In: ***"Modelling in Molecular Biology"*****, Springer Natural Computing Series**

Gabriel Ciobanu · Grzegorz Rozenberg (Eds.) pp. 1-19 (2004)

Stoichiometric Foundation of Large-Scale Biochemical System Analysis

Daniel A. Beard[1], Hong Qian[2], and James B. Bassingthwaighte[3]

[1] Bioengineering, University of Washington, Seattle, WA
`dbeard@bioeng.washington.edu`
[2] Applied Mathematics and Bioengineering, University of Washington.edu, Seattle, WA
`qian@amath.washington.edu`
[3] Bioengineering, University of Washington, Seattle, WA
`jbb@bioeng.washington.edu`

1 Introduction

The traditional approach to unraveling functions of a biochemical system is to study isolated enzymes and/or complexes, and to determine their kinetic mechanisms for catalyzing given biochemical reactions along with estimates of the associated parameter values [51, 44]. While this reductionist approach has been fruitful, the buzzwords of the present are *integration* and *systems*. One of the important tasks in current computational biology is to assimilate and integrate the behavior of interacting systems of many enzymes and reactants. Understanding of such systems lays the foundation for modeling and simulation of whole-cell systems, a defining goal of the current era of biomedical science.

In this paper we discuss approaches to modeling biochemical systems, with an emphasis on the basic concepts and techniques used in building large-scale integrated models of biochemical reaction networks. We consider the vices and virtues of the available methods; we speculate on what approaches are most reasonable for large-scale cellular modeling.

How far current technology is from a reasonable quantification of whole-cell biochemistry depends on what level of detail one considers. At the simplest level (considering only reaction stoichiometry), whole-genome metabolic models of several single-celled organisms have been developed [2, 48, 23, 47, 52]. At the more detailed level of kinetic modeling, models of the relatively simple metabolism of the red blood cell represent some of the most ambitious attempts to date at modeling whole cell metabolism [24, 57, 28, 29].

While there is no one single approach to biochemical reaction network modeling deemed superior, all models have to satisfy a set of basic criteria. Recently, one of us have proposed the concept of "sustainable conservative cell" [5]. It is argued that all biochemical systems models need to properly represent the basic stoichiometry, with balanced chemical reactions, and the conservation of mass, energy, and charge. It is along this line we carry on our discussion.

2 Stoichiometric Organization of Biochemical Systems

We group approaches to modeling and simulation of biochemical systems into three hierarchical levels of detail: (1) stoichiometric, in which only the stoichiometry of the reaction network is known; (2) kinetic, in which detailed kinetic mechanisms and associated parameters are known for a reaction system; and (3) distributed, in which, along with detailed kinetics, information on the heterogeneous spatial organization of a biological system is considered. Stoichiometric rules are outlined in this section; kinetic approaches are described in following sections. Spatially distributed systems, based on reaction-diffusion modeling, are reviewed elsewhere [3, 4].

The stoichiometry of a reaction network constrains the allowable metabolic fluxes according to mass balance and thermodynamics. In stoichiometric systems, such as a system of chemical reactions, the reactant concentrations change according to:

$$d\mathbf{c}/dt = \mathbf{SJ}, \tag{1}$$

where $\mathbf{S} \in \Re^{M \times N}$ is the stoichiometric matrix [10, 11], $\mathbf{c}$ is the vector of concentrations of M reactants in the systems, and $\mathbf{J}$ is the vector of N fluxes. Eq. (1) constrains the $d\mathbf{c}/dt$ vector to a subspace of $\Re^M$ as follows:

$$\mathbf{L}d\mathbf{c}/dt = 0, \tag{2}$$

where $\mathbf{L}$ is the left null space of $\mathbf{S}$, e.g., $\mathbf{LS} = 0$. Eq. (2) defines linear combinations of concentrations that are constant: $\mathbf{Lc} = \mathbf{k}$, where $\mathbf{k}$ is a constant vector of length equal to the dimension of the left null space. The matrix $\mathbf{L}$ is a so-called conservation matrix; the consequences of Eq. (2) have been extensively examined by Alberty [1].

In the steady state, $d\mathbf{c}/dt = 0$, and the fluxes obey flux balance:

$$\mathbf{SJ} = 0. \tag{3}$$

Eq. (3) is the mass balance constraint that serves as the centerpiece of flux-balance analysis (FBA). In FBA, which has been applied to large-scale metabolic systems with promising success [23, 47, 6], Eq. (3) is used in concert with some biological objective function (such as biomass production, or growth) that is assumed to be effectively optimized. Flux balance is discussed in greater detail in Section 3.4.

In addition to mass balance, the stoichiometry of a system constrains the fluxes according to the laws of thermodynamics. It has been shown that by introducing the right null space, $\mathbf{R}$, of the stoichiometric matrix $\mathbf{S}$ one can derive an energy-balance law [6, 35]:

$$\sum_j R_{jk} \sum_i S_{ij} \mu_i = \sum_j R_{jk} \Delta\mu_j = 0, \tag{4}$$

where μ_i is the chemical potential of the ith chemical reactant, and $\Delta\mu_j = \sum_i S_{ij}\mu_i$ is the chemical potential difference of the jth reaction. The right null space satisfies $\mathbf{SR} = 0$. The Second Law of Thermodynamics can be interpreted as:

$$J_j \Delta\mu_j \leq 0 \ , \tag{5}$$

where the equal sign holds when and only when the reaction is in equilibrium. The thermodynamic constraint on the flux is expressed as: for a flux vector $\mathbf{J}$ to be feasible, there must exist a vector $\boldsymbol{\mu}$ for which Eqs. (4) and (5) are satisfied. In practice, this constraint is difficult to implement. As an alternative, we have introduced an algorithm that is based on the sign structure of the null space, which is the subject of a forthcoming publication [7].

The stoichiometric conditions on thermodynamically allowable and mass-balanced fluxes are a set of mathematical rules that should be followed by any reaction system of a given stoichiometry. These rules alone, however, are not sufficient information to understand and predict the behavior of living metabolic systems, because the stoichiometric conditions fail to constrain systems to behave in a unique way. In fact, it is this inherent unconstrained flexibility of metabolic systems that contributes to their robust ability to maintain Claude Bernard's *milieu interior*.

To model, with specificity and accuracy, the dynamic variables—reactant concentrations and their rates of change—in a living system, requires adopting mechanistic models that relate these variables to the fluxes among them. In doing so, it is paramount to remember that the foundational principles, the basic stoichiometric rules, must apply, regardless of the formulation of the detailed reaction mechanisms. As "La vie..n'est autre chose qu'un phénomène physique"(life is nothing else than a physical phenomenon) [26], our models of living things should not violate fundamental principles of physical chemistry.

3 Theory and Modeling of Biochemical Systems

3.1 Enzyme Mechanisms

Standard approaches to modeling biochemical kinetics begin with mass action relationships [44]. For example, for the simple unimolecular reaction

$$S \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} P, \tag{6}$$

we have rate equations:

$$\frac{d[S]}{dt} = -\frac{d[P]}{dt} = -k_{+1}[S] + k_{-1}[P] \tag{7}$$

where $[S]$ and $[P]$ are the concentrations[1] of species S and P. It is important to remember that Eq. (7) is really based on two physiochemical principles. First, consider the stoichiometry:

$$\frac{d[S]}{dt} = -\frac{d[P]}{dt} = -J_1 + J_2 \ , \tag{8}$$

and second, the rate law according to mass action: $J_1 = k_1[S]$, $J_2 = k_2[P]$. As we shall show below, while the form of the rate law can vary, the stoichiometry is more fundamental.

The above case does not consider the existence of enzyme-catalyzed intermediates that can enhance the turnover between S and P. The well-known Michaelis-Menten model incorporates a substrate-enzyme intermediate step:

$$\xrightarrow{J} S + E \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} SE \underset{k_{-2}}{\overset{k_{+2}}{\rightleftharpoons}} E + P \xrightarrow{J} \ . \tag{9}$$

Assuming the system to be in steady state (the net turn over from S to P is constant and balanced by a flux J of S into and P out of the system) we arrive at the Michaelis-Menten law for the flux:

$$J = J_1 - J_2 = \frac{E_o\,(k_{+2}[S]/K_{M,S} - k_{-1}[P]/K_{M,P})}{1 + [S]/K_{M,S} + [P]/K_{M,P}}, \tag{10}$$

where E_o is the total concentration (free plus bound) of enzyme present, $K_{M,S} = (k_{+2} + k_{-1})/k_{+1}$, and $K_{M,P} = (k_{+2} + k_{-1})/k_{-2}$. This model predicts a chemical equilibrium ($J = 0$) when $[P]/[S] = J_1/J_2$ = $K_{eq} = (k_{+1}k_{+2})/(k_{-1}k{-}2)$. This is known as the Haldane relation [44].

In Eq. (9), as in Eq. (6), all steps are considered reversible, allowing for a physically realistic finite equilibrium. The action of a catalyst as in Eq. (9) does not change the overall thermodynamic equilibrium associated with the system represented in Eq. (6). When $k_{-2} = 0$, we have the familiar irreversible Michaelis-Menten scheme, for which the $K_{M,P}$ terms disappear from Eq. (10). In this case, the resulting equilibrium constant is infinite, and it is no longer possible to characterize the thermodynamics of the overall reaction $S \rightleftharpoons P$.

In practice, the kinetics of many enzyme-catalyzed biochemical reactions are more complex than the single-step Michaelis-Menten model described above. In 1956, King and Altman [25] introduced a systematic method for obtaining the form of steady-state flux laws from diagrams of a given enzyme mechanism (i.e., the collection of intermediate states and the various routes of transition between them). In a concise and accessible chapter, Cornish-Bowden [12] outlines the King-Altman method along with several approaches to simplifying the approach for complex mechanisms. A more comprehensive treatment is found in [44]. With modern computers and symbolic algebra packages, the exercise of deriving steady-state flux expressions from complex mechanisms may be done automatically.

[1] Strictly speaking, if the activity of a species changes with concentration, then the effective rate constants in Eq. (7) are no longer independent of concentration, resulting in a nonlinear system.

When nonlinear expressions such as Eq. (10) can be obtained for all of the reactions of a given system, a complete model in the form of Eq. (1) can be constructed. Large-scale examples of such systems can be found in [24, 57, 28, 29].

3.2 Biochemical Systems Theory: the S-System Approach

While the systems approach has entered the general consciousness of biochemistry only in recent years [9], the original idea can be traced to the early 70's or even earlier. M.A. Savageau and his coworkers, through Biochemical Systems Theory (BST), have continuously championed the cause [37, 38, 39, 40, 54]. Throughout its development, BST has been applied to many different biological research areas such as immunology, molecular genetics, even epidemiology and population dynamics. The approach, however, is not without controversy.

The most important contributions of the BST are inherent recognitions of: (1) the importance of a network of biochemical reactions to cellular functions, (2) the nonlinearity in the governing dynamics of biochemical systems, and (3) the complexities as emerging properties of a reaction network requiring an approach based on systems science.

BST grows out of the realization that the mechanism for an enzyme reaction is often hard to obtain. Even when the mechanism is clearly worked out, it is usually quite complex with many intermediate steps between the substrate binding and the product release. As an approximation, Michaelis-Menten theory assumes that all the intermediate forms of enzyme-substrate complexes remain in steady-state. (This assumption can be mathematically justified if the total enzyme concentration is sufficiently less than that of the substrate [30].) This assumption greatly simplifies the multiple steps of an enzyme reaction, and yields a compact rate law in terms of a ratio of polynomials in the substrate and modifier concentrations (rational functions, e.g., Eq. 10). Initial applications of BST revealed some interesting properties of the kinetic systems in terms of the roots of the polynomial functions [37]. A closely related approach called kinetic polynomial [27] also appears in the literature on reaction systems with heterogeneous catalysis [55] in which the reaction order for catalysts can be greater than unity (i.e., nonlinear).

Rational functions, however, are not necessarily easy mathematical objects to work with. In contrast, the so-called power-law representation can be easily manipulated mathematically; it is also a mathematical form consistent with the law of mass action. Hence, it has gradually become a core constituent of BST [53]. However, writing an enzymatic reaction in an aggregated from in a single rate equation, and using a power-law to approximate the turn-over rate, requires some justification. This justification is extensively discussed in [53], a very readable paper. Using Michaelis-Menten as the reference, it shows how the power-law approximation is more accurate than a linear approximation near a steady-state.

However, some difficulties are associated with the power law representation when dealing with reversible reactions catalyzed by an enzyme. In Michaelis-Menten theory, the product is considered to be a competitor for the substrate (the $[P]$ terms in the denominator of Eq. 10). Hence, the flux should be proportional to $(c_Y)^{\gamma}$ in the power law formulation, with $\gamma < 0$, where γ is a parameter and c_Y is the concentration of the product. This means one cannot represent initial or transient states where $c_Y \to 0$. In addition, it is not clear how to impose the Haldane relationship, which is necessary to develop a thermodynamically valid kinetic theory [44].

The promise of BST is as a theory for large-scale systems. Yet, further difficulties arise when dealing with networks of interacting biochemical reactions. In BST, the dynamics of a biochemical network can be represented by an elegant system of differential equations known as S-systems [41] where ‘S’ stands for *synergetic*.

For a simple reaction system,

$$X_2 \rightleftharpoons X_1 \rightleftharpoons X_3, \tag{11}$$

The S-system representation has the form:

$$\begin{aligned}\frac{dc_1}{dt} &= -J_{12} - J_{13} + J_{21} + J_{31} \\ &= -\alpha_2 c_1^{g_2} - \alpha_3 c_1^{g_3} + \beta_2 c_2^{h_2} + \beta_3 c_3^{h_3} \qquad (12) \\ &\approx -\alpha c_1^{g} + \beta c_2^{h_{12}} c_3^{h_{13}}, \qquad (13)\end{aligned}$$

where the sum of many power-law terms is further approximated by two single power-law terms, one positive and one negative [38].

While the S-system form is both elegant and convenient, there are at least two major disadvantages for the rate law as in Eq. (13): First, the identity of different reactions participating in J has disappeared. Therefore the stoichiometric structure of a reaction network is gone. Knowing everything about Eq. (13) still does not provide us any information on the individual fluxes in $X_2 \rightleftharpoons X_1$ and $X_1 \rightleftharpoons X_3$. Second, because of the stoichiometry has not been preserved, this system of differential equations in general will not satisfy the basic stoichiometric conservation laws, as described above. For the example in Eq. (11), the stoichiometric constraint $d(c_1 + c_2 + c_3)/dt = 0$ should be imposed. Although this constraint remains at the level of Eq. (12), it is lost in the final reduced power-law representation (Eq. 13). This means that computational simplicity is achieved only by loss of physiochemical accuracy.

3.3 Metabolic Control Analysis

Metabolic control analysis (MCA) is concerned with the perturbation of a steady-state of a metabolic network in response to changing enzyme concentrations (i.e., gene expression level) and substrate concentrations. MCA is not a tool for dynamic modeling of biochemical systems. More precisely, in MCA we are interested in characterizing the relationships between arbitrary enzyme i and the arbitrary flux J_k, in steady state. Even when the enzyme i is not catalyzing reaction k, its activity may influence the flux J_k through network interactions in a biochemical system.

To probe the quantitative relationship between enzyme i and flux J_k near a steady-state, the natural choice is sensitivity analysis from standard statistics:

$$C_i^k = \frac{\partial \ln J_k}{\partial \ln e_i} \ , \tag{14}$$

where e_i is the activity (or activity coefficient times concentration) of enzyme i. Because of the network connectivity, there are mathematical relations that these Flux Control Coefficients C_i^k have to satisfy. One of such relation, called summation rule, is the central piece of MCA.

For a given metabolic network, one can compute the flux control coefficients if all the enzyme mechanisms and rate constants are known or, alternatively, one can experimentally determine the flux control coefficient from perturbation and measurement on a biochemical reaction system. The former is a simple mathematical problem so we shall focus on the latter. More discussions of the details can be found in [17, 19, 20, 56].

While the traditional MCA emphasizes steady-states and mathematical derivations, recent developments by Reder, Delgado and Liao [36, 14] focus on obtaining flux control coefficients from measurements on transient, linear relaxation of concentrations to steady-state. The key result needed for the analysis is the summation rule:

$$\sum_{i=1}^{N} C_i^k \left(\frac{J_i(t)}{J_i^*} \right) = 1 \ , \tag{15}$$

where $J_i(t)$ is the flux in any linear transient and $J_i^* = J_i(\infty)$ is the corresponding flux in steady-state.

The key step in the dynamic MCA (dMCA) is to obtain the transient fluxes $J_i(t)$ from measurements on the rate of concentration change dc_k/dt. This problem is an under-determined linear algebra problem without unique solution since (Eq. 1),

$$\frac{dc_k}{dt} = \sum_{i=1}^{N} S_{ki} J_i \ , \tag{16}$$

and the right null space of the stoichiometric matrix $\mathbf{S}$ has non-trivial solution. This is precisely the mathematical problem confronted by the FBA.

In dMCA, the steady state $\mathbf{J}^*$ is assumed to be known. Based on this information, one can computationally obtain $\mathbf{J}$ from a given set of observed concentration changes $d\mathbf{c}/dt \neq 0$. An optimality condition assumes that the solution to the Eq. (16) occurs at the shortest possible distance from the steady-state $\mathbf{J}^*$. The resulting fluxes can be obtained by starting with arbitrary solution to Eq. (16), denoted $\widetilde{\mathbf{J}}$, and then constructing

$$\mathbf{J} = \widetilde{\mathbf{J}} - \mathbf{J}^{\perp} + \mathbf{J}^* \ , \tag{17}$$

where $\mathbf{J}^{\perp}$ is the projection of $\widetilde{\mathbf{J}}$ in the null space[2] of $\mathbf{S}$.

It has been shown [36, 14] that

$$\left\{ \frac{J_i(t) - J_i^*}{J_i^*} \middle| i = 1, ..., N; t \geq 0 \right\} \tag{18}$$

[2] The null space of $\mathbf{S}$ is defined as the subspace of $\Re^N$ for which $\mathbf{SJ} = 0$ is satisfied.

as a set of vectors in linear space, expands exactly the same subspace of the elasticity coefficients,

$$\epsilon_k^i = \frac{\partial \ln J_i^*}{\partial \ln c_k},$$

the second key quantity in MCA. Hence, the solution to Eq. (15) gives the complete set of flux control coefficients. This result in fact serves an alternative proof for Eq. (15), the central piece of dMCA.

3.4 Flux Balance Analysis and Stoichiometric Network Theory

The recent surge of the flux-balance based computational analyses of metabolic networks owes to the unifying ability in integrating null-space analysis of stoichiometric matrices, biological hypothesis-driven constraint-based optimization, and available bioinformatic data on cell metabolism.

The theoretical tools for mathematical analysis of the null space of stoichiometric matrices trace back to the work done by chemical kineticists in 70s and 80s [10, 11, 50]. A quite complete mathematical approach based on the flux cone and its generating vectors[3] (with certain modification called internal representation in [10], elementary modes in [43], and extreme pathways in [42]) was developed. Clarke's stoichometric network analysis (SNA) [10, 11] combines the null space analysis with the analysis of dynamic stability, representing a viable approach for extending FBA beyond steady-state applications.

A powerful algorithm for computing the generating vectors is presented by Schuster et al. [43, 42] and is based on a German reference [32]. The approach can be traced back to Fourier [16] and is known as Fourier-Motzkin double description method [13, 18]. We have summarized the basic idea in the Appendix to help understand the algorithm. We point out that R.J. Duffin [16], who had trained some of the greatest pure and applied mathematicians of 20th century, [31, 8, 46], had devoted his lifework to the theories of nonlinear electrical networks in terms of abstract algebraic topology and differential geometry, as well as practical results in mathematical programming and optimization. Duffin's work is a treasure which proves highly relevant to current constraint-based modeling of biochemical networks.

A key concept in stoichiometric analysis is setting a convention for what is a flux. If one takes all reaction fluxes to be unidirectional, $J_i \geq 0$, then reversible reactions must be represented using separated forward and backward fluxes. This approach, combined with the equality $\mathbf{SJ} = 0$, cogently defines a polyhedral flux cone in the positive quadrant. While this convention for fluxes is mathematically satisfactory, it obscures the nature of chemical reactions. In Fig. 1A the linear combination of the two cycles on the left is not identified as equivalent to the single cycle on the right. For a reversible reaction with forward and backward fluxes J_ℓ^+ and J_ℓ^-, the net flux which is what one needs in the FBA, is $J_\ell = J_\ell^+ - J_\ell^-$. So, this convention introduces unnecessary degeneracy and undeterminancy.

[3] The null space of $\mathbf{S}$ lies within an abstract mathematical entity called a feasible "cone" in linear analysis. The "generating vectors" are the edges of the feasible cone in $\Re^N$.

Different conventions for J's can leads to different conclusions. If one takes all the J's to be irreversible and unidirectional, then the cycles on the left-hand-side of Fig. 1A in fact represent a compound mode with two independent cycles (this is not an elementary mode, as defined by Schuster and Hilgetag [43], since one can eliminate the enzyme associated with the reversible step and still preserve a flux-balanced steady-state; neither it is an extreme pathway since it is a linear combination of two extreme pathways) and it is not equivalent to the single cycle on the right-hand side. However, if one takes all J's to be reversible as convention, then the sum of cycle I and II gives cycle III, which is itself elementary (Fig. 1B). Thus, in terms of the reversible flux convention, one will also be interested in finding a minimal cycle base for a null space [21, 22].

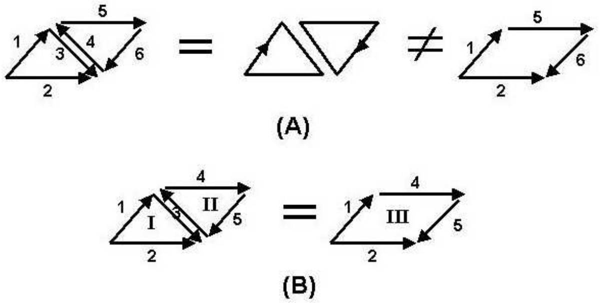

Fig. 1. In (A), the flux convention is set for all unidirectional fluxes. There are total six dimensions. Therefore we have $(1, 1, 1, 1, 1, 1) = (1, 1, 1, 0, 0, 0) + (0, 0, 0, 1, 1, 1) \neq (1, 1, 0, 0, 1, 1)$. The mathematics is quite different for (B) in which there are only five fluxes. Hence $(1, 1, 1, 0, 0) + (0, 0, -1, 1, 1) = (1, 1, 0, 1, 1)$.

Finally, when combining the null space analysis (i.e., characterization of the polyhedral flux cone) and a linear objective function, plus a corresponding second set of inequalities $J_\ell \leq J_\ell^*$, searching for optimal solution becomes a linear programming problem (LPP). (The second set of inequalities is necessary for the subspace on which the objective function is not a constant.) It is important to point out, however, that the extreme points of the polytope associated with the LPP: $\mathbf{SJ} = 0$, $0 \leq J_\ell \leq J_\ell^*, \max\left(\mathbf{b} \cdot \mathbf{J}\right)$, do not all coincide with the extreme rays of the cone (see Fig. 2 for a simple counter example). Furthermore, if a nonlinear objective function is introduced (e.g., the minimal heat dissipation of the network), then the optimal solution in general will lie inside the polytope. Therefore, having a minimal cycle base will be invaluable for nonlinear optimization problems.

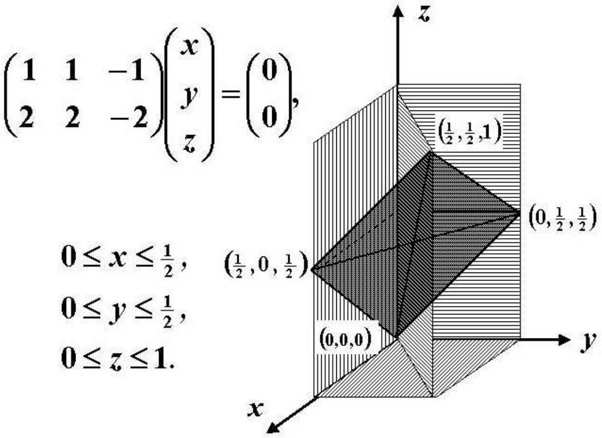
$\begin{pmatrix} 1 & 1 & -1 \\ 2 & 2 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \end{pmatrix}$,
$0 \le x \le \frac{1}{2}$,
$0 \le y \le \frac{1}{2}$,
$0 \le z \le 1$.
z
$(\frac{1}{2}, \frac{1}{2}, 1)$
$(0, \frac{1}{2}, \frac{1}{2})$
$(\frac{1}{2}, 0, \frac{1}{2})$
$(0,0,0)$
y
x

Fig. 2. The LPP shown on the left defines a polytope, a plane, with vertices $(0, 0, 0)$, $(1/2, 0, 1/2)$, $(0, 1/2, 1/2)$, and $(1/2, 1/2, 1)$. Note that the last vertex is not on the generating vectors, i,e., the edge of the cone defined by the equality in the positive octant.

4 Large-Scale Model Building

4.1 Modular Principles

The need to develop computational models encompassing major portions of biochemical and genetic regulatory networks and their control is evident. Integrative models are the key to providing context for the individual reactions and for eliciting an understanding of the influences of the various components upon whole system behavior. The methods for which theory was described in Section 3 are reduced to practice, for example, in formulating a virtual cell model, or in constructing models of more limited expanse, such as cellular energetics.

Small portions of a metabolic system can be described in detailed form. For a system such as glycolysis, it is useful to develop several representations providing different levels of complexity, or accuracy, or computational speed. Having the highest level of precise detail is not compatible with computing the solutions the fastest, yet speed is required to allow widespread exploration to gain insight, develop predictions, or optimize the fits of models solutions to experimental data. Making compromises for specific purposes is therefore essential.

Larger, more all-encompassing models are best composed of smaller modules, each of which has been previously validated by comparisons with data, and verified for computational accuracy. The individual modules must adhere to a pre-chosen standard and provide a scientifically accurate representation of the system, using semantics compatible with those of the larger system. Individual modules are best developed and maintained by individual investigators or groups who are expert in the particular science. Models are merely working hypotheses that must be kept at the forefront of the field if they are to be useful as tools for experiment design and for data analysis. Leaving them in model repositories tended by technical staff relegates them to obsolescence in short time. Certain principles and practices should be upheld in order that modules be actively sustainable. Here, we enumerate the essential principles that we attempt to adhere to in constructing the “eternal” or “sustainable cell” model [5]. The list begins with the science, and extends to matters of style, convenience, and dissemination to the scientific community.

1. Write model code to conserve mass, charge, volume, energy, and redox state.
2. Define all variables and parameters, along with name and units.
3. For linking purposes, identify all inputs and outputs.
4. Identify all assumptions and approximations.
5. Identify all information sources.
6. Write the code for maximal computational speed.
7. Provide operations manuals and tutorials for developed models.
8. Publish models on the web, so that they can be run or downloaded.
9. Establish open forum discussion of models and modules.

Speed, only number six on our list, is vital: one needs to compute at the speed of thought in order to facilitate exploration and the gain of insight into biological processes. Speed is also critical to the use of models as tools to analyze experimental data through automated optimization procedures.

4.2 Linking of Modules to Compose Larger Models

Of the list above, the first five are essential to using a module as a component in a more comprehensive system. The first, conservation, provides some assurance of self consistency; furthermore if each building block fulfills conservation requirements, then determination of conservation for a multi-component model is simplified. The second, definitions and units, defines content. The third, the list of all the inputs and outputs automatically provides the minimal list of links to other modules; the list of links must be extended to include variables which are common to two modules and variables which influence the behavior of a component of another module, e.g. the effect of cAMP on contractility. The fourth, identifying assumptions and approximations, introduces new possibilities into the composite model, such as the ability to ask “does the combining of modules allow for the elimination of certain approximations or assumptions?” and “what additional assumptions might become implicit in the combining of modules?” Such assumptions must be made explicit. The remaining three points are essential for documentation and dissemination.

The process of linking modules is suggested by the composite model in Figure 3. The upper panel shows a set of modules for intermediary metabolism that are linked with known stoichiometry. Consider each to be a separate model, each the evolutionary result of years of research. Linking them through the stoichiometric relationships of input substrates and output products appears straightforward, and results in the integrated model shown in the lower panel. It appears as if the sanctity of the individual modules remains intact, and in fact it is almost this simple. Problems arise, not so much in accounting for NADH and other slowly changing variables, but in accounting for the influences of rapidly changing calcium and hydrogen ion concentrations in some circumstances.

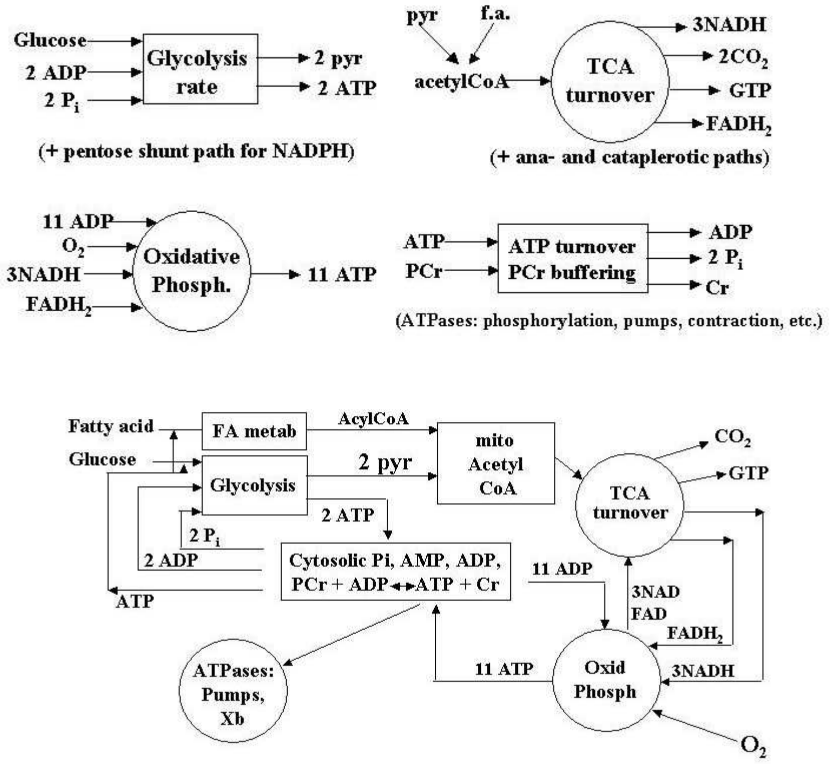

Fig. 3. Building composite modules from pre-built modules in intermediary metabolism. *Upper panel*: Individual models for glycolysis, Krebs (TCA) cycle, oxidative phosphorylation and nucleotide energetics showing their main inputs and outputs. *Lower panel*: Combining the modules from the upper panel, and adding fatty acid metabolism gives a composite model of intermediary metabolism. Stoichiometric balance is maintained, and the modules remain distinct from each other.

The argument that models are best developed and maintained by a group working in the particular field leads to contradictions when models become modules of larger system models. The expertise required to develop and maintain a given model may not be available in the group that chooses it as the best version of the desired component of a higher level integrative model. From the technical point of view, putting together two models from different sources is not too difficult when both submodels or modules can be described by ODEs (ordinary differential equations). Using a simulation system like JSim (from nsr.washington.edu), or Madonna (from www.berkeleymadonna.com/index.html), or SAAM (from www.saam.com), or Gepasi (from www.gepasi.org), one simply combines the two modules into a common piece of source code, combining all those equations which have common variables. The process sounds simple, but the problem is that one of the originating groups is not necessarily expert in the field of the other, so that unless all the common variables have been defined with the same names in both modules their identity will not be easily recognized. Clear semantics requires lists of synonyms.

The second issue is how to build a composite model out of modules while maintaining the identity of the code of the module so that it can be replaced automatically when the originating group improves the module. Ideally the composite model should be reconfigured whenever it is judged that improvements have been achieved in any particular module. Automating this is possible when common variables are named identically; it is also possible, but requires human intervention, to define equivalences, when the variable names in each module are not identical. There is a trade-off here: when the variable names are identical, the combining of the modules can be automated because the equations for the common variable can be automatically combined, as has been achieved by Gary Raymond in our laboratory. The cost is that the two source codes are now intermixed. For computational reasons this is good since it minimizes the numbers of different variables and facilitates solving the whole system simultaneously. But what it costs is that a composite model composed of a large set of modules must be entirely formed anew when a module is to be replaced.

Separating modules from one another can help to obtain computational speed. Solving a large set of simultaneous equations as a whole gives high accuracy when the system is non-stiff and linear, but is computationally costly when it is composed of non-linear equations and/or has dramatically different time constants in different parts of the model. Then one would like then to solve separately those submodels that have rate or time constants that are relatively slow and to solve at higher frequencies those submodels that have time constants orders of magnitude faster. Allowing difference in time steps from one module to another greatly reduces the “stiffness” of the overall system and increases computation speed. This argument favors keeping modules separated even while linked in the composite model, and enhances the incentive to use automated methods for composite model building. Last but not least, attentions should be directed to the numerical accurary of composite models for which fractional steps and exponential splitting could be introduced [49, 45].

A situation in which modular separation can be maintained occurs whenever the common variables change slowly relative to the internal rates of the modules to which they are relevant. An example is ATP, which is such large concentration normally that its concentration changes only very slowly even with dramatic changes in circumstances. By treating ATP as an external variable from the point of view of the individual modules its concentration can be considered constant during a time step; by preserving the fluxes of ATP into or out of the relevant modules, ATP concentration can be correctly represented still to a high degree of accuracy without solving for it at the high rates required for the fast modules. Then modules can be computed separately, bringing their solutions together at relatively long time intervals. This same approach lends itself to parallel computation of the different modules on different CPUs. Since computation time is a major factor for metabolic and electrophysiologic cell models and a huge issue for integrated organ models, such approaches need much further development to achieve maximal efficiency in computation and to enhance progress in investigation.

5 Summary and Conclusions

To summarize, we view biochemical systems as represented at the basic level as networks of given stoichiometry. Whether the steady-state or the kinetic behavior of a given system is explored, the stoichiometry constrains the feasible behavior, according to mass balance and the laws of thermodynamics. Study of the feasible behavior of a system, given a set of stoichiometric constraints, forms the basis of flux balance analysis, and more generally, stoichiometric network theory.

To obtain precise information beyond what stoichiometry can tell, one must develop a kinetic representation of the transformations occurring in a given system. The classical enzyme-kinetic approach is often used to build kinetic models that satisfy the stoichiometric rules. The drawback of this approach lies in the inherent uncertainly involved in assigning kinetic mechanisms to biochemical fluxes and in identifying the associated parameter values.

As an alternative to classical enzyme kinetics, the biochemical systems theory alleviates some of these drawbacks. At the heart of BST, the S-system is a versatile and useful phenomenological nonlinear dynamic equation system which has found many applications. However, as a modeling paradigm for metabolic reaction networks, it lacks the structure for introducing several fundamental, physiochemical elements of biochemical reactions: stoichiometry, mass and energy balance, and reversibility. These difficulties prevent a true integration of mechanistic studies of individual enzymes and simple metabolic reactions into a complex network system. Therefore, as it stands, BST has not fulfilled the need of current systems biochemistry.

Though many of the useful tools in biochemical analysis (enzyme kinetics, metabolic control analysis, stoichiometric network theory) have been well established in the literature for decades, new developments in theory continue to enhance the usefulness of these tools. In particular, recent additions of thermodynamic considerations into stoichiometric network theory [6, 35], and the development of a dynamic metabolic control analysis [36, 14], hold promise.

Whether or not these new tools provide the key to improving our understanding of the operation of whole-cell systems remains to be seen. In any case, we can be confident that whatever future technologies prove useful for biochemical systems analysis, those approaches will not conflict with the basic stoichiometric principles. As we move up the hierarchical ladder of system complexity, toward biophysically realistic representations of large-scale systems, the foundation must remain intact.

References

1. Alberty, R.A. (1991) Equilibrium compositions of solutions of biochemical species and heats of biochemical reactions. *Proc. Natl. Acad. Sci. USA*, **88**, 3268-3271.
2. Bailey, J.E. (1991) Toward a science of metabolic engineering. *Science*, **252**, 1668-75.
3. Bassingthwaighte, J.B. & Goresky, C.A. (1984) Modeling in the analysis of solute and water exchange in the microvasculature, in *Handbook of Physiology. Sect. 2, The Cardiovascular System Vol IV, The Microcirculation.* E.M. Renkin and C.C. Michel eds., Am. Physiol. Soc., Bethesda, MD. pp. 549-626.
4. Bassingthwaighte, J.B. & Goresky, C.A. & Linehan, J.H., eds. (1998) *Whole organ approaches to cellular metabolism: permeation, cellular uptake, and product formation.* Springer, New York.
5. Bassingthwaighte, J.B. (2001) The modeling of a primitive 'sustainable' conservative cell. *Phil. Trans. R. Soc. Lond. A.*, **359**, 1055-1072.
6. Beard, D.A., Liang, S.-D., & Qian, H. (2002) Energy balance for analysis of complex metabolic networks. *Biophys. J.*, **83**, 79-86.
7. Beard, D.A. & Qian, H. (2003) Thermodynamic Constraints for Cellular Metabolic Analysis. manuscript in preparation.
8. Bott, R. & Duffin, R.J. (1949) Impedance synthesis without use of transformers. *J. Appl. Phys.*, **20**, 816.
9. Chong, L. & Ray, L.B. (2002) Whole-istic biology. *Science*, **295**, 1661-1661.
10. Clarke, B.L. (1980) Stability of complex reaction networks. *Adv. Chem. Phys.*, **43**, 1-215.
11. Clarke, B.L. (1988) Stoichiometric network analysis. *Cell Biophys.*, **12**, 237-253.
12. Cornish-Bowden, A. (1976) *Fundamentals of Enzyme Kinetics* 3rd Ed., Butterworths, London. Chapter 4.
13. Dantzig, G.B. & Eaves, B.C. (1973) Fourier-Motzkin elimination and its dual. *J. Combin. Theo. A*, **14**, 288-297.
14. Delgado, J.P. & Liao, J.C. (1991) Identifying rate-controlling enzymes in metabolic pathways without kinetic parameters. *Biotechnol. Prog.*, **7**, 15-20.
15. Dines, L.L. (1919) Systems of linear inequalities. *Annals Math.*, **20**, 191-199.
16. Duffin, R.J. (1974) On Fourier's analysis of linear inequality systems. *Math. Progm. Study*, **1**, 71-95.
17. Fell, D.A. & Sauro, H.M. (1985) Metabolic control and its applications Eur. *J. Biochem.*, **148**, 555-561.
18. Fukuda, K & Prodon, A. (1996) Double description method revisited. In *Combinatorics and Computer Science* (Lect. Notes Comput. Sci., Vol. 1120, M. Deza, R. Euler, & I. Manoussakis, eds), Springer-Verlag, pp. 91-111.
19. Giersch G. (1988) Control analysis of metabolic networks. I. Homogeneous functions and the summation theorems for control coefficients. *Eur. J. Biochem.*, **174**, 509-513.
20. Giersch G. (1988) Control analysis of metabolic networks. II. Total differentials and general formulation of the connectivity relations. *Eur. J. Biochem.*, **174**, 515-519.
21. Golynski, A. and Horton, J.D. (2001) A polynomial time algorithm to find the minimal cycle basis of a regular matroid. Preprint, http://www.stfx.ca/academic/mathcs/apics2001/discrete/Joseph.pdf.
22. Horton, J.D. (1987) A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM J. Comput.*, **16**, 358-366.
23. Ibarra, R.U., Edwards, J.S., & Palsson, B.O. (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, **420**, 186-189.

24. Kauffman, K.J., Pajerowski, J.D., Jamshidi, N., Palsson, B.O., & Edwards, J.S. (2002) Description and analysis of metabolic connectivity and dynamics in the human red blood cell. *Biophys. J.*, **83**, 646-62.
25. King, E.L. & Altman, C. (1956). A schematic method of deriving the rate laws for enzyme-catalyzed reactions. *J. Phys. Chem.*, **60**, 1375-1381.
26. Lamarck, J.B. (1830) *Systeme analytique des connaissances positives de l'homme*, J.B. Baillire, Paris.
27. Lazman, M.Z. & Yablonskii, G.S. (1991) Kinetic polynomial: a new concept of chemical kinetics. In "*Patterns and Dynamics in Reactive Media*". R. Aris, D.G. Aronson, and H.L. Swinney, Eds., Springer-Verlag, New York, pp. 117-149.
28. Mulquiney, P.J., Bubb, W.A. & Kuchel, P.W. (1999) Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: in vivo kinetic characterization of 2,3-bisphosphoglycerate synthase/phosphatase using 13C and 31P NMR. *Biochem. J.*, **342**, 576-580.
29. Mulquiney, P.J. & Kuchel, P.W. (1999) Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: equations and parameter refinement. *Biochem. J.*, **342**, 581-596.
30. Murray, J.D. (2002) *Mathematical Biology I: An Introduction*. 3rd Ed., Springer-Verlag, New York.
31. Nasar, S. (1998) *A beautiful mind: a biography of John Forbes Nash, Jr*, Simon & Schuster, New York.
32. Nozicka, F., Guddat, J., Hollatz, H., & Bank, B. (1974) *Theorie der linearen parametrischen Optimierung*, Akademie-Verlag, Berlin.
33. Oliveira, J.S., Bailey, C.G., Jones-Oliveira, J.B., & Dixon, D.D. (2001) An algebraic-combinatoiral model for the identification and mapping of biochemical pathways. *Bull. Math. Biol.* **63**, 1163-1196.
34. Papadimitriou, C.H. & Steiglitz, K. (1998) *Combinatorial Optimization: Algorithms and Complexity*, Dover, New York.
35. Qian, H., Beard, D.A., & Liang, S.-D. (2002) Stoichiometric Network Theory for Nonequilibrium Biochemical Systems. *Eur. J. Biochem.*, **270**, 415-421.
36. Reder, C. (1988) *J. Theoret. Biol.*, **135**, 175-201.
37. Savageau, M.A. (1969) Biochemical Systems Analysis. I. *J. Theoret. Biol.*, **25**, 365-369.
38. Savageau, M.A. (1969) Biochemical Systems Analysis. II. *J. Theoret. Biol.*, **25**, 370-379.
39. Savageau, M.A. (1970) Biochemical Systems Analysis. III. *J. Theoret. Biol.*, **26**, 215-226.
40. Savageau, M.A. (1976) *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology.* Addison-Wesley, Reading, MA.
41. Savageau, M.A. & Voit, E.O. (1987) Recasting nonlinear differential equations as S-systems: a canonical nonlinear form. *Math. Biosci.*, **87**, 83-115.
42. Schilling, C.H., Letscher, D., & Palsson, B.O. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theoret. Biol.*, **203**, 229-248.
43. Schuster, S. & Hilgetag, C. (1994) On elementary flux modes in biochemical reaction systems at steady-state. *J. Biol. Syst.*, **2**, 165-182.
44. Segel, I.H. (1975) *Enzyme Kinetics.* Wiley Interscience, New York.
45. Sheng, Q. (1994) Global error estimates for exponential splitting. *IMA J. Numer. Anal.*, **14**, 27-56.
46. Smale, S. (1972) On the mathematical foundations of electrical circuit theory. *J. Diff. Geom.*, **7**, 193-210.

47. Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., & Gilles, E.D. (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190-193.
48. Stephanopoulos, G. (1994) Metabolic engineering. *Curr. Opin. Biotechnol.*, **5**, 196-200.
49. Strang, G. (1968) On the construction and comparison of differential schemes. *SIAM J. Numer. Anal.*, **5**, 506-517.
50. Strasser, P., Stemwedel, J.D., and Ross, J. (1993) Analysis of a mechanism of the chloride-iodide reaction. *J. Phys. Chem.* **97**, 2851-2861.
51. Stryer, L. (1981) *Biochemistry.* 2nd Ed., W.H. Freeman, San Francisco.
52. Van Dien, S.J. & Lidstrom, M.E. (2002) Stoichiometric model for evaluating the metabolic capabilities of the facultative methylotroph Methylobacterium extorquens AM1, with application to reconstruction of C(3) and C(4) metabolism. *Biotechnol. Bioeng.*, **78**, 296-312.
53. Voit, E.O. & Savageau, M.A. (1987) Accuracy of alternative representations for integrated biochemical systems. *Biochem.*, **26**, 6869-6880.
54. Voit, E.O. (2000) *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists.* Cambridge Univ. Press, New York.
55. Wei, J. & Prater, C.D. (1962) The structure and analysis of complex reaction systems. Adv. Catal., **13**, 203-392.
56. Westhoff, H.V. & Chen, Y.D. (1984) How do enzyme activity control metabolic concentrations? *Eur. J. Biochem.*, **142**, 425-430.
57. Wiback, S.J. & Palsson, B.O. (2002) Extreme pathway analysis of human red blood cell metabolism. *Biophys J.*, **83**, 808-818.

Appendix

Let $\mathbf{b}_1$, $\mathbf{b}_2$, ..., $\mathbf{b}_n$ be an n-dimensional, orthonormal base set. In terms of this base set, n-dimensional vectors $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_m$ can be written as column vectors. Matrix $\{a_{ij}|0 \leq i \leq n, 0 \leq j \leq m\}$ appended with the vectors $\mathbf{b}$'s is called a tableau [16]:

$$\left(\begin{array}{c|cccc} \mathbf{b}_1 & a_{11} & a_{21} & \ldots & a_{m1} \\ \mathbf{b}_2 & a_{12} & a_{22} & \ldots & a_{m2} \\ \mathbf{b}_3 & a_{13} & a_{23} & \ldots & a_{m3} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{b}_n & a_{1n} & a_{2n} & \ldots & a_{mn} \end{array}\right) . \tag{19}$$

If one carries out row operation for the above tableau, say

$$\left(\begin{array}{c|cccc} \mathbf{b}_1 & a_{11} & a_{21} & \ldots & a_{m1} \\ \mathbf{b}_2 + c\mathbf{b}_1 & a_{12} + ca_{11} & a_{22} + ca_{21} & \ldots & a_{m2} + ca_{m1} \\ \mathbf{b}_3 & a_{13} & a_{23} & \ldots & a_{m3} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{b}_n & a_{1n} & a_{2n} & \ldots & a_{mn} \end{array}\right) , \tag{20}$$

then the first column on the left is no longer a set of orthonormal vectors. However, the entries (i, j) of the matrix on the right still gives the inner product $(\mathbf{b}'_i, \mathbf{a}_j)$ where

$\mathbf{b}_i'$ is the current ith vector on the left, after the operation. If one carries out Gauss-Jordan elimination by row operation and reaches

$$\left(\begin{array}{c|cccc} \mathbf{b}_1' & 1 & 0 & \ldots & 0 \\ \mathbf{b}_2' & 0 & 1 & \ldots & 0 \\ \mathbf{b}_3' & * & * & \ldots & * \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{b}_n' & 0 & 0 & \ldots & * \end{array}\right), \tag{21}$$

then one has $(\mathbf{b}_1', \mathbf{a}_1) = 1$ and $(\mathbf{b}_1', \mathbf{a}_2) = (\mathbf{b}_1', \mathbf{a}_3) = \ldots = (\mathbf{b}_1', \mathbf{a}_m) = 0$. In other words, vector $\mathbf{b}_1'$ lies in the linear subspace defined by the intersection of hyperplanes $(\mathbf{x}, \mathbf{a}_2) = (\mathbf{x}, \mathbf{a}_3) = \ldots = (\mathbf{x}, \mathbf{a}_m) = 0$; and $(\mathbf{b}_1', \mathbf{a}_1) = 1$.

One notices that the above algorithm is closely related to that for inverting a matrix and computing its null space. For a system of linear equations, the null space plays a fundamental role in the solution to and characterization of the problem. The null space of a matrix $\mathbf{S}$ can be expressed in terms of the loop matrix $\mathbf{R}$: $\mathbf{SR} = 0$. Analogously, for a system of linear inequalities [15], the *solvent matrix* $\mathbf{Y} = \{Y_{ij} | Y_{ij} \geq 0, \sum_j S_{ij} Y_{jk} = 0\}$ plays the fundamental role in the solution to and characterization of the convex system.

The above method can be used to compute the generating vectors of the polyhedral cone defined by a set of linear inequalities $(\mathbf{x}, \mathbf{a}_\ell) \geq 0$, $(\ell = 1, 2, ..., m)$. In this case, proposed by Fourier, pairwise eliminations with positive multipliers and only additions are carried out [13, 16] to preserve the inequalities, and one stops when all the nonzero entries are positive. For a column with P, Q, and R number of positive, negative, and zero terms respectively, the number of pairwise eliminations is determined by the *expansion number* $PQ + R$ where $P + Q + R = n$ [15]. An efficient pairwise elimination algorithm is designed to have a rule for selection of an optimal order for the elimination steps [16, 32].

Every convex polyhedral cone has two representations, in terms of a set of inequalities (H-representation) or a set of extreme rays (V-representation), respectively. The double description method works with both sets similar to solving the minimal spanning tree problem (MSTP) for a graph. The MSTP can also be solved in terms of the maximal weight forest which belongs to the class of problems known as matroid programming [34, 33].