BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 8 2005, pages 1603–1609
doi:10.1093/bioinformatics/bti213

Systems biology

Exploring the diversity of complex metabolic networks

Vassily Hatzimanikatis*, Chunhui Li, Justin A. Ionita, Christopher S. Henry, Matthew D. Jankowski and Linda J. Broadbelt*

Department of Chemical and Biological Engineering, McCormick School of Engineering and Applied Sciences, Northwestern University, Evanston, IL, USA

Received on August 30, 2004; revised on November 16, 2004; accepted on December 8, 2004
Advance Access publication January 4, 2005

ABSTRACT

Motivation: Metabolism, the network of chemical reactions that make life possible, is one of the most complex processes in nature. We describe here the development of a computational approach for the identification of every possible biochemical reaction from a given set of enzyme reaction rules that allows the *de novo* synthesis of metabolic pathways composed of these reactions, and the evaluation of these novel pathways with respect to their thermodynamic properties.

Results: We applied this framework to the analysis of the aromatic amino acid pathways and discovered almost 75 000 novel biochemical routes from chorismate to phenylalanine, more than 350 000 from chorismate to tyrosine, but only 13 from chorismate to tryptophan. Thermodynamic analysis of these pathways suggests that the native pathways are thermodynamically more favorable than the alternative possible pathways. The pathways generated involve compounds that exist in biological databases, as well as compounds that exist in chemical databases and novel compounds, suggesting novel biochemical routes for these compounds and the existence of biochemical compounds that remain to be discovered or synthesized through enzyme and pathway engineering.

Availability: Framework will be available via web interface at http://systemsbiology.northwestern.edu/BNICE (site under construction).

Contact: vassily@northwestern.edu or broadbelt@northwestern.edu

Supplementary information: http://systemsbiology.northwestern.edu/BNICE/publications

INTRODUCTION

The study of metabolic networks entails two kinds of problems: the analysis and the synthesis. The analysis problem involves the study of a set of biochemical reactions and the identification of every possible pathway for the production of a biochemical compound from a given set of starting compounds through a set of known biochemical reactions and compounds. Computational approaches for this problem involve the use of artificial intelligence methods (Mavrovouniotis *et al*., 1992), stoichiometric analysis (Price *et al*., 2003; Schuster *et al*., 2000) and graph network analysis (Arita, 2000; Jeong *et al*., 2000).

The synthesis problem involves the identification of novel reactions and novel biochemical compounds that are feasible, given a set of enzyme reaction rules and a set of initial compounds and cofactors. One of the approaches to the synthesis problem employs knowledge-based expert systems for the identification of possible biodegradation pathways based on a set of rules constructed from observed biodegradation reactions (Darvas, 1988; Greene *et al*., 1999; Hou *et al*., 2003; Klopman *et al*., 1994, 1999; Talafous *et al*., 1994). This approach predicts novel biodegradation routes, but it does not provide a systematic framework for application to other metabolic pathways that involve more than 4000 biochemical reactions (Kanehisa *et al*., 2004). Another approach considers the identification of possible new reactions that can connect the existing metabolites in a metabolic network (Arita, 2000), but it does not allow the identification of possible new biochemical pathways through novel biochemical compounds.

On the other hand, computational chemistry methodologies allow the generation of every possible reaction for a given set of chemical reaction rules and starting compounds (Ihlenfeldt and Gasteiger, 1996). We adopted one of these methods (Broadbelt *et al*., 1994) and developed a computational framework that we have named BNICE (*B*iochemical *N*etwork *I*ntegrated *C*omputational *E*xplorer), which can address the synthesis problem in metabolic pathways.

SYSTEMS AND METHODS

The framework requires a graph-theoretic matrix (Ugi *et al*., 1979) representation of biochemical compounds and enzyme reaction rules. Molecules are represented using the bond-electron matrix (BEM) (Ugi *et al*., 1979). Each atom in a molecule is represented by a row (or equivalently, a column). The diagonal elements, ii, of the BEM denote the non-bonded valence electrons of atom i; the non-diagonal elements, ij, give the connectivity via bonding between different atoms and the bond order between atoms i and j. Enzyme-catalyzed reactions can be represented using similar notation. The reactive sites for each enzyme class are pre-defined as two-dimensional (2D) molecule fragments and coded in the BNICE. A set of molecules is given as input and every molecule is evaluated to find if it has the appropriate functionality to undergo reactions corresponding to the specified reaction classes. The reactions are then implemented through matrix addition. Negative numbers in the reaction matrix correspond to the cleavage of bonds and positive numbers correspond to the formation of bonds. When the matrix representing the enzyme-catalyzed reaction is added to the BEM for the substrate, the BEM formed specifies the products of the reaction (Fig. 1).

The automated pathway generation algorithm operates in an iterative manner. Once reacted, all the reactants are placed in a 'reacted' list, and all the products from these reactants will be placed in an 'unreacted' list if they are molecules that have not been specified or generated before. Next, each molecule in the 'unreacted' list will be checked for its reactivity, and the reaction matrix operations will be repeated. An iteration count is maintained as new molecules are created, keeping track of the generation number of each

*To whom correspondence should be addressed.

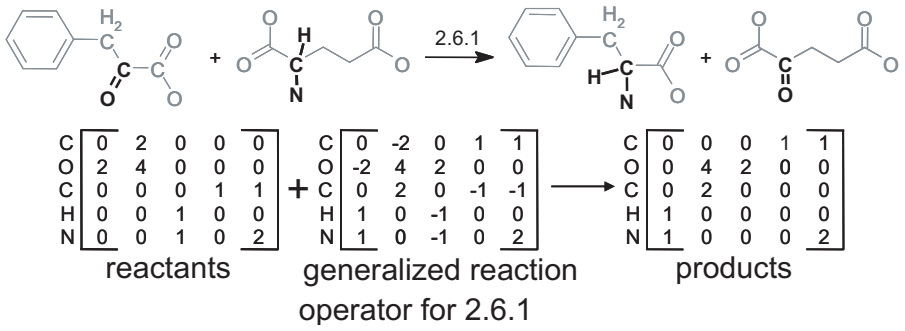

Fig. 1. Example of bond-electron matrix for the reacting portions of phenylpyruvate and glutamate (reactants) and phenylalanine and 2-oxoglutarate (products) along with the reaction operator matrix for the 2.6.1 generalized enzyme reaction. The reaction operator matrix is added to the BEM of the reactants to produce the BEM of the products. Negative number in the operator matrix represents bonds broken and positive number represents bonds created in the 2.6.1 reaction.

species, which is the number of steps required to create a given product from the original reactant(s). A maximum generation number can be specified, and thus the generation number can be used to determine if a given molecule from the 'unreacted' list is allowed to react in the next generation. If the generation number is above the specified maximum, the given molecule will be placed in the 'reacted' list and is a terminal point in the reaction network.

The BEMs of the input molecules are generated automatically after the 2D connectivity information is read by the program. All the reaction matrices for the enzyme-catalyzed reactions in the system under investigation are pre-formulated and stored.

Definitions of the generalized enzyme reactions

We utilized the Enzyme Commission (EC) classification system, developed by the International Commission on Enzymes (Tipton and Boyce, 2000), for a systematic formulation of these enzyme reaction rules. The EC established a classification scheme that involves a four-tier hierarchical classification: EC i.j.k.l. The first three numbers classify enzymes according to the chemical rules of the reaction they catalyze, and the fourth number corresponds to the participating substrates and products.

In the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa *et al*., 2004) there currently exist 4306 4th level reactions organized in 6 first level classes, 63 second level classes and 234 third level classes. We investigated all of the enzyme reactions in the reaction databases, and we found that the third level classification uniquely describes the chemistry associated with most of the corresponding enzymes in the fourth level. Based on this observation we defined the *generalized enzyme reaction*, which allowed the formulation of a unique matrix representation for every enzyme reaction within the same third level enzyme class. In order to define a generalized enzyme reaction based on the third level EC classes, we examined all the biochemical reactions within the corresponding fourth level. We can define the enzyme reaction by comparing the substrate and product structures. We also identified many cases of enzyme misclassification that could be reclassified based on the definition of the generalized enzyme reaction, and we introduced new third level classes. Therefore, the generalized enzyme reactions were defined according to the following four principles: (1) enzymes in the fourth level EC class follow a primary pattern of overall transformation that is consistent with the original classification rules used by the EC; (2) the overall reactions can be dissected using two or more different generalized enzyme reactions; (3) the enzyme reactions can be described by an operator different from the one that is characteristic of their third level classification but consistent with the reaction pattern of more than one biochemical reaction; (4) enzymes follow a unique reaction mechanism; in this case we use the pattern defined by the enzyme, and we classify it based on the possible chemistry or information about the mechanism of the enzyme. In addition, we augment the enzyme rules with known transformations that are consistent with organic chemistry rules or with non-enzymatic steps in enzyme-catalyzed mechanisms, such as keto-enol tautomerism and cyclizations. Applying these principles, we found that all the known enzyme-catalyzed reactions in KEGG database can be represented by fewer than 250 generalized enzyme reactions in place of the more than 4300 specific enzyme functions in the database, thus rendering the complexity of biological systems more manageable.

Generalized enzyme reactions in the aromatic amino acid pathways

We employed BNICE for the analysis of the biosynthetic pathways of the aromatic amino acids, phenylalanine, tyrosine and tryptophan, from chorismate and the cofactors and cosubstrates—glutamate glutamine, serine, NAD^+/NAD and 5-phospho-α-D-ribose-1-diphosphate (PRPP). The native pathways from chorismate to phenylalanine and tyrosine are each composed of three reactions, with two reactions common in both pathways, while the pathway to tryptophan is composed of five reactions. Out of these nine unique biochemical reactions, three can be mathematically described using a unique operator characteristic of their third level generalized enzyme reaction, and the other six can be described by combinations of third level generalized enzyme reactions (Supplementary information Tables S1 and S2).

The first enzyme, chorismate mutase (EC 5.4.99.5), in the pathway that leads to phenylalanine and tyrosine, performs a pericyclic reaction known as a Claisen rearrangement (Dosselaere and Vanderleyden, 2001). This is the only example of such an enzyme-catalyzed reaction, and we defined a new third level class generalized reaction, EC 5.4.4, for this reaction (Supplementary information Tables S1 and S2). In the phenylalanine pathway, prephenate dehydratase (EC 4.2.1.51) catalyzes a simultaneous dehydration, EC 4.2.1, and decarboxylation, EC 4.1.1. Thus, it would be consistent with both of these current third level EC classifications. Furthermore, if these two reactions are concurrent, the overall reaction requires an isomerase activity for double-bond rearrangement, characteristic of the third level class EC 5.3.3. We have therefore included all three of these generalized enzyme reactions in the phenylalanine pathway. Similarly, prephenate oxidoreductase (EC 1.3.1.12) in the tyrosine pathway was divided into three generalized enzyme reactions: EC 4.1.1, 5.3.3 and 1.3.1.

The first enzyme in the tryptophan biosynthetic pathway, chorismate pyruvate-lyase (EC 4.1.3.27), contains two subunits: alpha and beta (Miles, 2001). Subunit alpha is known to convert chorismate to anthranilate in two discrete steps. The first step converts chorismate to aminodeoxyisochorismate (ADIC) through a reversible amination reaction. The amino group comes from glutamine through the action of the beta subunit. The second step catalyzed by the alpha subunit converts ADIC to anthranilate through *cis*-elimination of pyruvate. Analysis of the functions of the two subunits led to the splitting of the overall reaction into three third level generalized enzyme reactions, EC 2.6.2, 4.2.4 and 5.3.2. EC 2.6.2 is a new third level transaminase class and corresponds to the first step catalyzed by the coordinated action of subunits, alpha and beta. EC 4.2.4 and 5.3.2 correspond to the second step

reaction catalyzed by the subunit alpha alone (Supplementary information Tables S1 and S2). N-(5-phospho-β-D-ribosyl)anthranilate ketol-isomerase (EC 5.3.1.24) is another interesting enzyme in the tryptophan pathway that does not appear to follow the generalized reaction characteristic of the 5.3.1 class. It catalyzes a tautomerization reaction known as an Amadori rearrangement (Creighton and Yanofsky, 1970; Isbell and Frush, 1958) for which we have defined the new class EC 5.3.5 (Supplementary information Tables S1 and S2).

With these decompositions and reclassifications, the three reaction steps in both the phenylalanine and tyrosine pathways were represented by five generalized enzyme reactions each, and the six reaction steps in the tryptophan biosynthetic pathway were represented by nine generalized enzyme reactions. Another major issue we resolved using the EC classification and database information as a guide was the participation of the cofactors and amino donors in the specified generalized enzyme reactions. For example, based on the observation that 41 of the 72 enzymes in the 2.6.1 class utilize glutamate as an amino donor, transaminase reactions, i.e. EC 2.6.1, were constrained to use glutamate as the sole amino-donor, despite the fact that the generalized enzyme reaction, EC 2.6.1, could permit every amino-carrier compound to act as an amino-donor indiscriminately. In addition, the cofactor compounds were not allowed to serve as substrates for reactions other than those where they serve as cofactors.

RESULTS AND DISCUSSION

Analysis of the aromatic amino acid pathways

The application of the three individual sets of generalized enzyme reactions to produce phenylalanine, tyrosine and tryptophan from chorismate and the corresponding cofactors and amino donors, resulted in a rich array of novel compounds and pathways. A total of 246 compounds were produced from chorismate and the five generalized enzyme reactions in the phenylalanine pathway; 289 compounds were produced from the five generalized enzyme reactions in the tyrosine pathway; 58 compounds were produced from chorismate and the nine generalized enzyme reactions in the tryptophan pathway (Table 1). Through automated comparison of the compounds created and those catalogued in the KEGG and the Chemical Abstracts Service (CAS) Registry, it was clear that the compounds fell into one of three categories: compounds that are a part of the original pathways, compounds that exist in the KEGG database and are not part of the original pathways and compounds that exist in the CAS database but not in the biological databases. The compounds that exist in the CAS but not the KEGG are intriguing compounds because there exists an organic synthesis route but not a reported biochemical synthesis route. A large number of compounds had no match in either of the two databases, suggesting the existence of compounds that remain to be discovered or synthesized through protein and pathway engineering (Table 1). Furthermore, the diversity of biochemical compounds was increased by at least one order of magnitude when we combined generalized enzyme reactions from two of the three different pathways, and the largest number of compounds—almost 35 000—was generated when we combined the 15 generalized enzyme reactions in all three pathways (Table 1).

Length distribution of the alternative biosynthetic routes

The generated metabolic networks were also analyzed with respect to the pathway length between chorismate and the individual aromatic amino acids. The pathway length between chorismate and the aromatic amino acid of interest in the novel biochemical reaction networks was established using a depth-first search algorithm. All reactions were assumed to be irreversible, proceeding in the direction defined by the generalized reaction rules, and all cofactors were removed from the network before the search was conducted since they do not act as logical intermediates in any of the pathways. The algorithm starts from the node corresponding to chorismate and performs a depth-first search of the network from this node until the target node (e.g. phenylalanine) is found. It reports this path length and backtracks to explore all the branching points in the pathway. If a node is ever encountered twice in the same pathway, the algorithm would immediately backtrack without logging the pathway to guarantee that no pathways with loops are counted.

We found that there exist 7, 23 and 15 alternative pathways from chorismate to phenylalanine, tyrosine and tryptophan, respectively, with the same number of generalized enzyme reactions as in the original pathways (Fig. 2 and Supplementary information Figure S3). Moreover, there exists a large number of pathways with lengths longer than the length of the native pathways in the cases of phenylalanine and tyrosine, exemplifying the flexible nature of the generalized chemistry, but there are no pathways of higher length than the original pathway in the tryptophan pathway (Fig. 3).

Regardless of length, multiple pathways of the same length are the result of the combinatorial nature of the biochemistry involved that allows the same overall chemistry to be performed in different sequences through different intermediates leading to the same products. However, combinatorics alone are a poor predictor of the viable pathways given a set of enzyme reactions. For example, if the five reactions comprising the native pathway from chorismate to phenylalanine could occur in any sequence, we would expect to have 120 (5!) different pathways of length five. However, we found only seven due to the conditional nature of the sequence of chemical reactions. This aptly demonstrates the usefulness of our computational framework that takes into account the dependency of each chemical

Table 1. Compounds and reactions in the three metabolic networks

	Phe[a]	Tyr[a]	Trp[a]	AAA[b]	Phe/Tyr[b]	Phe/Trp[b]	Tyr/Trp[b]
Total reactions	946	1318	80	98985	1601	19063	36988
Total compounds	246	289	58	34892	356	8252	13775
Exact matches in KEGG	9	8	19	39	19	34	28
(Percent total)	(3.66)	(2.77)	(33)	(0.112)	(5.34)	(0.247)	(0.339)
Exact matches in CAS	17	16	25	95	35	81	60
(Percent total)	(6.91)	(5.54)	(43)	(0.272)	(9.83)	(0.588)	(0.727)
Near matches in CAS[c]	55	51	1	181	67	156	106
(Percent total)	(22.4)	(17.6)	(1.7)	(0.519)	(18.8)	(1.13)	(1.28)
Novel compounds	229	273	33	34797	321	13694	8146
(Percent total)	(93.1)	(94.5)	(57)	(99.73)	(90.2)	(99.4)	(98.7)

[a]No new compounds were created after 10 iterations of the rules for phenylalanine and tyrosine and 14 iterations for tryptophan.
[b]Data obtained from 10 iterations of the reaction rules only.
[c]Near matches consist of tautomers and zwitterions. Tautomers are isomers that are rapidly inconverted. Often a double bond is shifted in the structure, such as in keto-enol tautomers. Zwitterions are molecules that are both cationic or anionic. Near matches in the CAS differ by either charge or a small structure feature such as a double bond shift.

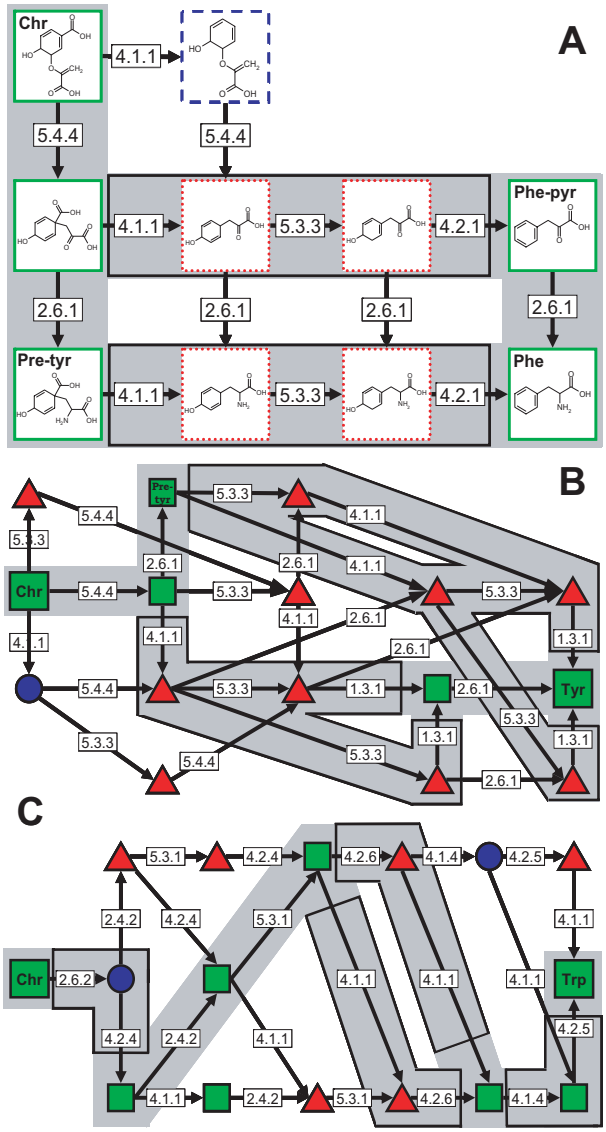

Fig. 2. Alternative native-length pathways to aromatic amino acids. All are pathways from chorismate to phenylalanine (**A**), tyrosine (**B**) and tryptophan (**C**) with length equal to the length of the original pathway. Chr, chorismate; Phe, phenylalanine; Tyr, tyrosine; Trp, tryptophan. Solid-line boxes (A) or squares (B and C) indicate compounds that exist in the KEGG database, dashed line boxes (A) or circles (B and C) indicate compounds that exist in the CAS database, and dotted-line boxes (A) or triangles (B and C) indicate novel compounds not found in KEGG or CAS databases. The structures of the compounds in the tyrosine (B) and tryptophan (C) pathways are provided in Figure S3 in the Supplementary information. Shaded pathways indicate the pathways that correspond to the native pathways; portions of the shaded native pathway outlined in black represent dissected reactions.

reaction on the history of the preceding chemical reactions, and the specific rules that govern enzyme chemistry where specific molecular substructures may be required in addition the target reactive site for the reactions to take place. The conditional nature of the sequence and specificity of the enzyme reactions is further exemplified in the tryptophan pathway, where we might expect 40 320 pathways of length eight, but only 15 pathways are generated.

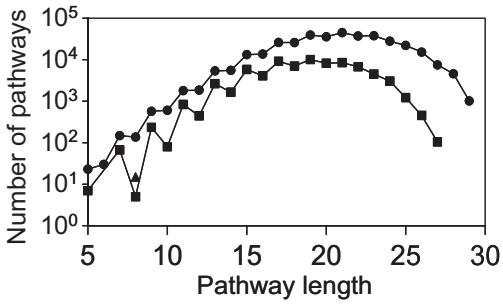

Fig. 3. Pathway length distribution in the novel reaction networks of the three aromatic amino acids: phenylalanine (squares), tyrosine (circles) and tryptophan (triangles). In phenylalanine there are no pathways with an even length until pathways reach a length of eight. This is due to the fact that the repetitive double bond shifting of the 5.3.3 transformation is responsible for increasing most of the pathways beyond the native length of 5. For paths of length less than eight, the structure of the substrates does not allow 5.3.3 to shift bonds to the configuration needed to produce phenylalanine in an even number of steps.

Network properties

In order to study the cluster organization of the predicted pathways and identify the 'hubs', i.e. highly connected metabolites in the networks, we used cluster analysis and visualization (Batagelj and Mrvar, 2002; Jeong *et al.*, 2000). The phenylalanine and tyrosine networks display a characteristic structure with two large clusters (Fig. 4A and B). One of the clusters consists of carbon compounds and the other consists of the corresponding aminated compounds. The cofactors, amino donors and small molecules, such as NAD^+/NADH, glutamate, CO_2 and H_2O, are the 'hubs' of the networks as already observed by Jeong *et al.* (2000). The tryptophan network appears to be sparse due to the small number of biochemical compounds with small molecules and cofactors as hubs (Fig. 4C).

Thermodynamic feasibility studies

The predicted pathways were further evaluated with respect to their thermodynamic feasibility. We used a group contribution method previously developed for the estimation of the Gibbs free energy of formation of biochemical compounds and the subsequent calculation of the standard Gibbs free energy of reaction (Mavrovouniotis, 1990, 1991). This method dissects a molecule into several groups of atoms and treats the thermodynamic properties of the whole molecule as the summation of the properties of these groups. The properties of these basic groups can be obtained from a training set of molecules with known thermodynamic data using a multiple linear regression analysis. If the set of groups obtained from molecules with known thermodynamic properties is comprehensive enough, group contribution methods can be used to calculate the thermodynamic properties of any biological molecule. Estimation of the Gibbs free energy of the reactants and products of an enzymatically catalyzed reaction then allows the estimation of the Gibbs free energy of the reaction. The Gibbs free energy contribution of cofactor transformations, such as NADH and ATP, is also added in reactions that involve cofactors.

We found that the native pathways follow the most thermodynamically favorable path relative to the alternative pathways with the same number of reaction steps for all three aromatic amino

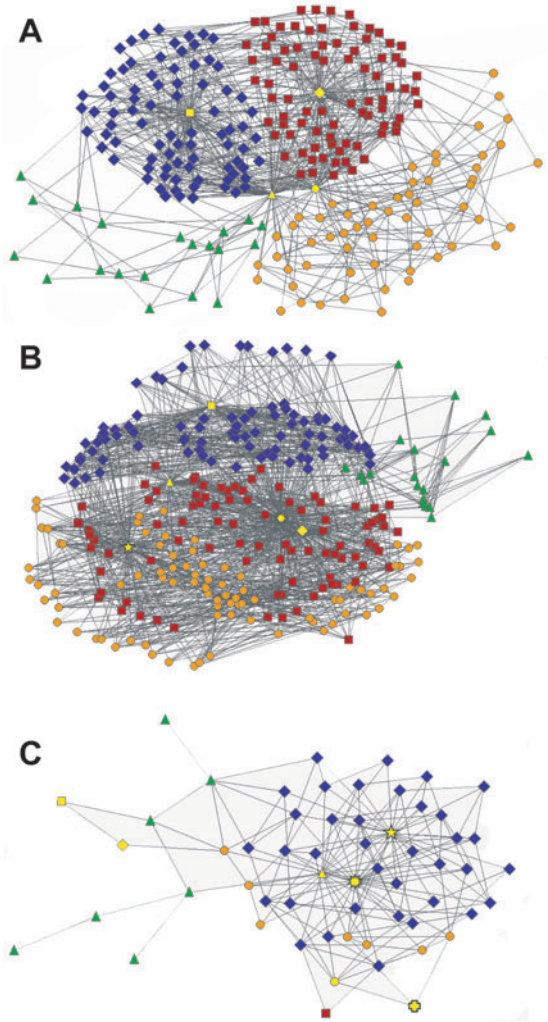

Fig. 4. Visual representation of the novel biochemical reactions and compounds generated from the generalized enzyme reactions. Nodes represent compounds and edges represent reactions. In the phenylalanine (**A**) and tyrosine (**B**) networks, carbon compounds (blue diamonds) act as amino acceptors in the 2.6.1 reaction producing amino compounds (red squares). Some carbon compounds (green triangles) cannot undergo transamination, whereas another large group of amino compounds (orange circles) is produced through further processing of the products of the 2.6.1 reaction. Glutamate (yellow square), 2-oxoglutarate (yellow diamond), CO_2 (yellow triangle), H_2O (yellow circle), NAD^+ (yellow star) and NADH (yellow asterisk) all act as hubs in the networks. The tryptophan network (**C**) contains carbon compounds (green triangle), amino compounds (orange circles), carbon compounds with a phosphate group (red squares) and amino compounds with a phosphate group (blue diamonds). PRPP (yellow star), pyrophosphate (yellow asterisk) and CO_2 (yellow triangle) are the hubs of the network, whereas H_2O (yellow circle), serine (yellow cross), glutamate (yellow square) and glutamine (yellow diamond), are not highly connected in this network. The network visualization was performed using the network analysis package Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

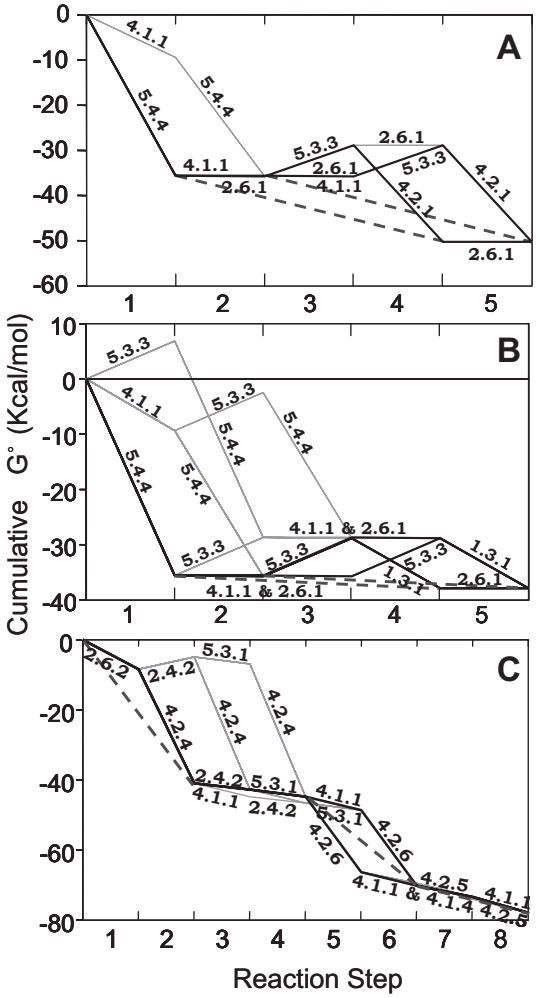

Fig. 5. Thermodynamic free energy landscape of alternative pathways. The changes in the reaction Gibbs free energy along the alternative pathways from chorismate to phenylalanine (**A**), tyrosine (**B**) and tryptophan (**C**). Only the free energy landscapes of pathways with length equal to the length of the original pathways are represented. Energy landscapes for longer pathways are shown in Figure S4 in the Supplementary information. Dashed lines denote the free energy change of the overall reaction in the native pathways prior to dissection into generalized enzyme reactions (Supplementary information Table S1).

acids (Fig. 5). The original pathways were also found to be the most thermodynamically favorable compared with any other longer alternative pathways (Supplementary information Figure S4), further suggesting the possible evolutionary design of these pathways toward thermodynamic efficiency. This might imply that the currently known native pathways have evolved under the pressure of thermodynamic constraints.

An interesting feature of the free energy landscapes of the native pathways is that certain transformations that appear to be thermodynamically unfavorable, i.e. $\Delta G^{\circ}_{rxn} > 0$, are components of the decomposed steps that appear to take place on multifunctional enzymes, suggesting that these enzymes have evolved to perform channelling reactions that pass the reaction intermediates to the next reaction site without releasing them into solution. Channelling, effectively, increases the apparent local concentration thereby driving the reaction forward by changing the reaction potential through apparent high concentration. In addition, the thermodynamic properties estimated here are for standard conditions ($T = 25°C$;

$P = 1$ atm; pH $= 7$; concentrations of reactants and products 1 M; zero ionic strength), and they might change under different conditions, as is particularly the case in extremophiles. While the potential chemical transformations will remain the same, it is possible that the different environmental conditions might favor different pathways in different organisms.

Concluding remarks

As the diversity of the sequenced organisms from extreme environments is increasing, novel and alternative biochemical pathways are expected (Falkowski and Vargas, 2004; Venter *et al*., 2004). The studies reported here display the wealth of biochemical reactions and biochemical compounds still to be discovered. While most of the compounds and pathways found through our computational studies have not been reported in the biochemical literature, they are feasible based on the concept of generalized enzyme reaction and the thermodynamics at standard conditions. Therefore, they might be already operative in nature, but we have not yet detected them.

Implementation of the predicted metabolic pathways in an organism will require that their mass, energy and redox demands be met by the metabolism of the host organism. The cellular feasibility of the predicted pathways can be further evaluated by implementing a constraints-based approach (Price *et al*., 2003). The mass balance equations of the predicted metabolites will provide the stoichiometric constraints that will augment the stoichiometric constraints of the host metabolic networks and will allow us to perform constraints-based studies for estimating important physiology aspects such as maximum theoretical yield of the novel metabolites, cost of their production on cell growth and optimization of process conditions.

In the case of the aromatic amino acid biochemistry the number of the metabolites and reactions converged after a relatively small number of successive iterations, with a maximum of 12 steps in the case of the tryptophan pathway. However, depending on the number and kind of the reaction rules and initial metabolites, the number of reactions and metabolites might grow exponentially. As new metabolites and reactions are being generated we can also control the growth rate of the reactions and metabolites by applying thermodynamic and cellular feasibility criteria, as well as other system-specific criteria, such as the maximum number of carbon atoms in each metabolite.

The developed framework can be applied to a large number of different systems of biotechnological importance. For example, the isoprenoid and polyketide synthesis pathways have been shown to produce an enormous number of structurally different compounds and their diversity has been expanded through metabolic engineering (Khosla and Keasling, 2003). Application of the computational framework will allow the identification of every possible chemical compound that can be produced by the generalized enzyme reactions in these pathways and the computational screening of compounds for potential antimicrobial activity will provide guidance for design of biosynthetic pathways for these compounds.

The application of the framework to the biochemistry of the central carbon pathways will also allow the identification of potential novel routes to important metabolic and biosynthetic compounds, or the synthesis of new compounds, based on renewable resources. For example, in the work reported in the paper we have identified that at least one of the chemicals predicted from the aromatic amino acid biochemistry is a valuable industrial chemical suggesting a possible new biosynthetic route for an industrial chemical. Such findings can have significant implications for the development of bioprocesses for industrial chemicals in a sustainable future technology (Lynd *et al*., 1999).

Another appropriate area of application is bioremediation of organic xenobiotics. Many of these manmade chemicals are compounds that have never been seen before by nature and application of the framework presented here will assist in investigating if the generalized enzyme reactions, present in the known biodegradation pathways could degrade them and suggest the possible biodegradation routes and products (Greene *et al*., 1999; Hou *et al*., 2003).

The systematic reclassification of the enzymes could also provide new insight into the correlation between enzyme structure and function (Rison and Thornton, 2002). Similar considerations might open new ways of addressing the problem of pathway reconstruction (Osterman and Overbeek, 2003) by suggesting alternative connections that close the gaps in pathways from organisms with missing enzymes. In addition, the approach presented here could provide a framework for the studies of the evolution of metabolism (Morowitz *et al*., 2000; Schmidt *et al*., 2003). Mining the wealth of the combinatorial biochemistry will require the development and application of approaches like the one described here.

ACKNOWLEDGEMENTS

We thank J. Widom, I. Radhakrishna, L. Shea, D. Young and R. Silverman for critically reviewing the manuscript prior to submission. The work is supported by the US Department of Energy, Genomes to Life Program.

REFERENCES

Arita,M. (2000) Metabolic reconstruction using shortest paths. *Simulat. Pract. Theory*, **8**, 109–125.

Batagelj,V. and Mrvar,A. (2002) Pajek—Analysis and visualization of large networks. *LNCS*, **2265**, pp. 477–478.

Broadbelt,L.J., Stark,S.M. and Klein,M.T. (1994) Computer generated pyrolysis modelling: on-the-fly generation of reactions, rates and predictions. *Ind. Eng. Chem. Res.*, **33**, 790–799.

Creighton,T.E. and Yanofsky,C. (1970) Chorismate to tryptophan (*Escherichia coli*). Anthranilate synthetase, N-5′-phosphoribosyltransferase, N-5′-phosphoribosylanthranilate isomerase, indole-3-glycerol phosphate synthetase, tryptophan synthetase. *Methods Enzymol.*, **17**, 365–380.

Darvas,F. (1988) Predicting metabolic pathways by logic programming. *J. Mol. Graph.*, **6**, 80–86.

Dosselaere,F. and Vanderleyden,J. (2001) A metabolic node in action: chorismate-utilizing enzymes in microorganisms. *Crit. Rev. Microbiol.*, **27**, 75–131.

Falkowski,P.G. and de Vargas,C. (2004) Shotgun sequencing in the sea: a blast from the past? *Science*, **304**, 58–60.

Greene,N., Judson,P.N., Langowski,J.J. and Marchant,C.A. (1999) Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. *SAR QSAR Environ. Res.*, **10**, 299–314, 292 plates.

Hou,B.K., Wackett,L.P. and Ellis,L.B.M. (2003) Microbial pathway prediction: a functional group approach. *J. Chem. Inf. Comput. Sci.*, **43**, 1051–1057.

Ihlenfeldt,W.-D. and Gasteiger,J. (1996) Computer-assisted planning of organic syntheses: the second generation of programs. *Angew. Chem. Int. Ed. Engl.*, **34**, 2613–2633.

Isbell,H.S. and Frush,H.L. (1958) Mutarotation, hydrolysis, and rearrangement reactions of glycosylamines. *J. Org. Chem.*, **23**, 1309–1319.

Jeong,H., Tombor,B., Albert,R., Oltval,Z.N. and Barabasi,A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Khosla,C. and Keasling,J.D. (2003) Metabolic engineering for drug discovery and development. *Nat. Rev. Drug Discov.*, **2**, 1019–1025.

Klopman,G., Dimayuga,M. and Talafous,J. (1994) META. 1. A program for the evaluation of metabolic transformation of chemicals. *J. Chem. Inf. Comput. Sci.*, **34**, 1320–1325.

Klopman,G., Tu,M. and Fan,B.T. (1999) META. Part 4. Prediction of the metabolism of polycyclic aromatic hydrocarbons. *Theoret. Chem. Accounts*, **102**, 33–38.

Lynd,L.R., Wyman,C.E. and Gerngross,T.U. (1999) Biocommodity engineering. *Biotechnol. Prog.*, **15**, 777–793.

Mavrovouniotis,M.L. (1990) Group contributions for estimating standard Gibbs energies of formation of biochemical-compounds in aqueous-solution. *Biotechnol. Bioeng.*, **36**, 1070–1082.

Mavrovouniotis,M.L. (1991) Estimation of standard Gibbs energy changes of biotransformations. *J. Biol. Chem.*, **266**, 14440–14445.

Mavrovouniotis,M., Stephanopoulos,G. and Stephanopoulos,G. (1992) Synthesis of biochemical production routes. *Comput. Chem. Eng.*, **16**, 605–619.

Miles,E.W. (2001) Tryptophan synthase: a multienzyme complex with an intramolecular tunnel. *Chem. Rec.*, **1**, 140–151.

Morowitz,H.J., Kostelnik,J.D., Yang,J. and Cody,G.D. (2000) The origin of intermediary metabolism. *Proc. Natl Acad. Sci. USA*, **97**, 7704–7708.

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

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

Rison,S.C.G. and Thornton,J.M. (2002) Pathway evolution, structurally speaking. *Curr. Opin. Struct. Biol.*, **12**, 374–382.

Schmidt,S., Sunyaev,S., Bork,P. and Dandekar,T. (2003) Metabolites: a helping hand for pathway evolution? *Trends Biochem. Sci.*, **28**, 336–341.

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

Talafous,J., Sayre,L.M., Mieyal,J.J. and Klopman,G. (1994) META. 2. A dictionary model of mammalian xenobiotic metabolism. *J. Chem. Inf. Comput. Sci.*, **34**, 1326–1333.

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

Ugi,I., Bauer,J., Brandt,J., Freidrich,J., Gasteiger,J., Jochum,C. and Schubert,W. (1979) New applications of computers in chemistry. *Angew. Chem. Int. Ed. Engl.*, **18**, 111–123.

Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.