem. Biol.* **4**: 653-657.
 - [4] Wittmann, C., Heinzle, E. (2001) Application of MALDI-TOF MS to lysine-producing *Corynebacterium glutamicum*: a novel approach for metabolic flux analysis. *Eur. J. Biochem.* **268**: 2441-2455.
 - [5] Jankowski, J., Stephan, N., Knobloch, M., Fischer, S., Schmaltz, D., Zidek, W., Schlüter, H. (2001) Mass-spectrometry-linked screening of protein fractions for enzymatic activities-a tool for functional genomics. *Analyt. Biochem.* **290**: 324-329.
 - [6] Schlüter, H., Jankowski, J., Rykl, J., Thiemann, J., Artsis, 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**(7-8): 1102-1107.
 - [7] Runge, D., Köhler, C., Kostrubsky, V.E., Jäger, D., Lehmann, T., Runge, D.M., May, U., Beer-Stolz, D., Strom, S.C., Fleig, W.E., Michalopoulos, G.K. (2000) Induction of cytochrome P450 (CYP)1A1, CYP1A2, and CYP3A4 but not of CYP2C9, CYP2C19, multidrug resistance (MDR-1) and multidrug resistance associated protein (MRP-1) by prototypical inducers in human hepatocytes. *Biochem. Biophys. Res. Commun.* **273**: 333-341.
-

STUDYING ENZYME KINETICS BY MEANS OF PROGRESS-CURVE ANALYSIS

HERMANN-GEORG HOLZHÜTTER

Humboldt-University Berlin, Medical School (Charité), Institute of Biochemistry,
Monbijoustr. 2, D-10117 Berlin, Germany

E-Mail: hergo@rz.hu-berlin.de

Received: 3rd March 2004 / Published: 1st October 2004

ABSTRACT

Almost all chemical reactions and transport processes in a cell are catalysed by specific enzymes and transport proteins, respectively. Kinetic characterization of these auxiliary proteins is a necessary prerequisite for understanding the dynamics and regulation of cellular reactions networks. Progress-curve analysis, i.e. estimation of kinetic parameters by fitting of integrated rate laws to the time-course of a biochemical reaction, allows an efficient kinetic characterization of enzymes. This article outlines the mathematical fundamentals of progress-curve analysis and provides examples for the application of this method in enzyme kinetics and system biology.

INTRODUCTION

The reaction network of a living cell comprises several thousands of chemical reactions and transport processes, most of them catalysed or facilitated by specific enzymes and transport proteins. On one hand, these auxiliary proteins act as catalysts accelerating the rate of the underlying process. On the other hand, and most importantly, they act as regulators in that their activity can be tuned to allow for an optimal functioning of the whole cellular network. Temporal gene expression, binding of allosteric effectors and reversible enzyme phosphorylation and dephosphorylation are prominent types of enzyme regulation.

Computational systems biology is a rapidly growing field of theoretical research aimed at the establishment of computer models that can be used to simulate the dynamics of the complete cellular reaction network.

In my opinion, this ambitious goal can only be reached if the kinetic properties of the participating enzymes - at least of those enzymes that are mainly involved in the regulation of the network - are known. Thus, there is an urgent need for the systematic kinetic characterization of enzymes belonging to a branch of the cellular network subjected to mathematical modelling.

The conventional method used to set up enzyme-kinetic models is to isolate the respective enzyme from the cell and to determine its kinetic properties monitoring the reaction kinetics under well-controlled *in vitro* conditions. This procedure usually ends up in the formulation of a rate law representing a mathematical expression relating the reaction velocity to the concentration of reactants and other metabolites that may function as effectors (activators or inhibitors) of the enzyme. The so called initial-rate method is usually applied to guess the mathematical structure of the rate law and to estimate numerical values of the kinetic parameters (affinity constants for the various ligands, cooperativity indices, maximal velocities etc.) entering it. The initial-rate method consists in measuring the initial rate of the reaction at various concentrations of all relevant ligands [1]. This method has two basic drawbacks. First, it presumes a steady-state regime of the reaction, i.e. the enzyme activity (maximal rate) is constant over the time period of the measurements. For enzymes exhibiting self-activation or inactivation this condition is obviously not met. Second, the initial-rate method is laborious as it requires a series of parallel assays differing in the initial concentrations of the reactants and effectors. A more elegant method is the so-called progress-curve analysis [2-5]. This method extracts information about the kinetic properties of an enzyme from the full time-course of the reaction - and not only its initial part. From the mathematical viewpoint, progress-curve analysis requires the calculation of the time-dependent solution of the kinetic equations governing any enzyme-catalysed reaction and to fit this solution to observed time-courses by an appropriate choice of the model parameters. Thus, progress-curve analysis combines methods for the solution of differential equation systems with non-linear regression methods. With the exception of the Michaelis-Menten equation, for which an explicit integrated form can be derived (expressing the reaction time through the concentration of the substrate) solution of the kinetic equations for more enzymes with complicated rate laws is only possible by using numerical integration methods. This relatively high mathematical effort may explain why progress-curve analysis so far does not enjoy widespread applications.

In the following I will briefly outline the mathematical fundamentals of progress-curve analysis and provide some examples from our recent research work demonstrating the usefulness of this method in reducing the experimental effort and in coping with situations where the initial-rate method cannot be applied.

PROGRESS-CURVE ANALYSIS: MATHEMATICAL BACKGROUND

To illustrate the essence of the progress-curve method let us consider the time-courses of substrate consumption and product formation of a monomolecular reaction $S \rightarrow P$ (see Fig. 1). The initial velocity $v_0 = v(S_0; P_0)$ of the reaction is defined by the slope of the quasi-linear phase of the time-course, expressed as amount of product formed or substrate consumed per time. Determining initial rates by varying the initial concentrations S of the substrate at fixed concentration of the product (or of all other ligands in more complex reactions) one arrives at the so-called rate diagram (or v - S -characteristics) showing the relationship between reaction rate and substrate concentration. Similar rate diagrams can be constructed for the other ligands. These rate diagrams are very helpful for guessing the enzyme mechanism and proposing an appropriate mathematical expression for the underlying rate law.

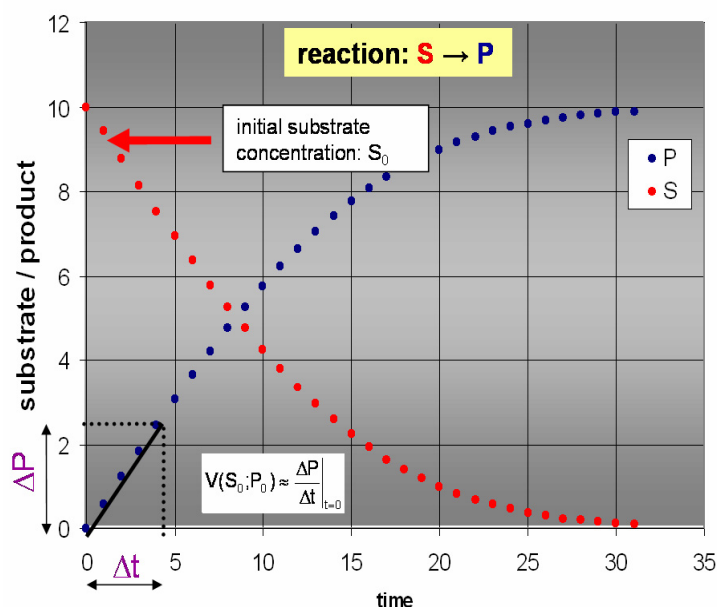


Figure 1. Determination of initial reaction rates. The initial rate is calculated by measuring over a short time period the consumption of substrate (or formation of product) and relating the change of concentration to the time elapsed. Mathematically speaking, this method consists in approximating the slope (= first derivative) of the progress curve at time $t = 0$ by the average slope of the quasi-linear initial part. The initial rate is assigned to the initial concentrations of the ligands.

The construction of rate diagrams does not necessarily require initial-rate measurements. They can be obtained more efficiently by determining reaction velocities from the first derivative of the progress curve. At a given time point of the reaction, the concentrations of the reactants can be assessed from their initial concentrations and the known stoichiometry of the reaction (see Fig. 2).

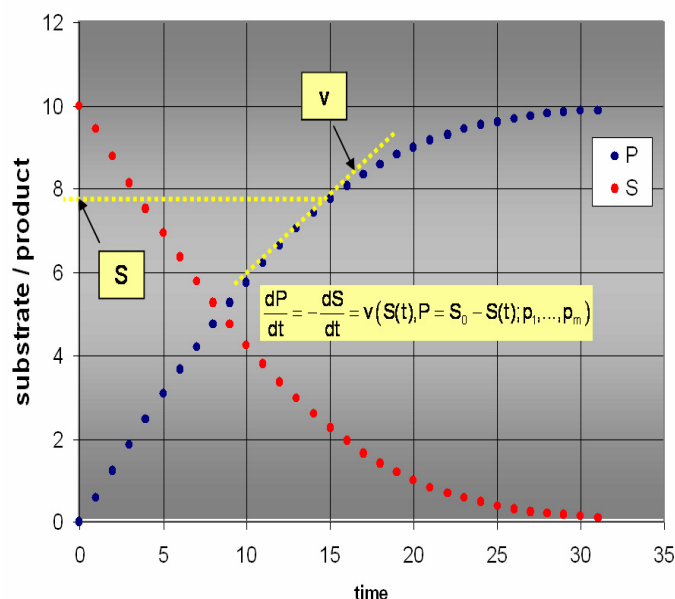


Figure 2. Determination of reaction rates by taking the slope of the progress curve. Reaction rates can be calculated as slopes of the progress curve at any time point of the reaction. The corresponding concentrations of the reactants can be assessed by their initial values and the known stoichiometry of the reaction.

The issue with determining slopes of the progress curve is that the data points typically display some experimental noise which even amplifies in the higher derivatives of the curve (see insert in Fig. 3A). Thus, some smoothing of the progress-curve data is required before taking the first derivative. A multitude of statistical curve-smoothing techniques have been developed [6] all of which are based on a local polynomial approximation of the curve. For the smoothing of polarographically or photometrically monitored progress curves comprising 500 - 5000 data points we have made good experience with the Golay-Savitzky polynomials.

Studying Enzyme Kinetics by Means of Progress-Curve Analysis

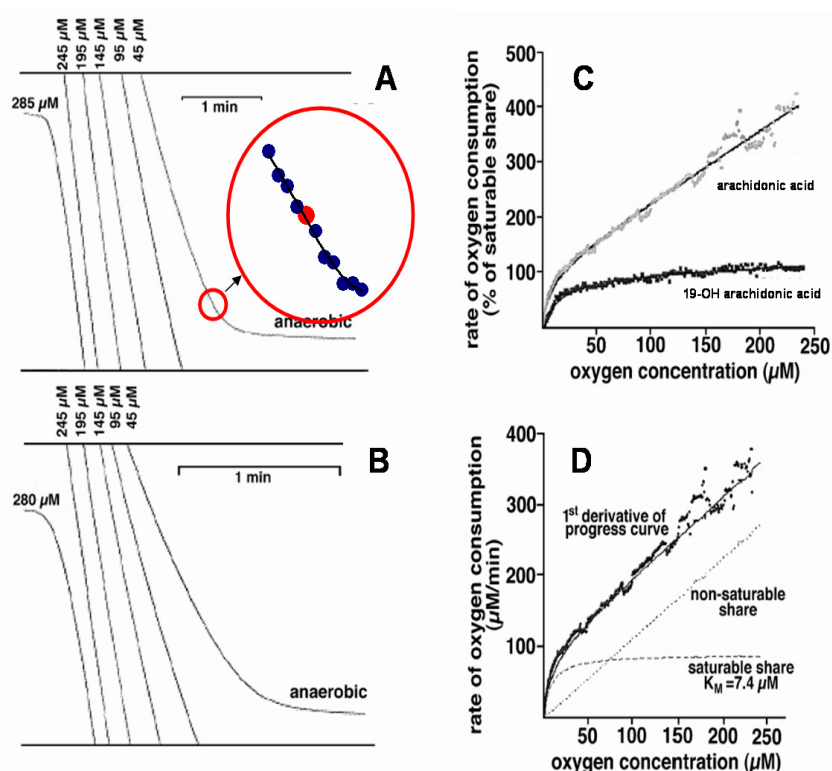


Figure 3. Polarographic progress curves of O_2 -consumption during lipoyxygenase reaction. The oxygen concentration during the lipoyxygenase-catalyzed dioxygenation of two different fatty acids was monitored by means of a Clark electrode. **Panel A:** dioxygenation of arachidonic acid, **panel B:** dioxygenation of 19-OH arachidonic acid. **Panel C** shows the reaction rates derived from the two progress curves by using a 5th order Golay-Savitzki filtering for the calculation of the first derivatives. **Panel D** illustrates that the rate diagrams in panel C can be well accounted for by splitting the reaction rate into a saturable share (Michaelis-Menten equation) and a non-saturable linear share.

Figures 3A-B depict polarographic progress curves of molecular oxygen monitored in a lipoyxygenase assay with two different fatty acid substrates. The red circle enlarges a small part of the progress curve. The solid line represents the slope of the curve calculated by means of a 5th order Golay-Savitzky filter (including 5 data points on both sides of the data point) marked in red. Panels in Fig. 3C, D show the rate diagrams constructed by plotting the slopes of the progress curves versus oxygen concentrations. For both fatty acid substrates the rate exhibits a non-saturable behaviour. As shown in Fig. 3D, the observed rate-versus-oxygen relationships can be well described by a phenomenological rate law consisting of a hyperbolic Michaelis-Menten function and a linear function. One way of determining the unknown kinetic parameters is to fit the empirical rate equation to the velocity data shown in Fig. 3C. However, these numerical estimates may be influenced by the bias of the velocity data that is inevitably associated with the local smoothing of the original progress-curve data.

Hence, a more feasible way to get numerical estimates of the model parameters is to directly fit the solutions of the rate equation to the progress-curve data.

This procedure requires the combination of a numerical integration routine for the solution of differential equations with non-linear regression methods. Figure 4 delineates the typical protocol for progress-curve analysis.

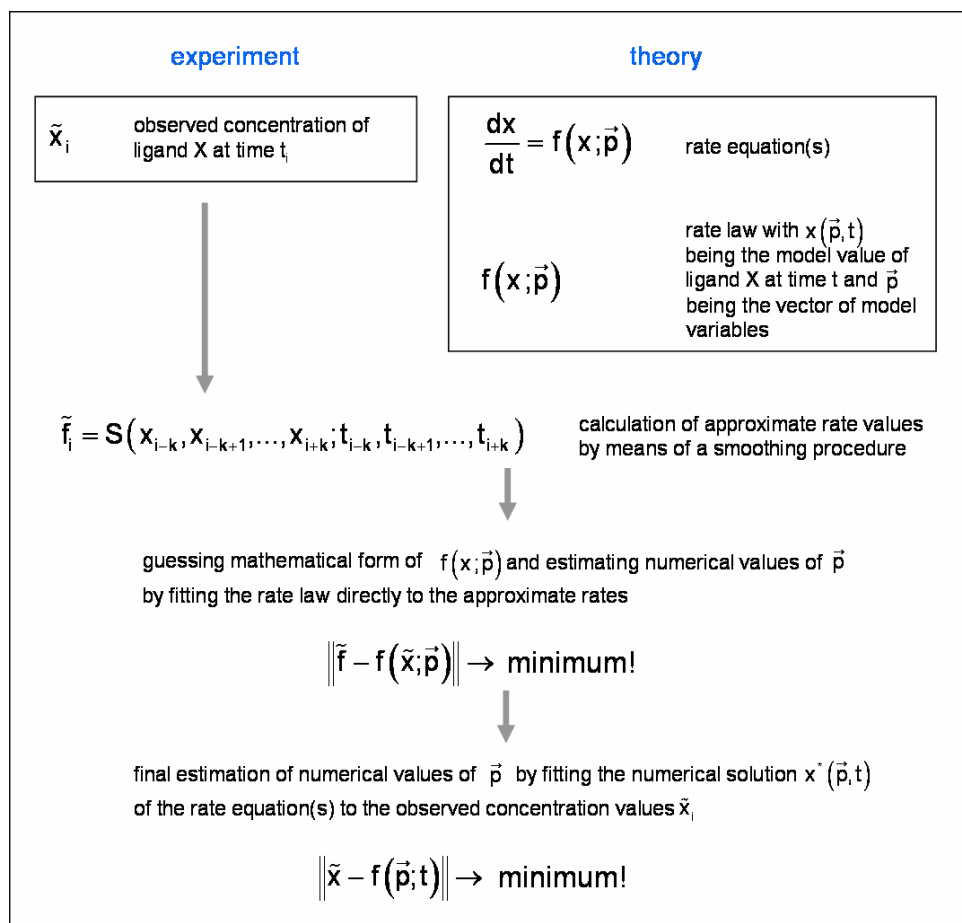


Figure 4. Main steps of progress curve analysis.

PROGRESS-CURVE ANALYSIS OF NON-STATIONARY ENZYME REACTIONS

A striking advantage of progress-curve analysis over the initial-rate method is that it allows the study of enzyme reactions under non-stationary conditions. As an example for such an application, I show here some recent results of kinetic studies with the 15-lipoxygenase (15LOX) from reticulocytes. This enzyme plays an important role in the formation of leucotriens [7].

The enzyme-radical intermediate may either react with molecular oxygen under formation of the hydroperoxy product (P) or, alternatively, may leave the oxygenation cycle (marked in yellow in Fig. 5) by decaying into the fatty acid radical and the free $\text{Fe}^{(2+)}$ -enzyme which may re-enter the dioxygenation cycle by transferring one electron to the hydroperoxy product.

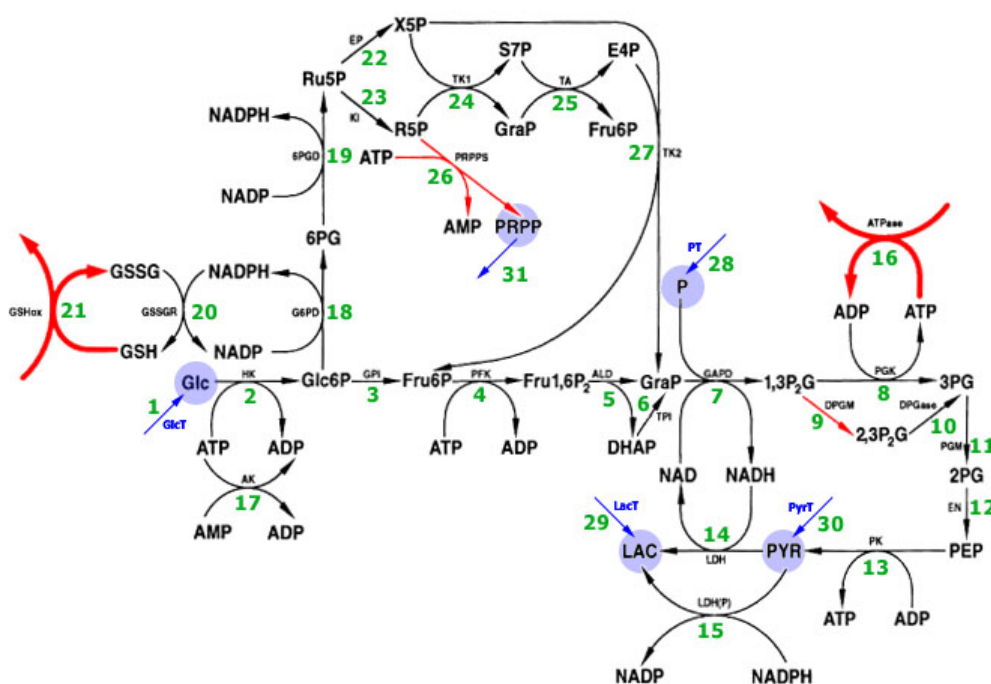


Figure 5. Metabolic reactions in red cells taken into account by the kinetic model [9] used for the analysis of systemic progress curves. *Enzymes:* HK - hexokinase [2.7.1.1]; PGI - phosphohexose isomerase [5.3.1.9]; PFK - phosphofructokinase [2.7.1.11]; ALD - aldolase [4.1.2.13]; TIM - triosephosphate isomerase [5.3.1.1]; GAPD - glyceraldehyde-3-phosphatedehydrogenase [1.2.1.12]; PGK - phosphoglycerate kinase [2.7.2.3]; DPGM - bisphosphoglycerate mutase [5.4.2.4]; DPGase - bisphosphoglycerate phosphatase [3.1.3.13]; PGM - phosphoglycerate mutase [5.4.2.1]; EN - enolase [4.2.1.11]; PK - pyruvate kinase [2.7.1.40]; LDH -lactate dehydrogenase [1.1.1.28]; AK -adenylate kinase [2.7.4.3]; G6PD - glucose 6-phosphate dehydrogenase [1.1.1.49]; 6PGD - phosphogluconate dehydrogenase [1.1.1.44]; GSSGR - glutathione reductase [1.8.1.7]; EP - phosphoribulose epimerase [5.1.3.1]; KI - ribose phosphate isomerase [5.3.1.6]; TK - transketolase [2.2.1.1]; TA - transaldolase [2.2.1.2]; PRPPS - phosphoribosylpyrophosphate synthetase [2.7.6.1].

Recently, we observed an unusual oxygen dependence of the LOX reaction when we used a hydroxylated fatty acid (19-OH arachidonic acid, 19-OH-AA) as substrate (see Fig. 6A). Up to oxygen concentrations of about 550 μM we could not see a saturation level of the initial rates. Moreover, the progress curves clearly showed biphasic behaviour. Our interpretation of the latter feature was that starting the reaction with all enzyme resident in the active $\text{Fe}^{(2+)}$ -form, a time-dependent transition into the inactive $\text{Fe}^{(3+)}$ -form occurs when an equilibrium has been established between inactivation (rate constant k_4) and re-activation (rate constant k_1).

To test this hypothesis and to determine the kinetic parameters of the LOX for this special substrate we applied the kinetic model [8] governing the reaction scheme shown in Fig. 5. Treating the reversible binding of the fatty acid substrate (S) and the hydroperoxy fatty acid product (P) as fast equilibrium reactions one may introduce the equilibrium pools

$$(X_1) = (E) + (ES) + (EP)$$

$$(X_2) = (E^*) + (E^*S) + (E^*P)$$

$$(X_3) = (ES\bullet)$$

which add up to the total enzyme concentration $(X) = (X_1) + (X_2) + (X_3)$.

The kinetic equations read

$$\frac{d(S)}{dt} = \angle f_2(X_2)$$

$$\frac{d(O_2)}{dt} = \angle f_3(X_3)$$

$$\frac{d(P)}{dt} = f_3(X_3) \angle f_1(X_1)$$

$$\frac{d(X_1)}{dt} = f_4(X_3) \angle f_1(X_1)$$

$$\frac{d(X_2)}{dt} = f_1(X_1) + f_3(X_3) \angle f_2(X_2)$$

$$\frac{d(X_3)}{dt} = f_2(X_2) \angle f_3(X_3) \angle f_4(X_3)$$

where the functions f_1 , f_1^* , f_2 , f_3 and f_4 have the following meaning:

$$f_1 = \frac{k_{1P}(P)}{K_{mP}(1 + (S)/K_{iS}) + (P)} \quad f_2 = \frac{k_2(S)}{K_{mS}(1 + (P)/K_{iP}) + (S)}$$

$$f_3 = k_3(O_2)$$

$$f_4 = k_4$$

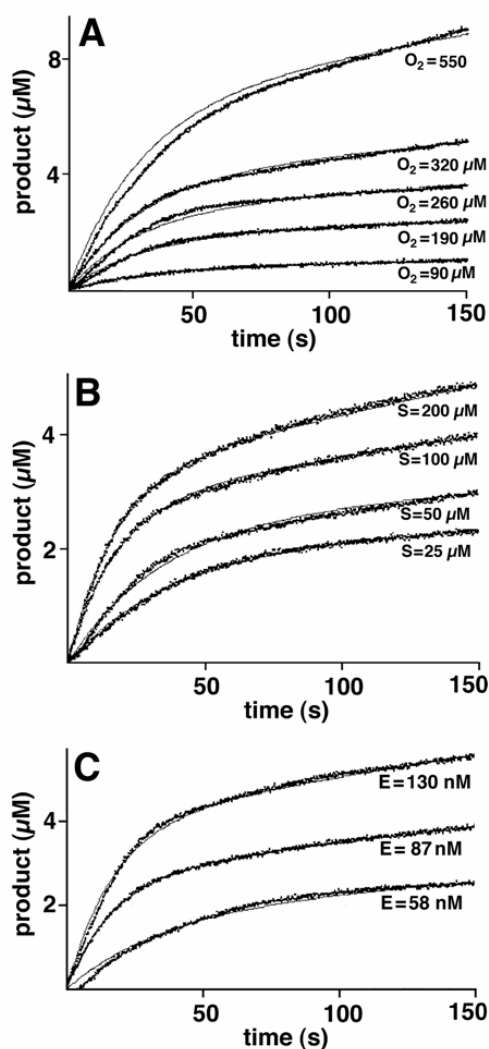


Figure 6. Photometric progress curves of the 15-lipoxygenase reaction using 19-OH-arachidonic acid as substrate. Bold (discontinuous) curves = experimental data. Thin (continuous) curves = model simulations.

The numerical values of the kinetic parameters entering the kinetic equations are shown in Table 1. They were obtained by progress-curve analysis, i.e. fitting the numerical solutions of the above equation system to the time-courses of product formation monitored at varying concentrations of oxygen, substrate and enzyme shown in Fig. 6.

Table 1. Numerical values of the model parameters.

Parameter	Meaning	Value	Unit
k_1	activation (product)	1.86E-02	s^{-1}
k_2	H-abstraction	1.45E+01	s^{-1}
k_3	O ₂ -insertion	9.58E-03	$s^{-1}\mu M^{-1}$
k_4	inactivation	5.79E-02	s^{-1}
K_{mp}	P-binding to inactive enzyme	37	μM
K_{ms}	S-binding to active enzyme	87	μM
K_{is}	S-binding to inactive enzyme	101	μM
K_{ip}	P-binding to active enzyme	n.d.	μM

The excellent correspondence between experimental and simulated progress curves demonstrates that the kinetic model is able to capture all those unusual features observed with 19-OH AA. In particular, the very low oxygen affinity can be accounted for by the high value of the inactivation rate k_4 causing the enzyme to permanently drop off from the catalytic cycle to the inactive Fe²⁺-form. With increasing concentration molecular oxygen is capable of out competing this inactivation step at the level of the enzyme-radical complex ES[•].

ANALYSIS OF SYSTEMIC PROGRESS-CURVES: ESTIMATION OF ENZYME PARAMETERS *IN VIVO*

In the previous example progress-curve analysis was applied to a single enzyme comprising a complex kinetics. Here I present an even more advanced application of this technique aimed at determining the temperature coefficients of enzymes under *in vivo* conditions. This study was performed with human erythrocytes. The objective was to determine the (maximal) activities of a larger group of erythrocyte enzymes at low temperature (4°C) in order to better understand the metabolic control of the red cells under blood preservation conditions. To this end, we have monitored the time-dependent metabolic changes in human erythrocytes induced by a drop of the incubation temperature from 27°C to 4°C.

Studying Enzyme Kinetics by Means of Progress-Curve Analysis

To prevent acidification of the medium by accumulating lactate, the pH value was buffered at constant value (7.0). Figure 7 shows some typical time-courses of red cell metabolite

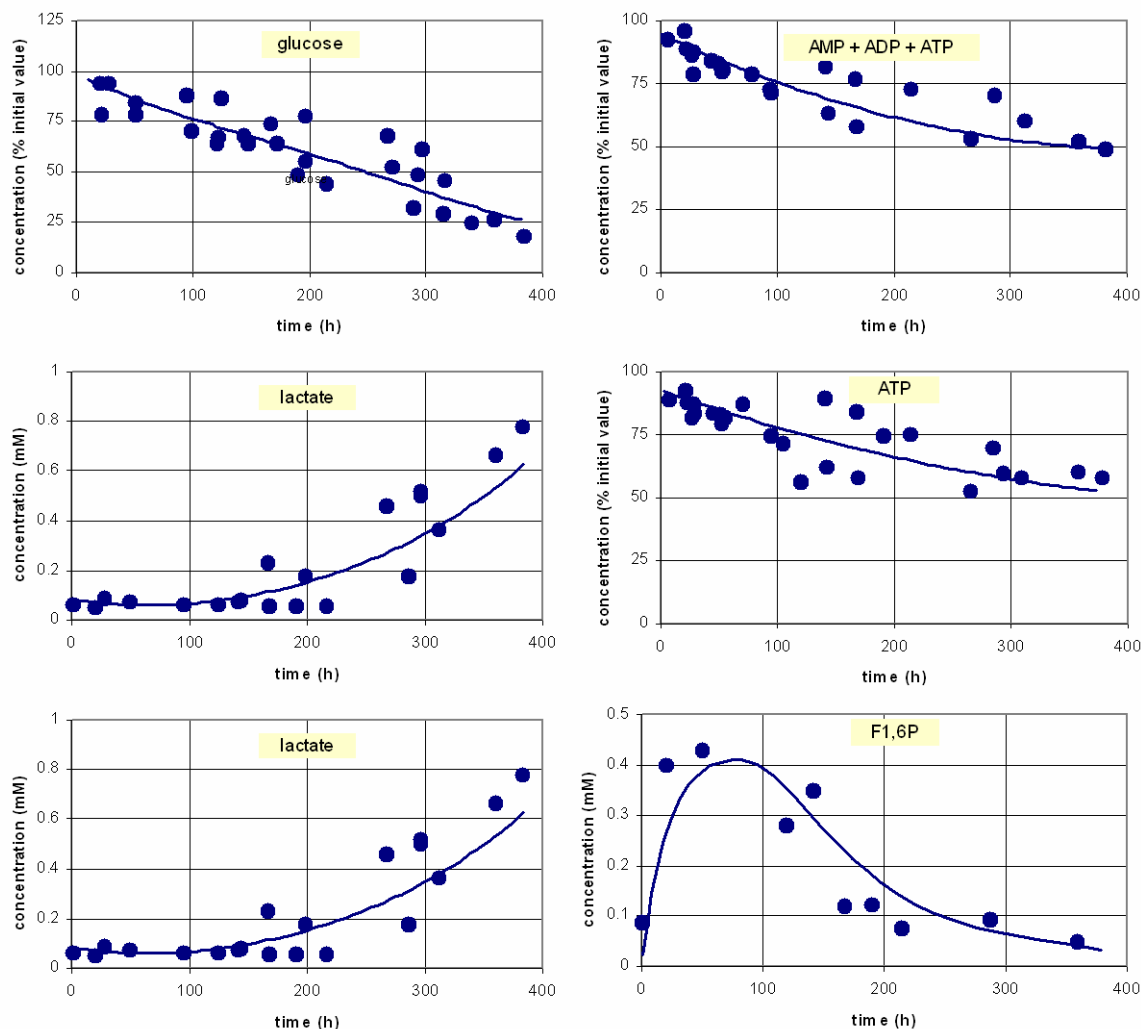


Figure 7. Time courses of selected red cell metabolites observed during a 15 day incubation at $T = 4^{\circ}\text{C}$ at $\text{pH} = 7.0$

From this information we tried to estimate the activity of red cell enzymes at 4°C ($v_{max}^{4^{\circ}}$) by simulating the observed time-courses by a comprehensive kinetic model of the red cell metabolism [9].

The reaction scheme of this model is shown in Fig. 8. It comprises two major pathways of the cell: (i) glycolysis, degrading glucose to lactate and pyruvate and representing the only source of ATP and (ii) the hexose monophosphate shunt, responsible for the inter-conversion of hexoses and riboses and the production of NADPH₂ needed for various reductive processes of the cell. The system of algebro-differential equations describing the time-evolution of this system is given in [9].

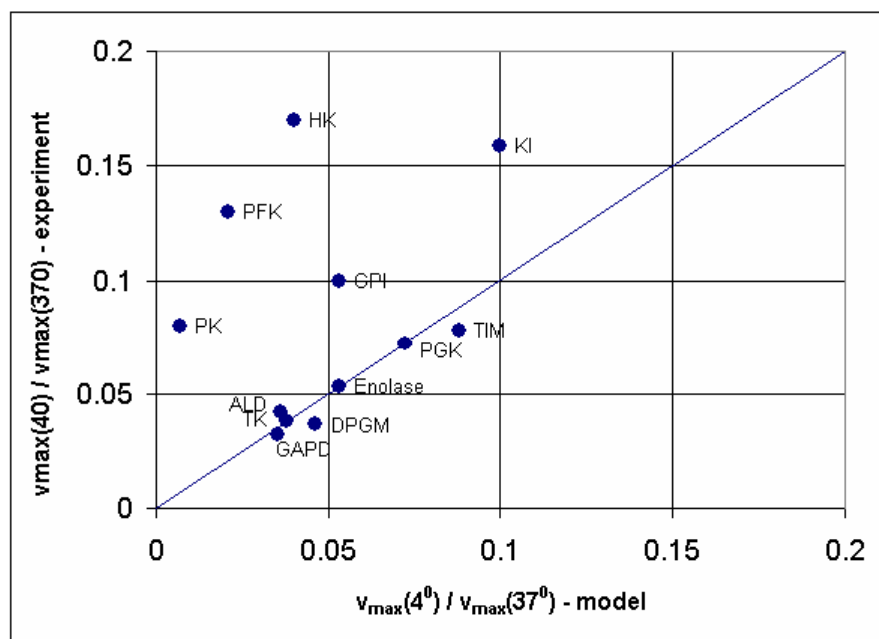


Figure 8. Metabolic reactions in red cells taken into account by the kinetic model [9] used for the analysis of systemic progress curves. Enzymes: HK - hexokinase [2.7.1.1]; PGI - phosphohexose isomerase [5.3.1.9]; PFK - phosphofructokinase [2.7.1.11]; ALD - aldolase [4.1.2.13]; TIM - triosephosphate isomerase [5.3.1.1]; GAPD - glyceraldehyde-3-phosphatedehydrogenase [1.2.1.12]; PGK - phosphoglycerate kinase [2.7.2.3]; DPGM - bisphosphoglycerate mutase [5.4.2.4]; DPGase - bisphosphoglycerate phosphatase [3.1.3.13]; PGM - phosphoglycerate mutase [5.4.2.1]; EN - enolase [4.2.1.11]; PK - pyruvate kinase [2.7.1.40]; LDH - lactate dehydrogenase [1.1.1.28]; AK - adenylate kinase [2.7.4.3]; G6PD - glucose 6-phosphate dehydrogenase [1.1.1.49]; 6PGD - phosphogluconate dehydrogenase [1.1.1.44]; GSSGR - glutathione reductase [1.8.1.7]; EP - phosphoribulose epimerase [5.1.3.1]; KI - ribose phosphate isomerase [5.3.1.6]; TK - transketolase [2.2.1.1]; TA - transaldolase [2.2.1.2]; PRPPS - phosphoribosylpyrophosphate synthetase [2.7.6.1].

Progress-curve analysis was performed by fitting the time-dependent solution of the kinetic model to the observed time-courses of red cell metabolites by treating the maximal activities of the enzymes as the only temperature-sensitive parameters of the model, i.e., neglecting temperature effects on other kinetic parameters as, for example, affinity constants.

Studying Enzyme Kinetics by Means of Progress-Curve Analysis

This simplification seems to be justified in the light of numerous enzymatic studies demonstrating that the catalytic rate constant of an enzymatic reaction exhibits a much more pronounced temperature dependence than other kinetic parameters. Figure 9 compares the relative temperature coefficients $v_{max}^{4^\circ} / v_{max}^{37^\circ}$ calculated by means of systemic progress-curve analysis with experimental values taken from the literature. Larger discrepancies result for the glycolytic enzymes hexokinase (HK), phosphofructokinase (PFK) and pyruvate kinase (PK), catalysing irreversible reaction steps and exerting the largest control over red cell glycolysis.

The temperature sensitivity of these enzymes as measured with the isolated enzyme under *in vitro* conditions is significantly lower than that predicted by progress-curve analysis. The most likely explanation for the overestimation of the temperature dependence predicted by systemic progress-curve analysis is due to the fact that important effectors down-regulating the activity of these enzymes at lower temperatures *in vivo* are lacking in the kinetic model. Hence, the drop in the activity of these enzymes needed to account for the observed time courses of these regulatory enzymes is attributed to the time-dependence of their v_{max} values.

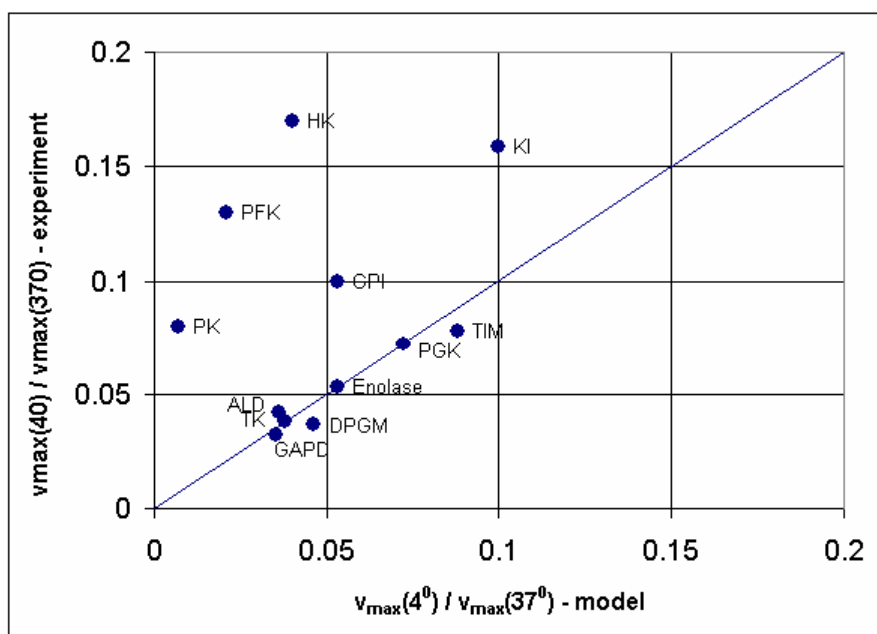


Figure 9. Relative temperature coefficients $v_{max}(4^\circ) / v_{max}(37^\circ)$ of selected red cell enzymes. x-Coordinate: values obtained by fitting a comprehensive kinetic model to time-courses of metabolite concentrations (systemic progress-curve analysis). y-Coordinate: *in vitro* experimental values measured with the isolated enzyme (data taken from references [10-22]).

However, it cannot be excluded that the *in vitro* data for these enzymes are wrong because some physiological effectors, making them more susceptible for changes in temperature, have not been considered in the *in vitro* assay. This points to a general problem arising when building enzyme-kinetic models on *in vitro* data and using them in the context of systems biology. The final decision to rely on mathematical models of complex cellular networks based on kinetic parameters obtained either from *in vitro* measurements or from systemic progress-curve analysis will ultimately depend on the goodness of testable predictions made with either types of models.

ACKNOWLEDGMENTS

The author would like to thank Dr. Hartmut Kühn, Dr. Igor Ivanov, Dr. Iris Rapoport and Dr. Ronny Schuster who have provided the experimental data used in this work and who have contributed to the establishment of the mathematical models.

REFERENCES

- [1] Segel, H.I. (1993) *Enzyme Kinetics : Behaviour and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems*, Wiley-Interscience
 - [2] London, J.W., Shaw, L.M., Garfinkel, D. (1977) Progress curve algorithm for calculating enzyme activities from kinetic assay spectrophotometric measurements. *Analyt. Chem.* **49**: 1716-1719.
 - [3] Duggleby, R.G. (1986) Progress-curve analysis in enzyme kinetics. Numerical solution of integrated rate equations. *Biochem. J.* **235**:613-615.
 - [4] Duggleby, R.G., Nash, J.C. (1989) A single-parameter family of adjustments for fitting enzyme kinetic models to progress-curve data. *Biochem. J.* **257**:57-64.
 - [5] Holzhütter, H.G., Henke, W. (1984) A new method of parameter estimation from progress curves. *Biomed. Biochim. Acta* **43**: 813-820
 - [6] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. (2002) *Numerical Recipes in C-The Art of Scientific Computing*. Cambridge University Press, Cambridge.
 - [7] De Caterina, R., Zampolli, A. (2004). From asthma to atherosclerosis: 5-lipoxygenase, leukotrienes, and inflammation. *New Engl. J. Med.* **350**:4-7.
 - [8] Ludwig, P., Holzhütter, H.G., Colosimo, A., Silvestrini, M.C., Schewe, T., Rapoport, S.M. (1987). A kinetic model for lipoxygenases based on experimental data with the lipoxygenase of reticulocytes. *Eur. J. Biochem.* **168**:325-337.
 - [9] 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.
-

Studying Enzyme Kinetics by Means of Progress-Curve Analysis

- [10] Noat, G., Ricard, J.(1968) Kinetic study of yeast hexokinase. 2. Analysis of the progress curve of the reaction. *Eur. J. Biochem.* **5**:71-72.
 - [11] Rutter, W.J., Richards, O.C., Woodfin, B.M. (1961) Comparative studies of liver and muscle aldolase. I. Effect of carboxypeptidase on catalytic activity. *J. biol. Chem.* **236**:3193-3197.
 - [12] Daddona, P.E., Kelley, W.N. (1977) Human adenosine deaminase. Purification and subunit structure. *J. biol. Chem.* **252**:110-115.
 - [13] Calvin, M.C., Blouquit, Y., Garel, M.C., Prehu, M.O., Cohen-Solal, M., Rosa, J., Rosa, R. (1990) Human bisphosphoglycerate mutase expressed in *E coli*: purification, characterization and structure studies. *Biochimie* **72**:337-343.
 - [14] Malmstrom, B.G. (1961) The inhibition of Mg (II)-enolase by other activating metal ions. *Biochim. Biophys. Acta* **51**:374-376.
 - [15] Kahana, S.E., Lowry, O.H., Schulz, D.W., Passonneau, J.V., Crawford, E.J. (1960) The kinetics of phosphoglucoisomerase. *J. biol. Chem.* **235**:2178-2184.
 - [16] <http://www.empproject.com>
 - [17] Gerber, G., Preissler, H., Heinrich, R., Rapoport, S.M. (1974) Hexokinase of human erythrocytes. Purification, kinetic model and its application to the conditions in the cell. *Eur. J. Biochem.* **45**:39-52.
 - [18] Krietsch, W.K., Bucher, T. (1970) 3-phosphoglycerate kinase from rabbit skeletal muscle and yeast. *Eur. J. Biochem.* **17**:568-580.
 - [19] Hove-Jensen, B., Harlow, K.W., King, C.J., Switzer, R.L. (1986) Phosphoribosylpyrophosphate synthetase of *Escherichia coli*. Properties of the purified enzyme and primary structure of the prs gene. *J. biol. Chem.* **261**:6765-6771.
 - [20] Miwa, S. (1987) Pyruvate kinase deficiency. *Nippon Ketsueki Gakkai Zasshi.* **50**:1445-1452.
 - [21] Eber, S.W., Krietsch, W.K. (1980) The isolation and characterization of the multiple forms of human skeletal muscle triosephosphate isomerase. *Biochim. Biophys. Acta* **614**:173-184.
 - [22] Mocali, A., Paoletti, F. (1989) Transketolase from human leukocytes. Isolation, properties and induction of polyclonal antibodies. *Eur. J. Biochem.* **180**:213-219.
-

MULTIFUNCTIONAL ENZYMES AND PATHWAY MODELLING

STEFAN SCHUSTER^{1,*} AND IONELA ZEVEDEI-OANCEA²

¹Friedrich Schiller University Jena, Faculty of Biology and Pharmaceutics, Section of Bioinformatics, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

²Humboldt University Berlin, Section of Theoretical Biophysics, Invalidenstr. 42, D-10115 Berlin, Germany

E-Mail: *schuster@minet.uni-jena.de

Received: 10th March 2004 / Published: 1st October 2004

ABSTRACT

The analysis of network properties of metabolic systems has recently attracted increasing interest. While enzymes are usually considered to be specific catalysts, many enzymes in living cells are characterized by broad substrate specificity. Here we discuss some aspects of the treatment of such multifunctional enzymes in metabolic pathway analysis, for example, their suitable representation. The fact that the choice of independent functions of multifunctional enzymes is non-unique is explained. We comment on the annotation of such enzymes in metabolic databases and give some suggestions to improve this. We then explain the proper definition of metabolic pathways (elementary flux modes) for systems involving multifunctional enzymes and discuss some ontological problems.

INTRODUCTION

Metabolic pathway analysis has become a widely used tool in biochemical modelling [1-5]. It is instrumental in metabolic engineering [6,7] and functional genomics [8,9]. The analysis is based on the decomposition of metabolic networks into their smallest functional entities - the metabolic pathways. One of its major advantages is that it does not require any knowledge of kinetic parameters. It only uses the stoichiometric coefficients and information about the directionality of enzymatic reactions.

A central concept in metabolic pathway analysis is that of elementary flux modes [10,11]. An elementary mode is a minimal set of enzymes that can operate at steady state, with all irreversible reactions used in the appropriate direction and the enzymes weighted by the relative flux they carry. Any flux distribution in the living cell is a superposition of elementary modes.

While enzymes are usually considered as specific biocatalysts, it should be acknowledged that many biochemical reactions in living cells are catalysed by enzymes with broad substrate specificity. That is, the same enzyme can alternatively convert various substrates. For example, 5'-nucleotidase (EC 3.1.3.5) can dephosphorylate AMP, IMP and other ribonucleotide monophosphates. Further examples are provided by transketolase (EC 2.2.1.1), hexokinase (EC 2.7.1.1) and branched-chain amino acid transaminase (EC 2.6.1.42). It has been argued that enzymes with high specificity have developed from low-specificity ancestors during biological evolution [12]. In pathway analysis and in biochemical modelling in general, only a few studies on enzymes with broad substrate specificity have been presented so far [12-14]. Such enzymes are often called multifunctional enzymes. However, the latter also include enzymes with more than one active site (e.g. multi-enzyme complexes). Here, we consider enzymes with only one active site, at which different substrates can be converted.

In bioinformatics and in the modelling of biochemical systems, an ever increasing role is played by online metabolic databases. Prominent examples are KEGG (<http://www.genome.ad.jp/kegg/kegg2.html>), BioCyc (<http://biocyc.org/>), ExPASy ENZYME (<http://us.expasy.org/enzyme/>), and BRENDA (www.brenda.uni-koeln.de). The former two include information on metabolic pathways, as well as information on individual enzymes. The latter has the advantage of including values of kinetic parameters of enzymes. KEGG has the helpful feature that in maps of metabolic networks, the enzymes present in particular organisms can be highlighted. Between these databases, numerous cross-links exist. By metabolic databases, the search for enzyme information is facilitated and accelerated in comparison to literature search. However, the data are sometimes less reliable.

In the present contribution, we discuss some aspects of the treatment of multifunctional enzymes in metabolic pathway analysis, for example, their suitable representation. The fact that the choice of independent functions of such enzymes is non-unique will be explained. We comment on the annotation of such enzymes in metabolic databases and give some suggestions to improve this. We then explain the proper definition of elementary modes for systems involving multifunctional enzymes and discuss some ontological problems.

REPRESENTATION OF REACTIONS CATALYSED BY ENZYMES WITH BROAD SUBSTRATE SPECIFICITY

Linearly independent functions

Let us first consider the example of transketolase. This enzyme can bind, and transfer two-carbon units between, glyceraldehyde-3-phosphate (G3P), erythrose-4-phosphate (E4P), ribose-5-phosphate (R5P), xylulose-5-phosphate (X5P), fructose-6-phosphate (F6P), sedoheptulose-7-phosphate (S7P), and several other substances which we will not consider here. Usually, two different functions are given in biochemistry textbooks:



with the equality sign denoting a reversible reaction (see, e.g. [15] and references given in [14]. However, the linear combination



is equally simple. From among the three functions Tkt1, Tkt2 and Tkt3, only two are linearly independent. Any two linearly independent functions can be chosen - one might use, alternatively, Tkt1 and Tkt3, or Tkt2 and Tkt3.

The question arises whether reaction Tkt3 really proceeds without the detour via Tkt1 and Tkt2. BRENDA provides the information that in *Oryctolagus cuniculus* (European rabbit), the reaction



occurs. The question mark means that the reaction product has not yet been verified experimentally. However, it is quite obvious that this reaction is just Tkt3, so that the products must be E4P and S7P. BRENDA is here "over-correct" in our opinion, because the products can often be inferred from other reactions, by linear combination of other functions of the same enzyme.

Nevertheless, the question remains whether for all multifunctional enzymes, all conceivable functions really occur. For non-catalysed reactions, often the assumption is made that such a transitive conclusion can be made: if reactions A and B proceed, then also reaction A-B proceeds [16].

However, this is subject to debate. Bauer [17] wrote: "Others insist, on purely statistical considerations, that all possible reactions should be included for the simple reason that since these could take place it is most likely they do, given the enormous number of molecular events involved. However, there are rational arguments in favour of a minimalist approach, to seek out the lowest number of dominant reactions required to account for the available data." For enzyme-catalysed reactions, steric hindrance may imply that reaction A-B virtually does not occur. The principle of microreversibility [18] says that the rate constants of the reactions around a cycle of reactions (such as Tkt1, Tkt2, Tkt3) fulfil the equation

$$\frac{k_{+A} \times k_{+B} \times k_{+C}}{k_{-A} \times k_{-B} \times k_{-C}} = 1.$$

It does not, however, say anything about the magnitude of the particular rate constants. It might be, for example, that k_C and k_{-C} are extremely small, so that reaction C does not proceed at a measurable rate.

Determining the rates of particular reactions in experiment meets with the difficulty of how to distinguish between the rate of reaction C itself and the "detour" reaction A-B. The convention to use Tkt1 and Tkt2 as "basic reactions" is originally due to the measurements of the entire pentose phosphate pathway by Horecker et al. [19]. They concluded that reactions Tkt1 and Tkt2 (together with transaldolase, EC 2.2.1.2, and the other monofunctional enzyme reactions of the pathway) provide the most plausible explanation for the measured exchange of radioactive label. Later, an alternative, the so-called L-type pentose phosphate pathway, has been proposed [20], which has not, however, become generally accepted. The correct sequence of events is very difficult to measure [15].

Some indications about the relative importance of the individual rates can be derived from the reaction mechanism of the enzyme and the chemical structure of the substrates. According to BRENDA, Tkt obeys a ping-pong bi-bi mechanism. This must be an ordered mechanism because Tkt catalyses a transfer reaction, which implies that the donor must bind first:



As any potential donor can be followed in the mechanism by any potential acceptor, steric hindrance should not play a role.

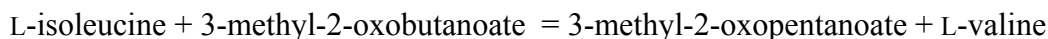
Moreover, the various substrates are very similar to each other (monophosphates of monosaccharides with 3-7 carbons). Therefore, also Tkt3 is likely to proceed at a measurable rate.

The branched-chain amino acid transaminase catalyses the reversible reaction



According to the database ExPASy ENZYME, instead of L-leucine, also L-isoleucine and L-valine can be used as substrates. BRENDA mentions, in addition, methionine, aspartate and several non-proteinogenic amino acids such as norvaline, aminopimelate, and aminobutanoate. Here, we restrict the analysis to leucine, isoleucine, valine, and glutamate. (The latter is a possible substrate as well since the reaction is reversible.) Thus, there are $4 \times 3/2 = 6$ different reactions in total. The number of linearly independent reactions equals three. Usually, the three reactions involving glutamate are taken as a "basis".

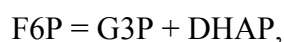
As far as chemical structure is concerned, the three branched amino acids are very similar, while glutamate differs in that it is not branched. BRENDA provides the information that in *Rattus norvegicus* and *Escherichia coli*, also the reaction



occurs. Thus, the detour via glutamate does not appear to be necessary. This makes it likely that also the direct, reversible reactions from isoleucine to leucine and from valine to leucine can occur.

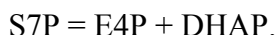
A likely reason for the usual choice of independent functions for branched-chain amino acid transaminase is the physiological role of glutamate as a molecule used for amino group transfer. Wagner and Fell [21] have shown by metabolic network analysis that glutamate is the most central metabolite except from cofactors such as ATP, ADP and NADH in that the mean pathway distance to all other metabolites is shortest. In fact, glutamate can be considered as a cofactor as well, since it transfers the amino group just as ATP transfers the phosphate group.

A further interesting example is aldolase (EC 4.1.2.13). It catalyses the reaction

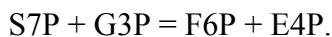


where DHAP stands for dihydroxyacetone phosphate.

In photo-autotrophic plants and some bacteria, aldolase also catalyses the reaction (see, e.g., BRENDA)



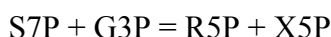
Subtracting these two reactions gives



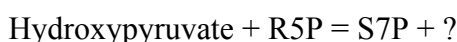
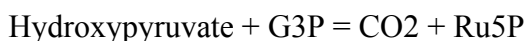
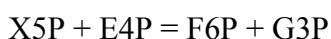
This reaction equation is not, however, as equally simple as the other two. Therefore, we will not consider such linear combinations here, although, depending on the reaction mechanism, such a reaction might be relevant as well.

Annotation in databases

It is interesting to look into metabolic databases and see which functions of enzymes with low substrate specificity are indicated. In BRENDA, for transketolase, the reaction



is given as the main reaction. Moreover, several other reactions such as



with the names of biological species in which they have been detected are listed.

In the KEGG database (and identically in the database ExPASy ENZYME), it is indicated that the main reaction is $S7P + G3P = R5P + X5P$ and the following comment is given:

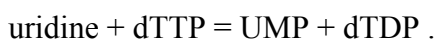
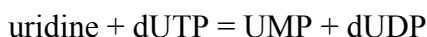
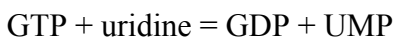
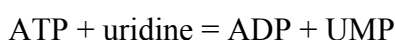
"A thiamine-diphosphate protein. Wide specificity for both reactants, e.g. converts hydroxypyruvate and R-CHO into CO_2 and R-CHOH-CO-CH₂OH. Transketolase from *Alkaligenes faecalis* shows high activity with D-erythrose as acceptor." (http://www.genome.ad.jp/dbget-bin/www_bget?enzyme+2.2.1.1).

KEGG includes a subdatabase called REACTION, which can be downloaded by anonymous ftp. REACTION comprises, for transketolase, the functions Tkt1 and Tkt2. Surprisingly, the information given in the clickable front-end of KEGG does not coincide with the data in the REACTION database.

BRENDA is much more comprehensive for this enzyme and most other enzymes in that it lists more reactions and because it includes information about the species in which they have been found.

In BioCyc, it is indicated that transketolase usually occurs in the form of two isoenzymes, TktA and TktB. According to BioCyc, TktA only performs function Tkt1, while TktB performs Tkt1 and Tkt2. According to our knowledge, however, TktA can also catalyse both reactions.

Let us now consider uridine kinase (EC 2.7.1.48). It catalyses reactions such as



The REACTION database of KEGG indicates these and 14 other reactions. The total number of 18 reactions arises from the fact that nine phosphate donors (ATP, GTP, UTP, ITP, dUTP, dTTP, dGTP, dCTP, and dATP) and two phosphate acceptors (uridine and cytidine) are indicated. However, only 10 of these reactions are linearly independent. (This number can be calculated as 22 metabolites minus 12 conserved moieties, see [14]). In contrast to transketolase, for uridine kinase, the REACTION database also gives linearly dependent reactions.

IS ANNOTATION IN TERMS OF HALF-REACTIONS AN OPTION?

Enzyme reactions can be described at two different levels. First, one may consider the half-reactions of formation, isomerization and decay of the enzyme-substrate complex. This has been done, for example, by Nuño et al. [13] and Alberty [22]. The much more common way, however, is to consider the overall reactions of conversion of substrate into product. In the former case, one needs to know the reaction mechanism. This information is not always available.

It is of interest to compare the number of half-reactions with the number of overall reactions of a multifunctional enzyme. Let us take uridine kinase as an example. Differing views on its reaction mechanism can be found in the literature.

If it is an ordered, sequential mechanism and the phosphate donor binds first [23], then there are 9 half-reactions of binding the various donors, $2 \times 9 = 18$ half-reactions of binding the phosphate acceptors to the 9 different enzyme-phosphate-donor complexes, and $9 + 18 = 27$ different half-reactions of release (Fig. 1A). In total, there are 54 different half-reactions.

If the enzymatic mechanism of uridine kinase is ordered ping-pong [24], then there are $2 \times (9 + 2) = 22$ half-reactions (Fig. 1B). In both cases, however, one can reduce the number of reactions to be considered by lumping sequential steps without branching in between. For example, the steps $\text{ATP} + \text{Urk} = \text{Urk-ATP}$ and $\text{Urk-ATP} = \text{ADP} + \text{Urk-P}$ can be combined into $\text{ATP} + \text{Urk} = \text{ADP} + \text{Urk-P}$ (where Urk stands for uridine kinase).

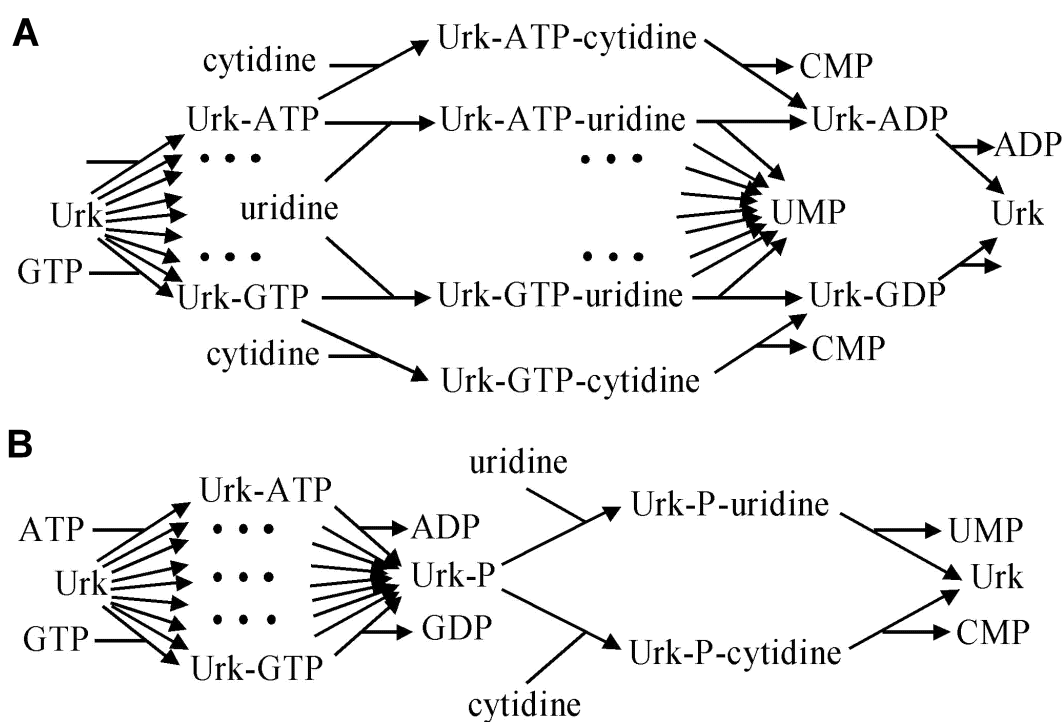


Figure 1. Possible reaction mechanisms for uridine kinase. Out of the 9 phosphate donors mentioned in the text, only two are shown explicitly for the sake of clarity. The others are referred to by "loose" arrows. Urk, uridine kinase. **A)** Ordered sequential mechanism. **B)** Ordered ping-pong mechanism.

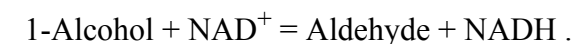
This reduces the total number to 11. So we see that for some enzymes, the list of half-reactions can be shorter than the complete list of overall reactions if sequential steps are lumped.

Nevertheless, annotation of multifunctional enzymes in terms of half-reactions in databases appears to be inappropriate because of the frequent uncertainty about the reaction mechanism and the generally high number of half-reactions.

Importantly, the number of linearly independent functions is always smaller than the number of half-reactions because it is equal to the number of independent fluxes, which cannot be greater than the number of reaction steps. Thus, an annotation in terms of linearly independent functions is certainly better suited in spite of the arbitrariness of choosing these functions.

ONTOLOGICAL PROBLEMS AND SPECIES-SPECIFIC INFORMATION

In enzyme databases as well as in the literature, substances are given at different levels of specificity. For example, in some databases, the term "branched-chain amino acids" is used while in others, the specific amino acids are mentioned separately. Another example is provided by alcohol dehydrogenase (EC 1.1.1.1). In REACTION (and similarly in several other databases), the following reactions (amongst others) are given:



If a multifunctional enzyme has been detected in an organism, it is not yet clear whether all functions of it are performed in that organism. Some substrates may be missing or the enzyme may be species-specific such that some functions are not performed. While information about the occurrence of particular functions of multifunctional enzymes in different organisms is missing in most databases, BRENDA provides much information about this. It gives lists of particular functions detected in particular organisms, e.g. for the particular functions of transketolase and branched chain amino acid transaminase in *Saccharomyces cerevisiae*, *E. coli*, *Triticum aestivum*, *Sus scrofa* and many other organisms.

This information is of importance for the reconstruction of metabolic maps. For example, the ability of aldolase to cleave S7P in some organisms may enable an alternative pentose phosphate pathway to work (for a theoretical analysis of potential monosaccharide pathways, though only up to six-carbon sugars, see [25]).

Of course, the substance class "primary alcohol" includes, for example, ethanol. This annotation makes automatic data-mining and usage for metabolic modelling difficult because common computer programs for metabolic modelling do not know that "ethanol" is a sub-concept of "alcohol". Thus, in the future, metabolic databases and/or simulation software should be extended to include a substance class ontology.

ARE THE ELEMENTARY MODES INDEPENDENT OF THE LEVEL OF DESCRIPTION?

The different functions of multifunctional enzymes should be treated as distinct reactions in the computation of elementary modes (pathways). Importantly, only a set of independent functions must be taken because linearly dependent reactions cause the occurrence of irrelevant cyclic elementary modes with no net transformation. For example, Tkt1 and Tkt2 have been used for the determination of the elementary modes of monosaccharide metabolism [11]. When using Tkt3 instead of one of the other two, the number of resulting elementary modes is the same.

As mentioned above, enzymes can be described at two different levels. The question arises as to whether the elementary modes are independent of the level of description. If all reactions are irreversible, this is really the case [14]. If some reactions are reversible, this is not necessarily so because two elementary modes may use the same reversible half-reaction in opposite directions, so that this reaction cancels when the modes are summed up. To resolve this discrepancy, it is helpful to examine the definition of elementary modes.

In the Introduction section, we have defined this concept as the "minimal set of enzymes" that fulfils certain properties. At the level half-reactions, one might be tempted to interpret this definition in terms of half-reactions rather than enzymes, and this is formally done by computer programs for computing elementary modes, e.g. METATOOL [26]. An example is the reaction system of monosaccharide metabolism studied by Nuño et al. [13]. In terms of overall reactions, this system gives rise to 296 elementary flux modes, while in terms of half-reactions, 866 modes would arise if the reaction steps were considered as basic units.

However, in the contradictory cases, several modes should be lumped because they are equivalent in terms of enzymes. If this is done, the elementary modes are indeed independent of the level of description [14]. The question of whether enzymes or reactions are the basic units in metabolism also occurs in Metabolic Control Analysis [27].

DISCUSSION

Using several example enzymes for illustration, we have outlined some problems in the modelling of enzymes with broad substrate specificity, as well as in the storage of data about such enzymes in online databases. We argue that in enzyme databases and for metabolic modelling, only independent functions of multifunctional enzymes need to be indicated, noting that linear combinations are possible.

Moreover, in many cases where not all reaction substrates or products are known, they can be inferred from other reactions by linear combination. It is worth investigating whether only such linear combinations need be considered that are equally simple (in terms of the number of reactants and products) as the "original" reactions

It is of interest to find criteria by which an appropriate choice of linearly independent functions can be made. One possibility is to choose the functions with the highest reaction rates, which requires, however, specific measurements. Another option is to choose the functions that occur in the convex basis [26] of the metabolic network in which the enzyme in question is embedded [14]. This is in line with the historic determination of the basic functions of transketolase within the pentose phosphate pathway [19]. In some cases, a third option is to seek for those reactions in which central metabolites, such as glutamate, energy currency metabolites and redox equivalents, are involved as much as possible.

In future modelling work, it will be worth analysing the evolutionary reasons why some biochemical reactions are catalysed by highly specific enzymes and others by less specific enzymes.

A compromise appears to have been found between efficient regulation and economy of the genome. For example, the branched-chain amino acid transaminase acts on very similar amino acids. If for each of these, a separate transaminase (with glutamate as amino group donor) existed, the synthesis of the amino acids could be regulated more specifically. However, more genes would be necessary in the genome. Another point may be that enzymes need to have a concentration greater than a certain lower bound in order to allow sufficiently frequent collision events. Since the synthesis of enzymes requires metabolic effort, two specialized enzymes need more effort than one less specific enzyme. A further interesting question is why enzymes exist (e.g. glucokinase, EC 2.7.1.2) the specificity of which covers part of the range of some other enzymes (e.g. hexokinase).

REFERENCES

-
- [1] Papin, J.A., Price, N.D., Edwards, J.S., Palsson, B.O. (2002a) The genome -scale metabolic extreme pathway structure in *Haemophilus influenzae* shows significant network redundancy. *J. Theor. Biol.* **215**: 67-82.
 - [2] Klamt, S., Stelling, J. (2003) Two approaches for metabolic pathway analysis? *Trends Biotechnol.* **21**: 64-69.
 - [3] Poolman, M.G., Fell, D.A., Raines, C.A. (2003) Elementary modes analysis of photosynthate metabolism in the chloroplast stroma. *Eur. J. Biochem.* **270**: 430-439.
 - [4] Gagneur, J., Jackson, D.B., Casari, G. (2003) Hierarchical analysis of dependency in metabolic networks. *Bioinformatics* **19**: 1027-1034.
 - [5] Dandekar, T., Moldenhauer, F., Bulik, S., Bertram, H., Schuster, S. (2003) A method for classifying metabolites in topological pathway analyses based on minimization of pathway number. *BioSystems* **70**: 255-270.
 - [6] Carlson, R., Fell, D.A., Sreenc, F. (2002) Metabolic pathway analysis of a recombinant yeast for rational strain development. *Biotechnol. Bioengng.* **79**: 121-134.
 - [7] Schuster, S. (2004) *Metabolic pathway analysis in biotechnology*. In: *Metabolic Engineering in the Post Genomic Era* (Kholodenko, B. N., Westerhoff, H. V., Eds) pp. 181-208. Horizon Scientific, Wymondham.
 - [8] Förster, J., Gombert, A.K., Nielsen, J. (2002) A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnol. Bioengng.* **79**: 703-712.
 - [9] Papin, J.A., Price, N.D., Palsson, B.O. (2002b) Extreme pathway lengths and reaction participation in genome-scale metabolic networks. *Genome Res.* **12**: 1889-1900.
 - [10] Schuster, S., Dandekar, T., Fell, D.A. (1999) Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.* **17**: 53-60.
 - [11] Schuster, S., Fell, D.A., Dandekar, T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnol.* **18**: 326-332.
 - [12] Kacser, H., Beeby, R. (1984) Evolution of catalytic proteins. On the origin of enzyme species by means of natural selection. *J. Mol. Evol.* **20**: 38-51.
 - [13] Nuño, J.C., Sánchez-Valdenebro, I., Pérez-Iratxeta, C., Meléndez-Hevia, E., Montero, F. (1997) Network organization of cell metabolism: monosaccharide interconversion, *Biochem. J.* **324**, 103-111.
 - [14] Schuster, S., Zevedei-Oancea, I. (2002) Treatment of multifunctional enzymes in metabolic pathway analysis. *Biophys. Chem.* **99**: 63-75.
 - [15] Flanigan, I., Collins, J.G., Arora, K.K., MacLeod, J.K., Williams, J.F. (1993) Exchange reactions catalyzed by group-transferring enzymes oppose the quantitation and the unravelling of the identify of the pentose pathway. *Eur. J. Biochem.* **213**: 477-485.
 - [16] Aris, R. (1965) Prolegomena to the rational analysis of systems of chemical reactions. *Archs Rational Mech. Anal.* **19**: 81-99.
-

- [17] Bauer, S.H. (1990) Comments on current aspects of chemical kinetics. *Int. J. Chem. Kinet.* **22**: 113-133.
 - [18] Lewis, G.N. (1925) A new principle of equilibrium. *Proc. Natl. Acad. Sci. USA* **11**: 179-183.
 - [19] Horecker, B.L., Gibbs, M., Klenow, H., Smyrniotis, P.Z. (1984) The mechanism of pentose phosphate conversion to hexose monophosphate. I. With a liver enzyme preparation. *J. biol. Chem.* **207**: 393-403.
 - [20] Williams, J.F., Clark, M.G., Blackmore, P.F. (1978) New reaction sequences for the non-oxidative pentose phosphate pathway. *Biochem. J.* **176**: 257-282.
 - [21] Wagner, A., Fell, D. A. (2001) The small world inside large metabolic networks. *Proc. R. Soc. Lond. B* **268**: 1803-1810.
 - [22] Alberty, R.A. (2001) Standard apparent reduction potentials for biochemical half reactions as a function of pH and ionic strength. *Archs Biochem. Biophys.* **389**, 94-109.
 - [23] Payne, R.C., Cheng, N., Traut, T.W. (1985) Uridine kinase from *Ehrlich Ascites Carcinoma*. Purification and properties of homogeneous enzyme. *J. biol. Chem.* **18**: 10242-10247.
 - [24] Orengo, A. (1969) Regulation of enzymic activity by metabolites. I. Uridine-cytidine kinase of *Novikoff ascites* rat tumor. *J. biol. Chem.* **8**: 2204-2209.
 - [25] Ebenhöf, O., Heinrich, R. (2003) Stoichiometric design of metabolic networks: multifunctionality, clusters, optimization, weak and strong robustness. *Bull. Math. Biol.* **65**: 323-357.
 - [26] Pfeiffer, T., Sánchez-Valdenebro, I., Nuño, J.C., Montero, F., Schuster, S. (1999) METATOOL: For Studying Metabolic Networks. *Bioinformatics* **15**: 251-257.
 - [27] Kholodenko, B.N., Molenaar, D., Schuster, S., Heinrich, R., Westerhoff, H.V. (1995) Defining control coefficients in non-ideal metabolic pathways. *Biophys. Chem.* **56**: 215-226.
-

JWS ONLINE CELLULAR SYSTEMS MODELLING AND THE SILICON CELL

JACKY L. SNOEP^{1,2,*}, BRETT G. OLIVIER¹ AND HANS V. WESTERHOFF²

¹Triple-J group, Department of Biochemistry, University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa

²Cellular BioInformatics, Free University, De Boelelaan 1087,
NL-1081 HV Amsterdam, The Netherlands

E-Mail: *jls@sun.ac.za

Received: 16th March 2004 / Published: 1st October 2004

ABSTRACT

Rapid developments in bioinformatics over the last decade, coupled with a dramatic increase in the amount of available, quantitative data has necessitated the need for good analysis tools to quantitatively understand the functioning of biological systems. Detailed kinetic models offer one such tool and while such models have been developed since the 1960s little attention has been paid to the presentation and conservation of such models. Here we focus on the JWS Online project (<http://jjj.biochem.sun.ac.za>) and its role in 1) offering a web based tool for analysis of kinetic models, 2) acting as a repository for published kinetic models and 3) facilitating the reviewing of new models. In addition we advocate the use of a specific type of kinetic models, the so-called "Silicon Cell" models (<http://www.siliconcell.net>). By elaborating on the process of constructing one such model, based on yeast glycolysis, we illustrate the approach of "modular modelling" and "model combining." This approach is presented as a preferred method to model biological systems, as opposed to the building of single large models.

INTRODUCTION

Models are tools that can be used to address many different questions. From the multitude of different kinds of models we focus here on detailed kinetic models using ordinary differential equations (ODEs) to describe biological systems. Our aim in using such models is to get a quantitative understanding of the functioning of the living cell.

The ultimate and very ambitious goal is to build a computer replica of a living cell. Our approach in doing so is a modular one. Build detailed kinetic models - modules - that can be studied, at the enzyme level, as isolated units of the complete system. The approach is a bottom up approach; the model is constructed on the basis of characterization of the individual components and their interactions. Such models are validated in independent experiments and posted in a model database. Interacting modules can be grouped together and a larger model can be constructed and revalidated. In this way the model will be gradually extended and upon each module addition a new validation can be done, eliminating the overwhelming parameter optimization that would be necessary if a large model were constructed in one step.

The level at which these detailed models are build is at the enzyme level, the enzymes being the catalytic units in the cell, the synthesis of which is regulated as a part of the cellular process. In addition many drugs are effective at the level of the enzyme activity and if our models are to be important in drug development strategies we need a good representation at the enzyme activity level. Of course here the modeller can use the huge body of knowledge that is available from the field of enzymology. At the core of the models lie the kinetic rate equations, which have been studied for over a century by enzymologists. However enzymologists and system biologists have a different topic of study, where for the former the enzyme is the system, the latter is interested in a set of enzymes working together in a cellular environment. This has important implications for the conditions under which the enzymes are characterized.

Of course an ambitious project such as attempting to build a silicon cell cannot be achieved in a single research group, but will be dependent on the collaboration between a large number of groups active in experimental as well as modelling and theoretical fields. Of crucial importance in such collaborations will be the standardization of experimental conditions for enzyme kinetic measurements, decision on model organism(s) and growth condition(s) that will be modelled.

Global initiatives such as the newly formed *Systems Biology for Yeast* might play an important role in coordinating such projects while organizations such as the Beilstein Institut could play an important role in catalysing the standardization procedures.

In these proceedings we will treat several of the aspects that are, in our view, important in the process of building the silicon cell and show how the JWS Online Cellular System modelling project can be used to achieve them.

We will start by introducing the JWS project and explain its function as an easy to use, web-based modelling tool, a model repository, and its role in model curating. We will finish by illustrating the bottom-up modelling approach as advocated in the Silicon Cell project using our detailed kinetic model of yeast glycolysis as an example.

MOTIVATION OF THE JWS PROJECT

ODE based models of biological systems have been used for over 40 years and many models have been published. Analysis of these models is usually heavily based on computer simulations due to the non-linear character of the ODE's, and whereas such analyses were limited by computer strength in the early years, these limitations have been largely overcome and it is now possible to simulate large sets of ODE's on today's desktop computers. Not only have there been rapid advances in the development of computer hardware but the development of good numerical algorithms for solving differential equations have made it easier to build and analyse kinetic models. In addition, several specialist software packages for simulations as well as general mathematical programmes now incorporate these algorithms, allowing the user to work in a high level environment. Still a certain amount of knowledge is necessary to build and analyse kinetic models and for someone that is not initiated in this theory and who would like to quickly check an existing model for its ability to describe a set of experimental data there is no off the shelf tool to do so. First, a new user would have to figure out what software tools are available and make a decision on which one to use. Second, the user would have to acquire the software (this is often easy as most of the packages are free of charge and can be downloaded from the internet), and install the software on his/her computer. Thirdly, the user would have to learn how to use the software and with many of the packages having a multitude of options this is not necessarily simple. After all this, there is still no model to run and this might be the biggest hurdle to overcome. The user is interested in running an existing model but where can such models be obtained? Of course the user may contact the author and in such a way receive the model in digital form but this might be in the dedicated format of whatever software package the model builder has used and this is not necessarily compatible with the easy to use package the user has selected. Often model descriptions only exist in literature, an electronic description of the model is no longer available and the user will have to code the model from the manuscript, a non-trivial task.

To overcome these problems we have started the JWS Online Cellular Systems modelling project, a repository of kinetic models that can be run over the internet using a standard web browser [1].

JWS AS A REPOSITORY OF MODELS

The JWS project was started as an effort to collect existing kinetic models of biological systems and present/preserve them in an easy to use format. It quickly became apparent that many of the models described in the literature run the risk of being lost. The description of models in manuscripts is often very poor and electronic versions of the models have been lost or formats have become outdated. No official repository of kinetic models existed and no standard way of presenting kinetic models in the literature has been agreed upon.

Two initiatives to standardize a model description of biological systems, using XML based exchange formats, should be mentioned: CellML (<http://www.cellml.org>) and SBML (<http://www.sbml.org>). Both these initiatives also have a list of kinetic models that can be downloaded in either CellML or SBML format. The number of programs that can load these file formats and that can be used to run these models is still limited (especially for CellML) but the list of programs supporting SBML is growing rapidly.

The focus of the JWS project is not so much on an exchange format but on building a repository of kinetic models. These models can be directly run on the web site (see below), thus no exchange format is necessary. However, in order to allow users to run the listed models on a stand-alone computer we are making a number of different formats available. Currently a download feature for models in SBML format is available for a number of models, illustrating the collaboration between the SBML group at CalTech and the JWS project in exchanging models. Future downloads will be made available as Mathematica notebooks Copasi, and PySCeS formats.

Currently the database holds 19 models in the Silicon Cell category, shown in Fig. 1. In addition to the Silicon Cell category two other types of models are stored, 1) Core models, minimal models that illustrate a specific hypothesis or idea and are not necessarily based on realistic kinetics, and 2) Demonstration models, simple models mostly used for teaching or demonstration purposes.

We try to include as many models as possible and invite users to submit to us models that they would like to include in the database. Please contact us via e-mail (jls@sun.ac.za) for submission formats.

The Silicon Cell: detailed metabolic models			
Detailed glycolytic model in <i>Lactococcus lactis</i> - model	Hoefnagel <i>et al.</i> - 2002	more	
Glycolysis in <i>Trypanosoma brucei</i> - model	Bakker <i>et al.</i> - 2001	more	sbml
A Computational Model for Glycogenolysis in Skeletal Muscle - model	Lambeth <i>et al.</i> - 2002	more	sbml
Pyruvate branches in <i>Lactococcus Lactis</i> - model	Hoefnagel <i>et al.</i> - 2002	more	sbml
Glycolysis in <i>Saccharomyces cerevisiae</i> - model	Teusink <i>et al.</i> - 2000	more	sbml
Sucrose accumulation in sugarcane - model	Rohwer <i>et al.</i> - 2001	more	
Bacterial phosphotransferase system - model	Rohwer <i>et al.</i> - 2001	more	
Threonine synthesis pathway in <i>E. coli</i> - model	Chassagnole <i>et al.</i> - 2001	more	
Kinetics of Histone Gene Expression - model	Koster <i>et al.</i> - 1988	more	
Glycolysis in <i>Saccharomyces cerevisiae</i> , 6 variables - model	Galazzo <i>et al.</i> - 1990	more	
Full scale model of glycolysis in <i>Saccharomyces cerevisiae</i> - model	Hynne <i>et al.</i> - 2001	more	
Quantification of Short Term Signaling by the Epidermal GFR - model	Kholodenko <i>et al.</i> - 1999	more	
Red Blood Cell Model - model	Mulquiney <i>et al.</i>		
Mechanism of protection of peroxidase activity by oscillatory dynamics - model	Olsen <i>et al.</i> - 2003	more	
Dynamic model of <i>Escherichia coli</i> tryptophan operon - model	Bhartiya <i>et al.</i> - 2003	more	
MCA of Glycerol Synthesis in <i>Saccharomyces cerevisiae</i> - model	Cronwright <i>et al.</i> - 2003	more	
Mathematical modelling of the urea cycle - model	Maher <i>et al.</i> - 2003		
A kinetic model of the branch-point between the methionine ... - model	Curien <i>et al.</i> - 2003	more	
Modelling Photosynthesis and its control - model	Poolman <i>et al.</i> - 2000	more	

Figure 1. Screen capture of the current (December 2003) models in the Silicon Cell category of the JWS repository of kinetic models. Each of the models can be interrogated by clicking on its specific model link. Literature references to the specific models are listed under 'more links'. Download options in SBML format is limited to only four models at present via the SBML link.

JWS AS AN EASY TO USE WEB BASED SIMULATOR

JWS allows users to run and interrogate kinetic models via any browser that is capable of running Java2 applets. Most modern browsers support the SUN Microsystems J2RE plug-in and modern versions have this plug-in pre-installed. The system has been tested using the Microsoft windows operating system 98 and higher using Internet Explorer 5+ and Netscape 6+, using Mac OS X and the SAFARI browser, and MOZILLA under Linux. Using any of these browsers the user can type in the URL for the JWS site (<http://jjj.biochem.sun.ac.za>) and after selecting the database link on the home page the user can interrogate any of the models listed in the database by clicking on the model link (see Fig. 1). Upon selection of a model a graphical interface is downloaded to the user as an applet, in which the user can change parameter values and select the type of simulation requested.

An example of such an applet is given in Fig. 2.

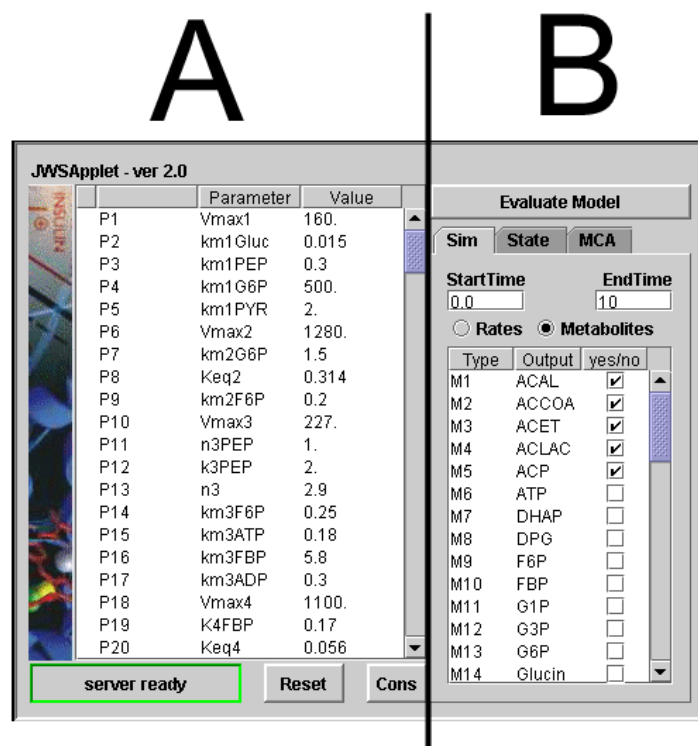


Figure 2. Screen capture of an example model applet. The applets consist of two panels: **A**) a scrollable table listing all parameter values and initial conditions, **B**) a control panel on which the user can select to either do a time simulation (Sim), a steady-state (State) analysis or a control analysis (MCA) by clicking on either of the three tabs at the top of the panel. The model simulation is started by pressing the "evaluate model" button.

In the left hand panel, (Fig. 2A) model parameters and initial conditions can be changed by the user in the scrollable table. The right hand panel (Fig. 2B) is used to select between the different simulations, i.e. time simulations (Sim), steady-state analysis (State) or metabolic control analysis (MCA), by clicking on the respective tabs. In Fig. 3 the respective panels are shown as if each of the three options were selected. In Fig. 3A the user has selected the Sim option, (default) and in this option the user can set the begin and end time of the simulation, and whether metabolites or rates should be plotted in the resulting graph. Figure 3B lists the options available for steady-state analysis. These options include: Steady State, N, K, or L matrix, Jacobian and Eigenvalues.

After the relevant options have been selected and the evaluate button pressed, a table with the steady-state metabolite concentrations, fluxes or the requested structural matrix will be displayed.

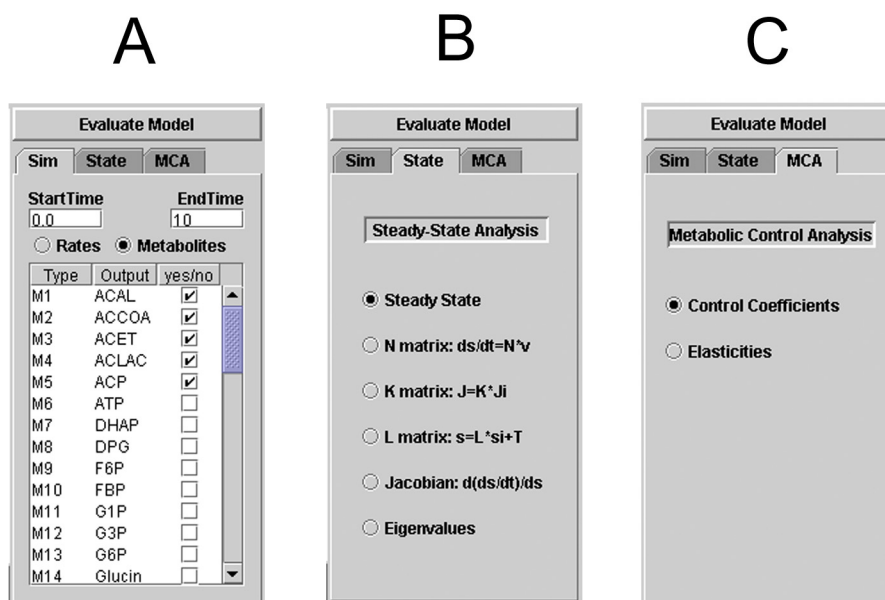


Figure 3. A), B), C), the control panel of the JWS applet showing the three types of simulation and their options.

In addition to the applet, the user also downloads a scheme of the model system. In this reaction scheme each of the catalytic steps is indicated and by moving the mouse cursor over the enzyme steps, the rate equation used in the model is shown in a panel below the applet.

Upon pressing the evaluate model button the user sends a request to the JWS server to analyse the model according to the selected options. After the analysis is complete a GIF file containing the result is send back to the user and displayed in a separate window (Fig. 4).

The best way to try the JWS simulation engine is of course to simply point your browser to the URL of any of the mirror sites and try it out yourself. Currently there are three sites for accessing the JWS models: the main site at the University of Stellenbosch, South Africa (<http://jij.biochemistry.sun.ac.za>), and two mirror sites, one at the Vrije Universiteit in Amsterdam (<http://www.jij.bio.vu.nl>) and at the Virginia Bioinformatics Institute, U.S.A. (<http://jij.vbi.vt.edu>). Try the site and give us feedback either via e-mail (jls@sun.ac.za) or via the model forums.

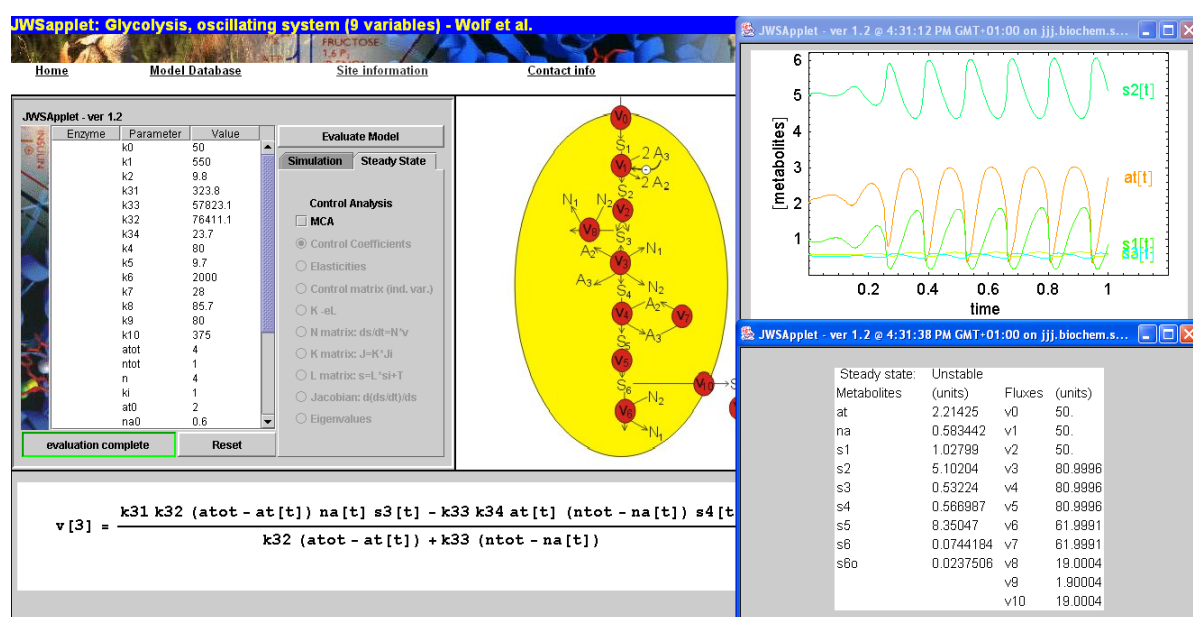


Figure 4. A typical screen capture with applet, scheme, a rate equation and some result windows.

JWS AS A CURATION TOOL FOR MODELS

Initially models to be included in the JWS repository were coded from the literature and a surprisingly high percentage of model descriptions were not complete, vague or lead to results different from those published. This indicated a weakness in the reviewing process of manuscripts containing kinetic models. Whereas it is assumed that a manuscript containing a model description is also checked on the correctness of the model, in practice such a check is usually not done. This is not surprising since it takes a lot of time to code a model from the literature and to rerun the simulations performed by the authors might not be a trivial task.

To improve the quality of model descriptions in manuscripts, and to assist in the reviewing process of manuscripts containing models, we recently started collaborating with the journals *Microbiology* and *European Journal of Biochemistry* to upload models onto a secure part of the JWS site. If a manuscript containing a suitable kinetic model is submitted to these journals the authors are requested to submit a model description in electronic form to JWS Online. Subsequently, the model will be made accessible to the authors and reviewers of the manuscript. After acceptance of the manuscript the model will be transferred to the public part of the database.

This collaboration with scientific journals assures a steady supply of new models to the database. In addition it makes models described in the literature more accessible, a direct link in the manuscript is given to the model applet, thus a reader of the manuscript has direct access to the model via the Internet.

THE SILICON CELL PROJECT

We initiated the Silicon Cell project at the Vrije Universiteit in Amsterdam to advocate the use of specific kinds of kinetic models (<http://www.siliconcell.net>). Such models, coined Silicon Cell models are intended to be replicas of (parts of) the cellular metabolism. What distinguishes this approach from other modelling strategies is that the parameters must be experimentally determined. The ultimate aim is to be able to make a quantitative description of the cellular metabolism using a kinetic model. Such models would be very powerful tools, for instance, in medical and metabolic engineering studies.

The Silicon Cell models are collected in the JWS database and separate models can easily be linked to bigger models (see below).

BUILDING A SILICON CELL TYPE MODEL OF YEAST GLYCOLYSIS

As an example model to illustrate the Silicon Cell approach we will use the detailed kinetic model on yeast glycolysis published by our group [2]. Although several kinetic models have been published on yeast glycolysis, we were interested in a specific question, can we describe yeast glycolysis on the basis of its enzyme kinetics as measured in isolated form *in vitro*?

This is an important question, rooted in a reductionist approach, can we describe the whole system on the basis of the characteristics of the isolated components? If this question can be answered positively this would help tremendously in building kinetic models of the living cell since we can then use the whole field of biochemistry in an integrated approach for studying systems biology.

The first step in the modular approach of building a Silicon Cell is to delineate the module. We were interested in studying yeast glycolysis under conditions where this pathway was isolated, as much as possible, from the rest of metabolism. Thus, in our first attempt we chose the linear pathway from glucose to ethanol as our system. We worked under anaerobic conditions so no carbon was converted via the citric acid cycle and since we worked under non-growing conditions we ignored anabolic routes.

Thus, we focused on the enzymes in the Embden-Meyerhof-Parnas pathway with as additional enzymes glucose transport, pyruvate decarboxylase and alcohol dehydrogenase.

As stated above, we should be careful in choosing the conditions under which to study the isolated components. In a systems biology approach it is important to study the systems components under the conditions under which they are active in the cell. This could be a very difficult task, it would be virtually impossible to mimic the cellular environment precisely and if the enzyme characteristics are crucially dependent on these conditions then erroneous results might be obtained.

We deliberately chose yeast glycolysis as a system because it has been studied extensively. Thus, many of the enzymes had been characterized in terms of kinetic mechanisms, and parameter values have been determined. As a first assumption we used the kinetic mechanism as has been published but noticed that for many of the enzymes the assay conditions were different from those prevailing in the cytosol. Thus we re-measured many of the enzyme kinetic parameters using the same assay buffer for all the enzymes [2].

After the enzyme kinetic information was thus collected, a kinetic model was build and analysed. This initial kinetic model did not lead to a steady state. When subsequently compared to experimental *in vivo* data (i.e. substrate consumption and product formation rates by a yeast culture) a number of branches were shown to be active in the system. This result shows the importance of delineating the system correctly; clearly we had not incorporated all the reactions that were active in the system. When branches to glycogen, trehalose, succinate and glycerol were added to the model (with simplified kinetics as only limited information on the kinetics of these reactions was available), a steady state was obtained.

Validation is an important aspect of the modelling process and should be done independent of the model construction, i.e. one cannot fit kinetic parameters on the same data set that will be used for the validation. In our silicon cell approach there is an even stronger restriction; kinetic parameters must be determined on enzymes and cannot be determined on system data sets. This does not mean that only *in vitro* measurements on purified enzymes can be used, with NMR techniques one can also obtain kinetic information on enzyme activities *in vivo* as a function of its substrate and product concentrations.

Validation of the yeast glycolytic model revealed that the model quite accurately described the pathway fluxes, i.e. within 10% of the experimentally determined values, but that some of the intermediate concentrations were not accurately described i.e. some were more than a factor 5 off. Important as a next step of validation is to go back to the model and check whether the model can describe the experimental data precisely given a certain error range of the experimentally determined parameter values. The model could, within a 5% error margin for the kinetic parameter values, give a precise description of the experimental data set. This result shows that the deviation observed between the model prediction and the experimental *in vivo* data set does not mean that the model is essentially wrong. A next important step would be to validate the model for different steady-state conditions. Validation should be an iterative process between model and experiment. Thus, after model construction it is tested against an experiment and differences are analysed, preferably going back to the isolated step. If a difference can be linked to a specific enzyme in the model but cannot be resolved within the measurement error of the kinetic parameters of that enzyme then this indicates that the rate equation used is not correct. Possibly a regulatory link is not included in the rate equation or the assay conditions under which the enzyme was measured were too different from the *in vivo* conditions. The problem can be addressed from both the modellers and the experimentalists side. The modeller can check what needs to be changed to the existing kinetic parameters to make the rate equation fit the experimental data and check experimentally whether such a value is realistic. The experimentalist can further characterize the enzyme *in vitro*. Also *in vivo* data, for instance other steady-state conditions might help to pinpoint errors in the rate equation.

In addition to validations using steady-state experimental data a more stringent validation test can be made on the dynamic behaviour of the model. Experimentally yeast has been observed to show limit cycle oscillations in the glycolytic pathway. Our model shows a stable steady-state solution and at first sight this might appear to be in disagreement with the dynamic experimental data. However here one should realize that the limit cycle oscillations in yeast glycolysis are only observed under specific experimental conditions. These include harvesting the yeast cells in a specific phase of the growth curve (several hours after glucose run-out) and after addition of cyanide and working at relatively high cell densities [3, 4]. It is known that yeast has a rapid change in expression of glucose transporters after the run out of glucose and our model could be made to oscillate after changing the activity of the glucose transport and the ATP hydrolysis reactions.

Thus, qualitatively the model is in agreement with the experimental dynamic behaviour in that it can show limit cycle oscillations. However, the description is not very precise; the frequency of the oscillations is 0.5 min^{-1} while the experimental frequency is 1.5 min^{-1} and the amplitude of the oscillations of the metabolites is too low.

To be able to observe the limit cycles experimentally the oscillations in different cells must be synchronized. Without synchronization small phase differences in the yeast cells would level out the oscillations of the population. Our hypothesis that synchronization works through acetaldehyde, a volatile compound that diffuses rapidly through the membrane and can thus act as a communicating molecule [5]. Cyanide complexes with acetaldehyde and oscillations are only observed in a narrow range of cyanide concentrations. The detailed kinetic model also showed synchronisation via acetaldehyde if the concentration of yeast cells chosen was high enough.

BUILDING THE SILICON YEAST CELL

One of the biggest challenges of systems biology is to construct detailed models of complete cells. The large number of parameters and limited information on interactions especially between macromolecules makes the construction of such models a daunting task. We would propose to tackle the problem by dividing the cell up in a large number of modules for each of which a model is build that is validated on its own. These models are collected in the JWS Online database and can be linked to other existing models on parts of yeast metabolism. When models are grouped together another round of validations needs to be made after which subsequent models can be added.

To illustrate the approach we have linked three independently build models of parts of yeast metabolism. We have taken the yeast glycolytic model as described by Teusink et al. [2] as the core and have replaced the simple description for glycerol formation by a detailed description as given by Cronwright et al. [6]. We have also added another branch of glycolysis, the Methylglyoxal pathway as described by Martins et al. [7]. Importantly, the overall description of the steady-state metabolite levels was significantly improved in the total model as compared to the glycolysis model on its own.

Of course such a project would need the support of a large number of experimental and modelling groups to work together. At the ICSB 2003 the international workshop for yeast was started and such a workgroup could possibly coordinate such a project.

REFERENCES

- [1] Olivier, B.G., Snoep, J.L. (2004) Web-based modelling using JWS Online. *Bioinformatics* (in press).
 - [2] Teusink, B., Passarge, J., Reijenga, C.A., Esgalhado, E., Van der Weijden, C.C., Schepper, M., Walsh, M.C., Bakker, B.M., Van Dam, K., Westerhoff, H.V., Snoep, J.L. (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur. J. Biochem.* **267**: 5313-5329.
 - [3] Richard, P., Bakker, B.M., Teusink, B., Westerhoff, H.V., Van Dam, K. (1993) *Synchronisation of glycolytic oscillations in intact yeast cells*. In: Modern Trends in Biothermokinetics (Schuster, S., Rigoulet M., Ouhabi, R., Mazat, J.P. Eds) pp.413-416. Plenum Press, London.
 - [4] Richard, P., Teusink, B., Westerhoff, H.V., Van Dam, K. (1994) Around the growth phase transition *S. cerevisiae's* make-up favours sustained oscillations of intracellular metabolites. *FEBS Lett.* **318**: 80-82.
 - [5] Richard, P., Teusink, B., Van Dam, K., Westerhoff, H.V. (1996) Acetaldehyde mediates the synchronization of sustained glycolytic oscillations in yeast-cell populations. *Eur. J. Biochem.* **235**:238-241.
 - [6] Cronwright, G.R., Rohwer, J.M., Prior, B.A. (2003) Metabolic control analysis of glycerol synthesis in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* **68**: 4448-4456.
 - [7] Martins, A.M., Mendes, P., Cordeiro, C., Freire A. P. (2001) In situ kinetic analysis of glyoxalase I and glyoxalase II in *Saccharomyces cerevisiae*. *Eur. J. Biochem.* **268**: 3930-3936.
-

CONTROLLED VOCABULARIES AND ONTOLOGIES IN ENZYMOLOGY

KIRILL DEGTYARENKO

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton,
Cambridge CB10 1SD, U.K.

E-Mail: kirill@ebi.ac.uk

Received: 8th January 2004 / Published: 1st October 2004

ABSTRACT

The diversity of objects and concepts in enzymology can be reflected in the number of possible classifications ('ontologies') needed to describe an 'elementary' biochemical event such as an enzymatic reaction: for example, the overall enzymatic reaction (including direction) taking place under physiological conditions; any other enzymatic reaction catalysed by the same enzyme observed *in vivo* or *in vitro*; the biochemical pathway of which the reaction is part; the mechanism of the enzymatic reaction; an enzyme itself; any of the subunits of a multimeric enzyme. These are all different classes of entities and as such have to be given their own terms and/or identifiers. In reality, the terminology used in publications or biological databases often is a mixture of terms borrowed from orthogonal (or contradicting) classifications.

In this respect, the Enzyme Nomenclature should provide the ultimate reference, whereas in fact it suffers the same problem. EC numbers form a strict hierarchy of *IsA* relationships and the enzymes often require re-classification. It is unlikely that in its current form, the Enzyme Nomenclature can cope with the growing demands of the biological and bioinformatics community in the 21st century. A more flexible, but at the same time a more strictly defined, approach has been pioneered by the Gene Ontology Consortium, which provides controlled vocabulary for *molecular functions* used to annotate *gene products*. I am going to discuss the building of an Enzyme Ontology. Here, novel relationships unique to chemical ontologies have to be introduced.

A FEW NOTES IN LIEU OF AN INTRODUCTION

The Hitch Hiker's Guide to the Galaxy is an indispensable companion to all those who are keen to make sense of life in an infinitely complex and confusing Universe, for though it cannot hope to be useful or informative on all matters, it does at least make the reassuring claim, that where it is inaccurate it is at least definitely inaccurate. In cases of major discrepancy it's always reality that's got it wrong.

(Douglas Adams)

'Ontology' is a formal definition of concepts (such as entities and relationships) of a given area of knowledge, described in a standardized form [1]. It can be organized as a structured vocabulary in the form of a directed acyclic graph or a network in which each term may be a 'child' of one or more 'parent' [2].

Naturally, sequences form the *core data* of biological sequence databases such as EMBL [3] and Swiss-Prot (SW) [4], while 3-D coordinates form the core data of structural databases such as the Protein Data Bank (PDB) [5]. The other data typically found in a database entry, such as the name of a gene or protein, the name of the organism, or literature references, are called *annotation*. In contrast, the Enzyme Nomenclature [6] and Gene Ontology (GO) [2] contain a fundamentally different kind of core data: terms and definitions. Thus GO contains free text definitions as in a dictionary, while in the Enzyme Nomenclature it is the chemical reactions themselves which characterize the enzymatic function.

Throughout this paper, I provide examples from the biological databases in the form (DatabaseName:AccessionNumber), e.g. SW:P15335, GO:0016610, PDB:1H4J. The database entries themselves contain comprehensive literature and database references.

VOCABULARIES AND ONTOLOGIES FOR BIOCHEMICAL REACTIONS

Enzyme nomenclature

The Enzyme Nomenclature provides the oldest controlled vocabulary for biochemical function. It was originally developed by the Enzyme Commission (EC) of the International Union of Biochemistry and now is published by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) [6]. The EC number serves as both unique identifier (ID) and descriptor of the enzyme place in hierarchy. Not surprisingly, EC numbers are often used (and misused) for annotation of gene products.

It is important to remember that the basis of the Enzyme Nomenclature is the *overall reaction catalysed* [7], but not the reaction mechanism or any other specific property of an enzyme. Nevertheless, other biological catalysts such as ribozymes [8], deoxyribozymes [9] or catalytic antibodies (abzymes) [10] do not form a part of the Enzyme Nomenclature. EC numbers form a *taxonomy*, a strict hierarchy of parent-child relationships known as *IsA* ('is kind of').

IntEnz

At the EBI, enzyme classification is collected in the Integrated relational Enzyme database (IntEnz) [11], a joint project with Trinity College Dublin, the Swiss Institute of Bioinformatics and the University of Cologne, supported by the NC-IUBMB. Currently, IntEnz contains enzyme data curated and approved by the members of NC-IUBMB. The goal is to create a single relational database containing all the relevant enzyme data, including those from the ENZYME [12] and BRENDA [13] databases.

GO

The Gene Ontology Consortium [2] has been developing controlled vocabularies which are widely used by the bioinformatics community. GO comprises three domains: 'molecular function', 'biological process' and 'cellular component'. In each domain, the GO terms are organized as a directed acyclic graph (DAG), which differs from a taxonomy in that a child term can have one or more parent terms. GO uses two parent-child relationships, *IsA* and *IsPartOf*. Throughout the text, I will use symbols for these relationships as specified in the GO File Format Guide [14]:

%	=	<i>IsA</i>
<	=	<i>IsPartOf</i>

GO is a part of the wider initiative known as OBO (Open Biology Ontologies). A list of freely available ontologies that are relevant to genomics and proteomics and which are structured in same way as GO can be found at the OBO website [15].

Specialized enzyme function resources

These include MEROPS, the Peptidase Database [16], REBASE, The Restriction Enzyme Database [17], CAZy, the database of carbohydrate-active enzymes [18], and UM-BBD,

The University of Minnesota Biocatalysis/Biodegradation Database [19]. The first three databases aim to provide comprehensive in-depth information about corresponding groups of enzymes. For example, REBASE (as on the 16 December 2003) contains information on 3631 restriction enzymes, which correspond to only three EC numbers: EC 3.1.21.3, EC 3.1.21.4 and EC 3.1.21.5. UM-BBD is not focused on any particular groups of enzymes. Currently, it lists 573 enzymes, many of which do not have EC numbers assigned, and 900 (enzymatic and non-enzymatic) reactions of xenobiotic biodegradation in micro-organisms.

WHAT IS THE MEANING OF AN ENZYME NAME?

Both in the literature and in biological databases, enzyme names and EC numbers are used to describe various concepts such as:

- an enzyme
- any subunit of a multimeric enzyme
- an enzyme system
- the overall enzymatic reaction (including the direction) taking place under physiological conditions
- any of the enzymatic reactions catalysed by the same enzyme observed *in vivo* or *in vitro*
- the reaction in the metabolic pathway context

Note that the first three concepts are biochemical objects while the last three are biochemical events. To illustrate the difference between these concepts, consider several database entries containing the term "nitrogenase" (EC 1.18.6.1), starting with the IntEnz entry (Example 1).

Controlled Vocabularies and Ontologies in Enzymology

Example 1. EC 1.18.6.1 in IntEnz

IUBMB Enzyme Nomenclature
EC 1.18.6.1

Common name:	Nitrogenase
Reaction:	$3 \text{ reduced ferredoxin} + 6 \text{ H}^+ + \text{N}_2 + n \text{ ATP} = 3 \text{ oxidized ferredoxin} + 2 \text{ NH}_3 + n \text{ ADP} + n \text{ phosphate}$
Systematic name:	reduced ferredoxin:dinitrogen oxidoreductase (ATP-hydrolysing)
Comments:	An iron-molybdenum protein. Acetylene can also act as acceptor; in the absence of other acceptors, H^+ is reduced to H_2 ; n is about 12 - 18. Formerly EC 1.18.2.1
Links to other databases:	BRENDA, EXPASY, GO, KEGG, WIT CAS registry number: 9013-04-1
References:	1. Zumft, W.G., Paneque, A., Aparicio, P.J. and Losada, M. Mechanism of nitrate reduction in <i>Chlorella</i> . <i>Biochem. Biophys. Res. Commun.</i> 36(1969) 980-986.[Medline UI: 70008605]

[EC 1.18.6.1 created 1978 as EC 1.18.2.1, transferred 1984 to EC 1.18.6.1]

The comment to this entry mentions that nitrogenase is "an iron-molybdenum protein". Now let us see how this enzyme is presented in a protein sequence database. Swiss-Prot entries are species-specific. For instance, there are 11 entries for EC 1.18.6.1 from *Azotobacter vinelandii* (Example 2). These constitute three enzymes: iron-molybdenum nitrogenase or dinitrogenase (SW:P07328, SW:P07329, SW:P00459), iron-vanadium nitrogenase or dinitrogenase 2 (SW:P16855-P16857, SW:P15335) and iron-iron nitrogenase or dinitrogenase 3 (SW:P16266-P16268). Note that only the first name in the description line provides an unequivocal name for a polypeptide in this species; the other names may be shared between several entries (e.g. eight instances of "Nitrogenase component I") while all of them are referred to as EC 1.18.6.1.

Degtyarenko, K.

Example 2. EC 1.18.6.1 from *Azotobacter vinelandii* in Swiss-Prot.

ID	AC	Description	Cofactor
NIFD_AZOVI	P07328	Nitrogenase molybdenum-iron protein alpha chain (EC 1.18.6.1) (Nitrogenase component I) (Dinitrogenase).	FeMoco
NIFK_AZOVI	P07329	Nitrogenase molybdenum-iron protein beta chain (EC 1.18.6.1) (Nitrogenase component I) (Dinitrogenase).	Fe ₈ S ₇
NIH1_AZOVI	P00459	Nitrogenase iron protein 1 (EC 1.18.6.1) (Nitrogenase component II) (Nitrogenase Fe protein 1) (Nitrogenase reductase).	Fe ₄ S ₄
VNFD_AZOVI	P16855	Nitrogenase vanadium-iron protein alpha chain (EC 1.18.6.1) (Nitrogenase component I) (Dinitrogenase 2 alpha subunit).	FeVco
VNFK_AZOVI	P16856	Nitrogenase vanadium-iron protein beta chain (EC 1.18.6.1) (Nitrogenase component I) (Dinitrogenase 2 beta subunit).	Fe ₈ S ₇ ?
VNFG_AZOVI	P16857	Nitrogenase vanadium-iron protein delta chain (EC 1.18.6.1) (Nitrogenase component I) (Dinitrogenase 2 delta subunit).	
NIH2_AZOVI	P15335	Nitrogenase iron protein 2 (EC 1.18.6.1) (Nitrogenase component II) (Nitrogenase Fe protein 2) (Nitrogenase reductase).	Fe ₄ S ₄
ANFD_AZOVI	P16266	Nitrogenase iron-iron protein alpha chain (EC 1.18.6.1) (Nitrogenase component I) (Dinitrogenase 3 alpha subunit).	Feco?
ANFK_AZOVI	P16267	Nitrogenase iron-iron protein beta chain (EC 1.18.6.1) (Nitrogenase component I) (Dinitrogenase 3 beta subunit).	Fe ₈ S ₇ ?
ANFG_AZOVI	P16268	Nitrogenase iron-iron protein delta chain (EC 1.18.6.1) (Nitrogenase component I) (Dinitrogenase 3 delta subunit).	
NIH3_AZOVI	P16269	Nitrogenase iron protein 3 (EC 1.18.6.1) (Nitrogenase component II) (Nitrogenase Fe protein 3) (Nitrogenase reductase).	Fe ₄ S ₄

Example 3. Nitrogenase in GO.

```

Molecular function
%oxidoreductase activity\, acting on reduced ferredoxin as donor\,
dinitrogen as acceptor ; GO:0016732 ; EC:1.18.6.-
%nitrogenase activity ; GO:0016163 ; EC:1.18.6.1 ;
MetaCyc:NITROGENASE-RXN ; UM-BBD_enzymeID:e0395
Cellular component
<intracellular ; GO:0005622 ; synonym:protoplasm
<nitrogenase complex ; GO:0016610
%iron-iron nitrogenase complex ; GO:0016611
%molybdenum-iron nitrogenase complex ; GO:0016612
%vanadium-iron nitrogenase complex ; GO:0016613

```

In GO, the EC number is mapped to its molecular function (Example 3).

This is reflected in the term 'activity' now customarily added to the enzyme name. In the GO molecular function ontology, *nitrogenase activity* (GO:0016163) is defined as:

Controlled Vocabularies and Ontologies in Enzymology

"Catalysis of the reaction: 8 reduced ferredoxin + 8 H⁺ + nitrogen + 16 ATP = 8 oxidized ferredoxin + 2 ammonia + 16 ADP + 16 phosphate". In the GO cellular component ontology, nitrogenase complex (GO:0016610) is defined as "An enzyme complex composed of two proteins, dinitrogenase and nitrogenase reductase; dinitrogenase is tetrameric with an alpha₂-beta₂ structure and nitrogenase reductase is a homodimer, and both are associated with metal ions, which differ between species. Both proteins are required for the enzyme activity of the complex, the formation of oxidized ferredoxin and ammonia from reduced ferredoxin and nitrogen". The only missing logical link is the one between the object (nitrogenase complex) and its function (nitrogenase activity).

Apart from recommended names, the enzyme names used in the literature and databases annotation often include various 'functional qualifiers' (see Example 4, underscored terms). Many organisms have more than one enzyme with the same EC number, and therefore such qualifiers are needed to provide a less ambiguous description of the enzyme function.

Example 4. Some functional qualifiers for enzyme names in Swiss-Prot.

ID	AC	Description	EC number
NASC_BACSU	P42434	<u>Assimilatory</u> nitrate reductase catalytic subunit	EC 1.7.99.4
SIR_DEVSVH	Q05805	Sulfite reductase, <u>assimilatory-type</u>	EC 1.8.-.-
DSVA_DESVH	P45574	Sulfite reductase, <u>dissimilatory-type</u> alpha subunit	EC 1.8.99.3
3DHQ_ACICA	Q59087	<u>Catabolic</u> 3-dehydroquinate dehydratase	EC 4.2.1.10
MBHL_ALCEU	P31891	<u>Uptake</u> hydrogenase large subunit	EC 1.12.99.6
OTCC-PSESH	P23752	Ornithine carbamoyltransferase, <u>phaseolotoxin-insensitive</u> , <u>catabolic</u>	EC 2.1.3.3
CN8B_HUMAN	O95263	<u>High-affinity cAMP-specific</u> and <u>IBMX-insensitive</u> 3',5'-cyclic phosphodiesterase 8B	EC 3.1.4.17
PBFB_VIBCH	Q9KUCo	Penicillin-binding protein 1B [Includes: <u>Penicillin-insensitive</u> transglycosylase <u>Penicillin-sensitive</u> transpeptidase]	EC 2.4.1.129 EC 3.4.-.-
ABP_HUMAN	P19801	<u>Amiloride-sensitive</u> amine oxidase [copper-containing] precursor	EC 1.4.3.6
NFSA-ECOLI	P17117	<u>Oxygen-insensitive</u> NADPH nitroreductase	EC 1.-.-.-
AOX2_ARATH	O22049	<u>Alternative</u> oxidase 2, mitochondrial precursor	EC 1.-.-.-
COOH-RHORU	P31895	<u>Carbon monoxide-induced</u> hydrogenase	
CDCLZ-ECOLI	P52095	Lysine decarboxylase, <u>constitutive</u>	EC 4.1.1.18

Some myths about EC numbers

- every EC number appearing in the literature is approved by NC-IUBMB
- every EC number is a unique and stable identifier for an enzymatic reaction
- EC numbers always correspond to reactions taking place under physiological conditions
- EC numbers can be predicted by similarity with known enzymes
- EC stands for European Community

I am sure that many participants of this workshop can deliver more fascinating examples of the 'EC number mythology', together with an exhaustive critique on each myth. However, my point is that we cannot be content with the current situation and should not mostly blame a user for misunderstanding what an EC number is. If the Enzyme Nomenclature assigns EC numbers to *proteins*, we should not really be surprised by bioinformaticists' claims such as "for functional annotation, 40% sequence identity can still be used as a confident threshold to transfer the first three digits of an EC number" [20], or that "the enzyme code similarity" can be used to increase the accuracy of protein fold recognition [21].

To clarify the meanings of the enzyme names and their usage in biological databases, I am going to revisit the three general principles of the Enzyme Nomenclature which can be summarized as follows [22]:

1. The names of enzymes, especially those ending in *-ase*, should be used only for single enzymes (single catalytic entities).
 2. The enzymes are principally classified and named according to the reaction they catalyse.
 3. The enzymes are divided into groups on the basis of the type of reaction catalysed, and this, together with the name(s) of the substrate(s) provides a basis for naming individual enzymes. It is also the basis for classification and code numbers.
-

Enzyme-polypeptide relationships

"The first *general principle* ... is that names purporting to be names of enzymes, especially those ending in *-ase*, should be used only for single enzymes, i.e. single catalytic entities. They should not be applied to systems containing more than one enzyme" [22]. I dwell on three problems with realization of this principle. The first is that there is no clear definition of what constitutes a "single enzyme". Let us come back to our example of nitrogenase. Component II of nitrogenase (an iron-sulfur protein, also known as "nitrogenase reductase") dissociates from component I (whether it is a molybdenum-iron, vanadium-iron or iron-iron protein) several times during its reaction cycle, but it reacts only with component I. The nitrogenase experts agree that it would be meaningless to classify the two proteins as separate enzymes: although the proteins do dissociate, they are part of a single enzyme (Richard Cammack, personal communication). The second problem is that there are no authoritative recommendations on naming systems apart from the notion that "the word *system* should be included in the name".

Finally, what is classified today as a single enzyme may be found to be a system tomorrow, and vice versa. For example, several P450 mono-oxygenases were classified as members of EC 1.14.13 sub-subclass ("With NADH or NADPH as one donor, and incorporation of one atom of oxygen"). Their corresponding reactions uniformly contain NADPH, while it is known that apart from P450_{nor} (nitric oxide reductase), P450s do not interact directly with NADPH. Eukaryotic P450s are reduced by either a flavoprotein (NADPH-cytochrome P450 reductase; EC 1.6.2.4), or an iron-sulphur protein (adrenodoxin), which in turn is reduced by a flavoprotein (NADPH-adrenodoxin reductase) [23]. Therefore, if one has to follow Principle 1, these enzymes should be re-classified as soon as there is enough evidence about their cognate reductases. The question is, why not assign an identifier to the enzyme system if there is a good reason to do so, as is done already for a number of 'systems'? For instance, fatty-acid synthase (EC 2.3.1.85) has a comment: "The animal enzyme is a multi-functional protein catalysing the reactions of EC 2.3.1.38 [acyl-carrier-protein] *S*-acetyltransferase, EC 2.3.1.39 [acyl-carrier-protein] *S*-malonyltransferase, EC 2.3.1.41 3-oxoacyl-[acyl-carrier-protein] synthase, EC 1.1.1.100 3-oxoacyl-[acyl-carrier-protein] reductase, EC 4.2.1.61 3-hydroxypalmitoyl-[acyl-carrier-protein] dehydratase, EC 1.3.1.10 enoyl-[acyl-carrier-protein] reductase (NADPH, B-specific) and EC 3.1.2.14 oleoyl-[acyl-carrier-protein] hydrolase".

The reaction catalysed by the system may be found also to be catalysed by a single fusion protein. The well studied example is P450 BM-3 from *Bacillus megaterium*, which is a fusion of P450 mono-oxygenase haem domain with NADPH-cytochrome P450 reductase (CPR). In plants and animals, P450 enzymes and CPR exist as separate proteins. Owing to its domain arrangement, P450 BM-3 has the highest catalytic activity of any known P450 mono-oxygenase system [24].

The *Nomenclature for multi-enzymes* [25] gives recommendations for proteins "possessing more than one catalytic function contributed by distinct parts of a polypeptide chain ('domains'), or by distinct subunits, or both." The concept of 'catalytic entity' of Principle 1 is thus extended towards domains and subunits. These recommendations are adopted in Swiss-Prot for the description of multi-enzymes. The description line of the previously mentioned P450 BM-3 (SW:P14779) contains: "[Includes: Cytochrome P450 102 (EC 1.14.14.1); NADPH-cytochrome P450 reductase (EC 1.6.2.4)]".

It is less clear how to deal with a novel 'monodomain' enzyme, which corresponds to a domain in a previously known enzyme. For instance, animal nitric oxide synthase (NOS; EC 1.14.13.39) includes an N-terminal haem-containing oxygenase domain and a C-terminal reductase domain, homologous to CPR. Bacterial nitric oxide synthase oxygenase protein (SW:O34453) is homologous to the NOS oxygenase domain and is able to synthesize nitric oxide upon reduction; however, its physiological reductase remains to be discovered. Table 1 summarizes the possible scenarios of enzyme-polypeptide relationships.

Table 1. Enzyme-polypeptide relationships.

Enzyme	Polypeptide	Protein	Example
1	1	Monomeric monofunctional enzyme	SW:P50578 (EC 1.1.1.2)
1	>1	Multisubunit enzyme	PDB:1H4J (EC 1.1.99.8)
>1	1	Multienzyme polypeptide	SW:Q9DCL9 (EC 6.3.2.6; EC 4.1.1.21)
>1	1	Polypeptide	SW:P08291 (EC 3.4.22.29; EC 3.4.22.28; EC 2.7.7.48)
>1	>1	Multienzyme complex	SW:Q04728 (EC 2.3.1.35 ; EC 2.3.1.1) SW:P40939 (EC 4.2.1.17; EC 1.1.1.35) and SW:P55084 (EC 2.3.1.16)

I) 1 enzyme: 1 polypeptide

The famous "one gene-one enzyme" hypothesis of Beadle and Tatum [26] is still very much with us, being found in about every biochemistry textbook around, even though neither "one gene-one polypeptide" nor "one polypeptide-one enzyme" models are always true. How much easier would life be for database authors if one enzyme always did correspond to one polypeptide! Many monomeric enzymes apparently have evolved from oligomeric enzymes via gene duplication/fusion events [27-29].

II) 1 enzyme: >1 polypeptide

This is the case of a multisubunit (oligomeric) enzyme. The difference between a monomeric enzyme and a subunit of an oligomeric enzyme is that the latter is not active on its own. Therefore, one can argue that an EC number should not be assigned to any of the subunits (as is customarily done in sequence databases) but only to a functional complex. The counter-argument is that to do this would be not helpful since in the protein sequence databases, as a rule, one entry corresponds to one polypeptide.

*III) >1 enzyme: 1 polypeptide**a) Multi-enzyme polypeptide*

A multi-enzyme polypeptide contains several domains associated with separate enzymatic activities (see EC 2.3.1.85 example above).

b) Polyprotein

In this case, one polypeptide is a precursor of more than one mature protein. Many viral polyproteins give rise to several separate enzymes. For example, POL polyprotein from human immunodeficiency virus type 1 is a precursor of HIV-1 retropepsin (EC 3.4.23.16), RNA-directed DNA polymerase (EC 2.7.7.49) and ribonuclease H (EC 3.1.26.4); all three enzymes were structurally characterized (SW:P03366).

IV) >1 enzyme: >1 polypeptide

Here, each subunit has one or more separate enzymatic activities. For instance, yeast fatty acid synthase (EC 2.3.1.86) is a multi-enzyme complex consisting of two types of multifunctional subunits organized as an $\alpha_6\beta_6$ hetero-oligomer (SW:P19097; SW:P07149).

Some other examples do not fit easily into the above four scenarios. The arginine biosynthesis bifunctional protein ARG7 (SW:Q04728) is a heterodimer of a large and a small subunit that are proteolytically processed from a single polypeptide precursor within the mitochondrion. In the case of mitochondrial fatty acid β -oxidation trifunctional protein, the α -subunit is bifunctional (EC 4.2.1.17 and EC 1.1.1.35) while the β -subunit is monofunctional (EC 2.3.1.16) [30].

Enzyme-reaction relationships

Let us now consider enzyme-reaction relationships (Table 2). Three possible scenarios are:

Table 2. Enzyme-reaction relationships.

Enzyme	Reaction	Protein	Example
1	1	Monofunctional enzyme	
1	>1	Promiscuous enzyme	EC 1.4.3.19
1	>1	Enzyme catalyses sequential reactions	EC 3.5.3.21
>1	1	Enzymes are assigned different EC numbers because of different mechanism or by "historical reasons"	EC 1.4.3.4 and EC 1.4.3.6 EC 3.4.16.4 and EC 3.4.17.14

I) 1 enzyme: 1 reaction

This is the 'typical' case, which is in accord with Principle 2 of Enzyme Nomenclature which states that "enzymes are principally classified and named according to the reaction they catalyse". In a way, this principle is responsible for confusion arising time and again when the enzyme (i.e. protein) is equated with an EC number.

II) 1 enzyme: >1 reaction

a) Promiscuous enzyme

One enzyme may catalyse more than one reaction. Example 5 shows the entry for glycine oxidase (EC 1.4.3.19) containing four alternative reactions, only one of which involves glycine. Cf. Rule 15 of [22]: "When an enzyme catalyses more than one type of reaction, the name should normally refer to one reaction only."

Example 5. Alternative reactions.

EC 1.4.3.19	
Common name:	glycine oxidase
Reaction:	(1) glycine + H ₂ O + O ₂ = glyoxylate + NH ₃ + H ₂ O ₂ (2) D-alanine + H ₂ O + O ₂ = pyruvate + NH ₃ + H ₂ O ₂ (3) sarcosine + H ₂ O + O ₂ = glyoxylate + methylamine + H ₂ O ₂ (4) N-ethylglycine + H ₂ O + O ₂ = glyoxylate + ethylamine + H ₂ O ₂
Systematic name:	glycine:oxygen oxidoreductase (deaminating)
Comments:	A flavoenzyme containing non-covalently bound FAD. The enzyme from <i>Bacillus subtilis</i> is active with glycine, sarcosine, N-ethylglycine, D-alanine, D-α-aminobutyrate, D-proline, D-pipecolate and N-methyl-D-alanine. It differs from EC 1.4.3.3, D-amino acid oxidase, due to its activity on sarcosine and D-pipecolate.

b) Sequential reactions

Example 6 shows the entry for methylenediurea deaminase (EC 3.5.3.21) containing three consecutive reactions; only reaction 1 is enzymatic while reactions 2 and 3 are described as 'spontaneous' (while the meaning is 'non-catalytic').

Example 6. Sequential reactions.**EC 3.5.3.21**

Common name:	mehtylenediurea deaminase
Reaction:	<p>(1) $\text{NH}_2\text{-CO-NH-CH}_2\text{-NH-CO-NH}_2 + \text{H}_2\text{O} =$ $\text{N-(carboxyaminomethyl)urea} + \text{NH}_3$</p> <p>(2) $\text{N-(carboxyaminomethyl)urea} = \text{N-(aminomethyl)urea} + \text{CO}_2$ (spontaneous)</p> <p>(3) $\text{N-(aminomethyl)urea} + \text{H}_2\text{O} = \text{N-(hydroxymethyl)urea} + \text{NH}_3$ (spontaneous)</p>
Systematic name:	methylenediurea aminohydrolase
Comments:	The methylenediurea is hydrolysed and decarboxylated to give an aminated methylurea, which then spontaneously hydrolyses to hydroxymethylurea. The enzyme from <i>Ochrobactrum anthropi</i> also hydrolyses dimethylenetriurea and trimethylenetetraurea as well as ureidoglycolate, which is hydrolysed to urea and glyoxylate, and allantoate, which is hydrolysed to ureidoglycolate, ammonia and carbon dioxide.

Cf. "Special problems attend the classification and naming of enzymes catalysing complicated transformations that can be resolved into several sequential or coupled intermediary reactions of different types, all catalysed by a single enzyme (not an enzyme system). Some of the steps may be spontaneous non-catalytic reactions, while one or more intermediate steps depend on catalysis by the enzyme" [22].

III) >1 enzyme: 1 reaction

Rule 16 of [22] states: "A group of enzymes with closely similar specificities should normally be described by a single entry. <...> Separate entries are also appropriate for enzymes having similar catalytic functions, but known to differ basically with regard to reaction mechanism or to the nature of the catalytic groups, *e.g. amine oxidase (flavin-containing)* (EC 1.4.3.4) and *amine oxidase (copper-containing)* (EC 1.4.3.6)." It is clear that Rule 16 is but a built-in violation of Principle 2, for the reaction mechanism does not change the overall reaction.

Enzyme classification

"A *third general principle* adopted is that the enzymes are divided into groups on the basis of the type of reaction catalysed, and this, together with the name(s) of the substrate(s) provides a basis for naming individual enzymes. It is also the basis for classification and code numbers."

Somewhat confusingly, Principle 3 in part repeats Principle 2, viz. "enzymes are principally classified and named according to the reaction they catalyse." However, if Principle 2 is mostly relevant to the naming of the enzyme, it is Principle 3 which is responsible for the current classification of enzymes.

Principle 3 allows one and only one way of classification. (Any one EC number belongs to one and only one sub-subclass, which belongs to one and only one subclass, which belongs to one and only one class.) Since an enzyme cannot be simultaneously a member of two different classes (subclasses, sub-subclasses), many enzymes keep being renamed over and over again [7] (Example 7).

Example 7. The history line of EC 3.4.21.103.

```
[EC 3.4.21.103 created 1992 as EC 3.4.23.27 (EC 3.4.23.6 created 1992
(EC 3.4.23.6 created 1961 as EC 3.4.4.17, transferred 1972 to EC
3.4.23.6, modified 1981 [EC 3.4.23.7, EC 3.4.23.8, EC 3.4.23.9, EC
3.4.23.10, EC 3.4.99.1, EC 3.4.99.15 and EC 3.4.99.25 all created 1972
and incorporated 1978], part incorporated 1992), transferred 2003 to EC
3.4.21.103]
```

Historically, the EC number served as both unique ID and descriptor of the enzyme place within the hierarchy. For example, in oxidoreductases (EC 1), the subclass indicates the type of donor (e.g. EC 1.1 "acting on the CH-OH group of donors") and the sub-subclass indicates the type of acceptor (e.g. EC 1.1.1 "with NAD⁺ or NADP⁺ as acceptor"). Of course, the placement of the donor in the second position and the acceptor in the third position is completely arbitrary. (For a reverse reaction they should swap places.) This dual function of EC numbers is fairly limiting because it requires the enzyme to have a *unique* place within the hierarchy. However, an enzyme can be *correctly* classified in more than one way. For example, intramolecular oxidoreductases (EC 5.3) are as much oxidoreductases (EC 1) as isomerases (EC 5). In every subclass of oxidoreductases, the acceptors form repeating series, and therefore the alternative grouping is feasible, e.g. EC 1.1.1, EC 1.2.1, ..., EC 1.18.1 can be classified in an 'EC 1.x.1' subclass of 'oxidoreductases with NAD⁺ or NADP⁺ as acceptor'.

The systematic names, in general, are created according to the same conventions as EC numbers. However, not every enzyme entry has a systematic name.

Even within one sub-subclass, some additional hierarchical *IsA* relationships can be suggested on the basis of a natural hierarchy of chemical compound classes.

For instance, 3-hydroxybenzyl-alcohol dehydrogenase reaction (EC 1.1.1.97) can be considered to be a kind of aryl-alcohol dehydrogenase (NADP^+) reaction (EC 1.1.1.91) that, in turn, can be considered as a kind of alcohol dehydrogenase (NADP^+) reaction (EC 1.1.1.2).

```
%EC 1.1.1.2 alcohol dehydrogenase (NADP+)
```

```
%EC 1.1.1.91 aryl-alcohol dehydrogenase (NADP+)
```

```
%EC 1.1.1.97 3-hydroxybenzyl-alcohol dehydrogenase
```

Another problem of enzyme classification concerns the classes of overall transformations themselves. As Table 3 shows, the six classes of enzyme-catalysed reactions are mostly based on fundamental overall transformations of organic chemistry [31]. On the one hand, all EC 3 (Hydrolases), many EC 1 (Oxidoreductases) and some EC 4 (Lyases) can be considered as EC 2 (Transferases). EC 6 (Ligases) also can be considered as a kind of EC 3 (Hydrolases) because all ligase reactions involve hydrolysis of a diphosphate bond in ATP or other triphosphate. On the other hand, there are no classes for some fundamental reaction types, e.g. addition not to double bonds.

Controlled Vocabularies and Ontologies in Enzymology

Table 3. Overall transformation classes in enzymology and organic chemistry.

Enzyme classes	Transformations in organic chemistry from [31]
EC 1, Oxidoreductases $AH_2 + B \rightleftharpoons A + BH_2$ $AH_2 + B^+ \rightleftharpoons A + BH + H^+$ EC 2, Transferases $A-X + B-H \rightleftharpoons A-H + B-X$ EC 3, Hydrolases $A-B + HOH \rightarrow A-H + B-OH$ EC 6, Ligases $A + B + XTP \rightarrow A-B + XDP + P_i$ $A + B + XTP \rightarrow A-B + XMP + PP_i$	Substitution $A-X + B-Y \rightarrow A-Y + B-X$ σ -bound atom or group is replaced by another σ -bound atom or group
EC 4 (synthases, or reverse lyases) $X=Y + Z \rightarrow X-Y-Z$	Addition $A + B \rightarrow A-B$ usually, one π bond in A replaced by two new σ bonds
EC 4, Lyases $X-Y-Z \rightarrow X=Y + Z$	Elimination $A-B \rightarrow A + B$ usually, two σ bonds in A-B replaced by a new π bond
EC 5, Isomerases $A \rightleftharpoons B$	Rearrangement $A \rightarrow B$

Example 8. Misclassified enzymes?

Enzyme	Reaction
EC 4.99 Other Lyases	
EC 4.99.1.1	
Ferrochelatase	protoporphyrin + Fe^{2+} = protoheme + $2 H^+$
EC 4.99.1.2	
alkylmercury lyase	$RHg^+ + H^+ = RH + Hg^{2+}$
EC 6.6.1 Forming coordination complexes	
Magnesium chelatase	$ATP + \text{protoporphyrin IX} + Mg^{2+} + H_2O =$ $ADP + \text{phosphate} + \text{Mg-protoporphyrin IX} + 2 H^+$

Some enzymes are classified together simply because there is no suitable (sub)subclass. Example 8 shows that EC 4.99 (Other Lyases) hosts EC 4.99.1.1 (ferrochelatase) and EC 4.99.1.2 (alkylmercury lyase), although neither enzyme fits the definition of lyases (Keith Tipton, personal communication). Ferrochelatase is involved in the biosynthesis of a coordination compound (*creates* metal-N bond) while alkylmercury lyase catalyses breakdown of an organometallic compound (*breaks* metal-C bond). On the other hand, EC 6.6.1.1 (magnesium chelatase) couples ATP hydrolysis (1) with creation of a metal-N bond (2):



The use of partial reactions provides a logical solution for building a classification of enzymatic reactions. In this particular example, both EC 4.99.1.1 and EC 6.6.1.1 should be classified as metallochelataes (create metal-N bond), while magnesium chelatase also belongs to the ATP hydrolases.

Ligases exemplify one remarkable feature of enzymes, viz. their ability to couple different reactions. Apart from EC 6, many EC 1 (e.g. $\text{NAD}^+/\text{NADP}^+$ -dependent) and EC 3 (e.g. EC 3.6.4 "Acting on acid anhydrides; involved in cellular and subcellular movement") enzymes catalyse coupled reactions. In my opinion, the Enzyme Nomenclature will benefit from explicit descriptions of coupled reactions. However, one cannot define coupled reactions while ignoring the reaction directionality.

In most of the Enzyme Nomenclature entries, "the direction in which the reaction occurs is not specified (i.e. an equals sign is used rather than an arrow) so, even if a reaction has only been observed in the reverse direction, it is usually written in the direction that is common to the subclass to which it belongs" [7]. (In the other entries, the verbal description of the reaction often does convey the direction, e.g. EC 3.1.6.7 "Hydrolysis of the 2- and 3-sulfate groups of the polysulfates of cellulose and charonin".) Of course, this practice contrasts with what a chemist would intuitively expect, i.e. that a reaction is written in the direction it occurs. It also affects the *systematic* names, for these are "derived from a written reaction, even though only the reverse of this has been actually demonstrated experimentally" [22].

Controlled Vocabularies and Ontologies in Enzymology

As *The Hitch Hiker's Guide to the Galaxy* says, "it's always reality that's got it wrong" [32]. But it is also in contrast with the higher order Enzyme Nomenclature itself. Three class names (EC 3 Hydrolases, EC 4 Lyases and EC 6 Ligases) and numerous subclass names (e.g. EC 6.4 "Forming Carbon-Carbon Bonds") *do* imply the direction of the reaction.

This poses little problem for irreversible reactions or when the reaction can be catalysed by the same enzyme in both directions. However, it matters in cases when the opposite reactions are catalysed by different enzymes which are nevertheless given the same EC number (Example 9).

Example 9. Same EC numbers for the reverse reactions.

Enzyme	Reaction
EC 1.3.5.1	
Succinate dehydrogenase	succinate + Q \rightarrow fumarate + QH ₂
Fumarate reductase	fumarate + QH ₂ \rightarrow succinate + Q
EC 1.18.1.2	
Ferredoxin–NADP ⁺ reductase	reduced ferredoxin + NADP ⁺ \rightarrow oxidized ferredoxin + NADPH + H ⁺
NADPH–adrenodoxin reductase	oxidized ferredoxin + NADPH + H ⁺ \rightarrow reduced ferredoxin + NADP ⁺

Therefore, the three fundamental principles of the Enzyme Nomenclature deal satisfactorily only with the simplest cases (one enzyme:one polypeptide; one enzyme:one reaction; one enzyme:one way of classification).

Other relationships between enzyme entities

EC numbers form a strict hierarchy of *IsA* relationships. However, other relationships can exist between EC numbers. The Enzyme Nomenclature includes a number of these relationships.

I have divided them into 'structural and functional', 'historical' and 'dodgy' relationships (Tables 4-6).

Table 4. Structural and functional relationships between EC entries.

I to J	Example	Comment
(reaction) I is kind of (reaction) J	EC 1.1.1.1 is kind of EC 1.1.1	<i>IsA</i> (is kind of) relationship
I is a component of J	EC 1.3.99.1 is a component of EC 1.3.5.1	<i>IsPartOf</i> relationship
I and J are involved in metabolic process K	EC 1.1.1.252 is involved with EC 4.2.1.94 in the biosynthesis of melanin in pathogenic fungi	(unspecified) metabolic relationship
product of I is a substrate of J	The substrate of EC 3.5.4.26 is the product of EC 3.5.4.25	direct metabolic interaction
I activates J	EC 3.4.24.29 activates EC 3.4.21.19	positive regulatory relationship
I inactivates J	EC 3.1.2.16 inactivates EC 4.1.3.6	negative regulatory relationship
I phosphorylates J	EC 2.7.1.110 phosphorylates and activates EC 2.7.1.109	enzyme-substrate relationship
EC I dephosphorylates EC J	EC 3.1.3.44 dephosphorylates and activates EC 6.4.1.2	enzyme-substrate relationship

Table 5. Historical relationships between EC entries.

I to J	Example	Comment
I is transferred to J	EC 1.6.4.2 transferred to EC 1.8.1.7	Entry transfer
(deleted) I reinstated as J	EC 1.1.1.249 reinstated as EC 2.5.1.46	Entry reinstatement
I is incorporated in J	EC 4.2.1.21 incorporated in EC 4.2.1.22	Entries are merged
I is part incorporated in J1 I is part incorporated in J2 ...	EC 1.1.1.182 created 1983, part incorporated in EC 1.1.1.198, EC 1.1.1.227 and EC 1.1.1.228	Entry is split

Table 6. "Dodgy" relationships between EC entries.

I to J	Example	Comment
I is not identical with J	EC 2.8.2.5 is not identical with EC 2.8.2.17	In theory, all non-deleted EC entries should be not identical
I may be identical with J	EC 1.2.3.1 may be identical with EC 1.2.3.11	If proven, one of the entries should be deleted as "identical with"
I is possibly identical with J	EC 3.9.1.1 is possibly identical with EC 3.1.3.9 or EC 3.1.3.16	If proven, one of the entries should be deleted as "identical with"
I is related to J	EC 3.1.8.2. related to EC 3.1.8.1	unspecified relationship

'Structural and functional relationships' are mostly found in the comments. There is no official or otherwise formalized way of relating the EC entries functionally. 'Historical relationships' are those derived from the history line found at the very bottom of the enzyme entry. Therefore they are rather official. 'Dodgy relationships' are found in comments in the enzyme entries. In many cases, no meaningful information can be derived.

Incomplete EC numbers

One unfortunate consequence of the dual nature of EC numbers is the continuing practice of providing *incomplete EC numbers* in literature and databases. For example, Swiss-Prot contains many entries described as enzymes that have not been assigned EC numbers by NC-IUBMB, but which are assigned a provisional class (subclass, sub-subclass), the final digit(s) being '-'. Such assignments are thought to be useful for providing some functional information about the enzymes. The problem is that incomplete EC numbers are non-identifiers. As Example 4 shows, one EC 1.-.-.- is not equal to another EC 1.-.-.- at all!

BIOREACTION ONTOLOGY

In the words of Robert Rankin, it is a tradition, or old charter, or something, that biologists divide all reactions occurring *in vivo* into two groups: 'enzymatic' and 'non-enzymatic'. Of course, this is an extreme simplification which nevertheless summarizes the view of the world according to the modern molecular biologist: whatever is not in the genome is irrelevant.

Although enzymology, not surprisingly, deals with enzymatic reactions, it would be useful to put such reactions in the context of other biological reactions.

Thus the reactions could be classified according to type of catalyst. Chemists deal with homogeneous and heterogeneous catalysis; examples of both are also abundant in living cell. Catalytic macromolecules include catalytic proteins (enzymes and abzymes) and catalytic nucleic acids (deoxyribozymes and ribozymes). Macromolecular catalysis has features of both homogeneous and heterogeneous catalysis, for, even though reactants and catalytic macromolecules usually exist in the same phase, the reactant adsorption is crucial for the reaction mechanism.

The orthogonal classification of bioreactions is by overall reaction class, as presented in Fig. 1. These reaction classes correspond to the type of bond involved (Table 7). Most of the reactions with which the Enzyme Nomenclature deals are biotransformations.

```
%biochemical reaction
  %by overall reaction
    %binding reaction
    %biotransformation rection
    %conformation change reaction
    %molecular transport reaction
    %electron transfer reaction
    %excitation-energy transfer reaction
  %by nature of catalyst
    %catalytic macromolecule reaction
      %enzymatic reaction
      %abzymatic reaction
      %deoxyribozymatic reaction
      %ribozymatic reaction
    %heterogenous catalytic reaction
    %homogenous catalytic reaction
  %non-catalytic reaction
    %photoinduced reaction
    %thermal reaction ('spontaneous' reaction)
  %intramolecular catalysis reaction
```

Figure 1. Ontology of biochemical reactions in GO file format.

Table 7. Fundamental bioreaction classes.

Bioreaction class	Reaction	Bond involved
Biotransformation	$A + B \rightarrow C + D$	Covalent
Binding	$A + M \rightarrow AM$ M = macromolecule	Ionic, hydrogen, van der Waals
Conformation change	$A \rightarrow B$ A, B = conformers	Hydrogen, van der Waals
Molecular transport	$A_{\text{compartment}} \rightarrow A_{\text{compartment Y}}$	—
Electron transfer	$A^n + B^{n-m} \rightarrow A^{n-m} + B^n$ m = number of transferred electrons	—
Exciton transfer reactions	$A^* + B \rightarrow A + B^*$	—

Note that these reactions, like those of the Enzyme Nomenclature, are overall transformations. All enzymatic reactions include binding and conformational change as intermediate steps. Many enzymatic reactions are coupled, e.g. active transport is coupled with ATP hydrolysis. Protein folding is mostly conformational change but also may involve covalent and ionic bonds.

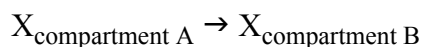
Electron transfer and excitation-energy transfer reactions

The term "pure electron transferase" was originally introduced as a name for a class of flavoproteins (exemplified by flavodoxins) where the flavin is reduced and reoxidized in one-electron steps [33, 34]. The meaning can be naturally extended to cover all the proteins that catalyse electron transfer reactions only, such as cytochromes, ferredoxins and cupredoxins. Although proteins involved in electron transfer are usually classified as oxidoreductases, none of the 'pure electron transferases' is assigned an EC number. Similarly, excitation-energy transfer processes, as in the antenna systems of photosynthetic organisms, are not covered by the Enzyme Nomenclature.

Analogous to 'traditionally understood' metabolic pathways that consist of separate enzymatic reactions, electron/exciton transfer reactions form electron/exciton transfer pathways, that form an integral part of the metabolic pathways.

Transmembrane Transport

A great number of fundamental biochemical reactions can be represented as:



Since solutes cannot cross biological membranes, the transmembrane transport has to be facilitated by specific carriers or transporters [35, 36]. Importantly, a distinct class (Energases) has been proposed to cover the enzymes that catalyse conversion of chemical energy into mechanical energy [37]. Energases include primary active transporters (directly utilizing covalent bond energy to transport solutes against a concentration gradient) and rotational molecular motors such as ATP synthase. However, there is no reason to deny the other transporters their place in the enzyme classification. Some electron transferases are also transmembrane transporters (class TC 5 in [35]).

Non-catalytic reactions

Examples of biologically relevant non-catalytic reactions include chain reactions of lipid peroxidation [38] and photoinduced transformation of ergosterol to previtamin D3 and its subsequent thermal isomerisation to vitamin D3 [39]. The non-catalytic thermal reactions often are referred to as "spontaneous" in the biological literature, including the Enzyme Nomenclature (in fact, every process with a negative change of Gibbs free energy at constant pressure is spontaneous, whether it is catalysed or not). Chain reactions of lipid oxidation can be initiated photochemically (UV radiation) or catalytically, e.g. via a Fenton mechanism, and can be terminated non-catalytically or catalytically [38].

Yet other reactions

Some biochemical reactions do not easily fit into 'catalytic' or 'non-catalytic' categories. For example, the comment for methylated-DNA-[protein]-cysteine *S*-methyltransferase (EC 2.1.1.63) reads: "Since the acceptor protein is the 'enzyme' itself and the *S*-methyl-L-cysteine derivative formed is relatively stable, the reaction is not catalytic." The reaction proceeds through suicidal alkyl transfer from the guanine O6 to the cysteine residue of the enzyme; therefore the enzyme should be present in stoichiometric, not catalytic, amounts. The reaction catalysed by EC 2.1.1.63 fits the IUPAC "Gold Book" definition of intramolecular catalysis [40].

On the one hand, the intramolecular catalyst is a kind of catalyst, since "the catalyst is both a reactant and product of the reaction" [41]. But if the direct result of the reaction is an inactivated catalyst, it makes the whole process non-catalytic (according to the Gold Book).

The term "autocatalytic reaction" is often used in a meaning not consistent with the Gold Book definition [42], e.g., "*autocatalytic quinone-methide mechanism of protein flavinylation*" [43] or "*autocatalytic formation of a thioether cross-link between the active-site residues*" in galactose oxidase [44]. These are in fact intramolecular catalysis events.

Mechanism of biochemical reaction

At least two orthogonal types of mechanism can be considered: one is inherent to the reaction or some reaction steps, another is specified by the catalyst. The 'inherent' mechanisms, such as polar reactions, free-radical reactions and pericyclic reactions [31], are shared with organic chemistry. The catalyst-specific mechanisms often feature particular metals or prosthetic groups (e.g. "copper-dependent" or "pterin-dependent" enzymes) or kinetic properties (e.g. "Ping Pong mechanism") [45]. Note that coenzymes, in contrast to prosthetic groups, do enter the overall transformations and therefore are covered by reaction classifications.

ONTOLOGY FOR ENZYME-CATALYSED REACTION: SOME WORKING PRINCIPLES

Organization: Similar to GO, terms should be organized as a directed acyclic graph (DAG). A child term can have many parent terms.

A unique identifier in form of accession number (AC) should be assigned to every node of DAG so we can be sure that we find data with AC even if the term is changed. When entries are merged, ACs do not disappear. Unlike EC numbers, ACs are devoid of any other meaning.

Traceability: Evidence should be assigned to every node (and possibly to every edge) of a DAG. Such evidence could be a literature reference, the database entry, or a curator judgement. The history line of an IntEnz entry in Example 9 provides some but not all of the information about an enzyme's history (it is important to know not only when, but also why and how the entries were merged, split or modified in any other way). Modern tools allow all events of this sort to be tracked (provided, of course, that editing is done within a database and not at an external source).

Scalability: "An architecture is considered to be scalable if, unchanged, it can handle increasingly complex problems that demand a greater amount of knowledge" [46]. The enzyme ontology should be designed in such a way that it can cope with the growing amount of data. Since our knowledge is always incomplete, the system should be adaptable for higher/lower levels of granularity.

It seems obvious that if the reactions are classified by overall transformation, considerations such as the nature of the catalyst should not be used; yet we saw exactly this in the Enzyme Nomenclature. So, the next principle is that the *orthogonal classifications* should not mix. The top level of the ontology of enzymatic reactions in a 'GO file format' is shown in Fig. 2.

```
%enzymatic reaction
  %by overall reaction (aka Enzyme Nomenclature)
  %by partial reaction
  %by nature of enzyme
  %by reaction mechanism
  %by enzyme regulation
    %activation
    %inhibition
```

Figure 2. Ontology of enzymatic reactions in GO file format

An important step towards development of a bioreaction ontology is made by introduction of the REACTION database, which is developed as an integral part of the LIGAND database [47]. The reactions are completely independent from the ENZYME entries. Therefore, it is possible to link one ENZYME entry to more than one REACTION entry as well as to include non-enzymatic reactions and, perhaps even more importantly, the reactions catalysed by yet unclassified enzymes. Currently, only 'parseable' reactions (i.e. those expressed by an equation) are included. REACTION does not take account of reaction directionality, the reactions being written as if they were reversible.

If the directed partial reactions are assigned their own ACs, the coupled reactions could be constructed as linear combinations of partial reactions; these new reactions are assigned their own ACs, etc.

Finally, in creating the ontology of enzymatic reactions, we *should avoid preconceived opinions regarding experimental conditions, efficiency or physiological role*. In many cases, the physiological role cannot be elucidated before we know enough about the physiology of the organism in question; however this should not prevent us from describing the enzyme function.

NEW RELATIONSHIPS

Let us have a look at Fig. 3. The universes of biochemical objects and of biochemical processes do not overlap. Most biochemical objects relate to each other via familiar *IsA* and *IsPartOf* relationships. The same is true for most biochemical processes. However, for some objects and processes novel relationships (e.g. *IsTransformedTo*, *Accelerates*, *SlowsDown*) have to be introduced. Since the reaction (process) uniquely specifies the relationship between 'reactant' and 'product', the logical way to link the two universes is via the overall reaction. Indeed, 'reactant', 'product' and 'solvent' are functional roles of chemical compounds in a particular reaction (in our example, $A \rightarrow B + C$ takes place in solvent D) and may play different roles elsewhere. All the relationships are directed, *including reactions*. As our knowledge of enzyme structure and function grows, further links relating active site (object) with catalytic mechanism (process), compound (object) with inhibition or activation (process), etc., may be required.

CONCLUSION

Although the Enzyme Nomenclature is the only accepted systematic nomenclature for biochemical reactions, it often violates its own fundamental principles. Moreover, it suffers from inherent confusion between enzyme (object) and enzymatic reaction (process) concepts that propagate into the scientific literature and biological databases. For instance, EC numbers are often used as synonyms of protein names. This strict hierarchy and depth limit may have been justified for a book layout but is otherwise too rigid. The limit of four levels does not allow additional hierarchical *IsA* relationships which otherwise can be introduced on the basis of a natural hierarchy of chemical compound classes. The extension of Enzyme Nomenclature beyond the traditional six classes of overall transformations will include, e.g., reactions affecting non-covalent bonds and transport phenomena.

Further modification of the Enzyme Nomenclature is required to enable multiple ancestry for enzymatic reactions. Ultimately, the Enzyme Nomenclature, cleansed of inconsistencies, should become a part of the new Enzyme Ontology.

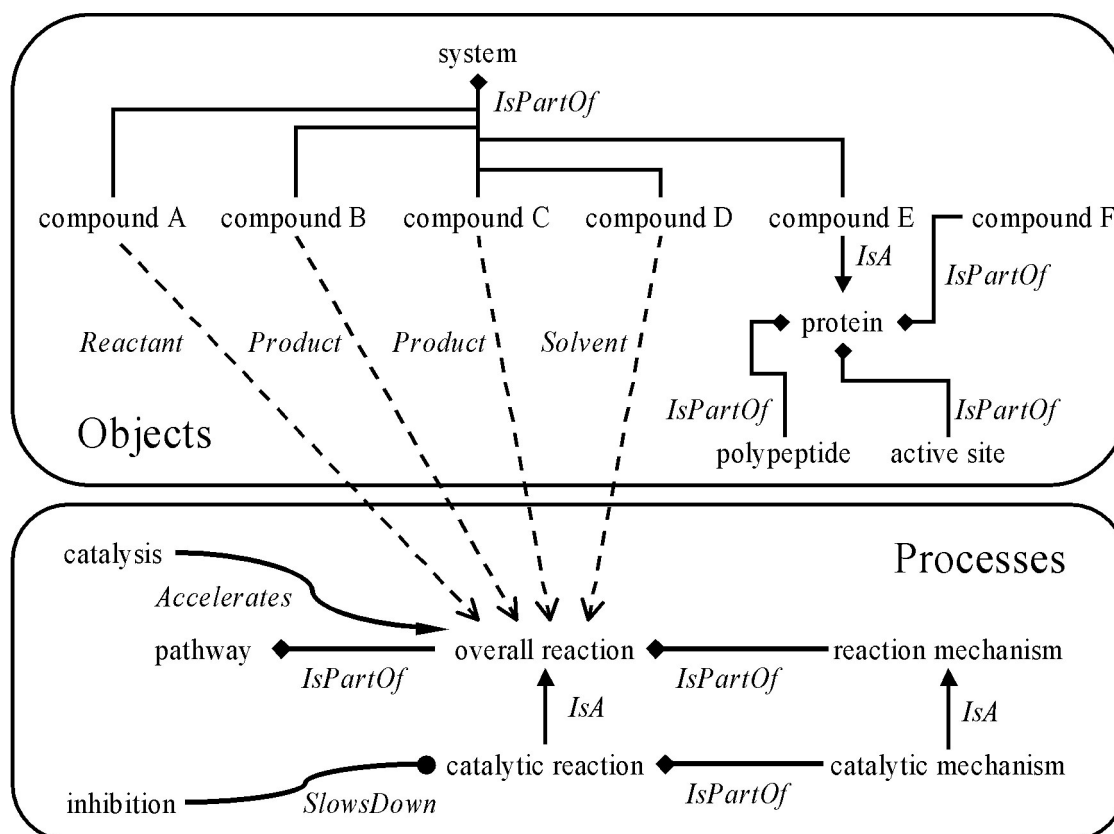


Figure 3. Some relationships between biochemical objects and biochemical processes.

ACKNOWLEDGMENTS

I wish to thank Rolf Apweiler, who suggested that I should submit an abstract for this workshop, and Carsten Kettner, who read through it and asked me to give this presentation. I thank my colleagues at the EBI, Evelyn Camon, Marcus Ennis and Jane Lomax, for their helpful comments and suggestions on the manuscript. This work is supported by the European Commission grant QLRT-2000-00981.

LIST OF ABBREVIATIONS

AC	accession number
CPR	NADPH-cytochrome P450 reductase
DAG	directed acyclic graph
EC	Enzyme Commission
GO	Gene Ontology
ID	identifier
NC-IUB	Nomenclature Committee of the International Union of Biochemistry
NC-IUBMB	Nomenclature Committee of the International Union of Biochemistry and Molecular Biology
NOS	nitric oxide synthase
PDB	Protein Data Bank

REFERENCES

- [1] Carugo, O., Pongor, S. (2002) The evolution of structural databases. *Trends Biotechnol.* **20**: 498-501.
 - [2] The Gene Ontology Consortium, <http://www.geneontology.org/>
 - [3] The EMBL Nucleotide Sequence Database, <http://www.ebi.ac.uk/embl/>
 - [4] The Swiss-Prot Protein Knowledgebase, <http://www.ebi.ac.uk/swissprot/>
 - [5] The Protein Data Bank, <http://www.pdb.org/>
 - [6] IUBMB (1992) *Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Academic Press, San Diego.
 - [7] Tipton, K., Boyce, S. (2000) History of the enzyme nomenclature system. *Bioinformatics* **16**: 34-40.
 - [8] Lilley, D.M.J. (2003) The origins of RNA catalysis in ribozymes. *Trends Biochem. Sci.* **28**: 495-501.
 - [9] Emilsson, G.M., Breaker, R.R. (2002) Deoxyribozymes: new activities and new applications. *Cell. Mol. Life Sci.* **59**: 596-607.
 - [10] Nevinsky, G.A., Buneva, V.N. (2003) Catalytic antibodies in healthy humans and patients with autoimmune and viral diseases. *J. Cell. Mol. Med.* **7**: 265-276.
 - [11] IntEnz: Integrated relational Enzyme database, <http://www.ebi.ac.uk/IntEnz/>
 - [12] The ENZYME database, <http://www.expasy.org/enzyme/>
 - [13] BRENDA: The Comprehensive Enzyme Information System, <http://www.brenda.uni-koeln.de/>
-

-
- [14] GO File Format Guide, <http://www.geneontology.org/doc/GO.format.html>
 - [15] Open Biology Ontologies, <http://obo.sourceforge.net/>
 - [16] MEROPS, the Peptidase Database <http://www.merops.ac.uk/>
 - [17] REBASE, The Restriction Enzyme Database, <http://rebase.neb.com/>
 - [18] CAZy, Carbohydrate-Active enZymes, <http://afmb.cnrs-mrs.fr/CAZY/>
 - [19] UM-BBD, The University of Minnesota Biocatalysis/Biodegradation Database, <http://umbbd.ahc.umn.edu/>
 - [20] Tian, W., Skolnick, J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* **333**: 863-882.
 - [21] Bindewald, E., Cestaro, A., Hesser, J., Heiler, M., Tosatto, S.C.E. (2003) MANIFOLD: protein fold recognition based on secondary structure, sequence similarity and enzyme classification. *Protein Engng* **16**: 785-789.
 - [22] NC-IUBMB (1992) Classification and Nomenclature of Enzyme-Catalysed Reactions. <http://www.chem.qmul.ac.uk/iubmb/enzyme/rules.html>
 - [23] Degtyarenko, K.N., Kulikova, T.A. (2001) Evolution of bioinorganic motifs in P450-containing systems. *Biochem. Soc. Trans.* **29**: 139-147.
 - [24] Munro, A.W., Leys, D.G., McLean, K.J., Marshall, K.R., Ost, T.W.B., Daff, S., Miles, C.S., Chapman, S.K., Lysek, D.A., Moser, C.C., Page, C.C., Dutton, P.L. (2002) P450 BM3: the very model of a modern flavocytochrome. *Trends Biochem. Sci.* **27**: 250-257.
 - [25] NC-IUB (1989) Nomenclature for multienzymes. Recommendations 1989. <http://www.chem.qmul.ac.uk/iubmb/misc/menz.html>
 - [26] Beadle, G.W., Tatum, E.L. (1941) Genetic control of biochemical reactions in *Neurospora*. *Proc. Natl. Acad. Sci. USA* **27**: 499-506.
 - [27] Fani, R., Liò, P., Lazcano, A. (1995) Molecular evolution of the histidine biosynthetic pathway. *J. Mol. Evol.* **41**: 760-774.
 - [28] Hawkins, A.R., Lamb, H.K. (1995) The molecular biology of multidomain proteins. Selected examples. *Eur. J. Biochem.* **232**: 7-18.
 - [29] Yasutake, Y., Watanabe, S., Yao, M., Takada, Y., Fukunaga, N., Tanaka, I. (2002) Structure of the monomeric isocitrate dehydrogenase: evidence of a protein monomerization by a domain duplication. *Structure* **10**: 1637-1648.
 - [30] Kamijo, T., Aoyama, T., Komiyama, A., Hashimoto, T. (1994) Structural analysis of cDNAs for subunits of human mitochondrial fatty acid β -oxidation trifunctional protein. *Biochem. Biophys. Res. Commun.* **199**: 818-825.
 - [31] Grossman, R.B. (1999) *The Art of Writing Reasonable Organic Reaction Mechanisms*. Springer-Verlag, New York.
 - [32] Adams, D. (1980) *The Restaurant at the End of the Universe*. p.35. Pan Books Ltd, London.
-

-
- [33] Hemmerich, P., Massey, V., Fenner, H. (1972) Flavin and 5-deazaflavin: A chemical evaluation of 'modified' flavoproteins with respect to the mechanisms of redox biocatalysis. *FEBS Lett.* **84**: 5-21.
 - [34] NC-IUB (1991) Nomenclature of electron-transfer proteins. Recommendations 1989. <http://www.chem.qmul.ac.uk/iubmb/etp/>
 - [35] Transport Protein Database, <http://tcdb.ucsd.edu/tcdb/>
 - [36] NC-IUBMB (2002) Membrane transport proteins. Recommendations 2002. <http://www.chem.qmul.ac.uk/iubmb/mtp/>
 - [37] Purich, D.L. (2001) Enzyme catalysis: a new definition accounting for noncovalent substrate- and product-like states. *Trends Biochem. Sci.* **26**: 417-421.
 - [38] Girotti, A.W. (1998) Lipid hydroperoxide generation, turnover, and effector action in biological systems. *J. Lipid Res.* **39**: 1529-1542.
 - [39] Holick, M.F. (1995) Defects in the synthesis and metabolism of vitamin D. *Exp. Clin. Endocrinol. Diabetes* **103**: 219-227.
 - [40] McNaught, A.D., Wilkinson, A., Eds. (1997) *Compendium of Chemical Terminology: IUPAC Recommendations ("The Gold Book")*, 2nd Edition, p.206. Blackwell Scientific Publications, Oxford.
 - [41] *Ibid.*, p. 58.
 - [42] *Ibid.*, p. 34.
 - [43] Edmondson, D.E., Newton-Vinson, P. (2001) The covalent FAD of monoamine oxidase: structural and functional role and mechanism of the flavinylation reaction. *Antioxid. Redox Signal.* **3**: 789-806.
 - [44] Firbank, S.J., Rogers, M., Hurtado-Guerrero, R., Dooley, D.M., Halcrow, M.A., Phillips, S.E.V., Knowles, P.F., McPherson, M.J. (2003) Cofactor processing in galactose oxidase. *Biochem. Soc. Trans.* **31**: 506-509.
 - [45] Silverman, R.B. (2002) *The Organic Chemistry of Enzyme-Catalyzed Reactions*. Academic Press, San Diego.
 - [46] Lemon, B., Pynadath, D., Taylor, G., Wray, R.E. Cognitive architectures: a hypertext analysis of architectures for intelligence. <http://ai.eecs.umich.edu/cogarch4/>
 - [47] LIGAND database of chemical compounds and reactions in biological pathways, <http://www.genome.ad.jp/ligand/>
-

PROFILES OF MOLECULAR FUNCTION - GENOMIC ENZYMOLOGY

JOHN L. ANDREASSI AND THOMAS S. LEYH*

Department of Biochemistry, The Albert Einstein College of Medicine, 1300 Morris Park Ave.,
Bronx, New York 10461-1926, U.S.A.

E-Mail: *leyh@acom.yu.edu

Received: 16th March 2004 / Published: 1st October 2004

ABSTRACT

A worldwide initiative, the goal of which is to place all of Nature's globular protein domains within modelling distance of a known three-dimensional structure, is underway. The tens of thousands of structures slated to be delivered to the scientific community by the Initiative over the ensuing decade will create an acute need for a complementary program to characterize the functions of these proteins. It is timely to consider the design of such a program.

INTRODUCTION

The rate at which protein structures are archived is remarkable (~ 5000/year in 2003) and continuing to increase. A primary aim of the global protein structural initiative has been to place all protein folds within modelling distance of a representative structure [1]. This endeavour is driven by a desire to understand the molecular functions of these structures. The coverage-of-protein-fold-space problem appears to be approaching a watershed from which the objectives of the Initiative are turning toward articulating the differences within a single domain type. With function in-hand, the domain architecture of a protein becomes a hierarchy of functional entities carefully positioned to accomplish a specific molecular task. As our understanding of the differences among domain-structures deepen, so does our need to understand how these differences produce changes in molecular function.

When a molecular biologist becomes interested in a specific protein, among the first pieces of information (s)he seeks are its structure, its conserved residues, and an understanding of the functions of those residues.

This information is fundamental in the sense that it is the information that the community naturally reaches for when it begins to pursue molecular explanations for biology - it is the information with which one begins to consider and to control (through mutagenesis) the molecular architecture that underlies biological function.

Often, though not exclusively, it is the conserved residues that lie at or near the surface of the protein that are of the greatest general interest. These residues, preserved through evolution, are the residues with which a protein family "senses" and interacts with its environment. Defining how these residues function establishes a signature, or *profile*, that describes the molecular operations of a given family. *This information is not generally available, and should be created for every family for which it is feasible.*

A database of family profiles would prove valuable in many areas of biology. For example, SNP (single nucleotide polymorphism) databases, currently under development for the human, mouse, fly and nematode genomes, are mined routinely to identify which mutations, from the millions that have been catalogued, are linked to disease and other interesting phenotypes. Cross-referencing disease-linked SNPs with a database of family-function profiles will associate the disease-causing mutations with specific alterations in protein function. In a second broad application, understanding how particular mutations effect function enables the geneticist to select the functional properties of the protein that will be bred into the genetic background of an organism. A database of profiles would provide an off-the-shelf resource for designing such transgenic experiments. As a third and final example, bioinformatics projects often aim at identifying the residues on which the features of a protein family pivot. Sophisticated models are being developed to scrutinize primary sequences and identify putative critical residues. A profile database will not only confirm the predictions of the models, but, importantly, will also provide data sets on which the models can be trained, tested and developed further.

Creating a database of profiles will require that the conserved residues positioned at the surface of a protein be identified (this can now be done computationally), that their functions be determined (through mutagenesis and functional characterization), and that the data be incorporated into the database. A database format that seems particularly intuitive for the user is one in which the conserved residues, presented in CPK on a rotatable structure, can be "clicked" to present a table describing the results of the functional studies on that residue (cf., Table 1, below). It would be valuable to have the table include links to the results of findings associated with a particular residue in other members of the same family.

Profiles of Molecular Function - Genomic Enzymology

Table 1. Effects of mutation on the initial-rate constants of the PMK-catalysed reaction.

Protein	K_i (P-mev) (μ M)	K_m (P-mev) (μ M)	K_i (ATP) (mM)	K_m (ATP) (mM)	k_{cat} (sec^{-1})
Wild Type	^a 22 (0.38) 1.0	2.4 (0.13) 1.0	2.85 (0.15) 1.0	0.31 (0.007) 1.0	18.3 (0.09) 1.0
D150E	39 (0.28) 1.8	2.4 (0.0085) 1.0	4.0 (0.016) 1.4	0.25 (0.0015) 0.8	1.54 (0.001) 0.084
^l S291A	1650 (172) 75	45 (1.0) 19	3.0 (0.074) 1.05	0.083 (0.009) 0.27	11.9 (0.083) 0.65
² A293T	^b ND	42,000 (5.8) 17,500	ND	4.2 (0.7) 14	0.28 (0.027) 0.015

^a A kinetic constant, and its standard error, are listed above the line; beneath it, the constant is normalized to that of the wild-type enzyme. ^bValue could not be determined.

Correlates

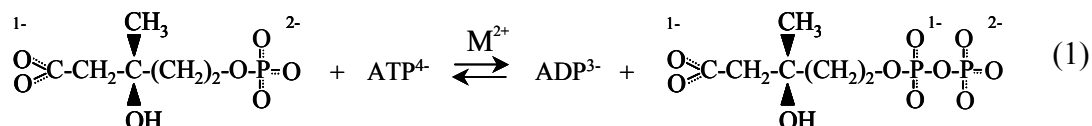
1) Mutation in human galactokinase at the position analogous to S291 in PMK causes type II galactosemia. (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>). [11]

2) Mutation in the human mavalonate kinase at the position analogous to A293 kinase causes mevalonate aciduria. (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>). [12]

RESULTS AND DISCUSSION**A case study-phosphomevalonate kinase (PMK)**

A preliminary exploration of the feasibility and efficacy of a large-scale effort to study protein function was carried out using the enzyme phosphomevalonate kinase, cloned from the R6 strain of *Streptococcus pneumoniae*. PMK was obtained from the Rockefeller University contingent of the New York Structural Genomics Research Consortium [2]. PMK was selected for structural characterization by the consortium because it is a member of the GHMP kinase family for which, at the time, there was no representative structure. The consortium uses sequence uniqueness as a criterion in selecting proteins for structure determination, which strongly biases the outcome toward novel structural folds, and directs the endeavour toward creating a diverse structural library that is representative of the folds found in Nature. As anticipated, the sequences unique to the GHMP kinase family manifest as novel structures which, in this case, are intimately involved catalysing phosphoryl-transfer [3, 4].

Phosphomevalonate kinase (ATP: 5-phosphomevalonate phosphotransferase, EC 2.7.4.2) catalyses transfer of the γ -phosphoryl group of ATP to (R)-5-phosphomevalonate, forming the pyrophosphoryl-linkage found in diphosphomevalonate (Reaction 1).



Reaction 1 is the secondstep in the so-called mevalonate pathway, which is comprised of a four-step sequence of reactions the end-product of which, isopentenyl diphosphate (IPP), provides the 5-carbon building blocks used in the biosynthesis of isoprenoids-a complex family of metabolites encompassing more than 23,000 compounds [5] including cholesterol, steroids, vitamin K₁₂, and the prenyl-moiety used to post-translationally modify and target proteins to the membranes [6]. In certain prokaryotes, IPP is linked into the linear undecaprenyl chains (C₅₅) that translocate peptidoglycans across the cell membrane during cell wall biosynthesis [5, 7].

Highly conserved, solvent-accessible GHMP kinase family residues were identified in several ways. The primary sequences of the GHMP kinase family were compared to identify both family-wide and PMK-specific conservation. These residues were mapped onto the PMK structure (PDB 1K47) to assess their surface accessibility. In an alternative strategy, GHMP kinase structures were aligned, using the *Vector Alignment Search Tool* (VAST) [8, 9], to identify surface accessible residues that are well-conserved in three-dimensional space, but not in sequence-space. These analyses identified a set of forty-seven candidate residues, twenty of which appear to be solvent accessible when mapped onto the PMK crystal structure (PDB ID 1k47) see Fig. 1. The structure suggested that nine of the twenty were integral to the structural core, and therefore removed from further consideration. The remaining eleven were characterized, and three of these are discussed here (D150E, S291A and A293T).

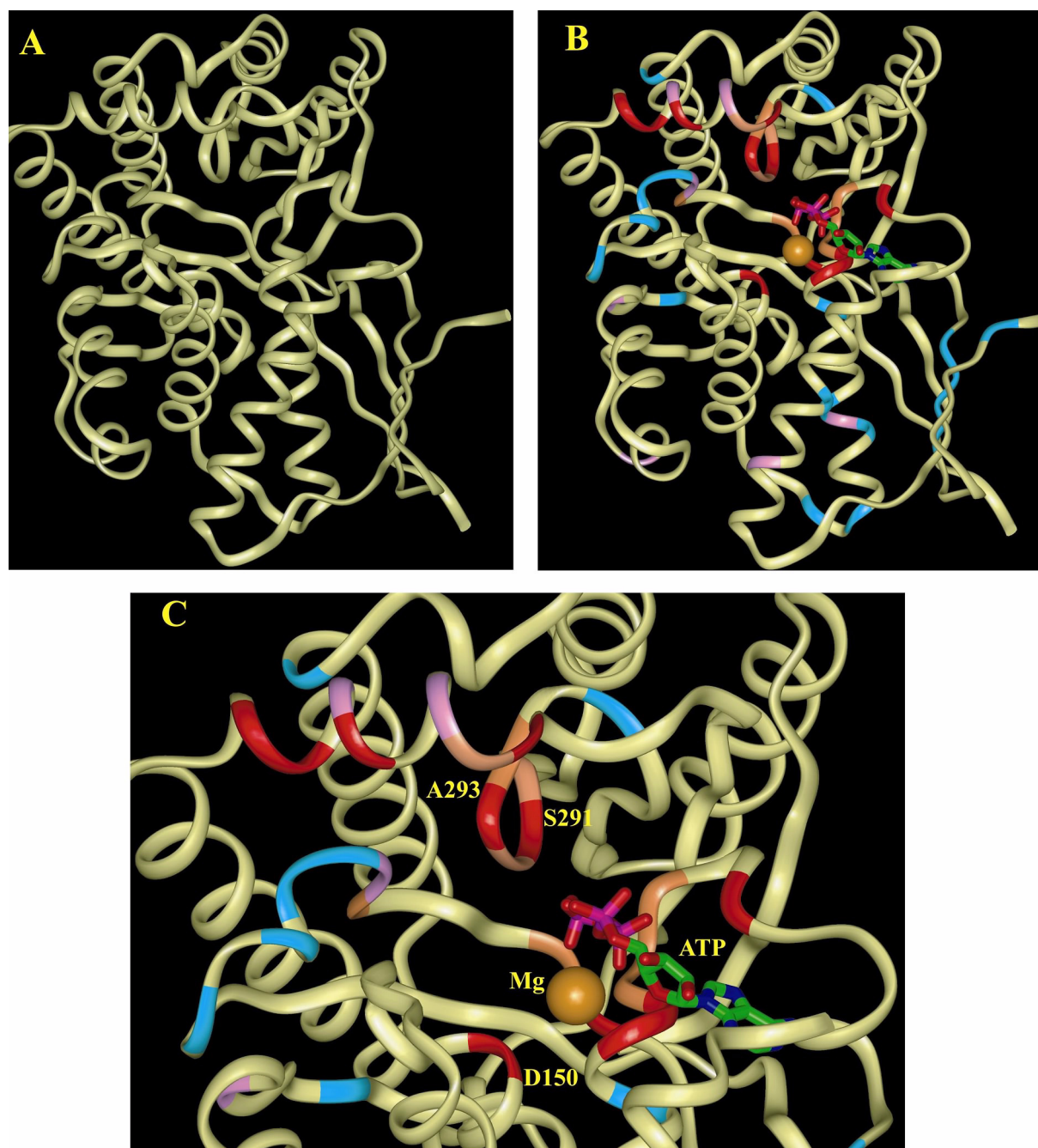


Figure 1. Conserved amino acids in PMK. (A.) A backbone representation of the structure of the wild-type, apo-PMK (PDB ID 1k47). (B.) The forty-seven, highly-conserved GHMP kinase residues are highlighted on the apo-PMK structure. Twenty of the forty-seven conserved residues are part of the hydrophobic core of the structure (Cyan), and 7 are involved in internal salt-bridge interactions. Of the remaining twenty residues, all of which were solvent accessible, nine appeared to be involved primarily in maintaining secondary structure (Orange), the remaining 11 residues (Red) were characterized. (C.) A close-up of the PMK active-site. The residues discussed in the text are labelled according to their sequence position. Mg^{2+} ·ATP was modelled into the PMK structure using the structure of the mevalonate kinase· Mg^{2+} ·ATP complex (PDB ID 1kvk) as a template. The figures were rendered using Insight II.

Aspartate 150

GHMP kinase structures support that this enzyme family uses a carboxylate-carrying residue (Asp or Glu) to catalytically position the divalent cation (Mg^{2+}) that is coordinated to the tripolyphosphate chain of ATP [10] - aspartic acid 150 in PMK. To explore whether and how this residue might participate in catalysis, a methylene group ($-\text{CH}_2-$) was inserted into the aliphatic chain from which the carboxylate "dangles" by replacing Asp150 with glutamic acid. The effects of lengthening the R-group on the kinetic parameters of the forward reaction are compiled in Table 1. The steady-state affinities of the substrates for various forms of the enzyme are influenced very little by the substitution (0.8 - 1.8-fold effects are observed). k_{cat} , however, decreases 12-fold when the chain is lengthened. Thus, the insertion selectively affects a rate-determining step(s) in the catalytic cycle - which is consistent with a repositioning of Mg^{2+} .

Serine 291

GHMP kinases active sites contain a small, glycine-rich loop that appears well positioned to interact with the substrate that receives the γ -phosphoryl group from ATP. To assess how this loop functions in catalysis, two of the loop's residues (S291 and A293) were altered, via mutagenesis, and the effects of the perturbations were evaluated. Replacing the hydroxyl group of S291 with a proton (S291A) causes $K_{\text{m(Pmev)}}$ to increase 19-fold. Clearly, the S291 hydroxyl is important in recognition of phosphomevalonate during steady-state turnover. Interestingly, the substitution enhances the steady-state affinity of $\text{Mg}^{2+}\cdot\text{ATP}$ for the enzyme ~ 5 -fold.

The γ -phosphoryl group is the closest ATP moiety to the S291 hydroxyl, and these two groups are separated by ~ 11 Å. Notwithstanding a surprising reorganization of the active-site during catalysis, it appears that the hydroxyl is involved in a communication network between the S241 side chain and the determinants involved in the steady-state recognition of ATP that spans a distance of at least ~ 11 Å. While the S291A substitution affects the steady-state affinities of both substrates, its influence on k_{cat} is small, 1.5-fold.

Position 291 is conserved across mevalonate kinases, homoserine kinases and galactokinases as either Ser or Thr. Interestingly, mutating threonine 334 (the Ser 291 homologue) in human galactokinase results in a galactokinase deficiency that causes cataract formation in children not maintained on a lactose free diet [11]. Our PMK findings suggest that the galactokinase deficiency may be caused by a decreased steady-state affinity of galactokinase for galactose.

Alanine 293

Alanine 293, also in the glycine-rich substrate-recognition loop, was mutagenized to a threonine, and the effect of this substitution on catalysis was studied. The influence on $K_{m(\text{Pmev})}$ was profound (17,000-fold increase) while the effect on $K_{m(\text{ATP})}$ was relatively small (14-fold). Thus, we again see that perturbing the loop causes highly selective effects on the substrate affinities, with the greater effect on the affinity of the non-ATP substrate. Unlike the S291A mutation, the A293T substitution resulted in a large decrease (64-fold decrease) in k_{cat} . While it is not clear whether A293 itself contributes to the energetics of rate-determining steps in the catalytic cycle, it is clear that alterations at this site can influence the relative energetics/structures of the ground- and transition-state(s) of such steps.

In addition to its conservation among PMKs, alanine 293 is also highly conserved in mevalonate kinase across species. Mutating the Ala 334 codon of the human mevalonate kinase (the PMK A293 homologue) causes a mevalonate kinase deficiency that results in elevated levels of mevalonic acid in plasma and can lead to early fatality [12-14]. A human recombinant mevalonate kinase A334T mutant exhibited a k_{cat} decrease comparable to that seen for the PMK A293T mutant [12]. The mutagenesis results support the notion that a second, highly conserved, loop in the active site of GHMP kinases functions to recognize and position the non-nucleotide substrate for catalysis.

CONCLUSION

The findings described above illustrate well the value of family profiling. The behaviour of two of the mutants, S291A and A293T, in conjunction with the PMK structure, suggests that the active sites of GHMP kinase family members exhibit a conserved, substrate-recognition loop, and that the loop participates in an allosteric network that establishes an energetic poise between the substrate binding pockets. Furthermore, these findings provide precise descriptions of the metabolic lesions that are caused by these disease-linked mutations. The remaining mutant, D150A, is remarkable in that its effects are nearly exclusively focused at the rate-determining step(s) in the reaction, and its position in the PMK structure suggests that the effects are likely to be mediated through coordination of the ATP-bound Mg^{2+} ion. D150A seems an excellent candidate for reverse genetic exploration of the effects of down-regulating flux through the mevalonate pathway in model organisms.

It is impossible to predict the myriad of ways that the information available in a *Database of Family Profiles* will be applied by molecular biologists to their individual situations. What is clear, however, is that the profiling method has worked, and that the impact of the information that it yields is considerable. There are thousand of proteins for which this work has yet to be done, and the literature suggests that it will not likely be accomplished using *status-quo* mechanisms. The creation of such a database will require a multidisciplinary program.

ACKNOWLEDGMENTS

Supported by the National Institutes of Health Grants GM54469 and the Albert Einstein College of Medicine.

REFERENCES

- [1] Burley, S.K., Bonanno, J.B. (2002) Structuring the universe of proteins. *Annu. Rev. Genomics Hum. Genet.* **3**: 243-262.
- [2] Romanowski, M.J., Bonanno, J.B., Burley, S.K. (2002) Crystal structure of the *Streptococcus pneumoniae* phosphomevalonate kinase, a member of the GHMP kinase superfamily. *Proteins* **47**(4): 568-571.
- [3] Burley, S.K., Bonanno, J.B. (2002) Structural genomics of proteins from conserved biochemical pathways and processes. *Curr. Opin. Struct. Biol.* **12**(3): 383-391.
- [4] Bork, P., Sander, C., Valencia, A. (1993) Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Sci.* **2**(1): 31-40.
- [5] Kharel, Y., Koyama, T. (2003) Molecular analysis of cis-prenyl chain elongating enzymes. *Nat. Prod. Rep.* **20**(1): 111-118.
- [6] Wang, K.C., Ohnuma, S. (2000) Isoprenyl diphosphate synthases. *Biochim. Biophys. Acta* **1529**(1-3): 33-48.
- [7] Fujihashi, M. et al. (2001) Crystal structure of cis-prenyl chain elongating enzyme, undecaprenyl diphosphate synthase. *Proc. Natl. Acad. Sci. USA* **98**(8): 4337-4342.
- [8] Gibrat, J.-F., Midej, T., Bryant, S.H. (1996) Surprising similarities in protein structure. *Curr. Opin. Struct. Biol.* **6**: 377-385.
- [9] Madej, T., Gibrat, J.-F., Bryant, S.H. (1995) Threading a database of protein cores. *Protein Struct. Funct. Genet.* **23**: 356-369.
- [10] Fu, Z. et al. (2002) The structure of a binary complex between a mammalian mevalonate kinase and ATP: insights into the reaction mechanism and human inherited disease. *J. biol. Chem.* **277**(20): 18134-18142.
- [11] Asada, M. et al. (1999) Molecular characterization of galactokinase deficiency in Japanese patients. *J. Hum. Genet.* **44**(6): 377-382.

- [12] Hinson, D.D. et al. (1997) Identification of an active site alanine in mevalonate kinase through characterization of a novel mutation in mevalonate kinase deficiency. *J. biol. Chem.* **272**(42): 26756-26760.
 - [13] Hoffmann, G. et al.(1986) Mevalonic aciduria--an inborn error of cholesterol and nonsterol isoprene biosynthesis. *New Engl. J. Med.* **314**(25): 1610-1614.
 - [14] Hoffmann, G.F et al. (1993) Clinical and biochemical phenotype in 11 patients with mevalonic aciduria. *Pediatrics* **91**(5): 915-921.
-

EXPERIMENTAL ENZYME DATA AS PRESENTED IN BRENDA - A DATABASE FOR METABOLIC RESEARCH, ENZYME TECHNOLOGY AND SYSTEMS BIOLOGY

**IDA SCHOMBURG, ANTJE CHANG, CHRISTIAN EBELING, GREGOR HUHN,
OLIVER HOFMANN, DIETMAR SCHOMBURG***

CUBIC (Cologne University Bioinformatics Centre), Institute of Biochemistry,
Köln, Germany

E-Mail: [*d.schomburg@uni-koeln.de](mailto:d.schomburg@uni-koeln.de)

Received: 15th April 2004 / Published 1st October 2004

ABSTRACT

BRENDA represents the most comprehensive information system on enzyme and metabolic information, based on primary literature. The database contains data from at least 83,000 different enzymes from 9800 different organisms, classified in approximately 4200 EC numbers. BRENDA includes biochemical and molecular information on classification and nomenclature, reaction and specificity, functional parameters, occurrence, enzyme structure, application, engineering, stability, disease, isolation, and preparation, links, and literature references. The data are extracted and evaluated from approximately 46,000 references, which are linked to PubMed as long as the reference is cited in PubMed. In the last year BRENDA underwent major changes including a large increase in updating speed with more than 50% of all data updated in 2002 or in the first half of 2003, the development of a new EC-tree browser, a taxonomy-tree browser, a chemical substructure search engine for ligand structure, the development of controlled vocabulary and an ontology for some information fields, and a thesaurus for ligand names. The database is accessible free of charge for the academic community at <http://www.brenda.uni-koeln.de>.

Analysis of the experimental data stored in BRENDA shows a number of problems that prohibit a systematic comparison and evaluation of experimental protein data. This is caused by the fact that on the one hand, many experimental data are determined in a non-systematic way and that - on the other hand - the existing recommendations on nomenclature are systematically ignored by most authors of biochemical and molecular-biological papers. Examples will be given.

INTRODUCTION

Enzymes represent the largest and most diverse group of all proteins, catalysing all chemical reactions in the metabolism of all organisms. They play a key role in the regulation of metabolic steps within the cell. With the recent development and progress of projects of structural and functional genomics and metabolomics, the systematic collection, accessibility and processing of enzyme data becomes even more important in order to analyse and understand biological processes.

The protein function database BRENDA [1] was founded in 1987 at the German National Research Centre for Biotechnology (GBF) and is continued at the Cologne University Bioinformatics Centre (CUBIC). Firstly, BRENDA was published as a series of books (Handbook of Enzymes, Springer [2]). The second edition was started in 2001. Eighteen volumes have been published so far, each containing about 500-600 pages encompassing 50-150 EC classes. By 2006, 15 more volumes will have been produced.

BRENDA contains a very large amount of enzymatic and metabolic data and is updated and evaluated by extracting information from the primary literature. BRENDA represents a comprehensive relational database containing all enzymes classified according to the EC system of the Enzyme Nomenclature Committee (IUBMB [3]). This classification is based on the type of reaction (e.g. oxidation, reduction, hydrolysis, group transfer) catalysed by the enzyme.

All data in BRENDA have a standard structure:

Value (or range of values), e.g. Turnover number

Protein (information on the exact protein, either organism or - if available - sequence)

Literature reference

Commentary (giving experimental conditions, isoform, etc.)

Additional information (e.g. Substrate for kinetic constants, reversibility and product for substrate fields, etc.)

Since 1998 all data are available on the internet in a relational database system. Since then the user interface has been developed intensively to provide a sophisticated access to the data. The user can choose from seven search modes:

Quick search, Full text search, Advanced search, Substructure search, TaxTree search, ECTree browser, and Sequence search. In 2003 a BRENDA discussion forum was started.

Access to BRENDA is free for the academic community at <http://www.brenda.uni-koeln.de>. An in-house version for academic users is available for a low handling fee. Commercial users are required to purchase a license.

PHILOSOPHY

In contrast to other databases, BRENDA is not limited to a specific aspect of the enzyme or to a specific organism. It covers organism-specific information on functional and molecular properties, enzyme names, catalysed reaction, occurrence, sequence, kinetics, substrates/products, inhibitors, cofactors, activators, structure and stability. Presently, BRENDA holds information on 4200 enzyme classes, which represent more than 83,000 different enzyme molecules. Since 2002 the annotation speed has been tripled to 1000 enzyme classes per year.

THE ANNOTATION PROCEDURE

The annotation procedure comprises the insertion of new data, the reallocation and reclassification of enzymes to their respective class and the removal of data which have been proven to be wrong. The data are annotated manually and are controlled for consistency via computer-aided and manual techniques. Special sections of BRENDA contain automatically annotated data which are indicated explicitly.

Step 1 Enzyme Names

The first step is the search for all the names with which the enzyme is associated. Enzyme names can be found at the IUBMB, from databases, or from the literature.

Step 2 Literature evaluation

In the literature evaluation step the major databases (CAS [4], PubMed [5], databases for specific protein classes) are searched for literature dealing with the respective enzyme class. The number of citations varies greatly with the enzyme class, some searches will result in only a single publication for an enzyme class, others may produce 10,000 hits. In a series of refinement cycles these search results are reduced to those references which will probably yield information that is suitable for at least one of BRENDA's information fields. Manual assessment of the title or the abstract is often necessary to make the right choice.

Step 3 Annotation and quality control

The annotation involves a high amount of manual work because the literature references rarely contain all relevant data in concise tables. Another great amount of work is required to sort out all the various names for enzymes and chemical compounds. Quite often in this stage of annotation, the literature reveals data for enzymes which are not yet classified in the EC number system. These are then collected, completed via an exhaustive search and assembled as a proposal for a new entry in the list of enzymes at the IUBMB. After approval a new EC number is awarded which is then integrated into BRENDA.

The revision and annotation process frequently reveals inconsistencies regarding the enzyme's classification. Then the respective enzyme will be allocated to a different enzyme class after the IUBMB has given approval.

AUTOMATIC CONTROL OF DATA (selected)

- no data-fields missing?
- EC-Number correct?
- all references cited?
- all organisms cited?
- entries in numeric fields in the correct range?
- all brackets, braces, parentheses correct?
- structure of commentary correct?
- journal abbreviations according to list?
- all organisms cited with their correct references and vice versa?
- names for organisms in accordance to *NCBI taxonomy*?
- CAS Number correct?
- correct terms in fields application, post-translational modification?

In addition for a number of fields a controlled vocabulary was introduced and is checked during processing time (application, cofactors, localization, organic solvent stability, post-translational modification, reaction type, source tissue, subunits).

Step 4 Processing the database

In consecutive final steps the data are processed for integration into the database.

Compilation of BRENDA database:

- Parsing of TEXT data, integration into non-organism-specific database, final automatic control
- Split up of database into multiple tables with organism-specific information.

Compilation of BRENDA LIGAND database:

- draw structures of new ligands (Mol-format)
- convert to SMILES
- create thesaurus
- convert mol-files to gif-images.

THE BRENDA DATA STRUCTURE

- Classification and Nomenclature
- Reaction & Specificity
- Functional Parameters
- Organism related Information
- Enzyme Structure
- Isolation and Preparation
- Literature References
- Application and Engineering
- Enzyme-Disease Relationship

CLASSIFICATION AND NOMENCLATURE

Since enzyme names have a long history they are not unique. In many cases the same enzymes became known by several different names, while conversely the same name was sometimes given to different enzymes. Many of the names conveyed little or no idea of the nature of the reactions catalysed, and similar names were sometimes given to enzymes of quite different types.

The International Commission on Enzymes was founded in 1956 by the International Union of Biochemistry. Since then the system of EC numbers with systematic names and recommended names has been established.

Currently there are 3741 active EC numbers plus 556 numbers for deleted or transferred enzymes. The old numbers have not been allotted to new enzymes; instead the place has been left vacant or comments are given concerning the fate of the enzyme (deletion or transfer).

In the EC number system an enzyme is not defined by its name but by the reaction it catalyses. In some cases where this is not sufficient, additional criteria are employed such as cofactor specificity or stereospecificity of the reaction. The 3741 active EC numbers currently account for 28,900 synonyms.

THE ENZYME NOMENCLATURE PROBLEM

Unlike other protein classes, a standard nomenclature and recommended names exist for enzymes. Unfortunately they are often not used by researchers in publications. Therefore, often many different names are in use for enzymes, EC 3.1.21.4, i.e. "type II site specific deoxyribonuclease" with 370 different names. Thus, if a researcher searches in literature databases (e.g. PubMed) only those references will be found which are stored with the synonym he uses. The particular name chosen may be in fact a rarely used synonym and thus he will retrieve only a fraction of the information. Table 1 contains examples of enzymes which are characterized by manifold synonyms.

One important aspect of BRENDA data input is to give the user complete information for an enzyme when he queries the database with a single synonym. Thus great effort is invested in the best possible completeness of enzyme names. The majority of the names are extracted manually from the original literature and completed by searching internet databases (e.g. CAS, PubMed, SwissProt).

Table 1. Enzymes with manifold names in BRENDA.

EC-Number	Recommended Name	Number of Synonyms
3.1.21.4	type II site-specific deoxyribonuclease	369
3.1.3.48	protein-tyrosine-phosphatase	169
1.6.5.3	NADH dehydrogenase (ubiquinone)	162
2.7.7.6	DNA-directed RNA polymerase	91
3.1.2.15	ubiquitin thiolesterase	81

Experimental Enzyme Data as Presented in BRENDA

2.7.1.69	protein-Npi-phosphohistidine-sugar phosphotransferase	80
5.2.1.8	peptidylprolyl isomerase	111
3.1.3.16	phosphoprotein phosphatase	74
3.2.1.4	cellulase	72
3.1.1.1	carboxylesterase	60
3.6.3.14	methylphosphotrioglycerate phosphatase	56

THE UNIQUENESS PROBLEM

Another problem in enzyme literature is that often identical names or abbreviations are applied for more than one enzyme thus creating confusion. Moreover the use of ambiguous names would create completely misleading results. In many cases names or abbreviations refer to more than one EC number (Figs 1-3), e.g. The name GTPase applies to 6 different EC numbers within the same subclass, the abbreviation FDH applies to 8 EC numbers in 3 different subclasses, or the name chondroitinase applies to 5 different EC numbers in 2 different EC classes. Thus the use of any arbitrary enzyme name can lead to great confusion and misleading results in the selection of enzyme data from a database.







EC Number	Recommended Name	Synonyms
 3.6.1.46	heterotrimeric G-protein GTPase	GTPase
 3.6.1.47	small monomeric GTPase	GTPase
 3.6.1.48	protein-synthesizing GTPase	GTPase
 3.6.1.49	signal-recognition-particle GTPase	GTPase
 3.6.1.50	dynamine GTPase	GTPase
 3.6.1.51	tubulin GTPase	GTPase

Figure 1. Enzyme classes carrying the name GTPase.









EC Number	Recommended Name	Synonyms
 1.1.1.1	alcohol dehydrogenase	FDH
 1.1.1.122	D-threo-aldose 1-dehydrogenase	FDH
 1.1.99.11	fructose 5-dehydrogenase	FDH
 1.2.1.1	formaldehyde dehydrogenase (glutathione)	FDH
 1.2.1.2	formate dehydrogenase	FDH
 1.2.1.43	formate dehydrogenase (NADP)	FDH
 1.2.1.46	formaldehyde dehydrogenase	FDH
 1.5.1.6	formyltetrahydrofolate dehydrogenase	FDH

Figure 2. Enzyme classes carrying the name FDH.






EC Number	Recommended Name	Synonyms
 3.1.6.4	N-acetylgalactosamine-6-sulfatase	chondroitinase
 3.1.6.12	N-acetylgalactosamine-4-sulfatase	chondroitinase
 3.2.1.35	hyaluronoglucosaminidase	chondroitinase
 4.2.2.4	chondroitin ABC lyase	chondroitinase
 4.2.2.5	chondroitin AC lyase	chondroitinase

Figure 3. Enzyme classes carrying the name chondroitinase.

In BRENDA the information on enzyme nomenclature can be retrieved from the section Classification & Nomenclature which is divided into the data-fields:

• Enzyme Names	34,509 entries
• EC Number	4293 entries
• Recommended/Common Names	4293 entries
• Systematic Names	3425 entries
• Synonyms	27,903 entries
• CAS Registry Number	3955 entries

Any query in BRENDA will have the EC number as the result, thus enabling the user to select the correct enzyme.

REACTION AND SPECIFICITY - METABOLITES AND LIGANDS

An enzyme is defined by the reaction it catalyses. Thus all proteins found to catalyse a specific reaction are summarized under one EC number. Apart from this an enzyme may have a wider substrate specificity and may accept different substrates. These appear in BRENDA in the section Substrate/Product and Natural Substrate/Natural Product.

Additional sections provide lists of inhibitors, cofactors and activating compounds. Since in biological sciences very often trivial names are used instead of IUPAC nomenclature many compounds are known by a variety of names. Thus even simple molecules may have a dozen or more names. For example, the inhibitor 2,2'-bipyridine is cited with 12 different names. BRENDA is equipped with a thesaurus for ligand names. This thesaurus is based on the generation of unique and chiral SMILES-strings [6, 7] for ligand structures in the database.

If the function of a compound is not known, it can be searched in the table LIGANDS. This will perform a search in all data-fields which contain ligand names (substrates, products, natural substrates, inhibitors, cofactors, activating compounds, K_M , K_i).

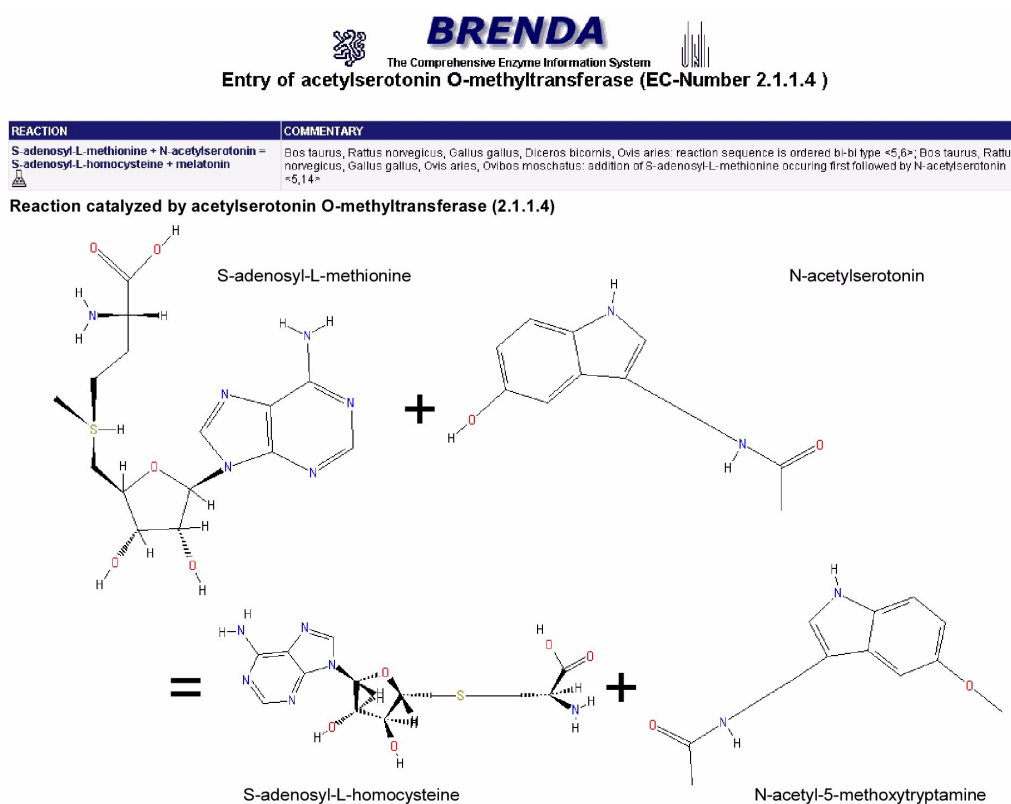


Figure 4. Display of enzyme-catalysed reactions in BRENDA.

The most exhaustive search for ligands is a full-text search of the complete database. This mode, however, does not apply the thesaurus for molecule names. Ligands can also be viewed as 2D-structures thus offering an unambiguous method to display a reaction. (Fig. 4).

Overview BRENDA ligand data

• enzyme/ligand relationships	537,293
• Cofactor	9572
• Activating Substance	10,600
• Metals/Ions	15,641
• Substrates	255,270
• Products	237,686
• Natural Substrate	16,148
• Inhibitors	72,982
• different ligand names	54,895
• ca. 5000 of these macromolecules, molecule classes etc.	
• ligand structures as mol-files	36,820
• Ligand name thesaurus	
• Grouped into 25198 different compounds	

ENZYME FUNCTION AND STABILITY

Functional Parameters

The BRENDA database contains a section for functional parameters of enzymes with these datafields:

• Functional Parameters	116,130
• K_M -value	47,299
• Turnover number	8010
• K_i -value	4441
• Temperature optimum	7762
• Temperature range	1826
• pH optimum	18,086
• pH range	4825
• pI -value (coming soon)	


Experimental Enzyme Data as Presented in BRENDA

Each of these data-fields is divided into subsections.

Example K_M -Value

- Value
- Substrate
- Organism
- Protein (Swissprot/Trembl Code if available)
- Commentary
 - Experimental conditions
 - Isoform
 - Method
 - Other commentaries
- Literature reference
- Date of last change

An example of entries for turnover numbers can be seen in Figure 5.



Entry of alcohol dehydrogenase (EC-Number 1.1.1.1)

TURNOVER NUMBER	TURNOVER NUMBER MAXIMUM	SUBSTRATE	ORGANISM	COMMENTARY	LITERATURE	IMAGE
18480	-	ethanol	Equus caballus	wild-type enzyme Adh 1 <92>	92	2D-image
12800	-	2-buten-1-ol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	2D-image
7980	-	cinnamyl alcohol	Equus caballus	wild-type enzyme Adh 1 <92>	92	2D-image
7380	-	ethanol	Equus caballus	wild-type enzyme Adh 1 <92>	92	2D-image
7080	-	butanol	Equus caballus	wild-type enzyme Adh 1 <92>	92	2D-image
6120	-	ethanol	Equus caballus	mutant enzyme W54L <92>	92	2D-image
5880	-	cinnamyl alcohol	Equus caballus	mutant enzyme W54L <92>	92	2D-image
5380	-	benzyl alcohol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	2D-image
3760	-	ethanol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	2D-image
3650	-	1-octanol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	2D-image
3396	-	hexaldehyde	Equus caballus	-	42	2D-image
3372	-	pentanol	Equus caballus	wild-type enzyme Adh 1 <92>	92	2D-image
3186	-	propan-2-ol	Equus caballus	wild-type enzyme Adh 1 <92>	92	2D-image
2930	-	1-butanol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	2D-image
2930	-	1-pentanol	Rattus norvegicus	isoenzyme ADH-1, pH 10.0 <49>	49	2D-image
2514	-	propionaldehyde	Equus caballus	-	42	2D-image
2124	-	butyraldehyde	Equus caballus	-	42	2D-image
2090	-	butanol	Homo sapiens	-	53	2D-image
1962	-	butanol	Equus caballus	mutant enzyme W54L <92>	92	2D-image
1908	-	acetaldehyde	Equus caballus	-	42	2D-image
1840	-	ethanol	Homo sapiens	-	53	2D-image
1776	-	hexanol	Equus caballus	wild-type enzyme Adh 1 <92>	92	2D-image
1770	-	benzaldehyde	Equus caballus	-	42	2D-image

Figure 5. Sample of turnover numbers in BRENDA.

The data are often obtained under very different experimental conditions. Since every laboratory carries out experiments on enzyme characterizations under individually defined conditions, and since they depend on the given experimental know-how, methods and technical equipment available, raw data for the same enzyme from different laboratories are not at all comparable. Therefore BRENDA not only contains individual values but very often the experimental conditions are also included. Because until now there has been no standardization for documenting these, the details are only given as text. Each entry is linked to a literature reference, thus for reproduction of data the researcher may have to go back to the original literature.

Example: K_M and pH optimum for the human enzymes of glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) (EC-Number 1.2.1.12).

K_M

0.002 3-phospho-D-glyceroyl phosphate, enzyme form E6.8, pH 7 <15>

0.03 3-phospho-D-glyceroyl phosphate, enzyme form E9.0, pH 9 <15>

0.14 3-phospho-D-glyceroyl phosphate, <2>

0.17 3-phospho-D-glyceroyl phosphate, enzyme form E6.8, pH 9 <15>

pH optimum

7 enzyme form E6.8, two pH optima: pH 7.0 and pH 8.5, with activity between pH 7.5 and pH 8.0 being rather low <15>

7.2-7.3 reaction with 3-phospho-D-glyceroylphosphate <34>

8.5 enzyme form E6.8, two pH optima: pH 7.0 and pH 8.5, with activity between pH 7.5 and pH 8.0 being rather low <15>

8-8.3 reaction with D-glyceraldehyde 3-phosphate <34>

9.8 enzyme form E8.5, D-glyceraldehyde 3-phosphate <15>

Experimental Enzyme Data as Presented in BRENDA

These data do not allow automatized access and are unsuitable for the modelling of sections of the cellular metabolism, the whole cellular metabolism or the interaction of cells within tissues and organs. Thus a new data model is needed which provides data which have been generated under standardized experimental conditions.

Stability Parameters

In BRENDA the stability of the enzymes is documented in six sections

• Stability parameters	27,154
• pH stability	3755
• Temperature stability	8841
• General stability	5702
• Organic solvent stability	452
• Oxidation stability	452
• Storage stability	7951

Stability data are especially difficult to put into an automatically interpretable format since the literature data are very inhomogeneous. Whereas one research group states an enzyme to be stable at a certain temperature another will find it to be highly unstable. The discrepancy may be due to the type of buffer, presence of substrates, cofactors, stabilizing or destabilizing ingredients, type of storage vial.

Even some of the purification steps can result in a lower or higher stability. Therefore the stability data in BRENDA are as detailed as possible, reproducing details from the literature.

The sections on General stability and Storage stability contain the organism, text describing conditions, a time and a reference. These two sections contain rather inhomogeneous information for which a standard format has not yet been found.

The sections on pH stability and Temperature stability contain a value, the organism, a commentary and a literature reference. Looking at the value alone will not give sufficient information because the enzyme may have varying stabilities depending on the presence of buffer components or stabilizing/destabilizing agents.

Standardization of experimental conditions

Standardization of experimental conditions is a prerequisite for two reasons:

1. In order to render kinetic or stability data comparable they must be obtained under identical experimental conditions. It is impossible to compare the efficiency of two enzymes if their reaction has been monitored at different pH values which may not even be the optima. Also the stability of enzymes can only be compared if the conditions are identical. In BRENDA ca. 50% of the K_M values are measured at physiological pH values, ca. 33% refer to natural substrates.
2. For the creation of metabolic networks the kinetic data must represent the enzyme's reaction under physiological conditions. These have to be defined regarding the temperature, the pH, ionic strength, or macromolecular crowding. Assay procedures and assay conditions need to be the same to obtain comparable data.

Organism-related information

For the organisms in BRENDA the taxonomy-lineage is given if the respective organism can be found in the NCBI taxonomy database. Using the TaxTree search mode the user can search for enzymes along the taxonomic tree and move to higher or lower branches to either get an overview or to restrict the search.

The tissue may be an important criterion for an enzyme. Sometimes enzymes are restricted to a single tissue or a tissue may express a tissue-specific isoenzyme. The BRENDA tissues are grouped into a hierarchical ontology which was developed especially for this database.

The localization terms are in accordance with the terms of the Gene Ontology [9] consortium.

Overview organism-related data:

• Organism/enzyme relationships	69,408
• from 6728 different organisms	
• Source Tissue/enzyme relationships	25,482
• for 1408 different tissues and cell-lines	
• Localization/enzyme relationships	10,973
• for 148 different subcellular locations	

ENZYME STRUCTURE

Whereas the SwissProt and PDB links are automatically generated, the sections molecular weight, subunits and post-translational modifications are extracted manually from the literature. As the accuracy of the value for the molecular weight or the size of the subunits is dependent on the method of determination, BRENDA gives the method in the commentary, if available. Of the 3741 EC classes sequences are only available for 2166 classes.

Overview enzyme structure

• SwissProt links	53,999
• PDB links	10,610
• Molecular weight	17,715
• Subunit	10,744
• Posttranslational modification	19,019

Isolation and Preparation

The isolation/purification section contains information on purification, crystallization, cloning and renaturation. Due to the inhomogeneity of the data, these are in non-structured text-format.

Overview isolation and preparation

• Purification	14,380
• Cloned	5491
• Renaturation	317
• Crystallization	1548

LITERATURE REFERENCES

For the BRENDA database all information except the sequence information and the enzyme-associated diseases is manually extracted from 50,300 scientific publications. The major drawback of this method is the low speed of annotation compared to automated methods. During the manual annotation procedure the scientist is able to assess the facts, compare the results of different research groups, and choose the data which he wants to include in BRENDA depending on the experimental conditions. For example data obtained with a crude cell extract have to be distinguished from data that were obtained with a purified enzyme.

Studying the literature of an enzyme sometimes reveals misclassification and thus leads to the transfer of an enzyme to another enzyme class.

METABOLIC DISORDER-RELATED INFORMATION

BRENDA contains a large section of data for metabolic disorders which are connected to a dysfunction of an enzyme. However, due to the rapid growth of information there is a widening gap between manually annotated data and information available in the literature. In order to alleviate the problem a tool to automatically extract enzyme-related information from the biomedical literature was developed. It is based on the co-occurrence of enzyme names and interesting phrases which are identified utilizing concepts from the Unified Medical Language System (UMLS) [12]. A variety of filters reduce the number of false extraction events, among them a classification of sentences based on their semantic context by a Support Vector Machine (KMO) [13].

A prototype of this concept based approach links 524 enzyme classes from the BRENDA database to more than 1400 disease related concepts, achieving a precision of more than 90% and a recall of 49% on a test-set of 1500 manually annotated sentences. Current work is focusing on expanding the scope of the tool to include other fields of interest, i.e. subcellular localization of enzymes or co-occurrences of enzyme names with pharmaceutical compounds.

Overview Disease-related data

- ca. 50,000 PubMed references with disease-term and enzyme name in title
- ca. 20,000 references selected by text-mining tool
- 506 EC numbers in disease-related papers
- 1407 disease terms related to enzymes

APPLICATION AND ENGINEERING

- | | |
|---------------|------|
| • Application | 1413 |
| • Engineering | 4531 |

Enzymes are widely applied in industry, pharmacology, medicine or for analytical purposes. BRENDA not only lists established applications but also putative future usages.

This data field is based on a controlled vocabulary, the comments are in text format. The engineering section displays the amino acid exchange in the engineered enzyme. The comments give as much detail on the properties of the mutated enzyme as available, which is mostly restricted to a short comment on the activity or stability. For mutants with kinetic constants these can be found in the functional parameters section.

SUMMARY AND PERSPECTIVES

The enzyme database BRENDA represents data for ~4000 enzyme classes defined in the EC system. The data give detailed information on nomenclature, specificity, structure, organism, functional parameters, enzyme stability and diseases related to dysfunction. All data are linked to primary literature references. Enzyme data are essential for understanding and predicting the biological chemistry of the cell. For a reliable interpretation of these values by computational methods standardization is indispensable:

1. All enzymes names must be in accordance to the IUBMB system of enzyme nomenclature.
2. Thermodynamic and kinetic data must be recorded under defined conditions, mimicking physiological conditions.
3. Metabolites must carry unequivocal names or identifiers.
4. Organisms and cell-types, tissues and cellular components must be named in accordance to defined ontologies.

REFERENCES

- [1] Schomburg, I., Chang, A., Ebeling, E., Gremse, M., Heldt, C., Huhn, G., Schomburg, D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucl. Acids Res.* **32** : D431-D433.
 - [2] Schomburg, D., Schomburg, I. (2001) *Springer Handbook of Enzymes*, 2nd Edn. Springer, Heidelberg, Germany.
 - [3] *Enzyme Nomenclature* (1992) Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, NC-IUBMB. Academic Press, New York.
 - [4] Ridley, D.D. (2002) *SciFinder and SciFinder Scholar*. J. Wiley & Sons, New York
-

-
- [5] Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L. Rapp, B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* **29**: 11-16.
- [6] Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**: 31-36.
- [7] Weininger, D., Weininger, A., Weininger, J. (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **29**: 97-101.
- [8] Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen E. (2003) The Chemistry Development Kit (CDK): An open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **43**(2):493-500.
- [9] Ashburner, C.A., Ball, J.A., Blake, D., Botstein, H., Butler, J.M., Cherry, A.P., Davis, K., Dolinski, S.S., Dwight, J.T., Eppig, M.A. et al. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**: 25-29.
- [10] Berman, H.M., Westbrook, J., Feng, Z., Gillilan, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000) The Protein Data Bank. *Nucl. Acids Res.* **28**: 235-242.
- [11] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., 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. *Nucl. Acids Res.* **31**: 365-370.
- [12] Bodenreider, O. (2003) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucl. Acids Res.* **32**(database issue): 267-270.
- [13] Kazama, J., Makino, T., Ohta, F., Tsujii, J. (2002) Tuning support vector machines for biomedical named entity recognition. *Proceedings of the workshop on natural language processing in the biomedical domain*, Philadelphia, pp.1-8.
-

SYSTEMATIC NAMES FOR SYSTEMS BIOLOGY

RICHARD CAMMACK

Department of Life Sciences, King's College London, UK

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

Received: 14th April 2004 / Published: 1st October 2004

ABSTRACT

The aim of systematic nomenclature is to provide a name for each entity, such as a metabolite, an enzyme, or a measured quantity. There are different requirements for biochemical nomenclature, depending on how the name or symbol is to be stored and communicated, by written, printed, or spoken word, as a diagram, or as computer-readable data. Names are often related to biological function, structure or evolutionary relationships; nomenclature follows classification. For interaction with computers and databases, identifiers should be searchable, and referred to an authoritative source. The requirements for nomenclature are distinct from those of a dictionary, where the criterion for inclusion of a word is that it is used. When proposing systematic nomenclature, timely intervention is important, and much effort should be devoted to ensuring acceptance of within the scientific community.

INTRODUCTION: THE NEED FOR NOMENCLATURE SYSTEMS

Communication in science relies on having consistent and recognizable terminology, units and symbols. This is particularly important in multidisciplinary areas such as systems biology. For mathematicians, biochemists, bioinformaticists, chemists and other scientists to communicate, we need a standard and unambiguous name for each entity or concept. Inevitably, new species, compounds and concepts will be given different names at first, by research workers in different subjects. There needs to be agreement as to what these names should be, and there needs to be a mechanism to connect these "preferred" terms to other terms found in the literature. This is an important function of databases, glossaries and dictionaries. Interoperability of databases also depends on consistent nomenclature. In other words, we need to know we are talking about the same thing.

There will be a large number of enzymes, proteins and other cell factors to be named in the future. In the genome of the well-understood organism *Escherichia coli*, there are about 36% of the predicted gene products for which the function is still completely unknown. Moreover some of the present annotations will probably prove to be inaccurate, and there are many new species that have hardly been explored. So we need a system in place to agree on good names for these entities.

Table 1. Criteria for a good name or descriptive phrase.

Essential	Advisable
unique	good search term
infinitely extendable	memorable
not be easily confused with anything else	reflects structure or function
open source	decided by international authority
not copyright or a trademark	not obscene or hilarious in any language

There are many different requirements for a system of nomenclature, and no system is perfect. However for the purposes of this discussion it is useful to list the desirable characteristics of an effective nomenclature system (Table 1).

WHAT MAKES A GOOD NAME?

There should be standards for nomenclature and symbols in systems biology. A good name for an entity or phenomenon can crystallize our thoughts about it.

What makes a good name ? This depends on:

- the medium in which it is to be presented (Table 2)
- recommended or informal nomenclature
- who is intended to use it:
 - laboratory specialists
 - specialist community
 - wider biochemical community
 - scientific community

Systematic Names for Systems Biology

Table 2. Different formats, and nomenclature issues.

Format	Problem	Example
Printed word	Character formats	l and 1, O and 0
Handwritten (lab notebook)	Legibility of symbols	v and V (kinetics)
Spoken word	Pronounceable	Sulfenate/sulfinate
Diagrams	Computer readability	Metabolic maps
Structures	Standard representation	
ASCII text, internet	Special characters	Greek letters, italics, subscripts etc.
Database	Generally only use ASCII	Consistent use, e.g. Unicode
Proprietary, Trademarks	Interoperability	Adrenaline/epinephrine

Some names can be helpful, because they invoke an analogy with another well-known term. "Polymerase chain reaction" (PCR) is such a name, as it suggests a similarity to other chain reactions of chemistry and nuclear physics: one molecule leads to more and more products. On the other hand some names can trip up the uninitiated. "Real-time" is a well-known term in computer science to describe a process that a computer monitors *as it occurs*. Someone reading about "real-time PCR" might expect it to be some sort of continuous or instantaneous measurement. In fact the most significant feature of real-time PCR is that it provides a *quantitative* measure of the amount of DNA. Add to this the frequent use of "RT-PCR" for both this method and "reverse-transcription PCR", and there is scope to confuse the uninitiated.

Names originating from laboratory jargon often cause problems.

- An example of informal notation is the letter "p" followed by a number. Often the number represents the apparent molecular mass, in kilodaltons, on SDS-polyacrylamide gel electrophoresis. An example is the intensively studied tumour-suppressor gene p53 (actually a tetramer, so its molecular mass is 212 kDa). Such names are not extendable, although a series of related proteins, p63 and p73 have been described. Sometimes the "P" (in capitals) stands for "pigment". The well-known cytochrome P450 is named for the Soret peak in its carbon monoxide difference spectrum. Being an enzyme and not just an electron-transfer protein, it is not recognized as a cytochrome in systematic nomenclature [1].
- X+number, where X stands for species: "H" could be horse, horseradish or Hansenula. There are simply too few letters in the alphabet!

- TLA's (three-letter abbreviations) and similar short names are a source of confusion.

Such abbreviations have a large number of hits when searched in PubMed, but this may conceal a range of different meanings.

Since there are only 17,576 permutations, the chances of having more than one meaning is high (Table 3). Often the meaning of the abbreviation is buried in the text of a paper, which makes it difficult for the reader to find.

Table 3. An example of ambiguity with TLA's.

ACF	ATP-utilizing chromatin assembly and remodeling factor
	APOBEC-1 complementation factor
	aberrant crypt foci of the colon
	anticoagulation factor
	2-[(2-amino-4-chloro-5-fluorophenyl)thio]-N,N-dimethyl-benzenmethanamine
	anterior corpectomy with fusion
	N-acetyl phenylalanine
	accessory colonization factor

Acronyms are useful if they are good search terms. They should preferably not be the same as common words, e.g. WAVE designed for DNA fragment analysis. This helps as a mnemonic, but makes them difficult to find in literature searches.

Names with complex syntax, such as capitals and small letters, mixed with numerals, have the problem that they are easy to forget, and mistakes are often made. One only has to think of the complex passwords that are required to log onto some secure data systems ("Forgotten your password again ? Click here..."). But biochemical nomenclature which has no systematic basis is also difficult to use consistently. An example shown in Table 4 is the proton-translocating ATP synthase of mitochondria, known as the F_0F_1 ATPase (class EC 3.6.3.14). This name dates from a time when manuscripts were typewritten, and there were inconsistent uses of characters, such as capital "O" or zero. In the original form the lower case subscript "o" stood for oligomycin-sensitive, and F_1 represented the large water-soluble part of the protein complex. Thus, many different variants have appeared in the literature.

Systematic Names for Systems Biology

Table 4. Synonyms in the literature for the proton-translocating ATPase.

	Frequency of use	
	Web of Science [®] *	Google TM *
F ₁ ATPase	531	7590
F-1 ATPase	508	4520
F ₀ F ₁ ATPase	225	911
FoF ₁ ATPase	30	426
F ₁ Fo ATPase	30	554
F ₀ /F ₁ ATPase	3	167
F ₀ F ₁	266	1750
FOF ₁	51	4680
F ₁ F ₀	279	3610
F ₁ Fo	40	772
ATP synthase	1365	63600
ATP synthetase	22	31500

This raises a number of points about the differences in usage of names in the written word and in computers. A human reader will easily recognize that all the terms in Table 4 probably represent the same thing. However character-matching software obviously regards them as distinct, since the hits in Table 4 were found by searching computer databases. In order for a database of enzymes to provide an accurate representation of the literature on the subject, it must include all the variant forms. In fact the number of variants is actually much greater than indicated in Table 4, because the search engines used to determine the number of "hits" take no account of upper and lower case, italics, subscripts and superscripts, greek letters and other symbols.

The requirements for a distinct written name, and a good search term, mean that compromises are being made in terminology. Databases may employ some form of encoding to distinguish variants in syntax, although there is no consistent practice. In the literature, features of punctuation, such as italics in species names and foreign phrases, are increasingly being omitted. A recent such recommendation is that the italics representing the source organism in symbols for restriction enzymes should be omitted for example *Eco*R1 would be EcoR1 [2].

SYSTEMATIC CHEMICAL NAMES

In chemistry, the most important characteristic of a compound for classification purposes is usually its structure. Chemical compounds were first given arbitrary names, as they were identified, but the number of these had become unsustainable during the 19th century. International efforts to create an acceptable system of nomenclature of organic compounds date back at least as far as the Geneva Convention of 1892, and have been extended and refined ever since [3]. This was followed by recommendations for inorganic, physical, organometallic and macromolecular chemistry. These systems of chemical nomenclature are principally aimed at providing a name, which can be written or spoken, that defines every compound. The names of compounds defined by the IUPAC systems have legal standing, for example in patents.

A useful introduction to the principles of chemical nomenclature is provided by the Guide to IUPAC recommendations [4]. Usually the name of a compound is derived from a parent compound, with substituents at positions defined by a numbering system. The formalisms are continually being reviewed and extended, to describe new classes of molecules such as fullerenes.

Computer databases are now an indispensable part of sciences such as organic chemistry, where enormous numbers of new compounds are synthesized. They allow information on structures, spectroscopic and other physical properties, to be assembled in an accessible way. The new areas of science such as systems biology would not be possible without computer databases. Databases such as the CAS (Chemical Abstracts Service) Registry (<http://www.cas.org/EO/regsys.html>) list all compounds, including biochemical compounds and gene sequences. The CAS number is an identifier, and can only be understood in the context of the database. It provides a means of cross-referencing different names for a compound. A biochemical compound such as glucose may have several CAS Registry numbers, reflecting the different enantiomers, open-chain and ring structures that interconvert spontaneously in solution.

Generally, chemists who are non-experts in nomenclature find it easier to visualize a chemical structure than to interpret a systematic chemical name. It is easy to make mistakes when deriving a chemical name. Changing a bond in a ring structure, for example, can completely change the numbering of the rest. Increasingly the task of converting structures to names, and names to structures, is being taken over by software, such as the programs used for drawing chemical structures, which implement the rules of chemical nomenclature.

As an alternative to systematic chemical names, linear notations for molecular structures have been developed. SMILESTM strings (<http://www.daylight.com/>) are a well-established notation for chemical structures. Based on a set of simple rules, they are readily generated for a particular molecule. This can also be done by proprietary software packages.

If the data on chemical compounds is to be stored on databases, it becomes less important that the name used is readable by humans. A recent innovation is the ICHI (IUPAC Chemical Identifier) or INCHI (IUPAC-NIST Chemical Identifier) [5,6]. This is an ASCII string, generated by a computer algorithm, that uniquely defines a compound. In contrast to the SMILES system, where often several valid strings can be written for a compound, every chemical structure yields a unique INCHI. The INCHI is open-source, whereas the software to create SMILES strings is proprietary, and even some of the strings themselves are copyright. The INCHI has the status of a IUPAC project at the moment, and software to use it has yet to be developed. However this identifier, if adopted widely, should be extremely useful in databases.

An important feature of the description of an entity in a database is its *identifier*. This is an invariant label for the entity within the data system. It should be extensible, that it has sufficient letters and digits to encompass all examples that could possibly be encountered. It is important to recognize that an identifier should be devoid of any other information. Numbers used as identifiers are never re-used within the database. If the entity is given another name, it can be traced back through the system. Often there are several ways of naming a single compound. These should all be linked to the same identifier. Databases also use preferred names for compounds, a feature known as *controlled vocabulary*.

SYSTEMATIC NAMES IN BIOCHEMISTRY AND MOLECULAR BIOLOGY

In bioinformatics, systematic chemical names would be unwieldy and non-intuitive. Moreover the experimental data on the structures and characteristics of biological molecules, is disparate, incompletely defined, and distributed among different online databases. When working with information on the Internet, interoperability is a watchword. This means that, while working on a database, information in other databases should be only a few clicks away. The need for any conversion software or password access slows the process down enormously [7].

Often the same or similar molecule is given a different name, for example if derived from a different species. Computer databases can manage this complexity by storing and manipulating lists of synonyms, as part of their controlled vocabulary.

Systematic, functional nomenclature implies a classification. A classification of a gene product may be on the basis of function, molecular structure, phylogeny or genes. There should be a hierarchy of such criteria, otherwise conflicts will arise where one criterion implies that an entity belongs in one class, and another criterion would put it in another. Genes often have a multiplicity of names. More than one name is used for similar gene products in different species, or even from the same species. Organizations such as the HUGO Gene Nomenclature Committee (<http://www.gene.ucl.ac.uk/nomenclature/>) aim to simplify this multiplicity as much as possible. They publish, and apply, guidelines for nomenclature, in parallel with those for the mouse and other genomes [8].

Biochemical nomenclature has been an ongoing activity since the 1950s. Its original purpose was to arrive at a more consistent terminology in the literature and the textbooks. It is recommended in the instructions to authors of biochemical journals. Editors have a part to play in encouraging authors to use recommended nomenclature and terminology. Currently these activities are coordinated by the Nomenclature committee of IUBMB (NC-IUBMB) and the Joint Commission on Biochemical Nomenclature (JCBN) [9]. The two committees work together, to set up panels for specific nomenclature. This has led to the publication of reports on the nomenclature of proteins, carbohydrates, nucleic acids and other compounds, published in book form [10] and more recently on the web (<http://www.chem.qmul.ac.uk/iubmb/>). Newsletters are posted on the website (<http://www.chem.qmul.ac.uk/iubmb/newsletter/>).

The EC list of enzymes, one of the activities of the committees, provides a good example of a functional classification and system of nomenclature (<http://www.chem.qmul.ac.uk/iubmb/enzyme/>). It is described in the article by Sinead Boyce et al. in this book. The basis of classification is the reaction catalysed. An entry in the EC list denotes simply that an enzyme has been shown to exist, that catalyses the approach to equilibrium of a specific biochemical reaction. The EC number identifies the reaction classified. The EC number lends itself naturally to computing, and there are databases that use it as the primary method of searching, e.g. INTENZ, part of the Expasy database of protein structure and function (<http://www.ebi.ac.uk/intenz/>).

The EC class of an enzyme is arrived at by application of a set of rules [11]. Other secondary criteria such as cofactor composition have occasionally been invoked to distinguish between enzymes, but in most cases they are not admitted since they may cause confusion. The EC list is a classification of enzymes that are demonstrated to exist, rather than a list of possible reactions. Because the EC class may change in the light of new knowledge about the enzyme, it is therefore not an identifier, for database purposes.

The nomenclature committees operate interactively with the biochemical community. The process of classification often begins with the submission of information from someone who is working on that enzyme. There is an exchange of information that leads to the checking of the enzyme details, creation and correction of a draft entry. To fulfil the requirements for public consultation, proposed entries or revisions of the enzyme list are displayed on the website at www.chem.qmul.ac.uk/iubmb/enzyme/newenz for two months, while biochemists (including those who proposed the entries) are invited to comment. After the consultation period the entry is corrected and put into the enzyme list. (Fig. 1). EC numbers are never re-used even if they are finally not approved or they become superseded. The progress of any changes to EC numbers is traceable through information on the website.

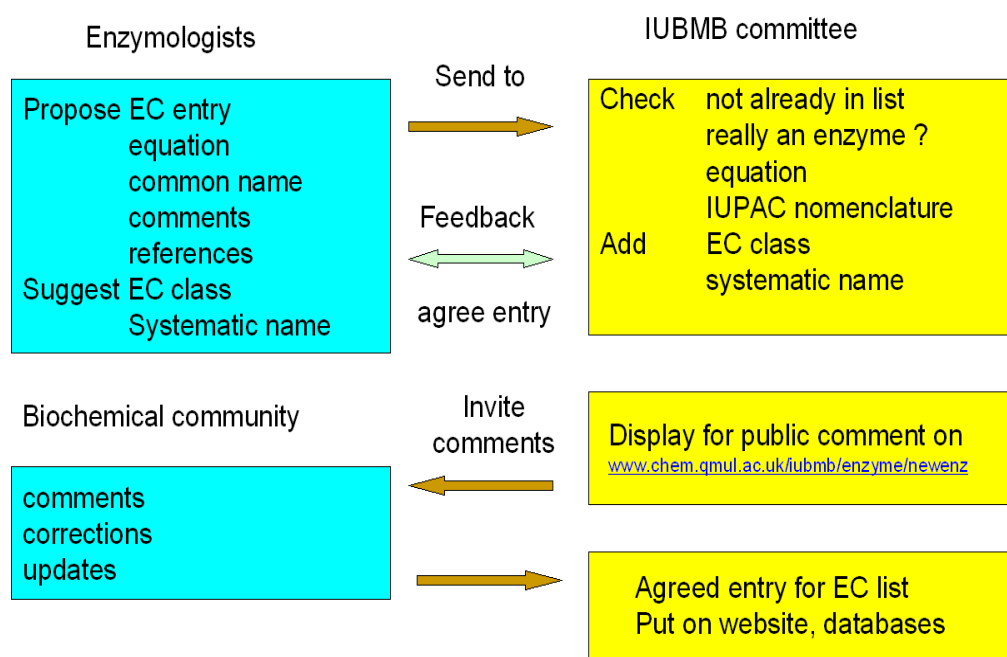


Figure 1. Flow chart for classification of an enzyme

DICTIONARIES AND GLOSSARIES

Along with the development of online bioinformatics databases, there is still a need for dictionaries and glossaries. Although an unfamiliar word can be found by using search engines such as Google and the literature databases, a dictionary describes how it is normally used [12]. The inclusion of a misleading or ambiguous term provides the opportunity for cross-references to alternative or recommended names.

The principal criterion for inclusion of a new biochemical term in a dictionary or glossary is that it is widely used in the laboratory and in the literature. Literature searches provide a means to check the frequency with which terms are used, by their prevalence in the titles, keywords and abstracts of the relevant journals in biochemistry and molecular biology. For the second edition of the *Oxford Dictionary of Biochemistry and Molecular Biology*, the informal criterion being applied is that, for inclusion, a neologism should be mentioned at least 10 articles per year in titles, keywords and abstracts of the relevant journals.

ACCEPTANCE

The development and acceptance of nomenclature standards has been a gradual process. It is human nature to be reluctant to abandon familiar names, albeit they are non-systematic or even misleading. The work of developing internationally agreed standards has been undertaken by international bodies, particularly IUPAC, which has a formal process of public review before they are accepted.

Finally, it is important to remember that setting a standard does not necessarily lead to compliance. It may be that everyone is talking about the same thing, but not using preferred nomenclature. It is not unknown for official recommendations to have such low levels of acceptance that they are finally forgotten. Generally this occurs when the usage of other names, units and symbols has become established and the scientific community does not perceive a need for new names, units and symbols. However if the new terms are easier to explain, and more intuitive to understand, new generations of students will accept them. Timely intervention is important: not too early when the compounds are inadequately understood; not too late when misleading terms have become embedded in the literature and databases.

REFERENCES

- [1] Palmer, G. (1989) Nomenclature of electron-transfer proteins. *J. biol. Chem.* **267**:665-677.
 - [2] Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., Bitinaite, J., Blumenthal, R.M., Degtyarev, S.K., Dryden, D.T.F., Dybvig, K., et al. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucl. Acids Res.* **31**:1805-1812.
 - [3] Panico, R., Richer, J.C., Powell, W.H. (1993) *A Guide to IUPAC Nomenclature of Organic Compounds*. Blackwell, Oxford.
 - [4] Leigh, G.J., Favre, H.A., Metanomski, W.V. (1998) *Principles of Chemical Nomenclature: a Guide to IUPAC Recommendations*. Blackwell, Oxford.
 - [5] Adam, D. (2002) Chemists synthesize a single naming system. *Nature* **417**:369.
 - [6] Stein, S.E., Heller, S.R., Tchekhovskoi, D.V. (2001) Toward the development of a standard chemical identifier. *Abstracts of Papers of the A.C.S.* **222**:5.
 - [7] Berendsen, H.J.C. (2003) Inter-union bioinformatics group report. *Acta Crystallogr. Section D-Biol. Crystallogr.* **59**:777-782.
 - [8] Wain, H.M., Bruford, E.A., Lovering, R.C., Lush, M.J., Wright, M.W., Povey, S. (2002) Guidelines for human gene nomenclature. *Genomics* **79**:464-470.
 - [9] Cammack, R. (2000) The biochemical nomenclature committees. *IUBMB Life* **50**:159-161.
 - [10] Liebecq, C. (1992) *Biochemical Nomenclature and related documents*. Portland Press, London.
 - [11] Webb, E.C. (1992) *Enzyme Nomenclature*. Academic Press, San Diego.
 - [12] Smith, A.D., Datta, S.P., Howard Smith, G., Campbell, P.N., Bentley, R., McKenzie, H.A. (eds) (1997) *Oxford Dictionary of Biochemistry and Molecular Biology*. Oxford University Press, Oxford.
-

BIOGRAPHIES

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 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 group 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 Ontology Annotation (GOA) and the Proteome Analysis database 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. Rolf Apweiler has also a long-standing interest in data standards and nomenclature as exemplified in his engagement in the IUBMB Nomenclature Committee and in the HUPO Proteomics Standards Initiative.

URLs:

<http://www.ebi.ac.uk/sprot/>

<http://www.ebi.ac.uk/trembl/>

<http://www.ebi.ac.uk/interpro/>

<http://www.ebi.ac.uk/proteome/>

<http://www.ebi.ac.uk/GOA/>

<http://www.ebi.ac.uk/embl/>

Richard Cammack

is a Professor of Biochemistry at King's College, University of London. He is the Chairman of the Nomenclature committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and Joint commission on Biochemical Nomenclature (JCBN), and also editor-in-Chief of the second edition of the Oxford Dictionary of Biochemistry and Molecular Biology.

http://www.kcl.ac.uk/kis/schools/life_sciences/life_sci/metals/Cammack.html

Athel Cornish-Bowden

carried out his undergraduate and post-graduate studies at Oxford, obtaining his D.Phil. with Jeremy R. Knowles in 1967 on the basis of studies of pepsin catalysis in the Dyson Perrins Laboratory. After spending three post-doctoral years in the laboratory of Daniel E. Koshland, Jr., at the University of California, Berkeley, he moved to a position as a Lecturer, and subsequently Senior Lecturer, in the Department of Biochemistry at the University of Birmingham, where he remained for 16 years. Since 1987 he has been Directeur de Recherche in three different laboratories of the CNRS at Marseilles. Despite having started his research career in a department of organic chemistry, essentially all of his research has been related to biochemistry in general and enzymes in particular, including pepsin, mammalian hexokinases, and bacterial enzymes involved in electron transfer. He is thus an enzymologist, with a major interest in kinetics, and has written several books in this area, including *Fundamentals of Enzyme Kinetics* (Portland Press, 1995) and *Analysis of Enzyme Kinetic Data* (Oxford University Press, 1995).

In the past 15 years his interests have been focussed on multi-enzyme systems rather than on the kinetics of single enzymes. This topic includes the regulation of metabolic pathways, and his long-term aim is to develop a modern and coherent theory of metabolic regulation. There has always been a major element of computer analysis in his work, which at different times has involved statistical analysis of data, construction of protein phylogenies, and, most recently, modelling of metabolic systems.

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).

Alisdair Fernie

Max-Planck-Institut für Molekulare Pflanzenphysiologie has worked in the fields of plant biochemistry and plant molecular biology for the last eight years. My major research focus has been understanding the regulation of primary metabolism paying particular interest to carbon metabolism. In recent years this has included both the application of metabolomic techniques to plant systems and the parallel analysis of transcriptional complements in plants and the development of techniques that should facilitate better understanding of spatial and temporal aspects of plant metabolism. The goal now is to utilize these techniques to understand more fully the regulation of, and by, metabolism under a range of environmental and developmental conditions.

Martin G. Hicks

is a member of the board of management of the Beilstein-Institut. He received an honours degree in chemistry from Keele University in 1979. There, he also obtained his PhD in 1983 studying synthetic approaches to pyridotropones under the supervision of Gurnos Jones. He then went to the University of Wuppertal as a postdoctoral fellow, where he carried out research with Walter Thiel on semi-empirical quantum chemical methods. In 1985, he joined the computer department of the Beilstein-Institut where he worked on the Beilstein Database project. His subsequent activities involved the development of cheminformatics tools in the areas of substructure searching and reaction databases, and products such as Current Facts and CrossFire.

After brief sojourns as the managing director of the Beilstein Verlagsgesellschaft in 1997 and subsequently the Beilstein GmbH from 1998 - 2000, he returned home to the Beilstein-Institut as head of the funding department in 2000.

He is particularly interested in furthering interdisciplinary communication between chemistry and neighbouring scientific areas and has been organizing the Beilstein Bozen Workshops since 1988.

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 60s/early 70s 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 Electrobiological".

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 funding, as well as development of new projects and products for the Beilstein-Institut.

Ekaterina Kostina

is a Scientific Assistant at the Interdisciplinary Center for Scientific Computing (IWR) of the Ruprecht-Karls-University of Heidelberg. Her present research work concentrates on numerical methods and software for large-scale nonlinear optimization and optimal control problems, including parameter estimation and design of optimal experiments, in application areas like chemical engineering, aerospace engineering, environmental sciences and finance. After graduating with a Diploma degree in Mathematics from the Byelorussian State University (Minsk) Kostina worked as a Research Scientist at the Institute of Mathematics of the Byelorussian Academy of Sciences, where in 1990 she obtained a Ph.D. in mathematics. In 1997 Kostina moved to IWR, and since then she is a member of the Optimization and Simulation Team. Kostina is an author and co-author of 28 refereed papers in journals, conference proceedings, and in books. She is a co-author of a pending patent on the methods for identification of stability parameters of enzymes.

Thomas S. Leyh

received a Ph. D. in biophysics from the University of Pennsylvania in 1983. He joined the faculty at the Albert Einstein College of Medicine in New York in 1989, where he is currently a Professor of Biochemistry. Prof. Leyh is a mechanistic enzymologist with a long-standing interest in sulfur biochemistry, GTPase function, and the conformational coupling of energetics. His group has recently demonstrated that enzymes in the cysteine biosynthetic pathway self-organize into a multifunctional protein complex out of which emerges new catalytic function that orchestrates the activities of the complex. John Andreassi, Ph. D., is a postdoctoral fellow working with Dr. Leyh to initiate the genomic enzymology program.

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
-

Biographies

- 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
- since 1996: Full Professor of Biochemistry, University of Cologne

Research Interests:

- Protein Structure and Function
 - Structural Biochemistry
 - Bioinformatics
 - Enzyme Information/Metabolic Networks
-

Biographies

Stefan Schuster

born 7. November 1961 in Meissen (Germany);

studies in biophysics at Humboldt University in Berlin; Dr. sc. nat. (PhD) in 1988;

1988-1991: Assistant at the same university, Dept. of Biology;

1991-93: Postdoc at the University of Bordeaux and the Cancer Institute of the Netherlands in Amsterdam;

1993-97: Lecturer at Humboldt-University in Berlin;

1997-2003: Group leader at the Max Delbrück Centre for Molecular Medicine Berlin-Buch;

1997 and 1998: three-month visits at the universities Maribor (Slovenia) and Stuttgart;

2003: Professor of bioinformatics at the University of Jena.

Main topics of research:

Structural analysis of metabolic and regulatory networks; Metabolic Control Theory; Evolution and optimization of enzyme systems; Modelling of calcium oscillations.

Jacky Snoep

received his PhD 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://jij.biochem.sun.ac.za>.

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.

Hans V. Westerhoff

did his Ph D (in 1983) with Karel van Dam at the University of Amsterdam on the Thermodynamics and Control of Biological Free-Energy Transduction. At this stage it was fairly unique for biochemistry to combine experimentation and modelling in a single set of studies. The 'Control' referred to the application of the Metabolic Control Analysis (MCA) developed by Kacser, Burns, Heinrich & Rapoport, and the 'Thermodynamics' to Mosaic Non Equilibrium Thermodynamics developed by Westerhoff and Van Dam. In hindsight, both these approaches were Systems Biology *avant la lettre*, in that they aimed at explaining the behavior of biochemical systems in terms of the interactions of their components. Westerhoff then spent 5 years as a visiting scientists at the (US) National Institutes of Health where he worked on stochastic aspects of proton-mediated free-energy transduction, on DNA supercoiling, on membrane active peptides. There he also began to develop hierarchical control analysis (HCA), the extension of MCA that incorporates gene expression and signal transduction. Back in Amsterdam, first as a group leader at the Netherlands Cancer Institute, then as Professor of Mathematical Biochemistry at the University of Amsterdam and as Professor of Microbial Physiology at the Free University in Amsterdam, Westerhoff and coworkers spearheaded a number of experimental and theoretical developments leading from HCA more and more to full blown Systems Biology.

The applications extended to the cascade regulation of ammonia assimilation in *E.coli*, glucose transport and glycolysis in *E. coli* and *S. cerevisiae*, bioenergetics of *E. coli* and *P. denitrificans*, multidrug resistance in tumor cells. In the Systems Biology field the work of the Westerhoff group is characterized by explicit links between new theoretical concepts, modelling, and quantitative experimentation, with an emphasis of systems where all three are feasible. Of necessity this has lead to an emphasis on microbial systems.

Biographies

In recent years the Systems Biology approach has led the group to 'vertical genomics' related to the production of and by bakers yeast, to network based drug design for the sleeping sickness agent *T. brucei*, to 'Integrative Bioinformatics' combining information from all levels of functional genomics, and to the 'Silicon Cell', a set of computer replica of parts of living cells (www.siliconcell.net). At present Westerhoff is the scientific director of the Centre for Research of BioComplex Systems and of the Institute for Molecular Cell Biology, both in Amsterdam. He is involved in the German, Finnish and Dutch Systems Biology program committees, and serves as a coordinator of a European umbrella initiative for Systems Biology called ESBIGH.



Rear row, from left to right: J. Barthelmes, K. Degtyarenko, T. Leyh, D. Krömker, J. Zügge, H. Westerhoff, C. Kettner, D. Schomburg, M. Kanehisa, R. Apweiler, H. Schlüter.

Middle and front row, left to right: K. Tipton, J. Snoep, F. Lottspeich, S. Schuster, D. Fell, E. Kostina, M. Poolman, H.-G. Holzhütter, H. Sauro, H. Grammel, P. Mendes, H. Bock, A. Fernie, R. Cammack.

Behind the camera: M. Hicks.

Index

INDEX

NUMERICS

15-lipoxygenase 104
15LOX 104

A

ACE 95
alcohol dehydrogenase 34, 123
algebraic differential equation 110
Alkaligenes faecalis 120
analysis of array data 78
angiotensin-converting enzyme 95
angiotensin-I 92
annotation 144, 188
Arabidopsis thaliana 71
Arrhenius kinetic terms 51
assay conditions 34
Azotobacter vinelandii 147

B

Bacillus megaterium 152
baker's yeast 7
biochemical modelling 115
 reaction 116
BioCyc 116, 121
bioinformatics 73, 129
biological
 database 2
 function 203
 reaction 163
 system 129, 132
bioreaction ontology 163
boundary value 53
 problem 46
Brassica napus 77
BRENDA 6, 33, 116, 119, 120, 121, 123,
 145, 185, 186

C

Candida antarctica 59
carbon isotope labelling 76
catalytic reaction 48
CellML 132
cellular metabolism 2
 network 5
 reaction network 99
chemical nomenclature 208
chimeric protein 76
classification 203
common name 19
computational biology 80
Corynebacterium glutamicum 88
cycle oscillations 139

D

database 207, 208
design optimization problem 65
drug development 130

E

E. coli 123
EC number 18, 150
 system 188
elementary mode 124
Embden-Meyerhof-Parnas pathway 6, 138
EMBL 144
environmental perturbation 74
enzymatic activity 92
 reaction 163
ENZYME 145
enzyme 186, 204
 activity 130
 classification 18, 156
 Commission 144
 kinetics 87
 List 18
 Nomenclature 4, 143, 144, 150, 160

Index

nomenclature problem 190
 ontology 169
 -kinetic model 100
 -polypeptide relationship 152
 enzymology 3, 130, 143
 ESCEC 14
Escherichia coli 7, 71, 119, 204
 ExPASy ENZYME 116, 120
 experimental
 biology 14, 71
 condition 10, 130
 costs 68
 data 139
 design 14, 61
 protein data 185
 experimentalists 71

F

fluorescence energy resonance transfer 76
 flux study 76
 F_0F_1 ATPase 206
 FRET 76
 functional
 characteristics 8
 data 5
 genomics 115
 parameter 194
 function-structure relationship 4

G

Gauss-Newton 45
 Gene Ontology Consortium 143, 145
 genechip 78
 genetic perturbation 74
 genomics 186
 GHMP kinase 177
 glycolysis 6, 81, 110, 137, 138
 GO 144
 growth condition 130
 GTD Thermodynamics of Enzyme-catalysed Reactions database 33
 GTPase 191

H

half-life 52
 hexose monophosphate shunt 110
 human
 heart 77
 renin 92

I

immobilization 92
 initial-rate method 100
 IUBMB 18, 187
 IUPAC 208

J

JWS simulation 135

K

KEGG 116, 120
 key enzyme 6
 kinetic
 model 132
 parameter 45, 139

L

laboratory jargon 205
 LC-MS 75
 L-type pentose phosphate pathway 118

M

MALDI-MS analysis 89
 mass spectrometry 88
 mathematical
 biology 14
 description 80
 modelling 46
 MCA 80, 134

Index

metabolic
 control analysis 80, 124, 134
 database 116
 disorder 200
 engineering 115
 flux analysis 80
 map 123
 modelling 125
 network 125
 networks 2, 115
 pathway 3, 115, 165
 pathway analysis 115
 metabolism 186
 metabolite
 analysis 71, 75
 profiling 73
 metabolomics 186
 METATOOL 124
 mevalonate pathway 178
 MFA 80
 Michaelis-Menten
 equation 100
 kinetics 40
 microarrays 78
 microreversibility 118
 model organism 130
 molecular interaction 5
 MS 88
 multi-enzyme complex 116
 multifunctional enzyme 116, 124, 125

N

NCBI taxonomy database 198
 nitrogenase 146, 147, 151
 NMR 76
 nomenclature 18, 190, 203
 non-stationary conditions 104

O

OBO 145
 ODE 129, 131
 ontology 143, 144
 Open Biology Ontologies 145

optimization problem 47
 optimum pH 35
 ordinary differential equation 129
Oryctolagus cuniculus 117

P

parameter estimation 59
 PCR 205
 PDB 144
 links 199
 phosphomevalonate kinase 177
 plant metabolism 80
 pleiotropic artefact 74
 porcine renin 92
 profile database 176
 progress-curve analysis 100
 protein 204
 biochemistry 73
 Data Bank 144
 function 177
 structural initiative 175
 structure 175
 -protein interactions 77
 proteome 75

R

radiometric assays 88
Rattus norvegicus 119
 reaction 121, 123
 condition 45
 kinetics 100
 regulatory
 control point 74
 network 73
 reticulocytes 104
 RNA profiling 73

S

Saccharomyces cerevisiae 7, 71, 72, 123
 SBML 132
 sensitivity analysis 58

silicon cell 130, 137
 single nucleotide polymorphism 176
 SMILES 189
 SNP 176
 spectrophotometric
 assay 75
 methods 87
 SQP method 45, 62
 stability data 197
 standardization 14, 130
Streptococcus pneumoniae 177
 substrate
 concentration 50
 specificity 193
 Support Vector Machine 200
Sus scrofa 123
 SwissProt 144, 199
 systematic chemical names 208
 systematic names 22
 systems biology 2, 5, 14, 71, 80, 99, 140,
 203, 208

W

WIT database 33

T

taurine metabolism 31
 taxonomy 145
Tetrahymena pyriformis 77
 theoreticians 71
 thermodynamic data 34
 thesaurus 193
 transcript pair 73
 transketolase 117, 120
Triticum aestivum 123

U

uniqueness problem 191
 universal assay mixture 41

V

VAST 178
 Vector Alignment Search Tool 178
