ELSEVIER

Metabolic networks: enzyme function and metabolite structure
Vassily Hatzimanikatis*, Chunhui Li, Justin A Ionita and Linda J Broadbelt

Metabolism is one of the most complex cellular processes. Connections between biochemical reactions via substrate and product metabolites create complex metabolic networks that may be analyzed using network theory, stoichiometric analysis, and information on protein structure/function and metabolite properties. These frameworks take into consideration different aspects of enzyme chemistry, enzyme structure and metabolite structure, and demonstrate the impact of metabolic biochemistry on the systemic properties of metabolism. The integration of these approaches and the systematic classification of enzyme function and the chemical structure of metabolites will enhance our understanding of metabolism, and could improve our ability to predict enzyme function and novel metabolic pathways.

Addresses
Department of Chemical and Biological Engineering, Northwestern University, 2145 Sheridan Road, E136, Evanston, Illinois 60208-3120, USA
*e-mail: vassily@northwestern.edu

Abbreviations
EC   Enzyme Commission
ETS   electron transfer system
SMM   small-molecule metabolism

Introduction
Small-molecule metabolism (SMM) has been the subject of intense study because of its importance in generating a plethora of small organic compounds (metabolites) and energy from relatively simple molecules, such as glucose. These metabolites are further used for the synthesis of every cellular component and more complex chemical molecules, such as secondary metabolites and vitamins. Moreover, SMM also includes reactions that are responsible for the biodegradation of toxic compounds.

SMM consists of a large number of enzyme-catalyzed reactions. Current knowledge about these biochemical reactions is organized into databases that contain information about the substrates and products of biochemical transformations, the enzymes that catalyze these reactions

(if known), the substrate specificity and kinetic and regulatory properties of these enzymes, and the mapping of enzymes onto genomes of sequenced organisms [1–5]. One of the most valuable aspects of these databases is the organization of biochemical reactions into 'metabolic pathways' or 'metabolic maps'. In general, a pathway consists of the sequence of biochemical reactions that produce a set of metabolites from a set of precursor metabolites and cofactors. The length of a pathway is the number of biochemical reactions between the precursor and the final metabolites of the pathway. The definition of a pathway is not unique and, therefore, the number and length of pathways in the different databases vary.

In the two databases most commonly used in the study of metabolic pathways, MetaCyc and KEGG [2,5], biochemical reactions are organized into maps based on information from experimental literature. For example, the glycolysis pathway is defined as the set of reactions that have been identified, through biochemical and genetic studies, as the most important steps for the conversion of glucose to pyruvate. The organization of these biochemical reactions into databases allows the mining of this information for the identification of alternative routes between metabolites (Figure 1). This possibility of multiple biochemical pathways between two metabolites, which is extremely common [6••,7], is the result of the combinatorial complexity of enzyme chemistry.

The study of enzyme chemistry and its organization into metabolic pathways involves two very challenging problems. The first problem deals with the identification of the biochemical function of a protein based on the amino acid sequence and the structure of the protein. Although functional aspects of a protein can be deduced from its sequence and structure, the exact catalytic activity of the enzyme is not easy to predict. The other main challenge is the discovery of possible new enzymatic activities that connect existent or produce novel biochemical compounds. We will discuss here three of the main approaches that have emerged to address these challenges in the analysis of the biochemical complexity of SMM. The first one considers currently known biochemical reactions and compounds, and studies the properties of the metabolic pathways as networks, particularly focusing on network connectivity or properties derived from the stoichiometry of the metabolic pathways comprising the network. The second approach considers biochemistry with respect to the structure of the enzymes participating in a reaction and a pathway. The third approach views the biochemistry from the substrate/product level.

Figure 1
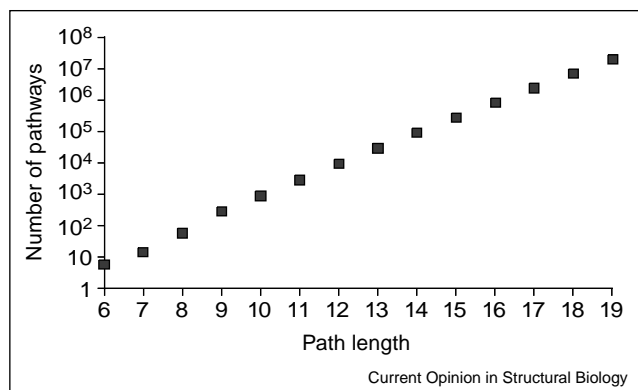
Distribution of the pathway length from glucose to pyruvate based on the *E. coli* biochemical reactions in the KEGG database. ATP, ADP, H_2O, redox carriers and cofactors are not considered as intermediates (i.e. they have been excluded from the nodes of the *E. coli* KEGG-based network). Reversibility is taken into account for those reactions that are classified as reversible in the KEGG database. The network was explored using the algorithm described in [6••,7].

We will also introduce a framework that could potentially bridge these approaches and offer a new method of analysis of metabolic biochemistry.

Pathways as networks: connectivity

A first-level analysis of SMM biochemical pathways employs graph theory methods and addresses their topological properties as networks of metabolites connected through biochemical reactions. In a graph theoretic representation of metabolism, metabolites are assigned to the nodes of a graph and the biochemical reactions interconverting the metabolites are the corresponding edges. Such a representation of metabolism allows the characterization of metabolic pathways with respect to two essential properties: the degree of metabolite connectivity, defined as the number of reactions in which a given metabolite participates; and the degree of interconnectivity or network diameter, defined as the shortest biochemical pathway averaged over all pairs of substrates. Jeong *et al.* [7] found that, with respect to the degree of connectivity, metabolic networks in 43 organisms (6 archaea, 32 bacteria and 5 eukaryotes) follow a power-law distribution characteristic of scale-free networks, with very few metabolites having a high degree of connectivity. With respect to the degree of interconnectivity, they found that any two metabolites in the system, in any of the 43 organisms, can be connected through a relatively short pathway of about three reaction steps. These results suggested that metabolism in every organism behaves like a small-world network.

Ma and Zeng [6••] revisited this analysis and pointed out that Jeong *et al.* had treated metabolites such as ATP, ADP, H_2O, phosphate and cofactors as nodes in their network, and that they had not considered the irreversibility of certain types of reactions. As a result, glucose was estimated to be two reaction steps away from pyruvate due to the connectivity of glucose to ATP, through hexokinase, which in turn is connected to ADP, which is connected to pyruvate through pyruvate kinase. These considerations are inconsistent with the definition of a pathway from glucose to pyruvate because the two postulated connections do not interconvert the carbon backbone of glucose. Ma and Zeng repeated the network analysis of 80 fully sequenced organisms after the removal of these small molecules and cofactors, and also accounted for irreversible reactions. Their results suggested that, in all organisms, the degree of connectivity of metabolites follows a scale-free distribution, but the degree of interconnectivity is greater than estimated previously. On average, it is different between organisms in different domains of life (eukaryotes 9.5, archaea 8.5, bacteria 7.2).

The results of both of these network studies provide useful insights into the properties and evolution of metabolism. They also demonstrate the importance of considering the specific properties of the enzyme chemistry and the need for the consistent definition of 'what is a pathway' when we analyze cell-wide metabolic networks.

Pathways as networks: stoichiometry and system properties

An analysis of metabolism that takes into account properties of the enzyme chemistry considers the balance of every metabolite and cofactor that participates in the metabolic reactions of a network. This approach, known as metabolic flux analysis or constraints-based analysis [8], is based on the basic principle of mass balance; this requires that, at steady state, the sum of the rates (fluxes) of the reactions that produce a metabolite must be equal to the sum of the rates (fluxes) of the reactions that consume this metabolite. This requirement imposes a set of linear constraints on the reaction rates that should be simultaneously satisfied when a network of metabolites is considered. The set of mass balance constraints defines the stoichiometric matrix of metabolic reaction networks. In addition, a set of inequality constraints is also included to constrain the directionality of irreversible reactions.

Constraints-based modeling of metabolism provides a unique framework for a network-based definition of metabolic pathways through the identification of 'elementary flux modes' [9–11] and 'extreme pathways' using convex analysis and linear algebra [12–14]. These are a subset of the biochemical reactions connecting a subset of metabolites in the network and they satisfy the mass balance equations of these metabolites. Each of these subsets can then define a pathway with substrate, intermediate and product metabolites, and removal of a

biochemical reaction from a subset will disrupt the synthesis of the product metabolites. These elementary pathways are unique in the sense that there is no other elementary pathway that consists of exactly the same set of metabolites and reactions. In addition, every possible set of reaction rates of the global metabolic network will be a linear combination of these elementary pathways. Therefore, through mathematical analysis of the constraints imposed by the biochemistry and the reversibility properties of the enzymes in a complex metabolic network, metabolic pathways can be defined in a systematic and consistent manner.

Constraints-based analysis also allows the investigation of a broad range of properties of metabolic systems, such as optimal growth rate patterns, physiological responses after gene deletion, growth medium requirements and network robustness [8]. One of the most interesting recent studies involves modeling of the metabolic pathways of *Saccharomyces cerevisiae* using a constraints-based modeling framework [15•]. This has been the first such systematic analysis of the metabolic properties of a eukaryotic model organism. Simulation of gene deletions using the model correctly predicted viable phenotypes and lethal phenotypes in 89.4% (496 out of 555) and 68.2% (30 out of 44), respectively, of the cases considered. This allowed the formulation of experimentally testable hypotheses for identifying the origins of the false predictions [16].

Palsson and co-workers have also demonstrated the importance of charge (electron and proton) balance in the constraints-based network analysis of metabolism. In a constraints-based model of *Escherichia coli*, they have shown that the predicted properties of the system, such as growth rate, depend on the proton exchange flux between the cell and the medium [17]. The most striking demonstration of the importance of charge balance comes from studies of the *S. cerevisiae* constraints-based model, whereby the pathways for the electron transfer system (ETS) in mitochondria are included in the electron balance [18••]. This balance revealed the direct coupling of metabolic pathways with the respiratory chain. One of the most interesting results of these studies is that the energetic efficiency of the ETS was estimated to be equivalent to a P/O ratio, that is, the relative amount of ATP molecules produced per pair of electrons donated to the ETS, of approximately 1. This is very close to previously estimated P/O ratios from physiology studies. This study is the first to demonstrate the coupling of cellular biochemistry with electron transfer energetic efficiency.

One of the challenges facing future study and development in this area is the connection between transcriptional and translational regulation; this will determine the presence or absence of a reaction in the network, and will determine the exact constraints of the system under different environmental and genetic conditions [14]. Although constraints-based approaches offer the best framework for the study of metabolic biochemistry, at the moment they are limited to the current knowledge of metabolic biochemistry and do not allow the investigation of the potential for novel biochemical connections through reactions of known or new metabolites.

Pathway biochemistry and protein structure

Several recent studies in structural biology have focused on the relationship between enzyme function and enzyme sequence and structure. One of the main objectives of these studies has been the investigation of alternative proposed mechanisms of pathway evolution [19]. A common theme of these studies is the comparison of both sequence identity and protein structure similarity with enzyme functionality, which is assigned based on the Enzyme Commission (EC) classification scheme [20]. According to the EC classification scheme for biochemical reactions, each enzyme is assigned a four-digit EC number (i.j.k.l). The first level of classification (i) denotes the type of chemistry: oxidoreductases comprise class 1, transferases comprise class 2, hydrolases comprise class 3, lyases comprise class 4, isomerases comprise class 5 and ligases comprise class 6. The second level (j) denotes the specific functional group acted on by the enzyme. For example, 1.1 denotes oxidoreductases acting on the CH–OH group of donors. The third level of classification (k) denotes the cofactors used by the enzyme. For example, 1.1.1 denotes oxidoreductases acting on CH–OH groups using NAD^+ or $NADP^+$ as acceptors. The fourth level (l) is specific to the substrates and products of the reaction. For example, enzyme 1.1.1.1 acts on primary alcohols, whereas 1.1.1.3 acts on homoserine. Thus, the first three digits (i.j.k) define the functional group transformations of the enzyme-catalyzed reaction, and all enzymes with the same designation should be consistent with one another if classified correctly; the fourth is a unique substrate identifier.

Early studies suggested that greater than 40% sequence identity between two enzymes could imply that these enzymes perform exactly the same function [21,22]. In a recent study, Tian and Skolnick [23••] revisited this problem; they found that a 40% global sequence identity threshold could imply similar function with respect to the third level EC class, but greater than 60% global identity is required for similar function with respect to the fourth EC level. They also highlight cases in which the EC classification number does not always reflect the exact chemistry performed by the corresponding enzyme and suggest that attention must be paid when computational approaches are used without consideration of these misclassifications.

All of these studies are aimed at the ultimate identification of protein function based on sequence and structure

similarity and they demonstrate that there exists a relationship between protein structure and biochemical function. It is also possible to interpret the findings about third level functional similarity by considering that enzymes do not act on the entire substrate molecule but rather on a specific active site. Hence, the third EC level corresponds to enzyme action on characteristic sites of a metabolite, not on the entire structure of a metabolite. Therefore, it appears that a 40% sequence identity represents enzyme specificity for active sites and therefore for the specific chemical action, rather than for the entire metabolite. The requirement of significantly higher identity for fourth level functional similarity can be attributed to the specificity of the enzyme for the global structure of the metabolite.

Pathway biochemistry and metabolome structure

The chemistry of metabolic reactions has recently been the subject of several studies that examine SMM through the structure and properties of the metabolites. These studies involve approaches that are commonly used in organic chemistry for the identification of structural similarity between organic compounds. In most of these studies, metabolites are represented as graphs, with atoms corresponding to the nodes and bonds corresponding to the edges of the graph. Comparison of metabolite structures can be then reduced to comparing two graphs and searching for the maximum number and size of common subgraphs (i.e. common substructures).

One of the earliest studies of metabolic biochemistry using metabolite structure was focused on metabolic reconstruction, that is, the identification of new reactions that can connect existing metabolites in a metabolic network [24]. This approach identifies similar metabolites based on the identification of common subgraphs and postulates possible reaction connectivities between them based on a list of 16 reaction rules common to enzymatic reactions. For example, a dehydratase reaction is postulated between two similar metabolites if a double carbon–carbon bond in one metabolite appears to come from the removal of a hydroxyl group and a hydrogen atom from the corresponding carbon atoms in a second metabolite. More recently, an alternative but similar approach has also been introduced for metabolic reconstruction based on known metabolites and enzymatic reactions [25]. In this approach, one derives a list of rules for atom mapping between metabolites involved in every reaction in a metabolic network based on the known biochemistry. However, these rules are not based on biochemical knowledge, but rather they are simply sets of atom and bond changes that do not take into account the environment of the atoms. Therefore, the number of such rules is large (6191 rules for a metabolic network of 2920 compounds and 3737 reactions), and the generated reconstructed pathways might be thermodynamically infeasible.

Two recent studies have examined metabolism through the diversity of the chemical structures in metabolic pathways [26[sup]••[/sup],27[sup]••[/sup]]. Borrowing concepts from protein classification based on hierarchical structural classification, these approaches introduce a hierarchical structural classification of metabolites. These approaches are based on the sound assumption that ''the restrictions imposed on proteins by evolution must be embedded in structural features that are shared among subsets of metabolites'' [26••]. In both of these studies, structural groups that are common in known metabolites and that form the building blocks of their overall chemical structure are identified first. Metabolites are then classified according to their similarity, based on the number of common structural groups they share and the relative arrangement of these groups in the overall molecular structure of the metabolites.

Nobeli *et al.* [26••] classified the metabolome, that is, the full complement of metabolites of an organism, according to molecule fragments they selected based on visual examination of the two-dimensional structure of a metabolite data set. They then assigned a similarity score to the metabolites based on the number of common fragments. Hattori *et al.* [27••] used a set of 'atom types' that classify the same atoms, such as carbon, according to their environment, such as types of bonds connected to the atom, types of adjacent atoms and participation in a ring. Based on the number of common fragments [26••] or common subgraphs defined by the atom types [27••], a similarity score was then assigned between metabolites, and metabolites were clustered according to their similarity. Both approaches identified some common major clusters, such as carbohydrate and amino acid like molecules, suggesting that, despite the diversity of metabolites, the molecular scaffolds are limited and diversity arises from the combinatorial nature of biochemical action.

Kanehisa and colleagues [27••] went further to investigate the relationship between metabolite chemical structure and metabolic pathway organization. They discovered that compounds that belong to the same similarity cluster (i.e. that share high similarity) also belong to the same metabolic maps in the KEGG database and are likely to be linked through sequential reaction steps. In a similar case study, Nobeli *et al.* [26••] also found that the metabolites in the TCA cycle belong to a small number of clusters and the similarity of connected metabolites is very high. These findings demonstrate the biochemical organization of metabolic pathways in which enzymes operate on specific active sites on their substrate metabolites and, through a sequence of small changes in the substrate/product similarity, create a diverse array of non-similar metabolites. Moreover, it also appears that enzymes in pathway modules with similar metabolites correspond to genes clustered in operons and operon-like structures in bacteria.

Although most studies in this area focus on the development of efficient algorithms for the unsupervised identification of the conserved molecular groups that comprise the structure of the metabolites [28], they all point to the need for considering the information hidden in their molecular structure and the insight this information can provide for protein function prediction, and identification of the relationship between metabolite structure and protein function.

Conclusions: enzyme chemistry and metabolome structure

The approaches discussed above suggest the need for connecting enzyme chemistry with metabolome structure and vice versa. Such a connection will open an additional dimension in understanding and exploring the wealth of metabolic biochemistry and could also provide the link between sequence similarity and metabolite structure through enzyme functionality. Despite the recognized

Figure 2
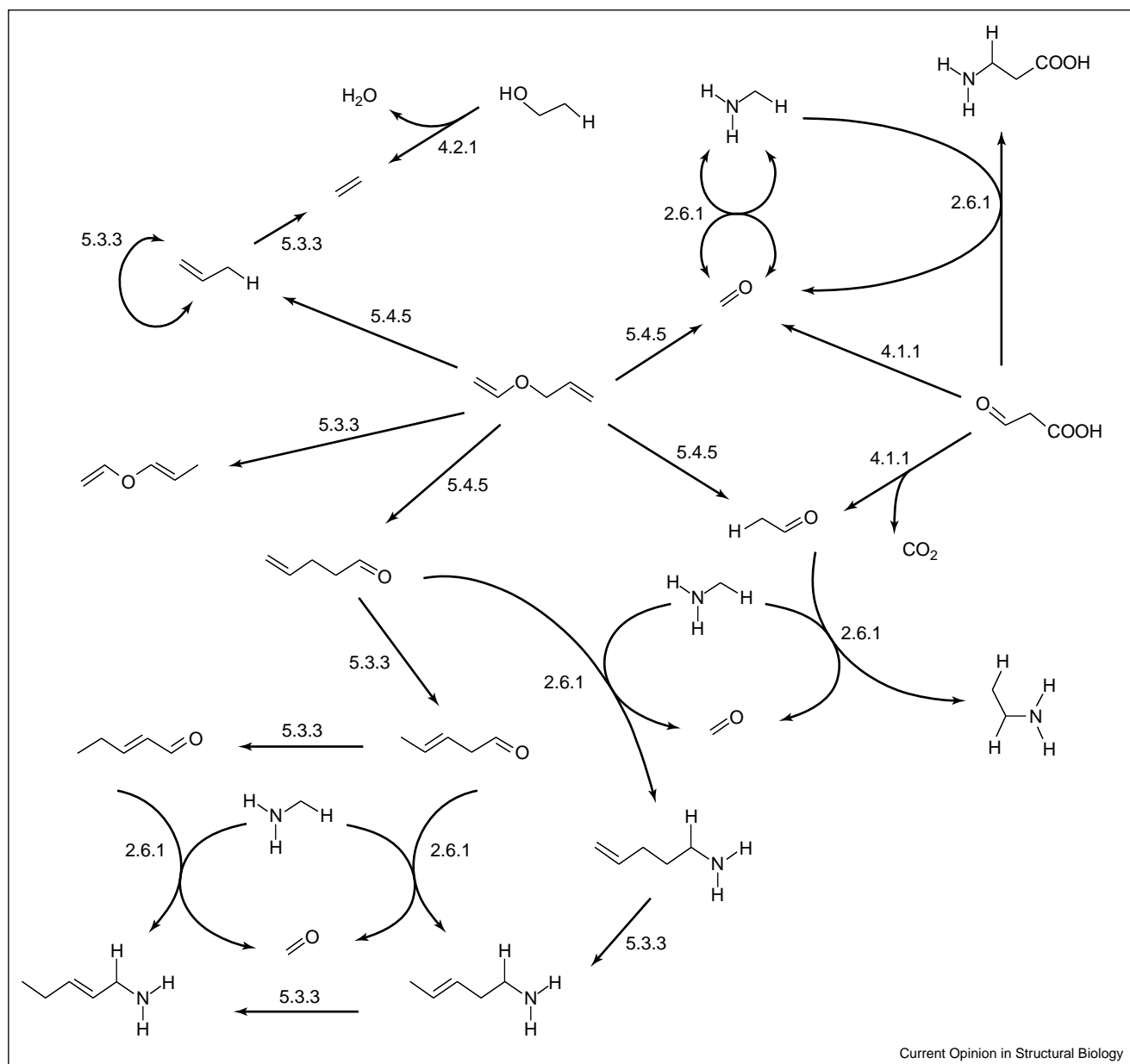
The elementary metabolic pathway of the enzyme actions in the phenylalanine pathway. The chemical structures shown are the metabolite substructures subject to enzyme transformation characteristic of the third level EC class indicated by the number next to the corresponding arrows. Some of the substructures that correspond to an enzyme action contain smaller substructures that can also serve as substrate structures for other enzyme actions in the network; some of these reactions have been included in the figure. Enzyme action 5.4.5 is a new third level class we introduced based on the reaction pattern of the enzyme chorismate mutase (EC 5.4.99.5) in the phenylalanine pathway.

limitation of the EC classification system, we believe that it offers an excellent starting point.

Enzymes are classified according to the active site or metabolite substructure they act upon or the metabolite substructure they produce. We can investigate the enzyme chemistry using the primary patterns of metabolite substructure changes identified within the fourth level enzymes of every third level class of the EC. The dominant pattern within the third level class will thus serve as the generalized enzyme action for that third level class, and we propose that these generalized enzyme actions be formulated as mathematical operators. For example, enzymes in class 4.2.1 are lyases (4) that break a carbon–oxygen bond (4.2) and release a water molecule (4.2.1). Investigation of the chemistry of the enzymes in the fourth level of the 4.2.1 class suggests that 37 out of 91 enzymes perform exactly the same active site transformation. Moreover, 19 of these enzymes can be described by the same active site transformation plus an additional enol transformation operator. The remaining 35 enzymes within the 4.2.1 class appear to be inconsistent with this generalized enzyme action, suggesting the misclassification of these enzymes.

This approach provides a systematic way to bridge enzyme chemistry and active sites on metabolites. It can also serve as a way to reconcile misclassifications and inconsistencies in the current EC classification, and to expand, in a systematic manner, the existing enzyme classes. Our investigation of currently available enzymes in KEGG suggested that this approach will allow us to formulate a set of less than 250 reaction rules. A large set of enzymes that can be described by one of these rules comprise the .99 third level enzymes. These are enzymes whose chemistry is consistent with the first two levels of the classification scheme, but do not have a unique third level action associated with their chemistry. There currently exist 29 .99 third level classes in the EC database, with a total of 202 (fourth class/specific) enzymes. We established that, within the .99 third level classes, many enzymes follow a reaction pattern that could be described using a unique mathematical operator.

It would be interesting to revisit studies of the relationship between protein structure and function [23••] with the reclassified set of EC classes. We can also use the active sites that are acted upon by the different third level classes to define groups for chemical structure comparison and study of the similarity of metabolome. This will provide an enzyme reaction description of metabolites and enable the monitoring of similarity change along a pathway consistent with the metabolic biochemistry. Ultimately, we can define the 'elementary metabolic pathways' in which active sites are the substrates of the generalized enzyme actions (Figure 2); this will provide an elementary description of metabolism and allow the generation of novel metabolites and pathways.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Galperin MY: **The molecular biology database collection: 2004 update**. *Nucleic Acids Res* 2004, **32**:D3-D22.

2. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome**. *Nucleic Acids Res* 2004, **32**:D277-D280.

3. Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen KB, Bairoch A, Schomburg D, Tipton KF, Apweiler R: **IntEnz, the integrated relational enzyme database**. *Nucleic Acids Res* 2004, **32**:D434-D437.

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

5. Krieger CJ, Zhang PF, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD: **MetaCyc: a multiorganism database of metabolic pathways and enzymes**. *Nucleic Acids Res* 2004, **32**:D438-D442.

6. Ma HW, Zeng AP: **Reconstruction of metabolic networks from
•• genome data and analysis of their global structure for various organisms**. *Bioinformatics* 2003, **19**:270-277.
This article presents a network analysis of metabolic networks in various organisms and identifies the importance of discriminating between core metabolites and cofactors in the analysis.

7. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks**. *Nature* 2000, **407**:651-654.

8. Price ND, Papin JA, Schilling CH, Palsson BO: **Genome-scale microbial *in silico* models: the constraints-based approach**. *Trends Biotechnol* 2003, **21**:162-169.

9. Schuster S, Hilgetag C: **On elementary flux modes in biochemical reaction systems**. *J Biol Syst* 1994, **2**:165-182.

10. Schuster S, Dandekar T, Fell DA: **Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering**. *Trends Biotechnol* 1999, **17**:53-60.

11. Schuster S, Fell DA, Dandekar T: **A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks**. *Nat Biotechnol* 2000, **18**:326-332.

12. Schilling CH, Palsson BO: **The underlying pathway structure of biochemical reaction networks**. *Proc Natl Acad Sci USA* 1998, **95**:4193-4198.

13. Schilling CH, Letscher D, Palsson BO: **Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from? A pathway-oriented perspective**. *J Theor Biol* 2000, **203**:229-248.

14. Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BO: **Metabolic pathways in the post-genome era**. *Trends Biochem Sci* 2003, **28**:250-258.

15. Forster J, Famili I, Fu P, Palsson BO, Nielsen J: **Genome-scale
• reconstruction of the *Saccharomyces cerevisiae* metabolic network**. *Genome Res* 2003, **13**:244-253.

This paper describes the first stoichiometric model of a model eukaryotic organism. This stoichiometric model could serve as a platform for the development of stoichiometric models of higher organisms.

16. Forster J, Famili I, Palsson BO, Nielsen J: **Large-scale evaluation of *in silico* gene deletions in *Saccharomyces cerevisiae***. *OMICS* 2003, **7**:193-202.

17. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR)**. *Genome Biology* 2003, **4**:R54.

18. Famili I, Forster J, Nielson J, Palsson BO: **Saccharomyces**
•• **cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network**. *Proc Natl Acad Sci USA* 2003, **100**:13134-13139.
The authors investigate the physiology of *S. cerevisiae* using a stoichiometric metabolic model of this organism. The model captures extremely well the physiology of the organism in aerobic and anaerobic chemostat conditions. The model includes the ETS with its exact stoichiometry and predicts a P/O ratio of around 1, which is in the range of the experimentally observed values. These results demonstrate the importance of charge balance when considering the stoichiometric and network analysis of metabolic pathways.

19. Rison SCG, Thornton JM: **Pathway evolution, structurally speaking**. *Curr Opin Struct Biol* 2002, **12**:374-382.

20. Tipton K, Boyce S: **History of the enzyme nomenclature system**. *Bioinformatics* 2000, **16**:34-40.

21. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective**. *J Mol Biol* 2001, **307**:1113-1143.

22. Wilson CA, Kreychman J, Gerstein M: **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores**. *J Mol Biol* 2000, **297**:233-249.

23. Tian WD, Skolnick J: **How well is enzyme function conserved as**
•• **a function of pairwise sequence identity?** *J Mol Biol* 2003, **333**:863-882.
The authors examine the relationship between sequence identity and enzyme function according to the EC classification scheme. They find that 60% sequence identity between two enzymes suggests that the enzymes are functionally similar with respect to their fourth level EC classification and that 40% sequence identity suggests functional similarity with respect to the third level EC classification.

24. Arita M: **Metabolic reconstruction using shortest paths**. *Simul Pract Theory* 2000, **8**:109-125.

25. Boyer F, Viari A: ***Ab initio* reconstruction of metabolic pathways**. *Bioinformatics* 2003, **19(suppl 2)**:II26-II34.

26. Nobeli I, Ponstingl H, Krissinel EB, Thornton JM: **A structure-**
•• **based anatomy of the *E. coli* metabolome**. *J Mol Biol* 2003, **334**:697-719.
See annotation to [27••].

27. Hattori M, Okuno Y, Goto S, Kanehisa M: **Development of a**
•• **chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways**. *J Am Chem Soc* 2003, **125**:11853-11865.
These two papers [26••,27••] present a novel approach to analyzing metabolic pathways based on the structure of metabolites. Characteristic substructures of metabolites are identified and used for the quantification of their similarity. Clustering of the metabolites according to their similarity reveals the relationship between metabolic pathways and metabolite structural similarity.

28. Akutsu T: **Efficient extraction of mapping rules of atoms from enzymatic reaction data**. In *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology; 2003* April 10–13; Berlin. New York: ACM Press; 2003:1-8.