Proceedings

of the

3rd International Beilstein Workshop

on

Experimental Standard Conditions

of

Enzyme Characterizations

September 23rd – 26th, 2007

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 Symposium, September 23rd – 26th 2007, Rüdesheim, Germany.

Bibliographic information published by the *Deutsche Nationalbibliothek*.
The *Deutsche Nationalbibliothek* lists this publication in the *Deutsche Nationalbibliografie*; detailed bibliographic data are available in the Internet at http://dnb.ddb.de.

ISBN 978-3-8325-2038-0

Layout by: Hübner Electronic Publishing GmbH
Steinheimer Straße 22a
65343 Eltville
Cover Illustration by: SEIBERT MEDIA GmbH
Söhnleinstr. 8
65201 Wiesbaden

Printed by Logos Verlag Berlin GmbH
Comeniushof, Gubener Str. 47
10243 Berlin
http://www.logos-verlag.de

Preface

The almost complete sequencing of the genomes from numerous organisms paved the way for the development and application of new experimental and instrumental techniques which contribute to the understanding of complex biological pathways and networks by providing apparently endless opportunities to generate massive amounts of data. Cell machinery is currently envisaged as an inter-relationship of enzymes, proteins and chemical compounds. However, both a large number of metabolic pathways and enzymes even in well-described pathways still remain unknown. It is therefore necessary to develop further experimental and mathematical methods to reconstruct unknown parts of the networks, to identify genes for missing enzymes and to characterize the kinetic behaviour of those enzymes that have been identified.

The post-genomic era is also characterized by the concept of systems biology. This has gained significant momentum and metabolic research is now being conducted on an integrated and cross-disciplinary platform pulling together its resources from diverse fields such as mathematics, computational biology, bioinformatics, functional genomics and proteomics, and structural biology.

The enormous growth in the computation speed and data storage capability has fuelled new opportunities for both the accumulation of massive amounts of sequence, expression and functional data and the characterization, analysis and comparison of larger biological systems. However, as long as the data quality of the in-put and the resulting modelling data cannot be improved, the chances of success for this young discipline to escape from the verbally overused *–omics*-sciences are poor.

Systems level investigation of genomic and proteomic scale information requires incomparably higher demands for data quality than in previous decades. Truly integrated databases that deal with heterogeneous data need to be developed to be able to retrieve properties of genes, for kinetics of enzymes, for behaviour of complex networks and for the analysis and modelling of complex biological processes. One perspective of the output can be the investigation of cellular pathways involved in disease biology and targeted by newer molecular therapeutics. The understanding of these processes will assist the development of early diagnosis, prognosis and the prediction of response to individual therapies.

Despite the fast paced global efforts in biological systems research, the current analyses are limited by the lack of available systematic collections of comparable functional enzyme data. Besides its reliability, these data have to provide defined minimum experimental information, they must be available from the literature along with their accepted enzyme names, and must be as comprehensive as possible.

The STRENDA commission, founded on the 1st ESCEC meeting in 2003, has worked out a number of checklists which are intended to improve the quality of reporting enzyme data and thus to support the comparability of *inter alia* enzyme kinetics. The commission has also spent much time and effort in the creation of an electronic data submission system which allows authors to deposit their data and to provide an interaction record accession number that can be quoted in publications.

This 3rd ESCEC symposium, organized by the Beilstein-Institut together with the STRENDA commission, provided a platform to discuss the checklists (see also http://www.strenda.org/documents). Further suggestions regarding the checklists have been collected and discussed. Questions such as how to organize and store these massive data sets in standard and easily accessible forms have been asked an the first running draft of an data acquisition tool considering the STRENDA guidelines has been presented.

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

Frankfurt/Main, August 2008

Carsten Kettner
Martin G. Hicks

Contents

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

Functional Genomics in *Escherichia coli*: Experimental Approaches for the Assignment of Enzyme Function

Nina V. Stourman[1], Megan C. Wadington[1], Matthew R. Schaab[1], Holly J. Atkinson[2], Patricia C. Babbitt[3], Richard N. Armstrong[1*]

[1]Departments of Biochemistry and Chemistry, Center in Molecular Toxicology, and the Vanderbilt Institute of Chemical Biology, Vanderbilt University, Nashville, TN, 37232 – 0146, U.S.A.

[2]Program in Biological & Medical Informatics, University of California, San Francisco, CA 94158 – 2330, U.S.A.

[3]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry and California Institute for Quantitative Biosciences, University of California, San Francisco, CA 94158 – 2330, U.S.A.

E-Mail: *r.armstrong@vanderbilt.edu

Received: 11th February 2008 / Published: 20th August 2008

Abstract

A major challenge in biochemistry is to understand the functional genomics of organisms. This is a staggering problem when one considers the fact that almost 40% of the genes in one of the best-understood organisms in the biosphere, *Escherichia coli*, have no experimentally verified function. In this paper we address the challenge of, and criteria for, assigning protein function in the context of the glutathione (GSH) transferase paralogues encoded in the *E. coli* genome. The *E. coli* genome harbors genes encoding nine GSH transferase homologues including YliJ, YncG, Gst, YfcF, YfcG, YghU, SspA and YibF as well as the membrane-bound enzyme YecN. Amazingly, only one of these genes has a reasonably well-defined function and it does NOT encode a protein with GSH transferase activity but rather a transcription factor, stringent starvation protein A, SspA.

Introduction

The protein world, as we currently understand it, is composed in large part, of groups of proteins commonly called superfamilies that share similarities in sequence and three-dimensional structure but can diverge considerably with respect to biological function. One central question in modern biology is the extent of functional diversity that can be realized in a given protein superfamily. A different but related question is, how can we experimentally define the biological function(s) of all genes in a given organism? This has yet to be done for any single organism. In this paper we illustrate some experimental approaches that, when applied in parallel, are designed to reveal the functions of members of the GSH transferase superfamily in *E. coli*.

Glutathione is the predominant redox-active thiol in most aerobic organisms where it plays a fundamental role in metabolic, catabolic and redox chemistry. GSH transferases are enzymes that typically catalyze the addition of GSH to electrophilic acceptors as illustrated in Scheme 1 [1]. In microorganisms, these enzymes often participate in the catabolism of xenobiotic molecules [2 – 5]. The name, however, obscures the diverse impact that this group of proteins has in mammalian and microbial biology. In the last several years it has become apparent that members of this superfamily also perform other very diverse functions that are not at all related to the reaction illustrated in Scheme 1. These other functions include the regulation of transcription [6 – 10] and translation [11] and the intracellular transport of ions [12, 13].

Scheme 1: $GSH + R\text{-}X \rightarrow GSR + X^- + H^+$

The canonical or soluble GSH transferases are typically dimeric proteins where each subunit is composed of a N-terminal thioredoxin-like domain and a C-terminal α-helical domain as illustrated in Figure 1. The eight GSH transferase homologues encoded in the *E. coli* genome share these same structural characteristics based on available structures determined to date or inferred from sequence alignments. The *E. coli* paralogues are also defined by a consensus sequence of 17 residues, most of which are involved in the hydrophobic core of the thioredoxin domain [14]. Only two of the seventeen conserved residues are implicated in the binding of GSH, a fact that suggests that some of the paralogues may not bind or utilize GSH.

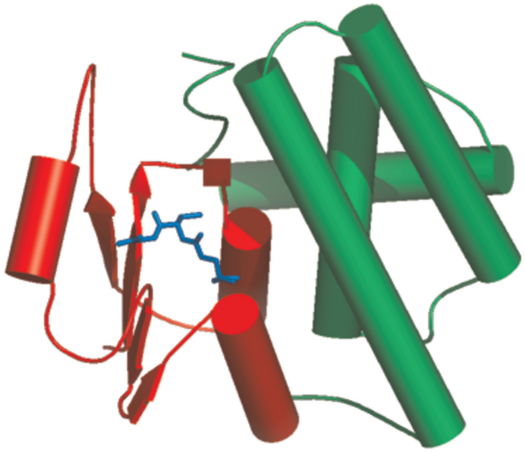


Figure 1. Diagram of the structure of a typical GSH transferase subunit. The all α-helical domain is shown in green and the thioredoxin domain is shown in red. Glutathione (shown in blue) is bound principally by the thioredoxin domain.

The question then arises, how can the biological functions of proteins with unknown function be divined? The intellectual and experimental approaches to this type of problem are numerous and must be used in concert, inasmuch as no single approach is likely to be definitive. Our combination of approaches includes; (i) analysis using informatics and genome context, (ii) response of gene expression to environmental stress, (iii) phenotypic response to gene knockouts, (iv) a search for protein partners, (v) structural biology, and of course (vi) functional assays of proteins. Other possible experimental avenues include metabolomics, proteomics and genetics. The point of this article is to illustrate how we are approaching this problem in the context of the GSH transferase superfamily in the biological context of *E. coli*.

The GSH transferase homologues encoded in the genome of *Escherichia coli* are quite diverse within the superfamily. A Cytoscape analysis [15] (Figure 2) of sequence similarity of both microbial and eukaryotic GSH transferase homologues indicates that at least five families are represented by the eight proteins including HSP26 (YibF, sspA); zeta, beta or tau (YfcF), beta (gst, YncG); phi or other (YliJ); and theta or other (YghU and YfcG). The diversity is remarkable and suggests a significant functional diversity as well. It is also remarkable that several nodes for the *E. coli* paralogues (YncG, YfcF, gst, and YghU) lie on the periphery of one main cluster and none are associated with the cluster that includes the alpha, mu, pi and sigma subfamilies that are common in eukaryotes.

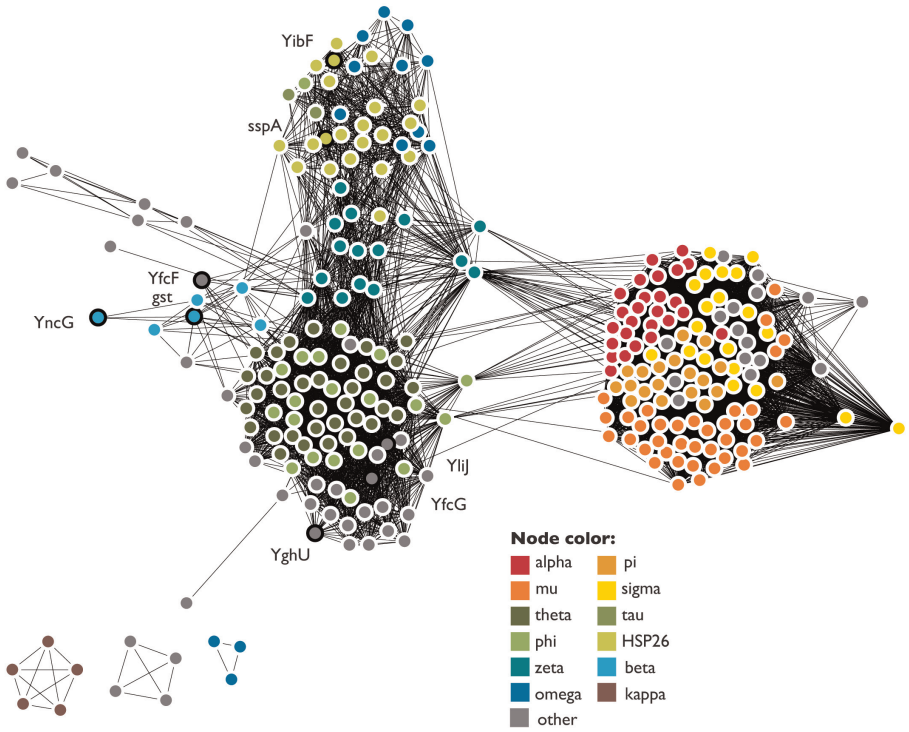


Figure 2. Sequence similarity network of 272 sequences longer than 80 residues from the SWISS-PROT GST superfamily, as well as proteins containing GST-like domains. For each pair of sequences, an edge is shown if the BLAST E-value for the pairwise alignment is better than 1×10^{-9}. Edges at this limiting E-value had a median residue identity of 25% over an alignment length of 180 amino acids. If two sequences are less similar, their nodes are not connected in this network. A thick black border denotes the *E. coli* GSH transferase paralogues. Node color indicates the family classification by SWISS-PROT.

Materials and Methods

Reduced glutathione (GSH), all buffer salts and other chemicals were obtained from commercial sources. Glutathionylspermidine (GspSH) was synthesized by incubating the C55A mutant of the *E. coli* glutathionylspermidine synthetase/amidase with GSH, ATP and spermidine. The GspSH was purified by ion exchange chromatography. The details of the synthesis will be published at a later date.

Chromosomal gene disruptions or knockouts in *E. coli* genome were made by the Wanner method [16] as follows. *E. coli* BW25113-*pKD46* cells were grown overnight at 30 °C in 5 mL SOC media containing ampicillin (100 µg/mL). The culture was diluted 100-fold into 20 mL of fresh SOB media supplemented with 1 mM arabinose and grown at 30 °C to

OD_{600}= 0.6. Cells were made electrocompetent by four consecutive washes with decreasing volumes of ice-cold 10% sterile glycerol. The final pellet was suspended in 40 µL of cold 10% sterile glycerol, flash-frozen and stored at – 80 °C. DNA fragments used for the gene disruption were amplified by PCR from the pKD3 vector carrying the chloramphenicol resistance gene. Each primer consisted of 40-nt homologous to the gene to be eliminated and 20-nt priming sequence for pKD3. Electroporation was done in 1 mm cuvettes using ElectroCell Manipulator BTX ECM 399 (BTX Harvard Apparatus, Holliston, MA) with charging voltage 1.4 kV and 5 ms pulse length. Electroporation cuvettes were chilled on ice before the addition of 40 µL of electrocompetent cells mixed with 2 µL of gel-purified PCR product containing 10 to 30 ng of DNA. Immediately after the pulse the cells were transferred to culture tubes containing 960 µL of SOC media and incubated at 37 °C for 1 hr. A 400 µL aliquot of the culture was plated on the LB/chloramphenicol plates. Plates were kept in the incubator at 37 °C for up to 24 hr allowing colonies to grow. The deletion of the gene was confirmed by colony PCR using specific gene primers. The primers for the *yghU* knockout using pKD3 were:

Forward: ATACTTATCA GCCCGCGAAA GTCTGGACGT GGGATAAATC *GTGTAGGCTGGAGCTGCTTC*
Reverse: TGACGCTTAT CTTCCGTATT CGTCTCGAAA TCACTGGCGT *CATATGAATATCCTCCTTAGT*
Specific primers for *yghU* (amplicon size 318 bp) were:
Forward: TATCTGGCGAGAAATTTGG
Reverse: CTCAGCGGCATCATACACAC
Specific primers for *gss* knockout using pKD3 were:
Forward: CAAAGGAACG ACCAGCCAGG ATGCCCCGTT CGGGACATTA *CATATGAATATCCTCCTTAGT*
Reverse: CTCTTTTTTG ATGACCAGTG ATTCATCACC GCGCAAACAC *GTGTAGGCTGGAGCTGCTTC*
Specific primers for *gss* (amplicon size 353 bp) were:
Forward: AAGTCCGTATTGCGGAACAG
Reverse: GGCACTCTCGGTAATGGTGT

Gene expression levels of the GSH transferase homologues were quantified by the reverse transcriptase polymerase-chain-reaction (RT-PCR) as follows. Total RNA was purified from 5 mL of *E. coli* cells with RNeasy Mini Kit (Qiagen Inc,, Valencia, CA) according to the manufacturer's protocol with an additional RNA clean-up step after the treatment with DNA-free kit (Ambion, Foster City, CA). After this procedure, the RNA was essentially free of genomic DNA. Total RNA was quantified by measuring absorbance at 260 nm. RT-PCR reactions were performed using SuperScript II reverse transcriptase from Invitrogene (Carlsbad, CA) and random hexamer primers following the manufacturer's protocol. For each reaction 1 µg of purified RNA was used. cDNA obtained in the first step was used for the following PCR reactions with the specific gene primers which was carried at 55 °C in the

linear range of amplification (22 – 25 cycles). A 10 µL aliquot from each reaction was run on 1.2% agarose gel containing ethidium bromide. The gel images were acquired with Molecular Imager Gel Doc XR System (BioRad) and analyzed with Quantity One 1-D analysis software and Adobe Photoshop.

Thiol analysis of *E. coli* was obtained as follows. An overnight culture grown of BW251137 in LB at 37 °C was diluted 100-fold into fresh LB or MM9 medium containing 0.4% glucose. For aerobic conditions cultures were grown in the shaker at 37 °C to OD 0.6 for exponential growth or overnight for stationary phase. For anaerobic conditions the cells were grown in the tubes with tight-screw caps filled with the media to the top and harvested without opening the tube. Cells were harvested by centrifugation at 6,000 g for 10 min. The pellet was washed with ice cold PBS, resuspended in ice cold 10% TCA, incubated for 10 min on ice and centrifuged at 10,000 x g for 10 min. For derivatization of thiols, 5 µL of the supernatant was mixed with 50 µL of 100 mM potassium phosphate buffer, pH 7.0 and 5 µL of 0.5 mM solution of monobromobimame in acetonitrile. After incubation for 40 min in the dark, 10 µL of the reaction mixture was injected into C 18 reverse phase HPLC column for analysis with a Varian Analytical Instruments (Walnut Creek, CA) HPLC system equipped with a Dynamax (Rainin Instrument Company, Inc.,Woburn, MA) fluorescence detector tuned to excitation at 380 nm and emission at 480 nm. The elution buffer was 140 mM ammonium acetate, pH 5.0 with the gradient of acetonitrile (15 – 25%) over 20 min.

Thiol Substrates in *Escherichia coli*

One of the first issues in elucidating the functions of the GSH transferase paralogues in *E. coli* is whether they interact with or utilize GSH. This question is complicated by the fact that there are two major forms of GSH in *E. coli*; GSH itself and glutathionylspermidine (GspSH), a condensation product of GSH and spermidine [17]. It is known that GSH is the predominant thiol under aerobic growth in log phase. However, in late stationary phase and particularly under anaerobic conditions most of the thiol is found as GspSH [18].

GspSH is formed by the enzyme glutathionylspermidine synthetase/amidase (GSS), which is a bifunctional enzyme that catalyzes the ATP-dependent condensation of the glycylcarboxylate of GSH with N1 (or the short arm) of spermidine as illustrated in Figure 3. The enzyme also catalyzes the hydrolysis of GspSH to give GSH and spermidine. The two opposing activities of the enzyme obviously need to be regulated but the mechanism of that regulation is not known [19, 20]. The fact that the *yghU* gene is located adjacent to the gene encoding GSS led us to the initial hypothesis that YghU might be a protein that regulates GSS activity [14]. A number of experiments using purified GSS and YghU revealed that YghU had no detectable influence on the activity of GSS either in the forward or reverse reactions.

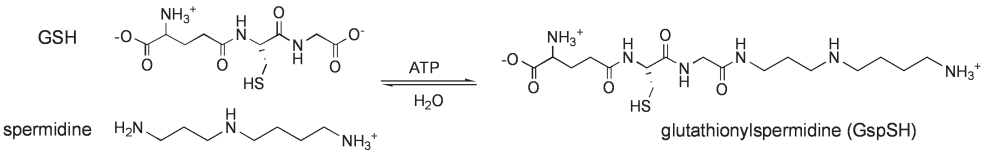


Figure 3. Reactions catalyzed by glutathionylspermidine synthetase/amidase (GSS).

Gene Knockouts

In order to examine the effect of YghU on GSS in a cellular assay we disrupted each gene in separate experiments. The loss of the gene is illustrated in Figure 4 where the absence of the *yghU* or the *gss* message is clear from the RT-PCR experiment as compared to the wild–type organism. The effect of the disruptions on the levels of GSH and GspSH under anaerobic conditions is illustrated in Figure 5. As anticipated from the published literature [18], in minimal media under anaerobic conditions, GspSH was the predominant thiol (81%). The *yghu* gene knockout decreased the amount of GspSH to 65% of the thiol total which is essentially within the error of the experimental measurement. Disruption of the *gss* gene essentially eliminated the GspSH (< 5%) from the thiol pool. The conclusion from these results is that the YghU protein does not regulate the GSH/GspSH tone in the cell to a significant extent under these conditions.

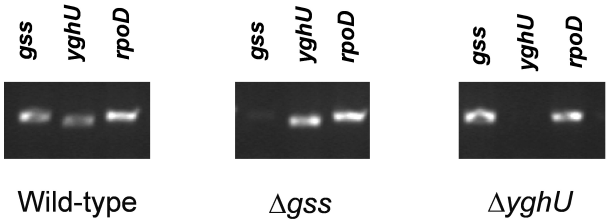


Figure 4. Reverse transcriptase polymerase chain reaction (RT-PCR) analysis of the messenger RNA for glutathionylspermidine synthetase/amidsase (gss) and YghU in wild-type *E. coli* cells BW25113 (left), *E. coli* cells BW25113 (D*gss*) where the *gss* gene has been disrupted (middle) and *E. coli* cells BW25113 (D*yghU*) where the *yghU* gene has been disrupted (right). Note the absence of messenger RNA in the two knockouts.

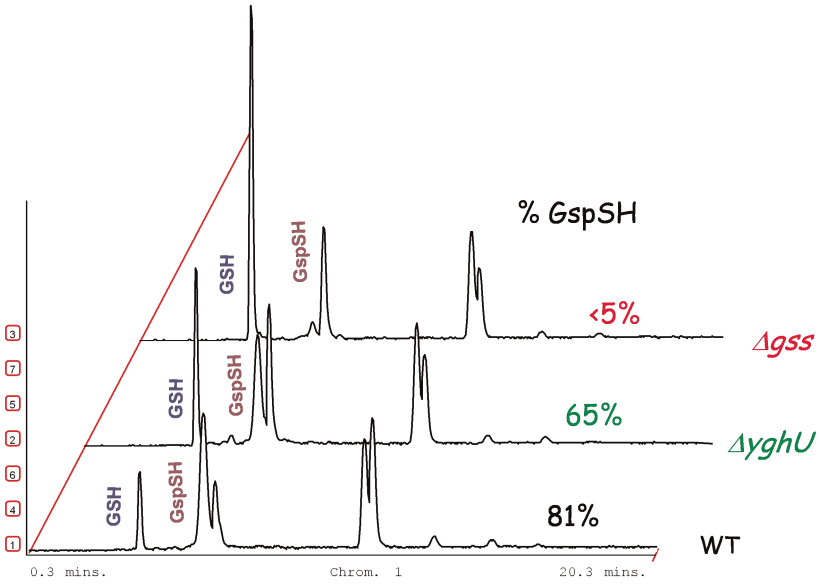


Figure 5. Analysis of the thiol content of *E. coli* cells BW25113 grown in M9 minimal media to stationary phase under anaerobic conditions. The bottom trace (wt) is for wild-type cells. The middle (D*yghU*) and top (D*gss*) traces are results for the *yghU* and *gss* gene knockouts, respectively. The unlabeled peaks are due to decomposition products of the fluorescence reagent.

Gene Expression Levels

The measurement of gene expression levels offers another view of how proteins influence the biology of a cell under particular environmental conditions. The conditions can represent points in the growth of the organism, nutrient status, or physical or chemical stress. Gene expression can be measured either by mRNA levels in a cell or by direct measurement of protein expression levels. The semi-quantitative measurement of mRNA levels by RT-PCR is a cost effective way of examining gene expression levels. Figure 6 illustrates the gene expression levels for all eight GSH transferase paralogues and glutathionylspermidine synthetase/amidase under normal aerobic growth conditions.

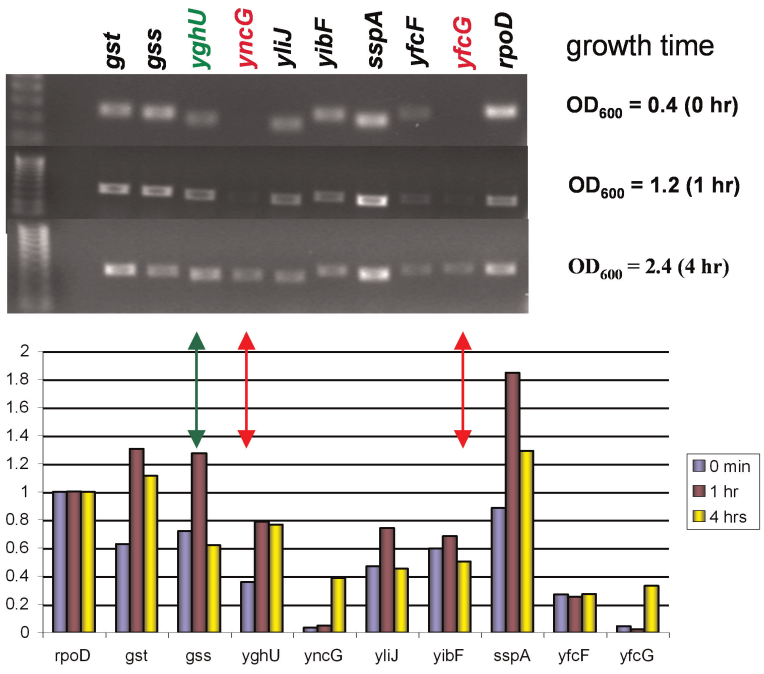


Figure 6. mRNA levels for *gss* and the eight GSH transferase paralogues as a function of growth time in LB media and determined by RT-PCR analysis. The bottom panel shows the mRNA levels normalized to the *rpoD* message. Note that two of the GSH transferase paralogues, *yncG* and *yfcG* (red arrows), exhibit large = 10-fold increases in message in late stationary phase while the *yghU* (green arrow) gene exhibits robust expression through out growth.

The normalized gene expression levels reveal only modest (= 2-fold changes) in gene expression as a function of growth time for most of the genes. The two exceptions are the *yncG* and *yfcG* genes that exhibit 10 to 30-fold increases in expression in late stationary phase. Three interesting observations can be made from these data. The first is that the increased expression level of both genes coincides with the elevation of GspSH in late stationary phase suggesting that there may be a connection between GspSH and the YncG and YfcG proteins. The second observation is that YncG and YfcG are distinctly different with respect to the GSH transferase families to which they belong (Figure 2). The enhanced expression of these two genes in late stationary phase has not been reported in microarray data. In contrast, *yfcG* has been reported to be down-regulated by the absence of the transcriptional regulator Fis (factor for inversion stimulation) in stationary phase [21].

If the *yncG* and *yfcG* genes are expressed in late stationary phase when the synthesis of GspSH takes place, it might be expected that the gene products would preferentially interact with GspSH as opposed to GSH. In preliminary work, we have found, by fluorescence

titration, that the YfcG protein binds GspSH 10-fold more tightly (K_d= 29 ± 7 μM) than it does GSH (K_d = 329 ± 6 μM). It would then appear that YfcG has a significant preference for binding GspSH and we conclude that the protein plays some role in the biochemistry of GspSH.

Criteria for the Assignment of Enzyme Function

The criteria for the assignment for protein or enzyme function are as varied as the function of any given protein is complicated. At a minimum, a protein needs to be characterized with respect to what other molecules it interacts with, including small molecules or substrates and other macromolecules. The latter can be accomplished with various types of pull-down assays using the protein of interest as bait. Ideally, the molecular interactions with large or small molecules should be characterized in as much structural detail as possible by X-ray crystallography or NMR spectroscopy.

The temporal or environmental influence on gene expression levels is also often a valuable piece of information as demonstrated above. The viability or sensitivity of an organism to gene knockouts can also reveal essential clues as to the biological role of a particular protein. These clues can be detected by with a variety of techniques including metabolomics (the appearance or loss of metabolites), proteomics (the appearance or loss of specific proteins) and genetics (the interaction of one gene with another). Needless to say, the stringency of the criteria for defining enzyme function can vary enormously from simply elucidating what kind of reaction an enzyme catalyzes to more global questions as to why a particular reaction is important to a given organism under specific circumstances.

Acknowledgement

This work was supported by National Institutes of Health Grants R01 GM030910, T32 ES 007028 and T32 GM008320, P30 ES 000267 and R01 GM060595.

References

[1] Armstrong, R.N. (1995) Structure, Catalytic Mechanism and Evolution of the Glutathione Transferases. *Chem. Res. Toxicol.* **10**:2 – 18.

[2] Vuilleumier, S., Pagni, M. (2001) The elusive roles of bacterial glutathione S-transferases: New lessons from genomes. *Appl. Microbiol. Biotechnol.* **58**:138 – 146.

[3] Oakley, A.J. (2005) Glutathione transferases: new functions. *Curr. Opin. Struct. Biol.* **15**:716 – 723.

[4] Stourman, N.V., Rose, J.A., Vuilleumier, S., Armstrong, R.N. (2003) Catalytic mechanism of dichloromethane dehalogenase from *Methylophilus* sp. strain DM 11, *Biochemistry* **42**:11048 – 11056.

[5] Thompson, L.C., Ladner, J.E., Codreanu, S., Harp, J., Gilliland, G.L., Armstrong, R.N. (2007) 2-Hydroxychromene-2-carboxylic acid isomerase: a Kappa class glutathione transferase from *Pseudomonas putida*. *Biochemistry* **46**:6710 – 6722.

[6] Williams, M.D., Ouyang, T.X., Flickinger, M.C. (1994) Starvation induced expression of SspA and SspB: The effects of a null mutation in sspA on *Escherichia coli* protein synthesis and survival during growth and prolonged starvation. *Mol. Microbiol.* **11**:1029 – 1060.

[7] Hansen, A.M., Qui, Y., Yeh, N., Blattner, F.R., Durfee, T., Jin, D.J. (2005) SspA is required for acid resistance in stationary phase by down regulation of H-NS in *Escherichia coli. Mol. Microbiol.* **56**:719 – 734.

[8] Hansen, A.M., Gu, Y., Li, M., Andrykovitch, M., Waugh, D.S., Jin, D.J., Ji, X. (2005) Structural basis for the function of stringent starvation protein A as a transcription factor. *J. Biol. Chem.* **280**:17380 – 17391.

[9] Bai, M., Zhou, J.M., Perett, S. (2003) The yeast prion protein Ure2 shows glutathione peroxidase activity in both native and fibrillar forms. *J. Biol. Chem.* **279**:50025 – 50030.

[10] http://www.ncbi.nlm.nih.gov/pubmed/16275904?ordinalpos=32&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum (2005) The transduction of the nitrogen regulation signal in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.* **102**:16537 – 16538.

[11] Koonin, E.V., Mushegian, A.R., Tatusov, R.L., Altschul, S.F., Bryant, S.H., Bork, P., Valencia, A. (1994) Eukaryotic translation elongation factor 1μa; contains a glutathione transferase domain-Study of a diverse, ancient protein superfamily using motif search and structural modeling. *Protein Sci.* **3**:2045 – 2054.

[12] Dulhunty, A., Gage, P., Curtis, S., Chelvanayagam, G., Board, P. (2001) The Glutathione Transferase Structural Family Includes a Nuclear Chloride Channel and a Ryanodine Receptor Calcium release Channel Modulator. *J. Biol. Chem.* **276**:3319 – 3323.

[13] Harrop, S.J., DeMaere, M.Z., Fairlie, W.D., Reztsova, T., Valenzuela, S.M., Mazzanti, M., Tonini, R., Qui, M.R., Jankova, L., Warton, K., Bauskin, A.R., Wu, W.M., Pankhurst, S., Campbell, T.J., Breit, S.N., Curmi, P.M.G. (2001) Crystal structure of the soluble form if the intracellular chloride Ion channel CLIC 1 (NC 27) at 1.4 Å resolution. *J. Biol. Chem.* **276**:44993 – 45000.

[14] Rife, C.L., Parsons, J.F., Xiao, G., Gilliland, G.L., Armstrong, R.N. (2003) Conserved structural elements in glutathione transferase homologues encoded in the genome of *Escherichia coli*. *Proteins: Struct. Func. Genetics* **53**:777 – 782.

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

[16] Datsenko, K.A., Wanner, B. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U.S.A.* **97**:6640 – 6645.

[17] Tabor, H., Tabor, C.W. (1975) Isolation, characterization and turnover of glutathionylspermidine from *Escherichia coli*. *J. Biol. Chem.* **250**:2649 – 2654.

[18] Smith, K., Borges, A., Ariyanayagam, M.R., Fairlamb, A.H. (1995) Glutathionylspermidine metabolism in *Escherichia coli*. *Biochem. J.* **312**:465 – 469.

[19] Bollinger, J.M. Jr., Kwon, D.S., Huisman, G.W. Walsh, C.T. (1995) Glutathionylspermidine metabolism in *Escherichia coli*. *J. Biol. Chem.* **270**:14031 – 14041.

[20] Lin, C.-H., Kwon, D.S., Bollinger, J.M. Walsh, C.T. (1997) Evidence for a glutathionyl-enzyme intermediate in the amidase activity of the bifunctional glutathionylspermidine synthetase/amidase from *Escherichia coli*. *Biochemistry* **36**:14930 – 14938.

[21] Bradley, M.D., Beach, M.B., Jason de Koning, A.P., Pratt, T.S., Osuna, R. (2007) Effects of Fis on *Escherichia coli* gene expression during different growth stages. *Microbiology* **153**:2922 – 2940.

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

Catalysis at the Membrane Interface: Cholesterol Oxidase as a Case Study

Nicole S. Sampson* **and Sungjong Kwak**

Department of Chemistry, Stony Brook University, Stony Brook, NY 11794 – 3400, U.S.A.

E-Mail: *nicole.sampson@stonybrook.edu

Received: 21st February 2008 / Published: 20th August 2008

Abstract

Interfacial enzymes present additional challenges in their study compared to enzymes with soluble substrates. Cholesterol oxidase is an interfacial enzyme that transiently associates with lipid membranes to convert cholesterol to cholest-4-en-3-one. As a case study to exemplify the issues that should be considered, we describe our structural and mechanistic understanding of cholesterol oxidase kinetic activity based on X-ray crystal structures and kinetic analysis.

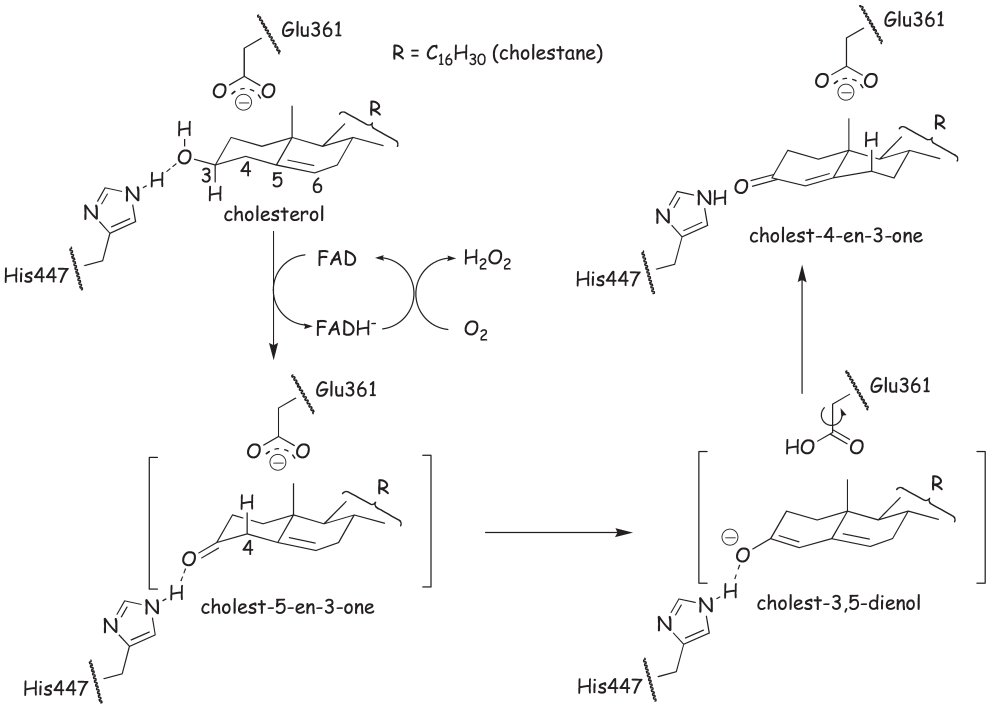
Introduction

Interfacial enzymes are water-soluble enzymes that catalyze reactions with membrane-soluble substrates. Kinetic characterization of these enzymes is made more complex by the necessity to consider the role of the interface and interactions with the interface in assessing the catalytic activity. Moreover, the interface can change during catalysis, further complicating the kinetic analysis. The interface used in assaying the enzyme influences the apparent substrate specificity measured. Thus, the assignment of a physiological role for an enzyme is dependent on the interface employed in enzymatic assays.

Cholesterol oxidase is one such water-soluble enzyme that is catalytically active at the membrane interface from which cholesterol, the substrate, is accessed. As a case study, we present work from our laboratory that characterizes what happens at the membrane interface. We delineate the kinetic issues in reporting the catalytic activity of such an enzyme.

Cholesterol Oxidase

The history of cholesterol oxidase derives from the discovery over 50 years ago that some actinomycetes can utilize cholesterol as a carbon source [8, 9]. They are believed to break down the side-chain and the ring of cholesterol to acetyl CoA and propionyl CoA through a multi-step process. The enzyme which catalyzes the first step is cholesterol oxidase. Cholesterol oxidase was isolated in bio-panning experiments when there was a search for an enzyme to use in clinical serum cholesterol assays [10 – 12].



Scheme 1: The reaction catalyzed by cholesterol oxidase. Active site residues are shown schematically.

The chemistry that is catalyzed by cholesterol oxidase occurs in one active site (Scheme 1). Cholesterol is oxidized to cholest-5-en-3-one by the flavin cofactor. The reduced cofactor is recycled by oxygen to form hydrogen peroxide. This product is the basis of the serum cholesterol assays, because hydrogen peroxide can be coupled to colorimetric assays using horseradish peroxidase. However, the cholest-5-en-3-one intermediate is not particularly stable. It is susceptible to radical oxygenation, and forms cholest-4-en-6-hydroperoxy-3-one that disproportionates to cholest-4-en-3,6-dione and cholest-4-en-6-hydroxy-3-one. Thus, the cholest-5-en-3-one is isomerized to cholest-4-en-3-one, the α,β-unsaturated ketone before being released from the enzyme [13].

The identity of general acids and bases to help catalyze the reaction may be surmised upon inspection of the active site model with a dehydroepiandrosterone bound [3]. Histidine 447 and asparagine 485 hydrogen bond to the alcohol of the substrate helping to position the steroid relative to the flavin cofactor (Scheme 1). Glutamate 361 is poised over the β-face of the steroid, to act as a base in the isomerization reaction.

Mutagenesis of glutamate 361 to glutamine turned the oxidase/isomerase into an oxidase-only enzyme [13]. The E361Q enzyme no longer isomerizes the intermediate, cholest-5-en-3-one. However, it is released from the mutant enzyme at a catalytically competent rate. The turnover of cholesterol is only 30 times slower than that of wild-type enzyme (Table).

Table. Catalytic parameters for wild-type and mutant cholesterol oxidases.

Enzyme	k_{cat} (s^{-1})	K_m^{app} (μM)[a]	Product formed	reference
Wild type	45 ± 3	3.2 ± 0.2	cholest-4-en-3-one	13
E361Q	1.4 ± 0.2	5.3 ± 1.5	cholest-5-en-3-one	13
H447E/E361Q	0.0015	n.d.[b]	n.d.	14

[a]Rates were assayed in triton X-micelles. K_m^{app} is the apparent Michaelis-Menten constant that includes a micelle binding term. [b]n.d.: not determined.

Mutation of histidine 447 in conjunction with glutamate 361 blocks oxidation as well as isomerization and provides a ‘‘dead’’ mutant that turns over cholesterol some 30,000-fold times slower than wild type [14]. The enzyme still folds correctly based on its behavior in solution [14] as well as its X-ray crystal structure (A. Vrielink, personal communication). This dead enzyme is an important tool for the study of an important aspect of catalysis by cholesterol oxidase: catalysis at the membrane interface.

In order to understand catalysis at the membrane interface, it is important to look at the three-dimensional crystal structures that have been solved by Prof Alice Vrielink and her laboratory. Several structures have been solved of wild-type and mutant enzymes. Some of the structures are at sub-Ångstrom resolution and allow hydrogen bonding within the enzyme to be visualized directly [3, 15]. However, those structures are not the focus for understanding interfacial catalysis. What is important is to examine how the steroid binds to the enzyme and to consider the changes that must occur upon binding to the membrane.

The unliganded structures reveal a deep, long pocket adjacent to the isoalloxazine ring of the flavin suitable for binding a steroid substrate. Dehydroepiandrosterone was used to obtain a substrate bound structure because the limited solubility of cholesterol precluded getting crystals in the presence of cholesterol. The steroid binds in the deep pocket as expected (Fig. 1A). The surprising observation is that the steroid is completely solvent inaccessible when bound (Fig. 1B). The protein encapsulates the A-D ring of the steroid. Dehydroepiandrosterone is of course lacking the 8-carbon tail of cholesterol.

If the larger steroid were to be bound it is not clear exactly how the protein would accommodate the steroid. What is proposed from inspection of the structure is that one or more loops of the protein must open at the membrane surface to allow sterol exit from the membrane and entry into the enzyme (Fig. 1C). The 8-carbon isoprenyl tail of cholesterol would pack with the loops and prevent them closing completely. The amphipathic nature of the loops would allow them to pack with the hydrophobic sterol on their inside face, and more polar headgroups of the lipid bilayer on their outside face. Our model of how the enzyme works is that it sits on the surface of the membrane and the loops provide a hydrophobic pathway for the substrate to partition from the membrane into the active site of the enzyme.

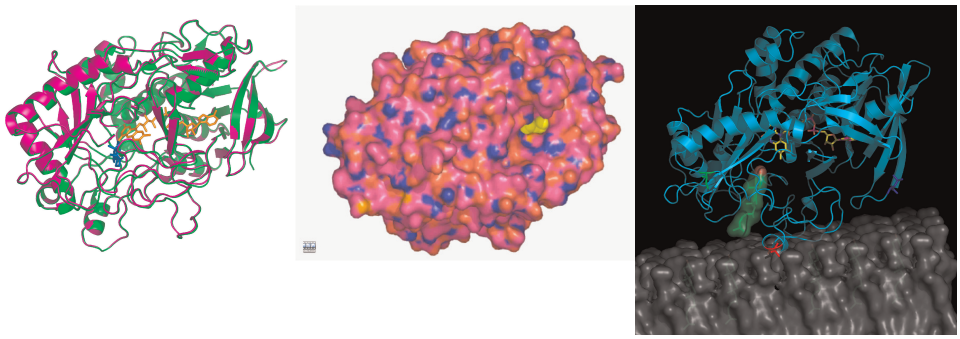


Figure 1. Cholesterol oxidase structure. (A) Ribbon diagram with steroid bound in active site (green) overlaid with unbound structure (magenta) [1 – 3]. (B) Solvent accessible surface on steroid-bound structure shown in A. Residues are colored by polarity: red, acidic; blue, basic; magenta, all other residues; yellow, flavin. (C) Model for how enzyme binds to the membrane interface. The loops that cover the active site have been modeled into an open conformation. The coordinates for the bilayer were obtained from Heller *et al.* [7].

We asked the question whether the formation of an enzyme-membrane complex results in perturbation of the membrane. This question was inspired by the Monsanto discovery that cholesterol oxidase is the biologically active component of bacterial fermentation broths that lyses boll weevil larval gut endothelial cells [16].

Addition of 10 µg/mL cholesterol oxidase to the larval feed results in disruption of the endothelial cell membranes. Using the active site mutants described above and vesicles that have a self-quenching dye encapsulated, we determined that leakage of vesicle contents only occurs if the cholesterol in the membrane is converted to cholest-4-en-3-one. That is, the chemical changes in the membrane catalyzed by cholesterol oxidase cause membrane structural changes rather than the physical interaction of the enzyme with the membrane. Our observation is consistent with what is known about cholesterol and the fluid phases of membranes. Cholesterol mixed with liquid-disordered phase phospholipids promotes order-

ing of the membrane to form a liquid-ordered phase. In contrast, mixing of cholest-4-en-3-one with liquid phase phospholipids maintains the liquid-disordered state [17]. This order-disorder effect occurs in both model membranes and in cell membranes.

Kinetics at the Membrane Interface

We used model membranes to establish how sensitive cholesterol oxidase activity is to membrane structure and lipid phase. We asked the question what are the relative catalytic activities with different membranes. Ultimately, the answer to this question is important for understanding the identity of the physiological substrate.

To address these questions, we used the binary phase diagram of dipalmitoylphosphatidylcholine (DPPC) and cholesterol as a starting point (Fig. 2) [4]. In the case of DPPC, the melting temperature for the gel (solid) phase to liquid-disordered phase transition is 41 °C in the absence of cholesterol. Above 30 mol% cholesterol, the phase transition is lost and the lipid phase is liquid-ordered above and below the DPPC melting temperature. In between 5 and 30 mol% cholesterol the gel phase is in coexistence with the liquid-ordered phase below the T_m, and the liquid-ordered and liquid-disordered phase coexist above the T_m.

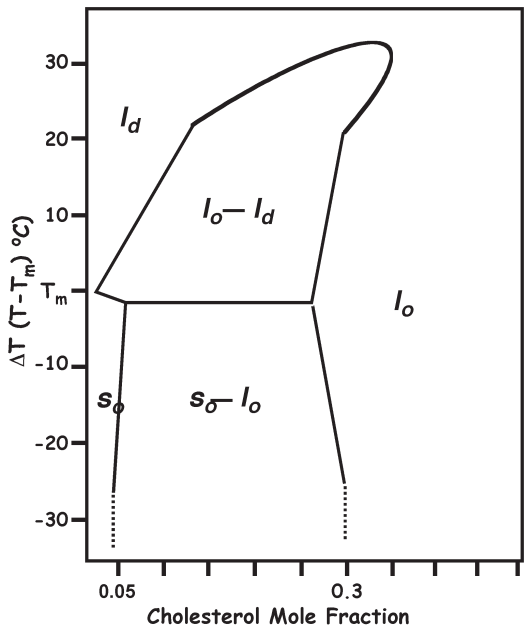
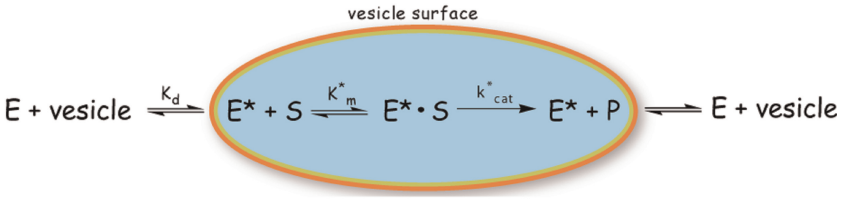


Figure 2. Binary phase diagram of DPPC:cholesterol adapted from Sankaram and Thompson [4]. The phase transition between s_o and l_d corresponds to the T_m of a lipid. For DPPC, the s_o to s_o-l_o transition is at 5 mol% cholesterol, and the s_o-l_o to l_o transition at 30 mol% cholesterol.

How does one measure the kinetics for an interfacial enzyme? Remember that the enzyme is soluble, but the substrate is a component of the membrane. The first step that must occur is association of the enzyme with the membrane surface (Scheme 2). The alternative is for the enzyme to wait for the substrate to dissociate from the membrane and then to bind the substrate from solution. The rate of cholesterol desorption has been measured for many different types of lipid bilayers. This rate is approximately 10^5 times slower than the turnover rate of the enzyme. Therefore, we conclude from a kinetic argument, that the enzyme must associate with the membrane in order for catalysis to occur. Moreover, measuring the change in intrinsic tryptophan fluorescence can follow the binding of the enzyme to the membrane surface [5]. Use of the catalytically inactive mutant H447E/E361Q enables binding to a substrate-containing vesicle to be measured [14]. Cholesterol oxidase binding to liquid-phase membranes shows little dependence on lipid composition [5, 18].



Scheme 2: Paradigm for determining interfacial steady-state rate constants. E: free enzyme; E*: membrane-bound enzyme; S: substrate in the membrane; P: product in the membrane; k_{cat}*: interfacial first-order rate constant; K_m*: interfacial Michaelis constant in units of mole fraction.

The next step is to measure the interfacial Michaelis-Menten constants. One common way of doing this is to measure the entire reaction progress curve and fit the integrated Michaelis-Menten equation. Recall that cholesterol orders the membrane and cholest-4-en-3-one disorders the membrane. Therefore, as you form more and more product, the structure of the membrane changes, and in fact the initial velocity of the catalyzed-reaction gets faster despite the decrease in mole fraction of substrate [5]. The net consequence is that we have to use initial velocities in order that we measure the rate corresponding to the initial, known, structure of the membrane.

The initial velocity measured depends on the fraction of enzyme that is actually bound to the vesicle surface. There are two substrate variables; the total concentration of lipid added which is proportional to number of vesicles, and the mole fraction of cholesterol in the vesicles. Increasing the concentration of vesicles pushes the first equilibrium to the right until all enzyme is bound to the vesicle surface. The initial velocity when all enzyme is bound (V_i^{sat}) is measured for a series of unilamellar vesicles of fixed size (prepared by extrusion) with varying mole fractions of cholesterol. The V_i^{sat} is then plotted versus mole

fraction of cholesterol and fit to the Michaelis-Menten equation in units of mole fraction. The rate dependence on mole fraction of cholesterol is expected to be hyperbolic, like an ‘‘ordinary’’ enzyme, as long as the substrate phase is not changing with mole fraction.

We first measured the initial velocities for cholesterol mixed with dioleoylphosphatidylcholine (DOPC), a lipid that is in the liquid-disordered phase regardless of mole fraction cholesterol. However, a hyperbolic dependence on cholesterol mole fraction was not observed (Fig. 3). That is, within the range of experimentally achievable mole fractions of cholesterol, the enzyme could not be saturated with substrate. The mole fraction dependence of V_i^{sat} was essentially linear. However, at 30 mol% cholesterol, the linear dependence of the rate has a discontinuity and above 30 mol% the slope is doubled. This increase in enzymatic activity above 30 mol% is consistent with the increase in chemical activity or potential that is observed in model membranes [19].

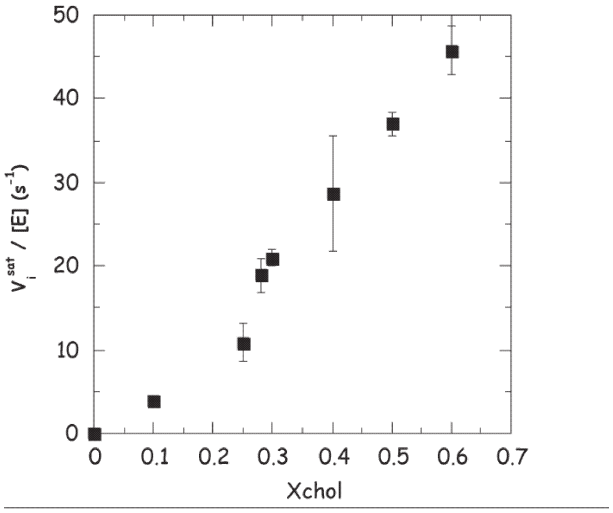


Figure 3. Detection of liquid-disordered cholesterol using cholesterol oxidase in a single phase vesicle composed of DOPC and varying mole fractions of cholesterol. The V_i^{sat} is the initial velocity of cholesterol oxidase turnover when all the enzyme is bound. The initial velocities were measured using 100 nm unilamellar vesicles at 31 °C. The error bars are standard deviations of three independent measurements. Adapted from Ahn and Sampson with a correction for concentration of active enzyme [5].

A similar lack of saturation was observed for vesicles composed of cholesterol mixed with dipalmitoylphosphatidylcholine (DPPC). Thus, we can only report k_{cat}^*/K_m^*. At 30 mol% cholesterol in the DPPC/cholesterol vesicles, the phase transitions to liquid-ordered. If we compare the k_{cat}^*/K_m^* for liquid-disordered membranes to liquid-ordered membranes in the same chemical potential regime, we observed that k_{cat}^*/K_m^* is two-fold slower with liquid-ordered membranes. However, if the DPPC is replaced with sphingomyelin (SM), the rate

drops more than 40-fold. Although DPPC/cholesterol is considered to be in the same phase as SM/cholesterol, the enzyme is sensitive to the precise packing of cholesterol with the lipids in the membrane. We surmise that the free energy of cholesterol in a sphingomyelin membrane is lower than in a DPPC membrane. It should be noted that cholesterol is predominantly localized with sphingomyelin in a cell membrane, and cholesterol in this environment is the worst substrate for the enzyme so far studied. Thus, the enzyme is specific for cholesterol that is in a low abundance form: liquid-disordered and high chemical potential.

In biphasic vesicles prepared with DOPC, SM and cholesterol, this rate difference elucidates the partitioning of cholesterol between the two phases (Fig. 4). At mole fractions less than 40 – 45% cholesterol, the phase is a liquid-ordered/disordered coexistence region. Above 40 – 45% cholesterol, the phase is liquid disordered [6]. We observed very little enzymatic activity below 45 mol% cholesterol, consistent with cholesterol portioning preferentially into the liquid-ordered phase. As the mole fraction increases, a steep increase in catalytic activity is observed. This increase correlates with the phase change to liquid-disordered. The shape of the activity dependence on mole fraction cholesterol indicates that despite the coexistence of liquid-ordered and liquid-disordered phases, the cholesterol resides primarily in the liquid-ordered region.

This experiment demonstrates the utility of cholesterol oxidase for this type of measurement. Many methods rely on microscopic observation of membranes that does not reveal which molecules are in which phase, or fluorescence spectroscopy that introduces reporter molecules. Cholesterol oxidase directly reports on the partitioning of cholesterol. However, it is not straightforward to quantitate precisely how much cholesterol is in the liquid-disordered region of coexisting phases at low mole fractions.

Substrate Specificity

What is substrate specificity for an interfacial enzyme? If one looks at the original work of Uwajima, who first isolated and characterized the *Brevibacterium sterolicum* (now *Rhodococcus equi*) cholesterol oxidase, the initial rate using cholesterol as a substrate is 10-fold higher than for sitosterol or stigmasterol, plant sterols [20, 21]. Consequently, the enzyme was named a cholesterol oxidase. However, these assays were run in triton X-100 detergent micelles. When we measured k_{cat}^*/K_m^* in liquid-disordered vesicles with 25 mol% cholesterol/palmitoyl, oleoylphosphatidylcholine, the specificity constant is the same for all three steroids [22]. In model membranes, cholesterol oxidase is not more specific for animal sterols over plant sterols.

In general, we know that detergent micelles are not physiologically relevant. Relative activities measured in micellar systems may not reflect relative substrate specificities under cellular conditions. The difficulty is identifying the correct physical state of potential sub-

strates to be utilized in assays of substrate specificity. If the activity of cholesterol oxidase had first been measured in vesicles of cholesterol and sphingomyelin, the predominant form of cellular cholesterol, turnover may never have been detected!

Conclusion

In comparison to characterizing enzymes with water-soluble substrates, interfacial enzymes require consideration of additional parameters. Cholesterol oxidase is just one example of an interfacial enzyme. Other classes are the well-known and studied phospholipases and lipases, as well as lipid-transforming enzymes and steroid biosynthetic enzymes. Characterization of integral membrane proteins also requires similar considerations.

The kinetics of an interfacial enzyme can be reliably determined if a measurement of membrane binding affinity (e.g., K_d for a standard state of model membrane) is included in the reported rate parameters. One way to access this binding parameter is use of a catalytically inactive mutant that still folds and associates with the membrane.

Rate constants must be reported in units of mole fraction as the use of bulk concentrations does not describe what the enzyme encounters at the membrane interface.

The structure and components of the membrane containing substrate can alter the kinetic activity, as well as the binding constant for the membrane surface. Moreover, the physical state of the substrate can be very temperature sensitive.

The kinetic characterization of interfacial enzymes is complicated by the uncertainties of knowing the physical state of the substrate that is relevant under physiological or *in vivo* conditions. In the case of integral membrane proteins, identification of the substrate structure may be simplified by identifying the location of the enzyme.

Acknowledgments

This work was supported by the National Institutes of Health (HL 53306, N.S.S) and the American Heart Association (0725861T, S.K.).

References

[1] Vrielink, A., Lloyd, L.F., Blow, D.M. (1991) Crystal structure of cholesterol oxidase from *Brevibacterium sterolicum* refinded at 1.8 Å resolution. *J. Mol. Biol*. **219**:533 – 554.

[2] Li, J., Vrielink, A., Brick, P., Blow, D.M. (1993) Crystal structure of cholesterol oxidase complexed with a steroid substrate: implication for flavin adenine dinucleotide dependent alcohol oxidases. *Biochemistry* **32**:11507 – 11515.

[3] Lario, P., Sampson, N.S., Vrielink, A. (2003) Sub-atomic resolution crystal structure of cholesterol oxidase: What atomic resolution crystallography reveals about enzyme mechanism and the role of the FAD cofactor in redox activity. *J. Mol. Biol*. **326**:1635 – 1650.

[4] Sankaram, M.B., Thompson, T.E. (1991) Cholesterol-induced fluid-phase immiscibility in membranes. *Proc. Natl. Acad. Sci. U.S.A*. **88**:8686 – 8690.

[5] Ahn, K.W., Sampson, N.S. (2004) Cholesterol oxidase senses subtle changes in lipid bilayer structure. *Biochemistry* **43**:827 – 836.

[6] Veatch, S.L., Keller, S.L. (2005) Miscibility phase diagrams of giant vesicles containing sphingomyelin. *Phys. Rev. Lett*. **94**:148101.

[7] Heller, H., Schaefer, M., Schulten, K. (1993) Molecular-Dynamics Simulation of a Bilayer of 200 Lipids in the Gel and in the Liquid-Crystal Phases. *J. Phys. Chem. B*. **97**:8343 – 8360.

[8] Tak, J. (1942) On bacteria decomposing cholesterol. *Antonie Leeuwenhoek* **8**:32 – 40.

[9] Turfitt, G.E. (1944) The microbiological degradation of steroids. 2. Oxidation of cholesterol by *Proactinomyces* spp. *Biochem. J*. **38**:492 – 496.

[10] Richmond, W. (1973) Preparation and properties of a cholesterol oxidase from *Nocardia* sp. and its application to the enzyme assays of total cholesterol in serum. *Clin. Chem*. **19**:1350 – 1356.

[11] Fukuda, H., Kawakami, Y., Nakamura, S. (1973) A method to screen anticholesterol substances produced by microbes and a new cholesterol oxidase produced by *Streptomyces violascens*. *Chem. Pharm- Bull. (Tokyo)* **21**:2057 – 2060.

[12] Smith, A.G., Brooks, C.J.W. (1974) Application of cholesterol oxidase in the analysis of steroids. *J. Chromat.* **101**:373 – 378.

[13] Sampson, N.S., Kass, I.J. (1997) Isomerization, but not oxidation, is suppressed by a single point mutation, E361Q, in the reaction catalyzed by cholesterol oxidase. *J. Am. Chem. Soc*. **119**-855 – 862.

[14] Ye, Y., Liu, P., Anderson, R.G.W., Sampson, N.S. (2002) Construction of a catalytically inactive cholesterol oxidase mutant: investigation of the interplay between active site residues glutamate 361 and histidine 447. *Arch. Biochem. Biophys*. **402**:235 – 242.

[15] Sampson, N.S., Vrielink, A. (2003) Cholesterol oxidase: A study of nature's approach to protein design. *Acc. Chem. Res*. **36**:713 – 722.

[16] Purcell, J.P., Greenplate, J.T., Jennings, M.G., Ryerse, J.S., Pershing, J.C., Sims. S.R., Prinsen, M.J., Corbin, D.R., Tran, M., Sammons, R.D., Stonard, R.J. (1993) Cholesterol oxidase: a potent insecticidal protein active aganst boll weevil larvae. *Biochem. Biophys. Res. Commun*. **196**:1406 – 1413.

[17] Xu, X., London, E. (2000) The effect of sterol structure on membrane lipid domains reveals how cholesterol can induce lipid domain formation. *Biochemistry* **39**:843 – 849.

[18] Chen, X., Wolfgang, D., Sampson, N.S. (2000) Use of the parallax-quench method to determine the position of the active-site loop of cholesterol oxidase in lipid bilayers. *Biochemistry* **39**:13383 – 13389.

[19] Radhakrishnan, A., McConnell, H.M. (2000) Chemical activity of cholesterol in membranes. *Biochemistry* **39**:8114 – 8124.

[20] Uwajima, T., Yagi, H., Nakamura, S., Terada, O. (1973) Isolation and crystallization of extracellular 3β-hydroxysteroid oxidase of *Brevibacterium sterolicum* nov. sp. *Agr. Biol. Chem*. **37**:2345 – 2350.

[21] Xiang, J., Sampson, N.S. (2004) Library screening studies to investigate substrate specificity in the reaction catalyzed by cholesterol oxidase. *Prot. Eng. Design & Select*. **17**:341 – 348.

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

Teaching Enzyme Kinetics and Mechanism in the 21st Century

Athel Cornish-Bowden

Unité de Bioénergétique et Ingénierie des Protéines, Centre National de la Recherche Scientifique, 31 chemin Joseph-Aiguier, B.P. 71, 13402 Marseille Cedex 20, France

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

Received: 4th April 2008 / Published: 20th August 2008

Abstract

The teaching of enzyme kinetics has been neglected in recent years, with the growth in influence of molecular biology, but its importance has not diminished. Elementary aspects of enzyme inhibition have always been central to the understanding and design of pharmacological agents and pesticides, and both kinetics and metabolism have acquired a new role for making sense of the flood of genome data that has appeared in the past decade. Although at one time it was hoped that sequence analysis alone would be sufficient for deducing phenotypic information from genomic data, it has become clear that it has to be combined with stoicheiometric analysis, knowledge of metabolic networks and analysis of enzyme regulation. Presentation of kinetics in general textbooks has always been very poor, and the decline of specialized teaching has made the inadequacy of these textbooks more serious than it already was in the past.

Introduction

The basis of enzyme kinetics as it is commonly taught today derives from the work of Henri [1] and Michaelis and Menten [2], and one may wonder why it should continue to be regarded as an essential component of biochemistry courses nearly a century later. However, despite a decline in interest in kinetics over recent decades the subject remains fundamental to several currently active fields of research. The development of genetic techniques for producing artificially modified enzymes has increased the importance of precise methods for

estimating kinetic parameters, because modern research requires the ability to quantify small differences in activity between mutant forms. Enzyme inhibition remains central to the development of new drugs, pesticides and other products of biotechnology, and needs to be well understood for such development to proceed efficiently. For these and other reasons to be developed in this chapter, satisfactory teaching of enzyme kinetics is no less important for the training of biochemists than it has ever been. Unfortunately, however, the majority of current general textbooks of biochemistry fall far short of providing an adequate treatment of the subject. At the same time although specialized textbooks exist [3 – 5] they are less numerous than they once were.

Drugs typically act by inhibiting enzymes, and this implies two things that must be important for designing them, first an understanding of the differences between the different kinds of inhibition (competitive, uncompetitive, etc.), and second, less obvious but at least as important, an understanding of why measurements of inhibition made in controlled conditions in a spectrophotometer may provide very little guide as to how much inhibition can be expected to occur (even at the same concentration of inhibitor) *in vivo*. Both points arise out of the same central fact: experiments in the spectrophotometer are typically done at concentrations of inhibitor, substrate, product etc., that are chosen and fixed by the experimenter, but concentrations of all metabolites (both substrates and products of enzyme-catalysed reactions) *in vivo* can vary by very large factors when conditions change; differences between inhibition types that are small enough to pass unnoticed when substrate and product concentrations are fixed can be very large when these concentrations are allowed to vary. Both points are considered in more detail in this chapter. For the moment it is sufficient to note that an inhibitor concentration that produces a large decrease in the rate of an enzyme-catalysed reaction considered in isolation will often produce no detectable effect on the rate *in vivo*. This problem does *not* arise primarily from an inability to achieve the same inhibitor concentration in a cell as one can readily achieve in the spectrophotometric (though that may well be a serious additional problem). Anyone setting out to design a pharmacological agent needs both to recognize the basic fact that kinetic behaviour *in vivo* is usually different from that *in vitro* (observed as well as theoretically predicted) and to understand why it happens.

Elementary Enzyme Kinetics in Current Textbooks

A major difficulty for understanding the (slightly) less elementary aspects of kinetics discussed later in this paper is that even the most elementary points are often presented very badly in general textbooks of biochemistry. It is a matter of simple arithmetic to calculate that if a reaction follows the Michaelis–Menten equation,

$$v = \frac{V_a}{K_m + a} \tag{1}$$

in which v is the rate at substrate concentration a, V is the limiting rate and K_m is the Michaelis constant, then $v = 10V/11$ when $a = 10K_m$, in other words v is nearly 10% below V at this concentration. As the calculation is so simple one can only be amazed that so many textbook authors are apparently unable to do it. Of general biochemistry textbooks published in the past decade, three [6 – 8] illustrate the dependence of v on a given by eqn. 1 with grossly inaccurate curves, three [9 – 11] show curves that are reasonably accurate on one page and inaccurate on another, and fewer than half [12 – 15] manage to present the relationship accurately. Notice that one of the textbooks that has it incorrect is now in its 6th edition, having appeared originally in 1975 [17]: evidently the correction of elementary errors is not a high priority when successful textbooks are revised. However, in the past even some authors of specialized books on enzyme kinetics were unable to draw the curve correctly [16], something that would be unthinkable today, so at least there has been some progress.

This is not a trivial matter, because if one thinks that v reaches V at a substrate concentration a few times greater than K_m it is impossible to understand why V cannot be determined by direct measurement, so that K_m could be obtained from a simple measurement of the a value at which $v = V/2$. Without this understanding it is impossible to understand why linear transformations of eqn.1 [18 – 21] played such a great role in the development of biochemistry, and why they remain important today for illustrating results and in teaching.

The presentation of the more elementary aspects of kinetics in the general biochemistry textbooks of today may be poor[1], but the treatment of more advanced aspects is nonexistent. Not since 1972 has any widely used general textbook [23] included any attempt to go beyond the Michaelis–Menten equation and competitive inhibition, and so anyone who wants to do this has no choice but to go to more specialized sources.

Deducing Phenotypes from Genotypes

When genome sequencing on a large scale first became practicable, and in particular when it first became realistic to expect the entire human genome to become known, there were hopes, not always clearly stated, that deducing phenotypes from gene sequences would be relatively straightforward. In the event, however, although the human genome project has certainly had and is having major effects on the diagnosis and treatment of many diseases, it has proved much more difficult than was widely expected to pass directly from genotype to phenotype, which is far from being a one-step transition.

[1] It is depressing to note that IUPAC's recommendations on quantities, units and symbols in physical chemistry [22], published as recently as 2007, assert that k_{-2}, defined as the rate constant constant for binding of the product to the free enzyme, is negligibly small. Astonishingly, the compilers of the recommendations did not wish to acknowledge that this was an error when it was brought to their attention. If experts on physical chemistry cannot understand that an algebraic expression is zero if any *one* of its factors (in this case the product concentration) is zero, without implying anything about the magnitudes of its other factors, what hope is there for undergraduate students?

In the first place the genes present in a genome need to be identified, and this is by no means an error-free process. When we discussed the steps involved in deducing phenotypes some years ago [24] we reported an estimate [25] that as many as half of the proteins in *Caenorhabditis elegans*, with a much more compact and non-repetitive genome than the human, might be incorrectly identified; even in as thoroughly studied an organism as *Escherichia coli* as many as 40% of its 4405 genes are still without experimentally determined functions [26].

So the first difficulty in trying to deduce a phenotype from genome information is the need to correct the errors of identification, and to identify the genes with unknown functions. As an example of how this can be done, consider the work of Schuster *et al.* [27], who examined the problem of identifying the genes of *Treponema pallidum*, the organism responsible for syphilis, for which the genome was known but was accompanied by extremely little biochemical information. Comparison with the genome of *E. coli* allowed many genes to be tentatively identified, and stoicheiometric analysis allowed some of the others to be identified. Assuming that the metabolism of *T. pallidum* is not totally different from that of *E. coli* and indeed from those of other organisms for which the information is available, one must explain the apparent absence of a gene for the enzyme transaldolase by supposing that it is indeed present but has not been recognized. The logic here is that in *E. coli* and all other known cases transaldolase occurs only in organisms that use the transketolase reaction, because stoicheiometric analysis shows that transketolase would have no function to fulfil in the absence of transaldolase. As two different genes for transketolase were found in *T. pallidum* there must be at least one transaldolase gene as well even if none had been found. It would of course be extremely laborious to analyse an entire genome in this way, but it can be very helpful nonetheless. However, there is also a more fundamental difficulty, that it offers away of detecting *similarities* between organism, whereas the main reason for studying an organism such as *T. pallidum* is to understand its *differences* from a more thoroughly studied organism like *E. coli*. Why does *T. pallidum* cause syphilis, whereas *E. coli* does not? We cannot learn this by noticing respects in which *T. pallidum* resembles *E. coli*.

Even if we brush aside all the problems inherent in establishing a list of putative proteins present in an organism, there remain several steps to be taken before we arrive at a real phenotype. For the first of these stoicheiometric analysis is again very helpful, as it can provide a metabolic map, from which one may deduce a possible phenotype. However, a possible phenotype is not a real phenotype, and there is no way in which purely stoicheiometric considerations allow one to proceed any further. It is not at all sufficient to know what reactions are present and hence what is stoicheiometrically possible. One also needs some information about kinetics and regulation. Perhaps in the future it may be possible to deduce this sort of information from genomic data, but it is certainly not possible now, and there is no alternative to undertaking some real biochemical experiments.

Characterizing Inhibition *in vitro*

As enzyme inhibition is central to the design of pharmacological agents[2], it is essential to consider how it should be measured and reported. Biochemists have long used *inhibition constants*, i. e. the parameters K_{ic} and K_{iu} that appear in the following equation:

$$v = \frac{V_a}{K_m\left(1+\frac{i}{K_{ic}}\right)+a\left(1+\frac{i}{K_{iu}}\right)} \qquad (2)$$

in which v is the rate at substrate concentration a and inhibitor concentration i, and V and K_m are the parameters that define the kinetics of the uninhibited reaction (as in eqn. 1). In the past it was common to consider only the competitive component of the inhibition, defined by K_{ic}, but it is now usual to recognize that the uncompetitive component, defined by K_{iu}, must also be considered. For reasons that will be discussed, this is actually essential for characterizing inhibition *in vivo*. Many methods are available for determining the inhibition constants: among the more widely known are the plots of $1/v$ against i for determining K_{ic} as the abscissa coordinate of the common intersection point of lines plotted at different values of a [28], and of a/v against i for determining K_{iu}, again as the abscissa coordinate of the common intersection point of lines plotted at different values of a [29].

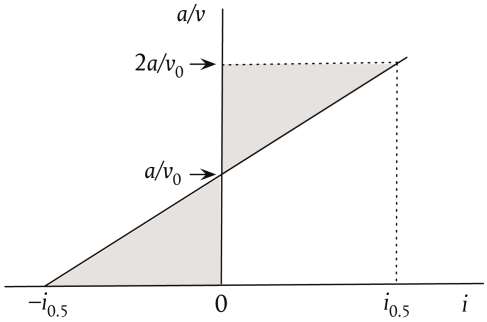


Figure 1. Determination of $i_{0.5}$ from standard inhibition plots. The value of a/v when i is extrapolated to zero is a/v_0 by definition, and likewise the value when $i = i_{0.5}$ is $2a/v_0$. Rotation of the shaded triangle 180 ° about the point $(0,a/v_0)$ gives a congruent triangle, from which it follows that the intercept of the line on the abscissa must occur at $i=-i_{0.5}$. Although the construction is illustrated for a/v as ordinate axis [29] the logic applies equally well to $1/v$ as ordinate axis [28].

In contrast, pharmacologists continue to characterize inhibition in terms of a parameter that has become largely obsolete in biochemistry, the *inhibitor concentration for half-inhibition*, the concentration needed to arrive at $v = 0.5V$, which can be symbolized as $i_{0.5}$. This would

[2] Enzyme activation is important only in some special circumstances, exemplified by liver hexokinase, as discussed later.

be a trivial difference if there were a simple relationship relating $i_{0.5}$ to K_{ic} and K_{iu}, but in reality no such relationship exists, as one may deduce from the fact that K_{ic} and K_{iu} are independent of the substrate concentration a whereas $i_{0.5}$ is not. For the limiting cases, competitive and uncompetitive inhibition, it is *never* true (even at a carefully selected substrate concentration) either that $i_{0.5}=K_{ic}$ or that $i_{0.5}=K_{iu}$. For mixed inhibition, when both competitive and uncompetitive components make significant contributions, it is possible to choose values of a for which $i_{0.5}=K_{ic}$ or $i_{0.5}=K_{iu}$, but these are special cases of no particular interest or importance. Only when $K_{ic}=K_{iu}$ do we find $i_{0.5}=K_{ic}=K_{iu}$, but this is also a special case, known as pure non-competitive inhibition. It is often given undeserved attention in elementary accounts of inhibition because, many years ago, Michaelis contrasted it with competitive inhibition in his studies of the inhibition of invertase [31] and maltase [32]. Except in the case of inhibition by protons it has very little importance in the real world.

So, if $i_{0.5}$ cannot be understood as an inhibition constant, how should it be understood in standard biochemical terms? In fact it is the abscissa intercept of a line plotted in either of the ways mentioned, i.e. either in a plot of $1/v$ against i or in one of a/v against i [30]. The logic of these relationships is illustrated in Fig. 1. In addition, plots of $1/i_{0.5}$ against v_0/V allow the type of inhibition to be determined from measurements of $i_{0.5}$ at different values of a: with all linear types of inhibition this plot gives a straight line, with a negative slope of $-1/K_{ic}$ and an abscissa intercept at $v_0/V = 1$ if the inhibition is competitive, a positive slope of $1/K_{iu}$ and a line passing through the origin if the inhibition is uncompetitive, and intermediate behaviour for mixed inhibition [30]. This approach gives results in accordance with theoretical expectation for inhibition of lactate dehydrogenase by different kinds of inhibitor [30].

Inhibition *in vivo*

The equation for competitive inhibition is a limiting case of eqn. 2 with the term in i/K_{iu} omitted:

$$v = \frac{V_a}{K_m(1+i/K_{ic})+a} \tag{3}$$

and the equation for uncompetitive inhibition is at the opposite extreme with the term in i/K_{ic} omitted:

$$v = \frac{V_a}{K_m+(1+i/K_{iu})} \tag{4}$$

These two equations are not only similar in appearance; they are also similar in the quantitative behaviour they predict when i is varied at fixed a. That is why experiments in the spectrophotometer often lead them to be distinguished poorly or not at all (Fig. 2), even in published work, and the uncompetitive component sometimes passes unnoticed. For this

reason one must always treat reports of competitive inhibition with suspicion: did the authors really exclude the possibility of an uncompetitive component, or did they just ignore it?

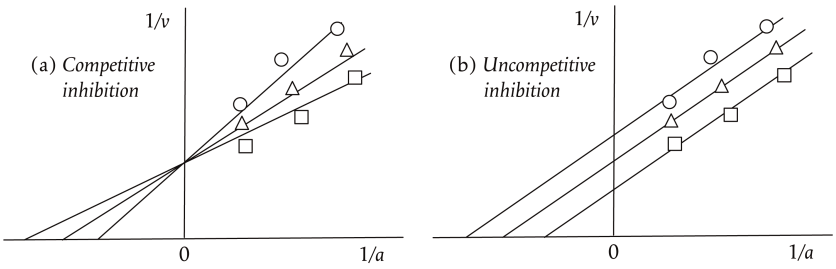


Figure 2. On a casual inspection the double-reciprocal plots in (**a**) made at three different inhibitor concentrations illustrates competitive inhibition, with the lines intersecting on the ordinate axis, where the plots in (**b**) show parallel lines, and hence uncompetitive inhibition. However, inspection of the data points reveals that they are identical in the two cases, so what the plots illustrate is not different kinds of inhibition but different *interpretations* of the same data.

One maybe tempted to dismiss the (invented) example in Fig. 2 as an exaggeration of what happens with real data, but in reality one can find published examples (e. g. [33]) that are much worse than what is illustrated here. The current tendency of journal editors to discourage the presentation of primary data is making the problem progressively less obvious, but that does not mean the problem is disappearing, only that it is becoming more difficult to recognize.

Matters are drastically different if i is varied at fixed v, with a allowed to vary freely. To see this the two equations need to be rearranged to show a/K_{ic} as a function of i in each case. For competitive inhibition, rearrangement of eqn. 3 yields

$$a \Big/ K_m = \frac{1+i/K_{ic}}{V/v-1} \tag{5}$$

from which it is immediately evident that a is a linear function of i under these conditions. Rearrangement of eqn. 4 shows that the corresponding behaviour with uncompetitive inhibition is very different, however:

$$a \Big/ K_m = \frac{1}{V/v-1-i/K_{iu}} \tag{6}$$

Not only is this a non-linear function, it is also one that allows a to become infinite when $i/K_{iu}=V/v-1$, a condition that is easily satisfied in practice: it means, for example, that at a substrate concentration of K_m in the absence of inhibitor it is sufficient for i to reach K_{iu} for the steady state to be lost [34], as illustrated in Fig. 3.

Although eqns. 5 – 6 are useful for understanding the great differences between kinetics at fixed rate and kinetics at fixed substrate concentration, they oversimplify the problem of understanding the differences between *in vivo* and *in vitro* kinetics, because one cannot equate *in vivo* conditions with fixed-rate conditions. For a minority of enzymes, those that act of substrates like glucose that are maintained at essentially constant concentrations by regulatory mechanisms, the conditions may, in fact, resemble those in the spectrophotometer; such enzymes can be expected to respond just as well to inhibition *in vivo* as they do *in vitro*. They are the exceptions, however. More often an enzyme finds itself in the middle of a metabolic pathway where it has little influence on the flux through the reaction that it catalyses, but considerable influence on the concentrations of its substrate and product. In the language of metabolic control analysis [4] it typically has a small flux control coefficient for the flux through its own reaction, but large concentration control coefficients (negative and positive respectively) for the concentrations of its product and substrate. Effectively it must process its substrate at the rate at which it arrives, but can adjust the concentrations to satisfy its kinetic equation. As expected, therefore, computer modelling of a ten-step pathway in which the fifth enzyme is inhibited by an added inhibitor gave results for both competitive and uncompetitive inhibitors that closely resembled the calculated behaviour for fixed-rate conditions and were very different from those for fixed-concentration conditions [34].

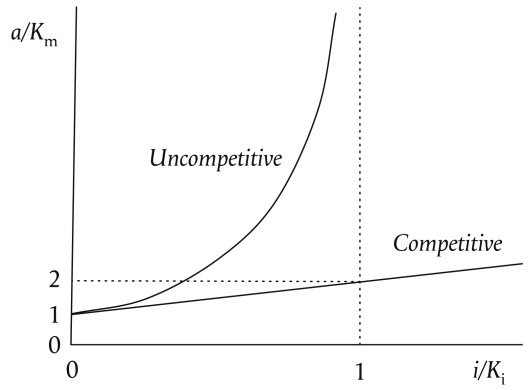


Figure 3. Inhibition at constant rate. When the concentration of an inhibitor is varied and the rate is held constant the substrate concentration needs to change in order for the rate equation to be satisfied. If the inhibition is competitive the substrate concentration varies linearly with the inhibitor concentration, and, starting from an initial state of $a/K_m = 1$ at $i = 0$ the substrate concentration is simply doubled at $i = K$i. For uncompetitive, however, the behaviour is very different, and the same condition $i = K_i$ is sufficient to incur loss of the steady state.

From the point of view of designing inhibitors as drugs or pesticides, the conclusion to be drawn from this discussion is that in most circumstances (i. e. when the enzyme to be inhibited acts under conditions of constant or nearly constant flux through its reaction) there needs to be an uncompetitive component in the inhibition if it is to be useful *in vivo*.

Practical examples of this principle include Li^+, an uncompetitive inhibitor of *myo*-inositol monophosphatase [35] used to treat manic depression, and Glyphosate ("Roundup"), an uncompetitive inhibitor of 3-phosphoshikimate 1-carboxyvinyltransferase [36] and the herbicide with the greatest commercial success in history.

Designing an Uncompetitive Inhibitor

A difficulty that will occur to anyone who wishes to apply the principle developed in the preceding section is that whereas it is easy to design a competitive inhibitor, as the structural characteristics of a typical competitive inhibitor are obvious, it is almost impossible to design an uncompetitive inhibitor, as there are no general structural properties that make a particular molecule likely to be an uncompetitive inhibitor of a particular enzyme. The reality, however, is not nearly as discouraging as this suggests. First of all it is not necessary for the inhibition to be purely uncompetitive, because as long as a mixed inhibitor displays a sufficient uncompetitive component it can have useful effects *in vivo*. In fact the uncompetitive component may dominate the behaviour *in vivo* even if the inhibition is predominantly competitive [37, 38]. So, although true uncompetitive inhibition is rare [34] mixed inhibition is not, and inhibitors with structural characteristics that suggest them to be competitive inhibitors often act as mixed inhibitors.

A second important point is that nearly all metabolic reactions have two or more substrates and two or more products, and an inhibitor that is competitive with respect to one substrate is often mixed or uncompetitive with respect to another. This is, in fact the case for Glyphosate: it is a structural analogue of one substrate of 3-phosphoshikimate 1-carboxyvinyltransferase, phospho*enol*pyruvate and is competitive with respect to it, but it is predominantly uncompetitive with respect to the other substrate, 3-phosphoshikimate.

In practice, therefore, designing an uncompetitive inhibitor may not be as difficult as it appears at first sight. In any case, the degree of difficulty is ultimately not the point: it is better to search for a difficult solution to a problem that has a chance of succeeding than to search for an easy solution that cannot work.

Does Enzyme Activation Ever Have a Useful Pharmacological Role?

The tendency of most enzymes to have small or negligible control coefficients for the flux through their own reactions not only means that inhibiting them will typically have little or no effect or this flux *in vivo* (though it may, if the inhibition has an uncompetitive component, have a major effect on some metabolite concentrations); it also means that activating them will also have little or no effect on the flux. In fact even activating them by large factors will usually have a negligible effect, because flux control coefficients typically decrease when the enzyme activity increases. There is, however, an important exception to

this generalization. As discussed elsewhere [39], the resistance of most metabolic fluxes to changes in enzyme activity, mainly due to the summation relationship [40], is in part due to the regulatory design of pathways in terms of supply and demand. As most pathways respond to changes in demand they resist changes in supply. An important exception, however, concerns uptake of glucose by the mammalian liver followed by phosphorylation to glucose 6-phosphate: this is not primarily regulated by the liver's need for glucose, but by the need to maintain homeostasis, and particular to maintain a constant blood-glucose concentration [41]. In other words it must be regulated according to supply: increases in glucose availability need to be followed by increased uptake in the liver. It follows that hexokinase D, the enzyme responsible for phosphorylation of glucose in the liver, has a high flux control coefficient for its own reaction [42], and as a result is capable of responding *in vivo* to activators. Hence the current commercial interest in finding good activators of this enzyme is much better founded than it would be for most other enzymes. The example is discussed in more detail elsewhere [43].

Computer Analysis of Kinetic Experiments

For many years an unsatisfactory aspect of practice in enzyme kinetics was the almost universal tendency to estimate kinetic parameters by visual inspection of double-reciprocal plots, even after it was pointed out that the deviation of $1/v$ from its theoretical value provided an extremely misleading indication of the corresponding deviation in v and satisfactory methods of calculation became available [44]. Practice has, however, changed drastically since desktop computers became generally available, and it is now almost universal to use commercial software for estimating rate constants. Unfortunately, this is not necessarily an improvement: when double-reciprocal plots appeared in every paper it was at least possible for a critical reader to understand what had been done, but this has become almost impossible when crucial details are often hidden by a cryptic sentence to the effect that a particular commercial program was used. Even when authors are aware of the statistical assumptions implicit in the software they rarely reveal this to their readers, who have no way of knowing whether the observations were appropriately weighted or not. More detail may be found in kinetics textbooks [3 – 5]: the important point is that computer fitting is an advance over fitting by eye only if it is correctly done, with the added disadvantage that it is difficult for the reader to judge.

Concluding Remarks

There are many aspects of enzyme kinetics that form a less visible part of biochemistry courses than they did 40 years ago, though they remain essential to the proper understanding of the subject, and in particular to its application to biotechnology. As noted, development of pharmacological agents, metabolic engineering, etc., typically involve knowledge of enzyme inhibition and the kinetics of multienzyme systems, but the first of these tends to be taught in a superficial way and the second often not at all.

References

[1] Henri,V. (1903) *Lois Générales de l'Action des Diastases.* Hermann, Paris.

[2] Michaelis, L., Menten, M.L. (1913) Kinetik der Invertinwirkung. *Biochem. Z.* **49**:333 – 369.

[3] Marangoni, A. (2003) *Enzyme Kinetics: a Modern Approach.* Wiley–Interscience, Hoboken.

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

[5] Cook, P.F., Cleland, W.W. (2007) *Enzyme Kinetics and Mechanism.* Garland Science, NewYork.

[6] Boyer, R.F. (2002) *Concepts in Biochemistry,* 3rd edn., p. 139,Wiley.

[7] Campbell, M.K., Farrell, S.O. (2006) *Biochemistry,* 6th edn., Brooks Cole.

[8] McKee, T., McKee, J.R. (2002) *Biochemistry: The Molecular Basis of Life.* 3rd edn., McGraw-Hill.

[9] Berg, J.M., Tymoczko, J.L., Stryer, L. (2006) *Biochemistry.* 6th edn., W.H.Freeman.

[10] Mathews, C.K., van Holde, K.E., Ahern, K.G. (1999) *Biochemistry,* 3rd edn., Prentice Hall.

[11] Horton, R., Moran, L.A., Scrimgeour, G., Perry, M. (2005) *Principles of Biochemistry,* 4th edn., Prentice Hall.

[12] Garrett, R.H., Grisham, C.M. (2004) *Biochemistry,* 2nd edn., Saunders.

[13] Metzler, D.E. (2002) *Biochemistry, the Chemical Reactions of Living Cells,* 2nd edn., Academic Press.

[14] Nelson, D.L., Cox, M.M. (2004) *Lehninger Principles of Biochemistry,* 4th edn., W.H. Freeman.

[15] Voet, D., Voet, J.G. (2004) *Biochemistry,* 3rd edn.,Wiley.

[16] Naqui, A. (1986) Where are the asymptotes of Michaelis-Menten? *Trends Biochem. Sci.* **11**:64 – 65.

[17] Stryer, L. (1975) *Biochemistry,* 1st edn., W.H.Freeman.

[18] Lineweaver, H., Burk, D. (1934) The determination of enzyme dissociation constants. *J. Amer. Chem. Soc.* **56**:658 – 666.

[19] Hanes, C.S. (1932) Studies on plant amylases. *Biochem. J.* **26**:1406 – 1421.

[20] Eadie, G.S. (1942) The inhibition of cholinesterase by physostigmine and prostigmine. *J. Biol. Chem.* **146**:85 – 93.

[21] Hofstee, B.H.J. (1952) Specificity of esterases. *J. Biol. Chem.* **199**:357 – 364.

[22] International Union of Pure and Applied Chemistry (2007) *Quantities, Units and Symbols in Physical Chemistry,* 3rd edn., RSC Publishing, Cambridge.

[23] Mahler, H.R., Cordes, E.H. (1972) *Biological Chemistry,* 2nd edn., Harper and Row, NewYork.

[24] Cornish-Bowden, A., Cárdenas, M. L. (2000) From genome to cellular phenotype – a role for metabolic flux analysis? *Nat. Biotechnol.* **18**:267 – 268.

[25] Claverie, J.-M. (2000) Do we need a huge new centre to annotate the human genome? *Nature* **403**:12.

[26] Stourman, N.V., Wadington, M.C., Schaab, M.R., Atkinson, H.J., Babbitt, P.C., Armstrong, R.N. (2008) Functional Genomics in *Escherichia coli:* Experimental Approaches for the Assignment of Enzyme Function**.** *In:* Proceedings of the 3rd International Beilstein Symposium on Experimental Standard Conditions of Enzyme Characterizations (Eds. M.G. Hicks, C. Kettner), Logos Verlag Berlin, p. 1–13

[27] Schuster, S., Fell, D., Dandekar, T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* **18**:326 – 332.

[28] Dixon, M. (1953) The determination of enzyme inhibitor constants. *Biochem. J.* **55**:170 – 171.

[29] Cornish-Bowden, A. (1974) A simple graphical method for determining the inhibition constants of mixed, uncompetitive and non-competitive inhibitors. *Biochem. J.* **137**:143 – 144.

[30] Cortés, A., Cascante, M., Cárdenas, M. L., Cornish-Bowden, A. (2001) Relationships between inhibition constants, inhibitor concentrations for 50% inhibition and types of inhibition: new ways of analysing data. *Biochem. J.* **357**:263 – 268.

[31] Michaelis, L., Pechstein, H. (1914) Über die verschiedenartige Natur der Hemmungen der Invertasewirkung. *Biochem. Z.* **60**:62 – 78.

[32] Michaelis, L., Rona, P. (1914) Die Wirkungsbedingungen der Maltase aus Bierhefe. III. Über die Natur der verschiedenartigen Hemmungen der Fermentwirkungen. *Biochem. Z.* **60**:79 – 90.

[33] Li, X.-Y., McClure, W.R. (1998) Characterization of the closed complex intermediate formed during transcription initiation by *Escherichia coli* RNA polymerase. *J. Biol. Chem.* **273**:23549 – 23557.

[34] Cornish-Bowden, A. (1986) Why is uncompetitive inhibition so rare? *FEBS Lett.* **203**:2 – 3.

[35] Pollack, S.J., Atack, J.R., Knowles, M.R., McAllister, G., Ragan, C.I., Baker, R., Fletcher, S.R., Iverson, L.I., Broughton, H.B. (1994) Mechanism of inositol monophosphatase, the putative target of lithium therapy. *Proc. Natl. Acad. Sci. U.S.A.* **91**:5766 – 5770.

[36] Boocock, M.R., Coggins, J.R. (1983) Kinetics of 5-enolpyruvylshikimate-3-phosphate synthase inhibition by glyphosate. *FEBS Lett.* **154**:127 – 133.

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

[38] Hofmeyr, J.-H.S., Cornish-Bowden, A. (1996) Predicting metabolic pathway kinetics with control analysis. *In:* BioThermoKinetics of the Living Cell (Eds.Westerhoff, H.V., Snoep, J.L., Wijker, J.E., Sluse, F.E., Kholodenko, B.N.), BioThermoKinetics Press, Amsterdam. pp. 155 – 158.

[39] Hofmeyr, J.-H. S., Cornish-Bowden, A. (2000) Regulating the cellular economy of supply and demand. *FEBS Lett.* **476**:47 – 51.

[40] Kacser, H., Burns, J.A. (1973) The control of flux. *Symp. Soc. Exp. Biol.* **27**:65 – 104.

[41] Cárdenas, M.L. (1995) *Glucokinase: its regulation and role in liver metabolism.* R.G. Landes Austin.

[42] Agius, L., Peak, M., Newgard, C.B. et al.(1996) Evidence for a role of glucose-induced translocation of glucokinase in the control of hepatic glycogen synthesis. *J. Biol. Chem.*, **271**:30479 – 30486.

[43] Cornish-Bowden, A., Nanjundiah, V. (2006) The basis of dominance. *In:* The Biology of Genetic Dominance (ed. Veitia, R. A.) Landes Bioscience, Georgetown, Texas, pp. 1 – 16.

[44] Wilkinson, G.N. (1961) Statistical estimations in enzyme kinetics. *Biochem. J.*, **80**:324 – 332.

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

How to Develop a Standard – the HUPO-PSI Experience

Sandra Orchard

EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

E-Mail: orchard@ebi.ac.uk

Received: 2nd June 2008 / Published: 20th August 2008

Abstract

The HUPO Proteomics Standards initiative has designed and implemented common data reporting and exchange standards to enable the transfer of proteomics data from originator to collaborator to a final public repository immediately prior to publication. This work has been undertaken with extensive community involvement at every stage of the process to ensure that the end product fulfils the users' needs. The scientific community is already benefiting from this work, with XML formats to exchange and import data into databases, allowing direct access and comparability irrespective of the originating instrumentation. Public repositories allow researchers to access and search published experimental data with the result that reference datasets are becoming available for benchmarking purposes. Collaborations between databases are exposing these datasets to an ever increasing audience and enabling exciting new science to be derived from existing data.

Introduction

Since 2002, the Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) have worked towards producing standards formats by which proteomics data can be collected, transferred from application to application, and finally submitted to a repository where it becomes available to the user community [1]. Within the field of proteomics, the ever-increasing amounts of data generated by mass spectrometers with ever-faster cycle

times, high sensitivity and high quality MS spectra required a move beyond the traditional route of printed publication. Whilst journal articles may still be the most practical method for disseminating the conclusions drawn from such experiments, they display only a small proportion of the actual results and almost none of the underlying data, such as peptides or spectra, from which this data was generated. When mass spectrometry data has been made available, for example via an author's website, it is generally only provided in vendor-specific formats. In most cases, the sheer size of the data files generated by a typical mass spectrometry experiment, result in raw data being erased once published, with only the final processed protein lists being retained and published.

To enable the community to effectively mine an increasingly rich data source, the experimental data needs to be collected in central repositories, in a single common format, from where it can be searched, and researched. As protein sequence databases, such as UniProtKB [2], improve in both coverage and sequence quality, high quality spectra produced at an earlier point in time can be rerun, and protein assignments made to previously orphan spectra, provided these spectra have been retained. To maximize the value which can be extracted from each experiment, the metadata needs to be both well annotated, and consistently annotated across different datasets. The aim of the work of the HUPO-PSI is to make it possible for labs across the world to run experiments and combine results, to distribute workloads, to enable the user to select the best tools for specific tasks rather than be forced to use only those which use a very specific format, and to derive new methodologies by comparing results. To this end a four-fold approach was taken, for each aspect of the potential workflow within a proteomics experiment:

1. Formal requirements specification with use cases submitted as many users as possible.
2. A checklist of minimum information with which each experiment should be annotated.
3. The development of an XML interchange format to enable data transfer which can be used by instrumentation, tools and repositories.
4. Controlled vocabularies to enable a rich annotation of the data within the schema.

Each of these have now been developed and released for gel electrophoresis, chromatographic techniques, mass spectrometry, protein identifications and molecular interactions.

Formal Requirements Specification – Standards as a Community effort

The best standard in the world will be of limited use if it is not adopted and employed by the majority of the community, it is not designed with ease of use and accessibility in mind and is not accompanied by a suite of tools which maximize its use. With that in mind, all

workgroups within the PSI have actively sought the input of as broad a group as possible, including hardware manufacturers, software developers, data producers, data repositories and bioinformaticians. Standards developments have centered around (bi-)annual workshops, which have been widely advertised and made freely available to all to attend. Development continued between workshops using all possible means of communications, including discussion groups, mailing lists and telephone conferences. All discussion documentation was made publicly available on the HUPO-PSI website (http://www.psidev.info/). Completed work enters a formal documentation process designed to ensure a good balance between expert design and public scrutiny [3]. Documents are first subjected to anonymous review followed by a set period of time in which the document is put forward for public comment. At all points, any feedback is responded to and, if appropriate, standards are updated and the document resubmitted into the review process. Any documents which are formally published are then additionally subject to standard journal review processes – *Nature Biotechnology* also established a community review page which provided further opportunity for potential users to comment on the standards prior to final publication. The PSI review process has also been used for the validation of documents not directly produced by the HUPO-PSI, for example the FuGE data model was reviewed through this proves prior to journal publication [4].

Minimum Information About a Proteomics Experiment (MIAPE)

Proteomics data should therefore ideally be accompanied by contextualizing metadata, making explicit both where samples came from and how analyses were performed. To that end, the Proteomics Standards Initiative develops guidance documents specifying the data and metadata that should be collected from various proteomics workflows, known collectively as the ”minimum information about a proteomics experiment” (MIAPE) guidelines [5]. These consist of a series of documents, linked by a single parent which makes explicit the scope, purpose and manner of use of the modular MIAPE guidelines that accompany it, and lay out the principles underlying module production. The first of these modules has already been published, the Molecular Interaction guidelines MIMIx [6], with those for gel electrophoreses and column chromatography to follow. Broadly speaking, MIAPE documents require that any description of a proteomics experiment should allow the user to understand, qualify and reproduce the work described in that paper. Each document provides a checklist of information and data to provide when an experiment is reported and acts as an aid to assessing quality control but does not attempt to tell a scientist how to run an experiment, or how to represent the final data and does not state how any quality judgment should be made. It is envisaged that these requirements will eventually be adopted and enforced by journals, repositories and funding agencies. Finally, these documents will ensure that the presentation of the information published will be in a style compatible with the HUPO-PSI data formats.

An increasing number of minimum requirements documents have now been produced, since the original, MIAME [7], produced by the micro-array community covering a broad range of biological disciplines. Such 'minimum information' checklists are usually developed independently, from within particular biologically- or technologically-delineated domains. Consequently, the full range of checklists can be difficult to find without intensive searching; they are also inevitably partially-redundant one against another, and where they overlap arbitrary decisions on wording and sub-structuring make integration difficult. This presents significant difficulties for the users of checklists; for example, in the area of systems biology, where data from multiple biological domains and technology platforms are routinely combined. A common portal to such MI checklists; to act as a 'one-stop shop' for those exploring the range of extant projects, foster collaborative development and ultimately promote gradual integration, MIBBI (http://mibbi.sourceforge.net/, http://www.mibbi.org/) has now been developed. All MIAPE modules are submitted to MIBBI, on publication.

XML Interchange Standards

To facilitate data management and exchange, the HUPO-PSI has developed data exchange formats for proteomics. For each work group/domain, these can minimally represent the data items specified in the MIAPE guidelines, but usually allow a much more detailed representation. Normally, the data exchange format is specified as a fully annotated XML schema. PSI schemas are developed to facilitate data exchange between databases as well as databases and end users. They explicitly do not propose any internal data representation for databases or tools. While XML is inherently verbose, standard compression algorithms typically reduce the file size by 50 – 90% of the original, and such compression normally is not the limiting factor on modern computer systems. On the plus side, XML is well supported by standard mechanisms for querying, native XML databases, and automated mappings to both relational databases and object models.

The molecular interaction interchange format, PSI-MI XML 2.5, is now used by all interaction databases and an increasing number of graphical visualisation and analysis tools. The mass spectrometry standard, mzData, was released in 2004 and was well accepted by many users. The format allowed the storage of proteomic-related mass spectral data, ranging from basic details about the sample, instrument details and data processing steps, through to the actual spectral lists of mass-to-charge values and intensities, using base64 encoding to represent the floating point mass-to-charge (m/z) and ion intensity. Following some refinement in response to their feedback, the format was rapidly implemented by several of these manufacturers and files containing spectral data were soon being generated by several large groups, most notably the HUPO tissue initiatives. A standards compliant repository, PRIDE (http://www.ebi.ac.uk/pride) was also established. However, in 2004 a second open, generic XML representation of MS data, was published by the Institute of Systems Biology, mzXML [8]. Whilst this was originally designed to be work-flow specific, other workers began to find wider uses for the format with the result that manufacturers were faced with

the prospect of having to implement two separate open-source formats. Rather than loose the initial good-will with which the manufacturers had entered this project, and to avoid user confusion, in 2006 the two groups decided to merge the two formats into a single, and much improved, XML schema was released in June 2008, mzML with full vendor support was a critical part of the design process and many open-source implementations.

Accompanying an XML representation of MS data, there is a need for a corresponding representation of the peptide and protein identifications made in any experiment to capture results from MS search engines and represent the input parameters for analysis algorithms, thus unifying results from different search engines. The development of AnalysisXML has proven far from straightforward, partly because the scope of the project has changed often in a fast moving field however Version 1.0 will be submitted to the PSI documentation process summer 2008. Quantitation will not be addressed until version 2.0, however version 1.0 documents will be backwards compatible with the 2.0 schema.

GelML, designed for the interchange of protocols and image data from 1D- and 2D-gel electrophoresis, completed the PSI document process and was released as a stable version 1.0 late in 2007. The interchange format for non-gel based separations, spML, is in development.

Controlled Vocabularies

While XML schemas provide a syntax for data exchange, they do not specify the semantics of data elements exchanged. As an example, the yeast two-hybrid technology might be designated by many different terms, most of which are sufficiently distinct to make automatic recognition impossible. Thus, the PSI either references external controlled vocabularies (CVs) or ontologies such as the NCBI taxonomy where possible, or develops its own controlled vocabulary, for example for protein interaction detection technologies, where necessary. The combination of reasonably stable XML schemas and regularly maintained controlled vocabularies has proven to allow quick adaptation to new terms and technologies, while providing the stability required for database and software development. All PSI CVs are written in OBO format and maintained at the Open Biomedical Ontologies website (http://www.obofoundry.org) and can be viewed using the Ontology Lookup Service (http://www.ebi.ac.uk/ontology-lookup/"\t"_parent) [9].

Interaction with the Publishing Community

The need for underlying experimental data to be available to the reader has long been recognised by the scientific publishing community, and this has been addressed in many ways – from the mandated submission of nucleotide and protein sequences to public domain repositories, with accession numbers then being given in the article, to the provision of Supplementary Material held by the journal itself. As the volume of data increases, the

journals will become more reliant on external repositories to collect and manage this data and, in recognition of this, are actively supporting the implementation of the required standards needed to enable deposition. Several journals have already published their own guidelines as to how proteomics data should be represented and these requirements have been incorporated in the MIAPE guidelines [5]. Increasingly, journals are starting to request data deposition: *Proteomics*, and *Nature Biotechnology* [10] and *Nature Methods* [11] have recently started to request that authors deposit proteomics and interaction data in HUPO-PSI standards-compliant databases prior to publication and it is anticipated that this trend will increase as both standards and databases mature.

Summary

The HUPO-PSI have produced a range of standards allowing the user to describe all aspects of a proteomics experiment, to exchange and compare data across collaborating groups, or use established datasets as standards, and to deposit the final results into standards-compliant repositories. All of these have been written in full consultation with an extensive user community to ensure as much by-in as possible. Where an existing standard, be it an interchange format or CV the PSI have strived to work with that facility rather than replace it or produce a redundant application. Cross-community efforts are actively supported to ensure that users working in multiple ''omics'', for example analysing the same sample by both transcriptomic and proteomic techniques can find the tools to assist in the data handling and resorting processes. The emphasis now is on developing the tools to make these standards usable and accessible to the bench scientist and to encourage the direct deposition of data into public domain repositories.

References

[1] Orchard, S., Hermjakob, H. (2008) The HUPO proteomics standards initiative–easing communication and minimizing data loss in a changing world. *Brief. Bioinformatics* **9**:166 – 73.

[2] The UniProt Consortium (2008) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **36**:D 190 – 195..

[3] Vizcaíno, J.A., Martens, L., Hermjakob, H., Julian, R.K., Paton, N.W. (2008) The PSI formal document process and its implementation on the PSI website. *Proteomics* **7**:2355 – 2357.

[4] Jones, A.R., Pizarro, A, Spellman, P., Miller, M., FuGE Working Group (2006) http://www.ebi.ac.uk/citexplore/citationDetails.do?externalId = 16901224&dataSource = MED FuGE: Functional Genomics Experiment Object Model. *OMICS* **10**:179 – 184.

[5] Taylor, C.F., Paton, N.W., Lilley,K.S., Binz, P.-A., Randall, J.R.K., Jr, Jones A.R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E.W., Dunn, M.J., Heck, A.J.R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T.A., Patterson S.D., Ping, P., Seymour, S.L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T.M., Whitelegge, J.P., Wilkins, M.R., Xenarios, I., Yates III J.R., Hermjakob, H. (2007) The Minimum Information About a Proteomics Experiment (MIAPE). *Nat. Biotechnol*. **25**(8):887 – 893.

[6] Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stümpflen, V., Ceol, A, Chatraryamontri, A, Armstrong, J., Woollard, P., Salama, J.J., Moore, S., Wojcik, J., Bader, G.D., Vidal, M., Cusick, M.E., Gerstein, M, Gavin, A.-C., Superti-Furga, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Mulder, N., Gilson, M., Niepmann, M., Burgoon, L., De Las Rivas, J., Prieto, C., Perreau, V.M., Hogue, C, Mewes, H.-W., Apweiler, R., Xenarios, I., Eisenberg, D. Cesareni, G., Hermjakob, H. (2007) The Minimum Information required for reporting a Molecular Interaction Experiment (MIMIx). *Nat. Biotechnol*. **25**(8):894 – 898.

[7] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C.P., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*. **29**:365 – 371.

[8] Pedrioli, P.G., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R., Cheung, K., Costello, C.E., Hermjakob, H., Huang, S., Julian, R.K., Kapp, E., McComb, M.E., Oliver, S.G., Omenn, G., Paton, N.W., Simpson, R., Smith, R., Taylor, C.F., Zhu, W., Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol*. **22**(11):1459 – 1466.

[9] Côté, R.G., Jones, P., Martens, L., Apweiler, R., Hermjakob H (2008) The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res*. In press.

[10] Nature Biotechnology (2007) Editorial Time for leadership. *Nat. Biotechnol* **25**(8):821.

[11] Doerr, A. (2007) Standardizing proteomics. *Nat. Methods* **4**:774.

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

Thermodynamic Property Values for Enzyme-catalyzed Reactions

Robert N. Goldberg

*Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, U.S.A.
and
Department of Chemistry and Biochemistry, University of Maryland, Baltimore County, Baltimore MD 21250, U.S.A.

E-Mail: robert.goldberg@nist.gov

Received: 24th April 2008 / Published: 20th August 2008

Abstract

This chapter deals with how one can obtain values of thermodynamic properties – specifically the apparent equilibrium constant K', the standard molar transformed Gibbs energy change $\Delta_r G'^\circ$, and the standard molar transformed enthalpy change $\Delta_r H'^\circ$ for biochemical reactions – and, in particular, for enzyme-catalyzed reactions. In addition to direct measurement, these property values can be obtained in a variety of ways: from thermochemical cycle calculations; from tables of standard molar formation properties; by estimation from property values for a chemically similar reaction or substance; by means of estimation by using a group-contribution method; by combining a known value of the standard molar enthalpy change $\Delta_r H^\circ$ and an estimated value for the standard molar entropy change $\Delta_r S^\circ$ in order to obtain the standard molar Gibbs energy change $\Delta_r G^\circ$ for a given reaction; and by use of computational chemistry.

* This is official contribution of the National Institute of Standards and Technology and is not subject to copyright in the United States.
Certain commercial items are identified in this paper. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology (NIST), nor is it intended to imply that these items are necessarily the best available for the purpose.

Introduction and General Principles

This chapter deals with how one can obtain values of thermodynamic properties – specifically the apparent equilibrium constant K', the standard molar transformed Gibbs energy change $\Delta_r G'^\circ$, and the standard molar transformed enthalpy change $\Delta_r H'^\circ$ for biochemical reactions – and, in particular, for enzyme-catalyzed reactions. Much of the interest in these property values arises from applications in bioprocess engineering, where the aim is to optimize product yield and energy utilization [1]. Another interest arises from the use of thermodynamics to model metabolic processes [2]. This approach has been expanded in recent years to also include kinetic considerations in the modeling calculations [3, 4].

Firstly, it is important to appreciate that, for biochemical reactions, thermodynamic quantities are, in general, functions of temperature T, pH, pX, and ionic strength I. Here, $pX = -\log_{10}[X]$, where [X] is the concentration of a species X, typically an ion, that binds to one or more of the reactants. This dependency on pH and pX arises because of the multiple states of ionization and metal ion binding in which the reactant molecules can exist. This point is illustrated by means of a generic reaction – the hydrolysis of adenosine 5′-triphosphate (ATP) to adenosine 5′-diphosphate (ADP) and phosphate (all reactions discussed in this chapter pertain to aqueous media unless indicated otherwise)

$$\mathrm{ATP} + \mathrm{H_2O} = \mathrm{ADP} + \mathrm{phosphate}. \tag{1}$$

The apparent equilibrium constant K' for this reaction is

$$K' = [\mathrm{ADP}][\mathrm{phosphate}]/[\mathrm{ATP}]. \tag{2}$$

By convention the concentration of water has been omitted in the expression for K'. The concentrations used in eqn. (2) are *total* concentrations of the various ionic and metal bound forms of the reactants and products. For example

$$[\mathrm{ATP}] = [\mathrm{ATP}^{4-}] + [\mathrm{HATP}^{3-}] + [\mathrm{H_2ATP}^{2-}] + [\mathrm{H_3ATP}^{-}] + [\mathrm{MgATP}^{2-}] + [\mathrm{MgHATP}^{-}] + [\mathrm{MgH_2ATP}] + [\mathrm{Mg_2ATP}], \tag{3}$$

$$[\mathrm{ATP}] = [\mathrm{ADP}^{3-}] + [\mathrm{HADP}^{2-}] + [\mathrm{H_2ADP}^{-}] + [\mathrm{MgADP}^{-}] + [\mathrm{MgHADP}], \tag{4}$$

$$[\mathrm{phosphate}] = [\mathrm{PO_4^{3-}}] + [\mathrm{HPO_4^{2-}}] + [\mathrm{H_2PO_4^{-}}] + [\mathrm{H_3PO_4}] \tag{5}$$

If calcium or other divalent metal ions are present, one must also consider additional, analogous species such as CaATP^{2-}. The essential point is that, because biochemical reactants such as ATP, ADP, and phosphate exist in several different ionic and metal bound

forms, there is a multiplicity of species that make up each of these reactants. This, in turn, leads to the aforementioned dependencies of thermodynamic quantities on pH and pX. Illustrations of these dependencies are shown in Fig. 1. These surface plots were calculated by using the equilibrium constant for the chemical reference reaction

$$ATP^{4-} + H_2O = ADP^{3-} + HPO_4^{2-} + H^+, \quad (6)$$

and equilibrium constants for the pertinent H^+ and Mg^{2+} binding constants:

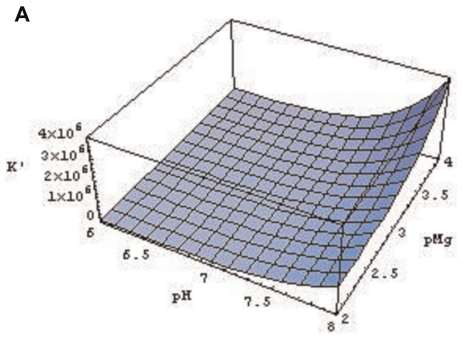
$$ATP^{4-} + H^+ = HATP^{3-}, \quad (7)$$

$$ATP^{4-} + Mg^{2+} = MgATP^{2-}, \quad (8)$$

$$HATP^{3-} + H^+ = H_2ATP^{2-}, \quad (9)$$

$$HATP^{3-} + Mg^{2+} = MgHATP^-, \text{ etc.} \quad (10)$$

It is important to recognize that the equilibrium constants K for reactions (6) to (10) pertain to specific chemical species. Clearly, these chemical reactions must balance both the number of atoms and the charges. While equilibrium constants K depend on temperature and ionic strength they do not depend on pH or pX as do apparent equilibrium constants K'. Thus, it is important to maintain a clear distinction between K and K' [5]. The book *Thermodynamics of Biochemical Reactions* [6] contains a definitive treatment of transformed thermodynamic properties and many examples involving biochemical reactions.



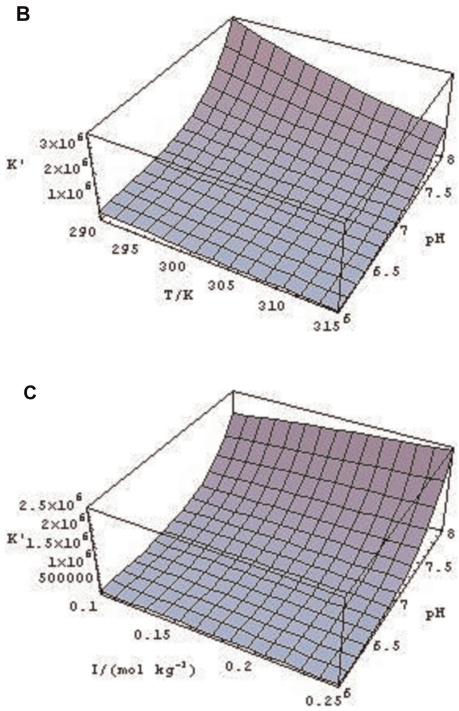


Figure 1. The apparent equilibrium constant K' for the hydrolysis reaction (ATP + H_2O = ADP + phosphate) as a function of temperature T, pH, pMg, and ionic strength I. Since it is not possible to represent a five dimensional surface in two dimensions, three 3-D projections with appropriate constraints are shown. These constraints are: T = 298.15 K and I = 0.25 mol·dm^{-3} (A); I = 0.25 mol·dm^{-3} and pMg = 3.0 (B); and T = 298.15 K and pMg = 3.0 (C).

Values of standard molar enthalpies of reaction $\Delta_r H^\circ$ can be used to adjust values of K from one temperature to another over a relatively narrow temperature range by means of the van't Hoff equation. If one is operating over a wider temperature range, one also needs values of the standard molar heat capacity changes $\Delta_r C_P^o$ for the reactions of interest. For this purpose, some useful formulas are [7]

$$\Delta_r G^\circ(T) = (T/T_{ref})\cdot\Delta_r G^\circ(T_{ref}) + \Delta_r H^\circ(T_{ref})\cdot(T_{ref} - T)/T_{ref} + \Delta_r C_P^o \cdot \{(T - T_{ref}) - T\ln(T/T_{ref})\} \quad (11)$$

$$K = \exp\{-\Delta_r G^\circ(T)/(RT)\} \quad (12)$$

Here, R is the gas constant, T is the temperature of interest, and T_{ref} is a reference temperature, typically 298.15 K. Also, since reactions are carried out at ionic strengths where non-ideality has to be accounted for, one also needs values of the activity coefficients γ of the species for use in the equilibrium equations. In the absence of values of activity coefficients γ for the vast majority of biochemical species, a common practice has been to estimate values of the activity coefficient γ_i of an ion i by using the extended Debye-Hückel equation

$$\log_e y_j = -A_m z_i^2 I^{1/2} / \left(1 + BI^{1/2}\right). \tag{13}$$

Here A_m is the Debye-Hückel constant ($A_m = 1.1758\,\mathrm{kg}^{1/2}\cdot\mathrm{mol}^{-1/2}$ at 298.15 K), z_i is the charge of species i, and B is an empirical constant, which has often been set at $1.6\,\mathrm{kg}^{1/2}\cdot\mathrm{mol}^{-1/2}$. Values of A_m have been tabulated by Clarke and Glew [8] for the temperature range 273 K to 423 K.

The principles discussed above provide a basis for performing calculations of the type that led to the results shown in Fig. 1. These calculations depend on having the necessary thermodynamic data available for both the chemical reference reaction {e. g., reaction (6)} and the pertinent H^+ and metal-ion binding constants {e. g., reactions (7) to (10)}. One must also be able to solve the simultaneous non-linear equations that describe these complex reaction systems. Fortunately, this numerical problem can be handled routinely by using computer algorithms to calculate the extent of reaction for each chemical reaction, the concentrations of all chemical species and of the biochemical reactants, and finally the value of K'. These algorithms can also be used to calculate how calorimetrically determined molar enthalpy changes $\Delta_r H$(cal) and changes in binding $\Delta_r N$ of H^+ and X also vary with pH, pX, T, and I.

As illustrated in Fig. 1, these calculations allow one to obtain all of this information as a function of T, pH, pX, and I. *Thus, the formalism outlined above makes it possible to obtain an essentially complete thermodynamic picture of these important reactions.*

Having a sound thermodynamic framework for dealing with complex reactions is essential. However, reliable property values are required for performing practical calculations. A substantial body of experimental results for enzyme-catalyzed reactions has accumulated over many years and has been systematized in several review articles [9 – 15] and on the web [16]. The equilibrium data has been obtained by using a variety of analytical methods, with chromatography, spectrophotometry, and enzymatic assays being the most commonly used. Molar enthalpies of reaction have been obtained either from calorimetric results or from equilibrium constants which have been measured as a function of temperature.

The essential point is that if one knows K' at a given T, pH, pMg, and I, it is possible to calculate a value of K for a chemical reference reaction. This value of K can then be used with the pKs of the chemical reactants and products in the reference reaction to calculate

values of K' as a function of T, pH, and I. A similar approach applies to calorimetrically determined molar enthalpy changes $\Delta_r H$(cal). In this case one needs, in addition to the pKs, values of standard molar enthalpies of reaction $\Delta_r H^\circ$ for the relevant proton and metal ion binding reactions. Many of these pK and $\Delta_r H^\circ$ values can be found in existing databases [17, 18]. However, if a pK or $\Delta_r H^\circ$ value has not been measured, it may be possible to estimate a value to a sufficient degree of accuracy by using property values for structurally similar substances [19]. To summarize, one has a substantial body of experimental data, K' and $\Delta_r H$(cal), that can be used to calculate values of K and $\Delta_r H^\circ$ for chemical reference reactions that correspond to overall biochemical reactions. These calculated values of K and $\Delta_r H^\circ$ can then be used together with values of K and $\Delta_r H^\circ$ for the relevant proton and metal ion binding reaction to calculate K' as a function of T, pH, pMg, and I. These calculations can be performed relatively conveniently by using published algorithms [20 – 22].

Computational and Estimation Methods

However, if K' has not been measured for a reaction of interest, it may still be possible to obtain a value of K' by a variety of means. Possible approaches are:

- Calculate K and/or K' by means of thermochemical cycle calculations or by using tables of standard molar formation properties.
- Estimate the desired property value by using property values for a chemically similar reaction.
- Estimate the desired property value by using a group-contribution or Benson approach.
- If a value of $\Delta_r H$(cal) has been measured, one can calculate $\Delta_r H^\circ$ for an appropriate chemical reference reaction. A value of $\Delta_r S^\circ$ can then be estimated and combined with $\Delta_r H^\circ$ to give a value of $\Delta_r G^\circ$ for the reference reaction. This value of $\Delta_r G^\circ$ can then be used in conjunction with values of K and $\Delta_r H^\circ$ for the relevant proton and metal ion binding reactions to obtain K' at the desired T, pH, pMg, and I.
- Use computational chemistry to calculate the desired property values.

A brief discussion of each of these approaches follows.

Tables of Standard Molar Formation Properties

The method by which one adds or subtracts chemical reactions and combines the thermodynamic properties of these reactions to obtain the thermodynamic property for the summed reaction is well-known and will not be described herein. The generalization of this method leads to tables of formation properties. Thus, while extensive tables of formation properties exist [23] for inorganic substances and for organic substances that have one or two carbons,

tabulations of the formation properties of biochemical substances is relatively limited. The earliest thermodynamic tables for biochemical substances appear to have been prepared by Krebs, Kornberg, and Burton [2]. These pioneering tables contain standard molar Gibbs energies of formation $\Delta_f G^\circ$ for 88 species – and the property values for 21 of these species came from the NBS thermochemical tables [24]. The 1969 tables of Wilhoit [25] cover a much larger number of substances than Krebs *et al.* [2] and include values of the standard molar enthalpy of formation $\Delta_f H^\circ$, the standard molar entropy S°, and the standard molar heat capacity C_P^o in addition to the standard molar Gibbs energy of formation $\Delta_f G^\circ$.

The 1989 tables of Goldberg and Tewari [26] are limited to the thermodynamic properties ($\Delta_f G^\circ$, $\Delta_f H^\circ$, S°, and C_P^o) of carbohydrates, specifically the pentoses and the hexoses, and their monophosphates. Goldberg and Tewari [26] constructed their tables by using a ''reaction catalog'' which consists of a table of experimental property values for the reactions and substances that are the basis for the calculation of the formation properties. The experimental property values were weighted according to their estimated accuracies. The advantage to this approach is that, as new experimental results become available, the reaction catalog can be updated relatively easily and new thermodynamic tables can be calculated promptly from the new reaction catalog.

The 1990 tables of Miller and Smith-Magowan [27] are limited in their coverage to the substances found in the Krebs cycle. It should be noted that their [27] tabulated values of $\Delta_f G^\circ$ pertain to an ionic strength of 0.1 mol·dm^{-3}, while the other thermochemical tables use the conventional thermochemical standard state based upon a hypothetical ideal solution of unit molality. The thermochemical tables of Alberty [21, 22] are the most extensive of those available. They contain values of $\Delta_f G^\circ$ for all of the species of 199 biochemical reactants (sums of species). Values of $\Delta_f H^\circ$ are also given for the species that comprise 94 of these biochemical reactants. Alberty [21, 22] also gives computer programs that allow one to use the property values in his tables to calculate standard transformed Gibbs energy changes $\Delta_r G'^\circ$, standard transformed enthalpy changes $\Delta_r H'^\circ$, and apparent equilibrium constants K' for biochemical reactions as a function of T, pH, and I. The user of any of these tables is cautioned about the risks in using data from two or more different tables to calculate property values for a given reaction. Specifically, while each of the aforementioned thermochemical tables is internally consistent, the values may not be consistent with the values given in another table. Thus, serious errors can result if values from two or more tables are combined to calculate property values for a given reaction. While substantial progress has been made in thermodynamic tables for biochemical substances since the early work of Krebs, Kornberg, and Burton [2], an inspection of the total amount of thermodynamic data available for enzyme-catalyzed reactions [9 – 16] shows that there are many substances for which values of standard molar formation properties can be calculated from existing data. In addition to property values for enzyme-catalyzed reactions, there are also substantial amounts of data for standard molar enthalpies of combustion, standard molar entropies, saturation molalities (solubilities), standard molar enthalpies of solution, and pKs and stan-

dard molar enthalpies of reaction for proton and metal-ion binding reactions that tie into the calculation of standard molar formation properties for biochemical substances. It will be a major challenge to pull together all of these thermodynamic property values together and to produce the equivalent of the tables that currently exist for inorganic substances and for small organic molecules in aqueous media.

Estimation of a Property Value by Using Property Values for a Chemically Similar Reaction

The correlation of property values with structure can be usefully exploited to obtain property values for reactions and substances that have not been the subjects of direct measurements. This approach will be illustrated by a few examples. Boerio-Goates *et al.* [28] used the structural similarity (see Fig. 2) between the inosine 5′-triphosphate series (ITP, IDP, and IMP) and the adenosine 5′-triphosphate series (ATP, ADP, and AMP) to estimate the p*K*s and $\Delta_r H^\circ$ values for the H^+ and Mg^{2+} binding reactions of the ITP series from the known values for the ATP series of similar reactions.

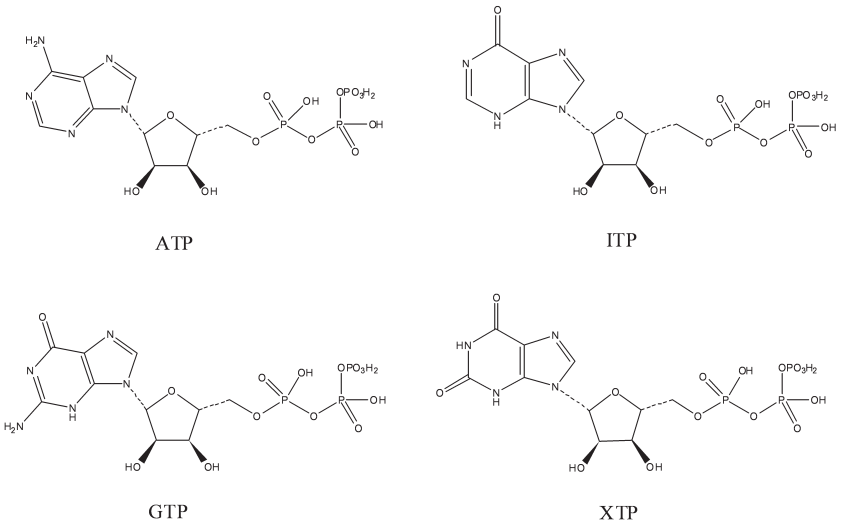


Figure 2. The structures of adenosine 5′-triphosphate (ATP), inosine 5′-triphosphate (ITP), guanosine 5′-triphosphate (GTP), and xanthosine 5′-triphosphate (XTP)

Alberty [29] later used this same structural similarity and assumed that $\Delta_r G^\circ$ for the hydrolysis reactions in the GTP series and the XTP series were the same as in the ITP series. Also, Goldberg *et al.* [30] measured standard molar enthalpy changes for the hydrolysis reactions of maltose, maltotriose, maltotetraose, maltohexaose, and maltoheptaose to form glucose at 298.15 K. It was found that the values of $\Delta_r H^\circ/N$, where N is the number of α-1,4 linkages in the aforementioned substances, had a constant value equal to $-(4.53 \pm 0.04)$ kJ·mol^{-1}. Consideration of additional calorimetric and equilibrium measurements [30] showed that,

for reactions involving the making/breaking of N saccharide linkages, the assignment of characteristic values of $\Delta_r H^\circ/N$ or $\Delta_r G^\circ/N$ or $\Delta_r S^\circ/N$ for a specified linkage, was accurate in predicting the values of $\Delta_r H^\circ$, $\Delta_r G^\circ$, and $\Delta_r S^\circ$ for reactions involving saccharides containing multiples or combinations of such linkages.

Finally, Tewari and Goldberg [31], in their summary of the results of the hydrolysis reactions of several six-carbon disaccharides, found that the values of $\Delta_r S^\circ$ for these reactions were in the range $32\,J\cdot K^{-1}\cdot mol^{-1}$ to $48\,J\cdot K^{-1}\cdot mol^{-1}$. This relatively narrow range of values has a practical use in that a value of $\Delta_r G^\circ$ for a disaccharide hydrolysis reaction can be estimated from its measured value of $\Delta_r H^\circ$ by using a typical value of $\Delta_r S^\circ \approx 40\,J\cdot K^{-1}\cdot mol^{-1}$ based on the aforementioned range of values.

Clearly, there are many other biochemical reactions where knowledge of the property values for one reaction can lead to reliable property values for a structurally similar reaction.

Estimation of the Desired Property Value by Using a Group-Contribution Method

Closely related to the aforementioned method based on structural similarity is its generalization which assigns characteristic values to distinct chemical groups. In this approach, a chemical substance is broken down into its characteristic groups, values are assigned to each group, and a property value for the substance of interest is obtained by summation of these values.

This approach is commonly referred to as the group-contribution or Benson method, named after Sidney W. Benson, the principal developer of this approach [32 – 34]. This method is made possible by the availability of a large array of data for the standard molar enthalpies of formation $\Delta_f H^\circ$, standard molar heat capacities C_P^o, and standard molar entropies S° of organic compounds. Thus, in their 1993 review, Domalski and Hearing [35] were able to provide ≈ 3700 comparisons between measured and calculated values for $\Delta_f H^\circ$, S°, and C_P^o for organic compounds in the gas, liquid, and solid phases.

However, if one wishes to utilize the group-contribution values tabulated by Domalski and Hearing [35] to estimate a value of $\Delta_f H^\circ$, $\Delta_f G^\circ$, S°, and C_P^o for a substance in the aqueous phase, one must first use their [35] tabulated group-contribution values to calculate these property values for the substance (to take an example) in the solid phase. The values of $\Delta_f H^\circ$ and S° can then be used to calculate a value of $\Delta_f G^\circ$ for this substance. In order to calculate $\Delta_f H^\circ$ and $\Delta_f G^\circ$ for this substance in aqueous solution, one then needs values for the standard molar enthalpy of solution $\Delta_{sol} H^\circ$ and for the standard molar Gibbs energy of solution $\Delta_{sol} G^\circ$. A value of $\Delta_{sol} G^\circ$ can be obtained from measurements of the saturation molality (or solubility) along with a value of the activity coefficient of the substance in solution at the saturation molality.

If the condensed phase is hydrated, one also needs the number of waters of hydration associated with the condensed phase and a value of the activity of the water. A value of $\Delta_{sol}H^\circ$ can be obtained either from a direct calorimetric measurement or by use of the van't Hoff equation with values of $\Delta_{sol}G^\circ$ at several temperatures.

Domalski [36] has discussed an approach to develop an estimation method for molar enthalpies of formation of organic compounds in water. This same and similar approaches can also be applied to the estimation of molar entropies and Gibbs energies of formation of organic compounds in water.

A limited set of such group-contribution values was developed by Cabani *et al.* [37] for 58 chemical groups. Mavrovouniotis [38, 39] has also developed a table of group-contribution values for biochemical substances in water. It should be noted that the table of Mavrovouniotis [38] appears to be based on values of apparent equilibrium constants obtained at a variety of conditions (T, pH, pMg, and I).

Thus, the tabulated group-contribution values may have inconsistencies caused by variations in the conditions of the measurements on which the group-contribution values are based. Clearly, the construction of extensive tables of group-contribution values such as the tables of Domalski and Hearing [35] were enabled by the existence of a large array of property data, most of which came from measurements of standard molar enthalpies of combustion and third law molar entropies. Thus, the existence of an extensive table of Δ_fH°, Δ_fG°, and S° values for aqueous biochemical species is a prerequisite for the development of accurate group-contribution values for biochemical substances.

Estimation of K′ and $\Delta_rG'^\circ$ from Calorimetric Results

For some reactions, the position of equilibrium lies too far in a given direction to permit the practical measurement of an apparent equilibrium for that reaction. In such cases, and in the absence of either a thermochemical pathway or tabulated values of Δ_fG° for the reactants and products, it may still be possible to obtain a reasonable estimate of K' and $\Delta_rG'^\circ$ for the reaction of interest if one has a value of the calorimetrically determined molar enthalpy change Δ_rH(cal) for the overall biochemical reaction. Specifically, one can use Δ_rH(cal) to calculate Δ_rH° for a suitable chemical reference reaction. This calculation will most likely involve a correction for the enthalpy of protonation of the buffer [40] and will also require values of K and Δ_rH° for the proton and metal-ion binding for the biochemical reactants. Once Δ_rH° has been calculated, one must then estimate the standard molar entropies S° of the chemical reactants and products in the chemical reference reaction. These estimates are best done on the basis of structural similarity and are greatly aided if there are reasonably extensive tables of standard molar entropies available. In the absence of reliable property

values for similar substances, one must rely on a Benson or group-contribution method. This approach is useful if an apparent equilibrium constant is too large (or too small) to measure and if there are no thermochemical pathways to $\Delta_r G^\circ$ for the reaction of interest.

Computational Chemical Results

Computational chemistry continues to make considerable progress [41] and is approaching a state where some practically useful results can be obtained for thermodynamic properties for biochemical reactions in aqueous media in which the reactant molecules are not too large. Clearly, the principal bottleneck is treating the hydration of the molecules, which, in turn, impacts conformational equilibria and energies. What has been done in some cases [41, 42], as an alternative to the explicit inclusion of waters in the calculations, is to use a polarizable continuum model (PCM). In the cited studies [42, 43], comparisons were made between measured and calculated values of $\Delta_r H^\circ$ for two biochemical reactions:

$$\text{chorismate}^{2-} = \text{prephenate}^{2-} \tag{14}$$

$$\text{chorismate}^{2-} = \text{pyruvate}^{-} + \text{4-hydroxybenzoate}^{-} \tag{15}$$

For reaction (14), the experimental value for the standard molar enthalpy change was $\Delta_r H^\circ = -(55.4 \pm 2.3)$ kJ·mol^{-1} and the computed (PCM) value was $\Delta_r H^\circ = -46.4$ kJ·mol^{-1}. For reaction (15), the experimental value was $\Delta_r H^\circ = -(144 \pm 7)$ kJ·mol^{-1} and the computed (PCM) value was $\Delta_r H^\circ = -154$ kJ·mol^{-1}. While it was not possible to assign uncertainties to the computed $\Delta_r H^\circ$ values, it is clear that any reasonable assignment of possible error to the computed values would demonstrate that they were in agreement with the experimental values. Clearly, the calculation of $\Delta_r S^\circ$ is more difficult than the calculation of $\Delta_r H^\circ$ for a given reaction. Nevertheless, the value $\Delta_r S^\circ = 3$ J·K^{-1}·mol^{-1} was obtained for reaction (14) by computation in which PCM was again used. In the absence of any experimental value for either $\Delta_r G^\circ$ or for $\Delta_r S^\circ$ for reaction (14), a comparison of a computationally derived and a Benson estimate for $\Delta_r S^\circ$ was made for the reaction

$$\text{chorismate(g)} = \text{prephenate(g)}. \tag{16}$$

The Benson estimate led to $\Delta_r S^\circ = 9$ J·K^{-1}·mol^{-1} for the above reaction and the computationally derived value was $\Delta_r S^\circ = 20$ J·K^{-1}·mol^{-1}. It is hoped that advances in computational chemistry will eventually make it possible to have values of $\Delta_r H^\circ$, and, eventually, $\Delta_r G^\circ$ and $\Delta_r S^\circ$ for biochemical reactions that are comparable in accuracy to experimentally derived values. Clearly, at the present time, accurately measured property values along with property values calculated by means of thermochemical pathways remain at the top of the hierarchy for the reliability of property values obtained by the various methods discussed herein.

Acknowledgments

I thank Drs. Robert A. Alberty and Yadu B. Tewari for many helpful discussions.

Glossary of Symbols and Terminology

A_m	Debye-Hückel constant (1.1758 $kg^{1/2}$ $mol^{-1/2}$ at 298.15 K)	$kg^{1/2}\cdot mol^{-1/2}$
B	empirical constant in the extended Debye-Hückel equation	$kg^{1/2}\cdot mol^{-1/2}$
C_P^o	molar heat capacity	$J\cdot K^{-1}\cdot mol^{-1}$
$\Delta_r C_P^o$	standard molar heat capacity change[a]	$J\cdot K^{-1}\cdot mol^{-1}$
$\Delta_r G°$	standard molar Gibbs energy change[a]	$kJ\cdot mol^{-1}$
$\Delta_r G'°$	standard molar transformed Gibbs energy change	$kJ\cdot mol^{-1}$
$\Delta_r H°$	molar enthalpy change for a reaction[a]	$kJ\cdot mol^{-1}$
$\Delta_r H'°$	standard molar transformed enthalpy change	$kJ\cdot mol^{-1}$
$\Delta_r H$(cal)	calorimetrically determined molar enthalpy change	$kJ\cdot mol^{-1}$
I_c	ionic strength, concentration basis	$mol\cdot dm^{-3}$
I_m	ionic strength, molality basis	$mol\cdot kg^{-1}$
K	equilibrium constant	dimensionless
K'	apparent equilibrium constant	dimensionless
N	number of entities	dimensionless
$\Delta_r N_X$	change in binding of species X	dimensionless
P	pressure	Pa
pH	$-\log_{10}[H^+]$	dimensionless
pK	$-\log_{10}K$	dimensionless
pMg	$-\log_{10}[Mg^{2+}]$	dimensionless
R	gas constant (8.314 472 $J\cdot K^{-1}\cdot mol^{-1}$)	$J\cdot K^{-1}\cdot mol^{-1}$
T	temperature	K
z_i	charge number of species i	dimensionless
γ	activity coefficient	dimensionless

[a] The subscript "r" denotes a reaction. Subscripts "f" and "sol" are used, respectively, to denote formation and solution reactions.

References

[1] Goldberg, R.N., Kishore, N., Tewari, Y.B. (1999) Thermodynamic studies of enzyme-catalyzed reactions. *In*: Chemical Thermodynamics for the 21st Century. (Letcher, T., Ed), Blackwell Science, Oxford; pp. 291 – 300.

[2] Krebs, H.A., Kornberg, H.L., Burton, K. (1957) *A Survey of the Energy Transformations in Living Matter.* Springer-Verlag, Berlin.

[3] Fell, D.A. (1992) Metabolic control analysis: A survey of its theoretical and experimental development. *Biochem J.* **286**:313 – 330.

[4] Kashiwaya, Y., Sato, K., Tsuchiya, N., Thomas, S., Fell, D.A., Veech, R.L., Passonneau, J.V. (1994) Control of glucose-utilization in working perfused rat-heart. *J. Biol. Chem.* **269**:25502 – 25514.

[5] 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. (1994) Recommendations for nomenclature and tables in biochemical thermodynamics. *Pure Appl. Chem.* **66**:1641 – 1666.

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

[7] Clarke, E.C.W., Glew, D.N. (1966) Evaluation of thermodynamic functions from equilibrium constants. *Trans. Faraday Soc.* **62**:539 – 547.

[8] Clarke, E.C.W., Glew, D.N. (1980) Evaluation of Debye-Hückel limiting slopes for water between 0 and 150 °C. *J. Chem. Soc., Faraday Trans. 1* **76**:1911 – 1916.

[9] Goldberg, R.N., Tewari, Y.B., Bell, D., Fazio, K., Anderson, E. (1993) Thermodynamics of enzyme-catalyzed reactions: Part 1. Oxidoreductases. *J. Phys. Chem. Ref. Data* **22**:515 – 582.

[10] Goldberg, R.N., Tewari, Y.B. (1994) Thermodynamics of enzyme-catalyzed reactions: Part 2. Transferases. *J. Phys. Chem. Ref. Data* **23**:547 – 617.

[11] Goldberg, R.N. Tewari, Y.B. (1994) Thermodynamics of enzyme-catalyzed reactions: Part 3. Hydrolases. *J. Phys. Chem. Ref. Data* **23**:1035 – 1103.

[12] Goldberg, R.N., Tewari, Y.B. (1995) Thermodynamics of enzyme-catalyzed reactions: Part 4. Lyases. *J. Phys. Chem. Ref. Data* **24**:1669 – 1698.

[13] Goldberg, R.N., Tewari, Y.B. (1995) Thermodynamics of enzyme-catalyzed reactions: Part 5. Isomerases and ligases. *J. Phys. Chem. Ref. Data* **24**:1765 – 1801.

[14] Goldberg, R.N. (1999) Thermodynamics of enzyme-catalyzed reactions: Part 6 – 1999 update *J. Phys. Chem. Ref. Data* **28**:931 – 965.

[15] Goldberg, R.N., Tewari, Y.B., Bhat, T. N. (2007) Thermodynamics of enzyme-catalyzed reactions: Part 7 – 2007 Update. *J. Phys. Chem. Ref. Data* **36**:1347 – 1397.

[16] Goldberg, R.N., Tewari, Y.B., Bhat, T.N. (2004) Thermodynamics of enzyme-catalyzed reactions – a database for quantitative biochemistry. *Bioinformatics* **20**:2874 – 2877; http://xpdb.nist.gov/enzyme_thermodynamics/

[17] Pettit, L.D., Powell, K.J. (2000) *The IUPAC Stability Constants Database*. Academic Software, York, U.K.; http://www.acadsoft.co.uk/

[18] Martell, A.E., Smith, R.M., Motekaitis, R.J. (2003) *NIST Critically Selected Stability Constants of Metal Complexes Database, NIST Standard Reference Database 46, Version 8.0*. National Institute of Standards and Technology, Gaithersburg, MD; http://www.nist.gov/srd/nist46.htm

[19] Perrin, D.D., Dempsey, B., Serjeant, E.P. (1981) *pKa Prediction for Organic Acids and Bases.* Chapman and Hall, London.

[20] Akers, D.L., Goldberg, R.N. (2001) BioEqCalc: A package for performing equilibrium calculations on biochemical reactions. *Mathematica J.* **8**:86 – 113.

[21] Alberty, R.A., *Basic Data for Biochemistry*: http://library.wolfram.com/infocenter/MathSource/5704/

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

[23] Wagman, D.D., Evans, W.H., Parker, V.B., Schumm, R.H., Halow, I., Bailey, S.M., Churney, K.L., Nuttall, R.L. (1982) The NBS tables of chemical thermodynamic properties. *J. Phys. Chem. Rev. Data* **11,** Supplement 2.

[24] Rossini, F.D., Wagman, D.D., Evans, W.H., Levine, S.; Jaffe, I. (1952) *Selected Values of Chemical Thermodynamic Properties.* National Bureau of Standards (U.S.) Circular 500, Washington, D.C.

[25] Wilhoit, R.C. (1969) *Thermodynamic Properties of Biochemical Substances*. *In:* Biochemical Microcalorimetry. (Brown, H.D., Ed), Academic Press, New York; pp. 33 – 81, 305 – 317.

[26] Goldberg, R.N., Tewari, Y.B. (1989) Thermodynamic and transport properties of carbohydrates and their monophosphates: the pentoses and hexoses. *J. Phys. Chem. Ref. Data* **18**:809 – 880.

[27] Miller, S.L., Smith-Magowan, D. (1990) The thermodynamics of the Krebs cycle and related compounds. *J. Phys. Chem. Ref. Data* **19**:1049 – 1073.

[28] Boerio-Goates, J., Francis, M.R., Goldberg, R.N., Ribiero da Silva, M.A.V., Ribiero da Silva, M.D.M.C., Tewari, Y.B. (2001) Thermochemistry of adenosine. *J. Chem. Thermodyn.* **33**:929 – 947.

[29] Alberty, R.A. Thermodynamic properties of enzyme-catalyzed reactions involving guanine, xanthine, and their nucleosides and nucleotides. *Biophys. Chem.* **121**:157 – 162.

[30] Goldberg, R.N., Bell, D., Tewari, Y.B., McLaughlin, M.A. (1991) Thermodynamics of hydrolysis of oligosaccharides. *Biophys. Chem.* **40**:69 – 76.

[31] Tewari, Y.B., Goldberg, R.N. (1991) Thermodynamics of hydrolysis of disaccharides. Lactulose, α-D-melibiose, palatinose, D-trehalose, D-turanose, and 3-*O*-β-D-galacto-pyranosyl-D-arabinose. *Biophys. Chem.* **40**:59 – 67.

[32] Benson, S.W., Buss, J.H. (1958) Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *J. Chem. Phys.* **29**:546 – 572.

[33] Benson, S.W. (1968) *Thermochemical Kinetics*, J. Wiley & Sons, New York.

[34] Benson, S.W., Cruickshank, F.R., Golden, D.M., Haugen, G.R., O'Neal, H.E., Rodgers, A.S., Shaw, R., Walsh, R. (1969) Additivity rules for the estimation of thermochemical properties. *Chem. Rev.* **69**:279 – 324.

[35] Domalski, E.S., Hearing, E.D. (1993) Estimation of the thermodynamic properties of C-H-N-O-S-halogen compounds at 298.15 K. *J. Phys. Chem. Ref. Data* **22**:805 – 1159.

[36] Domalski, E.S. (1998) Estimation of enthalpies of formation of organic compounds at infinite dilution in water at 298.15 K. *In:* Computational Thermochemistry – Prediction and Estimation of Molecular Thermodynamics (Irikura, K.K., Frurip, D.J. (Eds). ACS Symposium Series No. 677, American Chemical Society, Washington, D.C.

[37] Cabani, S., Gianni, P., Mollica, V., Lepori, L. (1981) Group contributions to the thermodynamic properties of non-ionic organic solutes in dilute aqueous solution. *J. Solution Chem.* **10**:563 – 595.

[38] Mavrovouniotis, M.L. (1991) Estimation of standard Gibbs energy changes of biotransformations. *J. Biol. Chem.* **266**:1440 – 1445.

[39] Mavrovouniotis, M.L. (1991) Group contributions for estimating standard Gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. Bioeng.* **36**:1070 – 1082.

[40] Alberty, R.A., Goldberg; R.N. (1993) Calorimetric determination of the standard transformed enthalpy of a biochemical reaction at specified pH and pMg. *Biophys. Chem.* **47**:213 – 223.

[41] Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Montgomery, Jr., J.A., Vreven, T., Kudin, K.N., Burant, J.C., Millam, J.M., Iyengar, S.S., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani, G., Rega, N., Petersson, G.A., Nakatsuji, H., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene, M., Li, X., Knox, J. E., Hratchian, H.P., Cross, J.B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R.E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J.W., Ayala, P.Y., Morokuma, K., Voth, G.A., Salvador, P., Dannenberg, J.J., Zakrzewski, V.G., Dapprich, S., Daniels, A.D., Strain, M.C., Farkas, O., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Ortiz, J.V., Cui, Q., Baboul, A.G., Clifford, S., Cioslowski, J., Stefanov, B. B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Martin, R.L., Fox, D. J., Keith, T., Al-Laham, M.A., Peng, C.Y., Nanayakkara, A., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Gonzalez, C., and Pople, J.A. (2004) *Gaussian 03, Revision C.02*, Gaussian, Inc., Wallingford, CT.

[42] Kast, P., Tewari, Y.B., Wiest, O., Hilvert, D., Houk, K.N., Goldberg, R.N. (1997) Thermodynamics of the conversion of chorismate to prephenate: Experimental results and theoretical predictions. *J. Phys. Chem. B* **101**:10976 – 10982.

[43] Tewari, Y.B., Chen, J., Holden, M.J., Houk, K.N., Goldberg, R.N. (1998) A thermodynamic and quantum chemical study of the conversion of chorismate to (pyruvate + 4-hydroxybenzoate). *J. Phys. Chem. B* **102**:8634 – 8639.

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

Effects of pH in Biochemical Thermodynamics and Enzyme Kinetics

Robert A. Alberty

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

E-Mail: alberty@mit.edu

Received: 3rd November 2007 / Published: 20th August 2008

Abstract

In biochemical thermodynamics, the apparent equilibrium constants of enzyme-catalyzed reactions have been represented by $K' = K_{ref}10^{n\mathrm{pH}}f(\mathrm{pH})$, where K_{ref} is a reference chemical reaction, n is the number of hydrogen ions in the reference reaction, and f(pH) is a function of pH that brings in the pKs of the substrates. This equation suggests that hydrogen ions are involved in two different ways in biochemical thermodynamics. If this is true in thermodynamics, it has to be true in kinetics. However, the choice of reference reaction in thermodynamics is arbitrary, and so n cannot be determined from equilibrium measurements. However, when hydrogen ions are consumed in the rate-determining reaction, the experimental limiting velocity of the forward reaction is given by $V_{fexp} = 10^{n\mathrm{pH}}V_f$. V_f is the limiting velocity in the forward direction when $n = 0$, or V_f can be calculated from experimental data using $V_f = 10^{n\mathrm{pH}}V_{fexp}$. V_f brings in the pKs of the enzyme-substrate complex that reacts in the rate-determining reaction. When hydrogen ions are consumed in the rate-determining reaction, the Haldane equation yields $K' = K_{ref}10^{n\mathrm{pH}}f(\mathrm{pH})$. Since n can be -8 (EC 1.7.7.1), the effects of pH on kinetic and thermodynamic properties can be very large. webMathematica can provide the thermodynamic properties of enzyme-catalyzed reactions that are difficult to calculate and require a database without having Mathematica® in a personal computer or knowing how to use it.

Introduction

Since 1992 I have been concentrating on biochemical thermodynamics, and so when I attended ESCEC 2005 I looked at it as a biochemical thermodynamist. But that conference got me interested in enzyme kinetics again; I say "again" because I had worked on enzyme kinetics in 1950 – 1966. My recent experience in biochemical thermodynamics has made me look at enzyme kinetics in a very different way than I did in 1950 – 1966. As a bridge between biochemical thermodynamics and enzyme kinetics, the first thing I did was to write a paper "Relations between Biochemical Thermodynamics and Enzyme Kinetics" [1]. Since I was interested in pH effects, it was natural to use the rapid-equilibrium assumption because so many p*K*s and chemical equilibrium constants have to be included in the derivation of the rate equation. My respect for enzyme kinetics grew when I realized that enzyme kinetics includes all of biochemical thermodynamics as a special case, that is, when the reaction rate is zero.

Now I want to tell you about some new ideas in enzyme kinetics that in a sense have their origin in biochemical thermodynamics. When I was working on the kinetics of the conversion of fumarate to malate a long time ago, we recognized that the pH dependence of the apparent equilibrium constant for the fumarase reaction is given by $K' = K_{ref} f(\mathrm{pH})$, where K_{ref} is the equilibrium constant for a chemical reaction written in terms of species and f(pH) brings in the p*K*s of fumarate and malate. Later, when I became interested in the thermodynamics of the hydrolysis of ATP to ADP, I recognized that the pH dependence of the apparent equilibrium constant for that reaction is described by

$$K' = K_{ref} 10^{\mathrm{pH}} f(\mathrm{pH}) \qquad (1)$$

where the reference reaction is $ATP^{4-} + H_2O = ADP^{3-} + H_2PO_4^- + H^+$ [2]. This equation can be generalized to $K' = K_{ref} 10^{n\mathrm{pH}} f(\mathrm{pH})$, where n is an integer: positive if produced, negative if consumed. This type of expression for an apparent equilibrium constant has been used many times. Tewari and Goldberg used it in 1988 [3] for the conversion of penicillin G to 6-aminopenicillanic acid in 1988. Athel Cornish-Bowden and I used it in an article about pH effects in *Trends Biochem. Sci*. in 1993 [4].

But I want to caution you about this equation. In thermodynamics, the choice of reference reaction is arbitrary. Therefore, you can use the reference reaction $H^+ + ATP^{4-} + H_2O = HADP^{2-} + H_2PO_4^-$, which leads to $K' = K_{ref} 10^{-\mathrm{pH}} f(\mathrm{pH})$, rather than equation 1. Thus, if n means anything in kinetics, it has to be determined by rate measurements. Thermodynamics is independent of mechanism, and so it cannot be used to determine n in a mechanistic sense. However, the equation $K' = K_{ref} 10^{n\mathrm{pH}} f(\mathrm{pH})$ is important because it suggests that there are two different ways that hydrogen ions can be involved in biochemical thermodynamics and enzyme kinetics. This is especially important because the $10^{n\mathrm{pH}}$ effect is huge, since npH is in the exponent and the effect extends over the whole range of pH [5].

Derivation of Rate Equations when Hydrogen Ions are Consumed in the Rate-determining Reaction

Now I want to show that when the rapid-equilibrium assumption is used, *n* can be determined from rate measurements. One of the reasons this can be done is that enzyme kinetics is something like electrochemistry because you can use the idea of half reactions. In rapid-equilibrium enzyme kinetics, there is a forward half reaction and a reverse half reaction. I developed this point of view in writing a recent article for the *Journal of Chemical Education* [6] on a faster way to derive rapid-equilibrium rate equations (incidentally, that article contains rapid-equilibrium rate equations and Haldane equations fifteen different mechanisms). In studying kinetics of the forward reaction, you are learning about the properties of half reactions.

When a hydrogen ion is consumed in a reduction reaction, the hydrogen ion shows up in the chemical reference reaction. For example, consider the chemical reference reaction for EC 1.1.1.1.

$$H^+ + NAD_{red}(H_{27})^{2-} + \text{acetaldehyde}(H_4) = \text{ethanol}(H_6) + NAD_{ox}(H_{26})^- \quad (2)$$

In enzyme kinetics this hydrogen ion has to be consumed in the rate-determining reaction because its pH effects extend over the whole pH range of interest.

The following mechanism for ordered A + B = ordered P + Q includes two p*K*s for each reactant, the enzymatic site, and each enzyme-substrate complex. The equal signs indicate reactions that are assumed to be equilibrated rapidly and ↔ indicates the rate-determining reaction where a single hydrogen ion is consumed.

$$\begin{array}{ccccc} E^- & & HA^- & & H_2EA^- \\ \| & & \| & & \| \\ HE & + & H_2A & = & H_3EA \\ \| & & \| & & \| \\ H_2E^+ & & H_3A^+ & & H_4EA^+ \end{array} \quad (3)$$

$$\begin{array}{ccccc} H_2EA^- & & B^- & & H_3EAB^- \\ \| & & \| & & \| \\ H_3EA & + & HB & = & H_4EAB \\ \| & & \| & & \| \\ H_4EA^+ & & H_2B^+ & & H_5EAB^+ \end{array} \quad (4)$$

$$\begin{array}{ccc} H_3EAB^- & & H_4EPQ \\ \| & & \| \\ H^+ + H_4EAB & \underset{k_r}{\overset{k_f}{\rightleftharpoons}} & H_5EPQ^+ \\ \| & & \| \\ H_5EAB^+ & & H_6EPQ^{2+} \end{array} \tag{5}$$

$$\begin{array}{ccccc} H_4EPQ & & HEQ & & H_2P^- \\ \| & & \| & & \| \\ H_5EPQ^+ & = & H_2EQ^+ & + & H_3P \\ \| & & \| & & \| \\ H_6EPQ^{2+} & & H_3EQ^{2+} & & H_4P^+ \end{array} \tag{6}$$

$$\begin{array}{ccccc} HEQ & & E^- & & Q \\ \| & & \| & & \| \\ H_2EQ^+ & = & HE & + & HQ^+ \\ \| & & \| & & \| \\ H_3EQ^{2+} & & H_2E^+ & & H_2Q^+ \end{array} \tag{7}$$

The expressions for the pH dependencies of the four Michaelis constants for this mechanism are quite complicated (according to this mechanism, there are seven kinetic constants for each Michaelis constant, but the effects due to the p*K*s of substrates can be taken out. The p*K*s of substrates can be determined by using acid titrations of the substrates, and taking them out of the pH dependencies of the Michaelis constants simplifies the determination of the other p*K*s in the mechanism.

The rapid-equilibrium rate equation for this mechanism is

$$v = \frac{\dfrac{10^{-pH}V_f[A][B]}{K_{IA}K_B} - \dfrac{V_r[P][Q]}{K_PK_{IQ}}}{1 + \dfrac{[A]}{K_{IA}} + \dfrac{[A][B]}{K_{IA}K_B} + \dfrac{[Q]}{K_{IQ}} + \dfrac{[P][Q]}{K_{IQ}K_P}} = \frac{\dfrac{V_{fexp}[A][B]}{K_{IA}K_B} - \dfrac{V_r[P][Q]}{K_PK_{IQ}}}{1 + \dfrac{[A]}{K_{IA}} + \dfrac{[A][B]}{K_{IA}K_B} + \dfrac{[Q]}{K_{IQ}} + \dfrac{[P][Q]}{K_{IQ}K_P}} \tag{8}$$

This is a new rate equation because of the 10^{-pH}. This factor is a result of the consumption of one hydrogen ion in the rate-determining reaction. I use the symbol V_{fexp} in the second form of the rate equation because V_{fexp} is the property obtained in making a Lineweaver-Burk plot at a specified pH. Equation 8 shows that the experimental limiting velocity in the forward direction is $V_{fexp} = 10^{-pH}V_f$, where

$$V_f = \frac{k_f[E]_t}{1 + 10^{pH - pK1EAB} + 10^{pK2EAB - pH}} \tag{9}$$

Note $pK_1 > pK_2$. V_{fexp} and the other kinetic parameters can be obtained from rate measurements, but V_f has to be calculated using $V_f = 10^{pH} V_{fexp}$ for this mechanism.

According to this mechanism, V_f yields a bell-shaped plot, but V_{fexp} increases as the pH is reduced, as shown by $V_{fexp} = 10^{-pH} V_f$. Mechanism 3 to 7 can be generalized by replacing 10^{-pH} with 10^{npH} so that one, or more hydrogen ions can be consumed in the rate-determining reaction.

Use of Mathematica to Derive Rapid-equilibrium Rate Equations

Mathematica® [7] is very useful for deriving equations for various kinetic properties and making plots and tables. An example of a program is **derordAB** (see Appendix) for the forward reaction for the ordered mechanism of A + B → *products* when hydrogen ions are consumed in the rate-determining reaction [8]. This program is for a reaction like alcohol dehydrogenase where none of the reactants have p*K*s in the range pH 5 – 9. The required input is specified, and the program derives the expressions for V_{fexp}, V_f, K_{IA}, and K_B as functions of pH. It also derives the initial velocity v as a function of [A], [B], and the pH. The pH dependencies for V_f/K_B, and $V_f/K_{IA}K_B$ that are also produced by the program are useful for making bell-shaped plots of experimental data for the determination of kinetic parameters.

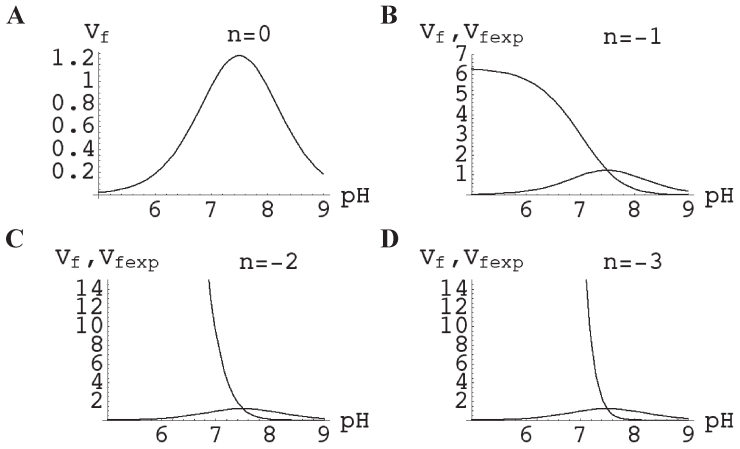


Figure 1. (A) Plot of V_f when $pK_{1EAB} = 8$, $pK_{2EAB} = 7$, $k_f[E]_t = 2$, and $n = 0$. (B) Composite plot of V_f and V_{fexp} when n=-1. (C) Composite plot of V_f and V_{fexp} when n=-2. (D) Composite plot of V_f and V_{fexp} when n=-3.

Figure 1 shows the pH dependencies of V_f and V_{fexp} when n= 0, -1, -2 and -3. V_f produces a bell-shaped plot with pK_{1EAB}= 8 and pK_{2EAB}= 7. The calculation of this figure has used $V_{fexp}= 10^{n(pH-7.5)}V_f$ so that the curves cross at pH 7.5; otherwise the two functions have very different magnitudes. Changing the 7.5 does not alter conclusions about the pH dependence of V_{fexp}.

Equation 8 shows that n can be determined from the initial rate of the forward reaction. If hydrogen ions are produced in the rate determining reaction, they do not affect the rate equation for the initial forward velocity.

The Haldane equation obtained from equation 8 is

$$K' = \frac{[P][Q]}{[A][B]} = \frac{10^{-pH}V_f K_P K_{IQ}}{V_r K_{IA} K_B} = \frac{V_{fexp} K_P K_{IQ}}{V_r K_{IA} K_B}$$
$$= \frac{K_{ref}10^{-pH}(1+10^{pH-pK1P}+10^{pK2P-pH})(1+10^{pH-pK1Q}+10^{pK2Q-pH})}{(1+10^{pH-pK1A}+10^{pK2A-pH})(1+10^{pH-pK1B}+10^{pK2B-pH})} \quad (10)$$

This shows the form of f(pH) that occurs in equation 1. It also shows that for this mechanism n=-1. This is a new type of Haldane equation. It can be generalized by replacing 10^{-pH} with 10^{npH}. The Haldane equation yields an expression for K' that is of the form of K' $=K_{ref}10^{npH}f$(pH).

When V_{fexp}, V_r, K_{IA}, K_B, K_P, and K_{IQ} are determined using the initial reaction rates of the forward and reverse reactions at a specified pH, the right value of K' is obtained at each pH. However, to determine the pKs and $k_f[E]_t$ in the expression for V_f, it is necessary to use $V_f= 10^{npH}V_{fexp}$. This is the way that n can be obtained from kinetic measurements.

An Ordered Mechanism for a Type of Oxidoreductase Reaction

The pH effects due to the consumption of hydrogen ions can be very large for some oxidoreductase reactions as, for example, the nitrite-ferredoxin reductase reaction:

$$\text{EC 1.7.7.1 Nitrite} + 6\text{ferredoxin}_{red} = \text{ammonia} + 6\text{ferredoxin}_{ox} + 2\,H_2O \quad (11)$$

This reaction consumes 8 hydrogen ions (4 to make ammonia at pHs below about 9 and 4 to make 2 H_2O). Apparent equilibrium constants for this reaction have been calculated using existing tables of standard Gibbs energies of formation of species [9, 10]. The apparent equilibrium constant for reaction 11 decreases extremely rapidly with increasing pH. At pH

5, it is 1.2×10^{91} and at pH 9 it is 1.9×10^{59}. The change in binding of hydrogen ions calculated using $\Delta_r N_H = -d\log K'/d\mathrm{pH}$ is essentially 8 across this range of pH. The mechanism for this and related oxidoreductase reactions can be represented by [11]

$$\mathrm{E} + m\mathrm{R} = \mathrm{ER}_m \qquad K^m_{\mathrm{IR}m} = \frac{[\mathrm{E}][\mathrm{R}]^m}{[\mathrm{ER}_m]} \tag{12}$$

$$\mathrm{ER}_m + \mathrm{O} = \mathrm{ER}_m\mathrm{O} \qquad K_\mathrm{O} = \frac{[\mathrm{ER}_m][\mathrm{O}]}{[\mathrm{ER}_m\mathrm{O}]} \tag{13}$$

$$\mathrm{ER}_m\mathrm{O} \rightarrow \text{products} \tag{14}$$

where R is the reductant and O is the oxidant. Equal signs indicate equilibria that are adjusted rapidly, and so the Michaelis constants $K_{\mathrm{IR}m}$ and K_O are apparent equilibrium constants that are functions of pH. The Michaelis constant $K_{\mathrm{IR}m}$ is used, rather than $K^m_{\mathrm{IR}m}$ because it has the units of a concentration. The rapid-equilibrium rate equation for the forward reaction is

$$v = \frac{V_{\mathrm{fexp}}}{1 + \frac{K_\mathrm{O}}{[\mathrm{O}]} + \frac{K^m_{\mathrm{IR}m} K_\mathrm{O}}{[\mathrm{R}]^m[\mathrm{O}]}} = \frac{10^{-8\mathrm{pH}} V_\mathrm{f}}{1 + \frac{K_\mathrm{O}}{[\mathrm{O}]} + \frac{K^m_{\mathrm{IR}m} K_\mathrm{O}}{[\mathrm{R}]^m[\mathrm{O}]}} \tag{15}$$

Note that the pH factor is 10^{-40} at pH 5 and 10^{-72} at pH 9. This sensitivity to pH will make it almost impossible to determine the kinetic parameters for reaction 11.

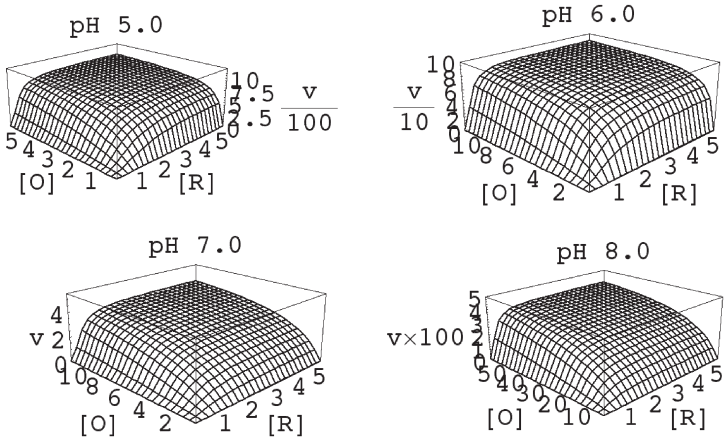


Figure 2. Three-dimensional plots of initial reaction velocities for O + 2R → *products* versus reactant concentrations at specified pHs for arbitrary constants. For this ordered mechanism n is -2 and m= 2. The scale of the ordinate has been changed by 10^4 between pH 5 and pH 9. Each plot is made up of 20 Michaelis plots at constant [A] and 20 Michaelis plots at constant [B].

Mathematica® is very useful for displaying initial reaction velocities as functions of the reactant concentrations and the pH. A computer program was written to derive the initial rate equation for mechanism 12 – 14. The chemical equilibrium constants and the value of *n* have to be specified. This rate equation can be treated like an actual reaction system in the sense that the initial velocity *v* can be calculated at various [R], [O], and pH. Using Mathematica®, *v* can be presented as a surface in a 3-dimensional plot at a specific pH [11]. Figure 2 shows these plots for the catalyzed reaction O + 2R → *products* as functions of [O] and [R] at four pHs, for *n*=-2 and arbitrary p*K*s and other constants.

Note that the plots of *v* versus [R] are sigmoid. A sigmoid plot of *v* versus the concentration of a reactant is usually taken to be an indication of allosterism. This can arise when there is positive cooperativity between active sites of a polymeric enzyme. But a sigmoid plot can have other origins. In this case, it is caused by the stoichiometric number of R in the biochemical equation for the forward reaction O + 2R → *products*.

Three-dimensional Lineweaver-Burk Plots for the Forward Reaction Ordered A + B

Mathematica® can also be used to plot reciprocal velocities 1/*v* versus 1/[A] and 1/[B]. This is illustrated for the ordered mechanism for A + B → *products* [12]. These plots can be called three-dimensional Lineweaver-Burk plots. These plots can be made for various pHs and for *n*= 0, -1, -2, ... They are given in Figure 3 only for arbitrary constants and pHs 5, 6, 7, and 8, with *n*=-1.

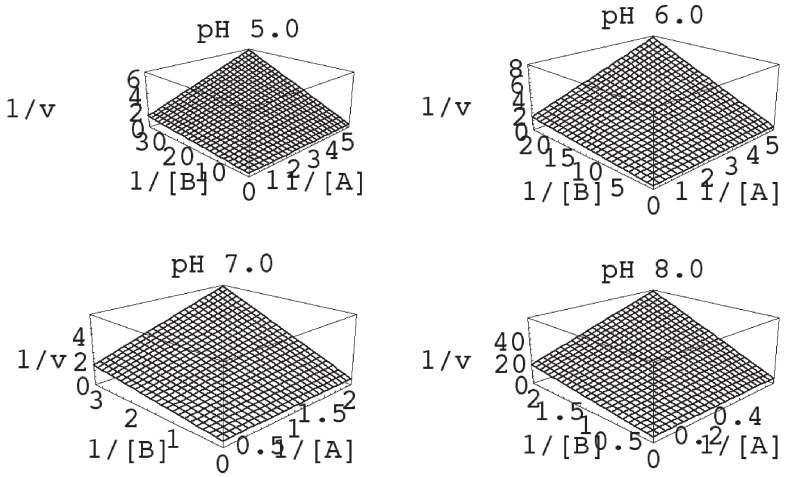


Figure 3. Reciprocal initial velocities for the ordered mechanism for A + B → *products* versus 1/[A] and 1/[B] at pHs 5, 6, 7, and 8 for *n*=-1 and arbitrary constants.

The plots in Figure 3 are referred to in mathematics as ruled surfaces, and quite a bit has been written about ruled surfaces. Each of the surfaces in Figure 3 is characterized by three constants, the experimental limiting velocity V_{fexp}, and Michaelis constants K_{IA} and K_B at the specified pH. These surfaces are not planes, but the lines are all straight. When I first made one of these plots, I recognized that each of these surfaces is really determined by the ordinates of three of the corners: one corner can be taken a reference because the concentration of the enzyme is arbitrary. And the velocities depend on three kinetic constants. Therefore, you should be able to calculate the three kinetic constants from three velocity measurements.

webMathematica

webMathematica provides a way to take advantage of the computing power of Mathematica® to perform calculated calculations without having Mathematica® in your computer or knowing how to use it. An example of such a calculation is the calculation of the apparent equilibrium constant *K*' for an enzyme-catalyzed reaction at a desired temperature, pH, and ionic strength. This requires a database of species properties. BasicBiochemData3 [13, 14] provides this data. WebMathematics presents a screen with boxes for inputting the enzyme-catalyzed reaction, desired temperature, pH, and ionic strength. This information is sent to a server that has Mathematica® and a database on it so the server can calculate the apparent equilibrium constant (or other thermodynamic property) and present it on the screen of the user.

webMathematica programming requires knowledge of Mathematica® and HTML, and it has been done by Dr. Violeta Ivanova of MIT's Office of Educational Innovation and Technology. This program can be extended to calculate the change in binding of hydrogen ions in an enzyme-catalyzed reaction. BasicBiochemData3 [13, 14] makes it possible to calculate these properties for about 300 different reactions at 298.15 K. For about 100 reactions these calculations can be made at other temperatures in the range 273.15 K to about 313.15 K. webMathematica can be used to calculate other properties in biochemical thermodynamics and enzyme kinetics.

Acknowledgements

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

Appendix

derordAB[pK1e_,pK2e_,pK1ea_,pK2ea_,pK1eab_,pK2eab_,kfEt_,n_,kHEA_,kHEAB_]:= -
Module[{eafactor,efactor,abfactor,vf,vfexp,kia,kb,v},(*Calculates kinetic parameters of the forward enzyme-catalyzed reaction ordered A+B = products as functions of pH. The output is a list of 6 functions of pH: vfexp,vf,kia,kb,vf/kb, and vf/kiakb. v is a function of [A]= a, [B]= b, and pH. The 7.5 in (pH-7.5) can be changed because it is equivalent to changing the value of kfEt.*)
efactor = $1+10^{pK2e-pH}+10^{pH-pK1e}$;
eafactor = $1+10^{pK2ea-pH}+10^{pH-pK1ea}$;
eabfactor = $1+10^{pK2eab-pH}+10^{pH-pK1eab}$;
vf = kfEt/eabfactor;
vfexp =(10^(n*(pH-7.5)))*vf;
kia = kHEA*efactor/eafactor;
kb = kHEAB*eafactor/eabfactor;
v = vfexp/(1+(kb/b)*(1+(kia/a)));
{vfexp,vf,kia,kb,vf/kb,vf/(kia*kb),v}]

References

[1] Alberty, R.A. (2006) Relations between Biochemical Thermodynamics and Biochemical Kinetics. *Biophys. Chem.* **124**:11 – 17.

[2] Alberty, R.A. (1968) Effect of pH and metal ion concentrations on the equilibrium hydrolysis of ATP to ADP. *J. Biol. Chem.* **243**:1337 – 1342.

[3] Tewari, Y.B. and Goldberg, R.N. (1988) Thermodynamics of the conversion of penicillin G to 6-aminopenicillanic acid. *Biophys. Chem*. **29**:245.

[4] Alberty, R.A. and Cornish-Bowden, A. (1993) On the pH dependence of the apparent equilibrium constant *K*' of a biochemical reaction. *Trends Biochem. Sci*. **18**:288 – 291.

[5] Alberty, R.A. (2007) Two Different Ways that Hydrogen Ions are Involved in the Thermodynamics and Rapid-Equilibrium Kinetics of the Enzyme-Catalysis of S = P and S + H_2O = P. *Biophys. Chem.* **128**:204 – 209.

[6] Alberty, R.A. (2007) Rapid-Equilibrium Rate Equations. *J. Chem. Educ*. In press. Wolfram Research, 100 World Trade Center Drive, Champaign, IL. http://www.wolfram.com

[7] Alberty, R.A. (2007) Effects of pH in Rapid-Equilibrium Enzyme Kinetics, *J. Phys. Chem. B*. In press.

[8] Alberty, R.A. (2007) Change in the Binding of Hydrogen Ions in Biochemical Reactions. *Biophys. Chem.* **125**:328 – 333.

[9] Alberty, R.A. (2006) Changes in the Binding of Hydrogen Ions in Biochemical Reactions. http://library.wolfram.com/infocenter/MathSource/6386

[10] Alberty, R.A. (2007) Three Mechanisms and Rapid-Equilibrium Rate Equations for a Type of Reductase Reaction. *Biophys. Chem.* In press.

[11] Alberty, R.A. (2007) Rapid-Equilibrium Rate Equations for the Enzymatic Catalysis of A + P = P + Q over a Range of pH. *Biophys. Chem.* In press.

[12] Alberty, R.A., BasicBiochemData3 (2005) http://library.wolfram.com/infocenter/MathSource/5704

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

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

The KineticsWizard: a Data Capture Tool for the Submission of Enzyme Kinetics Data

Neil Swainston

Manchester Centre for Integrative Systems Biology, University of Manchester, Manchester, M1 7DN, U.K.

E-Mail: neil.swainston@manchester.ac.uk

Received: 25th June 2008 / Published: 20th August 2008

Introduction

There are a number of resources containing enzyme kinetics data. Two widely used databases are BRENDA [1] and SABIO-RK [2]. While these databases contain kinetic constants, the key to ensuring that these resources can be usefully employed in a systems biology environment is in the richness of the metadata associated with these values.

Obvious requirements for these metadata include the environmental conditions, such as pH and temperature, under which these constants were measured. However there are other, more subtle, metadata that must also be captured and recorded along with the kinetic parameters to allow the database to be utilised correctly in modelling and simulation studies. This article describes these metadata and also introduces the KineticsWizard, a data capture tool that allows the experimentalist to specify these data in an intuitive manner.

The data capture tool is designed to be used by experimentalists, and as such, it is intended to hide much of the more technical aspects of data management from the user and present an intuitive, biologist-focussed interface from which the necessary metadata can nevertheless be input.

The tool itself is part of a larger system in which kinetic constants are determined through analysis and fitting of spectrophotometric data, and then automatically submitted to a data resource. In addition, the original raw data are also archived, providing the facility to map kinetic parameters back to their original source data, which could then be reanalysed or refitted if necessary.

Enzyme kinetics studies are a subset of the experimental programme of the Manchester Centre for Integrative Systems Biology (MCISB), and will be supplemented by both quantitative metabolomics and proteomics studies. The informatics infrastructure designed by the MCISB is one of a distributed, loosely-coupled system [3], in which a number of independent data resources are populated, and then later queried via web service interfaces in order to parameterise SBML models [4, 5].

The key to the development of such a distributed system is to ensure a consistent means of identifying species, reactions, and parameters across each of these data resources.

Problems of Currently Published Enzyme Kinetics Data

Kummer and Sahle highlighted a number of issues with enzyme kinetics data currently published in existing resources [6]. Some of these problems are summarised below.

Importance of the kinetic equation

In addition to specifying kinetic parameters, such as V_{max} and K_M, it is also necessary to specify the kinetic equation that is either assumed or was used to determine the constant. This additional information is crucial to ensure that modellers use the constant in the intended manner.

Furthermore, it is insufficient to specify the equation as a textual description, such as ''Michaelis-Menten'', as these terms can be used inconsistently and do not unambiguously describe the intended kinetic equation.

Additionally, in cases where the reaction involves species with different stoichiometries, it is important to specify to which participant the parameter applies.

The V_{max} parameter

V_{max} parameters are often specified without any indication of the enzyme concentration contained within the term. While many V_{max} parameters are generated from *in vitro* studies, the enzyme concentration is commonly either unspecified or poorly estimated. This reduces the usability of the parameter in modelling studies, as in these cases the enzyme concentration must be estimated, introducing unnecessary imprecision into the system.

Coherent unit notation

For parameters to be used reliably in modelling and simulation studies, standard units must be specified. Taking the example of enzyme concentration, it is common to see these values described by units such as ''enzyme mass per dry weight of protein'', which dramatically reduces the usability of the parameter.

Further Issues

In addition to the problems highlighted by Kummer and Sahle, there also remains the issue of inconsistent naming of reaction participants. It has been reported that relying on textual descriptions of small molecules and enzymes can result in inconsistencies, as the naming of such species is largely subjective and can differ greatly from individual to individual [7].

Capture of Enzyme Kinetics Data

A submission tool is introduced here which has been designed to provide solutions to the problems specified above. The philosophy of the tool is to provide the facility for experimentalists to submit the required metadata without unduly burdening them with some of the more tedious tasks that this entails.

To provide this, the tool draws heavily on the use of existing data resources which are relevant to the task, and queries these resources via web service interfaces where possible. Exploiting existing data resources greatly reduces the volume of data that the experimentalist must submit. For example, consider the case of specifying a publication. Rather than providing a user with an interface to specify journal name, authors, paper title, etc., the philosophy followed here is to prompt the user for a PubMed identifier (id). Once this id is associated with a given parameter, web services can be queried from which the above fields can be extracted.

Furthermore, specifying and storing metadata as external database links greatly enhances the usability of the data resource, providing it with the facility to be queried with standard terms, and greatly enhancing its usability in a distributed, web service-driven infrastructure.

To illustrate this principle, a data submission tool is introduced here to show practical implementations of this approach.

Specifying the reaction components

The tool facilitates the consistent specification of reaction components by providing an interface to the KEGG database web service [8]. The user is presented with the following interface, which allows them to specify an organism and a gene name. Upon specifying these terms, the KEGG web service is queried and all reactions catalysed by the supplied

gene are returned. (Fig. 1). An individual reaction can then be specified, and as this reaction is an entry in the KEGG database, the entry can be queried to harvest a number of terms that would otherwise have to be specified by the user.

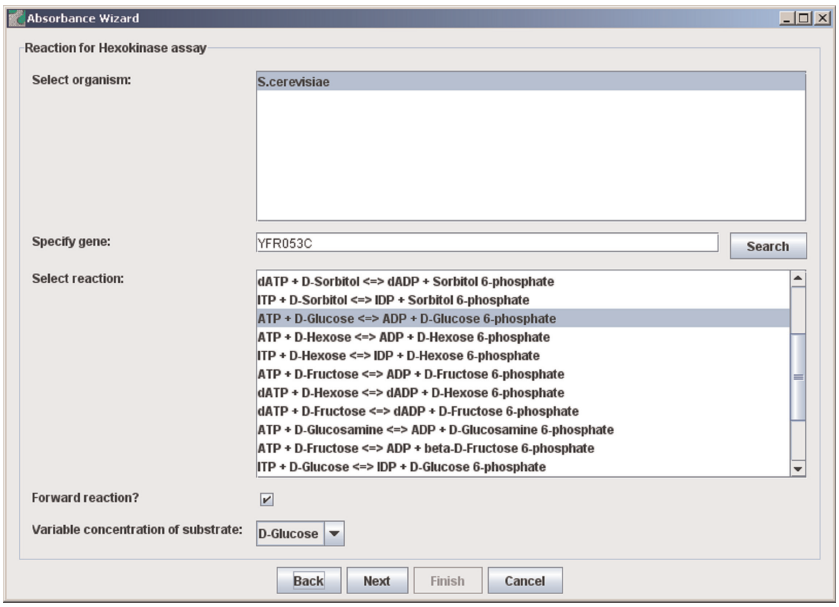


Figure 1. Specifying the reaction components.

The first of these is the enzyme itself. As the user specified a gene name, KEGG can be queried to determine the UniProt id of the enzyme. From this reference, a range of additional information can be subsequently harvested, such as protein name, synonyms, molecular weight, protein sequence, EC number and via external links to other resources, perhaps even also the protein structure. Specifying and ultimately storing metadata as external database references, rather than simply a textual name, can therefore greatly enhance the subsequent usability of this data.

The same principle applies to the reactants and products. By selecting a KEGG reaction, the web service can be queried to determine the reaction participants as entries in either KEGG or the ChEBI database[9], and also the stoichiometry of each of the participants. Consequently the selected reaction, textually presented to the user as "ATP + D-Glucose <–> ADP + D-Glucose 6-phosphate", will be defined computationally as an unambiguous set of reactants (CHEBI:15422 and CHEBI:17634), products (CHEBI:16761 and CHEBI:15954) and enzyme (UniProt:P04806).

It is also important to specify the reactant to which the parameter applies. The interface allows the user to specify this term upon selecting a reaction.

Specifying the Kinetic Equation

The specification of the kinetic equation once again exploits a web service to an existing data resource: in this case, the Systems Biology Ontology (SBO) [10]. SBO consists of a number of vocabularies, one of which – mathematical expression – can be used to describe reaction mechanism.

The user is presented with a tree of these SBO terms, and upon selection of a term, the underlying kinetic equation can be attained. (Fig. 2). This provides two advantages. The first is that the selected SBO term can be used to map to an appropriate data fitting algorithm used in the next step of the data analysis and submission pipeline. The second is that, upon submission, the kinetic parameters are immediately annotated with an unambiguous, standard term that defines the assumed mechanism of reaction and fitting. Rather than defining the reaction mechanism textually as "Michaelis-Menten", the standard SBO term SBO:0000029 is used. Using such a defined term specifies the mechanism precisely: in this case it indicates that the mechanism is considered to be irreversible. Furthermore, the user is not prompted with the requirement of specifying the kinetic equation directly – a task that would be difficult to standardise and in many cases would be prone to error. This, along with other metadata such as synonyms and a description of the term, can be accessed directly from the SB ontology.

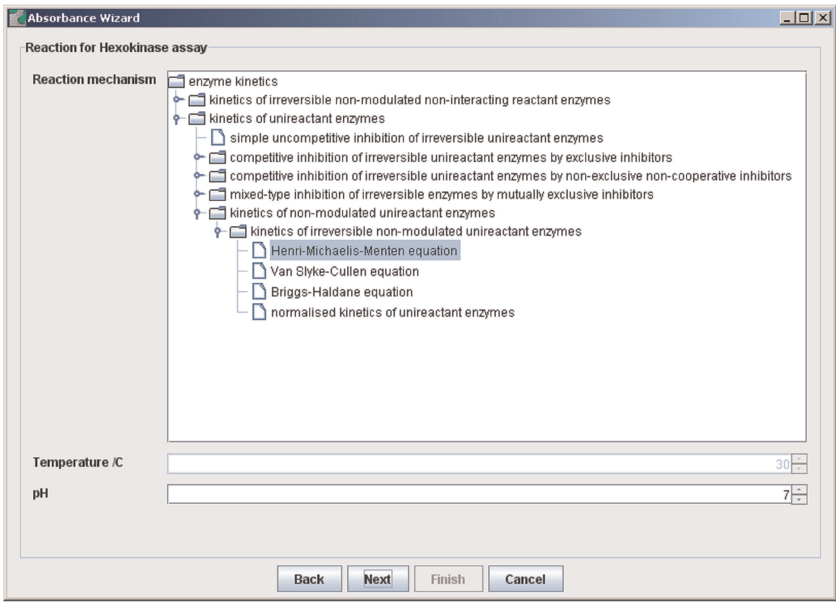


Figure 2. Specifying the kinetic equation.

A further advantage of the use of SBO is that the kinetic equations themselves refer to SBO terms to define the kinetic constants. Therefore, the constants themselves are not referred to in potentially ambiguous terms such as K_M and k_{cat}, but by the standard SBO terms SBO:0000027 and SBO:0000025 respectively.

The V_{max} parameter

Due to the nature of the kinetic assay experiments being performed, the enzyme concentration is known accurately at the time of experiment. As such, the user is prompted for this value and as such, upon data fitting, the k_{cat} value is calculated and submitted to the data resource, rather than the V_{max}.

Coherent unit notation

The user is presented with an interface that prevents the user from supplying units for the required terms. Substrate and enzyme concentrations are assumed to be accurately measured and their input is forced to be in mM and nM respectively. Dictating the units at the point of data submission ensures that these values will be consistent and not contain any ''fuzzy'', imprecise or inconsistent terms.

Discussion

The benefit of this approach is that kinetic parameters stored in data resources are associated to standardised terms, which greatly facilitates to querying of such resources. If a user then has an unparameterised but MIRIAM-compliant [11] SBML file, the task of parameterising the SBML file with kinetic parameters is greatly simplified, as both the SBML file and the data resource are annotated with consistent terms for metabolites, enzymes, EC codes, etc. Therefore, the task of automatically parameterising an annotated SBML becomes far easier, as the ambiguity that may exist between mapping metabolites or enzymes in the model with those in the database is removed.

It is the intention of submitting kinetic parameters, and their associated metadata terms, to send them to SABIO-RK automatically upon completion of the submission wizard. SABIO-RK stores kinetic parameters with references to the database terms that are collected through the submission process. Furthermore, SABIO-RK contains a web service interface that allows queries to be formed in terms of these database terms. (Fig. 3).

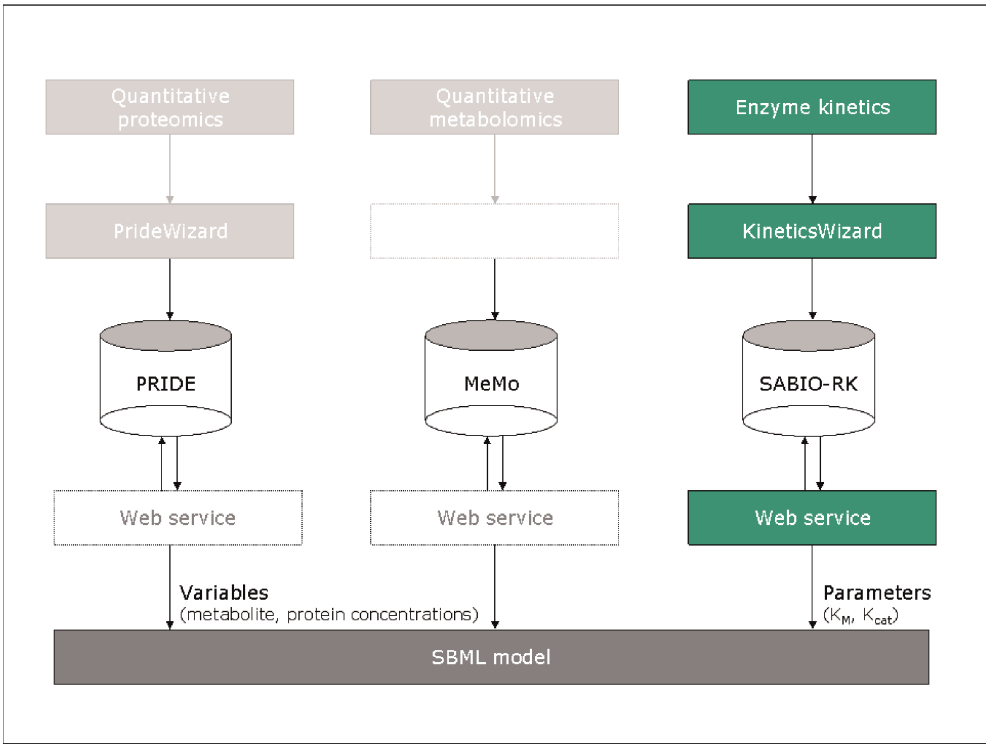


Figure 3. Overview of the MCISB informatics infrastructure. (MeMo is the MCISB metabolomics database [12]).

Whilst the importance of associating kinetics parameters with standardised database terms has been discussed, it is also necessary to ensure that this can be performed in a manner that is intuitive to the experimentalist. In order to facilitate this, existing web services are exploited such that options can be presented to users in human-readable format, while behind the scenes, many database references are gathered which allows the concept to be unambiguously specified using standardised terms.

Acknowledgements

I thank the EPSRC and BBSRC for their funding of the Manchester Centre for Integrative Systems Biology (http://www.mcisb.org/). I am also very grateful for the help and collaboration of colleagues at EML Research, specifically Isabel Rojas, Martin Golebiewski, Renate Kania, Olga Krebs, Saqib Mir and Ulrike Wittig.

References

[1] Schomburg, I., Hofmann, O., Baensch, C., Chang, A., Schomburg, D. (2000) Enzyme data and metabolic information: BRENDA, a resource for research in biology, biochemistry, and medicine. *Gene Funct. Dis*. **3 – 4**:109 – 18.

[2] Wittig, U., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann, A., Anstein, S., Saric, J., Rojas, I. (2006) SABIO-RK: Integration and Curation of Reaction Kinetics Data. *In:* Proceedings of the 3rd International workshop on Data Integration in the Life Sciences 2006 (DILS'06), Hinxton, UK. *Lecture Notes in Bioinformatics* **4075**:94 – 103.

[3] Kell, D.B. (2006) Metabolomics, modelling and machine learning in systems biology: towards an understanding of the languages of cells. The 2005 Theodor Bücher lecture. *FEBS J.* **273**:873 – 894.

[4] Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., and the rest of the SBML Forum: Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.-H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novère, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and J. Wang (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**:524 – 531.

[5] Li, P., Oinn, T., Stoiland, S., Kell, D.B. (2008) Automated manipulation of systems biology models using libSBML within Taverna workflows. *Bioinformatics* **24**:287 – 289.

[6] Kummer, U., Sahle, S. (2006) Problems of currently published enzyme kinetic data for usage in modelling and simulation. *In*: Proceedings of the 2nd International Beilstein Workshop on Experimental Standard Conditions of Enzyme Characterizations, Beilstein-Institut. Logos Verlag Berlin, pp 129 – 136.

[7] Herrgård, M, Swainston, N, *et al*. (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol.* Submitted.

[8] Kanehisa, M., Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**:27 – 30.

[9] Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**:D 344 – D 350.

[10] Le Novère, N. (2006) Model storage, exchange and integration. *BMC Neurosci*. Oct 30; 7 Suppl 1:S 11.

[11] Le Novère, N., Finney, A., Hucka, M., Bhalla, U.S., Campagne, F., Collado-Vides, J., Crampin, E.J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J.L., Spence, H.D., Wanner, B.L. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* **12**:1509 – 15.

[12] Spasiæ, I., Dunn, W.B., Velarde, G., Tseng, A., Jenkins, H., Hardy, N., Oliver, S.G, Kell, D.B. (2006) MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. *BMC Bioinformatics* **7**:281.

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

Integration and Annotation of Kinetic Data of Biochemical Reactions in SABIO-RK

Ulrike Wittig*, Renate Kania, Martin Golebiewski, Olga Krebs, Saqib Mir, Andreas Weidemann, Henriette Engelken and Isabel Rojas

Scientific Databases and Visualization Group, EML Research gGmbH, Heidelberg, Germany

E-Mail: *ulrike.wittig@eml-r.villa-bosch.de

Received: 30th April 2008 / Published: 20th August 2008

Abstract

SABIO-RK is a curated database for the systems biology community containing biochemical reactions and their kinetic properties, the latter being manually extracted from literature sources. This information is crucial for the quantitative understanding of biological systems. Modellers and wet-lab scientists alike require reliable information about reaction kinetics which is normally contained in publications generated worldwide. In SABIO-RK kinetic data are related to reactions, organisms and biological locations. The type of kinetic mechanism and corresponding rate equations are presented together with their parameters and experimental conditions. In order to enable comprehensive understanding, integration and comparison of data it is necessary to provide annotations and links to community resources, such as external databases and ontologies that augment the content and the semantics of the SABIO-RK database entries. In this short paper we will present SABIO-RK (http://sabio.villa-bosch.de/SABIORK) and our approach towards integration and annotation of the kinetic data and their respective biochemical context.

Introduction

For the understanding of complex biochemical processes, the simulation and modelling of biochemical reactions and complex networks require reliable kinetic information about the individual reaction steps within the process. Information such as the kinetic laws describing the dynamics of the reactions with their respective parameters determined under certain experimental conditions is of paramount importance. These kinetic data are mainly found in the literature and described in many different formats. Within the publications no controlled vocabulary is used for the representation of biochemical reactions, kinetic data and environmental conditions. The integration of these data needs the definition and use of standards for reporting and exchanging.

SABIO-RK extends and supplements the information content of other databases containing kinetic information (e. g. BRENDA [1], BioModels [2], and JWS Online [3]) by storing highly interrelated information about biochemical reactions and their kinetics. It includes reactants and modifying compounds (enzymes, cofactors, inhibitors or activators) of reactions. The kinetic laws with their parameters and information about experimental conditions are connected with the reaction information. The data about biochemical reactions, their rate equations and parameters can be exported in SBML (Systems Biology Markup Language) [4] file format.

The STRENDA [5] (Standards for Reporting Enzymology Data) commission is working on the definition of a standard for reporting on enzyme activity. The standard should contain the minimum amount of information that should accompany any published enzyme activity data. The use of references to controlled vocabularies and ontologies is also of great importance for the implementation of the STRENDA guidelines.

In this paper we describe new implementations in SABIO-RK and in the input interface to match the defined requirements of STRENDA.

Data Integration

The data contained in SABIO-RK are extracted from different sources, in order to establish a broad information basis. Most of the reactions, their associations with biochemical pathways and their enzymes are automatically extracted from the KEGG database [6]. Information about chemical compounds is extended additionally by data from ChEBI [7] or PubChem [8].

Kinetic data in SABIO-RK are mainly extracted from literature and inserted manually using a web-based input interface to enter the data into a temporary database. The temporary database is used by biological experts to curate the data and to insert them into the final SABIO-RK database. The main objective of the input interface is to supply a uniform format

for users entering and curating the data found in publications. Apart from this, automatic control mechanisms have been implemented to check for errors and inconsistencies during the integration process. The systems check, for example: whether all reaction participants of a reaction equation have to be defined as compound species; if all parameters in a rate equation have to be defined in the parameter list independent of whether or not the parameter value is known; if the rate equations are mathematically correct; and which (if applicable) parameter type should be related to a compound species (for example, a K_m value always should be related to a compound, however V_{max} or k_{cat} do not need a species assignment).

To avoid errors and inconsistencies within the SABIO-RK database during the first input process, the interface offers lists of controlled vocabularies for the selection of values for the following data:

- Species (compounds search function)
- Species roles (e. g. substrate, product, modifier...)
- Biochemical reactions (search functions)
- Organisms, tissues and cellular locations
- Kinetic law types (e. g. Competitive inhibition or Sequential ordered Bi Bi)
- Parameter types (e. g. K_m, k_{cat}, V_{max}, K_i, K_d, rate constant)
- Parameter units (e. g. mM, µM, 1/s, nmol/min)

Based on information in the literature, detailed information about the protein catalysing the reaction is inserted. This includes information about a specific isoenzyme or mutations used in the experiments (e. g. wildtype, mutant E540K, wildtype isoenzyme A), UniProt [9] accession numbers and information about the composition of subunits (e. g. (Q 6UG02)*4 for a homotetramer).

The data format defined in the input interface fits most of the current requirements of the STRENDA commission for reporting enzymology data. So SABIO-RK offers the structure for inserting and storing data required for a complete description of an experiment (Level 1, List A) which comprise, for example, information about the enzyme (EC number, biological location, isoenzyme information etc.) and the assay conditions. The description of enzyme activity data (Level 1, List B) includes kinetic parameters necessary for enzyme function definition (V_{max}, k_{cat}, K_m etc.), information about enzyme inhibition or activation and the description of the kinetic mechanisms including rate equations.

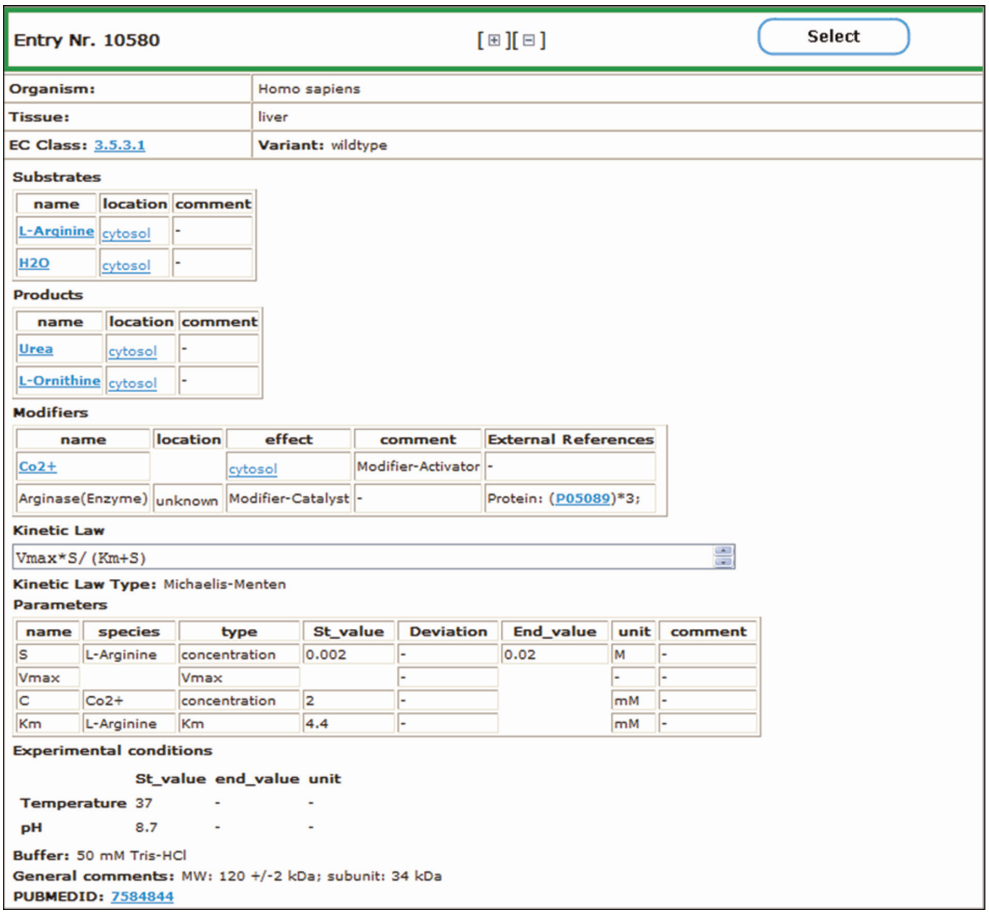
Another possible way to insert kinetic data into SABIO-RK is by using an XML-based integration tool to import a higher amount of kinetic data automatically.

Data Annotation

Scientific communication needs standards and a shared vocabulary to avoid misinterpretations. Such standards are especially important when gathering data from different sources. To identify entities or terms unambiguously and to facilitate the search, interpretation and comparison of data, these are standardized to a uniform format and structure. This implies the usage and development of controlled vocabularies. Entities and expressions in SABIO-RK are annotated to other resources and biological ontologies to clarify the biological terms used and to embed the data into its context. This enables users to collect further information through links to external databases and facilitates the integration of different database entries into kinetic models.

Biological ontologies used in SABIO-RK are ChEBI, Gene Ontology (GO) [10], Systems Biology Ontology (SBO) [11], and NCBI taxonomy [12]. ChEBI is a dictionary and ontological classification of small chemical compounds. Gene Ontology comprises controlled vocabularies for molecular functions, biological processes and cellular components of gene products. Systems Biology Ontology defines controlled vocabularies for systems biology, especially in the context of computational modelling. NCBI taxonomy is a controlled vocabulary and complex classification system of organisms. These controlled vocabularies and ontologies are the basis for the definition of the selection lists for compounds, organisms, kinetic law types, parameter types etc. in the SABIO-RK input interface.

Links to external databases in the user interface (Fig. 1) allow the user to connect to other data sources to get further information about the selected entry. Annotations are shown for chemical compounds to KEGG, ChEBI and PubChem. Enzymes are linked through their EC number to ExPASy ENZYME [13], KEGG, IntEnz [14], IUBMB [15] and Reactome [16]. Further links to KEGG are implemented for the reaction equation and the information source is referenced to PubMed [17]. Proteins or protein complexes catalysing the reactions are linked to the specific UniProt accession number(s). These annotations of proteins or protein complexes to UniProt are done manually by biological experts based on information in the publication.



Entry Nr. 10580 [⊞][⊟] Select

Organism:	Homo sapiens
Tissue:	liver
EC Class: 3.5.3.1	**Variant:** wildtype

Substrates

name	location	comment
L-Arginine	cytosol	-
H2O	cytosol	-

Products

name	location	comment
Urea	cytosol	-
L-Ornithine	cytosol	-

Modifiers

name	location	effect	comment	External References
Co2+		cytosol	Modifier-Activator	-
Arginase(Enzyme)	unknown	Modifier-Catalyst	-	Protein: (P05089)*3;

Kinetic Law

Vmax*S/(Km+S)

Kinetic Law Type: Michaelis-Menten

Parameters

name	species	type	St_value	Deviation	End_value	unit	comment
S	L-Arginine	concentration	0.002	-	0.02	M	-
Vmax		Vmax		-		-	-
C	Co2+	concentration	2	-		mM	-
Km	L-Arginine	Km	4.4	-		mM	-

Experimental conditions

	St_value	end_value	unit
Temperature	37	-	-
pH	8.7	-	-

Buffer: 50 mM Tris-HCl

General comments: MW: 120 +/-2 kDa; subunit: 34 kDa

PUBMEDID: 7584844

Figure 1. Database entry containing links to external databases and controlled vocabularies.

Many annotations to external databases and controlled vocabularies defined in the user interface are also stored in the SMBL file for exporting the information (Fig. 2).

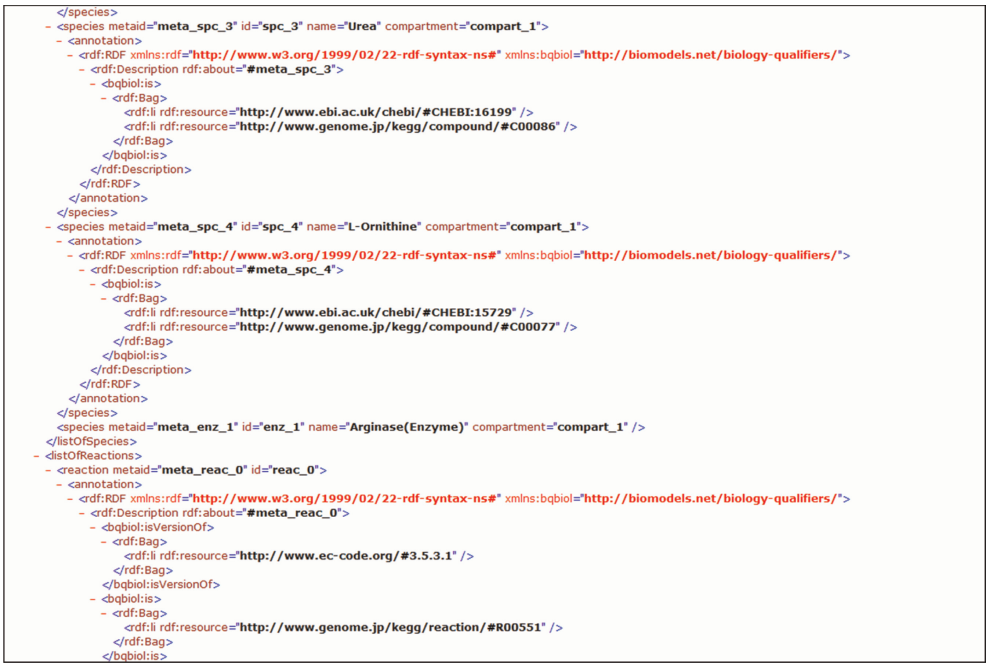


Figure 2. SBML file containing annotations to external databases and controlled vocabularies.

Future Directions

In order to offer the users more references to additional information we are working on the cross-linking and annotation of the database content to more database resources and ontologies.

One of the next steps for further SABIO-RK developments is the incorporation of detailed information about the mechanisms of the reaction to allow the user to obtain information about the kinetic properties of sub-reactions or binding mechanisms of enzymes and substrates. This includes the visualization of reaction mechanism information in the user-interface.

In the future the input interface will be further adapted to the current and changing requirements of the STRENDA commission which could comprise, for example, more detailed and structured information about experimental conditions and the methodology of measurement, purification etc. SABIO-RK will participate in further discussions about STRENDA requirements and work on their implementation. SABIO-RK could serve as the basis for a general data input and storage system for experimentalists and modellers.

Summary

SABIO-RK is a curated data resource for modellers of biochemical networks to assemble information about reactions and their kinetics. It also offers experimentalists the opportunity to obtain information about biochemical reactions and their kinetics, within the context of cellular locations, tissues and organisms. The data are extracted automatically from different databases and kinetic information is manually extracted from literature. The database uses controlled vocabularies and links to other ontologies or external databases to allow the comparison of data and to extract additional information from other sources.

Acknowledgements

We would like to thank the Klaus Tschira Foundation as well as the German Research Council (BMBF) for their funding. We thank all the student helpers, who have contributed to the population of the database.

References

[1] Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., Schomburg, D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* **32**:D431-D433.

[2] Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J.L., Hucka, M. (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.* **34:**D689-D691.

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

[4] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novère, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**:524 – 531.

[5] STRENDA: http://www.strenda.org

[6] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**:D 354 –D 735.

[7] ChEBI: http://www.ebi.ac.uk/chebi/

[8] PubChem: http://www.ncbi.nlm.nih.gov/sites/entrez? db=pccompound

[9] The UniProt Consortium. The Universal Protein Resource (UniProt)**.** (2008) *Nucleic Acids Res*. **36**:D 190 –D 195.

[10] Gene Ontology: http://www.geneontology.org/

[11] Systems Biology Ontology: http://www.ebi.ac.uk/sbo/

[12] NCBI taxonomy: http://www.ncbi.nlm.nih.gov/sites/entrez? db=taxonomy

[13] ExPASy ENZYME: http://www.expasy.org/enzyme/

[14] IntEnz: http://www.ebi.ac.uk/intenz/

[15] IUBMB: http://www.chem.qmul.ac.uk/iubmb/enzyme/

[16] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E., Stein, L. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**:Database Issue:D 428 –D 432.

[17] PubMed: http://www.pubmed.gov

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

Considerations for the Specification of Enzyme Assays Involving Metal Ions

Richard Cammack[1*] and Martin N. Hughes[2]

[1]Pharmaceutical Sciences Research Division, King's College London, 150 Stamford Street, London SE1 9NH, UK

[2]Centre for Hepatology, Royal Free & University College London Medical School, Royal Free Campus, Rowland Hill Street, Hampstead, London NW3 2PF, UK

E-Mail: *richard.cammack@kcl.ac.uk

Received: 25th July 2008 / Published: 20th August 2008

Introduction

The recommendations of the STRENDA Commission (Version 1.2 June 16th, 2006) of standard requirements for reporting enzyme activity data (http://www.strenda.org/) include the proposal that the specification of assay conditions should include any metal salts to be added. They also require the definition of some other parameters which, as will be seen later, may have a bearing on the activity of metal ion-dependent enzymes. These include assay pH, buffer type and concentrations, and other assay components such as EDTA or dithiothreitol that will coordinate to metal ions [1, 2].

This chapter is intended to provide a guide to issues that are relevant to the determination of accurate kinetic data for the reactions of metal-dependent enzymes. Of particular importance are factors relating to the speciation and availability of metal ions in the assay medium. The interaction of the metal ions in the added metal salts with compounds present in the medium may result in the formation of a number of metal-ligand complexes. These may activate the enzyme to different extents at different rates. In extreme cases, metal ions may be precipitated out of solution and be unavailable to function in enzyme activation. We will further discuss the relevance of the metal ions in modelling the activity of the enzyme in the cell.

Essential Metallic Elements in Enzymes and Other Proteins

Eleven metallic elements are known to be essential for most types of living cells; two of these occur in Group 1 of the Periodic Table of the Elements, two in Group 2 and the remaining seven are d block elements (Fig. 1). Another four metals are, or may be, necessary for some organisms. Many of these metals are required for the activity of enzymes. The function of the metal ion may be catalytic, structural or regulatory.

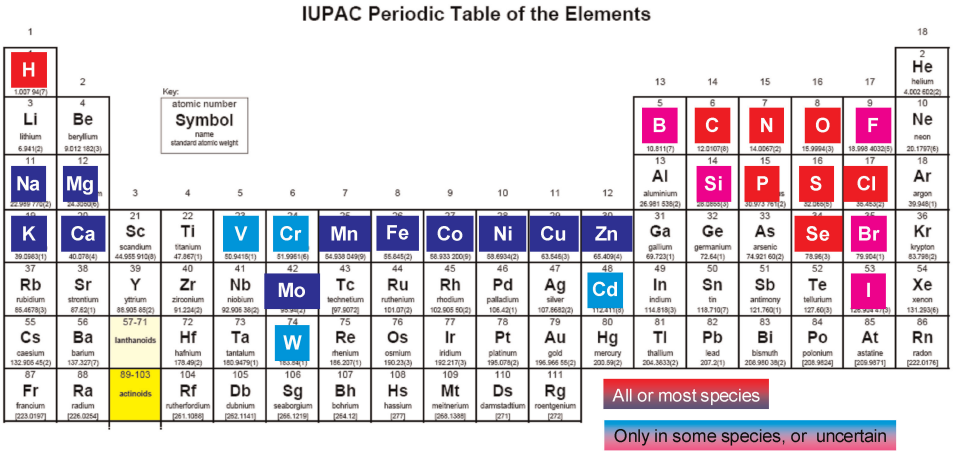


Figure 1. The periodic table, showing metals and non-metals that are required for life

Generally, according to the rules of the EC classification, enzymes that differ in their metal content or requirement have the same EC number if they catalyse the same reaction. However for some enzymes, the "Comments" field of the EC list specifies the presence of a metal; in other cases it may note a requirement for a metal ion for activity. Examples are given in Table 1.

Table 1. Examples of ways in which a metal requirement is specified in the EC List of enzymes.

"A copper protein"	EC 1.14.17.1 dopamine β-monooxygenase
"A zinc metallopeptidase"	EC 3.4.15.4 peptidyl-dipeptidase B
"Requires magnesium"	EC 6.3.1.8 glutathionylspermidine synthase
"Requires Ca^{2+}"	EC 2.7.11.18 myosin-light-chain kinase

Much more detailed information about metal ion requirements may be found in databases, notably BRENDA (http://www.brenda-enzymes.info/). Many enzymes contain metal ions in their structures, which can be detected by analytical methods such as particle-induced x-ray

emission (PIXE). The metal ions may be essential for activity or maintenance of protein structure. Occasionally they may be adventitiously bound to the protein, for example if a His-tag is used to aid purification, and the metal ion is bound to it [3].

The affinity with which metal ions bind to an enzyme varies greatly. Metal ions such as zinc and iron are usually tightly bound, and remain in position during isolation of the enzyme. However in some cases the metal ion can dissociate with loss of activity. In this case, addition of metal salts may reconstitute the protein and restore activity, though, as discussed later, this may not necessarily lead to a duplication of the normal environment of the metal in the cell.

The distinction between essential metal ions that are firmly bound to the enzyme, and those that can dissociate from the enzyme after the reaction cycle, is somewhat analogous to that between *prosthetic groups* such as pyridoxal phosphate which remain with the enzyme, and *cosubstrates* such as coenzyme A, which dissociate.

The requirement for loosely-bound metal ions in the activity of enzymes has often been inferred from observing the effects of added metal salts on the rate of reaction. A range of different salts may be incubated with the enzyme before assay, or included in the assay medium. Sometimes more than one element will confer activity, while others are inhibitory. Alternatively, the presence of metal ions may be inferred from the effects of treating the enzyme with chelating agents. Chelators display some degree of specificity for binding of a particular type of metal ion. A well-known example of the use of specific chelators to establish a role for a specific metal ion is the siderophore desferrioxamine, a specific and strongly binding chelator (log $K_d \approx 31$) for Fe^{3+}. Desferrioxamine cannot penetrate into an iron-containing enzyme such as ribonucleotide reductase, and can complex Fe(III) only when it dissociates from the protein [4]. More rapid complexation may be achieved with small ligands such as cyanide or nitric oxide, or the hydroxypyridones, [5] which can penetrate to some extent into the protein. Loss of iron from the enzyme may be facilitated by changing the protein conformation, e. g. by reduction of Fe(III) to Fe(II), or by lowering the pH. The rate of chelation of the metal ion depends on the spontaneous dissociation of the metal ion from its binding site, rather than on the thermodynamic stability of the final complex.

Properties of Metal Ions

Oxidation states

The oxidation state of a metal in a complex may be defined as the effective charge left on the metal ion when the ligands attached to the metal centre have been removed in their normal charged form. Thus, in aqueous solution, for a cation such as $Fe(H_2O)_6^{2+}$, the oxidation state is Fe(II), and the cation may be described by its charge number, i. e. iron(2+) or Fe^{2+} [6]. It

should be noted that while oxidation states of metal ions in complexes may usually be calculated straightforwardly, some "non-innocent" ligands may not allow this to be done. An example is nitric oxide, as this molecule may be bound to the metal centre as either NO^+ or NO^-. In complex structures such as iron-sulfur clusters, the oxidation state is more difficult to define.

"Hard" versus "soft" metal ions

A useful guide to the nature and extent of metal-ligand binding comes from the concept of hard and soft acids and bases [7]. Chemically-hard metal ions (acids) are small and not easily polarized, while soft metal ions are large and easily polarized. Ligands with highly electronegative donor atoms (O or N centres) are hard bases, while polarisable ligands such as those with sulfur donors are soft bases. In general stable complexes are formed between hard acids and hard bases and between soft acids and soft bases (Fig. 2).

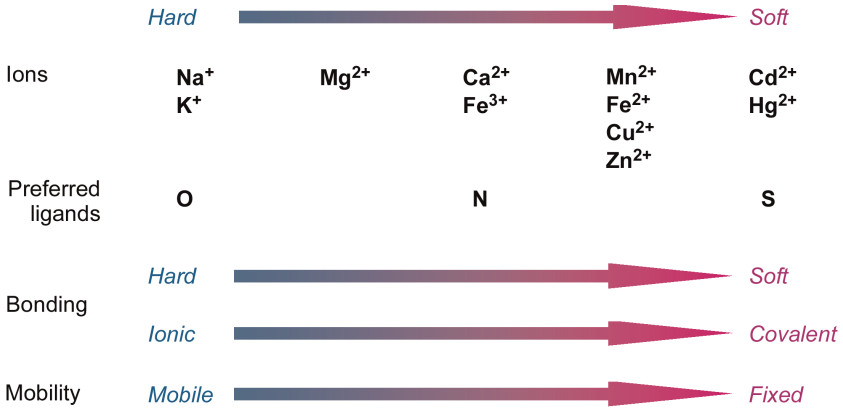


Figure 2. General properties of "hard" and "soft" metals and ligands.

There is an important gradation of properties between the sets of metals comprising (a) the Group 1 metal ions Na^+ and K^+, where the metal ions bind reversibly and weakly to the protein, (b) Group 2 metal ions Mg^{2+} and Ca^{2+} (and the transition metal Mn^{2+}), which bind more strongly than do Na^+ and K^+, and (c) the transition metal ions and Zn^{2+}, where the metal ion is bound firmly to the protein. In general, the "hard" metal ions such as K^+ and Mg^{2+} are considered to be freely mobile in the aqueous phase. Salts of the hard metal ions are added to the assay medium. By contrast soft metals form stronger, covalent interactions with their ligands. These complexes are more stable kinetically, and so the metal ion is not readily released, and so they do not usually need to be added to the assay medium.

The distinction implicit in the activity of these groups of metals is essentially a consequence of the polarizing powers of their metal cation, that is, the charge/radius ratio. A metal cation of high charge and small size has a high density of positive charge. It thus has high

polarizing power and so the interaction between the metal ion and a ligand is likely to be strong. A metal ion of large size and low charge polarizes ligands only weakly and binds to a lesser extent to ligands. Thus, the doubly charged cations Mg^{2+} and Ca^{2+} interact more strongly with ligands than do the Group 1 cations Na^{+} and K^{+}, but weakly compared to the transition metal ions.

Ionic radii decrease from left to right across the Periodic Table, so that the complexing power of the divalent transition metal cations increases from left to right, as expressed by the well known Irving-Williams series for the stability of high-spin octahedral metal complexes, $Mn^{2+} < Fe^{2+} < Co^{2+} < Ni^{2+} < Cu^{2+} > Zn^{2+}$ [8]. In accord with this, formation constant data show that Cu^{2+} binds most strongly of all the divalent transition metal ions. Thus, it complexes glycine with approximately a thousand-fold greater affinity than does the $3d^{10}$ ion Zn^{2+}. The strength of binding of Cu(II) to ligands is rather unexpected if is just considered in terms of the electronic configuration of Cu(II) and its ligand field stabilization energy, but other factors also contribute to values of formation constants. In the case of Cu(II) the structural distortion resulting from the Jahn-Teller effect leads to Cu(II) having two long axial bonds and four short bonds in the plane. Thus the overall binding of ligands to Cu(II) is controlled by the strong binding of the first four ligands in the plane. However, Cu^{2+} cannot compete with Fe^{3+} for binding of glycine, in accord with their relative polarizing power as discussed above.

Common toxic metal ions such as those of Cd^{2+}, Pb^{2+} and Hg^{2+} are soft, and form very strong bonds with sulfur ligands in particular, and so they can displace essential metal ions from their binding sites on proteins. The bound form is usually inactive, but in some cases it may have some activity. In one remarkable case, cadmium seems to have been adopted as the catalytic metal in place of zinc, in the carbonic anhydrase of diatoms; this is presumably the result of the availability of the metal in the environment [9].

Ligand types

When discussing coordination chemistry in biology, it should be noted that there are two different conventions for the use of the term "ligand". In biochemistry, any group that is bound to a protein, including a metal ion, might be termed a ligand. In inorganic chemistry, and for the purposes of this chapter, a ligand is a group that coordinates to a metal ion. A range of ligand groups are available for binding metal ions in biology, from proteins, nucleic acids and carbohydrates, together with a variety of smaller, specialized ligands. These include carboxylate, thiolato, amino, imidazolato and phosphate groups.

Substantial collections of formation constants for various metal ions and ligands are available that allow the likely speciation of a metal ion in a solution containing various ligands to be assessed.

Chelate effect

The chelate effect refers to the greater stability of complexes between a transition metal ion and bidentate or multidentate ligands compared to complexes of that metal ion with monodentate ligands of similar chemical character. Thus, the bidentate ligand ethylenediamine ($NH_2CH_2CH_2NH_2$) will displace ammonia from a metal complex. The sites for binding metal ions in proteins will usually be multidentate, and the binding will be correspondingly stronger.

Coordination geometry

In addition to variations in polarizing power, different metal ions have different preferences for ligand type and coordination geometry, which provide proteins with further selectivity for binding particular metal ions. This is particularly important for proteins involved in transport and storage of metal ions, which must show a high level of selectivity for the appropriate metal ion. In enzymes, the binding geometry provided for the metal by a protein is often distorted from the ideal geometry normally associated with these metal ions, resulting in a fine tuning of the reactivity of the metal. This is the so-called ''entatic state'' of Vallee and Williams [10]. For example, binding sites for copper centres, which undergo oxidation-reduction during catalytic reaction, are intermediate in geometry between square planar (optimum for Cu(II)) and tetrahedral (optimum for Cu(I)). This is enforced by the protein and so limits the extent of reorganization that has to take place in the localized geometry of the metal ion during the redox reaction. In the ''blue'' copper centres, sulfur occurs as a soft ligand, which will tend to stabilize the lower oxidation state.

Roles for Metals in Biology

The biochemical functions of metal ions reflect their chemistry (see Figure 2). In general, some metal ions are an essential component of the enzyme-catalysed reaction, while others are found in special sites in the protein, for structural or other reasons. Generally, the ionic strength affects protein-protein interactions, but in some cases a particular metal ion is involved. In the crystal structures of enzymes, special sites are sometimes found to be occupied by Na^+ or K^+ or Mg^{2+} ions.

Sodium and potassium ions are usually present as free ions; K^+ has the higher concentration within cells, while Na^+ is present in higher concentrations as an extracellular cation. These ions make major contribution to the osmotic balance. In enzyme-catalysed reactions their functions include acting as counter-ions to negatively charged amino-acyl residues or phosphates. The compensation of electrostatic charge is necessary for the energetics of enzyme-catalysed reactions.

Magnesium ions are often required for complexes with substrates, especially nucleotides and nucleic acids. For reactions with nucleotides the ratio $[Mg^{2+}]/[NTP]$ is important. Mg is required in stoichiometric amounts with ATP, but in many enzymes, (e. g. pyruvate kinase [11]) excess Mg^{2+} is inhibitory, owing to the formation of Mg_2-ATP complex.

Calcium is well known to have a special role in regulation of metabolism e. g. protein kinases, protein phosphatases and certain peptidases. The intracellular concentration of ions fluctuates considerably in response to metabolic changes, e. g. in muscle contraction and nerve transmission. Ca^{2+} is fairly hard, so can bind and dissociate readily as the intracellular concentration of Ca^{2+} changes. Often it binds to calmodulin domains, which undergo structural changes on binding.

Zinc is the most common 3 d block metal ion in a typical animal or plant cell. Zn^{2+} is chemically softer than Mg^{2+}or Ca^{2+} and forms more stable complexes. Zinc is a component of hundreds of enzymes [12, 13]. Because of its strong Lewis acidity, Zn^{2+} may polarize a substrate or lower the pK_a of coordinated water, leading to the Zn-hydroxide pathway in hydrolases for example. It occurs in the catalytic centres, and in other parts of the protein where it stabilizes the protein structure. In yeast alcohol dehydrogenase for example there are two zinc ions, one catalytic and one structural; removal of the structural zinc does not affect catalysis, but renders the protein less stable [14].

Iron exists in heme and nonheme proteins, including many oxidoreductases. It has special roles such as binding gases such as O_2, NO or CO. Iron is assembled into complex metalloclusters, sometimes with other metals such as nickel, molybdenum or vanadium, in enzymes for the metabolism of H_2 or N_2 [15]. Iron metabolism is complex because of the different properties of its oxidation states. Fe(II) and Fe(III) are the most common; Fe(IV) may be formed in oxidases and peroxidases by reaction of Fe(III) with O_2 or H_2O_2.

Transition metal ions such as Fe(II) are known to be cytotoxic under aerobic conditions, forming the reactive oxygen species superoxide and hydroxyl radical [16]. An example is the Fenton reaction:

$$Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + OH^- + \bullet OH$$

In the cell, the concentration of free iron ions is maintained at a very low level, except under pathogenic conditions such as chronic iron overload. Some nonheme iron oxygenases use the oxygen chemistry of iron to catalyse the hydroxylation of their substrates. Since the substrate and reactive oxygen species must bind simultaneously to the iron, the iron sites in these enzymes have a rather open configuration, often with only three ligands from the protein, for example by the triad His_2Asp [17]. The harder Fe(III) may tend to dissociate from the enzyme during purification steps, such as column chromatography, that involve dilution [18]. Some nonheme iron-containing oxygenases are routinely assayed with iron

salts added to the reaction medium. Although the oxidation state of the iron in these enzymes is often Fe(III), Fe^{3+} salts will hydrolyze and precipitate at neutral pH, forming the mineral ferrihydrite [19]. In order to keep Fe(III) in solution it is necessary to have a tight, soluble complex such as Fe(III) citrate, where the iron may be unavailable for interaction with enzymes. Therefore, it is preferable to add Fe^{2+} salts, as a freshly prepared solution, under anaerobic conditions.

Copper: Cu^{2+} is a good Lewis-acid catalyst, but is seldom used as such by enzymes, presumably because of the possibility of forming reactive oxygen species in a similar way to iron. It tends to bind strongly to the peptide chains of proteins, as in the biuret reaction. Copper is found in a considerable number of oxidoreductases. In the rare cases where it needs to be added to an enzyme assay, it is usually provided as the cupric Cu^{2+} salt. In the cell, it is transported by metallochaperones in the cuprous state, Cu^{+}.

Metal Ion Homeostasis in the Cell

In a mixture of compounds including metal ions in aqueous solution, the distribution of ions is driven by simple thermodynamic equilibria. The outcome of this is the ''speciation'' of metal ions. The ''free'' metal ion, M^{n+}(aq) is in equilibrium with each protein or a small-molecule ligand L, with binding constant, K_d. It was assumed for a long time that a similar equilibrium existed in the cell. This implied that the cell buffers the concentrations of the free metal ions, by analogy with a pH buffering system, so that the free metal concentration would remain relatively constant as total metal concentration changed. A development from this was the concept of the ''labile metal pool'', or ''biologically available metal pool'' instead of free metal pool. This implies that metals complex with abundant small molecules in the cytoplasm, but are still available to sites of stronger complexing power.

An example is the distribution of zinc in a bacterium such as *E. coli*. In the growth medium, the zinc concentration is approximately 100 nM. It is actively transported into the cell. The total amount of zinc corresponds to a concentration of approximately 0.2 mM, but this is bound to proteins and not freely available. The level of available zinc is controlled by the expression of the genes for two transporters: Zur, which is responsible for uptake, and ZntR, which is responsible for efflux. The expression is in turn controlled by metalloregulatory proteins binding to DNA. In consequence, the concentration of ''available'' zinc is around 1 femtomolar [20]. An even more extreme example is copper, for which the concentration of free ions in the cell has been estimated to be in the zeptomolar range, around 2.10^{-21} M [21].

Table 2. Enzyme assays that require activation by a metal ion

Enzyme	Metal ion	Conditions	Ref
EC 1.14.12.11 Toluene dioxygenase	Fe^{2+}	0.36 mM Fe^{2+} in assay medium	[33]
EC 2.7.1.40 Pyruvate kinase	Mg^{2+}	10 mM Mg^{2+}, 2 mM ADP	[34]
ADAM8 peptidase	Zn^{2+}	Pre-incubate with 0.1 mM Zn^{2+}	[35]
EC 1.14.17.3 Peptidylglycine monooxygenase	Cu^{2+}	Pre-treat with 25 μM Cu^{2+}	[36]

However, when metal ions such as iron, copper or zinc are added to enzyme assays, much higher concentrations, typically micromolar to millimolar, are found to be optimal (Table 2). This immediately suggests that the binding equilibria in the enzyme assay do not reflect the conditions in the cell, where there are specific systems for import, trafficking and storage of metal ions. These have become apparent from the discovery of genes, in addition to the structural genes for the enzyme proteins, which are required for expression of the metalloenzymes in their active forms. The insertion of metal ions such as iron, zinc or copper is a post-translational modification, which requires the presence of ion transporters, storage proteins, and ''metallochaperones'' [22]. The term ''metallochaperones'' applies to proteins with a variety of functions [23, 24]. A principal function is to act as a type of ''escort protein'', which selectively binds a particular metal, and sequesters it inside the cell. Metallochaperones protect metal ions from adventitious binding to metabolites and macromolecules and, in doing so, they protect the cell from potential damage, for example from radicals produced from reaction of transition metal ions with O_2 and H_2O_2. By means of specific protein-protein interactions, the metallochaperones accept the ions only from an appropriate donor such as a membrane transporter, and transfer them to a specific acceptor such as an apoenzyme. Thus, they ensure that the metal ions are delivered efficiently to their targets, while the intracellular concentration of free metal ions is kept extremely low (Fig. 3). Like the chaperonins that assist protein folding, the metallochaperone may partly unfold and refold the donor and/or acceptor proteins to assist the release and binding of metal ions. They can ensure that the metal ion is only transferred to the right protein when it is in the right state [25, 26]

For copper, a number of proteins involved in trafficking have been identified and characterized. There are different systems for delivery of copper to Cu/Zn superoxide dismutase, cytochrome oxidase, ceruloplasmin and other enzymes.

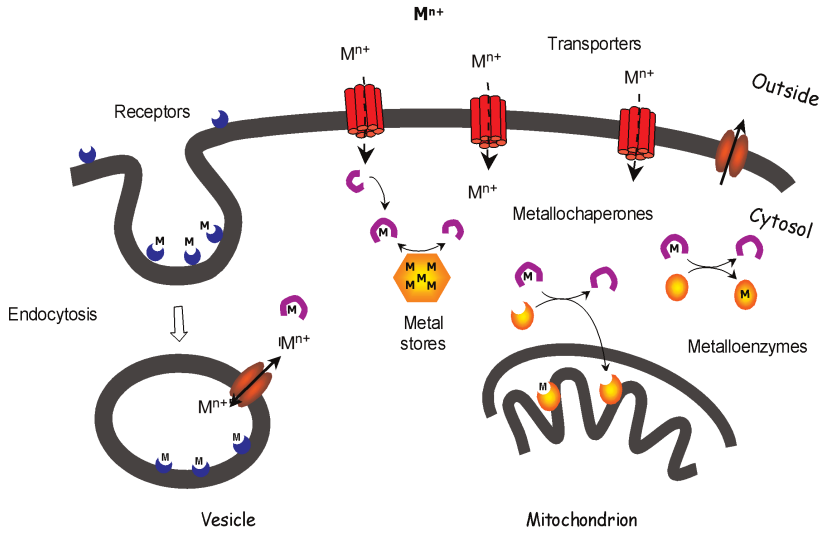


Figure 3. General model of intracellular transport, trafficking and delivery of metal ions such as Fe, Cu and Zn. Free metal ions are bound by, and transferred between various transporters, metallochaperones and storage proteins. They may also be stored in intracellular organelles such as the Golgi apparatus, vacuoles or plastids. The concentration of the free unbound metal ion is generally extremely low, though it may still be accessible to chelators.

Zinc is required for hundreds of enzymes and over a thousand transcription factors and other regulatory proteins [12, 13]. Metallochaperones, transporters and other zinc-dependent factors have been identified, but their precise role in the insertion of Zn^{2+} is under investigation.

For iron, the systems are less well defined. It undergoes oxidation-reduction reactions as it is transported through the cell. Usually it is presented at the cell surface as Fe(III); it is reduced to Fe(II), and it is transported in this form through membranes, but for transport in the blood as transferrin, or storage in the cell in ferritin, it is oxidized again to the less toxic Fe(III) [27]. For iron, little is known about the precise role of metallochaperones, apart from a recent report of a selective iron metallochaperone for transfer to ferritin [28].

Relevant Parameters for Deriving Activity Data of Metal-Dependent Enzymes

A complete description of the activity of metalloenzymes in the cell would include the systems for insertion and maintenance of metal ions in enzymes. For assays of most enzymes containing metal ions such as zinc, iron or copper, it is not necessary to consider these processes because the metals are present in the isolated enzymes, and do not readily

dissociate. However, as already noted, the metal ion in some cases can dissociate, and has to be re-supplied in order for the enzyme to show activity. Some examples of enzyme assays where these metal ions are included are listed in Table 2.

In contrast to the situation in the cell, where the specific metal ion is often selected by protein-protein interactions between the apoprotein and a metallochaperone, *in vitro* it may be possible to insert different metal ions directly into the apoenzyme. If the non-physiological metal ion is inserted, the enzyme may be inactive, but in some cases several different metallic elements may restore some activity. In some cases they are inhibitory, but in others they may be active, and may change the specificity of the enzyme. Xylose isomerase (D-xylose ketol-isomerase; EC. 5.3.1.5), otherwise known as glucose isomerase, catalyzes the reversible isomerization of D-glucose and D-xylose to D-fructose and D-xylulose, respectively. The enzyme is widely used in the food industry because of its application in the production of high-fructose corn syrup. The enzyme has an absolute requirement for bivalent cations with ionic radii < 0.08 nm; Mg^{2+}, Co^{2+}, or Mn^{2+} are effective. There are two metal-binding sites, which have to be occupied for catalytic activity. Different metal ions in these sites change the specificity of the enzyme for sugars, as well as the stability of the protein [29, 30].

Additional parameters to be considered in assays of metalloenzymes may include the following factors, which relate largely to complications that arise from chemical interactions of the metal centre with components of the assay medium.

1. *Ligands in the medium that will bind to the metal ion that is necessary for activation of the enzyme under study.* Tables of formation constants will be helpful at this point. If ligands bind to the metal ion, then the metal may be present in the assay medium as several different complexes, possibly with different stoichiometry, overall charge, shape and lipophilicity. These different species may not activate the enzyme to the same extent.
2. *pH buffers that will bind the metal ion*. This is an important matter as the concentration of buffer may be relatively high compared to other components of the medium. Citrate buffers should be avoided if possible. The ‘‘Good buffers’’ are designed to avoid metal coordination, and should be preferred; at least the assay should be tried with one of these buffers [31].
3. *Precipitation by other constituents of the medium*. If phosphate, for example, is present in the medium, then metal ions may be lost from the solution by precipitation as metal phosphates. Tables of solubilities of inorganic compounds show that many common metal phosphates are very poorly soluble in water. For example, the solubilities of phosphates of Zn(II) and Fe(II) are very low, while the phosphate of Fe(III) is only slightly soluble.

4. *Precipitation as a function of pH*. Added transition metal and zinc ions in the assay medium may be precipitated as oxides and hydroxides at pH values over 6. The pK_a values of the aqua cations of Ni(II), Zn(II) and Cu(II) are 9.9, 9.0 and 8.0 respectively; precipitation normally begins to occur about 2 – 3 pH units below the pK_a value of the aquo complex and, as the pH is raised, occurs in the order of the Lewis acidity of the metal ion, taking place first for Cu(II), the smallest ion. Precipitation is therefore much more pronounced for aquo cations $M(H_2O)_6^{3+}$ such as $[Fe(H_2O)_6]^{3+}$, as already noted.
5. *Chelating ligands such as citrate that are added to assay solutions, to ensure that precipitation does not occur*. It is important to check if this added ligand also binds the essential metal ion to an extent that lowers its bioavailability, leading to a decrease in rate constant for the enzyme reaction.
6. *Strong chelators, e.g. EDTA, that are added to scavenge heavy metals in the medium*. Examples were given by Boyce *et al.* [32]. These can also remove metal ions from proteins, as mentioned above.
7. In order to maintain low "free ion" concentrations, chelating agents with lower binding affinities can be used to buffer the metal ion concentration. For example, *N,N,N',N'*-tetrakis(2-pyridylmethyl)ethylenediamine (TPEN) was used by O'Halloran *et al.* to maintain femtomolar concentration of zinc [20].
8. *Reconstitution with a metal ion.* If this is necessary because a metal normally present in the enzyme has dissociated, the conditions used to restore the metal ion must be stated.
9. If the concentration of the metal ion such as calcium fluctuates with time, and with position in the cell, the activity of the enzyme should be explored over the appropriate range of concentrations.

Conclusions

In this article we have attempted to bring together and discuss the factors that underlie the accurate determination of enzyme kinetic parameters in cases where metal ions are involved. Recently there have been advances in the understanding of the assembly of complex metal ion clusters such as iron-sulfur clusters [33]. Each of these processes may require a highly-organized sequence of scaffolds and other accessory proteins [33]. These represent metabolic pathways in their own right. For future studies of metabolic reconstruction, it might be necessary to consider their involvement in enzyme activity.

References

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

[2] Kettner, C. (2007) Good publication practice as a prerequisite for comparable enzyme data? *In Silico Biology.* **7**, S 57-S 64.

[3] Cammack, R. (2007) EPR spectroscopy in genome-wide expression studies. In *Spectral Techniques in Proteomics* (Sem, D. S., ed.). pp. 391 – 405, Taylor & Francis, London.

[4] Cooper, C.E., Lynagh, G.R., Hoyes, K.P., Hider, R.C., Cammack, R. and Porter, J.B. (1996) The relationship of intracellular iron chelation to the inhibition and regeneration of human ribonucleotide reductase as studied by EPR spectroscopy. *J. Biol. Chem.* **271**:20291 – 20299.

[5] Kayyali, R., Porter, J.B., Liu, Z.D., Davies, N.A., Nugent, J.H., Cooper, C.E. and Hider, R.C. (2001) Structure-function investigation of the interaction of 1-and 2-substituted 3-hydroxypyridin-4-ones with 5-lipoxygenase and ribonucleotide reductase. *J. Biol. Chem.* **276**:48814 – 48822.

[6] Connelly, N.G., Damhus, T., Hartshorn, R.M. and Hutton, A.T. (2005) Nomenclature of Inorganic Chemistry, IUPAC Recommendations 2005. RSC Publishing, London.

[7] Pearson, R.G. (1963) Hard and Soft Acids and Bases. *J. Am. Chem. Soc.* **85**:3533 – 3539.

[8] Housecroft, C. and Sharpe, A.G. (2007) *Inorganic Chemistry.* Prentice Hall, New Jersey.

[9] Xu, Y., Feng, L., Jeffrey, P.D., Shi, Y. and Morel, F.M.M. (2008) Structure and metal exchange in the cadmium carbonic anhydrase of marine diatoms. *Nature* **452**:56 – 61.

[10] Vallee, B.L. and Williams, R.J.P. (1968) Metalloenzymes – Entatic Nature of Their Active Sites. *Proc. Natl. Acad. Sci. U. S. A.* **59**:498 – 505.

[11] Holmsen, H. and Storm, E. (1969) Adenosine Triphosphate Inhibition of Pyruvate Kinase Reaction and Its Dependence on Total Magnesium Ion Concentration. *Biochem. J.* **112**:303 – 316.

[12] Vallee, B.L. and Auld, D.S. (1990) Zinc Coordination, Function, and Structure of Zinc Enzymes and Other Proteins. *Biochemistry* **29**:5647 – 5659.

[13] Maret, W. (2005) Zinc coordination environments in proteins determine zinc functions. *J. Trace Elem. Med. Biol.* **19**:7 – 12.

[14] Magonet, E., Hayen, P., Delforge, D., Delaive, E. and Remacle, J. (1992) Importance of the Structural Zinc Atom for the Stability of Yeast Alcohol-Dehydrogenase. *Biochem. J.* **287**:361 – 365.

[15] Rees, D.C. (2002) Great metalloclusters in enzymology. *Annu. Rev. Biochem.* **71**:221 – 246.

[16] Halliwell, B. and Gutteridge, J.M.C. (1992) Biologically Relevant Metal Ion-Dependent Hydroxyl Radical Generation – an Update. *FEBS Lett.* **307**:108 – 112.

[17] Koehntop, K.D., Emerson, J.P. and Que, L. (2005) The 2-His-1-carboxylate facial triad: a versatile platform for dioxygen activation by mononuclear non-heme iron(II) enzymes. *J. Biol. Inorg. Chem.* **10**:87 – 93.

[18] Axcell, B.C. and Geary, P.J. (1975) Purification and some properties of a soluble benzene-oxidizing system from a strain of Pseudomonas. *Biochem. J.* **146**:173 – 183.

[19] Michel, F.M., Ehm, L., Antao, S.M., Lee, P.L., Chupas, P.J., Liu, G., Strongin, D.R., Schoonen, M.A.A., Phillips, B.L. and Parise, J.B. (2007) The structure of ferrihydrite, a nanocrystalline material. *Science* **316**:1726 – 1729.

[20] Outten, C.E. and O'Halloran, T.V. (2001) Femtomolar sensitivity of metalloregulatory proteins controlling zinc homeostasis. *Science* **292**:2488 – 2492.

[21] Changela, A., Chen, K., Xue, Y., Holschen, J., Outten, C.E., O'Halloran, T.V. and Mondragon, A. (2003) Molecular basis of metal-ion selectivity and zeptomolar sensitivity by CueR. *Science* **301**:1383 – 1387.

[22] Finney, L.A. and O'Halloran, T.V. (2003) Transition metal speciation in the cell: Insights from the chemistry of metal ion receptors. *Science* **300**:931 – 936.

[23] Hausinger, R.P. (1997) Metallocenter assembly in nickel-containing enzymes. *J. Biol. Inorg. Chem.* **2**:279 – 286.

[24] O'Halloran, T.V. and Culotta, V.C. (2000) Metallochaperones, an intracellular shuttle service for metal ions. *J. Biol. Chem.* **275**:25057 – 25060.

[25] Luk, E., Jensen, L.T. and Culotta, V.C. (2003) The many highways for intracellular trafficking of metals. *J. Biol. Inorg. Chem.* **8**:803 – 809.

[26] Puig, S. and Thiele, D.J. (2002) Molecular mechanisms of copper uptake and distribution. *Curr. Opin. Chem. Biol.* **6**:171 – 180.

[27] Crichton, R. (2008) Iron Metabolism: From Molecular Mechanisms to Clinical Consequences. John Wiley & Sons, Chichester.

[28] Shi, H., Bencze, K.Z., Stemmler, T.L. and Philpott, C.C. (2008) A cytosolic iron chaperone that delivers iron to ferritin. *Science* **320**:1207 – 1210.

[29] Van Bastelaere, P., Vangrysperre, W. and Kersters-Hilderson, H. (1991) Kinetic Studies of Mg^{2+}-Activated, Co^{2+}-Activated and Mn^{2+}-Activated D-Xylose Isomerases. *Biochem. J.* **278**:285 – 292.

[30] Epting, K.L., Vieille, C., Zeikus, J.G., Kelly, R.M., Kelly, R.M., Zeikus, J.G. and Vieille, C. (2005) Influence of divalent cations on the structural thermostability and thermal inactivation kinetics of class II xylose isomerases. *FEBS J.* **272**:1454 – 1464.

[31] Good, N.E., Winget, G.D., Winter, W., Conolly, T.N., Izawa, S. and Singh, R.M.M. (1966) Hydrogen Ion Buffers for Biological Research. *Biochemistry.* **5**:467–477.

[32] Boyce, S., Tipton, K.F. and McDonald, A. (2004) Extending enzyme classification with metabolic and kinetic data: some difficulties to be resolved. In *Experimental Standard Conditions of Enzyme Characterizations* (Hicks, M.G. and Kettner, C., eds.). pp. 17 – 44, Beilstein Institut, Frankfurt

[33] Lill, R., Dutkiewicz, R., Elsasser, H.P., Hausmann, A., Netz, D.J.A., Pierik, A.J., Stehling, O., Urzica, E. and Muhlenhoff, U. (2006) Mechanisms of iron-sulfur protein maturation in mitochondria, cytosol and nucleus of eukaryotes. *Biochim. Biophys. Acta –Mol. Cell Res.* **1763**:652 – 667.

[34] Wackett, L.P. (1990) Toluene Dioxygenase from *Pseudomonas putida* F1. *Methods Enzymol.* **188**:39 – 45.

[35] Cardenas, J.M. (1982) Pyruvate-Kinase from Bovine Muscle and Liver. *Methods Enzymol.* **90**:140 – 149.

[36] Naus, S., Reipschlager, S., Wildeboer, D., Lichtenthaler, S.F., Mitterreiter, S., Guan, Z. Q., Moss, M.L. and Bartsch, J.W. (2006) Identification of candidate substrates for ectodomain shedding by the metalloprotease-disintegrin ADAM8. *Biol. Chem.* **387:**337 – 346.

[37] Jaron, S. and Blackburn, N.J. (2001) Characterization of a half-apo derivative of peptidylglycine monooxygenase. Insight into the reactivity of each active site copper. *Biochemistry* **40**:6867 – 6875.

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

From The Enzyme List to Pathways and Back Again

Andrew G. McDonald, Keith Tipton* and Sinéad Boyce

Department of Biochemistry, Trinity College, Dublin 2, Ireland

E-Mail: *ktipton@tcd.ie

Received: 25th September 2007 / Published: 20th August 2008

Abstract

The IUBMB Enzyme List is widely used by other databases as a source for avoiding ambiguity in the recognition of enzymes as catalytic entities. However, it was never designed for activities such as pathway tracing, which have become increasingly important in systems biology. This is because it often relies on generic or representative reactions to show the reactions catalysed by enzymes of wide specificity. It is necessary to go to databases such as *BRENDA* to find further, more detailed, information on what is known about the range of substrates for any particular enzyme. In order to provide a framework for tracing pathways involving any specific enzyme or metabolite, we have created a *Reactions Database* from the material in the Enzyme list. This allows reactions to be searched by substrate/product and pathways to be traced from any selected starting/seed substrate. An extensive synonym glossary allows searches by any of the alternative names, including accepted abbreviations, by which a chemical compound may be known. This database was necessary for the development of the application *Reaction Explorer* (http://www.reaction-explorer.org/), which was written in REALbasic to search the *Reactions Database* and draw metabolic pathways from reactions selected by the user. Having input the name of the starting compound (the ''seed''), the user is presented with a list of all reactions containing that compound and then selects the product of interest as the next point on the ensuing graph. The pathway diagram is then generated as the process iterates. A contextual menu is provided,

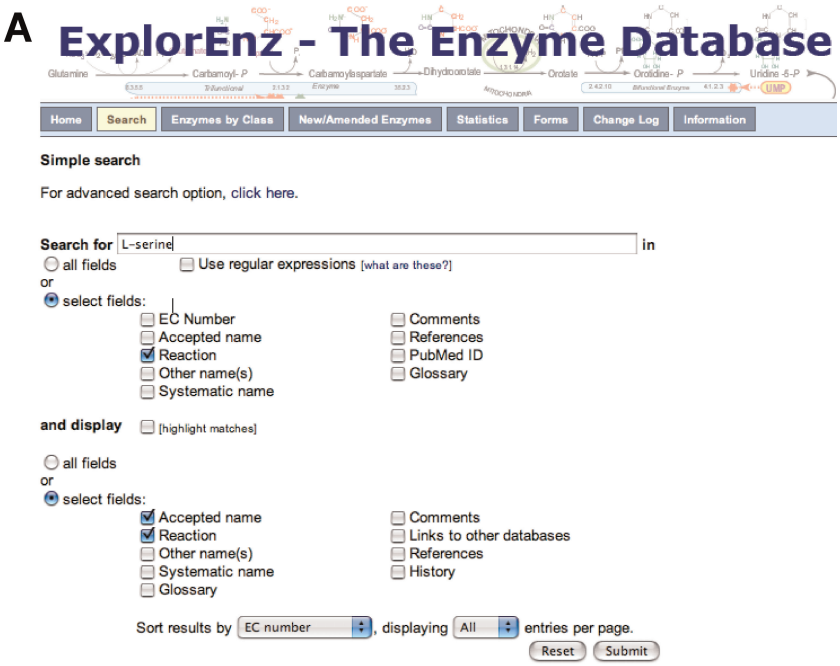
which allows the user to (i) remove a compound from the graph, along with all associated links; (ii) search the reactions database again for additional reactions involving the compound and (iii) search for the compound within the Enzyme List.

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. Because of this strictly defined function, its application to tracing metabolic systems is, perforce, limited. Although much of the data required for this application are there, the structure makes it difficult to access. Since it would be undesirable to alter the Enzyme List to meet other functions if that were to diminish its core utility, this account will discuss what can be achieved using the list itself and derivatives of it.

What the Enzyme List Can Do, the Enzyme Centric Approach

A



(B)

EC: 1.4.1.7: serine dehydrogenase
Reaction: L-serine + H_2O + NAD^+ = 3-hydroxypyruvate + NH_3 + NADH + H^+
EC: 2.3.1.30: serine O-acetyltransferase
Reaction: acetyl-CoA + L-serine = CoA + *O*-acetyl-L-serine
EC: 2.3.1.50: serine C-palmitoyltransferase
Reaction: palmitoyl-CoA + L-serine = CoA + 3-dehydro-D-sphinganine + CO_2
EC: 2.6.1.45: serine-glyoxylate transaminase
Reaction: L-serine + glyoxylate = 3-hydroxypyruvate + glycine
EC: 2.6.1.51: serine-pyruvate transaminase
Reaction: L-serine + pyruvate = 3-hydroxypyruvate + L-alanine
EC: 2.7.1.80: diphosphate-serine phosphotransferase
Reaction: diphosphate + L-serine = phosphate + *O*-phospho-L-serine
EC: 2.7.8.4: serine-phosphoethanolamine synthasee
etc.

Figure 1. A specific search of *ExplorEnz* (2) for enzymes having L-serine as a substrate or product (A) with a part of the output (B).

The Enzyme List can be searched to find all of those enzymes that catalyse reactions involving a given substrate or product. This is illustrated in figure 1 using serine as an example. From the output (Fig. 1b), it is possible to construct the simple pictorial representation shown in figure 2. Such ''enzyme-centric'' searching can be useful in predicting the possible effects of drugs that are targeted against a specific enzyme, since they will show other enzymes that might also be affected.

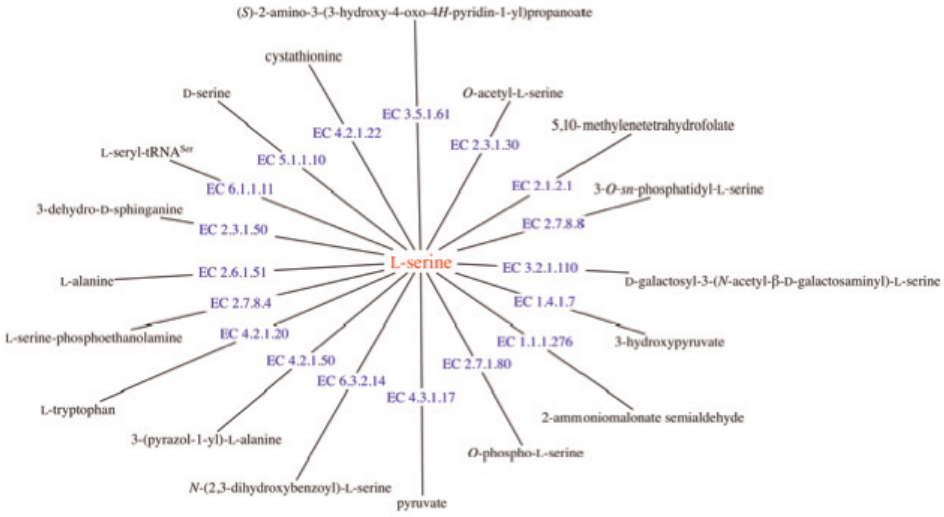


Figure 2. The substrate-centric approach: a redrawn view of the enzyme having L-serine as a substrate or product.

It is also possible to list groups of enzymes linked by common substrates and products, as shown in figure 3. The results of attempts to display these is graphical form can, however, appear quite complex because of the multiplicity of edges that occurs when forming connections between enzymes.

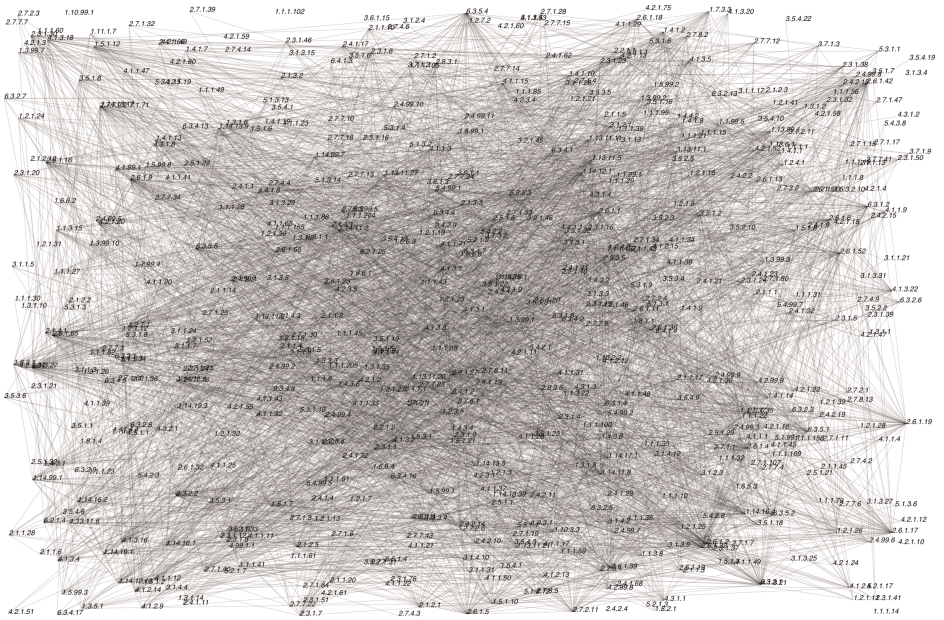


Figure 3. A network of 604 enzymes with links representing shared metabolites. There are 4,062 possible connections made amongst the enzymes in this subset, from which H^+, H_2O, and common cofactor pairs (e.g., ATP/ADP) were excluded.

From combinatorics, the general formula for the number of ways r items can be taken from n is:

$$^{n}\mathrm{C}_{r} = n!/[r!(n\text{-}r)!]$$

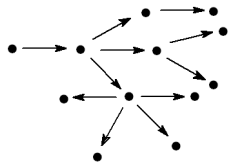
$^{n}C_2$ therefore represents the total number of possible enzyme (node) pairs, where each pair shares the same metabolite ($r = 2$). Thus 4 enzymes sharing a common metabolite will have to be connected by 6 edges and 380 enzyme nodes, as would be required to show the number of reactions in the database that involve O_2, will require 72,010 edges.. Clearly, the situation can become more complicated than this if one considers the possibility of having several shared metabolites for each enzyme. Whilst such a representation can readily be searched for any given enzyme and has the advantage that each enzyme only occurs once in the diagram, as opposed to the hand-crafted, artistic, versions, such as the Nicholson metabolic pathways charts (see [3]), where the separation of different metabolic systems in the display can result in the same enzyme occurring in several different places.

Reactions Database: the Substrate-centric Approach

The *Reactions Database* was initially prepared by extracting all of the reactions in *ExplorEnz*, the MySQL version of the Enzyme List [2]. These were used to create a separate database (also MySQL). A web application created with PHP was developed to provide the query interface to the database. As with *ExplorEnz*, the query engine supports both case-insensitive and regular-expression substring searches. This, at least in theory, should all reactions involving any given substrate or product to be displayed. It should also allow one to trace the number of reactions n-steps from any given starting substrate, so that its metabolic fates can be better appreciated. However, just as in the case of edge multiplicity in the ''enzyme-centric'' approach, the system rapidly gains complexity. This is illustrated for some simple metabolites in the table below and by the illustrative tree structure.

Table: Some query results for successive reactions of different compounds.

Compound	Paths found		
	1 Step	2 Steps	3 Steps
L-Ascorbate	5	89	
L-Cysteine	5	25	577
L-Tyrosine	13	676	
Ribitol	2	50	2776



One major cause of this rapidly expanding complexity is the involvement of a reactant that is used or produced by several different enzymes. For example, if a reaction produces or uses ATP it will be linked to many other reactions (the kinases etc). This can be addressed by specifically excluding some compounds, such as H_2O, H^+, ATP, ADP, AMP, phosphate, diphosphate, $NADP^+$, NADPH, NAD^+, NADH, $NAD(P)^+$, NAD(P)H, A, AH_2, acceptor and reduced acceptor from the search. However, if one were to exclude, for example, NAD^+, that would eliminate ADP-ribosylation reactions as well as oxidoreductases and excluding ATP would eliminate several adenylyltransferase reactions. This problem can be better addressed by selective elimination of reactant pairs rather than single reactants from the search. These might include:

```
ATP <=> ADP
ATP <=> AMP
AH2 <=> A
NAD+ <=> NADH
NADP+ <=> NADPH
NAD(P)+<=> NAD(P)H
CoA <=> acetyl-CoA
```

Representations: *Reaction Explorer*

Although the *Reactions database* can provide lists of reactions, an additional tool is needed for display purposes. This is provided by *Reaction Explorer* [4], which is a multi-platform application, written in REALbasic, for constructing metabolic network graphs. Versions are currently available for the following operating systems: Mac OS X 10.1 or higher, Mac OS 9.x, Linux x86 and Windows 95 or higher. Selecting any product from a reaction will automatically draw a line connecting it to its parent substrate from where one can proceed to the next step in the pathway and so on, to provide a pictorial representation of the process.

The output is designed to be basic because its purpose is to convey information and not to construct works of art. Thus, it is not designed as a competitor for representational systems, such as GraphViz [5], or the craftsman-designed Nicholson metabolic pathways charts, but rather to display the essential information quickly and easily.

What Reaction DB & Explorer Cannot Do

As indicated above, *Reaction Explorer* is an aid to drawing pathways so that interactions may be visualized. In fact one can generate searchable connection graphs with any dataset that is entered in the *Reaction Explorer* file format, such as those shown in figures 2 and 3. As will be discussed later, it was also found to be of value as an aid to trouble-shooting the Enzyme List.

However, there are limitations imposed not by the programme but by the nature of the system involved. As discussed in connection with the table previously shown, a simple tree that describes all reactions proceeding for n-steps from any named reactant is not to be expected. Similarly, a unique pathway connecting two distant metabolites does not usually occur in metabolite space. Thus the question "find the pathway from glucose to lactate" might be expected to yield glycolysis. Indeed it does, together with very many other pathways. That is because there are very many ways in which glucose can be converted into lactate, including the synthesis and breakdown of compounds such as cholesterol. This might be addressed by specifying the number of steps allowed, but not all major pathways necessarily use the minimum number of steps.

Problems with Using Reactions of the Enzyme List

It is perhaps unreasonable to expect the Enzyme List to have functions that it was not designed for. However, there are several aspects that make it unsuitable for simple adaptation to reaction pathway tracing through systems such as *Reaction Database* & *Explorer*.

(a) Not all reactions catalysed by a given enzyme are listed

In the past, the Enzyme List has often used a representative reaction for enzymes with broad specificities. It is intended to add a field to include additional reactions where appropriate. Reactions involving non-physiological substrates are not listed except in the case of donors and acceptors where the physiological factor has not yet been identified. However, this can result in judgements about what is, and what is not, physiologically important.

For example aromatic-L-amino-acid decarboxylase (EC 4.1.1.28) is given as catalysing two reactions

(1) 3,4-dihydroxy-L-phenylalanine = dopamine + CO_2

(2) 5-hydroxy-L-tryptophan = 5-hydroxytryptamine + CO_2

but it will also catalyse the decarboxylation of L-tyrosine, L-tryptophan and L-phenylalanine. Although these reactions may be of lesser physiological significance, they are not unimportant and can, indeed, have major significance in the responses to therapy involving some antidepressant drugs.

In some cases additional information on the specificity is also given in the ''comments'' associated with the Enzyme-List entry. For example, the 6-phosphofructokinase (EC 2.7.1.11) reaction is given as

ATP + D-fructose 6-phosphate = ADP + D-fructose 1,6-bisphosphate

but the comments also state: ''D-Tagatose 6-phosphate and sedoheptulose 7-phosphate can act as acceptors. UTP, CTP and ITP can act as donors''. Clearly such material, although readily accessible in an *ExplorEnz* search, needs to be incorporated into the *Reactions Database*. The comprehensive lists of substrates provided by *BRENDA* [6], which also contains, somewhat arbitrary, listings of ''natural substrates'', can be most valuable for this purpose.

In the case of L-threonine 3-dehydrogenase (EC 1.1.1.103), only one reaction is listed

L-threonine + NAD^+ = L-2-amino-3-oxobutanoate + NADH + H^+

but a check of the *BRENDA* entry reveals that it also catalyses the reaction:

L-serine + NAD^+ = L-2-amino-3- oxopropionate + NADH + H^+

(b) General/Markush terms

In the case of enzymes with broad substrate specificities, such as alcohol dehydrogenase (EC 1.1.1.1), where the number of substrates, or potential substrates, is very large, the Enzyme List gives a single generic reaction, and *BRENDA* is an essential source of detail. For example, the alcohol dehydrogenase reaction is given as

$$\text{an alcohol} + NAD^+ = \text{an aldehyde or ketone} + NADH + H^+$$

and that of L-amino-acid oxidase (EC 1.4.3.2) is given as:

$$\text{an L-amino acid} + H_2O + O_2 = \text{a 2-oxo acid} + NH_3 + H_2O_2$$

The problem with this approach is that it does not indicate which alcohols, or L-amino acids, are not substrates. Again, *BRENDA* is invaluable in such cases, although it should always been borne in mind that the absence of a compound from the substrate/product list does not necessarily mean that it is not a substrate, but may simply mean that nobody has tried it.

Markush terms are also used for some reactions. For example the reaction catalysed by amine oxidase (copper-containing) (EC 1.4.3.6) is given as:

$$RCH_2NH_2 + H_2O + O_2 = RCHO + NH_3 + H_2O_2$$

and that of carbonyl reductase (NADPH) (EC 1.1.1.184) as:

$$R\text{-}CHOH\text{-}R' + NADP^+ = R\text{-}CO\text{-}R' + NADPH + H^+$$

Although such formulations are somewhat more informative than the general reactions above, and the Markush terms are searchable in *ExplorEnz* and *Reaction Explorer*, it is still necessary to revert to *BRENDA* for information on the exact substrates that are known to be used.

Another complexity occurs where it is not possible to describe the reaction catalysed by a simple reaction equation, without ambiguity. Examples of this are the reactions catalysed by (α-amylase) (EC 3.2.1.1), which is given as

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

1,4-α-glucan branching enzyme (EC 2.4.1.18), where the reaction is described as:

> Transfers a segment of a 1,4-α-D-glucan chain to a primary hydroxy group in a similar glucan chain and exodeoxyribonuclease I (EC 3.1.111)

Exonucleolytic cleavage in the 3′- to 5′-direction to yield nucleoside 5′-phosphates.

(c) Internal synonym inconsistencies

The Enzyme List has been in operation since 1952. Since then there have been many changes in nomenclature. Normally these have been made to correct or rationalize nomenclature, for example few now remember the furore caused by the change from DPN (even before that, it was called coenzyme I) to NAD, whereas the change from fructose 1,6-diphosphate to fructose 1,6-bisphosphate went relatively smoothly. Generally, the Enzyme List is punctilious about correcting entries but as will be discussed below, a few escape the notice of the IUBMB-IUPAC Joint Commission on Biochemical Nomenclature and those who use the Enzyme List. As discussed below, pathway tracing can be of value in finding such inconsistencies.

(d) No indication of **in vivo** *directionality*

Because the Enzyme List is restricted to providing factual data on the reaction catalysed, the reaction presented is, as far as possible, a mass-balancing equation. It is not meant to indicate the preferred equilibrium of the reaction or the direction in which the reaction is believed to operate *in vivo*. By convention, the direction chosen for the reaction in any given sub-subclass is the same for all enzymes. Systematic names are based on this written reaction and, therefore, carry no information about the favoured reaction direction. Although, this might seem to be less helpful than it might, it must be remembered that, for some enzymes, such as glutamate dehydrogenase [$NAD(P)^+$] (EC 1.4.1.2) and fructose-bisphosphate aldolase (EC 4.1.2.13), the preferred reaction direction varies with cellular conditions. Furthermore, the equilibrium constant of the reaction may be misleading in terms of the direction in which it actually operates *in vivo*.

For example, the equilibrium oxidation of ethanol dehydrogenase

$$CH_3CH_2OH + NAD^+ = CH_3CHO + NADH + H^+$$

very much favours ethanol formation under physiological conditions, but ethanol oxidation is the dominant direction *in vivo* because acetaldehyde (ethanal) is rapidly removed in the essentially irreversible reaction catalysed by aldehyde dehydrogenase (NAD^+) (EC 1.2.1.3)

$$CH_3CHO + NAD^+ + H_2O = CH_3COOH + NADH + H^+$$

Thermodynamic data for many enzymes can be found in the *GTD* Thermodynamics of Enzyme-catalysed Reactions database [7] and kinetic data are included in the *BRENDA* database, which may provide detail to determine reaction equilibria through Haldane relationships. It should be emphasised that only data that refer to "physiologically relevant conditions" should be used and that it is the thermodynamic properties of the overall metabolic system, not of the individual reaction, that are important in determining the flux direction [8, 9].

(e) No species information

In general the Enzyme List does not give information on the species, tissue or cell compartment in which the enzyme is found. Some information may be found in the references associated with each entry and the ''comments'' may refer to species in terms of behaviour that may not apply to the enzyme from all sources. For example, the entry for alcohol dehydrogenase contains the comment ''Acts on primary or secondary alcohols or hemiacetals; the animal, but not the yeast, enzyme acts also on cyclic secondary alcohols''. The *BRENDA* database, however, contains extensive species data that can be used in this context, and gene and protein databases may also provide valuable information about the species in which an enzyme might be expressed.

(f) Spontaneous (uncatalysed) reactions are not listed

Although one would not expect the Enzyme List to include reactions that are not enzyme-catalysed, such reactions do occur *in vivo* and will break a metabolic chain if not added to systems, such as the *Reactions Database* and *Explorer.* For example, the Enzyme List entry for L-threonine 3-dehydrogenase (see above) includes the comment ''The product spontaneously decarboxylates to aminoacetone''. This may be essential information for tracing the metabolic fates of L-threonine, since aminoacetone is known to be a substrate for the copper-containing amine oxidase (EC 1.4.3.6), (R)-aminopropanol dehydrogenase (EC 1.1.1.75) and glycine *C*-acetyltransferase (EC 2.3.1.29).

Maltose phosphorylase (EC 2.4.1.8) catalyses the reaction
maltose + phosphate = D-glucose + β-D-glucose 1-phosphate

but it would be difficult to trace the metabolic fates of the product with the information that β-D-glucose 1-phosphate spontaneously mutarotates to form α-D-glucose 1-phosphate.

(g) Overlapping specificities

It is not uncommon to find that more than one enzyme may be capable of catalysing the same reaction. These will be treated differently by the Enzyme List if they have sufficiently different substrate specificities. For example, an aldehyde may be a substrate for alcohol dehydrogenase (EC 1.1.1.1), alcohol dehydrogenase ($NADP^+$) (EC 1.1.1.2), aldehyde reductase (EC 1.1.1.21) and aldehyde oxidase (EC 1.2.3.1), among many other enzymes.

Carnitine *O*-palmitoyltransferase (EC 2.3.1.21) catalyses the reaction:

palmitoyl-CoA + L-carnitine = CoA + L-palmitoylcarnitine

and the comments indicate that it has a "Broad specificity to acyl group, over the range C_8 to C_{18}; optimal activity with palmitoyl-CoA". The related enzyme carnitine *O*-octanoyltransferase (EC 2.3.1.137) catalyses octanoyl-CoA + L-carnitine = CoA + L-octanoylcarnitine.

Thus both these enzymes will use octanoyl-CoA to extents that will depend on their respective activity levels, distribution and kinetic parameters.

In such cases, the necessary data are in the Enzyme List, supplemented by the additional information in *BRENDA*. The problem is simply one of ensuring that all enzymes that may work with a given metabolite are considered.

(h) Trouble-shooting through pathway reconstruction

Two examples will be used to illustrate how pathway tracing may be used to reveal deficiencies in the Enzyme List data. It has been known for many years that it was not possible to use the Enzyme List data to reconstruct the citric-acid cycle because the reaction catalysed by aconitase (aconitate hydrolase; EC 4.2.3.1) was given as:

$$\text{citrate} = \textit{cis}\text{-aconitate} + H_2O$$

However, in the operating citric-acid cycle, the overall process catalysed includes a second reaction:

$$\text{cis-aconitate} + H_2O = \text{isocitrate}$$

leading to an overall process of

$$\text{citrate} = \text{isocitrate}$$

Although this was known at the time the enzyme was first included in the Enzyme List (1961), only the first reaction was given because the equilibrium of the reaction very much favours citrate (the equilibrium mixture is 91% citrate, 6% isocitrate and 3% aconitate) and the existence of the second reaction was indicated only in the comments. Thus this is another example of the misleading inferences that can be drawn from considering isolated enzyme thermodynamics rather than system thermodynamics. The second reaction as well as the overall reaction are now included in the reaction field of the Enzyme List entry for aconitase.

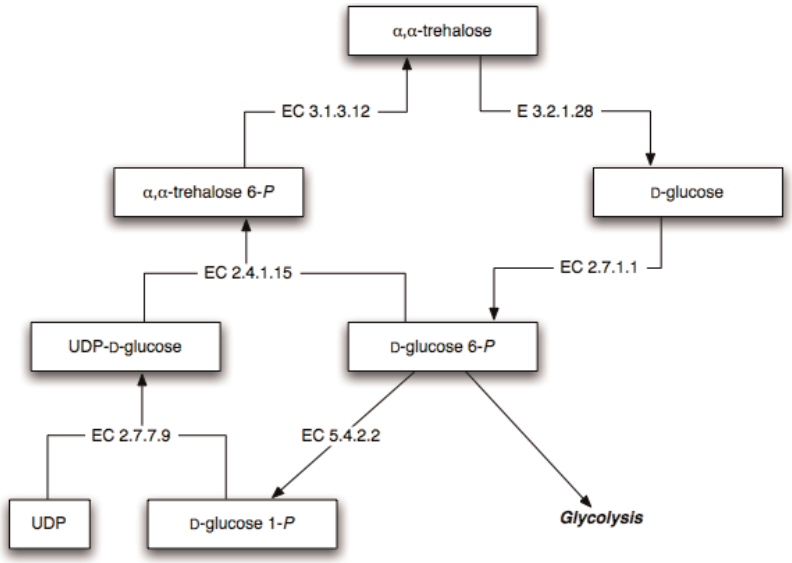


Figure 4. Trehalose metabolism in *M. grisea*.

The second example concerns the metabolism of trehalose [10], as shown in Fig. 4. The enzyme EC 2.4.1.15, α,α-trehalose-phosphate synthase (UDP-forming), was listed as catalysing the reaction:

$$\text{UDP-glucose} + \text{D-glucose 6-phosphate} = \text{UDP} + \alpha,\alpha\text{-trehalose 6-phosphate}$$

However, the enzymes that might use this product, such as trehalose-phosphatase (EC 3.1.3.12) were shown as using trelalose 6-phosphate:

$$\text{trehalose 6-phosphate} + H_2O = \text{trehalose} + \text{phosphate}$$

rather than α,α-trehalose 6-phosphate. Thus, any search for α,α-trehalose 6-phosphate would not reveal this, or other enzymes in the process. This was, in fact, an example of changes in nomenclature. In earlier formulations of the Enzyme List, some common enantiomeric designations were omitted. For example, it was assumed that all amino acids were L-amino acids unless otherwise specified. Similarly inositol was synonymous with *myo*-inositol and trehalose was α,α-trehalose. Since many biochemists were not familiar with these arcane conventions, the omitted enantiomeric designations have been added in more recent formulations of the Enzyme List but, somehow, this was not done for all the relevant trehalose entries. This has now been rectified.

(i) Conclusions

Not all of the problems discussed above concern the Enzyme List. Lacunae, such as those mentioned in the previous section, are filled as quickly as possible after they are discovered. While the Enzyme List primarily shows the enzyme-catalysed reaction, it is sometimes appropriate to include details of a spontaneous reaction that follows or precedes the enzyme-catalysed reaction, especially in cases where there would otherwise be a gap in a metabolic pathway.

Synonyms are important to allow the compounds to be found. The Enzyme List includes commonly used synonyms (other names) for each enzyme but it is not its function to include synonyms for all possible substrates. Synonyms are needed for searching *Reaction Database* and *Explorer* because many people use different names for the same compound and few use the often-unwieldy IUPAC-approved names. There are excellent small molecule databases, such as *ChEBI* [11] and *KEGG LIGAND* [12]. However, for convenience and since it is not uncommon to find that chemists prefer different alternative names from those favoured by biochemists and pharmacologists, *ChemFinder* [13] was searched for names to add to the *Reactions Database* and these were supplemented with information from the *Merck Index* [14]. Synonyms were then linked to the corresponding primary term for each compound, which were generally those used by the Enzyme List.

A big remaining challenge is to populate the *Reactions Database* with additional reactions that are not found in the Enzyme List, such as those provided by *BRENDA*. It will also be necessary to address the species problem, but at least for now, the problem of thermodynamic information may be best served by links to other sources.

References

[1] International Union of Biochemistry and Molecular Biology (IUBMB) Enzyme List, http://www.chem.qmw.ac.uk/iubmb/enzyme/.

[2] McDonald, A.G., Boyce, S., Moss, G.P., Dixon, H.B. and Tipton, K.F. (2007) ExplorEnz: a MySQL database of the IUBMB enzyme nomenclature. *BMC Biochem.* **8**:14, [http://www.enzyme-database.org/].

[3] Nicholson, D. (2000) The evolution of the IUBMB-Nicholson maps. *IUBMB Life* **50**(6):341 – 344, [http://www.iubmb-nicholson.org/].

[4] *Reaction Explorer,* http://www.reaction-explorer.org/

[5] GraphViz, http://www.graphviz.org/

[6] Schomburg, I., Chang, A., Hofmann, O., Ebeling, C., Ehrentreich, F. and Schomburg, D. (2002) BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem. Sci.* **27**:54 – 56, http://www.brenda-enzymes.info/.

[7] Goldberg, R.N., Tewari, Y.B., Bhat, T.N. (2004) Thermodynamics of enzyme-catalyzed reactions – a database for quantitative biochemistry. *Bioinformatics* **20**:2874 – 2877, http://xpdb.nist.gov/enzyme_thermodynamics/.

[8] Alberty, R.A. (2006) Calculation of equilibrium compositions of systems of enzyme-catalyzed reactions. *J. Phys. Chem. B* **110**:24775 – 24779.

[9] Alberty RA. (2003) Thermodynamics of Biochemical Reactions, John Wiley, U.S.A..

[10] Wang, Z.Y., Jenkinson, J.M., Holcombe, L.J., Soanes, D.M. *et al.* (2005) The molecular biology of appressorium turgor generation by the rice blast fungus *Magnaporthe grisea*. *Biochem. Soc. Trans*. **33**:384 – 388.

[11] ChEBI, http://www.ebi.ac.uk/chebi/.

[12] Goto, S., Okuno, Y., Hattori, M., Nishioka, T. and Kanehisa, M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30**:402 – 404, http://www.genome.jp/ligand/.

[13] ChemFinder, http://chemfinder.cambridgesoft.com/.

[14] The Merck-Index, http://www.merckbooks.com/mindex/ .

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

Protein Species – the Future Challenge for Enzymology

Hartmut Schlüter[1*], Maria Trusch[1] and Peter R. Jungblut[2]

[1]Core Facility Protein Analysis, Charité – University Medicine Berlin, Charitéplatz 1, 10117 Berlin, Germany,

[2]Central Core Facility Protein Analysis, Max Planck Institute for Infection Biology, Charitéplatz 1, 10117 Berlin, Germany

E-Mail: *hartmut.schlueter@charite.de

Received: 21st May 2008 / Published: 20th August 2008

Abstract

Protein species – this term was originally introduced by Jungblut *et al.,* 1996 [1] – to name protein variants, which differ in their exact chemical composition. The term protein species differentiates between splicing variants, truncated proteins and posttranslational modified proteins. The exact chemical composition critically determines the function of a protein. Phosphorylation can activate or inactivate enzymatic activities. There are already many other posttranslational modifications, which are known to regulate enzymatic activities. Truncations also play an important role in activating enzymes. Therefore the knowledge of the identity comprising 100% sequence coverage and every posttranslational modification at its exact position is fundamental for assigning a function to an individual protein species. However, knowledge about the relationship of the function of a protein and its exact chemical composition is still not yet fully taken into account in many investigations of enzymes. In most of the proteomics approaches protein identification is based on sequence coverage significantly below 100% and posttranslational modifications are more or less ignored. Also in studies investigating single enzymes, a total analysis of the chemical structure of the enzyme of interest is not usually performed. Therefore it is recommended that this issue should be addressed in biochemical and biological investigations. The total analysis of the chemical composition

of an enzyme is quite a big challenge; however it is even more challenging to develop strategies, which allow the validation of the correctness of the function–chemical composition relationship.

Introduction

During the period of the human genome-sequencing project in the nineties of the last century a holistic view developed in life sciences resulting in the ‘‘ome’’ terminology. The expression ‘‘proteome’’, which was introduced in 1995 by Wasinger *et al*.[2], is a hybrid from the words **prote**in and gen**ome**, thus implicating the description of the entire protein complement encoded by the DNA of an organism. There is a huge difference in complexity between the genome and the proteome, because the genome is more or less static in contrast to the proteome, which is continuously changing from the beginning of the life of an individual organism until its death, thereby reflecting the different stages of development as well as the interaction of the organism with its environment. The proteome of an organism is many orders of magnitude larger than its genome. It is estimated that the human genome consists of 20,000 to 25,000 genes, which code more than 500,000 protein species [1]. The large number of proteins arising from the significantly smaller number of genes, is caused by at least 7 protein structure modifying steps following the transcription of a gene and resulting in the individual protein species, a term which was originally defined by Jungblut *et al.* [1] by its exact chemical structure comprising posttranslational modifications. The RNA processing and protein structure modifying steps, thus being multiplicators of the number of products originating by one single gene, are the cause of the huge number of protein species arising from a relatively small quantity of genes. In eukaryotes the RNA-transcript originating from a gene is processed yielding mRNA. Alternative splicing multiplies the number of mRNA coded by one single gene (multiplicator 1) [3]. After protein synthesis at the ribosome translating the information from the mRNA into the protein sequence, the protein will be folded into its 3-dimensional structure. Many proteins are then directed to particular locations in the cell guided by their N-terminal signal sequences. On their journey to their destination two further steps can occur to change the chemical structure of the protein. The N-terminal signal sequence will be removed by proteases (multiplicator 2) and further proteolytical processing may occur to the protein thus activating or inactivating it (multiplicator 3). For example, protease-activated receptors (PAR) are a group of proteins activated by this process. On the path from the protein synthesis at the ribosomes to the mature protein, disulfide bonds may be formed and/or posttranslational modifications such as carbohydrate chains may be added (multiplicator 4). It is estimated that several hundred posttranslational modifications (PTM) exist in eukaryotes [4]. Two basic forms of PTMs can be distinguished. The static PTMs such as oligosaccharides are known to have a key role in protein targeting. Dynamic PTMs like phosphate groups critically determine the activity of a protein. Furthermore, alternative splicing can also occur on the protein level (multiplicator 5). Protein activity is not only determined by PTMs or the action of proteases but also by the interaction of a defined protein species with other biomolecules or ions, which bind non-covalently to the

protein species (multiplicator 6). For example metallo-proteases require adequate metal ion to be active. Many enzymes of energy metabolism need cofactors like NADH. Other enzymes will be activated by forming complexes with defined proteins. Further changes to the structure of a protein happen at the end of its life-time (multiplicator 7). Ubiquitinylation for example starts its degradation [3]. The ''record file'' shown in Fig. 1 summarizes the parameters which are necessary for a comprehensive description of an individual protein species. Beside all the protein modifying steps listed above being responsible for the exact chemical structure of an individual protein species, the number of copies (the concentration of a protein species), the time of occurrence and the localization of protein species are three further fundamental parameters for fully describing an individual protein species in the context of the proteome (Fig. 1). The localization of a protein species in space in the organism is described exactly by information about its coordinates in the cell, in the tissue and in the organ. Furthermore, each individual protein species is determined by the time interval of its appearance measured on a time scale starting with the beginning of life of the organism and ending with its death. All parameters summed up in Fig. 1 determine the function of a protein species. A change of one of these parameters results in a different protein species and can be associated with an alteration of the function, as will be shown in the following examples. For several decades the phosphorylation and dephosphorylation of proteins have been known to switch enzymatic activities on and off [5]. Nitrosylation radically alters the function of GAPDH. Nitrosylated GAPDH is a switch for apoptosis [6]. Truncations of the amino acid chain can activate proteolytic activities of proteases (Fig. 2) [7]. Binding partners modulate substrate specificity as known from thrombin [7]. Proteins being the product of different splicing on the mRNA level may differ drastically in their functions as reported from the products of the angiotensin-converting enzyme gene. The protein species, which is synthesized e. g. in endothelial cells, is part of the blood pressure regulating system, whereas the splicing product present in the testis is involved in male fertility [8].

Identity: **PROTEIN SPECIES A** coded by **GENE XY**			
Structure	Amino acid sequence	3D structure	PTMs
Interactions	Metal ions	Small organic molecules	proteins
Localisation	intracellular	type of cell	tissue
Concentration			
Time interval of appearance			
Function: Enzymatic properties			

Figure 1. ''Record file'' of a protein species: Parameters describing all aspects of an individual protein species.

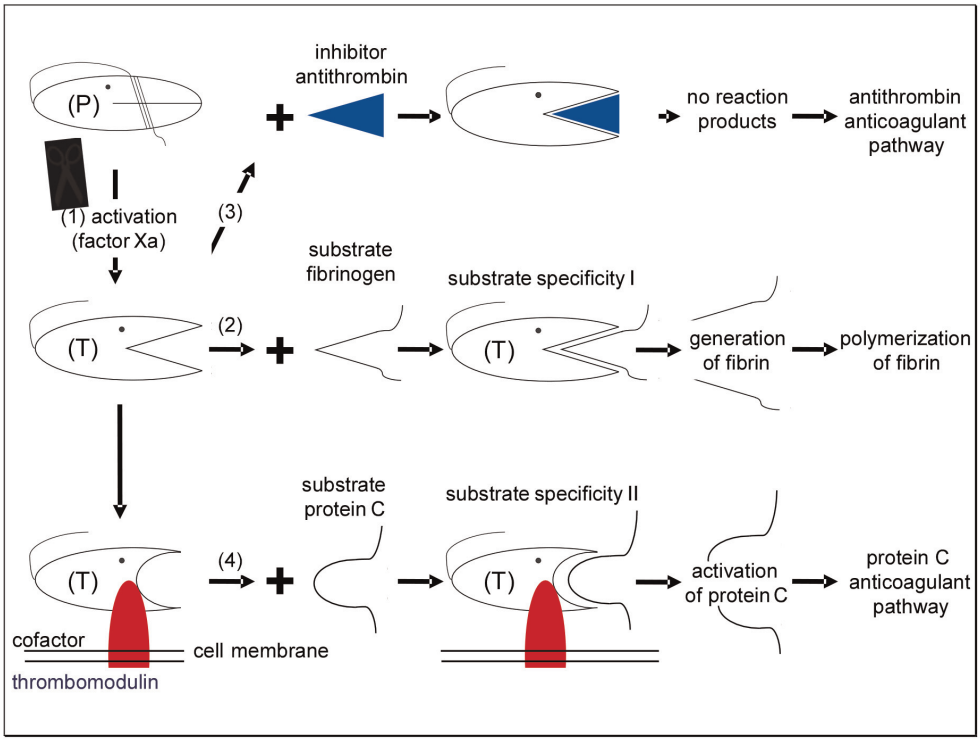


Figure 2. General scheme illustrating the dependency of identity of a protein species, defined by its exact chemical composition, and its functional status exemplified for thrombin. Thrombin (T) is generated from prothrombin (P) by factor Xa (1). Active thrombin cleaves its substrate fibrinogen into the fibrinopeptides (not shown) and fibrin (2) which polymerizes and therefore plays a central role in clog formation. Binding of thrombin to antithrombin inhibits the proteolytic activity of thrombin (3). Interaction of thrombin with the membrane bound cofactor thrombomodulin (4) results in a change of the substrate specificity of thrombin thus being able to activate protein C by proteolysis which leads to the protein C anticoagulant pathway.

In summary the function of a protein species depends on its exact chemical structure, its interactions with other biomolecules or ions (Fig. 2), its concentration as well as time and site of appearance in an organism (Fig. 1). To assign a defined function to a protein species unambiguously, knowledge of all the parameters listed in Fig. 1 is necessary. In this review the following question will be considered, which parameters that describe protein species can be yielded, by applying the strategies and techniques so far available for proteomics.

Access to the Protein Species Level

For the analysis of proteomes the "2-DE approach" and the "shot-gun approach" are the most established analytical procedures (Fig. 3). One of the big advantages of the 2-DE is that the presence of different individual protein species, coded by one single gene, can be resolved because of the very high resolution of the 2-DE and can be recognized (Fig. 3) – in this case these protein species will occur at different positions on the 2-DE gel [9, 10]. In the two-dimensional electrophoresis (2-DE) approach, the proteins are separated first by the two electrophoresis steps by their isoelectric point and their size. Many of the past studies using 2-DE and cataloguing the proteins present on the 2-DE gel, identified the proteins after enzymatic digestion via MALDI-MS peptide mass fingerprint (PMF) analysis followed by database searches [11]. Today the analysis of the enzymatic digests of protein spots cut from the 2-DE gel are performed preferably with PMF combined with tandem mass spectrometry (MS/MS) or pure MS/MS approaches yielding fragment spectra of the individual enzymatic peptides, thus ensuring a much higher grade of confidence concerning the correctness of the identity. In usual proteomics approaches full sequence coverage of the analysed proteins is not achievable with either the PMF approach or with the MS/MS analysis, as those peptides, which are modified by PTMs, are ignored or, even worse, wrongly interpreted in the case of the PMF analysis. Furthermore, some peptides may get lost during the LC steps or may not be resolved by the mass spectrometer, because they do not desorb into the gas phase, are not ionisable, or are too small or too big. As a result it is difficult to yield the identity of the protein species including the exact chemical composition, comprising 100% sequence coverage and all posttranslational modifications. However, recently Okkels *et al.* [12] resolved the complete chemical structure of 6 ESAT-6 protein species from a 2-DE gel of *Mycobacterium tuberculosis* and Myung demonstrated that 100% sequence coverage is achievable [13], if a multi-enzyme digestion strategy using trypsin, AspN, LysC and chymotrypsin is applied.

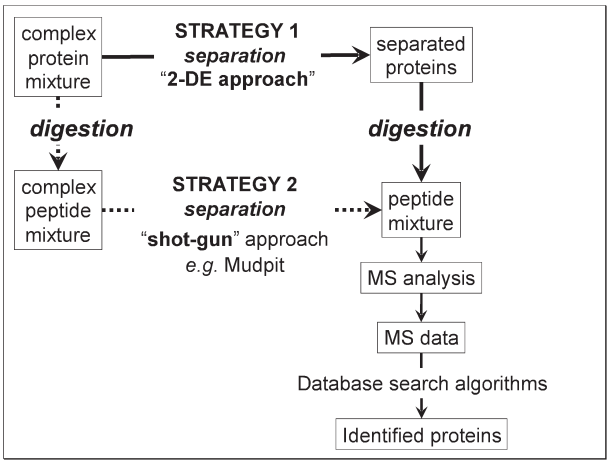


Figure 3. The two main strategies of proteomics.

In contrast to the 2-DE approach the recognition of the presence of different protein species originating from one gene is not possible using the ”shot-gun approach”. Here, many peptides of these protein species resulting from enzymatic digestion have the same amino acid sequence, thus eluting in one single peak (Fig. 4). The 2-DE pattern shows, that there are different protein species stemming from a single gene. In contrast, in the case of the shot-gun approach the origin of the enzymatic cleavage peptides in the LC (liquid chromatography) chromatogram cannot be assigned to the 4 protein species. From the data, recognizing the presence of different protein species is not feasible. As a result no statement about the quantitative relationship of the 4 protein species is possible (Fig. 5). More sophisticated shot-gun approaches such as MudPIT (Multidimensional Protein Identification Technology) [14, 15], also start with the enzymatic digestion of complex protein mixtures, followed by the separation of the generated peptides and have the same limitation.

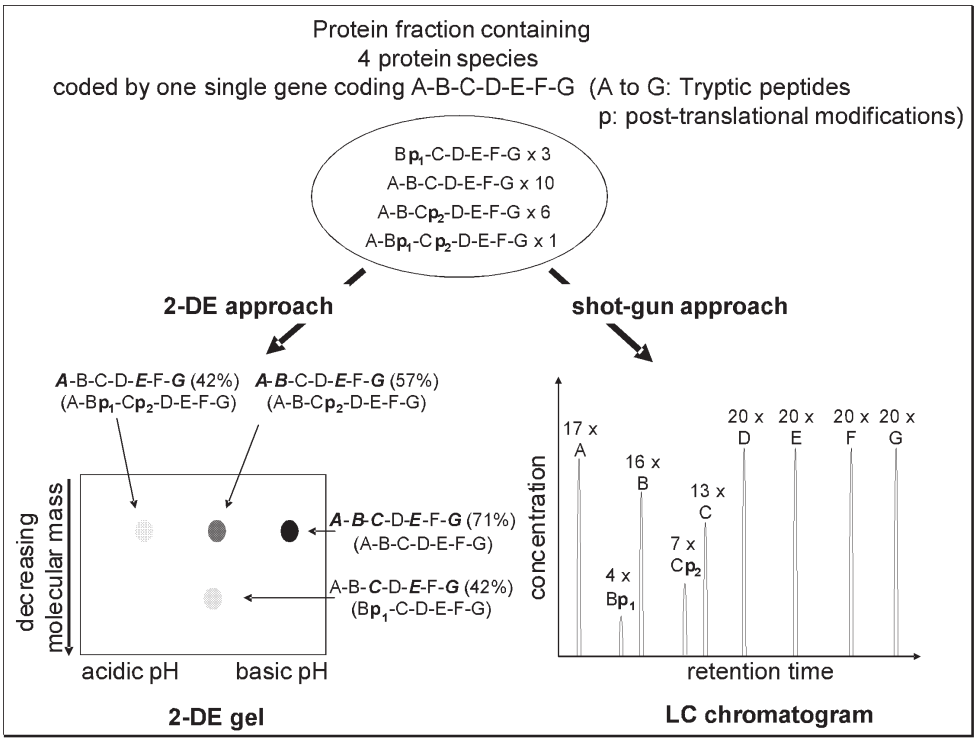


Figure 4. Simplified model of typical results of the analysis of a protein fraction containing 4 protein species present in 4 different concentrations (1x, 3x, 6x, 10x), coded by a single gene, performed either by the 2-DE approach or the shot-gun approach. The letters A to F represent peptides generated by enzymatic cleavage. Posttranslational modifications: p_1, p_2. 2-DE approach: The different quantities of the protein species are represented by the different greys. In this model only the tryptic peptides A, B, C, E, F and G (bold italic letters) yield data for the identification of the protein via data base searches resulting in 1 protein name (here named A-B-C-

D-E-F-G) for 4 different protein species (A-Bp_1-Cp_2-D-E-F-G; A-B-Cp_2-D-E-F-G; A-B-C-D-E-F-G; Bp_1-C-D-E-F-G). The peptide coverage (in percent) differs from protein species to protein species.

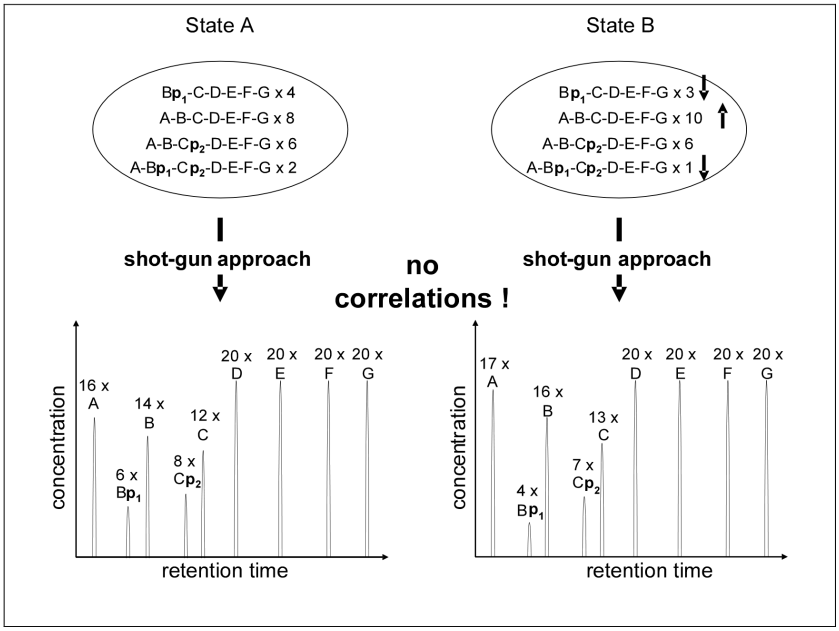


Figure 5. Simulated quantitative analysis by liquid chromatography of a protein fraction from state A and state B, differing in the concentrations of 4 protein species coded by one single gene. In the idealized chromatograms "absolute quantities" were yielded. A-B-C-D-E-F-G in state A is down-regulated 20% compared to state B. A-Bp-Cp-D-E-F-G in state A is up-regulated 100%. Bp-Cp-D-E-F-G in state A is up-regulated 25%.

Quantification of Protein Species with Labelling Reagents

Several quantitative approaches employing chemical labelling with stable isotopes are often combined with the shot-gun strategy [16], for instance ICAT (Isotope-Coded Affinity Tag) [17], ITRAQ (Isobaric Tag for Relative and Absolute Quantification) [18] and SILAC (Stable Isotope Labeling by Amino acids in Cell culture) [19]. These methods have already proved their applicability in the determination of relative protein amounts in different states of biological systems [20]. However in the case of the presence of several protein species which are all coded by one single gene, the shot-gun approach may yield false results since the enzymatic cleavage peptides have identical amino acid sequences, but originate from different protein species the concentrations of which usually change individually, can no longer be distinguished after enzymatic digestion (Fig. 4). The detection of the up- or down-

regulation of individual protein species, which may be of importance concerning their function, is not possible, as exemplified in Fig. 5. This problem can be circumvented by first labelling the intact proteins with reagents for differential quantification, followed by the separation of the intact protein species either by electrophoresis or by liquid chromatography or combinations of both. After their purification to near homogeneity the protein species will be digested enzymatically. This quantitative analytical approach can be performed with the ICAT reagent [21], the ITRAQ reagent (isobaric tags for relative and absolute quantification) [22], the ICPL reagent (Isotope-Coded Protein Labeling) [23] and the MECAT reagent [24] to mention only some of the possible labelling procedures.

In the ICAT method, proteins from the two states to be compared are labelled at cysteine residues with heavy and light tags, respectively, carrying a biotin moiety. The labelled proteins are then mixed, separated by 2-DE and digested. After a cation-exchange chromatography step, the mixed peptides are affinity purified via immobilized avidin. Peaks corresponding to the same peptide are identified as doublets in mass spectra due to the mass difference between light and heavy isotopes. The peak intensities of the peptides correlate directly with the relative abundance of the proteins in the two states.

The ITRAQ technology makes use of a four-plex and eight-plex set of amine reactive isobaric tags to derive peptides at the N-terminus and the lysine side chains, thus labelling all peptides in a digest mixture. In MS, peptides labelled with any of the isotopic tags are indistinguishable (isobaric). Upon fragmentation in MS/MS, signature ions (m/z from 114 to 117) are produced, which provide quantitative information upon integration of the peak areas [18].

The ICPL label is an isotope coded nicotinoyl group coupled to an amino reactive N-hydroxysuccinimide (NHS). NHS targets unmodified amino groups of lysine and the N-terminus. The benefit of lysine labelling is the more frequent occurrence in proteins as compared to cysteine (≈ 6% vs. 1.5%). A weakness of lysine labelling is that trypsin does not cleave ICPL-modified lysine sites. Trypsin digestion of ICPL labelled proteins results in rather long peptides, because cleavage occurs solely at arginine residues. Because of the loss of basic amino groups, proteins are shifted to the acidic side of the gel, which is an advantage for basic proteins.

The MECAT reagent contains a chelating group, which binds lanthanoid metal ions with high affinities and can be coupled covalently to proteins. With the MECAT reagent ‘‘loaded’’ with different lanthanoid cations, proteins from different states can be labelled. Since the metal elements are detected by the ICP-MS even an absolute quantification of MECAT- labelled proteins is possible. A further advantage of the ICP-MS based quantification is the large linear detection range of 7 orders of magnitude. In comparison to all other quantification techniques based on labelling, MECAT is the only quantification method which allows the detection of quantitative differences without the need for digesting the

proteins. Focusing on those proteins, which occur in different concentrations, will reduce significantly the number of proteins which have to be analysed for their identity. MECAT is also compatible with the mass spectrometric top-down approach, which introduces intact protein ions into the gas phase of the mass spectrometer by an electro-spray ionization (ESI) source, followed by determining the molecular mass of the protein and thereafter fragmenting the proteins, thus yielding more specific data for the characterization of sequence and posttranslational modifications than by peptides from the protein's digestion.

Intact protein identification by MS/MS – the top-down approach – was demonstrated first by Mortz *et al.* 1996 [25]. In the top-down approach, intact protein ions are introduced into the mass spectrometer and then fragmented, yielding the molecular masses of both the protein and the fragment ions. If a complete set of informative fragment ions are detected, this analysis can supply a complete description of the amino acid sequence of the protein and reveal all of its posttranslational modifications. The main challenge of the top-down approach is given by the problem of producing wide-ranging gas-phase fragmentation of intact protein ions. By pumping large amounts of energy into the ionized protein in the gas-phase, Han *et al.* significantly improved the applicability of the top-down approach towards the analysis of protein species [26]. The authors reported that they could obtain very informative fragmentation for proteins with molecular masses larger than 200 kDa. The essential fragmentation element of the top-down approach appears within reach since the current introduction of two other very helpful methods for fragmenting proteins – electron transfer dissociation [27] and electron capture dissociation [28]. Nevertheless top-down still is a technique for studying single purified proteins. Other challenging problems wait to be overcome before the top-down approach can be regarded as really strong for proteomics studies. The need to separate complex mixtures of proteins prior to mass spectrometric top-down analysis remains a central challenge. The extremely different physico-chemical properties of the individual proteins make them difficult to handle as mixtures without gaining awesome losses of certain proteins or leaving the proteins incompatible with mass spectrometry. Even more demanding is the need to separate the protein species which are only slightly different. Sensitivity is also a major challenge, because effective fragmentation of a high-molecular-mass protein implies that the protein will fracture in a large number of different ways. Therefore, the intensity of the resulting fragments will be weak compared to that of small peptides. The most practicable approach today is the top-down 2-DE separation of protein species combined with the bottom-up identification of the peptides by MS.

Label-Free Quantification of Protein Species

Is it possible to quantify protein species without labelling? This question can be accepted for hypothesis-driven investigations, if an enzymatic cleavage peptide of a protein species, which belongs to a family of the protein species derived from one single gene, is known, which is unique according to its chemical structure. This unique peptide can be quantified by

the selected-reaction-monitoring method (SRM, also called multiple-reaction monitoring) [29]. This type of mass spectrometric experiment is very common in quantifying drugs and their metabolites [30]. SRM experiments are designed for obtaining the maximum sensitivity for detection of target compounds, as shown by Onisko *et al*., who quantified attomole amounts of the prion protein with the SRM method in the brains of terminally ill Syrian hamsters [31]. Knowing the mass and structure of the peptide, it is possible to predict the precursor m/z and a fragment m/z (SRM transition) [32]. The advantage of measuring the fragment ions is the reduction of background interferences, especially for complex mixtures such as plasma. Fragmentation in the majority of cases offers one or more unique fragment ions. The combination of the specific parent mass and the unique fragment ion is generally an unambiguous method to monitor and quantify selectively the peptides of interest. The SRM method has already been applied for identifying and quantifying protein posttranslational modifications [29] thus demonstrating that the SRM method is suitable for quantifying protein species via their tryptic peptides which are unique within the tryptic peptides of the protein species family. Quantification of a protein by the SRM method via its tryptic peptides is reported by Berna *et al*., who analysed the time course of the protein myosin light chain 1 in rat serum following a 50 mg/kg subcutaneous dose of isoproterenol, a β-adrenergic receptor agonist known to induce cardiac injury [33].

Conclusion and Perspectives

Since the determination of the exact chemical composition in relation to its function will be time consuming, the question arises as to which data are really needed to achieve progress in life science. In the past the discovery and identification of key players of molecular processes in organisms helped in the understanding of fundamental aspects in biochemistry and molecular biology. This knowledge was also the basis for the development of new drugs. Therefore it should be beneficial to apply the new tools of proteomics for the identification of protein species linked to defined functions. This aim can be achieved by comparing quantitatively the protein species composition of two biological systems in the quiescent (control) state and in a defined activated state thereby identifying those protein species, the concentrations of which have changed. An alternative, yet classical way of identifying the function and the exact chemical composition of a protein species comprises its purification towards near homogeneity guided by a functional assay. An example following this strategy is given in Rykl *et al.,* [34] where a protease with a defined function was purified using a system (PPS [35]) for the determination of optimum chromatographic purification steps and detecting the protease via its catalytic properties with a mass spectrometric assay (MES [36]). The purified protease was identified via mass spectrometric analysis of the tryptic peptides of the purified active fraction. Although in this study the active protease was identified without determining the exact chemical structure, the workflow comprises the potential to elucidate the exact chemical structure.

Since the protein function is critically associated with the exact chemical composition of the protein species it should become obligatory in future analysing and publishing of the complete amino acid sequence and identifying every posttranslational modification of the individual enzyme under investigation. By applying targeted protein analysis to those key protein species will make it possible to follow these proteins over a longer period of time since the restriction of the analysis towards a few protein species will enable the analyses of a larger quantity of samples collected at many different time points. This reinforces the need for flexibility in our interpretation of sequence data and also illustrates the importance of more traditional approaches such as genetics and biochemistry for discovering gene functions.

References

[1] Jungblut, P., Thiede, B., Zimny-Arndt, U., Muller, E., Scheler, C., Wittmann-Liebold, B., Otto, A. (1996) Resolution power of two-dimensional electrophoresis and identification of proteins from gels. *Electrophoresis* **17**:839 – 847.

[2] Wasinger, V.C., Cordwell, S.J., Cerpa-Poljak, A., Yan, J.X., Gooley, A.A., Wilkins, M.R., Duncan, M.W., Harris, R., Williams, K.L., Humphery-Smith, I. (1995) Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis* **16**:1090 – 1094.

[3] Berg, J.M., Tymoczko, J.L., Stryer, L. *Synthesizing the Molecules of Life. Biochemistry,* 5th Ed.; W.H. Freeman and Company, U.K. 2002.

[4] Collins, M.O., Yu, L., Husi, H., Blackstock, W.P., Choudhary, J.S., Grant, S.G. (2005) Robust enrichment of phosphorylated species in complex mixtures by sequential protein and peptide metal-affinity chromatography and analysis by tandem mass spectrometry. *Sci. STKE* 2005, pl6.

[5] McCarty, K.S., McCarty, K.S.J. (1974) Protein modification, metabolic controls, and their significance in transformation in eukaryotic cells. *J. Natl Cancer Inst.* **53**:1509 – 1514.

[6] Hara, M.R., Snyder, S.H. (2006) Nitric oxide-GAPDH-Siah: A novel cell death cascade. *Cell Mol. Neurobiol.* **26**:527 – 538.

[7] Lane, D.A., Philippou, H., Huntington, J.A. (2005) Directing thrombin. *Blood* **106**:2605 – 2612.

[8] Franke, F.E., Pauls, K., Metzger, R., Danilov, S.M.(2003): Angiotensin I-converting enzyme and potential substrates in human testis and testicular tumours. *APMIS* **111**:234 – 243.

[9] Klose, J., Nock, C., Herrmann, M., Stuhler, K., Marcus, K., Bluggel, M., Krause, E., Schalkwyk, L.C., Rastan, S., Brown, S.D.,Büssow, K., Himmelbauer, H., Lehrach, H. (2002) Genetic analysis of the mouse brain proteome. *Nature Genet.* **30**(4):385 – 393.

[10] Scheler, C., Muller, E., Stahl, J., Muller-Werdan, U., Salnikow, J., Jungblut, P. (1997) Identification and characterization of heat shock protein 27 protein species in human myocardial two-dimensional electrophoresis patterns. *Electrophoresis* **18**:2823 – 2831.

[11] Gras, R,. Muller, M. (2001) Computational aspects of protein identification by mass spectrometry. *Curr. Opin. Mol. The.* **3**:526 – 532.

[12] Okkels, L., Muller, E., Schmid, M., Rosenkrands, I., Kaufmann, S., Andersen, P., Jungblut, P. (2004) CFP10 discriminates between nonacetylated and acetylated ESAT-6 of Mycobacterium tuberculosis by differential interaction. *Proteomics* **4**:2954 – 2960.

[13] Myung, J.K., Frischer, T., Afjehi-Sadat, L., Pollak, A., Lubec, G. (2007) Mass spectrometrical analysis of the processed metastasis-inducing anterior gradient protein 2 homolog reveals 100% sequence coverage. *Amino Acids* Epub ahead of print.

[14] Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., Yates, J.R.I.(1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**:676 – 682.

[15] Emmett, M.R., Caprioli, R.M. (1994) Micro-electrospray mass spectrometry: ultra-high-sensitivity analysis of peptides and proteins. *J. Am. Soc. Mass. Spectrom.* **5**:605 – 613.

[16] MacCoss, M.J., Matthews, D.E.(2005) Quantitative MS for proteomics. *Analyt. Chem.* **77**:295A – 302A.

[17] Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., Aebersold, R.(1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**:994 – 999.

[18] Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Marin, S., Bartlet-Jones, M., Feng He, Jacobsen, A., Pappin, D.J. (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3**:1154 – 1169.

[19] Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., Mann, M.(2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**:376 – 386.

[20] Frohlich, T., Arnold, G.J. (2006) Proteome research based on modern liquid chromatography–tandem mass spectrometry: separation, identification and quantification. *J. Neural Transm.* **113**:973 – 994.

[21] Smolka, M., Zhou, H., Aebersold, R. (2002) Quantitative protein profiling using 2-DE, ICAT, and mass spectrometry. *Mol. Cell. Proteomics* **1**:19 – 29.

[22] Wiese, S., Reidegeld, K..A., Meyer, H.E., Warscheid, B. (2007) Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics* **7**:340 – 350.

[23] Schmidt, A., Kellermann, J., Lottspeich, F. (2005) A novel strategy for quantitative proteomics using isotope coded protein labels. *Proteomics* **5**:4 – 15.

[24] Krause, M., Scheler, C., Bottger, U., Weisshoff, H., Linscheid, M. (2006) Method and reagent for the specifically identifying and quantifying one or more proteins in a sample. *Patent US 2006/0246530 A1* (10/518,727).

[25] Mørtz, E., O'Connor, P.B., Roepstorff, P., Kelleher, N.L., Wood, T.D., McLafferty, F.W., Mann, M. (1996) Sequence tag identification of intact proteins by matching tanden mass spectral data against sequence data bases. *Proc. Natl Acad. Sci. U.S.A.* **93**(16):8264 – 8267.

[26] Han, X., Jin, M., Breuker, K., McLafferty, F.W. (2006) Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science* **314**(5796):109 – 112.

[27] Coon, J.J., Ueberheide, B., Syka, J.E., Dryhurst,D.D., Ausio, J., Shabanowitz, J., Hunt, D.F. (2005) Protein identification using sequential ion/ion reactions and tandem mass spectrometry. *Proc. Natl Acad. Sci. U.S.A.* **102**(27):9463 – 9468.

[28] Zubarev, R.A., Horn, D.M., Fridriksson, E.K., Kelleher, N.L., Kruger, N.A., Lewis, M.A., Carpenter, B.K., McLafferty, F.W. (2000) Electron capture dissociation for structural characterization of multiply charged protein cations. *Analyt. Chem.* **72**(3):563 – 573.

[29] Cox, D.M., Zhong, F., Du, M., Duchoslav, E., Sakuma, T., McDermott, J.C. (2005) Multiple reaction monitoring as a method for identifying protein posttranslational modifications. *J. Biomol. Tech.* **16**(2):83 – 90.

[30] Bruins, A.P., Covey, T.R., Henion, J.D. (1987) Ion spray interface for combined liquid chromatography/atmospheric pressure ionization mass spectrometry. *Analyt. Chem.* **59**:2642 – 2646.

[31] Onisko, B, Dynin I., Requena, J.R., Silva, C.J., Erickson, M., Carter, J.M. (2007) Mass spectrometric detection of attomole amounts of the prion protein by nanoLC/MS/MS. *J. Am. Soc. Mass. Spectrom.* **18**:1070 – 1079.

[32] Tamvakopoulos, C. (2007) Mass spectrometry for the quantification of bioactive peptides in biological fluids. *Mass. Spectrom. Rev.* **26**(3):389 – 402.

[33] Berna, M.J., Zhen, Y., Watson, D.E., Hale, J.E., Ackermann, B.L. (2007) Strategic use of immunoprecipitation and LC/MS/MS for trace-level protein quantification: myosin light chain 1, a biomarker of cardiac necrosis. *Ana.l Chem.* **79**(11):4199 – 4205.

[34] Rykl, J., Thiemann, J., Kurzawski, S., Pohl, T., Gobom, J., Zidek, W., Schlüter, H. (2006) Renal cathepsin G and angiotensin II generation. *J. Hypertens.* **24**(9):1797 – 1807.

[35] Thiemann, J., Jankowski, J., Rykl, J., Kurzawski, S., Pohl, T., Wittmann-Liebold, B., Schlüter, H. (2004) Principle and applications of the protein-purification-parameter screening system. *J. Chromatogr. A* **1043**:73 – 80.

[36] Schlüter, H., Jankowski, J., Rykl, J., Thiemann, J., Belgardt, S., Zidek, W., Wittmann, B., Pohl, T. (2003) Detection of protease activities with the mass-spectrometry-assisted enzyme-screening (MES) system. *Anaytl. Bioanaytl. Chem.* **377**:1102 – 1107.

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

Symbolic Control Analysis of Cellular Systems

Johann M. Rohwer*, Timothy J. Akhurst and Jan-Hendrik S. Hofmeyr

Triple-J Group for Molecular Cell Physiology, Department of Biochemistry, Stellenbosch University, Private Bag X1, ZA-7602 Matieland, South Africa

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

Received: 19th May 2008 / Published: 20th August 2008

Abstract

Metabolic Control Analysis (MCA) is a powerful quantitative framework for understanding and explaining the relationships between the global steady-state properties of a cellular system in terms of control coefficients, and the local properties of the individual components of the system in terms of elasticities. The elasticities are apparent kinetic orders, which derive directly from the kinetic properties of the enzymes. Since MCA relates elasticities to control coefficients through a matrix inversion, it allows one to predict and to quantify how the kinetics of individual enzymes affect the systemic behaviour of biological pathways. Most often this problem has been solved numerically, with algebraic and symbolic control analysis having been tackled less frequently. We present here a general implementation of the symbolic matrix inversion of MCA through symbolic algebraic computation. The algebraic expressions thus generated allow an in-depth analysis of where the control within a system lies and which parameters have the greatest effect on this control distribution, even if the exact values of the elasticities or control coefficients are unknown.

Introduction

Metabolic Control Analysis (MCA; see [1, 2]) is a powerful quantitative framework for analysing and quantifying the control and regulation of cellular pathways. It was developed independently by Kacser and Burns [3] and Heinrich and Rapoport [4] in the early 1970 s. One of the fundamentals of MCA is that it quantifies the steady-state behaviour of a system in terms of global properties (termed control coefficients) and local properties (termed elasticities). Mathematically, a control coefficient is defined as:

$$C_{v_i}^{y} = \frac{v_i}{y} \cdot \left(\frac{\partial y}{\partial v_i}\right)_{ss} = \left(\frac{\partial y/y}{\partial v_i/v_i}\right)_{ss} = \left(\frac{\partial \ln y}{\partial \ln v_i}\right)_{ss} \tag{1}$$

where y is any steady-state variable of the system (e. g. flux or species concentration) and v_i is the local rate of step i. The control coefficient thus quantifies how sensitive the variable y is to changes in local rate v_i. The subscript ss indicates that the entire system relaxes to a new steady state after the perturbation in v_i; hence, control coefficients are systemic properties.

An elasticity on the other hand, quantifies the effect of any molecular species or parameter that directly affects a unit step on the local rate through this step in isolation:

$$\varepsilon_{s_j}^{v_i} = \frac{s_j}{v_i} \cdot \left(\frac{\partial v_i}{\partial s_j}\right)_{s_k,s_l,\ldots} = \left(\frac{\partial v_i/v_i}{\partial s_j/s_j}\right)_{s_k,s_l,\ldots} = \left(\frac{\partial \ln v_i}{\partial \ln s_j}\right)_{s_k,s_l,\ldots} \tag{2}$$

where v_i is the local rate of unit step i in the system and s_j the concentration of any molecular species (substrate, product or effector) or parameter (e. g. K_m) that affects the step directly. The subscript $s_k,s_l,\ldots$ indicates that the concentrations of all other substrates, products and effectors are kept constant at their prevailing values while s_j is varied. Elasticities are apparent kinetic orders, which derive directly from the kinetics of the enzymes; their properties are local to the particular step and do not depend on the rest of the system.

A particular strength of MCA lies in the fact that it relates control coefficients to elasticities through a number of summation and connectivity relationships [3 – 5], thus enabling one to calculate systemic behaviour and control from the local properties of each of the components of the reaction network. These relationships have been combined into a variety of matrix equations (e. g. [6 – 8]), of which a particularly elegant form is that of Hofmeyr and Cornish-Bowden [9]:

$$\mathbf{C^i} \times \mathbf{E} = \mathbf{I} \qquad \Rightarrow \qquad \mathbf{C^i} = \mathbf{E}^{-1} \tag{3}$$

where $\mathbf{C}^{\mathbf{i}} = [\,\mathbf{C}^{\mathbf{J_i}} \quad \mathbf{C}^{\mathbf{s_i}}]^T$ is a matrix of independent flux- and concentration-control coefficients, and $\mathbf{E} = [\mathcal{K} \quad -\varepsilon_s\mathcal{L}]$ is a matrix of all structural and local properties of the system. $\mathcal{K}$ is the scaled kernel matrix relating dependent fluxes to independent fluxes, $\varepsilon_\mathbf{s}$ is a matrix of elasticities, and $\mathcal{L}$ is the scaled link matrix relating the time-derivatives of the dependent species to the independent species – for further details, the reader is referred to [9, 10].

From Equation (3) it follows that the control coefficients can be calculated by inversion of the local matrix **E**. Computational methods have traditionally focused on the numerical solution of this matrix inversion of MCA. A great number of simulation packages exist for computational systems biology (see http://sbml.org/), many of which can also do numerical MCA. Algebraically and symbolically, the problem has been tackled less frequently; the notable examples include control-pattern analysis [11] and the MetaCon program [12]. In both cases the aim was to derive algebraic relations between the control and elasticity coefficients. As it is often experimentally difficult to determine all parameter values for large models, it would be useful to develop a general approach for identifying key system parameters without knowing their actual values.

In this paper we present SymCA, a general implementation through symbolic algebraic computation of the matrix inversion of MCA (Equation (3)). The symbolic algebra manipulations are performed with Maxima, a powerful open-source algebra software. We have developed a Python interface to Maxima to access its symbolic capabilities from within PySCeS [13], the computational systems biology software developed in our group. This enabled the successful implementation of the $\mathbf{C}^{\mathbf{i}} = \mathbf{E}^{-1}$ inverse problem for systems of any size and complexity. As will be shown below, the algebraic expressions generated, which are factorized for easier interpretation, allow an in-depth analysis of where the control within a system lies. The rest of the paper is organized as follows: first, the design of the SymCA software is described, then its application is illustrated with two examples, and finally, the results are discussed in a broader context.

The SymCA Software

Maxima (http://maxima.sourceforge.net/) is an open-source symbolic algebra software licenced under the GNU GPL. It runs on multiple operating systems and is capable of performing symbolic calculus, linear algebra, matrix manipulations and simplifications, factorization, etc. Maxima is a descendant of Macsyma, the famous algebra software developed at the MIT in the 1960 s. The program is implemented in common Lisp. Because of its open-source licence and multi-platform capabilities, we chose Maxima to integrate symbolic control analysis capabilities into PySCeS.

SymCA (Symbolic Control Analysis) refers to the combination of Python with PySCeS and Maxima. The code is released under an open-source licence together with the PySCeS project at http://pysces.sourceforge.net/. The operation of the software is summarized in Fig. 1.

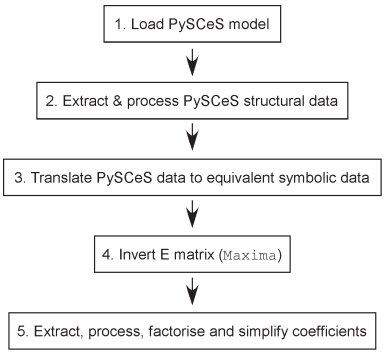


Figure 1. Operation of the SymCA software.

The symbolic algebra capabilities of Maxima are accessed through a Python interface, which we have developed using the subprocess module. The analysis is started by loading a model into PySCeS and performing a structural (stoichiometric) analysis. This extracts and processes the structural data and generates the $\mathcal{K}$, ε_s and $\mathcal{L}$ matrices, from which the **E** matrix (Equation (3)) is assembled. SymCA subsequently translates these data into equivalent *symbolic* data by substituting text strings for the elasticity and species names into the matrix entries. The symbolic **E** matrix is then passed to Maxima using our newly developed interface, where it is symbolically inverted and the results passed back to Python. In the final step, SymCA extracts, processes and simplifies the symbolic control coefficient expressions, which can then be output in various formats. The whole process, including the details of this simplification, are illustrated with an example (see SymCA by Example below).

Additional features

The SymCA software has a number of additional features built in for error checking and user convenience:

- The control coefficient expressions can be output in $\LaTeX$ [14] format for easy typesetting and incorporation into documents. The text format for loading the expressions into Maxima is also saved.
- The expressions are automatically *checked for correctness* by substituting the numerical values of the elasticities from the PySCeS model, calculating the control coefficient values, and comparing these to the control coefficients calculated numerically directly by PySCeS.
- SymCA allows *numerical substitution* (e. g. 0 or 1) of selected elasticity values to generate a simplified control coefficient expression subject to certain assumptions

about the model, such as a particular reaction being saturated with substrate (i. e. following zero-order kinetics) or operating in the first-order range.

- Expressions for *parameter-response coefficients* can be calculated from the partitioned response property [3]: $C_p^y = e_p^{v_i} \times C_{v_i}^y$, where p is an external parameter acting on step i. In fact, this merely involves multiplying the control coefficient expression by the parameter elasticity of step i towards p.

SymCA by Example

The operation of the SymCA software will be illustrated with two examples: first, a small "core" model, which has been designed to emphasize the salient features, and next, a more realistic model of fermentation pathways in *Lactococcus lactis*.

Simple metabolic pathway

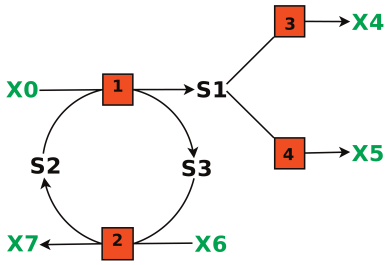


Figure 2. A simple metabolic pathway with a branch-point and a moiety-conserved cycle [10].

Figure 2 shows a scheme of a "minimal" system containing an example of the most important structural features observed in metabolic pathways, i. e. a branch-point and a moiety-conserved cycle. This example has been treated in detail to explain the matrix method of MCA [10] and will also be used here to demonstrate the workings of the SymCA software.

For the system in Fig. 2, Equation (3) reads:

$$\begin{bmatrix} C_3^{J3} & C_4^{J3} & C_1^{J3} & C_2^{J3} \\ C_3^{J4} & C_4^{J4} & C_1^{J4} & C_2^{J4} \\ C_3^{s1} & C_4^{s1} & C_1^{s1} & C_2^{s1} \\ C_3^{s2} & C_4^{s2} & C_1^{s2} & C_2^{s2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\varepsilon_{s1}^{v3} & 0 \\ 0 & 1 & -\varepsilon_{s1}^{v4} & 0 \\ \frac{J_3}{J_1} & \frac{J_4}{J_1} & -\varepsilon_{s1}^{v1} & (\varepsilon_{s3}^{v1}\frac{s_2}{s_3} - \varepsilon_{s2}^{v1}) \\ \frac{J_3}{J_2} & \frac{J_4}{J_2} & 0 & (\varepsilon_{s3}^{v2}\frac{s_2}{s_3} - \varepsilon_{s2}^{v2}) \end{bmatrix}^{-1} \tag{4}$$

For Equation (4) to be computed by SymCA, the **K** and **L** matrices (which are provided by PySCeS from the stoichiometric analysis) first have to be scaled to $\mathcal{K}$ and $\mathcal{L}$ as described in [10]. Subsequently, the matrix product $-\varepsilon_s\mathcal{L}$ is computed. Before proceeding with the symbolic matrix inversion, dependent fluxes are expressed in terms of independent ones according to $\mathbf{J} = \mathbf{K}\mathbf{J_i}$, which for the system in Fig. 2 reads:

$$\begin{bmatrix} J_3 \\ J_4 \\ J_1 \\ J_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} J_3 \\ J_4 \end{bmatrix} \tag{5}$$

Equation (5) shows that J_1 and J_2 are both equivalent to $J_3 + J_4$; thus, in the RHS of Equation (4) J_2 is replaced with J_1. The symbolic matrix **E** is then passed to Maxima for inversion. The determinant of **E** (also computed by Maxima) gives the common denominator of all the control coefficient expressions. For the example model in Fig. 2, this reads:

$$\begin{aligned} \sum = {} & \frac{J_4\varepsilon^4_{s1}\varepsilon^2_{s3}}{s_3} - \frac{J_4\varepsilon^4_{s1}\varepsilon^2_{s2}}{s_2} - \frac{J_4\varepsilon^4_{s1}\varepsilon^1_{s3}}{s_3} + \frac{J_4\varepsilon^4_{s1}\varepsilon^1_{s2}}{s_2} + \frac{J_3\varepsilon^3_{s1}\varepsilon^2_{s3}}{s_3} \\ & - \frac{J_3\varepsilon^3_{s1}\varepsilon^2_{s2}}{s_2} - \frac{J_3\varepsilon^3_{s1}\varepsilon^1_{s3}}{s_3} + \frac{J_3\varepsilon^3_{s1}\varepsilon^1_{s2}}{s_2} - \frac{J_1\varepsilon^1_{s1}\varepsilon^2_{s3}}{s_3} + \frac{J_1\varepsilon^1_{s1}\varepsilon^2_{s2}}{s_2} \end{aligned} \tag{6}$$

To obtain the individual control coefficient expressions, the denominator (Equation (6)) is first factored out of the symbolic inverse of **E**. Next, the expressions are rearranged and simplified so that:

- elasticity terms are only *multiplied* by fluxes (i.e. not divided), and
- elasticities towards species in conserved moieties are always divided by the concentration of that *same species*.

As an example, we show the expressions thus generated for the control coefficients on flux J_3:

$$C^{J_3}_1 = \left(\frac{J_1\varepsilon^3_{s1}\varepsilon^2_{s3}}{s_3} - \frac{J_1\varepsilon^3_{s1}\varepsilon^2_{s2}}{s_2} \right) \Big/ \sum \tag{7}$$

$$C^{J_3}_2 = \left(\frac{J_1\varepsilon^3_{s1}\varepsilon^1_{s2}}{s_2} - \frac{J_1\varepsilon^3_{s1}\varepsilon^1_{s3}}{s_3} \right) \Big/ \sum \tag{8}$$

$$C^{J_3}_3 = \left(\frac{J_4\varepsilon^4_{s1}\varepsilon^2_{s3}}{s_3} - \frac{J_4\varepsilon^4_{s1}\varepsilon^2_{s2}}{s_2} - \frac{J_4\varepsilon^4_{s1}\varepsilon^1_{s3}}{s_3} + \frac{J_4\varepsilon^4_{s1}\varepsilon^1_{s2}}{s_2} - \frac{J_1\varepsilon^1_{s1}\varepsilon^2_{s3}}{s_3} + \frac{J_1\varepsilon^1_{s1}\varepsilon^2_{s2}}{s_2} \right) \Big/ \sum \tag{9}$$

$$C_4^{J_3} = \left(-\frac{J_4 \varepsilon_{s1}^3 \varepsilon_{s3}^2}{s_3} + \frac{J_4 \varepsilon_{s1}^3 \varepsilon_{s2}^2}{s_2} + \frac{J_4 \varepsilon_{s1}^3 \varepsilon_{s3}^1}{s_3} - \frac{J_4 \varepsilon_{s1}^3 \varepsilon_{s2}^1}{s_2} \right) \Big/ \sum \qquad (10)$$

with Σ as in Equation (6). The automatic rearrangement and factorization of the terms has the consequence that the generated control coefficient expressions such as in Equations (7) – (10) always end up in a standard format that is easily interpretable: every term is a *control pattern* [11], which can be visualized as a "chain of local effects" through the pathway.

Fermentation pathways in Lactococcus lactis

To illustrate the application of SymCA with a more realistic model, we have analysed the fermentation pathways of *Lactococcus lactis*. The pathway shown in Fig. 3 is a simplified version of the model reported by Hoefnagel *et al.* [15] with the branch to acetolactate omitted.

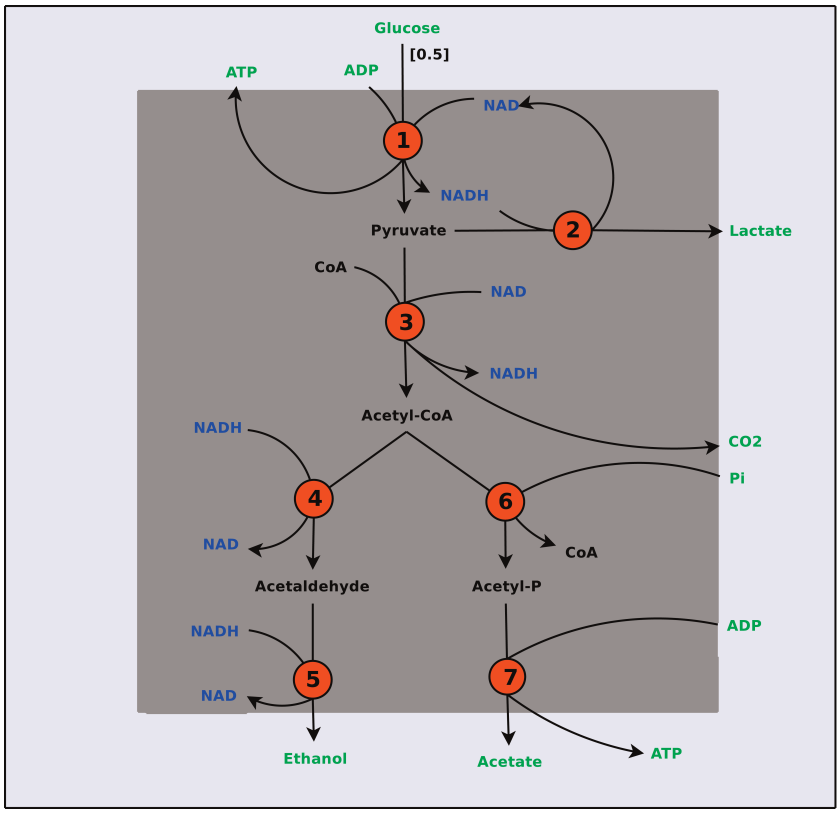


Figure 3. Fermentation pathways in *Lactococcus lactis.* Abbreviations for Equations (11) and (12): *PYR,* Pyruvate; *ACCOA,* Acetyl-CoA; *ACAL,* Acetaldehyde; *ACP,* Acetyl-P.

When processing this model with SymCA, the following expression for $C_1^{J_2}$ is generated, which we show by way of example:

$$C_1^{J_2} = \left(-\frac{J_1 J_5 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^5 \varepsilon_{PYR}^2}{COA\ NADH} + \frac{J_1 J_5 J_6 \varepsilon_{ACAL}^5 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^4 \varepsilon_{PYR}^2}{COA\ NADH} \right.$$
$$+ \frac{J_1 J_5 J_6 \varepsilon_{ACAL}^5 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^2 \varepsilon_{PYR}^3}{COA\ NADH} - \frac{J_1 J_5 J_6 \varepsilon_{ACAL}^5 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^3 \varepsilon_{PYR}^2}{COA\ NADH}$$
$$- \frac{J_1 J_3 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^2 \varepsilon_{PYR}^3}{COA\ NADH} + \frac{J_1 J_3 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^3 \varepsilon_{PYR}^2}{COA\ NADH} \qquad (11)$$
$$+ \frac{J_1 J_5 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACCOA}^6 \varepsilon_{ACP}^7 \varepsilon_{NADH}^5 \varepsilon_{PYR}^2}{ACCOA\ NADH} - \frac{J_1 J_5 J_6 \varepsilon_{ACAL}^5 \varepsilon_{ACCOA}^6 \varepsilon_{ACP}^7 \varepsilon_{NADH}^4 \varepsilon_{PYR}^2}{ACCOA\ NADH}$$
$$+ \ldots$$
$$\left. - \frac{J_1 J_3 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACCOA}^6 \varepsilon_{ACP}^7 \varepsilon_{NAD}^2 \varepsilon_{PYR}^3}{ACCOA\ NAD} + \frac{J_1 J_3 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACCOA}^6 \varepsilon_{ACP}^7 \varepsilon_{NAD}^3 \varepsilon_{PYR}^2}{ACCOA\ NAD} \right) \Big/ \sum$$

(24 terms)

with the denominator Σ given by:

$$\sum = \frac{J_3 J_5 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^5 \varepsilon_{PYR}^3}{COA\ NADH} + \frac{J_2 J_5 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^5 \varepsilon_{PYR}^2}{COA\ NADH}$$
$$- \frac{J_1 J_5 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^5 \varepsilon_{PYR}^1}{COA\ NADH} - \frac{J_3 J_5 J_6 \varepsilon_{ACAL}^5 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^4 \varepsilon_{PYR}^3}{COA\ NADH}$$
$$- \frac{J_2 J_5 J_6 \varepsilon_{ACAL}^5 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^4 \varepsilon_{PYR}^2}{COA\ NADH} + \frac{J_1 J_5 J_6 \varepsilon_{ACAL}^5 \varepsilon_{ACP}^7 \varepsilon_{COA}^6 \varepsilon_{NADH}^4 \varepsilon_{PYR}^1}{COA\ NADH} \qquad (12)$$
$$+ \ldots$$
$$+ \frac{J_2 J_3 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACCOA}^6 \varepsilon_{ACP}^7 \varepsilon_{NAD}^2 \varepsilon_{PYR}^3}{ACCOA\ NAD} - \frac{J_1 J_3 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACCOA}^6 \varepsilon_{ACP}^7 \varepsilon_{NAD}^1 \varepsilon_{PYR}^3}{ACCOA\ NAD}$$
$$- \frac{J_2 J_3 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACCOA}^6 \varepsilon_{ACP}^7 \varepsilon_{NAD}^3 \varepsilon_{PYR}^2}{ACCOA\ NAD} + \frac{J_1 J_3 J_6 \varepsilon_{ACAL}^4 \varepsilon_{ACCOA}^6 \varepsilon_{ACP}^7 \varepsilon_{NAD}^3 \varepsilon_{PYR}^1}{ACCOA\ NAD}$$

(56 terms)

The above expression is rather unwieldy and contains a very large number of terms, which makes it difficult to interpret. However, three aspects are noteworthy:

1. The expression is *mathematically correct,* providing an analytical solution for the control coefficient in terms of elasticities.
2. The expression is arranged and factorized so that all terms are similar and conform to the *standard format* described above.
3. The expression can be *simplified* subject to certain assumptions about the pathway.

To illustrate the last point, assume, for example, that $\varepsilon_{PYR}^1 = \varepsilon_{PYR}^2 = 0$ (i.e. that both reactions 1 and 2 are insensitive to changes in the pyruvate concentration). The expression for $C_1^{J_2}$ then simplifies to (automatically calculated by SymCA):

$$C_1^{J_2} = \frac{-J_1J_5\varepsilon_{ACAL}^5\varepsilon_{NAD}^2 + J_1J_3\varepsilon_{ACAL}^4\varepsilon_{NAD}^2 + J_1J_5\varepsilon_{ACAL}^5\varepsilon_{NADH}^2 - J_1J_3\varepsilon_{ACAL}^4\varepsilon_{NADH}^2}{J_3J_5\varepsilon_{ACAL}^4\varepsilon_{NADH}^5 - J_3J_5\varepsilon_{ACAL}^5\varepsilon_{NADH}^4 - J_2J_5\varepsilon_{ACAL}^5\varepsilon_{NADH}^2 + J_1J_5\varepsilon_{ACAL}^5\varepsilon_{NADH}^1}$$
$$+J_2J_3\varepsilon_{ACAL}^4\varepsilon_{NADH}^2 - J_1J_3\varepsilon_{ACAL}^4\varepsilon_{NADH}^1 - J_3J_5\varepsilon_{ACAL}^4\varepsilon_{NAD}^5 + J_3J_5\varepsilon_{ACAL}^5\varepsilon_{NAD}^4 \quad (13)$$
$$+J_2J_5\varepsilon_{ACAL}^5\varepsilon_{NAD}^2 - J_1J_5\varepsilon_{ACAL}^5\varepsilon_{NAD}^1 - J_2J_3\varepsilon_{ACAL}^4\varepsilon_{NAD}^2 + J_1J_3\varepsilon_{ACAL}^4\varepsilon_{NAD}^1$$

The number of terms has been reduced from 24 to 4 in the numerator and from 56 to 12 in the denominator. Clearly, Equation (13) is of a more manageable size so that direct algebraic analysis and interpretation is possible.

Discussion and Conclusion

The framework of MCA has been a driving force in dispelling the notion of the ''rate-limiting step'' in metabolic pathways, as was elegantly illustrated by the pioneering experimental work of Groen and co-workers [16] on the control of mitochondrial oxidative phosphorylation. However, MCA goes beyond the mere quantification of control coefficients – in our view, the real power of MCA lies in its relation of control coefficients to elasticities, thus allowing us to infer systemic properties from the characteristics of the isolated system components.

In this paper we have described SymCA, a software that implements the symbolic matrix inversion of MCA in a generalized way and generates analytical expressions for the control coefficients of a pathway in terms of the elasticities. This has a number of uses:

- Computational analysis of biochemical pathways is becoming increasingly important in the burgeoning field of computational systems biology. An inspection of on-line model databases such as JWS Online ([17], http://jjj.biochem.sun.ac.za/) or BioModels ([18], http://www.ebi.ac.uk/biomodels) reveals that the number of available models grows monthly, but that they are also increasing in size and complexity. Here, MCA can become an increasingly important analysis tool by dissecting where the control in a pathway lies and identifying the factors that determine this control. For example, the expressions generated with SymCA can be used to identify key elasticities that are responsible for a particularly large (or small) value of a certain control coefficient. Furthermore, the analysis can easily be extended to parameter-response coefficients (see *Additional features* above), thus addressing the question of which parameters in a kinetic model have the largest effect on a particular observed behaviour, and how this effect is transmitted.
- *Fermentation pathways in Lactococcus lactis* (above), shows that the MCA expressions quickly become unwieldy once the model size grows beyond a few reactions. In such cases, assumptions can be introduced to simplify the expres-

sions, as illustrated above. Very often, it is known that a particular reaction is saturated with substrate or insensitive to product under cellular conditions (allowing the elasticity to be set to zero) or is operating in the first-order range because the substrate concentration is lower than the K_m (allowing the elasticity to be set to one). Moreover, certain elasticities can be set numerically to vary within bounds to explore how the control distribution in the network would be affected.

- The development of kinetic models is often hampered by the fact that not all of the kinetic parameters are known. The significance of symbolic MCA is that it is valid in general, and as such does not depend on the availability of particular parameter values. All that is required is the stoichiometry and mapping of the allosteric modifier interactions. Thus, general conclusions about the control structure of a pathway may be drawn even if not all the kinetic parameter details are known.
- As was already pointed out by Hofmeyr almost 20 years ago [11], the individual terms of a control coefficient expression can be visualized on the network as a "control pattern" that describes a "chain of local effects" corresponding to a particular route of regulation. Symbolic control analysis can thus help identify such routes of regulation in a complex network and quantify their relative importance (e. g. comparing feedback inhibition along the main chain of a pathway *vs.* an allosteric feedback loop).

It should be mentioned that programmatic symbolic control analysis is not new. The problem has been tackled by Thomas and Fell with the MetaCon [12] computer program. However, their approach is not completely general in the sense that the analysis of branched pathways always requires the manual selection of a reference flux before the expressions can be generated and selection of different reference fluxes leads to different expressions, whereas the matrix method on which SymCA is based does not have this limitation. Moreover, the integration of SymCA data structures within PySCeS means that it is easy to substitute some or all of the elasticities with numerical values, allowing for further analysis, simplification or validation of the expressions.

In conclusion, as the field of computational systems biology grows it can be anticipated that the complexity of models will increase, approaching the level of complexity of the modelled systems themselves. Analysis tools will become ever more essential for making sense of huge amounts of model data. Symbolic control analysis is one such tool the SymCA software presented here facilitates and contributes to this analysis.

Acknowledgements

This work was supported by the National Bioinformatics Network (South Africa).

References

[1] Fell, D.A. *Understanding the Control of Metabolism*; Portland Press: London, 1996.

[2] Heinrich, R., Schuster, S. *The Regulation of Cellular Systems;* Chapman & Hall, New York, 1996.

[3] Kacser, H., Burns, J.A. (1973) The control of flux. *Symp. Soc. Exp. Biol.* **27**:65 – 104.

[4] Heinrich, R., Rapoport, T.A. (1974) A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur. J. Biochem*. **42**:89 – 95.

[5] Westerhoff, H.V., Chen, Y.-D. (1984) How do enzyme activities control metabolite concentrations? An additional theorem in the theory of metabolic control. *Eur. J. Biochem.* **142**:425 – 430.

[6] Sauro, H.M., Small, J.R., Fell, D.A. (1987) Metabolic control and its analysis: extensions to the theory and matrix method. *Eur. J. Biochem.* **165**:215 – 221.

[7] Small, J.R., Fell, D.A. (1989) The matrix method of metabolic control analysis: Its validity for complex pathway structures. *J. Theor. Biol.* **136**:181 – 197.

[8] Westerhoff, H.V., Hofmeyr, J.-H.S., Kholodenko, B.N. (1994) Getting to the inside of cells using metabolic control analysis. *Biophys. Chem.* **50**:273 – 283.

[9] Hofmeyr, J.-H.S., Cornish-Bowden, A. (1996) Co-response analysis: A new experimental strategy for metabolic control analysis. *J. Theor. Biol.* **182**:371 – 380.

[10] Hofmeyr, J.-H.S. Metabolic control analysis in a nutshell. In *Proceedings of the 2nd International Conference on Systems Biology*, 2001, (Yi, T.-M., Hucka, M., Morohashi, M. and Kitano, H., Eds) Omnipress, Madison, WI, 2001; pp. 291 – 300.

[11] Hofmeyr, J.-H.S. (1989) Control-pattern analysis of metabolic pathways. Flux and concentration control in linear pathways. *Eur. J. Biochem.* **186**:343 – 354.

[12] Thomas, S., Fell, D.A. (1993) A computer program for the algebraic determination of control coefficients in metabolic control analysis. *Biochem. J.* **292**:351 – 360.

[13] Olivier, B.G., Rohwer, J.M., Hofmeyr, J.-H.S. (2005) Modelling cellular systems with PySCeS. *Bioinformatics* **21**:560 – 561.

[14] Lamport, L. *LaTeX: A Document Preparation System: User's guide and reference,* 2nd ed; Addison-Wesley Professional, Reading, MA, 1994.

[15] Hoefnagel, M.H.N., Starrenburg, M.J.C., Martens, D.E., Hugenholtz, J., Kleerebezem, M., Swam, I.I.V., Bongers, R., Westerhoff, H.V., Snoep, J.L. (2002) Metabolic engineering of lactic acid bacteria, the combined approach: kinetic modelling, metabolic control and experimental analysis. *Microbiology* **148**:1003 – 1013.

[16] Groen, A.K., Wanders, R. J.A., Westerhoff, H.V., van der Meer, R., Tager, J.M. (1982) Quantification of the contribution of various steps to the control of mitochondrial respiration. *J. Biol. Chem.* **257**: 2754 – 2757.

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

[18] LeNovère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J.L., Hucka, M. (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.* **34**:D689 –D691.

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

JWS Online: a Web-accessible Model Database, Simulator and Research Tool

Jacky L. Snoep[1,2,3*], Carel van Gend[1], Riann Conradie[1], Franco du Preez[1], Gerald Penkler[1] and Cor Stoof[2]

[1]Triple-J group for Molecular Cell Physiology, Department of Biochemistry, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa;

[2]Cellular BioInformatics, Vrije Universiteit, De Boelelaan 1087, NL-1081 HV Amsterdam, The Netherlands;

[3]Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, Manchester University, 131 Princess Street, Manchester M1 7ND, U.K.

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

Received: 14th April 2008 / Published: 20th August 2008

Abstract

In previous contributions to the ESCEC proceedings we focused on the functionality of JWS Online and we made a comparison between JWS Online and other model database initiatives. In the current chapter an update is given on new developments for JWS Online and we illustrate the functionality of JWS Online web services in workflows.

Introduction

A number of Systems Biology tools have been made available in the JWS Online project [1]: 1) a database of curated kinetic models, 2) an easy to use, web-based simulator for those models and 3) a tool to help scientific journals with the reviewing of manuscripts that contain kinetic models. The project was initiated because there was a need for a model repository; most model descriptions in the literature are incomplete and for larger models it is not practical to re-code the model from a manuscript, even if a complete description were available. In 2000 we started building a repository for kinetic models of biological systems. The models are accessible via a web-based interface that enables the users to simulate the

models in a browser. Although the functionality of such a simulator is necessarily limited to queries that are not too computer intensive, it gives easy access for a first interaction to the model and, finally, it is a great tool for teaching.

For more elaborate or customer specific simulations the models can be downloaded in SBML and PySCeS format. The models can be accessed via three mirror sites: http://jjj.biochem.sun.ac.za/ (Stellenbosch University, South Africa); http://jjj.bio.vu.nl/ (Vrije Universiteit Amsterdam, the Netherlands) and http://jjj.mib.ac.uk/ (Manchester University, UK).

Currently (April 2008), 85 curated models can be accessed via the JWS Online model database. In 2005 we started collaborating with the Biomodels database, and we have made most of the Biomodels available via the JWS simulator (http://jjj.biochem.sun.ac.za/biomodels and http://jjj.bio.vu.nl/biomodels).

When we first recognized that kinetic models are very poorly described in the literature we also contacted a number of journals to point this out and we offered to help in the reviewing of manuscripts that contain kinetic models by making these models available on a secure site. We are now collaborating with four scientific journals: FEBS Journal, Microbiology, IET Systems Biology and Metabolomics for each of which separate web sites have been set-up to give reviewers access, via a secure site, to kinetic models described in submitted manuscripts.

JWS Online is actively used in a number of research projects, the Silicon Cell project (SiC) [2], the Yeast Systems Biology Network (YSBN) [3], and a number of new initiatives such as Systems Biology for Micro Organisms (SysMO) [4], and a seventh framework EU program on Systems Biology of eukaryotic unicellular organisms (UniCellSys) [5]. From the collaborations in these research projects it has become clear that a number of improvements needed to be made to the functionality of JWS Online to make it a good research tool, in addition to a service and educational tool. In this chapter we describe the functionality of the JWS Online simulator, with an emphasis on the newly added functions, but our focus is on a completely new functionality: web services. The importance of web services in research projects will be illustrated in an example workflow.

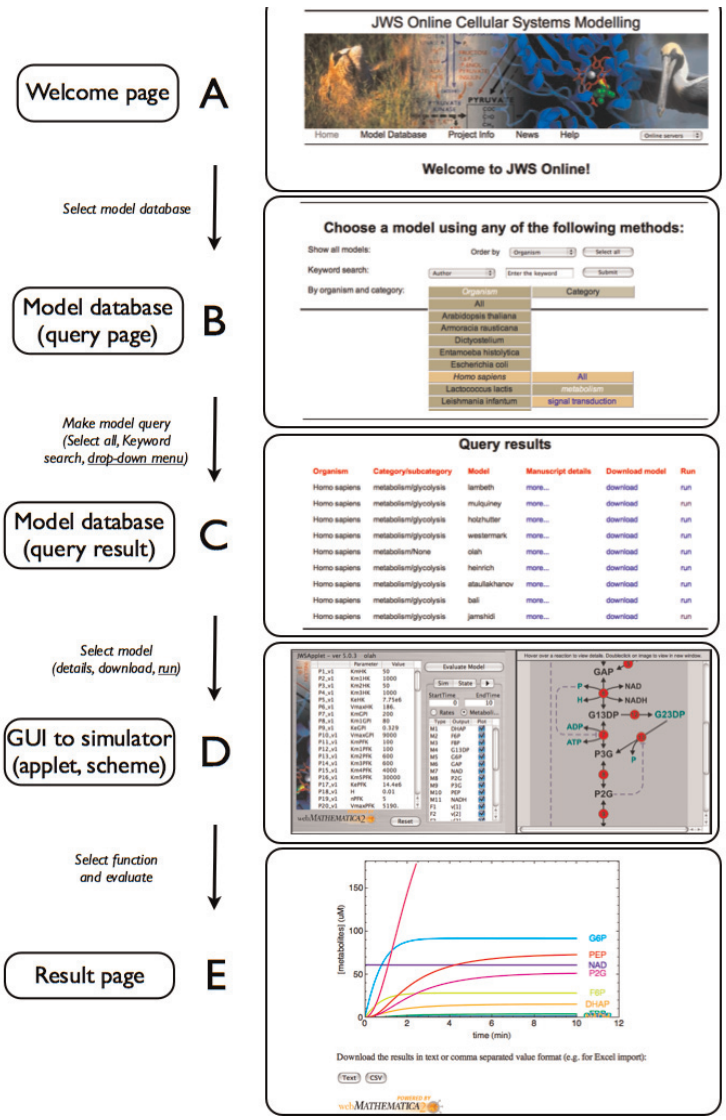


Figure 1. A typical example of a JWS Online simulation session.

(A) the user enters JWS Online via the home page, (http://jjj.biochem.sun.ac.za); (B) upon selecting the "model database" link, the database query page is loaded. The user selects a set of models via the query methods (drop down menu method is shown in the figure); (C) the result of the query (*Homo sapiens*, metabolism), is shown in the query result page (the user selects to run the olah model); (D) the olah model applet is loaded (a time simulation is evaluated); (E) the result page for the time simulation.

The kinetic models of JWS Online are stored in a PostgresQL database and the selection of a specific model and queries of the models can be made via menu selections (running Python scripts in the background). A typical sequence of actions leading to a simulation results are shown as a flow chart in Fig 1: on the welcome page of JWS Online (e.g. [6], Fig. 1A), a choice must be made between, ''Model Database'', ''Project Info'', ''News'', and ''Help''. In Fig. 1A, the ''Model Database'' option is selected and this leads to the Model database query page (Fig. 1B), where three query methods are available for model selections: select all, select via a key-word search (''author'', ''title'', ''journal'' (referring to respectively the first author on the manuscript in which the model is described, the title and journal name of that manuscript), ''organism'', ''category'' (e.g. metabolism), ''subcategory'' (e.g. glycolysis), ''model type'' (e.g. demonstration)), or select via a drop-down menu on the basis of ''organism'' and ''category''. In Fig. 1B the drop down menu is used to select ''*Homo sapiens*'' as organism and ''metabolism'' as category, leading to Fig. 1C, the query result page. On the query result page links to ''manuscript details'' and ''model downloads'' (in SBML and PySCeS format) can be selected in addition to the ''run'' option, which leads to the user interface of the JWS Online simulator. This interface consists of an applet and a metabolic scheme, Fig. 1D. The applet is used to change parameter values, and to select the simulation query (see next section), and executes the evaluation, calculated by the Mathematica® [7] kernel on the JWS Online server. The result is shown as a pop-up window in the browser (Fig. 1E, for an example of a time simulation). Notice the ''text'' and ''csv'' buttons at the bottom of the result window. This is a new option that allows the user to save the simulation result in either of the two formats. The ''csv'' format can be directly loaded into spreadsheet programs such as Microsoft Excel®.

JWS Online: Simulator Functionality

The original functionality of the JWS Online simulator consisted of: 1) time simulations (time integration of the models and options for plotting metabolites or rates; see Fig. 2A for the interface for this function); 2) steady state analyses, (steady state solution of the model, structural analyses; N matrix, K matrix and L matrix (see e.g. [8]), and analysis in terms of Jacobian matrix and eigenvalues (Fig. 2B and 3) metabolic control analysis, control and elasticity coefficients for the reactions in the model (Fig. 2C).

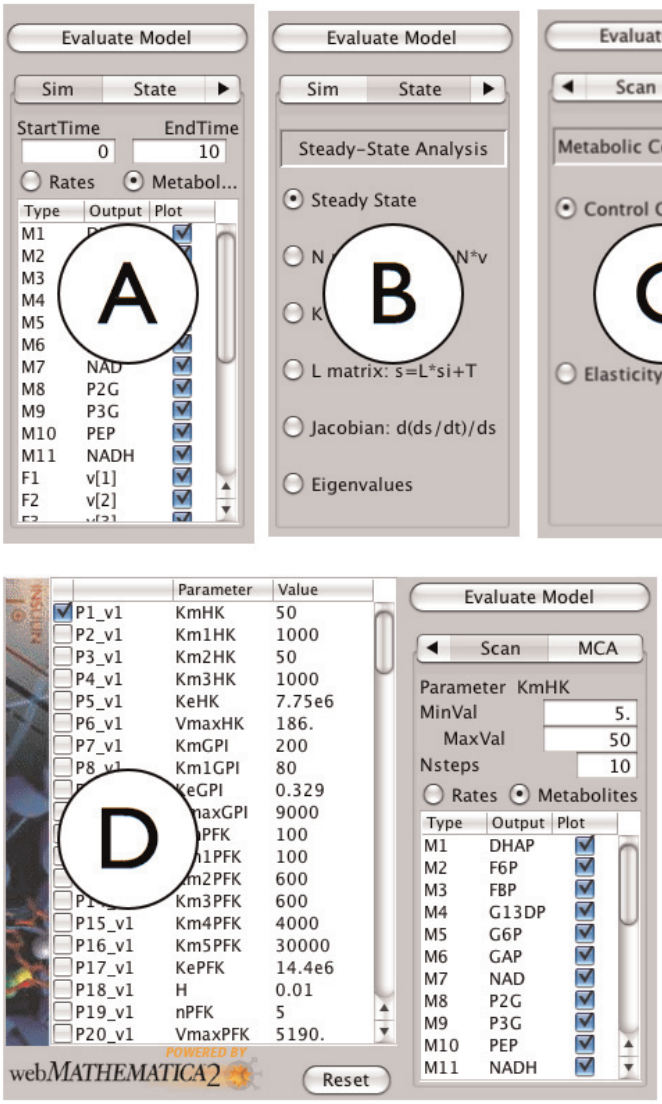


Figure 2. The control panels for the different simulation options.

(A) the time simulation panel, with options to set the time period, and select the metabolite or rates options; (B) the state panel, with options to calculate the steady state solution, and some structural and stability analysis functions; (C) the metabolic control analysis option with a selection of elasticity or control coefficients, (D) the scan options, for which the parameter table from which the scanning parameter must be selected, is also shown. For the scanning option the minimal and maximal value for the scanning parameter and the number of scanning points must be indicated in the control panel.

We have recently added a fourth functionality to vary a model parameter and plot the effect on the steady state values of the selected model variables. The interface to this functionality is shown in Fig. 2D, where in the parameter table on the left the scanning parameter can be selected by checking the box in front of it. In the example shown in Fig. 2D the parameter ''P1_v1, KmHk'' is selected (by default the first parameter in the list is selected). The name of the scanning parameter is also indicated in the control panel of the scanning function, together with the options, ''MinVal'', ''MaxVal'' and ''Nsteps'', specifying respectively the lower and upper value for the parameter to be scanned and the number of scans that will be made. In addition the user needs to chose whether steady state ''Metabolite'' concentrations or ''Rates'' must be plotted by selecting the respective radio button (a further differentiation on which fluxes and metabolites must be plotted can be made, in the table by (de)selecting the variables of choice).

JWS Online: A Research Tool

Up to now, most of the functionality that we discussed makes JWS Online a good service tool; a user can access models and make queries and download the models. A serious limitation in the functionality of JWS Online was the absence of a mechanism to save the simulation results. This has now been added (see JWS Online: An Overview), and the file that is saved includes the parameter values that were used for the simulation such that a complete record of the simulation is available to the user.

How can the functionality of JWS Online as a research tool further be improved?

We thought that the possibility to use simulation results in a workflow would be a significant improvement, certainly, if this could be done in an environment in which an analysis of the result can be made and used in a subsequent query. Web services are ideal to be used in such workflows and they will become an important tool for Systems Biology projects, because they make it possible to link databases and other information sources and to automate standard analyses methods. We here first introduce web services and show how they can be used (implemented in Taverna) to connect different database initiatives. Subsequently we show an example workflow in Mathematica® where JWS Online web services are used.

Web Services

Web services are a means of providing remote access to program functionality over a network, commonly the World Wide Web. This enables a client to use the services of programs running on servers situated in geographically diverse locations. A simple and standardised interaction specification means that each service may be accessed using the same basic protocol. Clients send service requests as XML messages encoded according to the Simple Object Access Protocol (SOAP) standard. These are usually transmitted using the ubiquitous HTTP protocol, and so can be received by any web-service enabled web server.

The server processes the request and returns the results, also over HTTP, as SOAP encoded XML. It is usual for servers to make available a Web Service Description Language (WSDL) file describing the web services available on the server and details of the request and response messages used to access them.

Since no knowledge is required of the underlying implementation of a web service, users may concentrate their attention only on the content of these services. Similarly, groups or organisations which have developed a tool that may be useful to the community may easily make this available, without worrying about end-users needing specialised tools or knowledge of protocols to utilise this. Access to a number of databases relevant to systems biology has been provided in this way, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [9], BRENDA [10], Biomodels [11], and SABIO-RK [12], a database containing information on biochemical reaction, reaction kinetics and parameters, and annotation of these detailing the experimental conditions under which they were measured.

As web services follow an open standard, anyone may write a client to access a particular service. Nevertheless, it is important that the SOAP encoded XML messages have the correct fields, are transmitted correctly, and that the response messages are correctly dissected. A number of platforms provide extensions which simplify the creation and transmission of these messages. Web browsers such as Firefox allow web service requests to be entered in the location bar, with the response displayed in the browser main window. Platforms such as Mathematica® similarly include a means of accessing web services, and additionally the results may be processed using Mathematica®'s powerful symbolic manipulation tools (see below).

The software package Taverna [13, 14] workbench is a tool that allows the design and execution of workflows involving any number of web services. A graphical user interface allows the user to drag the required method of a service to the workflow, and link its inputs and outputs to those of other services. Taverna then allows the user to run the workflow, displaying the results graphically or saving them to a file. The scheduling, creation and transmission of the SOAP messages is handled automatically by Taverna, as is the reception and unpacking of the web service response.

Taverna Workflow Example

Here, we describe a Taverna workflow that retrieves a list of EC numbers for the enzymes associated with a specified pathway from the KEGG web service, and then uses the SABIO-RK web service to retrieve the names of these enzymes. Although this is a trivial example, it illustrates the use of Taverna to retrieve information from multiple web services, as well as the use of Taverna's built in local transformation utilities and the Java bean shell.

The pathway takes as initial input the KEGG identifier for a particular pathway (we have used the glycolysis pathway of *Saccharomyces cerevisiae*), and queries the `get_enzymes_by_pathway` method of the KEGG web service for the Enzyme Classification (EC) numbers of all enzymes associated with this pathway (Fig. 3A). We then use Taverna's list selection local service (here labelled `Select_enzymes_from_enzyme_list`) to select a subset of the list. This takes, in addition to the list of enzyme EC numbers, a start and end index, defining the subset.

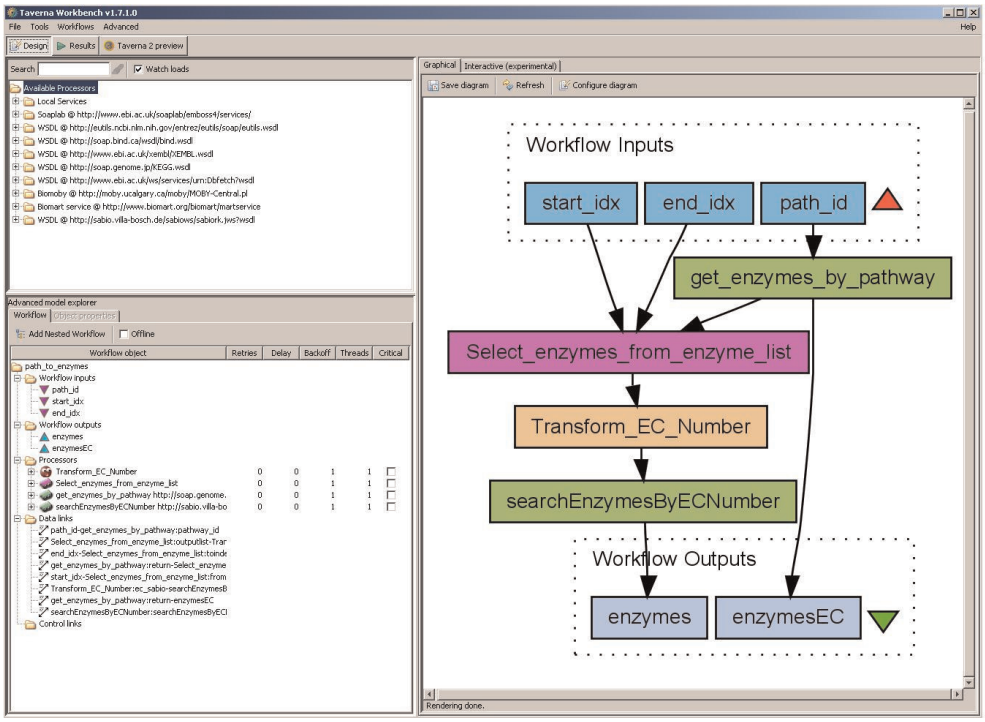


Figure 3A. Screenshots of a Taverna workflow example.

The Taverna workbench interface, with on the left the available processors and the workflow explorer, and on the right the graphical representation of the workflow.

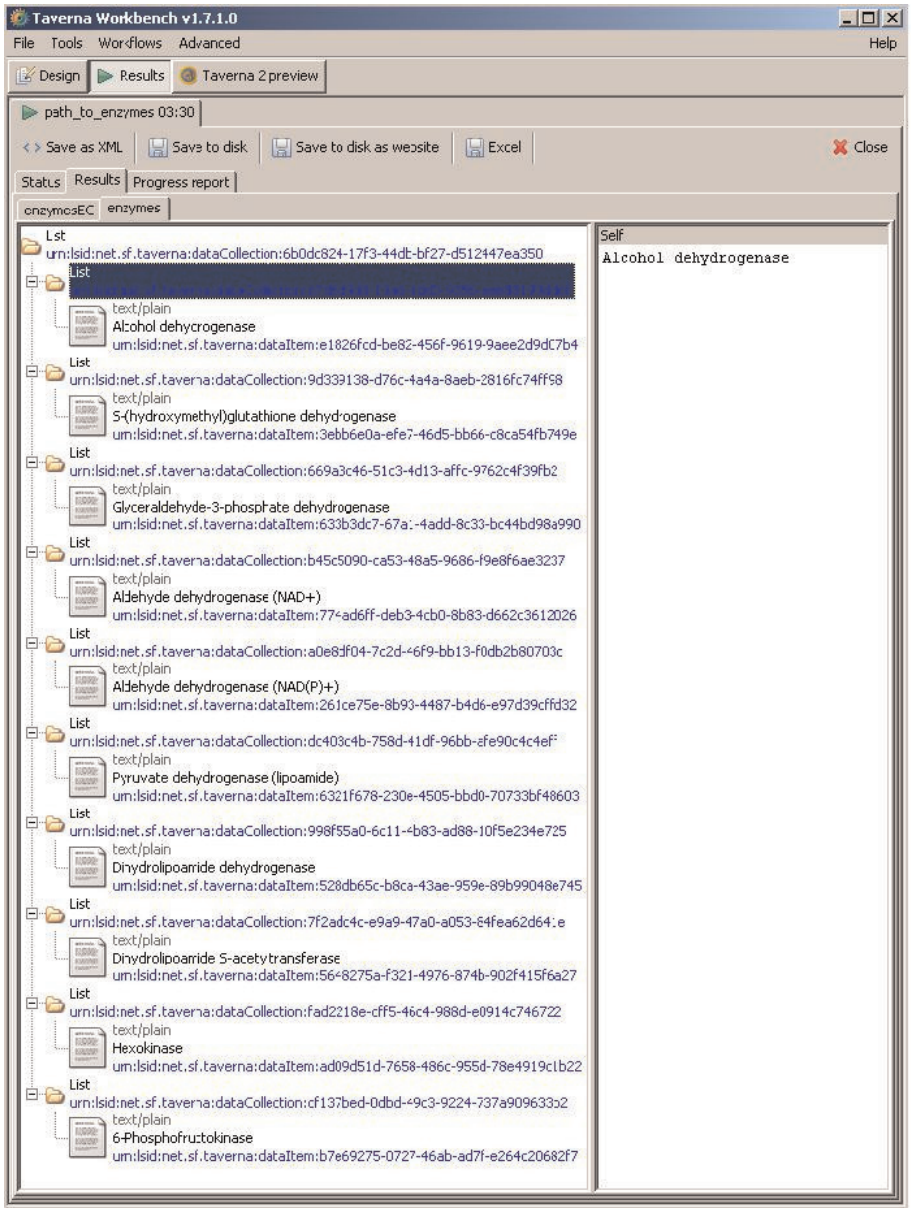


Figure 3B. Screenshots of a Taverna workflow example.

The Taverna results output for the query. See the text for more detail on the workflow.

We will then request from the SABIO-RK web service the names of all of these enzymes, but since the EC number format output by KEGG is somewhat different from the format used by SABIO-RK, we use the Java beanshell supplied with Taverna to transform these

(labelled `Transform_EC_Number`). Having done this the `searchEnzymesByEC-Number` method of the SABIO_RK web service returns the list of enzyme names (labelled `enzymes` in the workflow diagram). A feature of Taverna is that the output of a particular processor may be directed to multiple sinks; in this case the `get_enzymes_by_path-way` processor also sends its output to the `enzymesEC` output, so that we may view the results of this step.

When the pathway is started, a window appears allowing the user to input the various parameters required, and on completion the results (with separate tabs for separate outputs) as well as the workflow status and progress report are displayed in the results pane of Taverna (Fig. 3B).

JWS Online Web Services

JWS Online now offers a web service interface to the underlying functionality. This allows users to include JWS Online in workflows. The JWS Online web service is implemented in Java using Apache Axis which enables rapid development of functionality and automates much of the otherwise tedious coding. Each web service function is implemented as a Java method which connects to the JWS Online database through the JDBC Postgresql connector, or the Mathematica® kernel through JLink. The latter means that the existing JWS Online functionality, encoded in Mathematica® model files, can easily be accessed by the web service, and also makes the provision of new functionality straightforward. Functions exist which allow the database to be queried for a list of all models, or only models of a particular type or for specific organisms.

Mathematica® Workflow Example

The following workflow will analyse the stoichiometric matrix of all the models in the JWS Online database and calculates the number of reactions that each of the metabolites in the model is connected to. Subsequently a plot is generated showing the chance of having a number of connections as a function of the number of connections is generated. These analysis are standard in network analysis to check what type of network structure the system has; is there a random distribution of the number of connections or does the network show a scale free structure (see e. g. [15]).

Installs the web services and queries the wsdl file:

```
InstallService[''http://jjj.biochem.sun.ac.za/axis/ser-
vices/QueryJWS?wsdl'']
```

Returns the web services that are currently available for JWS Online:

```
{getRates, getAllModels, getAllBiomodels, getAllBiomodel-
sIds, getModelsByOrganism, getModelsByCategory, getModelIn-
fo, getNmat, getKmat, getLmat, getSteadyStateTable, getTime-
course, getJacob, getEigenv, getCmat, getEmat, getRateEqua-
tions, getRateEquationFormulae, getExtVar, hasFunction}
```

Retrieve the current model names from JWS Online and assign the names to the variable JWSmodels:

```
JWSmodels = getAllModels[]
```

The output of the query (a list with all model names) is not shown.

Define a function "queryFunction" that checks whether a given function is present in a model:

```
queryFunction[modelname_,  function_]:=(response=Invoke-
ServiceOperation[hasFunction,  ToString[modelname],  ToS-
tring[function]]; response[[2, 3, 1, 3, 1, 3, 1, 3, 1]])
```

Checks which of the JWS Online models have the stoichiometric analysis function:

```
queryJWS = queryFunction[#, Nmat] &/@ JWSmodels
```

Define a function to retrieves the stoichiometric matrix and returns a list of the variables, stoichiometry matrix and the rate names:

```
Nmat[modelname_]:=(response=InvokeServiceOperation[getN-
mat, ToString[modelname]]; rates=Table[response[[2, 3, 1, 3,
2, 3, 4, 3, i, 3, 1]], {i, Length[response[[2, 3, 1, 3, 2, 3, 4,
3]]]}]; stochmatraw=response[[2, 3, 1, 3, 2, 3, 5, 3]]; smrow-
s=Length[stochmatraw];      smcols=Length[stochmatraw[[1,
3]]]; stochmat=Table[stochmatraw[[i, 3, j, 3, 1]], {i,
smrows}, {j, smcols}]; varsraw=response[[2, 3, 1, 3, 2, 3, 6,
3]]; numvars=Length[varsraw]; vars=Table[varsraw[[i, 3,
1]], {i, numvars}]; {vars, stochmat, rates})
```

An example of the use of the N matrix function for the Teusink model:

```
nmat = Nmat[teusink]
```

And its output:

```
{{(ACE d)/dt, (BPG d)/dt, (d F16P)/dt, (d F6P)/dt, (d G6P)/dt,
(d GLCi)/dt, (d NAD)/dt, (d NADH)/dt, (d P2G)/dt, (d P3G)/dt, (d
PEP)/dt, (d Prb)/dt, (d PYR)/dt, (d TRIO)/dt},

{{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, -2, 0, -1, 0, 0}, {0, 0, 0, 0, 0,
0, 1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0}, {0, 0, 0, 0, 1, -1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0}, {0, 1, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0},
{1, -1, -1, -2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}, {-1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0}, {0, 0, 0, 0, 0, 0, -1, 0, 0, 0,
```

```
0,0,-3,0,1,1,0},{0,0,0,0,0,0,1,0,0,0,0,0,3,0,-1,-
1,0},{0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0,0,0},{0,0,0,
0,0,0,0,1,-1,0,0,0,0,0,0,0,0},{0,0,0,0,0,0,0,0,0,
1,-1,0,0,0,0,0,0},{-1,0,-1,-1,-1,0,0,1,0,0,1,0,0,
0,0,0,-1},{0,0,0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0},{0,
0,0,0,0,2,-1,0,0,0,0,0,0,0,0,-1,0}},

{vGLK,vPGI,vGLYCO,vTreha,vPFK,vALD,vGAPDH,vPGK,vPGM,
vENO,vPYK,vPDC,

vSUC,vGLT,vADH,vG3PDH,vATP}}
```

Defining a function to determine the number of reactions that each of the metabolites in a given model is connected to, note that this function calls the Nmat[] function defined above:

```
degreedistribution[modelname_]:=(nmat=ToExpression[N-
mat[modelname]];  Table[Length[Cases[nmat[[2,  i]],  Ex-
cept[0]]],{i,Length[nmat[[2]]]}])
```

And the application of the function to all the models:

```
degree=(degreedistribution[#]&)/@querymodels
```

Now we make a bin count for each of the number of connections in all the models:

```
dataDegreeJWS=N[BinCounts[Flatten[Join[degree]], {1, 40,
1}]]
```

And we determine the total number of connections in all the models:

```
totalNodes=Total[dataDegreeJWS]
```

Finally we can plot the chance of having a number of connections P(k), against the number of connections:

```
ListPlot[dataDegreeJWS/totalNodes,Frame->True,FrameLabel
-> {''k'', ''P(k)''}, PlotStyle -> {Red, PointSize[0.02]},
BaseStyle -> {FontSize ->20}, PlotRange -> {{0, 20}, {0,
0.5}},ImageSize->500]
```

The resulting plot is shown in Fig. 4A, together with a plot in double logarithmic space in Fig. 4B. From the plot in Fig. 4A it can be depicted that the connections do not show a normal distribution which would be indicative for a random connected network, rather the connections appear to have more of a scale free type of structure (Fig. 4B), which would result in a linear relation in double log space [15].

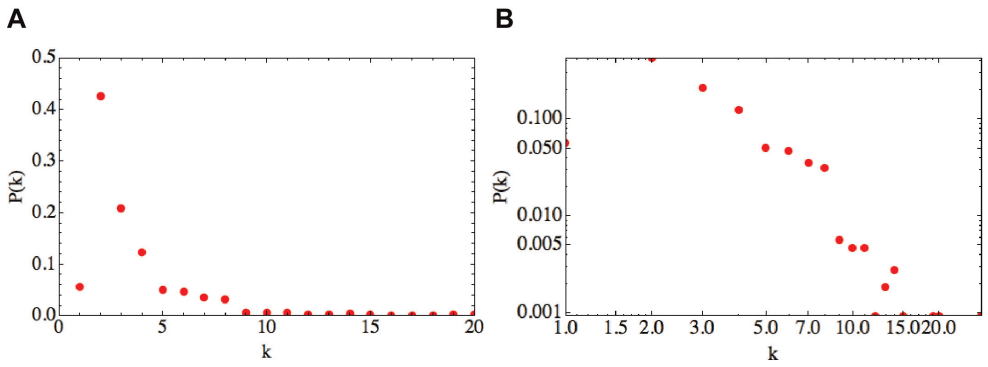


Figure 4. The results of a Mathematica® workflow example.

(A) the distribution of the chance for a number of connections for the metabolites in the JWS Online models is indicated. (B) the same distribution but now plotted in double logarithmic space. See the text for details on the workflow; 82 models were analysed with a total number of connections of 1081.

Clearly, this is a very specific question that is addressed, but the example serves to show that it is very simple to integrate workflows in an environment that is web service enabled. The standardisation of the queries and responses make it possible to automate requests, even to multiple databases and to make updates, for instance when new models are added to the database, or when a new DNA sequence needs to be analyzed via a certain workflow, trivial.

Conclusions

In this contribution we have highlighted some new developments in the JWS Online project. In addition to further extensions of JWS Online simulations functionality, such as the implementation of a scanning option and the saving options for simulation results, we have focused on web services. Web services form a very useful tool, by which one can process standardized requests, which can easily be automated and extended to large numbers of models, as is illustrated in two simple workflow examples in this chapter.

References

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

[2] Silicon Cell Project, SiC, http://www.siliconcell.net

[3] Yeast Systems Biology Network, YSBN, http://www.ysbn.org

[4] Systems Biology for Micro Organisms, SysMO, http://www.sysmo.net

[5] 7th Framework EU program on Systems Biology of eukaryotic unicellular organisms, UniCellSys, http://www.unicellsys.eu

[6] JWS Online, http://jjj.biochem.sun.ac.za

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

[8] Hofmeyr, J-H. S. (2001) Metabolic control analysis in a nutshell. In: *2nd International Conference on Systems Biology*, (eds. T.-M. Yi, M. Hucka, M. Morohashi and Kitano, H.) Omnipress, Madison, pp. 291 – 300.

[9] Kyoto Encyclopedia of Genes and Genomes, KEGG, http://www.genome.jp/kegg

[10] Braunschweig Enzyme Database, BRENDA, http://www.brenda-enzymes.info/

[11] Biomodels, http://www.ebi.ac.ul/biomodels/

[12] SABIO-RK, http://sabio.villa-bosch.de/SABIORK/

[13] Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A. and Li, P. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**:3045 – 3054.

[14] Li, P., Oinn, T., Soiland, S. and Kell, D.B. (2008) Automated manipulation of systems biology models using libSBML within Taverna workflows. *Bioinformatics* **24**:287 – 289

[15] Barabasi, A.-L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**:101 – 113.

Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

Validity and Combination of Biochemical Models

Wolfram Liebermeister

Computational Systems Biology, Max-Planck-Institut für molekulare Genetik, Ihnestraße 63 – 73, 14195 Berlin, Germany

E-Mail: lieberme@molgen.mpg.de

Received: 14th April 2008 / Published: 20th August 2008

Abstract

The merging of mathematical models (either manually or assisted by computer programs) is an important requisite for creating large mathematical models of cells. A kinetic model describes biochemical quantities such as concentrations and reaction rates by explicit differential and algebraic equations. We can regard it as a list of model statements, each comprising a biochemical quantity (e. g. a substance concentration), the corresponding mathematical object (e. g. a variable or parameter), and a mathematical equation that makes it possible to compute its numerical value. When two such models are merged, typical conflicts have to be detected and resolved: (i) incompatible names or identifiers; (ii) incompatible physical units; (iii) duplicate elements with contradicting assignments; (iv) conflicting (“semantically dependent”) quantities; (v) cyclic dependencies between model equations. To define and judge whether merging algorithms are trustworthy, we need formal criteria for the validity of models; such criteria can be classified into the categories “syntax”, “computation”, “biochemical semantics”, “physical laws and empirical knowledge” and “model relevance”.

Merging of Biochemical Models

Living cells can be described by mathematical models in order to test biological hypotheses by computer simulations and mathematical analysis. The mathematical elements in biochemical models (e. g. variables and equation terms) refer to chemical substances and processes such as transport, binding and reactions. In publications, models are described verbally and by mathematical formulae, but researchers also publish them in computer-readable formats like SBML [1] (systems biology markup language) and models become increasingly available in databases [2, 3]. One intention behind this is that models should be reusable; model reuse is further facilitated if standards [4] are respected – as put forward in the MIRIAM proposal [5] (“minimal information requested in the annotation of biochemical models”).

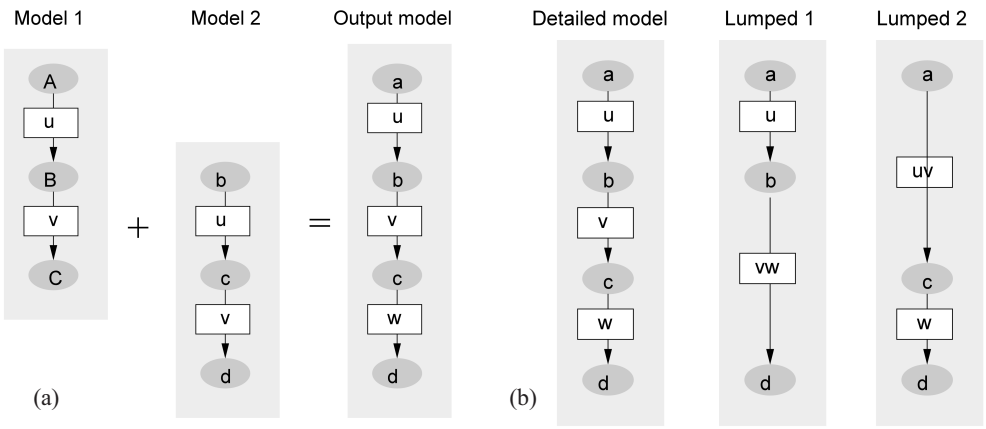


Figure 1. Merging of structural models. (a) Structural models of two metabolic pathways are added by taking the set union of all model elements (metabolites and reactions shown as ellipses and boxes, respectively). The graph topology is coded by stoichiometric coeffcients, contained as additional information in the reaction elements (not shown). The names of model elements can differ from model to model, so elements must be compared by their annotations (not shown) and a consistent set of names has to be chosen for the output model. (b) Models can contain lumped reactions (e. g. the lumped reaction VW in model “Lumped 1” contains reactions V and W from the detailed model). If lumped reactions overlap partially (like VW and UV in the models “Lumped 1” and “Lumped 2”), they do not fit to each other and the models cannot be directly combined. Similar conflicts would occur with lumped metabolites (not shown).

With a number of models already available, large dynamic models may be built by combining existing models of biochemical reactions [6] or cellular pathways [7]. Model merging can be straightforward if the input models originate from the same modelling framework, share the same naming conventions, and are based on a common set of non-conflicting biochemical quantities. In general, however, models will originate from different sources, so conflicts may easily occur. Model merging could be facilitated by computer programs that

execute uncritical steps and perform validity checks, but such tools and the theory to support them are still in their infancy. Model combination-whether manually or assisted by computer programs – requires that models are appropriately prepared: experiments and model formats must be standardized [5, 4], and all model elements need to have a clear biochemical meaning.

In publications, the elements are usually described in words (e. g. ''cellular concentration of ATP in mM''), while in computer-readable formats like SBML, they can be annotated with references to public databases (e. g. ATP may be represented by the identifier `C00002` in the KEGG database [8]). A chemical reaction can be annotated by an identifier or by specifying its substrates and products.

Merging of model structures

The aim in biochemical model merging is to combine several models describing reactions or biochemical pathways in order to obtain a valid model of the combined system. Before considering dynamic models, let us first have a look at simple structural models as shown in Fig. 1a. A structural model consists of a list of elements representing biochemical entities (e. g. metabolites and chemical reactions specified by annotations); in different models, the same entity can bear different names. Model elements may be linked to further information (e. g. pictures, comments, or mathematical expressions). Figure 1a shows how two overlapping pathway structures are combined: the resulting pathway contains all elements of the original models and pairs of duplicate elements (B = b, v = u and C = c) are merged into single elements, respectively. Merging of structural models involves the following steps:

1. The model elements have to be compared-either by a human expert or automatically-to detect duplicates. For automatic comparison, model elements have to bear annotations (i. e. standardized substance names or links to biological databases) that unambiguously determine their biochemical meaning. A simple string comparison between element names would not suffice because models may follow different naming conventions.
2. If duplicate elements are found, their accompanying information needs to be merged; if the two elements contain contradicting information (e. g. two models assign different concentrations to the same metabolite), some of the information has to be discarded.
3. Severe conflicts can arise if an element in one model (e. g. the lumped reaction VW in Fig. 1b) corresponds to several elements in another model (reactions V and W), or if several elements partially overlap in their meaning (e. g. the lumped reactions UV and VW). Such overlaps are a notorious source of conflict: in particular, if elements in a model are linked to mathematical expressions (e. g. chemical reactions are described by kinetic rate laws), the expressions for over-

lapping entities will probably not fit to each other. Therefore, overlapping elements generally should be avoided in model merging.

In this article, I shall discuss some basic theoretical concepts behind model merging. Merging of model structures will be our starting point: the same scheme also applies, *mutatis mutandis*, to dynamical models from different mathematical frameworks as long as all mathematical equations are given in the form of explicit assignments. Firstly, I shall focus on kinetic models as a special case and discuss the following questions: what are the basic elements in such models, in analogy to the structural elements shown in Fig. 1? How can we compare the biochemical meaning of elements and how can we detect conflict between them? Which additional problems can occur in dynamical models? Secondly, a general merging algorithm for explicit biochemical models is presented; it is applicable both in manual and automatic model merging. Finally, I shall classify some general validity criteria for biochemical models and discuss how models should be prepared to allow for model merging and other kinds of model reuse.

Mathematical Models and their Biochemical Semantics

Explicit biochemical models

Mathematical models allow the simulation of the dynamics of biochemical processes; depending on the system studied and on the questions to be answered, various mathematical frameworks can be used, including kinetic models, reaction–diffusion models, particle-based stochastic models, or constraint-based flux models. Despite their different forms, all such models describe a number of mathematical elements (variables, parameters...,) that are associated with biochemical objects (e. g. molecules) or quantities (e. g. concentrations). In addition, they contain mathematical statements supposed to hold for these quantities (e. g. ordinary differential equations, equality constraints, maximal postulates). The list of statements may either be an *ad hoc* collection (e. g. a number of the constraints used for flux balance analysis) or a complete description that allows for predictive simulations (e. g. a system of rate equations); in the latter case, mathematical solutions of the model should correspond, approximately, to the possible behaviour of the biological system under study.

Kinetic models

As a well-known example, we shall consider kinetic models comprising independent substance concentrations $c_i(t)$, dependent substance concentrations $c_j^{\mathrm{dep}}(t)$, external substance concentrations c_l^{ext} , reaction velocities $\nu_k(t)$, and kinetic constants p_m. The values are determined by explicit equations:

$$
\begin{aligned}
& p = (p_1, p_2, \ldots)^{\mathrm{T}} && \text{(constant numbers)} \\
& c_{\mathrm{ext}} = (c_1^{\mathrm{ext}}, c_2^{\mathrm{ext}}, \ldots)^{\mathrm{T}} && \text{(constant numbers)} \\
& c(0) = (c_1(0), c_2(0), \ldots)^{\mathrm{T}} && \text{(constant numbers)} \\
& c_j^{\mathrm{dep}}(t) = g_j(c(t)) \\
& \mathrm{v}k(t) = f_k(c(t), c^{\mathrm{dep}}(t), c^{\mathrm{ext}}, p) \\
& \frac{dc_i}{dt} = \sum_l N_{il} \upsilon_l(t)
\end{aligned}
\qquad (1)
$$

for all values of k, i, and j, where N is the stoichiometric matrix, the functions f_k denote kinetic rate laws, and the functions g_j relate the dependent concentrations to the independent concentrations. The variables and parameters represent biochemical quantities and are described by explicit algebraic or differential equations; models with these two properties (e. g. kinetic models, reaction–diffusion models, but also certain stochastic models) will be called "explicit biochemical models".

Computational cycles

For explicit biochemical models, computations become much more simple if all equations can be evaluated one after the other. To illustrate this point, let us consider an equation system with parameters p_1, p_2,..., differential equations of the form $\mathrm{d}x_i/\mathrm{d}t = f_i(x,y,p)$, and algebraic equations of the form $y_k(t) = g_k(x,y,p)$. If the mathematical formula for $g_k(x,y,p)$, corresponding to variable y_k, contains another variable y_l, then y_k is said to be computed from y_l. If each variable y_k in the model is only computed from variables y_l with smaller $l < k$, then the model is sequentially computable: the equations can be directly evaluated one after the other in each integration step. To check whether a model is sequentially computable, we can build a graph with nodes corresponding to the variables y_k and directed edges representing the "computed-from" relation. Cycles in the graph are called "computational cycles". If this graph is acyclic, the variables can be ordered such that the model is sequentially computable. If a model contains computational cycles (e. g. $y_1 = g_1(y_2)$, $y_2 = g_2(f_1)$), computations can become difficult and solutions may be non-unique or they may not even exist. Cycles between the differential equations, on the other hand, will not cause such problems.

Models as statement lists

In the following, we shall only consider explicit biochemical models consisting of algebraic equations $x = f(...)$ or differential equations $\mathrm{d}x/\mathrm{d}t = f(...)$. In such models, each quantity (or its time derivative) can be directly computed if the values of other quantities are known. Other

kinds of mathematical statements, such as inequalities $x < f(...)$, maximal requirements $x \stackrel{!}{=} \arg\max_y f(y, ...)$, or probabilistic assignments will not be considered here. Formally, an explicit biochemical model can be regarded as a list of model elements (also called "model statements"), each consisting of a biochemical quantity, a mathematical object and a mathematical assignment:

- The biochemical quantity (e. g. a concentration, reaction rate, compartment volume, or kinetic constant) is defined by a type (e. g. `concentration`), a unit (e. g. mM), a biochemical entity (e. g. a certain metabolite), and possibly, a location (e. g. a certain cell compartment). A quantity can also be related to several entities (e. g. a Michaelis constant refers to both an enzyme and a substrate metabolite).
- The corresponding mathematical object (e. g. a variable or a parameter) has a name or a unique identifier and a certain type (e. g. non-negative real number, time-dependent function $c(t)$, field $c(x,t)$).

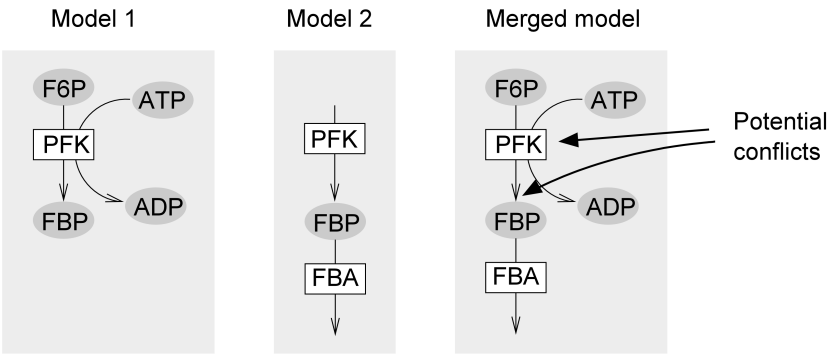


Figure 2. Merging of two small example models (equations see Table 1). Model 1 describes the PFK reaction rate at given substrate and product levels. Model 2 describes the mass balance of FBP resulting from production and degradation. The two models make different statements about the quantities representing PFK and FBP (thick arrows), so concatenating the models leads to conflicts. Abbreviations: ATP: adenosine triphosphate; ADP: adenosine diphosphate; F6P: fructose-6-phosphate, FBP: fructose-1,6-bisphosphate, PFK: phosphofructokinase, FBA: fructose-bisphosphate aldolase.

Table 1. The models from Fig. 2 are shown as statement lists. Each row represents a model statement (i. e. a model element). The numerical values in the example have been chosen arbitrarily. Simple concatenation of statements would lead to conflict because of duplicate biochemical quantities (marked by stars). When merging the models, one of the statements for PFK (and one for FBP) has to be chosen.

Model 1

Quantitiy	Math. Object	Assignment	Conflict
ATP concentration [mM]	c_{ATP}	$c_{ATP} = 1$	
ADP concentration [mM]	c_{ADP}	$c_{ADP} = 0.2$	
F6P concentration [mM]	c_{F6P}	$c_{F6P} = 0.5$	
FBP concentration [mM]	c_{FBP}	$c_{FBP} = 0.5$	*
PFK velocity [mM/s]	v_{PFK}	$v_{PFK} = f_{PFK}\ (c_{ATP}, c_{ADP}, c_{F6P}, c_{FBP})$	*

Model 2

Quantitiy	Math. Object	Assignment	Conflict
PFK velocity [mM/s]	v_{PFK}	$v_{PFK} = 0.1$	*
FBA velocity [mM/s]	v_{FBA}	$v_{FBA} = f_{FBA}\ (c_{FBP})$	
FBP concentration [mM]	c_{FBP}	$dc_{FBP}/dt = v_{PFK} - v_{FBA},\ c_{FBP}(0) = 0.2$	*

- The numerical values of the quantity are determined by mathematical assignments. In our terminology, all mathematical assignments for a quantity are regarded as a combined assignment. A differential equation, for instance, needs to be accompanied by an initial condition; both equations form a combined assignment and appear in the same model element.

Kinetic models like Equation (1) can be written in the form of statement lists: two small example models are shown graphically in Fig. 2 and as statement lists in Table 1.

Relations between biochemical quantities

In general, two biochemical quantities can be (i) identical, (ii) equivalent, i. e. identical up to conversion (e. g. concentration versus amount, same quantity measured in different units), (iii) semantically dependent (e. g. ATP concentration in cytoplasm and entire cell; concentration of glucose and hexoses; lumped metabolic pathway or single reaction in this pathway), or (iv) semantically independent (e. g. a concentration and a reaction velocity, concentrations of two unrelated substances).

Table 2. Possible relations between biochemical quantities. Quantities are specified by four characteristics: type, unit, entity and location. The rows describe conditions for the four possible relations. Several rows for the same relation denote alternative possibilities; bars (-) denote arbitrary entries.

Status	Type	Unit	Entity	Location
(i) Identical	Identical	Identical	Identical	Identical
(ii) Convertible	Identical or related	Different	Identical	Identical
(iii) Dependent	Identical or related	-	Identical or overlapping	Overlapping
	Identical or related	-	Overlapping	Identical or overlapping
(iv) Semantically Independent	Unrelated	-	-	-
		-	No overlap	-
		-	-	No overlap

Conflict within or between models can arise if the same biochemical quantity appears twice or if different quantities are semantically dependent. Two quantities are semantically dependent if their mere definition implies mathematical constraints or dependencies between their numerical values. For instance, the ATP amount in the cell and the ATP amount in the mitochondria are semantically dependent because the ATP amount in the mitochondria can never be larger than the total ATP amount. Another example is the velocities of the lumped reactions in Fig. 1b, which must have identical values. Besides semantic dependence, there may be other dependencies due to empirical laws (e. g. thermodynamic dependencies between rate constants or physiologically required concentration ratios).

To compare two given biochemical quantities, we need to describe them in a formal way: as previously stated, a biochemical quantity is specified by its type, a unit, one or more biochemical entities, and possibly, a location. Different quantity types are related if they refer to the same information: for instance, `concentration` and `amount` may be linked via the definition `concentration = amount/volume`; in different models, a compartment size may be described by a length, an area, or a volume, so these three types are related. Quantities referring to different kinds of objects (e. g. concentrations and reaction velocities) are unrelated.

Entities and locations have to be specified by annotations (e. g. a link to a database entry representing a biochemical substance). The biochemical entities and locations can be seen as notions: the entity `glucose`, for instance, comprises all glucose molecules in a system under consideration. If two different entities (e. g. `hexose` and `glucose`) share, by definition, common instances (in this case, specific glucose molecules), they are called *overlapping*. If instances of one entity (e. g. `ATP`) necessarily contain instances of another entity (e. g. `phosphate group`) as physical parts, the entities are also overlapping. A similar criterion holds for locations: locations are overlapping if they include common spatial regions (e. g. `cell` and `mitochondria`). Based on these relationships, one can compare different biochemical quantities by comparing their four characteristics, as shown in Table 2.

Meaning and reference of models

To merge models in a plausible way, we have to consider their biochemical interpretation. To do so, we have to link mathematical objects to biochemical quantities. But moreover, we also need to specify what are the basic statements that constitute a model. To illustrate this, let us consider two models containing the assignments $a = f(b)$, $b = g(c)$ (in model 1) and $a = h(c)$, $b = g(c)$ (in model 2, where the function h is defined by $h(x) := f(g(x))$ for all x). The mathematical relationship between c and a is identical in both models, so the models are mathematically equivalent; semantically, however, they make different statements (a depends on b in model 1, while it depends on c in model b).

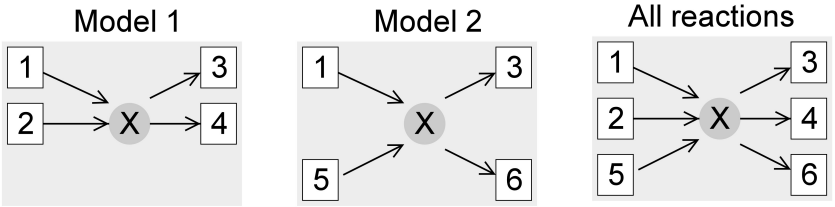


Figure 3. What is a basic model statement? According to model 1, a metabolite X participates in reactions 1, 2, 3 and 4. In model 2, it participates in reactions 1, 3, 5 and 6. In a merged model, one could either accept one of the two rate equations (regarding a rate equation as a basic statement), or one could assume a new rate equation comprising all six reactions (regarding each single term as a basic statement).

Following Frege's distinction between sense ("Sinn") and reference ("Bedeutung") [9] used in theological analysis of phrases, one may say that models 1 and 2 have the same reference (the same overall relation between numerical values), but a different sense (i. e. presumed direct relations between quantities). This difference does not play a role as long as the models are considered in their original form; however, it becomes apparent if the equation for b is changed during model merging: in this case, the mathematical behaviour in model 1 will change, while in model 2, it will not be affected.

For another example, let us consider two models containing contradictory statements for the same metabolite concentration,

$$\text{Model 1: } dc/dt = \nu_1 + \nu_2 - \nu_3 - \nu_4 \tag{2}$$

$$\text{Model 2: } dc/dt = \nu_1 + \nu_5 - \nu_3 - \nu_6. \tag{3}$$

The variable names are assumed to be non-conflicting and ν_1, ν_2, ν_3, ν_4, ν_5 and ν_6 denote the rates of different reactions (see Fig. 3). When merging the two models, we could accept one of the two rate equations (2) or (3) as our assignment for $c(t)$. This would imply that we

regard an entire rate equation as a basic model statement, which makes sense if we fit a model globally to concentration time series. Alternatively, we could merge the two rate equations and use:

$$dc/dt = v_1 + v_2 + v_5 - v_3 - v_4 - v_6. \quad (4)$$

With this choice, we assume implicitly that each of the terms on the right-hand side represents a basic statement, the fact that the metabolite is involved in a certain reaction. This point of view makes sense if models are built by combining individual reactions, possibly measured *in vitro*. It is also the rationale behind the structure of SBML.

Model Merging

Conflicts between statements

A naive way to merge explicit biochemical models would be to concatenate their statement lists (if necessary, after adjusting the variable names); the concatenated model would cover all quantities and statements from both input models. If all statements in the input models are true, then the concatenated model will be true as well, because correctness of a basic statement does not depend on the other statements around it. On the other hand, if models describe completely unrelated quantities, merging them should not create any conflict either. But if the two models contain identical, equivalent or semantically dependent quantities, the concatenated model may contain contradictions – especially if the original models are inaccurate or fitted to different experimental situations. Typical possible conflicts are as follows:

1. *The concatenated model contains different statements for the same quantity.* For instance, the two models in Fig. 2 make different statements about the PFK reaction rate: in model 1, the value depends on other quantities, while in model 2, the value is fixed. The concatenated statement list would be logically inconsistent because only one of the statements can be correct. Accordingly, the combined model would have no mathematical solution (except for rare cases in which both statements yield the same numerical value). Thus for each duplicate pair, one of the two statements has to be omitted. The combined resulting model will still be complete (no variable is missing), but it may contain computational cycles.
2. *The concatenated model contains semantically dependent quantities*. A model with semantically dependent quantities may be valid, but the corresponding mathematical assignments need to be fine-tuned to satisfy certain restrictions. If two semantically dependent quantities originate from different models, these restrictions will not be stated in either of the models, and it is likely that they will be violated after merging. At the same time, none of the quantities can be omitted

because both may be needed to compute other quantities, so the conflict cannot be resolved. Thus, models with semantically dependent elements should not be directly merged.

3. *The combined model may violate a physical law.* An example is the Wegscheider condition (see, e. g. [10]), which constrains the kinetic parameters along a circle in a metabolic network. If the merging of two models leads to a new circle and if the models were not especially prepared, the newly arising Wegscheider condition will probably not be satisfied. In the case of Wegscheider conditions, safe merging could be ensured by an appropriate parametrization of the kinetic rate laws [11, 12, 10].

A simple merging algorithm

In principle, merging of acyclic explicit biochemical models resembles the merging of structural models shown in Fig. 1. As before, we need to match elements and find identical and conflicting pairs: but now the elements (boxes and ellipses) represent model statements, they are compared according to their biochemical quantities, and the mathematical assignments are treated as additional information.

In the present context, explicit biochemical models are called valid if they satisfy the following criteria: (V1) correct syntax including consistent use of variable names; (V2) all elements are properly annotated; (V3) for each quantity, there is an assignment that allows it (or its time derivative) to be computed from the other quantities; (V4) the model can be sequentially computed, i. e. it does not contain computational cycles; (V5) each assignment agrees with the definition of the biochemical quantity described (e. g. only positive values for concentrations); (V6) all quantities are semantically independent (i. e. there are no pairs of identical, convertible, or semantically dependent quantities).

For the merging of two models (called ‘‘model 1’’ and ‘‘model 2’’), we assume that they are both valid. Ideally, the merged model should contain all statements from the input models; if this is not possible, some of the statements may be left out. In any case, the merged model has to be valid. This can be achieved by the following algorithm [13], which is actually quite similar to the merging scheme for structural models:

1. Convert all element names and physical units to a common set of names and standard units.
2. Compare all pairs of quantities from both models (see Fig. 4, left). Because of the previous conversion, pairs of quantities will either be identical (i. e. duplicates), semantically dependent, or semantically independent.

3. If semantically dependent quantities have been detected, stop the merging process and raise a warning message.
4. For each pair of duplicate quantities, choose one of the alternative model statements. The choice can be made either by the user or automatically (e. g. according to a rule like "always choose assignments from model 1").
5. Certain combinations of choices may lead to computational cycles (see Fig. 4, right), but it is always possible to avoid them by an appropriate choice of statements (e. g. by always choosing the assignments from model 1). In the algorithm, cycles are detected (by analysing the graph of dependencies between algebraic equations) and removed by revising some of the earlier choices.

This algorithm will return either a valid output model or stop with a warning.

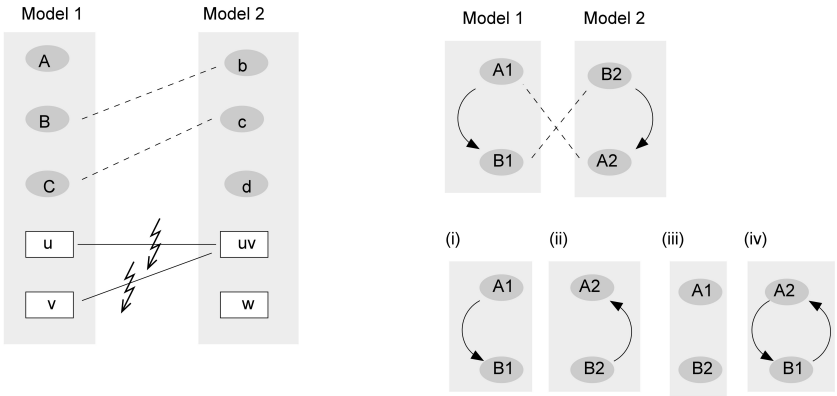


Figure 4. Merging of annotated biochemical models. Left: result of the pair-wise comparison between model 1 and model "lumped 2" from Fig. 1. Some elements are found to be identical (dotted lines) or semantically dependent (solid lines). Due to the semantic dependencies, merging should be abandoned in this case. Right: removal of computational cycles. Solid arrows within the models show that quantities are computed from each other; e. g. in model 1, quantity B is computed from quantity A. After matching the duplicate pairs (A1 = A2, B1 = B2), there are four possible choices: (i) keeping both elements from model 1; (ii) keeping both elements from model 2; (iii) keeping the two independent elements; (iv) keeping the dependent elements. The last choice creates a computational cycle and should therefore be avoided.

Merging of SBML models

SBML [1] is a widely used, XML-based format for biochemical models. SBML is tailored for kinetic models and describes simultaneously the mathematical form and the biochemical interpretation of a model. The main elements of a model represent compartments, substances and reactions including stoichiometries and kinetic laws and parameters. For simulations, the

`species` tags, which refer to substance amounts or concentrations, are translated into mathematical variables. By default, amounts and concentrations are assumed to follow a kinetic rate equation, but it is possible also to specify algebraic and differential equations for them.

Model elements (substances, compartments, reactions etc.) in SBML are not denoted by standard names, but by identifiers defined *ad hoc* within each model. However, elements can also be annotated by references to databases; a recommendable format is the MIRIAM-compliant RDF syntax with BioModels qualifiers [5, 14, 15]. It allows the annotation of model elements (e. g. a `species` tag describing a substance) with biological entities listed in databases. Besides exact equality, the qualifiers make it possible to specify different kinds of relationship: the `isVersionOf` qualifier, for instance, indicates that a substance described in a model (e. g. glucose) belongs to a substance class (e. g. hexoses) listed in the database.

Syntactically, SBML does not have the form of a statement list; however, a valid and properly annotated SBML file corresponds to an explicit biochemical model, so the above merging algorithm is in principle applicable to SBML models; we have implemented similar merging algorithms in the tools SBML merge [13] and SemanticSBML [16]. SemanticSBML allows the annotation, checking and merging of SBML models. It helps the user to annotate model elements with unique identifiers from various databases including KEGG [8], Reactome [17] and ChEBI [18]. These annotations are used for comparing the elements in model merging: the tool aligns presumably identical model elements and indicates conflicts between them; the user can then revise the alignment and resolve the conflicts. Following the structure of SBML, the chemical reactions (corresponding to individual terms in the rate equations) and not the entire rate equations are treated as basic model statements.

Validity Criteria for Biochemical Models

A main task in model merging is to ensure or to check the correctness of the merged model. Correctness, however, is a matter of definition: according to George Box [19], "essentially, all models are wrong, but some are useful", so even the best cell model is only a rough approximation of reality. Thus, instead of requiring correctness in an absolute sense, we shall ask whether a model satisfies certain *validity criteria*; which of the criteria is relevant in a specific case depends on the type and the purpose of the model. Even if a validity criterion is almost trivial, it may become an issue when models are merged automatically. The criteria can be grouped into five categories:

1. **Syntax**. Syntactical correctness ensures that a model can be read and processed, which is a basic requirement for all further validity checks and for model reuse in general. Syntactic problems, such as typos or missing tags in an SBML file, can be detected automatically from the model alone without any reference to a math-

ematical or biochemical interpretation; an automatic validation tool for SBML files can be found at [20]. In a broader sense, we can also regard verbal descriptions in a paper as syntactically incorrect if they are unclear or incomplete.

2. **Mathematics and calculation**. Depending on the intended sorts of calculations, a model should have certain mathematical properties, in particular, existence or uniqueness of mathematical solutions. For kinetic models, for instance, one may require that (i) there is one explicit equation per variable, (ii) the right-hand sides are defined for all allowed values of the function arguments (e. g., non-negative values for all concentration variables; real values for all flux variables) and (iii) there are no computational cycles.
3. **Biochemical semantics**. In this category, we consider the biochemical meaning of model elements, but only regarding simple ontological facts (''glucose is a hexose'', ''mitochondria are part of the cell'', ''reaction VW contains reactions V and W as parts''). Possible validity requirements are: (i) all elements must be correctly annotated; (ii) individual model statements must agree with their semantics (e. g. a variable representing a concentration must be non-negative); (iii) statements must agree with each other (or, more strictly, all quantities must be semantically independent).
4. **Empirical facts**. In addition, one may require that a model respects certain laws of physics (e. g. second law of thermodynamics), chemistry (e. g. conservation of atom numbers), or biochemistry (e. g. realistic values for concentrations). Testing these criteria may require semantic annotations and additional information (for instance, about molecule structures, energies etc.).
5. **Relevance**. Even if a model is free from conflict, it will not automatically be useful; in fact, a model should be based on plausible assumptions, represent a biological system of interest, bring out its basic mechanism, contain only relevant processes and agree with available data. These requirements do not concern the model alone, but also its relationship to available data and to other competing models. It is hard to test them automatically, and they are possibly beyond the realm of automatic checking and merging.

In my point of view, a model is wrong if it fails to fulfil a validity criterion *that it should fulfil*. A didactic model (e. g. a prototypic oscillator model) must have a mathematical solution, but it need not refer to a specific system, so criteria regarding biochemical semantics and realistic numerical values do not play a role. On the other hand, a model that describes a specific pathway should meet these requirements. For automatic model checking, it would be helpful to state explicitly the scope of a model, i. e. which cell types and experimental situations are described, which validity criteria should be fulfilled, or which calculations should be possible; to date, however, there is no formal way to state such requirements in SBML files.

How to prepare reusable models

Besides model merging, there are also other situations in which models are reused: models may be refitted to new data, expanded, simplified, or used as examples to build models for other cell types. Modellers should bear in mind that their models might be reused later, possibly by other people, and should ensure reusability of models right from the beginning (a strategy that could be termed ''sustainable model development''). So when constructing a model (and even when designing the experiments that will lead to a model), one should think of typical problems that might occur later, for instance: (i) if the experimental conditions (e. g. the microbial strain used) are not standardized or not well documented, the resulting models may not be compatible; (ii) lumped reactions and metabolites may cause problems and should be avoided, used in a systematic manner, or at least be described unambiguously; (iii) globally fitted parameters may become meaningless after merging and will have to be estimated again.

To avoid such problems, experimentalists and modellers should support standardization efforts (e. g. SBML, MIRIAM, and STRENDA [21]); models should be published (as required in MIRIAM) with all information necessary to reproduce the simulations and model fitting; they should be accessible in a standard (preferably free) format like SBML and be submitted to repositories such as BioModels [2] or JWS online [3]. The meaning of model elements has to be specified unambiguously: in publications, standardized identifiers or names should be used to describe the model elements.

Acknowledgments

I would to thank Falko Krause and Jannis Uhlendorf for their comments on this manuscript. This work was funded by the European integrated project BaSysBio.

References

[1] Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgemn, T.C., Hofmeyer, J.H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novère, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J. (2003) The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* **19**(4):524 – 531.

[2] http://www.ebi.ac.uk/biomodels.

[3] Olivier, B., Snoep, J. (2004) Web-based kinetic modelling using JWS online. *Bioinformatics*, **20**(13):2143 – 2144.

[4] Klipp, E., Liebermeister, W., Helbig, A., Kowald, A., Schaber, J. (2007) Systems biology standards-the community speaks. *Nat. Biotechnol*. **25**:390 – 391.

[5] Le Novère, N., Finney, A., Hucka, M., Bhalla, U.S., Campagne, F., Collado-Vides, J., Crampin, E.J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J.L., Spence, H.D., Wanner, B.L. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* **23**(12):1509 – 1515.

[6] Teusink, B., Passarge, J., Reijenga, C.A., Esgalhado, E., vanderWeijden, 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.

[7] Snoep, J.L., Bruggeman, F., Olivier, B.G., Westerhoff, H.V. (2006) Towards building the silicon cell: A modular approach. *Biosystems* **83**:207 – 216.

[8] Kanehisa, M., Goto, S., Kawashima, S., Nakaya., A. (2002) The KEGG databases at genomenet. *Nucleic Acids Res.* **30**:42 – 46.

[9] Frege, G. (1892) Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* **100**:25 – 50.

[10] Ederer, M., Gilles, E.D. (2007) Thermodynamically feasible kinetic models of reaction networks. *Biophys. J.* **92**:1846 – 1857.

[11] Liebermeister, W., Klipp, E. (2005) Biochemical networks with uncertain parameters. *IEE Proc Systems Biology* **152**(3):97 – 107.

[12] Liebermeister, W., Klipp, E. (2006) Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor. Biol. Med. Mod.* **3**:41.

[13] Schulz, M., Uhlendorf, J., Klipp, E., Liebermeister, W. (2006) SBMLmerge, a system for combining biochemical network models. *Genome Informatics Series* **17**(1):62 – 71.

[14] RDF/XML syntax specification (revised) (2004) http://www.w3.org/TR/rdf-syntax-grammar/.

[15] http://www.ebi.ac.uk/compneur-srv/miriam-main/mdb?section=qualifiers.

[16] http://sysbio.molgen.mpg.de/semanticsbml/.

[17] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio,P., Schmidt, E., deBono, B., Jas-sal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E., Stein, L. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*. **33** Database Issue:D428 –D432.

[18] http://www.ebi.ac.uk/chebi/.

[19] Box, G.E.P., Draper, N.R. *Empirical Model-Building and Response Surfaces*. Wiley, New York, 1987.

[20] http://www.sbml.org/Facilities/Validator.

[21] http://www.strenda.org/.

Biographies

Robert A. Alberty

worked on enzyme kinetics 1950 – 1963, was an administrator at the University of Wisconsin 1963 – 1967, was an administrator at MIT 1967 – 1982, did research on the thermodynamics of petroleum processing 1982 – 1992, research on biochemical thermodynamics 1992 – 2006, and has recently been working on enzyme kinetics. He has been a coauthor of a textbook on physical chemistry for 50 years, and has written two books on biochemical thermodynamics. He feels fortunate that his experience in petroleum thermodynamics prepared him to work on biochemical thermodynamics, and his work on biochemical thermodynamics prepared him to work on enzyme kinetics. Sometimes it is a good idea to get out of your field to work in a related field because when you come back to your earlier field you will see it with new eyes.

Rolf Apweiler

is a Team Leader and Senior Scientist at the European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. He studied Biology with a focus on Biochemistry and Molecular Biology in Heidelberg, Germany and Bath, UK, and worked in drug discovery in the pharmaceutical industry. He became involved in Bioinformatics through the Swiss-Prot project in 1987. He received his PhD in 1994 from the Center for Molecular Biology, University of Heidelberg, Germany and joined the European Bioinformatics Institute the same year. Dr Apweiler has coordinated the Swiss-Prot work at the European Bioinformatics Institute since 1994. He also started, among other projects, the TrEMBL protein database, the Integrated resource of protein families, domains and functional sites (InterPro), Gene Ontoloy Annotation (GOA), the Integr8 web portal, the Genome Reviews, and the UniProt resource (the sucessor of the Swiss-Prot, TrEMBL and PIR projects). These projects have organised large amounts of protein information, provided comparisons between proteomes and aim to produce dynamic, controlled vocabularies that can be applied to all organisms. In addition, Dr Apweiler has been in charge of the EMBL nucleotide sequence database since 2001. Dr Apweiler served on many review and editorial boards and published more than hundred peer-reviewed articles and numerous book chapters. Rolf Apweiler has also a long-standing interest in data standards and nomenclature as exemplified in his engagement in the IUBMB Nomenclature Committee, the HUGO gene nomenclature committee, and in the HUPO Proteomics Standards Initiative.

Richard N. Armstrong

Education

1970:	B.S. Western Illinois University, Chemistry, Macomb, Illinois
1975:	Ph.D. Marquette University, Organic Chemistry Milwaukee, Wisconsin (with Prof. N. E. Hoffman)
1976 – 1978:	Postdoctoral Fellow, University of Chicago, Chicago, Illinois (with Prof. E. T. Kaiser)
1990:	Sabbatical Leave, Center for Advanced Research in Biotechnology, Rockville, MD, Protein Crystallography (with Gary Gilliland)

Positions Held

since	1997: Professor, Department of Chemistry, Vanderbilt University, College of Arts & Sciences
since 1995:	Professor, Department of Biochemistry and the Center in Molecular Toxicology, Vanderbilt University School of Medicine
1990 – 1995:	Professor, Department of Chemistry & Biochemistry, UMCP
1988 – 1990:	Chair, Biochemistry Division, UMCP
1985 – 1990:	Associate Professor, Department of Chemistry & Biochemistry, UMCP
1980 – 1985:	Assistant Professor, Department of Chemistry University of Maryland, College Park
1978 – 1980:	Staff Fellow, Laboratory of Bioorganic Chemistry, NIAMDD, National Institutes of Health, Bethesda, Maryland
1976 – 1978:	Postdoctoral Fellow, Department of Chemistry, University of Chicago, Chicago, Illinois

115 publications (refereed articles, book chapters, reviews), more than 120 invited lectures.

Research Interests

Functional genomics. Enzymatic basis of antibiotic resistance. Mechanism and stereochemistry of enzyme-catalyzed reactions. Metabolism and detoxification of drugs and toxic compounds. Protein structure and function and engineering. Protein crystallography. Applications of physical organic chemistry to biochemical and biotechnological problems. Stereochemistry and conformations of strained molecules.

Richard Cammack

is Professor of Biochemistry at King's College, University of London. He was a Major Open Scholar at Jesus College, University of Cambridge, from where he graduated with a BA in Natural Sciences in 1965 and PhD in Enzymology in 1968, under the supervision of Malcolm Dixon. His research has centred on the use of spectroscopic techniques to study mechanisms of electron transfer and enzyme catalysis, particularly in complex iron-sulfur proteins. He is past Chairman (2000 – 2005) of the Nomenclature committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and Joint commission on Biochemical Nomenclature (JCBN), and Editor-in-Chief of the second edition of the Oxford Dictionary of Biochemistry and Molecular Biology. He has published two books, and over 220 research papers. He is currently investigating aspects of the role of iron in health and disease.

Athel Cornish-Bowden

carried out his undergraduate studies at Oxford, obtaining his doctorate with Jeremy R. Knowles in 1967. After three post-doctoral years in the laboratory of Daniel E. Koshland, Jr., at the University of California, Berkeley, he spent 16 years as Lecturer, and later Senior Lecturer, in the Department of Biochemistry at the University of Birmingham. Since 1987 he has been Directeur de Recherche in three different laboratories of the CNRS at Marseilles. Although he started his career in a department of organic chemistry virtually all of his research has been in biochemistry, with particular reference to enzymes, including pepsin, mammalian hexokinases and enzymes involved in electron transfer in bacteria. He has written several books relating to enzyme kinetics, including Analysis of Enzyme Kinetic Data (Oxford University Press, 1995) and Fundamentals of Enzyme Kinetics (3 rd edition, Portland Press, 2004).

Since moving to Marseilles he has been particularly interested in multi-enzyme systems, including the regulation of metabolic pathways. At present his main interest is in the definition of life and the capacity of living organisms for self-organization. In addition his principal areas of research, he has long had an interest in biochemical aspects of evolution, and his semi-popular book in this field, The Pursuit of Perfection, was published by Oxford University Press in 2004.

Robert N. Goldberg

received his Bachelor of Arts (Chemistry Major) from Johns Hopkins University and his Doctor of Philosophy (Physical Chemistry) from Carnegie-Mellon University. After completion of a post-doctoral research fellowship at Mellon Institute and the University of Pittsburgh, he joined the National Institute of Standards and Technology (formerly the National Bureau of Standards) in 1969. His primary areas of expertise include chemical

thermodynamics, calorimetry and equilibrium measurements, data evaluation, thermodynamics of solutions, biochemical thermodynamics, analytical microcalorimetry, and chromatography. A major focus of his research has been on the thermodynamics of enzyme-catalyzed reactions. This has resulted in the determination of the thermodynamic parameters for a large number of such reactions – including a substantial number of the most important reactions pertinent to physiology and to metabolism as well as reactions that are of major industrial interest. These studies involve the combined use of equilibrium and calorimetric measurements coupled with thermodynamic modeling calculations. The information obtained allows for the prediction of the position of equilibrium of the studied reaction(s) over wide ranges of temperature, pH, and ionic strength.

Recent research has included studies of biochemical reactions in non-aqueous solvents, redox reactions, and reactions in the shikimate and chorismate pathways.

Codes for performing equilibrium calculations on systems of biochemical reactions have also been developed and published. The entire field involving the thermodynamics of enzyme-catalyzed reactions has been surveyed and the data extracted and made available on the web: http://xpdb.nist.gov/enzyme_thermodynamics/. He has been active in IUPAC, ASTM, and in the U.S. Calorimetry Conference. He is a recipient of the Measurement Services Award of the National Institute of Standards and Technology.

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.

Carsten Kettner

studied biology at the University of Bonn and obtained his diploma at the University of Göttingen in the group of Prof. Gradmann which had the pioneering and futuristic name – ''Molecular Electrobiology''. This group consisted of people carrying out research in elec-

trophysiology 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 (Beilstein-Symposium and ESCEC), research (proposals) and development of new products such as MedPhyt, a medicinal plants database. Since 2004 he coordinates the work of the STRENDA commission and promotes along with the commissioners the proposed standards of reporting enzyme data. In 2007 he became involved in the invention of a program for the establishment of Beilstein Endowed Chairs for Chemical sciences and related sciences.

Sandra Orchard

Originally trained as a biochemist, Sandra Orchard spent a number of years working as an enzymologist for Roche Products Ltd, eventually becoming the Team Leader of a group looking at protein kinases as potential drug targets. She moved to the European Bioinformatics Institute in 2002, originally as a UniProtKB/Swiss-Prot curator but also worked on the InterPro and Gene Ontology databases before initiating the curation of the molecular interaction database IntAct of which she is now the curation coordinator.

Sandra has been involved in the work of the Human Proteome Organisation Proteomics Standards Initiative since its initiation, and is now a member of the Steering Committee.

Johann Rohwer

is Associate Professor in the Department of Biochemistry at Stellenbosch University, South Africa. He obtained his Ph.D. in 1997 from the University of Amsterdam, working on the control and regulation of the bacterial phosphotransferase system under the supervision of Hans Westerhoff. He then joined Stellenbosch University, where he and his colleagues Jannie Hofmeyr and Jacky Snoep constitute the ‘‘Triple-J Group for Molecular Cell Physiology’’, a research group that studies the control and regulation of cellular processes using theoretical, numerical and experimental approaches.

Johann has contributed to the theoretical development of metabolic control analysis, to its experimental application, and to the development of software tools for computational systems biology. His main research interests are the construction of kinetic models of cellular

function with a particular emphasis on plant central carbon metabolism, and the application of NMR spectroscopy to the non-invasive study of metabolism in vivo. He has received the President's Award from the South African National Research Foundation and the Silver Medal of the South African Society of Biochemistry and Molecular Biology. Together with the other Triple-Js, he has chaired the BTK: International Study Group for Systems Biology, and he represents his university on the South African National Bioinformatics Network.

Nicole S. Sampson

was born in Indianapolis, Indiana and acquired her B.S. degree in chemistry at Harvey Mudd College in 1985. She obtained her Ph.D. in the laboratory of Paul A. Bartlett at UC Berkeley in 1990 and then carried out postdoctoral research in the laboratory of Jeremy R. Knowles at Harvard University. Nicole joined the faculty at Stony Brook University in 1993 and is currently a Professor of Chemistry, as well as a member of the graduate programs in Pharmacology, Biochemistry & Structural Biology, and Biophysics. Her research interests are in the areas of mechanistic enzymology and chemical biology. Her work presently focuses on catalysis by cholesterol–modifying enzymes, how they modify the lipid bilayer and their role in bacterial pathogenesis, and probing protein–protein interactions in mammalian fertilization using synthetic molecules, in particular functionalized polymers.

Sampson's honors and awards include the Camille and Henry Dreyfus New Faculty Award, an NSF CAREER Award, the ACS Arthur C. Cope Scholar Award and the Pfizer Award in Enzyme Chemistry.

Her research program has been supported by the National Institutes of Health, the Petroleum Research Fund, the National Science Foundation, the American Heart Association, the Dreyfus Foundation, Biogen, and Bristol-Myers Squibb.

Hartmut Schlüter

Institution and Position:

Charité – University Medicine Berlin

Senior Scientist and Head of the Core-Facility Protein-Purification

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

1988:	Diploma (= M.Sc.) in Biochemistry, Faculty of Chemistry, University of Münster
1991:	Ph.D. (Dr. rer. nat.) in Biochemistry, University of Münster, Faculty of Chemistry, Thesis supervisor: Prof. Dr. H. Witzel
1994:	Heinz Maier-Leibnitz prize
1995:	Gerhard Hess award (DFG)
1995:	Bennigsen-Foerder prize
1991 – 1996:	Postdoctoral fellowship at the Medical Faculty of the University of Münster
1996:	Habilitation (Dr. rer. nat. habil.) in Pathobiochemistry at the Medical Faculty of the University of Münster
1996 – 2000:	Group leader at the Medical Faculty of the Ruhr-University of Bochum
2000 – current:	Senior Scientist and Head of the Bioanalytical Laboratory of Nephrology, Charité – University Medicine Berlin, Campus Benjamin-Franklin, Joint Facility of the Free University of Berlin and the Humboldt-University of Berlin
2003 – current:	Professor at the Charité, Campus Benjamin-Franklin, University-Medicine Berlin
2003 – current:	Member of the board of the Center for Functional Genomics – Berlin-Brandenburg (http://www.cffg.de/)
2004:	Head of the Core-Facility Protein Purification
2005:	Founding member of the Mass-Spec-Net Berlin-Brandenburg (www.mass-specnet.de)

Working field:

- Biochemistry and physiology of diadenosine polyphosphates
- Mass spectrometry of biomolecules
- Functional Proteomics – Enzymology
- Protein Purification
- Liquid Chromatography of biomolecules Automation

Dietmar Schomburg

1974:	Diplom in Chemistry at the Technical University "Carolo-Wilhelmina" in Braunschweig
1976:	Dr. rer. nat. in Chemistry (Structural Chemistry of Organo-phosphorus compounds)
1985:	Habilitation (Dr. rer. nat. habil.) for Structural Chemistry

Scientific Career:

1976 – 1978:	Post-Doc in the Chemistry Department at Technical University Braunschweig.
1978 – 1979:	Research Fellow at Harvard University in Cambridge, Mass., U. S. A. in Professor W.N. Lipscomb's and Professor F.H. Westheimer's groups.
1979 – 1981:	Post-Doctoral Fellow in the Chemistry Department at Braunschweig Technical University
1981 – 1983:	Assistant Professor (Hochschulassistent), Braunschweig Technical University
1983 – 1986:	Head of the x-ray lab at the German Centre for Biotechnology – GBF (Gesellschaft für Biotechnologische Forschung), Braunschweig
1987 – 1996:	Head of the GBF Department of "Molecular Structure Research."
1989 – 1995:	Head of CAPE (Center of Applied Protein Engineering)
1990 – 1996:	(apl.) Professor at the Technical University Braunschweig
1996 – 2007:	Full Professor of Biochemistry, University of Cologne
since 2007:	Full Professor of bioinformatics & biochemistry, Technical University Braunschweig

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 later together with Prof Westerhoff at the Netherlands Cancer Instuitute working on theoretical and modelling aspects of biological systems.

Currently Snoep is appointed at the University of Stellenbosch (Biochemistry), at the Vrije Universiteit in Amsterdam (Cellular Bioinformatics) and at the University of Manchester (Integrative Systems Biology).

His research aim is to get a quantitative understanding of cellular physiology, i. e. to attribute systemic (cellular) properties to characteristics of the underlying components (enzymes). Due to the complexity (non-linear interactions) and complicatedness (multitude of interactions) the research subjects necessitate a combined approach of precise and quantitative experimentation, computer modelling and a robust theoretical framework such as Metabolic Control Analysis. Research topics range from simple ecosystems to metabolic engineering of lactic acid bacteria, to the control of metabolic pathways such as glycolysis, to the control of DNA supercoiling.

Christoph Steinbeck

was born in Neuwied, Germany, in 1966. He studied chemistry at the University of Bonn, where he received his diploma and doctoral degree in the workgroup of Prof. Eberhard Breitmaier at the Institute of Organic Chemistry. Focus of his Ph. D. thesis was the program LUCY for computer assisted structure elucidation. In 1996, he joined the group of Prof. Clemens Richert at Tufts University in Boston, MA, USA, where he worked in the area of biomolecular NMR on the 3D structure elucidation of peptide-nucleic acid conjugates. In 1997 Christoph Steinbeck became head of the Structural Chemo- and Bioinformatics Workgroup at the newly founded Max-Planck-Institute of Chemical Ecology in Jena, Germany. In fall 2002 he moved to Cologne University Bioinformatics Center (CUBIC) as head of the Research Group for Molecular Informatics. His research focuses on methods for Computer-Assisted Structure Elucidation in Metabolomics and Natural Products Research. In December 2003 Christoph Steinbeck received his Habilitation in Organic Chemistry from Friedrich-Schiller-University in Jena, Germany.

His group develops a number of the leading open source software packages in Chemo- and Bioinformatics, including the Chemistry Development Kit (CDK), a Java library for chemo- and bioinformatics, NMRShiftDB, an open content database for chemical structures and their NMR data, and Bioclipse, an Eclipse-based Rich Client for everything and nothing in particular. Dr. Steinbeck is chairman of the Computers-Information-Chemistry (CIC) division of the German Chemical Society, trustee of the Chemical Structure Association (CSA) Trust, a lifetime member of the World Association of Theoretically Oriented Chemists (WATOC) and member of various editorial boards and committees. Today, Dr. Steinbeck is a lecturer in Chemoinformatics at the University of Tübingen, an Evangelist for Open Data, Open Standards and Open Source, and works as an independent consultant in Chemo- and Bioinformatics.

Neil Swainston

Education

10/01 – 08/05: Bioinformatics modules, University of Manchester,

09/97 – 10/98: MSc Computing Science, University of Newcastle-upon-Tyne. IRISA (06 – 10/98), Campus de Beaulieu, Rennes, France. MSc thesis: bioinformatics. Production of a bioinformatics web application to determine local alignments in nucleotide sequence data.

09/92 – 06/96: BSc (Hons) Chemistry with Industrial Experience, University of Manchester. First class honours. Industrial experience with Dow Deutschland Inc.

Cancer Research Campaign (01 – 05/96) Paterson Institute of Cancer Research, Christie Hospital, Manchester. BSc thesis: physical organic chemistry.

Employment

04/06 – present: Manchester Centre for Integrative Systems Biology, University of Manchester, Manchester, Experimental Officer (Information Management).

04/99 – 04/06: Waters Corporation, Micromass MS Technologies Centre, Manchester, Bioinformatics Team Manager.

10/98 – 03/99: AstraZeneca, Formally at: Hexagon House, Blackley, Manchester. Graduate trainee: IT problem management.

12/96 – 08/97: The British Council, Bridgewater House, Manchester. Website editorial assistant.

09/94 – 09/95: Dow Deutschland Inc. Werk Stade, Postfach 1120, 21677 Stade, Germany.

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.

Ulrike Wittig

is a research associate in the Scientific Database and Visualisation group of the EML Research gGmbH in Heidelberg, Germany. She studied biochemistry at the University of Leipzig, Germany and received her Ph.D. in biology from the University of Heidelberg, Germany in 1998. The Ph.D. thesis on mechanisms of apoptosis and oxidative stress was developed in a close collaboration between the University Hospital of Heidelberg and the German Cancer Research Center in Heidelberg. With the background of wet-lab work she joined a new founded group for Bioinformatics at the European Media Laboratory (EML) in Heidelberg and worked at the development of databases for biochemical pathways. Her research interests include modelling and visualisation of biochemical pathways, information extraction from biological data sources and data integration and standardisation in biological databases.

Author's Index

Index