Metabolic Pathway Analysis: Basic Concepts and Scientific Applications in the Post-genomic Era

Christophe H. Schilling,[†] Stefan Schuster,[‡] Bernhard O. Palsson,*[,†] and Reinhart Heinrich[§]

Department of Bioengineering, University of California, San Diego, La Jolla, California 92093-0412, Department of Bioinformatics, Max Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, D-13125 Berlin-Buch, Germany, and Institute for Biology/Theoretical Biophysics, Humboldt University—Berlin, Invalidenstrasse 42, D-10115 Berlin, Germany

This article reviews the relatively short history of metabolic pathway analysis. Computer-aided algorithms for the synthesis of metabolic pathways are discussed. Important algebraic concepts used in pathway analysis, such as null space and convex cone, are explained. It is demonstrated how these concepts can be translated into meaningful metabolic concepts. For example, it is shown that the simplest vectors spanning the region of all admissible fluxes in stationary states, for which the term elementary flux modes was coined, correspond to fundamental pathways in the system. The concepts are illustrated with the help of a reaction scheme representing the glyoxylate cycle and adjacent reactions of aspartate and glutamate synthesis. The interrelations between pathway analysis and metabolic control theory are outlined. Promising applications for genome annotation and for biotechnological purposes are discussed. Armed with a better understanding of the architecture of cellular metabolism and the enormous amount of genomic data available today, biochemists and biotechnologists will be able to draw the entire metabolic map of a cell and redesign it by rational and directed metabolic engineering.

Metabolism: The Chemical Engine Driving Cellular Functions

Metabolism is the chemical engine that drives the living process. Through the utilization of a vast repertoire of enzymatic reactions and transport processes, unicellular and multicellular organisms can process and convert thousands of organic compounds into the various biomolecules necessary to support their existence. In switchboard-like fashion the cell and, at a higher level, the organism direct the distribution and processing of metabolites throughout its extensive map of pathways.

While continuously striving to understand how the factory of life operates, as cellular engineers we also attempt to manipulate and exploit these pathways which the cell has at its disposal. At one extreme we seek to develop strategies to effectively eliminate pathways through antibiotics, halting the growth of microorganisms and curbing bacterial infection. And at the other extreme we seek to enhance the performance of certain pathways or introduce entirely novel routes for the production of various biochemicals of commercial interest. Whether the goal is to destroy, create, or enhance the production capabilities of an organism, it is necessary to develop an understanding of how the cell meets its metabolic objectives through the analysis of its metabolic pathways. An understanding of the structural design and capabilities of the cellular metabolic network clearly places the biochemical engineer in an advantageous position to manipulate the cell for various purposes. Like many areas of biological study today, the future of metabolic pathway analysis may depend greatly upon its ability to capitalize on the wealth of genetic and biochemical information currently being generated from the fields of genomics, and similarly proteomics.

Currently there are 20 completely sequenced and annotated microbial genomes that are publicly available and published, and this number continues to rise. Small genome sequencing is becoming routine, and in the future, studies of nearly every organism will be aided by the knowledge and availability of the complete DNA sequence of their genomes (*1*). Using the tools of bioinformatics it is possible to identify the coding regions of a genome and assign function to these regions on the basis of sequence similarities against genes of known function (*2*). Currently this process of genome annotation has resulted in the assignment of 45%−80% of the coding regions in the fully sequenced microbial genomes (*3*). In addition, the majority of genes in microbial cells encode products with metabolic functions (*4*) and the biochemical functions of most metabolic gene products are known. Thus, once the genome has been annotated, the entire metabolic map representing the stoichiometry of all the metabolic reactions taking place in the cell can be constructed. In fact there are now a number of extensive on-line databases detailing the metabolic content and pathway diagrams for most of the fully sequenced microbes (*5*−*7*). With such detailed information available about an organism's arsenal of metabolic reactions, it is now possible to perform detailed studies of the metabolic pathway structure for entire organisms. Undoubtedly,

[†] University of California, San Diego. Telephone: (619) 534-5668. Fax: (619) 822-0240.

[‡] Max Delbrück Center for Molecular Medicine. Fax: 49-30-94062834.

[§] Humboldt University—Berlin. Fax: 49-30-2093 8813.

10.1021/bp990048k CCC: $18.00
Published on Web 05/07/1999

studies of this nature hold potential value to research in various fields, which include metabolic engineering for bioprocesses and therapeutics, bioremediation, and antimicrobial research.

In contrast to detailed simulation studies of metabolic networks, which date back to the early 1970s, metabolic pathway analysis concentrates on the stoichiometric rather than kinetic properties of metabolic networks. In this article we will cover the rather short history of this pathway analysis which finds its theoretical beginning in the early part of the 1980s. The discussion will first begin with the development of various pathway synthesis algorithms based on artificial intelligence and then follow the evolution of a theory for the study of metabolic pathways and the structural aspects of metabolic reaction systems. The future holds bright for the application of these existing theories and concepts to the massive amounts of metabolic content information supplied through genomics.

Programs for the Synthesis of Metabolic Pathways

On the basis of metabolic maps identified by traditional biochemical methods, artificial intelligence has been used to detect special routes of biotransformations. Seressiotis and Bailey developed a computer algorithm to conduct complex searches through reaction networks for the identification/synthesis of biochemical pathways (*8*). Through the implementation of a series of heuristic rules that allow the search algorithm to converge on solutions in an efficient manner, the approach yields "genetically independent" pathways that convert a given substance into a target metabolite. For a pathway to be genetically independent means that there exist no other pathways that only utilize a subset of the reactions used in the pathway of interest, thus answering the question, "Does each pathway generated define an independent genotype?"

To illustrate the capabilities of this approach, a classical scheme of central metabolism was analyzed, which included the Embden−Meyerhof, pentose phosphate, and Entner−Doudoroff families of pathways. Additionally the algorithm was applied to a reaction system for the conversion of pyruvate into L-alanine, for which a pathway was identified that did not utilize alanine aminotransferase. While effective in analyzing reaction systems of moderate size, the computational complexity of the search algorithm limited the size of the reaction systems that could be analyzed with this approach. While the algorithm did lack a sound mathematical foundation, it provided the first attempt to develop a package for metabolic pathway analysis and in particular established the important concept of genetic independence.

This work was shortly followed up by a different approach developed by Mavrovouniotis et al. for the synthesis of biochemical pathways based on stoichiometric constraints (*9*). In this approach, reactions are classified as either being allowed, required, or excluded from the pathways to be generated and reversible reactions are decomposed into forward and backward reactions that are prohibited from participating together in the same pathway. In a refined version of the algorithm, reversible reactions need not be decomposed (*10*). In addition metabolites are also classified as either required reactants, allowed reactants, required products, allowed (by)products, or intermediates. The problem is thus defined by combining the constraints on all of the reactions and metabolites with the natural stoichiometric constraints of each enzymatic reaction. Pathways are continuously synthesized by introducing the constraints on each metabolite one at a time, resulting in the final set of pathways that satisfy all of the constraints placed on the metabolites and reactions. The pathways generated in this approach are genetically independent and constitute all of the pathways satisfying the constraints.

A case study focusing on the synthesis of lysine from glucose and ammonia has been presented to illustrate the utility of this algorithm in identifying fundamental limitations that govern biochemical pathways and processes. In this detailed analysis of the hundreds of pathways that were synthesized using the algorithm, it was shown that oxaloacetate is a key intermediate in the production of lysine from glucose and ammonia. In addition it was clear that the maximum molar yield of lysine to glucose is 67% in the absence of reactions to recover carbon lost as CO_2.

Both of these different approaches illustrated the potential use of the detailed analysis of metabolic pathways for the purposes of metabolic engineering. While the applications of pathway analysis were becoming clearer, there still remained a need for a sound mathematical theory for the study of metabolic pathways to combine with these initial attempts to synthesize metabolic pathways.

Stoichiometry and Structure of Biochemical Reaction Networks

The interconnectivity of metabolites within a network of biochemical reactions is given by reaction equations defining the stoichiometric conversion of substrates into products for every reaction. Enzymatic reactions as well as the transport of metabolites across system boundaries constitute fluxes, which serve to dissipate and generate metabolites. In following the law of conservation of mass, material balances describing the activity of a particular reactant through each reaction can be written where the difference between the rate of production and consumption of a particular metabolite is equivalent to the change in concentration of that metabolite over time. This yields the following equation for every metabolite in a system:

$$\frac{\mathrm{d}[X_i]}{\mathrm{d}t} = \sum_j S_{i,j} v_j \tag{1}$$

where v_j are the fluxes that produce and consume the metabolite in the system and the stoichiometric coefficient S_{ij} stands for the number of moles of metabolite X_i formed in reaction j. It is negative if X_i is a substrate of this reaction. At steady state, the concentration of metabolites in a network is constant and the activity of those fluxes that generate a metabolite must be equivalent to the activity of the fluxes that consume the metabolite. Since the time constants associated with growth are much larger than those associated with individual reaction kinetics, it is reasonable to place the metabolic system in a steady state when investigating aspects of metabolism related to growth. This reduces the above system of equations to a system of homogeneous linear equations, which in matrix notation is

$$\mathbf{S}\cdot\mathbf{v} = 0 \tag{2}$$

The stoichiometric matrix **S** is an $m \times n$ matrix where m corresponds to the number of metabolites and n is the number of reactions or fluxes taking place within the network. The vector **v** refers to the activity of each flux.

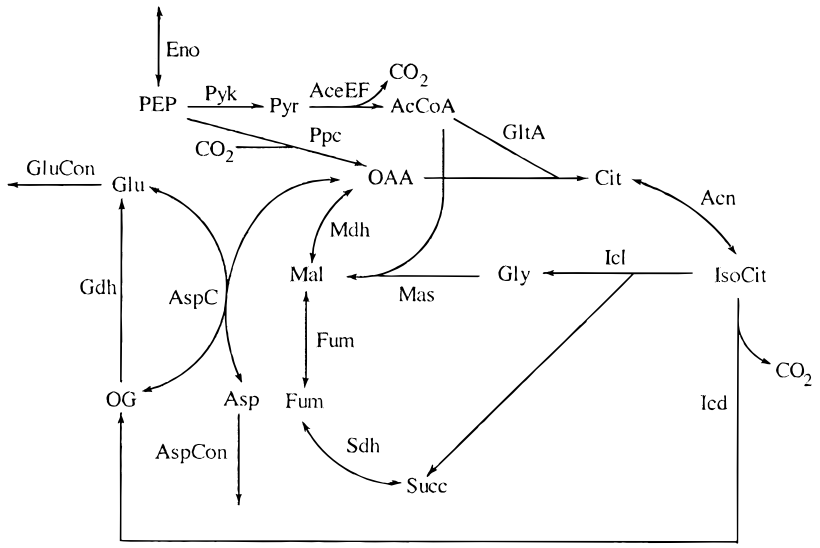

Figure 1. Reaction scheme of the glyoxylate cycle and adjacent reactions. Abbreviations of metabolites: AcCoA, acetyl-CoA; Asp, aspartate; Cit, citrate; Fum, fumarate; Glu, glutamate; Gly, glyoxylate; IsoCit, isocitrate; Mal, malate; OAA, oxaloacetate; OG, 2-oxoglutarate; PEP, phosphoenolpyruvate; Pyr, pyruvate; Succ, succinate. Enzymes are abbreviated by their gene names: AceEF, pyruvate dehydrogenase; Acn, aconitase; AspC, aspartate aminotransferase; Eno, enolase; Fum, fumarase; Gdh, glutamate dehydrogenase; GltA, citrate synthase; Icd, isocitrate dehydrogenase; Icl, isocitrate lyase; Mas, malate synthase; Mdh, malate dehydrogenase; Ppc, PEP carboxylase; Pyk, pyruvate kinase; Sdh, succinate dehydrogenase; AspCon, GluCon, consumption of aspartate and glutamate, respectively. Reversible reactions are indicated by double arrowheads.

Consider, for example, the glyoxylate cycle together with several adjacent reactions as shown in Figure 1. α-Ketoglutarate dehydrogenase and succinyl-CoA synthetase, which would complete the tricarboxylic acid cycle, are assumed to be inhibited. A more extended scheme including these two enzymes has been studied elsewhere (*11*). The concentrations of 2-phosphoglycerate (the substrate of enolase), CO_2, NH_3, and all cofactors such as ATP, NADP, and so on (which are omitted in Figure 1) are assumed to be buffered. By assigning, for example, reaction index 1 to enolase and metabolite index 1 to PEP, element $S_{1,1}$ of **S** would be 1 because 1 mol of PEP is produced in the enolase reaction. The stoichiometric coefficient of PEP in the pyruvate kinase reaction is −1 and in the aconitase reaction, for example, zero. As the stoichiometric matrix links the vector of time derivatives of the concentrations (which is the null vector at steady state) with the vector of reaction rates, it is not usually square, in contrast to system matrices widely used in systems analysis.

As it may pertain to genome-scale metabolic studies of organisms, the stoichiometric matrix can be directly constructed from knowledge of an organism's metabolic genotype, which may now be realistically determined from the results of genome annotation. The stoichiometric matrix thus contains all of the information about how substances are linked through reactions within the network. It indicates the topological structure and architecture of the network, and a knowledge of its properties is a prerequisite for any simulation of biochemical reaction networks (*12*). The matrix representation of an entire metabolic genotype or any biochemical reaction network such as the one shown in eq 2 lends itself to further analysis centered around the concepts of linear algebra and allows us to translate knowledge and concepts directly from mathematics into biology and vice versa.

As the matrix is constructed to represent homogeneous linear equations, a corresponding "null space" can be described (*13*). Within the null space lies all of the possible solutions and hence flux distributions or flows under which the system can operate. The concept of a null space in terms of biochemical reaction networks has been understood for over a decade (*14*, *15*), and these authors demonstrated that an analysis of the stoichiometric structure may reveal a number of fundamental system properties. In contrast to this structure, the kinetic properties of systems are subject to frequent change as a result of regulation. These regulatory interactions and control of biochemical networks are built around their overall structural properties, which remain relatively invariable in a certain time scale.

Since every solution or operating mode of the system is contained within the null space, it logically follows that the entire null space represents the capabilities of a given metabolic genotype. Thus the null space clearly defines what a genotype can and cannot do; what building blocks can be manufactured; how efficient the energy extraction and conversion of carbohydrates into biomolecules can be for a given substrate; and where the critical links in the network are (*16*, *17*). If the answers to these questions and any others related to the basic structural capabilities of the network lie within the null space, the goal must then be to develop a way to describe and interpret our position within this space from an overall metabolic perspective. In other words, we must find the best way to navigate through this solution space using a map of biochemical reactions. It is therefore essential that we explicitly define the null space and hence capabilities of a system in a manner that provides us with informative answers to our biologically oriented questions.

Metabolic Pathways Spanning the Null Space

From the Rank theorem of linear algebra, the dimension of the null space is given by the following equation:

$$\dim Nul(\mathbf{S}) = n - rank(\mathbf{S}) \qquad (3)$$

where n is once again the number of reactions in the system. In the case that the rank of **S** equals the number of its rows (m), the dimension of the null space simply equals the difference between the number of fluxes and metabolites. The rank of **S** is less than m whenever conservation relations (such as $ATP + ADP = \text{const}$) hold

Table 1. Null Space Matrix K to the Reaction System Shown in Figure 1[a]

Eno	Acn	Sdh	Fum	Mdh	AspC	Gdh	Pyk	AceEF	GltA	Icd	Icl	Mas	AspCon	Ppc	GluCon
2	1	1	1	2	1	1	2	2	1	0	1	1	1	0	0
1	1	1	1	2	0	0	2	2	1	0	1	1	0	−1	0
3	2	1	1	2	0	1	3	3	2	1	1	1	0	0	1

[a] For reasons of layout, the transposed matrix is given, that is, columns are here shown as rows. The enzyme abbreviations are the same as used in Figure 1. Note that the CO_2 production flux is the summation of the flux through AceEF and Icd less the flux through Ppc, as this reaction incorporates a molecule of CO_2 in the forward reaction whereas AceEF and Icd produce a molecule of CO_2.

in the system. Usually the dimension of the null space is greater than zero and there exist linear dependencies among the columns of **S** that lead to the existence of nontrivial solutions to the homogeneous equation system. All of these solutions will lie in the null space. To explicitly define the null space, it is necessary to generate a set of vectors that can be used to span the vector space (i.e., a spanning set). The most efficient way to span a vector space is through the use of a minimum number of linearly independent vectors that together form a basis, with the minimum number equal to the dimension of the vector space. Each basis vector forms a column of the null space matrix **K**, which satisfies the following equation:

$$\mathbf{SK} = 0 \tag{4}$$

The null space matrix to the system shown in Figure 1 is given in Table 1. The numbers in this table indicate the relative fluxes carried by the enzymes. The first row, for example, corresponds to a flux distribution where two units of flux go through enolase while one unit is carried by aconitase and so on. With the help of other examples, it has been previously demonstrated how the null space can be spanned by a series of independent vectors that are theoretically and biochemically meaningful, where each basis vector translates into a functional pathway operating within the network (*18*, *19*). Using the integrated metabolic network of the human erythrocyte as an example (see refs 20−22), it was illustrated how these basis vectors that represent functional pathways can be used to interpret the operational activities and capabilities of the red blood cell (*18*).

In this case the first basis vector/pathway in Table 1 corresponds to a pathway for the production of aspartate (4-carbon), while the second pathway would correspond to a cycling of the glyoxylate cycle resulting in the complete reduction of PEP (3-carbon) to 3 mol of CO_2 /mol of PEP, with the third pathway producing glutamate (5-carbon). From this perspective it should be possible to interpret the systemic functional activities of a metabolic genotype in terms of an underlying pathway structure in addition to an individual reaction-based perspective derived from simply assigning and cataloguing genes. More importantly, one can interpret the capabilities of a metabolic genotype in terms of a set of metabolic pathways that span its null space providing complete coverage of the solution space.

The null space matrix **K** provides us with a link of metabolic pathway analysis with metabolic control analysis (MCA) developed in refs 23 and 24 (see also refs 12 and 25). In the latter theoretical framework the flux control coefficients and concentration control coefficients, being the elements of a flux control matrix $\mathbf{C}^J$ and a concentration control matrix $\mathbf{C}^X$, respectively, obey the relations

$$\mathbf{C}^J\mathbf{K} = \mathbf{K} \tag{5a}$$

$$\mathbf{C}^X\mathbf{K} = 0 \tag{5b}$$

which are known as generalized summation theorems (*12, 14, 26*). The elements $C^J_{i,j}$ and $C^X_{k,j}$ of $\mathbf{C}^J$ and $\mathbf{C}^X$, respectively, express the control exerted by enzyme j on flux J_i and metabolite concentration X_k, respectively. The control coefficients can be determined uniquely from the algebraic eqs 5a,b together with the connectivity theorems of MCA. Whereas the connectivity relations are dependent on the kinetic properties of the reactions, eqs 5a,b represent constraints for the control coefficients reflecting exclusively stoichiometric properties of the network. While MCA in its general form is restricted to the study of steady states, attempts have been made to extend it so as to cope with time-dependent processes (*27, 28*). The approach presented in ref 28 is based on the concept of control coefficient, which expresses the control exerted by an enzyme (or a reaction), while in ref 27, sensitivities to parameter perturbations are considered.

This rather elementary first-hand approach to explicitly define the algebraic properties of a metabolic network illustrates the functional relevance of the basis vectors that span the null space. However, as with any coordinate transformation, the selection of basis vectors is nonunique, as there are many sets of vectors that can be used to span the null space that may be both theoretically and biochemically feasible. For example, the sum of any two rows in Table 1 also lies in the null space and could be taken as a basis vector.

An algorithm for the construction of a set of basis vectors was first published in ref 29 as well as in the form of a computer program CONTROL (*30*), which was developed for the determination of control coefficients of MCA. In this approach the null space matrix **K** is computed to have the following form, where **I** denotes the square identity matrix whose dimension is equivalent to the dimension of the null space given in eq 3:

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_o \\ \mathbf{I} \end{bmatrix} \tag{6}$$

The determination of $\mathbf{K}_o$ then follows from the use of the Gaussian elimination method. The representation given in eq 6 is of importance for finding a block diagonal structure of **K** (*29*). This approach for determining basis vectors has also been used for the determination of group control coefficients (*31*). Once again this approach to determine the basis vectors meets with the pitfall that the composition of the basis vectors is arbitrary and hence nonunique. Any linear combination of the column vectors of matrix **K** given in eq 6 is again a basis vector.

Convex Analysis and Elementary Modes

Similar to stoichiometry, the pathway structure of a metabolic system should also be invariant, as it is a direct result of the stoichiometry. Since any set of basis vectors is nonunique the pathway structure as determined by these vectors/pathways cannot be an invariant property of the network. To overcome this obstacle of nonuniqueness, we turn to the mathematics associated with the study of convex spaces for an elegant solution to the problem of determining metabolic pathways (*32*). Convex

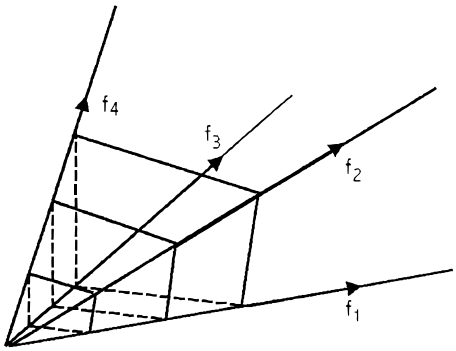

Figure 2. Schematic representation of the convex flux cone belonging to the reaction system shown in Figure 1. $\mathbf{f}_1-\mathbf{f}_4$ denote the generating vectors of the cone, which correspond to the four elementary modes of the system. The quadrangles are to visualize the three-dimensional structure of the cone.

analysis has been applied earlier in the study of nonbiological chemical systems (*33*). If one decomposes reversible reactions into both forward and backward reactions, the activity of every reaction must be either positive or zero. Another interesting case is that the net flux is positive because the backward reaction is, under all physiological conditions, slower than the forward reaction. In these situations the solution to the steady-state eq 2 must lie in the nonnegative orthant of the space spanned by all the individual reactions due to the nonnegativity constraints on the fluxes. The basis vectors of the null space as computed by eq 6 do not necessarily satisfy this condition, as can be seen in the second basis vector given in Table 1. PEP carboxylase, which is assumed to be irreversible, is used by the corresponding flux distribution in the wrong directionality. The solution space simply becomes the intersection of the null space with the positive orthant. Furthermore the solution space for such a problem takes the shape of a convex polyhedral cone with a finite number of edges (see Figure 2) (*32*). By virtue of the cone being convex, any vector within the cone can be represented as a nonnegative linear combination of the generating vectors of the cone, which correspond to its edges. Moreover, the edges of the cone are unique except for arbitrary scaling and correspond to biochemically feasible pathways. The mathematical description of such a cone, F, is as follows, with $\mathbf{f}_k$ referring to all of the generating vectors of the cone:

$$F = \left\{ \mathbf{v} \in \mathrm{R}^n | \mathbf{v} = \sum_k \alpha_k \mathbf{f}_k, \alpha_k \geq 0 \right\} \tag{7}$$

It should be noted that the number of edges of the cone can and often does exceed the dimensions of the null space. However, while the edges of the cone then are linearly dependent in the mathematical sense, they will all constitute genetically independent pathways as previously defined and mentioned herein (*8*). However, it should be noted that the term genetic independence may be slightly misleading, as there is not necessarily a one-to-one correspondence between reactions and gene products. Table 2 gives the four generating vectors belonging to the system shown in Figure 1. Their number is actually greater than the dimension of the null space to this system, which is three (see Table 1). These vectors are genetically independent because none of them can be written as a linear combination of the other three with nonnegative coefficients. Biologically, this means that each of them corresponds to a different set of enzymes involved, with none of these sets being a subset of another set. The four vectors can be brought in relation with different biological functions: aspartate synthesis via PEP carboxylase (Ppc), aspartate synthesis via the glyoxylate cycle, glutamate synthesis via Ppc, and glutamate synthesis via the glyoxylate cycle. The precise relationship between the three basis vector ($\mathbf{b}_1,...,\mathbf{b}_3$) and the four generating vectors ($\mathbf{f}_1,...,\mathbf{f}_4$) is outlined in eq 8.

$$\begin{aligned} \mathbf{f}_1 &= \mathbf{b}_1 - \mathbf{b}_2 \\ \mathbf{f}_2 &= \mathbf{b}_1 \\ \mathbf{f}_3 &= \mathbf{b}_3 - \mathbf{b}_2 \\ \mathbf{f}_4 &= \mathbf{b}_3 \end{aligned} \tag{8}$$

With pathways defined in this nature, it becomes clear from eq 7 that to achieve any particular flow through the network the pathways representing the frame of the cone must be either "switched-off" or "turned-on". This can provide a valuable biological interpretation of the underlying pathway structure of a metabolic genotype and may shed light on the regulatory logic implemented by the cell to control its metabolic network.

Clarke et al. have used this set of generating vectors for various studies on steady states and reaction dynamics, keeping in mind that these vectors are solely determined by stoichiometry (*33, 34*). In addition they developed two different algorithms for the calculation of these edges for a given reaction network.

Later a simple yet highly efficient algorithm for finding all of the admissible nonnegative steady-state flux distributions in reaction systems was developed (*35*). In addition to finding the edges of the cone like other preceding algorithms, this program was capable of calculating all of the admissible flux distributions with certain flux values fixed. This tool is valuable for the calculation of flux values from radioactive tracer or NMR labeling experiments (*36*−*38*).

While not providing a detailed mathematical basis for their development, Leiser and Blum were the first to define "fundamental" modes, which corresponded to steady-state pathways through a system (*39*). It was stated that any steady-state flux pattern for a system can be decomposed as a linear superposition of these modes. Included as fundamental modes were purely cyclic pathways. Cyclic pathways and/or "futile" cycles are an important aspect of reaction networks (*25, 40*). In all the above approaches based on convex analysis and studies of the null space, cyclic pathways are readily identified in the solutions. However, this is not the case with the pathway synthesis programs previously discussed.

Building on the early definition of "fundamental" modes, a theory behind the identification of steady-state pathways in metabolic reaction networks was developed (*41*). By providing sound mathematical proofs and definitions as well as biological definitions for what has been termed "elementary" modes, many issues related to the applications of convex analysis to metabolic pathways have been tackled. An elementary mode is a minimal set of enzymes that can operate at steady state with all irreversible reactions proceeding in the appropriate direction. An illustrative operational definition can be given as follows. Block some enzyme in the metabolic system under study by the addition of an excess amount of an enzyme-specific inhibitor and determine whether there is still some flow going through the system. Next, block a second enzyme and so on. An elementary mode is reached when the inhibition of a further, still active, enzyme leads to breakdown of any steady-state flux in

Table 2. Generating Vectors to the Flux Cone[a] Belonging to the Reaction System Shown in Figure 1[b]

Eno	Acn	Sdh	Fum	Mdh	AspC	Gdh	Pyk	AceEF	GltA	Icd	Icl	Mas	AspCon	Ppc	GluCon
1	0	0	0	0	1	1	0	0	0	0	0	0	1	1	0
2	1	1	1	2	1	1	2	2	1	0	1	1	1	0	0
2	1	0	0	0	0	1	1	1	1	1	0	0	0	1	1
3	2	1	1	2	0	1	3	3	2	1	1	1	0	0	1

[a] Which here coincide with the elementary modes. [b] Note that the CO_2 production flux is the summation of the flux through AceEF and Icd less the flux through Ppc as this reaction incorporates a molecule of CO_2 in the forward reaction where as AceEF and Icd produce a molecule of CO_2.

the system. In mathematical terms, an elementary mode can be represented by a flux vector that cannot be decomposed into two flux vectors that would have additional zero components. In this approach, reversible reactions are considered as single fluxes, so that a decomposition into forward and backward reactions is not necessary. The flux modes used to span cones are no longer confined to the positive orthant due to the relaxation of inequality constraints on the flux values of those reactions that are reversible (*26, 41*). An algorithm for detecting the elementary modes of systems containing reversible reactions was also developed (*42*). Equation 7 is still valid, with $\mathbf{f}_k$ denoting elementary modes. However, there may be more of these modes than needed to span the cone. A slightly different approach was adopted in the study of the evolution of sugar metabolism (*43*), where only a subset of the modes were taken so as to span the cone by a minimal number of vectors. However, these are then not uniquely determined.

As for the system shown in Figure 1, the generating vectors given in Table 2 form a complete set of elementary modes. Here, the irreversible reactions force the reversible reactions to operate in a unique directionality. In other systems, however, such as in the pentose phosphate pathway, some reactions may operate in either direction in different elementary modes. In principle, it is even possible that an entire elementary mode is reversible; however, it seems that, in metabolism, irreversible reactions such as those catalyzed by many kinases and phosphatases are located in sufficient number and at appropriate positions so that reversible flux modes rarely occur.

The Current Theory and State of Affairs

At this point it is clear that we can use genomics to define an organism's metabolic genotype and hence the stoichiometric matrix and interconnectivity of metabolites involved. Within the null space of this matrix lies all of the possible flux distributions and functional capabilities of the system. A particular point within the null space or a particular solution to the system represents a metabolic phenotype that the cell can express. Importantly this phenotype can be described in terms of the individual reactions or in terms of the genetically independent pathways operating within the system.

To explicitly define the null space one may choose to span the null space using the traditional methods of linear algebra to generate a basis that is nonunique yet biochemically feasible and does have the potential mathematical advantage that the vectors are all linearly independent. However, having chosen a basis, we do not know whether we missed an important fundamental pathway. On the other hand, using the concepts of convex analysis, one may generate a unique set of pathways that correspond to the edges of the admissible region of flux vectors. In this case the set of pathways are not linearly independent; however, their construction is unique and this has the advantage that all the vectors are biologically or genetically independent and are invariant much like the stoichiometry itself. Both cases offer explicit definitions of the null space and can be used as tools to interpret the capabilities and structural design of the metabolic genotype, the choice is simply linear independence and nonuniqueness versus genetic independence and a unique and comprehensive pathway structure. The benefit of determining a unique pathway structure in most cases outweighs the advantage of basis vectors being linearly independent and thus suggests the use of convex analysis versus traditional linear algebra in the investigation of metabolic systems. Figure 2 helps to illustrate many of these concepts discussed above. Tables 1 and 2 help to show the one-to-one correspondence between established mathematical concepts and those that pertain to the biochemistry of metabolic networks.

Regardless of the approach used to define functional pathways, the determination of pathways is entirely dependent upon the metabolic system constructed. Thus it is important to have confidence in the functional assignments that have been made to a genome. If errors exist in this functional assignment and annotation or if genes are included with broad specificity, associated pathways will appear that are based upon this error in defining the metabolic network. This is a problem common to any approach that is based on genomic sequencing and the results of its annotation. However, from pathway analysis, it is possible to determine on the basis of genome annotation if an organism is capable of synthesizing a component of the biomass. In the case that it is shown through pathway analysis that the organism is unable to synthesize a molecule (e.g., an amino acid) that has been shown to be produced from biochemical assays, then we can conclude that a functional assignment has been missed, or perhaps lend confidence to an assignment with a low level of certainty attached to it that was not included in the original annotation.

Pathway analysis is an essential prerequisite of studies on optimal properties of metabolic networks. In studying optimization in metabolic engineering, it is important to realize that biochemical pathways are the result of an *evolutionary* optimization (cf. refs 12, 26, and 44). Thus, studies on the optimal structural design of these pathways can be used as a guideline for a further optimization in biotechnology. On the other hand, there is a difference in that biotechnological optimization is generally aimed at the improvement of a few specialized objectives, whereas biological evolution has acted to achieve a well-tuned balance between several functions. An example of an objective that is relevant both in biotechnology and evolution is the maximization of the stoichiometric yield of biosynthetic processes. This maximization problem can be solved by linear programming (*45, 46*), by analyzing the elementary modes in the system (*11*), or in combination with kinetic optimization (*47*).

In our example, mode 1 gives a molar yield of one mole of aspartate formed per mole of PEP used, while mode 2 gives only a yield of 1:2. The yields of glutamate synthesis in the two modes 3 and 4 differ as well (1:2 and 1:3,

respectively). As any real flux distribution is a nonnegative linear combination of elementary modes, there cannot be any flux pattern that would give a better yield than realized by one of these modes. Thus, mode 1 is the optimal solution with respect to the yield of aspartate synthesis in the framework of the network considered.

It is worth noting that applications of linear programming and optimization to metabolism, such as flux balance analysis (FBA) (*16, 48*), are based on many of the same concepts of convexity mentioned above. Beginning with an unbounded polyhedral cone (solution space) determined from stoichiometric constraints, the objective functions and additional constraints of a linear programming problem slice through the cone so as to create a bounded polyhedron. The particular solution to the problem then lies in the corners or possibly the edges of this bounded polyhedron. FBA can be used as a method to analyze, interpret, and predict metabolic phenotypes from metabolic genotypes derived from genomic data (*49*).

Future of Metabolic Pathway Analysis

With a solid conceptual framework developed and a growing list of applications for the study of metabolic pathways, the time appears ripe to develop further applications and analytical techniques to better characterize the relation between an organism's metabolic genotype and phenotype.

A limitation to the approaches reviewed above is that no predictions about the dynamic behavior of the system can be made. Nevertheless, utilizing a pathway-oriented approach to gain insight into the regulatory logic implemented by the cell could prove to be a fruitful endeavor. While there is great attention given to the mathematical foundations of regulation and control through the study of the linearly independent pathways spanning the null space (*12, 50*), relatively little effort has been directed at assessing the role that genetically independent pathways and elementary modes may play in the regulatory process. Are these the true entities which the cell must control? It is, for example, an intriguing question whether each regulatory activation or inhibition loop may be assigned to one of the elementary modes. This may be a future subject in MCA, which considers both structural and kinetic properties. Moreover, it is an interesting task to extend the concepts of elementary modes and pathways to the study of signal transduction pathways, an area of modeling that remains relatively untouched. Preliminary ideas in this direction have been put forward (*42*).

Liao et al. were the first to utilize metabolic pathway analysis for the development of high-efficiency production of biochemicals (*51*). By applying the theory of convex analysis for the determination of metabolic pathways, the authors examined all of the optimal and suboptimal flux distributions and elementary modes operating through central metabolism that redirected carbon flow to the pathways for aromatic amino acid production. Through genetic manipulation they constructed an *E. coli* strain that successfully channeled carbohydrates down the aromatic pathway at theoretical yields, offering a shining example of the application of pathway analysis to industrial bioprocesses.

With recent advances such as oligonucleotide chip technology and DNA microarrays, it is possible to analyze the dynamic events which occur in gene expression as a consequence of operational shifts in cellular systems. In the future we might expect to see the utilization of a variety of DNA-chip-based studies to track the expression of the various metabolic genes and pathways existing within an organism. Already a recent article has shown how DNA microarray technology can be applied to study the metabolic and genetic control of gene expression on a genomic scale (*52*). In this study, temporal changes in genetic expression profiles were observed for virtually every known gene in *Saccharomyces cerevisiae* during the diauxic shift and the up regulation and/or down regulation of key metabolic pathways was observed.

Armed with a complete understanding of the pathway structure and capabilities of a metabolic genotype and the tools of genetic engineering, the pathway engineer is poised to dismantle or fine-tune the metabolic engine as required to achieve their cellular objective.

References and Notes

(1) Ash, C. Year of the genome. *Trends Microbiol.* **1997**, *5* (4), 135−139.
(2) Huynen, M. A.; Diaz-Lazcoz, Y.; Bork, P. Differential genome display. *Trends Genet.* **1997**, *13*, 389−390.
(3) Pennisi, E. Laboratory workhorse decoded. *Science* **1997**, *277*, 1432−1434.
(4) Ouzounis, C.; et al. Computational comparisons of model genomes. *Trends Biotechnol.* **1996**, *14* (8), 280−285.
(5) Selkov, E., Jr.; et al. MPW: the metabolic pathways database. *Nucleic Acids Res.* **1998**, *26* (1), 43−45.
(6) Karp, P. D.; et al. EcoCyc: Encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Res.* **1998**, *26* (1), 50−53.
(7) Kanehisa, M. A database for post-genome analysis. *Trends Genet.* **1997**, *13*, 375−376.
(8) Seressiotis, A.; Bailey, J. E. MPS: An artificially intelligent software system for the analysis and synthesis of metabolic pathways. *Biotechnol. Bioeng.* **1988**, *31*, 587−602.
(9) Mavrovouniotis, M. L.; Stephanopoulos, G.; Stephanopoulos, G. Computer-aided synthesis of biochemical pathways. *Biotechnol. Bioeng.* **1990**, *36*, 1119−1132.
(10) Mavrovouniotis, M. L. Synthesis of reaction mechanisms consisting of reversible and irreversible steps. 2. Formalization and analysis of the synthesis algorithm. *Ind. Eng. Chem. Res.* **1992**, *31*, 1637−1653.
(11) Schuster, S.; Dandekar, T.; Fell, D. A. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.* **1999**, *17*, 53−60.
(12) Heinrich, R.; Schuster, S. *The regulation of cellular systems*; Chapman & Hall: New York, 1996.
(13) Lay, D. C. *Linear algebra and its applications*, 2nd ed.; Addison-Wesley Longman Inc.: Reading, MA, 1997.
(14) Reder, C. Metabolic control theory: A structural approach. *J. Theor. Biol.* **1988**, *135*, 175−201.
(15) Clarke, B. L. Stoichiometric network analysis. *Cell Biophys.* **1988**, *12*, 237−253.
(16) Varma, A.; Palsson, B. O. Metabolic flux balancing: basic concepts, scientific and practical use. *Biotechnol. Bioeng.* **1994**, *12*, 994−998.
(17) Edwards, J. S.; Palsson, B. O. How will bioinformatics influence metabolic engineering. *Biotechnol. Bioeng.* **1998**, *58*, 162−169.
(18) Schilling, C. H.; Palsson, B. O. The underlying pathway structure of biochemical reaction networks. *Proc. Natl Acad. Sci. U.S.A.* **1998**, *95*, 4193−4198.
(19) Fell, D. A. The analysis of flux in substrate cycles. In *Modern Trends in Biothermokinetics*; Schuster, S., et al., Eds.; Plenum: New York, 1993; pp 97−101.
(20) Rapoport, T. A.; Heinrich, R.; Rapoport, S. M. The regulatory principles of glycolysis in erythrocytes in vivo and in vitro. A minimal comprehensive model describing steady states, quasi-steady states and time dependent processes. *Biochem. J.* **1976**, *154*, 449−469.
(21) Werner, A.; Heinrich, R. A kinetic model for the interaction of energy metabolism and osmotic states of human erythrocytes. Analysis of the stationary "in vivo" state and of time dependent variations under blood preservation conditions. *Biomed. Biochim. Acta* **1985**, *44*, 185−212.

(22) Joshi, A.; Palsson, B. O. Metabolic dynamics in the human red cell. *J. Theor. Biol.* **1989**, *141*, 515−528.
(23) Kacser, H.; Burns, J. A. The control of flux. *Symp. Soc. Exp. Biol.* **1973**, *27*, 65−104.
(24) Heinrich, R.; Rapoport, T. A. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur. J. Biochem.* **1974**, *42*, 89−95.
(25) Fell, D. *Understanding the control of metabolism*; Portland Press: London, 1997.
(26) Heinrich, R.; Schuster, S. The modelling of metabolic systems. Structure, control, and optimality. *Biosystems* **1998**, *47*, 61−77.
(27) Kohn, M. C.; Whitley, L. M.; Garfinkel, D. Instantaneous flux control analysis for biochemical systems. *J. Theor. Biol.* **1979**, *76*, 437−452.
(28) Heinrich, R.; Reder, C. Metabolic control analysis of relaxation processes. *J. Theor. Biol.* **1991**, *151*, 343−350.
(29) Schuster, S.; Schuster, R. Detecting strictly detailed balanced subnetworks in open chemical reaction networks. *J. Math. Chem.* **1991**, *6*, 17−40.
(30) Letellier, T.; Reder, C.; Mazat, J.-P. CONTROL: software for the analysis of the control of metabolic networks. *Comput. Appl. Biosci.* **1991**, *7* (3), 383−390.
(31) Stephanopoulos, G.; Simpson, T. W. Flux amplification in complex metabolic networks. *Chem. Eng. Sci.* **1997**, *52* (15), 2607−2627.
(32) Rockafellar, R. T. *Convex analysis*; Princeton Landmarks in Mathematics; Princeton University Press: Princeton, NJ, 1970.
(33) Clarke, B. L. Complete set of steady states for the general stoichiometric dynamical system. *J. Chem. Phys.* **1981**, *75* (10), 4970−4979.
(34) Von Hohenbalken, B.; Clarke, B. L.; Lewis, J. E. Least distance methods for the frame of homogeneous equation systems. *J Comput. Appl. Math.* **1987**, *19*, 231−241.
(35) Schuster, R.; Schuster, S. Refined algorithm and computer program for calculating non-negative fluxes admissible in steady states of biochemical reaction systems with or without some flux rates fixed. *Comput. Appl. Biosci.* **1993**, *9* (1), 79−85.
(36) Salon, C.; Raymond, P.; Pradet, A. Quantification of carbon fluxes through the tricarboxylic acid cycle in early germinating lettuce embryos. *J. Biol. Chem.* **1988**, *263* (25), 12278−12287.
(37) Chance, E. M.; et al. Mathematical analysis of isotope labeling in the citric acid cycle with applications to ^{13}C NMR studies in perfused rat hearts. *J. Biol. Chem.* **1983**, *258* (22), 13785−13794.
(38) Sauer, U.; et al. Metabolic fluxes in riboflavin-producing *Bacillis subtilis*. *Nature Biotechnol.* **1997**, *15* (5), 448−452.
(39) Leiser, J.; Blum, J. J. On the analysis of substrate cycles in large metabolic systems. *Cell Biophys.* **1987**, *11*, 123−138.
(40) Stein, R. B.; Blum, J. J. On the analysis of futile cycles in metabolism. *J. Theor. Biol.* **1978**, *72*, 487−522.
(41) Schuster, S.; Hilgetag, C. On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.* **1994**, *2* (2), 165−182.
(42) Schuster, S.; et al. Elementary modes of functioning in biochemical networks. In *Computation in Cellular and Molecular Biological Systems*; Cuthbertson, R., Holcombe, M., Paton, R., Eds.; World Scientific: London, 1996; pp 151−165.
(43) Nuño, J. C.; et al. Network organization of cell metabolism: monosaccharide interconversion. *Biochem. J.* **1997**, *324*, 103−111.
(44) Heinrich, R.; Schuster, S.; Holzhütter, H. G. Mathematical analysis of enzymic reaction systems using optimization principles. *Eur. J. Biochem.* **1991**, *201*, 1−21.
(45) Varma, A.; Palsson, B. O. Metabolic capabilities of *Escherichia coli*: Synthesis of biosynthetic precursors and cofactors. *J. Theor. Biol.* **1993**, *16*, 477−502.
(46) Varma, A.; Boesch, B. W.; Palsson, B. O. Biochemical production capabilities of *Escherichia coli*. *Biotechnol. Bioeng.* **1993**, *42*, 59−73.
(47) Stephani, A.; Heinrich, R. Kinetic and thermodynamic principles determining the structural design of ATP-producing systems. *Bull. Math. Biol.* **1998**, *60*, 505−543.
(48) Bonarius, H. P. J.; Schmid, G.; Tramper, J. Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends Biotechnol.* **1997**, *15* (8), 308−314.
(49) Schilling, C. H.; Edwards, J. S.; Palsson, B. O. Toward metabolic phenomics: analysis of genomic data using flux balances. *Biotechnol. Prog.* **1999**, *15*, xxx.
(50) Kacser, H.; Acerenza, L. A universal method for achieving increases in metabolite production. *Eur. J. Biochem.* **1993**, *216*, 361−367.
(51) Liao, J. C.; Hou, S.-Y.; Chao, Y.-P. Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnol. Bioeng.* **1996**, *52*, 129−140.
(52) DeRisi, J. L.; Iyer, V. R.; Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **1997**, *278*, 680−686.

Accepted April 1, 1999.

BP990048K