CHAPTER

9

Identification of Qualitatively Feasible Metabolic Pathways

Michael L. Mavrovouniotis

1. Introduction

Cells function as organized chemical engines carrying out a large number of transformations, called bioreactions or biochemical reactions, in a coordinated manner. These reactions are catalyzed by enzymes and exhibit great specificity and rates much higher than the rates of non-enzymatic reactions. Enzymes are neither transformed nor consumed, but that facilitate the underlying reactions by their presence. The coordination of the extensive network of biochemical reactions is achieved through regulation of the concentrations and the specific activities of enzymes. Single enzyme-catalyzed steps in succession form long chains, called biochemical pathways, achieving the overall transformation of substrates to far removed products.

Biochemical pathways are often described in symbolic terms, as a succession of transformations of one set of molecules (called reactants) into

another set (called products); reactants and products are collectively referred to as metabolites.

In the construction of metabolic pathways one uses enzyme-catalyzed bioreactions as building blocks, to assemble pathways that meet imposed specifications. A class of specifications can be formulated by classifying each available building block, i.e., each metabolite and each bioreaction, according to the role it can play in the synthesized pathways. For example, a set of specifications may include some metabolites designated as required final products of the pathways, other metabolites as allowed reactants or by-products, and some bioreactions as prohibited from participating in the pathways.

Non-obvious alternative pathways, including those that are not known to be present in any single strain, are especially interesting, because they might prompt new discoveries. The complexity and density of intermediary metabolism generally permit a large number of pathways.

Many distinct pathways can be constructed to include the same bioreactions but achieve different transformations. For example, the reactions $A \rightarrow B+C$ and $2B+C \rightarrow D$ can form the pathways $2A \rightarrow \rightarrow D+C$ and $A+B \rightarrow \rightarrow D$, depending on whether the reactions are combined in 2:1 or 1:1 proportions. Thus, a fully specified pathway must include a *coefficient* for each bioreaction, to indicate the proportions at which the constituents are combined.

Systematic synthesis of pathways that satisfy a set of such specifications is relevant in the early steps of the conception and design of a bioprocess, where a pathway must be chosen for the production of the desired product. For the synthesis of desired bioproducts, the operating pathway is a crucial factor in the feasibility of the process and the selection of appropriate cell lines and media

The synthesis of pathways can identify fundamental limitations in the e *anabolism* (synthesis of biomolecules) and *catabolism* (breakdown of biomolecules, e.g. for digestion) of any given cell, because the pathways determine what transformations are possible in principle. Consider, for example, the problem of identifying a mutant strain lacking a particular enzyme. One must define the set of combinations of substrates on which the mutant cell is able to grow by identifying suitable pathways (despite the lack of a particular enzyme) to consume the substrates in question. Such pathways may differ significantly from the standard routes. Thus, systematic generation of pathways is a more reliable way to predict the ability or inability of a mutant strain to grow on specified sets of substrates. Consequently, it can have a significant impact on the identification of mutant strains lacking a certain bioreaction. Conversely, if the target is not elimination of an enzyme but absence of growth on specific sets of substrates, one must pinpoint the enzymes that should be eliminated to *block* all the pathways for the catabolism

of the substrates. The selection of an appropriate set of enzymes depends on the correct generation of all relevant pathways.

The rates of enzymatic reactions are influenced by factors such as the pH and the concentration of metabolites. Furthermore, the regulation of gene expression determines what enzymes are synthesized by a cell, and therefore what bioreactions are available to participate in pathways. Without detailed information on these phenomena, one cannot identify pathways that will definitely be active under given conditions. One can only identify *potential* metabolic pathways which are qualitatively feasible. The qualitative feasibility is confined here to two attributes. First, each bioreaction participating in a pathway must be feasible in the direction in which it is used. Second, the overall stoichiometry (the quantitative relationship between the metabolites involved, expressed as ratios) of the pathway, which derived from a linear combination of the stoichiometries of the constituent bioreactions, must satisfy imposed constraints.

In the following sections, I present an AI method for addressing both the question of judging whether a particular reaction is feasible and the process of taking a collection of reactions and set of constraints and then finding all of the feasible pathways. The next section focuses on thermodynamic feasibility and describes a group-contribution technique that allows the estimation of equilibrium constants of biochemical reactions. This method was implemented in Symbolics Lisp, but is not currently available in executable form; a future version will be implemented in Common Lisp. The remainder of the chapter describes a symbolic approach to the construction of pathways that satisfy stoichiometric constraints.

2. Thermodynamic Feasibility

The feasibility and reversibility of a bioreaction is determined by its equilibrium constant and the concentrations of its reactants (also called substrates) and products. Because intracellular concentrations vary within limited ranges (e.g., 1μM to 5mM), the equilibrium constant alone is sufficient for reaching a qualitative conclusion on a bioreaction's feasibility. In general, a feasible and irreversible reaction is characterized by an equilibrium constant, K, much larger than 1. A feasible and reversible (i.e., feasible in both the forward and reverse directions) reaction is characterized by an equilibrium constant of the order of 1. A reaction that is infeasible in the forward direction but feasible in the reverse direction is characterized by an equilibrium constant much smaller than 1. The quantitative interpretation of these criteria depends on the range of permissible intracellular concentrations for metabolites.

The thermodynamic analysis can also be carried out using the standard Gibbs energy of reaction, $\Delta G^{\circ\prime}$ which is closely related to the equilibrium constant, K:

$$\Delta G^{\circ\prime} = - RT \ln K \quad (1)$$

Here, R is the ideal-gas constant and T the temperature. The standard state for $\Delta G^{\circ\prime}$ is a dilute aqueous solution at T =25°C, pH = 7, and concentrations of compounds (other than H^+, OH^-, and H_2O) equal to 1 M. The standard Gibbs energy of reaction is related to the standard Gibbs energies of formation of its reactants and products. Letting V_i be the stoichiometric coefficient of compound S_i, we can write a reaction (or any transformation with known stoichiometry) as:

$$\sum V_i S_i = 0 \quad (2)$$

Here, V_i is positive for products and negative for reactants. Let $\Delta G_i^{\circ\prime}$ be the Gibbs energy of formation of S_i. The Gibbs energy of reaction, $\Delta G^{\circ\prime}$, is then given by the equation:

$$\Delta G^{\circ\prime} = \sum V_i \, \Delta G_i^{\circ\prime} \quad (3)$$

From the Gibbs energies of formation of a set of compounds one can calculate the Gibbs energy for *any* biochemical transformation within this set of compounds.

Group-Contribution methods [Benson, 1968; Benson *et al.*, 1969; Domalski and Hearing, 1988; Joback and Reid, 1987; Mavrovouniotis *et al.*, 1988; Mavrovouniotis, 1990b; Reid *et al.*, 1977; Reid *et al.*, 1987] have been widely used to estimate numerical values of properties of pure compounds. To estimate a property, one views the compound as composed of functional groups and sums amounts contributed by each group to the overall value of the property.

A given group-contribution method must provide a set of functional groups, which serve as the building blocks for the compounds of interest. The contribution of each group to the thermodynamic property of interest must also be provided, along with the *origin*, a starting value that is used in the estimation (and is constant for all compounds). To estimate the property of a particular compound, one decomposes the compound into groups, and adds the contributions of the groups to the constant origin.

Let C_o be the origin for the property C, and let C_i be the contribution of group g_i which is used N_i times in the compound. The property C for the whole compound is calculated as:

$$C = C_o + \sum N_i \, C_i \quad (4)$$

A group contribution method can be developed using data (i.e., the value of the property for several compounds), to estimate the contributions of the groups to the property of interest. In effect, if the data consist of values of C for a set of compounds, and the molecular structures (hence the N_is) of the compounds are known, then a number of equations of the form of Equation (4) are available, and the unknown origin C_o and contributions C_i can be determined. It is generally desirable to have as many data points as possible

and obtain values for the contributions by minimizing the sum of the square of the errors (multiple linear regression); the error is defined for each data point as the difference between the given value and the value estimated by Equation (4).

If C is a property applicable to reactions, in the linear manner suggested by Equation (3), then data on reactions and data on compounds can be treated uniformly. Reactions can be viewed as collections of groups by subtracting the number of occurrences of each group in reactants from its occurrences in products, because both Equation (3) and Equation (4) are linear. The same linear-combination treatment must also be applied for the contribution of the *origin* and all additional corrections.

We have recently developed a comprehensive group-contribution method [Mavrovouniotis, 1990b, 1991] for the estimation of the Gibbs energies of formation of biochemical compounds, and hence the Gibbs energies and equilibrium constants of biochemical reactions.

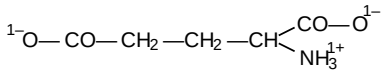
The data used in the regression were taken from several sources [Thauer et al.,1977, Barman, 1969, Barman, 1974, Hinz, 1986, Lehninger, 1975, Lehninger, 1986, Sober, 1970, Edsall, and Gutfreund, 1983] and were screened for gross errors. A large set of groups was used, to cover most biochemical compounds and achieve good accuracy. In addition, corrections were introduced to account for certain group-interactions.

Special groups were used for certain complex compounds with important metabolic roles. For example, the pair NAD^+/NADH was represented by a single group, which represents the structural differences between the two compounds which are relevant for a large number of biochemical reactions. Finally, it should be noted that all compounds and groups were represented in their common state in aqueous solution. The determined contributions of groups have been presented by Mavrovouniotis [1990b, 1991].

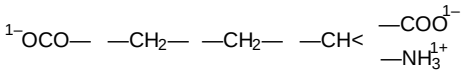
Examples. A few example calculations will be provided here to illustrate the use of the group-contribution method. Consider the estimation of the Gibbs energy of formation of glutamate, whose syntactic formula is shown in Figure 1a. It can be broken down into groups in a straightforward manner, as shown in Figure 1b. The calculation entails the addition of the contributions (multiplied by the number of occurrences of each group) to the fixed contribution of the origin, as shown in Table 1. In this example, no special corrections are needed. The final result is -164.7 kcal/mol, which deviates by 2.4 kcal/mol from the literature value -167.1 kcal/mol [Thauer et al., 1977].

As an example involving a complex cyclic compound, consider next the estimation of the Gibbs energy of formation of ATP, whose structure is shown in Figure 2. Table 2 shows the calculation of the Gibbs energy from the contributions.

Another example is provided in Figure 3 and Table 3 for a biochemical reaction, catalyzed by *alcohol dehydrogenase.* The reaction is decomposed into



(a)

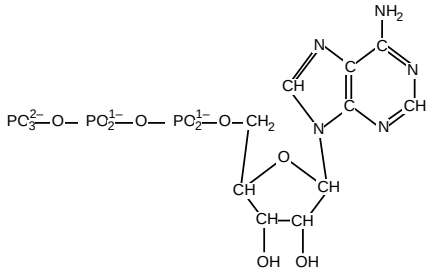


(b)

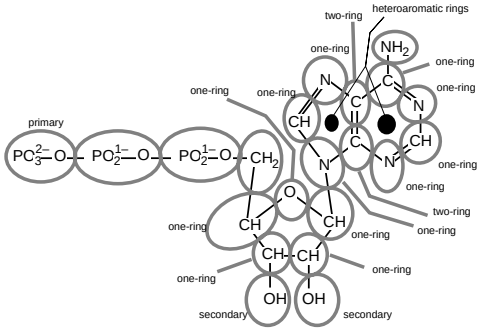
Figure 1: (a) The structure of glutamate. (b) Decomposition of the structure into groups

Group or Correction	Number of Occurrences	Contribution (kcal/mol)	Total Contribution
Origin	1	–23.6	–23.6
$-NH_3^{1+}$	1	4.3	4.3
$-COO^{1-}$	2	–72.0	–142.0
$-CH_2-$	2	1.7	3.4
–CH<	1	–4.8	–4.8
			–164.7

Table 1: Calculation of the Gibbs energy of formation of glutamate from contributions of groups. The contributions are those given by Mavrovouniotis (1991).



(a)



(b)

Figure 2: (a) The structure of ATP (b) Decomposition of the structure of ATP into groups

Group or Correction	Number of Occurrences	Contribution (kcal/mol)	Total Contribution
Origin	1	–23.6	-23.6
$–NH_2$	1	10.3	10.3
$–OPO_3^{1–}$ primary	1	–29.5	-29.5
–OH secondary	2	–32.0	-64.0
$–CH_2–$	1	1.7	1.7
$–OPO_2^{1–}–$	2	–5.2	-10.4
ring –O–	1	–24.3	-24.3
ring –CH<	4	–2.2	-8.8
ring –N<	1	7.6	7.6
ring –CH=	2	9.6	19.2
ring =N–	3	10.4	32.2
ring >C=	1	8.2	8.2
two-ring >C=	2	16.8	33.6
heteroaromatic ring	2	–5.9	-11.8
			–60.8

Table 2: Calculation of the Gibbs energy of formation of ATP from contributions of groups. The contributions are those given by Mavrovouniotis (1991).

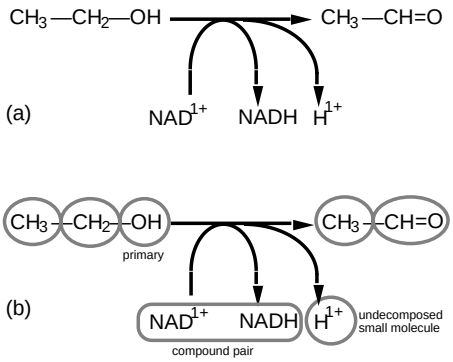


Figure 3 (a) The reaction catalyzed by alcohol dehydrogenase. (b) The reaction is decomposed into groups, so that its Gibbs energy can be estimated.

groups in Figure 3; note that the pair NADH / NAD^+ is considered a single group. The calculation in Table 3 ignores the contributions of the *origin* and the group $–CH_3$, because they are the same for ethanol and acetaldehyde, i.e., they have net number of occurrences equal to zero. The result is 4.8 kcal/mol, which compares well with literature values of 5.5 kcal/mol [Barman, 1969] and 5.9 kcal/mol [Hinz, 1986].

Group or Correction	Number of Occurrences	Contribution (kcal/mol)	Total Contribution
Origin	0		
H^{1+}	1	-9.5	-9.5
NADH *minus* NAD^+	1	4.7	4.7
$–CH_3$	0		
$–CH_2–$	–1	1.7	-1.7
–OH primary	–1	–28.6	28.6
–CH=O	1	–17.3	-17.3
			–4.8

Table 3: Calculation of the Gibbs energy of the reaction catalyzed by alcohol dehydrogenase, from contributions of groups.. The contributions are those given by Mavrovouniotis (1991).

Discussion. Using the contributions and corrections of given by Mavrovouniotis [1990b, 1991], one can estimate the standard Gibbs energy of formation of a biochemical compound, provided that the molecular structure of the compound is known. The standard Gibbs energy and the equilibrium constant of any biochemical reaction can be estimated from the molecular structures of its reactants and products.

The method has broad applicability because it provides the contributions for a comprehensive set of groups. The error is usually less than 2 kcal/mol. Thus, the method provides an acceptable first approximation to the thermodynamics of biochemical systems. Mavrovouniotis [1990b, 1991] provides precomputed values for common metabolites. For compounds that have only small structural differences from the precomputed ones, only the contributions of groups describing the differences need be considered [Mavrovouniotis, 1990b].

A fundamental difficulty in the group-contribution methods for biochemical compounds is that there are often strong interactions among groups due to *conjugation.* The *conjugates* of a compound are alternative arrangements of the valence electrons; a compound that is strongly influenced by conjugation cannot be properly decomposed into groups.

We are currently investigating a new property-estimation framework, named ABC [Mavrovouniotis, 1990a], which is based on using contributions of Atoms and Bonds for properties of Conjugates of a compound. This approach has been enhanced by approximate quantum-chemical analysis and has been demonstrated for simple compounds in the ideal-gas state [Mavrovouniotis, 1990a]. It is expected that the ABC framework will be of great value in estimating their properties of biochemical compounds.

Information on Metabolites
• Set of groups that comprise the molecule
• Standard Gibbs Energy of formation in aqueous solution
• Typical concentrations (only for currency metabolites)
• List of reactions that consume the metabolite
• List of reactions that produce the metabolite
Information on Bioreactions
• Stoichiometry of the reaction
• Standard Gibbs Energy of reaction
• Physiological information on reversibility of the reaction
• List of metabolites the reaction consumes
• List of metabolites the reaction produces

Table 4: Information useful in the synthesis of biochemical pathways from the database of metabolites and bioreaction.

2 Synthesis of Pathways

An approach for the synthesis of biochemical pathway [Mavrovouniotis, *et al* 1990a] is presented here. This section gives the formulation of the problem and the developed algorithm, which is complete and sound. An example showing the step-by-step operation of the algorithm in a small abstract problem is provided, along with a discussion of computational issues. The next section in this chapter presents a case study on the biosynthesis of lysine from glucose and ammonia.

Biochemical pathway synthesis is here the construction of pathways which produce certain target bioproducts, under partial constraints on the available substrates (reactants), allowed by-products, etc. In connection with this formulation of the synthesis problem, it should be noted that:

- A pathway must include all reactions responsible for the conversion of initial substrates to final products, and not merely the steps leading from the intermediary metabolism to the product.
- The pathways sought are not restricted to the already known routes found in textbooks. New pathways are quite acceptable and present the most interest.

In order to construct pathways from bioreactions, one needs a database of metabolites and bioreactions. The information stored in the database for each bioreaction and each metabolite is shown in Table 4. The database included roughly 250 bioreactions and 400 metabolites..

Stoichiometric Constraints. A whole class of specifications in the synthesis of biochemical pathways can be formulated by classifying each building block (each metabolite and each reaction from the database) according to the role it is required or allowed to play in the synthesized pathways.

A given metabolite can participate in a pathway in any of three capacities: (a) as a net *reactant* or substrate of the pathway; (b) as a net *product* of the pathway; and (c) as an *intermediate* in the pathway, i.e., participating without *net* consumption or production. One can impose constraints on pathways by stating which metabolites are *required* and which are *prohibited* to participate in the synthesized pathways in each of the above three capacities. Not all metabolites need be strictly constrained as required or prohibited. Some may simply be *allowed* to participate in the pathways.

For example, metabolites (from a database of biochemical reactions and metabolic intermediates) can be classified according to whether they are allowed to be net reactants in the pathways:

1. *Required reactants* (or desired reactants) *must* be consumed by the pathway;
2. *Allowed reactants* may or may not be consumed by the pathway; and
3. *Excluded reactants* (or prohibited reactants) *must not* be consumed by the pathway.

In a realistic synthesis problem, the default characterization for each metabolite is specification (3): The bulk of the metabolites in the database are *excluded* from being net reactants of the synthesized pathways.

Specification (1) underlies a strict inequality, i.e., stoichiometric coefficient of the metabolite (in the pathway) less than zero, while specification (2) underlies a loose inequality, i.e., stoichiometric coefficient less than or equal to zero. Thus, the first constraint is strict, while the second one is loose. This distinction is relevant in the description of the algorithm, because strict constraints are initially satisfied only in their loose form.

The classification of metabolites as potential products or intermediates of the pathways is quite similar. For intermediates, however, the default characterization differs, as most of the metabolites would normally be classified as *allowed* intermediates. It should be noted that constraints on intermediates are generally not motivated by physiological considerations. They are usually a device for selecting a particular subset of the synthesized pathways.

The constraints on different roles of the same metabolite are not independent. For example, a metabolite that is required as a net product *must* be excluded as a reactant.

A given bioreaction can participate in pathways in either (a) its *forward*, or (b) its *backward* direction. Thus, one can impose constraints by stating which bioreactions are required, which are allowed, and which are prohibited to participate in the synthesized pathways in each of the two directions.

Many constraints designating bioreactions excluded in the backward direction will be present, stemming from knowledge about the (thermodynamic or mechanistic) irreversibility of bioreactions. To this end, one can introduce some kind of constraint on the equilibrium constants (or the Gibbs Energies) of the reactions that can be used; one possible constraint is a simple upper bound on the Gibbs Energy (or a lower bound on the equilibrium constant). Note that the same kind of constraint ought to apply to both the forward direction and reverse (backward) direction of the reaction, because the nominal direction of a reaction in the database is often arbitrary.

The constraints imposed on the two directions of a bioreaction are not completely independent. For example, it is not meaningful to specify a reaction as required in both the forward and reverse direction; a reaction that is required in one direction *must* be excluded in the other direction. Thus, there are in total 5 possible designations (out of a total of 3x3=9 simple-minded combinations) for a reversible reaction:

- Allowed in both directions
- Required in the forward direction and excluded in the backward direction
- Required in the backward direction and excluded in the forward direction
- Allowed in the forward direction and excluded in the backward direction
- Allowed in the backward direction and excluded in the forward direction

3 Description of the Algorithm.

The synthesis algorithm [Mavrovouniotis *et al* 1990a] is devoted to the satisfaction of the above kinds of constraints imposed on the participation of metabolites and bioreactions in biochemical pathways.

Given a set of stoichiometric constraints and a database of biochemical reactions, the developed algorithm synthesizes all biochemical pathways satisfying the stoichiometric constraints. The algorithm is based on the *iterative* satisfaction of constraints, and the stepwise transformation of the initial set of available bioreactions (which can be thought of as one-step pathways that, in general, do not satisfy the constraints), into a final set of pathways, which satisfy all imposed constraints.

To facilitate the description and analysis of the algorithm, some definitions are given here. A *combination* of a set of *constituent pathways* is a pathway whose coefficients are linear combinations of the coefficients of the constituent pathways. In order to retain the original direction of each constituent pathway, the linear combination may involve only *positive combination coefficients*. Let P and Q be pathways derived from the same reaction database. The pathway P is a *subpathway* of Q if and only if every reaction

that has a positive coefficient in P also has a positive coefficient in Q. Equivalently Q is called a *superpathway* of P.

Reaction-Processing Phase. In order to account for the reversibility of reactions, each thermodynamically reversible reaction is decomposed into a forward and a backward reaction. From this point on, we prohibit the participation of both the forward and the reverse reaction in the same pathway, because such a pathway would be redundant.

The constraints placed on the original reactions are then easily transformed into constraints on the new reactions. For a reaction R_k, and its coefficient a_k:

- R_k may occur in the pathway, i.e., $a_k \geq 0$.
- R_k must occur in the pathway, i.e., $a_k > 0$.
- R_k must not occur in the pathway, i.e., $a_k = 0$.

Constraints dictating that certain reactions should not participate in the constructed pathways can be satisfied right from the start. Such reactions are simply eliminated and removed from the active database.

The remaining reactions can be thought of as *one-step pathways* which will be combined in subsequent phases of the algorithm to form longer and longer pathways satisfying more and more constraints.

Metabolite-Processing Phase. The main body of the algorithm tackles one constraint at a time, by transforming the set of particular pathways . Thus, at each iteration stage in the synthesis algorithm, the problem state, often called the *state of the design* [Mostow, 1983, Mostow, 1984] consists of the following elements:

- The set of constraints (on the stoichiometry) or metabolites that still remain to be processed.
- The set of incomplete pathways constructed so far. These are pathways that satisfy the already-processed constraints.
- Back-pointers which show, for each remaining metabolite, the pathways in which it participates. These data-structures must be initially created by passing over each of the initial one-step pathways.

At each pathway-expansion step, the set of active pathways is modified to satisfy a constraint. For example, if the constraint designates a metabolite as an excluded reactant and excluded product, all possible combination-pathways must be constructed by combining one pathway consuming the metabolite and one pathway producing it, such that the metabolite is eliminated from the overall net stoichiometry. Once the combinations are constructed, all pathways consuming or producing the metabolite are deleted.

More generally, the algorithm finds a modification of the set of pathways

such that all surviving pathways satisfy the requirement. This involves the construction of new pathways as combinations of existing ones, as well as deletion of pathways. More specifically, for S the metabolite being processed and using the backward-pointers readily available in each metabolite, two subsets of the current pathway set, L, are assembled:

- The list of pathways that produce the metabolite: $L_p=\{P_i|$ S participates in P_i with a net stoichiometric coefficient $a_i>0\}$.
- The list of pathways that consume the metabolite: $L_c=\{P_i|$ S participates in P_i with a net stoichiometric coefficient $a_i<0\}$.
- The list of pathways in which the metabolite participates as an intermediate: $L_r=\{P_i|$ S participates in some reaction R with coefficient $r_i\neq0$, but S does not participate in the net transformation of P_i, i.e., $a_i=0\}$.
- The list of pathways in which the metabolite does not participate at all $L_n=\{P_i|$ the coefficient of S in each reaction R of P_i is $r_i=0\}$.

The pathways that may, at this step of the algorithm, be deleted from the current set will be pathways from the lists L_p, L_c, and L_r, depending on the nature of the constraint. The pathways that may be constructed are linear combinations using exactly one pathway from L_c and exactly one pathway from L_p:

- Combination pathways: $L_e=\{a_kP_i-a_iP_k \mid P_i\in L_c; P_k\in L_p;$ P_i and P_k do not involve the same reaction in different directions; and a_i and a_k are the net coefficients with which S participates in P_i and $P_k\}$. Since $P_i\in L_c$, $a_i<0$ and the combination $a_kP_i-a_iP_k$ has positive coefficients; thus, it is a legitimate combination of pathways. The net coefficient of S in $a_kP_i-a_iP_k$ is $a_ka_i-a_ia_k=0$. Thus, for all pathways in L_e, S is only an intermediate; it is neither a net reactant nor a net product. As was noted earlier, we exclude combinations of two pathways that involve the same reaction in different directions.

For constraints on reactants and products, the construction of the new set of active pathways is delineated below. The different cases are listed by priority, i.e., in the order in which they should be applied. Once a particular case applies then the remaining cases are automatically excluded[1].

- If S is an excluded product and a required reactant (i.e., $a_k < 0$), all combination pathways are constructed, and the producing pathways are deleted. In effect: $L\leftarrow L\cup L_e-L_p$
- If S is an excluded reactant and a required product (i.e., $a_k > 0$), then: $L\leftarrow L\cup L_e-L_c$.
- If S is an excluded product and an allowed reactant (i.e., $a_k \leq 0$), then: $L\leftarrow L\cup L_e-L_p$

- If S is an excluded reactant and an allowed product (i.e., $a_k \geq 0$), then: $L \leftarrow L \cup L_e - L_c$
- If S is an excluded reactant and an excluded product (i.e., $a_k = 0$), then: $L \leftarrow L \cup L_e - L_c - L_p$
- If S is an excluded intermediate, then: $L \leftarrow L - L_c - L_p - L_r$, or, equivalently, $L \leftarrow L_n$
- If S is a required intermediate, then: $L \leftarrow L$.
- If S is an allowed reactant, an allowed product, and an allowed intermediate, then the set of active pathways is carried intact to the next iteration: $L \leftarrow L$
- If S is a required intermediate , then: $L \leftarrow L$.

After the processing of the constraint, there is a new set of active pathways which satisfy the constraint, with the exception that for strict-inequality constraints, i.e., required products ($a_k > 0$), required reactants ($a_k < 0$), and required intermediates, only the corresponding loose-inequality constraints are guaranteed to be satisfied. The strict-inequality constraints will receive additional consideration in the last phase of the algorithm.

In addition to the set of active pathways, L, the whole *state of the design* that was described earlier must also be properly updated after each constraint is processed. For example, to update the back-pointers that point from each metabolite to the pathways in which it participates, pointers corresponding to deleted pathways must be removed and pointers corresponding to new pathways must be added.

Pathway-Marking Phase. At the end of the metabolite-processing phase, there is a final set of active pathways satisfying all of the requirements, except the strict-inequality constraints for which only the corresponding loose inequalities are satisfied. Because of the linear nature of the requirements, all combinations of pathways also satisfy the constraints.

If each pathway is marked with the strict-inequality constraints it satisfies, the final answer to the synthesis problem can be obtained:

- The pathways satisfying the original stoichiometric constraints of the synthesis problem are all those combinations of pathways which include:
  - ◊ at least one pathway consuming each required reactant,
  - ◊ at least one pathway producing each required product,
  - ◊ at least one pathway containing each required intermediate, and
  - ◊ at least one pathway in which each required reaction participates.

Naturally, a single constituent pathway may possess many of the strict-inequality constraints and can serve to satisfy many of the above requirements.

Thus, if a pathway from the final active set satisfies all of the strict inequalities, then that pathway itself is acceptable as one solution to the overall synthesis problem; *any* combination of that pathway with other pathways from the final set is also acceptable. If, on the other hand, there is a strict-inequality requirement which is not satisfied by any of the pathways in the final set, then there is no solution to the original synthesis problem.

One may wonder whether there are, in the final set, pathways which do not satisfy any of the strict-inequality requirements, and whether there is any reason to construct or keep such pathways. There are indeed such pathways, called *neutral* pathways, generated by the algorithm. Since these pathways do not contribute to the satisfaction of any strict-inequality requirements, it is not *necessary* to use them in combinations constructed from the final set, but they may by freely included in such combinations as they neither prevent any requirements from being satisfied nor introduce additional requirements.

The algorithm that was presented above is correct (it generates *only* feasible pathways) and complete (it generates *all* pathways satisfying the requirements). The performance of the current implementation of the algorithm is quite efficient for well formulated problems. These mathematical and computational properties of the algorithm are discussed in detail below.

Correctness. If a combination of pathways from the final set (produced by the synthesis algorithm) contains at least one constituent pathway satisfying each of the strict inequality requirements (referring to required reactants, products, intermediates, or reactions), then the combination pathway satisfies all of the initial stoichiometric requirements.

The algorithm is correct because each of the original requirements is satisfied in one of the three phases (and after each constraint is satisfied it cannot be subsequently violated):

- Excluded reactions are removed during the reaction-processing phase.
- Excluded intermediates, reactants, or products are eliminated in the metabolite-processing phase. This happens because the pathways that violate the constraints are removed, and any new combination-pathways satisfy the constraints (by their construction).
- Constraints on required reactants or products are satisfied in two phases. In the metabolite-processing phase of the algorithm, after the processing of any particular metabolite, the current set of active pathways satisfies the stoichiometric constraints imposed on that metabolite at least in their loose inequality form. In the pathway-marking phase of the algorithm, a combination of pathways from the final set satisfies the union of the strict-inequality requirements satisfied by its constituent pathways, because the stoichiometries of a combination-pathway are linear combinations of its constituent pathways (with positive coefficients), and those constituent

pathways that do not satisfy the strict inequalities satisfy the corresponding loose inequalities. Thus, acceptable final solutions will contain required reactants and products.

- Constraints on required intermediates and reactions are similarly satisfied in the pathway-marking phase.

Completeness. The synthesis algorithm creates a final set of pathways such that: Any pathway satisfying the original stoichiometry constraints is a combination of pathways from the final set, with one constituent pathway satisfying each strict-inequality constraint.

Incompleteness could only arise in the metabolite-processing phase. At the beginning of the phase, the algorithm has an initial set of (one-step) pathways. Since that set contains all the feasible reactions (unless they have been designated as excluded) any feasible pathway is (by definition) a combination of pathways from that set. The metabolite-processing phase processes each metabolite and its constraints, transforming the set of active pathways. Therefore, it must be shown that:

if *before* processing a particular metabolite there exists a pathway that: (a) Satisfies the constraints on the metabolite; and (b) can be constructed as a combination pathway from the current set of active pathways

then, *after* processing the metabolite, the pathway can still be constructed from the (changed) active set.

This holds because of the way each kind of requirement is handled. Consider, as an example, a metabolite S whose constraint is that it may not occur at all in the stoichiometry of the pathway (excluded reactant and excluded product). As defined in the description of the metabolite-processing phase, let L be the initial active set, L_c the set of the partial pathways (in the initial active set) that consume it, and L_p the set of partial pathways that produce it. Let $L_e=\{a_kP_i–a_iP_k \mid P_i\in L_c, P_k\in L_p$, and a_i and a_k are the net coefficients with which S participates in P_i and $P_k\}$ be the set of new combination pathways created. The net coefficient of S in a pathway $P_e=a_kP_i–a_iP_k$ from L_e is $a_ka_i\text{-}a_ia_k=0$. Processing the metabolite will lead to a new set of active pathways: $L\cup L_e–L_c–L_p$. It will be shown that any pathway Q that can be constructed from L to satisfy the constraints on S can also be constructed from $L\cup L_e–L_c–L_p$.

If the composite pathway Q does not involve any pathways from L_c or L_p, then it can be constructed after the processing exactly the way it was constructed before, since its constituent pathways remain unaffected by the processing of the metabolite.

If Q involves constituent partial pathways from L_c and L_p, then for each of these partial pathways P_i let x_i be the coefficient of S in P_i and y_i the (non-negative) coefficient with which the constituent pathway P_i participates

in Q. If the constraint on S is satisfied, its coefficient in Q must be zero. Thus:

$$\sum_i x_i y_i = 0 \tag{5}$$

Let Y be the total consumption and total production of the metabolite in Q:

$$Y = \sum_{(x_i>0)} x_i y_i = -\sum_{(x_i<0)} x_i y_i \tag{6}$$

By defining

$$f_i = \left| x_i y_i / Y \right| \tag{7}$$

The net stoichiometric coefficient of S in Q can be written as:

$$\sum_i x_i y_i = \sum_{i:(x_i>0)} \sum_{j:(x_j<0)} \left(x_i y_i f_j + x_j y_j f_i \right) \tag{8}$$

Note that, since $x_i>0$ and $x_j<0$ for each of the right-hand summation terms:

$$x_i y_i f_j + x_j y_j f_i = 0 \tag{9}$$

After the parameters f_i are determined, an equation similar to Equation (8) holds for any metabolite. Specifically, if a_i is the coefficient of another metabolite T in P_i, and a_Q is the coefficient of T in Q, then:

$$a_Q = \sum_i a_i y_i = \sum_{i:(x_i>0)} \sum_{j:(x_j<0)} \left(a_i y_i f_j + a_j y_j f_i \right) \tag{10}$$

where f_i and f_j are still derived from Equation (7), i.e., from the coefficients of S—the metabolite being processed. An identical equation holds for the coefficients of reactions in the pathways: If a_i is the coefficient of a reaction in pathway P_i and a_Q is the coefficient of the same reaction in Q, then Equation (10) holds.

Thus, the transformation in Equation (10) denotes that the composite pathway Q can be written as a sum of pairs of constituent partial pathways (with f_i and f_j the coefficients used in combining P_j and P_i), such that for each pair the metabolite has zero total coefficient, as Equation (9) states. To demonstrate that these pairs are exactly the combinations (i.e., the pathways of L_e) created by the algorithm, Equation (11) can be used to eliminate f_i and f_j from Equation (10):

$$a_Q = \sum_{i:(x_i>0)} \sum_{j:(x_j<0)} \left[y_i y_j \left(a_j x_i - a_i x_j \right) \right] Y^{-1} \tag{11}$$

The term $a_j x_i - a_i x_j$ refers precisely to a combination pathway from L_e, while the factor $y_i y_j Y^{-1}$ provides the coefficients of combination that construct Q from pathways in L_e. Hence, a composite pathway that satisfied the constraint *before* the metabolite was processed can still be constructed *after* the

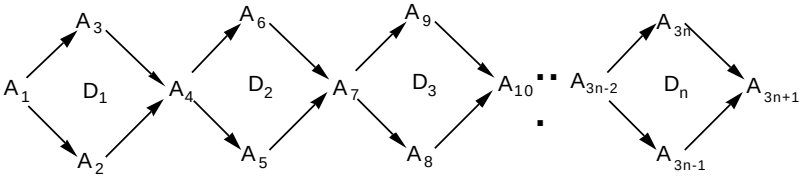


Figure 4: A set of reactions giving rise to an exponential number of pathways

metabolite is processed, using the combination-pair partial pathways created in the processing.

Computational Complexity Issues. The number of pathways that satisfy a set of stoichiometric constraints is, in the worst case, exponential in the number of reactions. Consider the reactions depicted in Figure 4. For each diamond (numbered as D_1, D_2, etc.) consisting of two parallel branches, a pathway can follow either the upper branch or the lower branch. If there are n diamonds (and $4n=m$ reactions), there are n junctions where these choices occur. Thus, there are $2^n=2^{m/4}$ distinct pathways. These are all genotypically independent: Since no two of them involve the same set of choices (at the junctions), it follows that no two of them involve the same set of enzymes.

Since the algorithm described here constructs all genotypically independent pathways, the algorithm would require time (and storage space) exponential in the number of reactions. Thus, the algorithm's worst-case complexity is at least exponential. In practice, however, the metabolism contains long sequences of reactions but few parallel branches of the type of Figure 4. Thus, with careful design of the computer programs it is possible to obtain results more efficiently than the worst-case complexity suggests.

It is useful to discuss, in the context of computational complexity, why the metabolite-processing phase of the pathway does not necessarily start from metabolites that are required reactants and may instead start from other intermediates. In the formulation of the problem, constraints are imposed on all metabolites. When the algorithm selects the next constraint to satisfy, it picks the one that appears easiest to process (an approach reminiscent of *greedy algorithms*); this would be the metabolite that participates in the smallest number of reactions, regardless of whether the metabolite is a required reactant or an excluded reactant.

The fact that the algorithm processes not only designated required reactants (or products) is an important factor in guaranteeing the completeness of the algorithm and guarding it against computational complexity. Consider the simple pathway of Figure 5, and suppose that the whole reaction database consists of the two reactions in the figure, and the objective is to convert pyruvate to oxaloacetate. An algorithm that searches from substrates towards

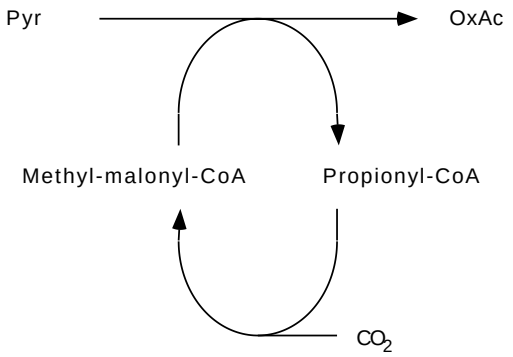


Figure 5. Carboxylation of pyruvate through an alternative pathway, involving Methyl-malonyl-CoA and Propionyl-CoA

products cannot start the construction of a pathway from the reaction:

PYRUVATE + METHYL-MALONYL-COA → OXALOACETATE + PROPIONYL-COA

because this reaction uses methyl-malonyl-CoA, which is not available as a reactant. Likewise, that algorithm can not start from:

PROPIONYL-COA → METHYL-MALONYL-COA + CO_2

because the reaction requires propionyl-CoA which is also not available. Thus, this type of algorithm fails to see that, taken as a cluster, these two reactions achieve the desired transformation. The algorithm presented here, on the other hand, considers the constraint that designates propionyl-CoA as an excluded reactant and excluded product, and immediately constructs the pathway of Figure 5 to satisfy the constraint.

Implementation. The algorithm was implemented in LISP, on Symbolics 3640 and 3650 computers. The performance of the implementation of the algorithm greatly varies with the exact formulation of the problem, but it is generally proportional to the cardinality of the final set of pathways.

- The requirements for setting up the initial data structures are proportional to the size of the database. Rough requirements per database object are 0.05 s of elapsed time (with garbage-collection suppressed), 70 list-words, and 70 structure-words[2].
- The requirements for the main body of the algorithm appear proportional to the number of solutions for those cases in which results were obtained. The program needs 0.15 seconds (elapsed time), 200 list-words, and 100 structure-words per synthesized pathway.

A typical problem requires 35s, 40k list words, and 40k structure words for the initial set-up, and 8 minutes, 1M list words, and 500k structure words

for the construction of pathways (based on 5000 final pathways).

4. An Example of the Operation of the Algorithm.

A step-by-step application of the algorithm for a synthesis problem is presented here. The set of reactions under consideration is:

a. $A \rightarrow B$
b. $B \leftrightarrow C$
c. $C \leftrightarrow D$
d. $C + D \leftrightarrow F + K$
e. $F + K \leftrightarrow H + E$
f. $H + D \leftrightarrow E + F$
g. $A \leftrightarrow E$
h. $E \rightarrow F + G$
k. $F \leftrightarrow G$
m. $G \rightarrow L$

All metabolites are designated as excluded reactants and excluded products, with the exception of A, which is a required reactant, and L, which is a required product.

We first construct the reverse reactions, for reactions b, c, d, e, f, g, and k. We designate the reverse reactions as -b, -c, -d, -e, -f, -g, and -k respectively. We also list separately each metabolite and the pathways in which it participates. Representing only the reactions from which a pathway is constructed, an expression like [2a, 2-g, b] denotes a pathway that is constructed as a linear combination of the reactions a, -g, and b, with coefficients 2, 2, and 1, respectively. To represent instead the overall transformation accomplished by this pathway, the expression $2E \rightarrow \rightarrow B + C$ is used. The two expressions can be combined into $2E \rightarrow[2a, 2\text{-}g, b]\rightarrow B + C$. Using this notation the initial pathways are:

$A \rightarrow[a]\rightarrow B$
$B \rightarrow[b]\rightarrow C$
$C \rightarrow[-b]\rightarrow B$
$C \rightarrow[c]\rightarrow D$
$D \rightarrow[-c]\rightarrow C$
$C + D \rightarrow[d]\rightarrow F + K$
$F + K \rightarrow[-d]\rightarrow C + D$
$F + K \rightarrow[e]\rightarrow H + E$
$H + E \rightarrow[-e]\rightarrow F + K$
$H + D \rightarrow[f]\rightarrow E + F$
$E + F \rightarrow[-f]\rightarrow H + D$
$A \rightarrow[g]\rightarrow E$
$E \rightarrow[-g]\rightarrow A$

E →[h]→ F + G
F →[k]→ G
G →[-k]→ F
G →[m]→ L

The set of metabolites with the pathways in which they participate is:

A: [a], [g], [-g]
B: [b], [-b], [a]
C: [b], [-b], [c], [-c], [d], [-d]
D: [c], [-c], [d], [-d], [f], [-f]
E: [e], [-e], [f], [-f], [g], [-g], [h]
F: [d], [-d], [e], [-e], [f], [-f], [h], [k], [-k]
G: [h], [k], [-k], [m]
H: [e], [-e], [f], [-f]
K: [d], [-d], [e], [-e]
L: [m]

Following the algorithm, the metabolites that participate in fewer pathways must be processed first. L, which participates in only one pathway, is a required product (and an excluded reactant). Since L is produced by one partial pathway and is not consumed by any partial pathway, processing the constraints on this metabolite does not change any pathway.

The next metabolite that is processed must be either A or B; the order in which these two metabolites are processed does not affect the results and we arbitrarily choose B. One new pathway are constructed: [a,b] as a combination of [a] and [b]; this operation is denoted as [a]+[b]=[a,b]. Note that it is not permissible to construct the pathway [b]+[-b], because it would involve the same reaction in both the forward and reverse directions. The pathways [a], [b], and [-b] are then deleted. The set of active pathways is now:

A →[a, b]→ C
C →[c]→ D
D →[-c]→ C
C + D →[d]→ F + K
F + K →[-d]→ C + D
F + K →[e]→ H + E
H + E →[-e]→ F + K
H + D →[f]→ E + F
E + F →[-f]→ H + D
A →[g]→ E
E →[-g]→ A
E →[h]→ F + G
F →[k]→ G
G →[-k]→ F
G →[m]→ L

The updated set of metabolites becomes:

A: [a, b], [g], [-g]
C: [a, b], [c], [-c], [d], [-d]
D: [c], [-c], [d], [-d], [f], [-f]
E: [e], [-e], [f], [-f], [g], [-g], [h]
F: [d], [-d], [e], [-e], [f], [-f], [h], [k], [-k]
G: [h], [k], [-k], [m]
H: [e], [-e], [f], [-f]
K: [d], [-d], [e], [-e]

The metabolite A is processed next. Since A is a required reactant and excluded product, a new combination pathway are constructed as [-g]+[a,b]=[-g,a,b], and only pathway [-g] is deleted. For the next step G is selected arbitrarily among the metabolites G, H, and K (which participate in the same number of reactions). In processing G, there are two pathways consuming it ([-k] and [m]) and two pathways producing ([h] and [k]). Hence, four combinations would be constructed, except that [k] cannot be combined with [-k]. Three legitimate combinations remain, namely: [h]+[-k]=[h, -k]; [h]+[m]=[h, m]; [k]+[m]=[k, m]. The original four pathways in which G participated are deleted.

After the processing of A and G, the active pathways are:

A $\rightarrow$[a, b]$\rightarrow$ C
C $\rightarrow$[c]$\rightarrow$ D
D $\rightarrow$[-c]$\rightarrow$ C
C + D $\rightarrow$[d]$\rightarrow$ F + K
F + K $\rightarrow$[-d]$\rightarrow$ C + D
F + K $\rightarrow$[e]$\rightarrow$ H + E
H + E $\rightarrow$[-e]$\rightarrow$ F + K
H + D $\rightarrow$[f]$\rightarrow$ E + F
E + F $\rightarrow$[-f]$\rightarrow$ H + D
A $\rightarrow$[g]$\rightarrow$ E
E $\rightarrow$[-g, a, b]$\rightarrow$ C
E $\rightarrow$[h, -k]$\rightarrow$ 2 F
E $\rightarrow$[h, m]$\rightarrow$ F + L
F $\rightarrow$[k, m]$\rightarrow$ L

The set of metabolites becomes:

C: [a, b], [c], [-c], [d], [-d], [-g, a, b]
D: [c], [-c], [d], [-d], [f], [-f]
E: [e], [-e], [f], [-f], [g], [-g, a, b], [h, m], [h, -k]
F: [d], [-d], [e], [-e], [f], [-f], [h, m], [k, m], [h, -k]
H: [e], [-e], [f], [-f]
K: [d], [-d], [e], [-e]

The metabolite K, participating in four pathways, is processed next. The combinations [d]+[e]=[d, e], and [-e]+[-d]=[-e, -d] are created, and the pathways [d], [-d], [e], and [-e] are deleted. The set of active pathways becomes:

A →[a, b]→ C
C →[c]→ D
D →[-c]→ C
C + D →[d, e]→ H + E
H + E →[-e, -d]→ C + D
H + D →[f]→ E + F
E + F →[-f]→ H + D
A →[g]→ E
E →[-g, a, b]→ C
E →[h, -k]→ 2 F
E →[h, m]→ F + L
F →[k, m]→ L

The set of metabolites becomes:

C: [a, b], [c], [-c], [d, e], [-e, -d], [-g, a, b]
D: [c], [-c], [d, e], [-e, -d], [f], [-f]
E: [d, e], [-e, -d], [f], [-f], [g], [-g, a, b], [h, m], [h, -k]
F: [f], [-f], [h, m], [k, m], [h, -k]
H: [d, e], [-e, -d], [f], [-f]

Processing H in a very similar fashion, two combination pathways are constructed, namely [-f]+[-e,-d]=[-f,-e,-d] and [d,e]+[f]=[d,e,f]. The pathways now become:

A →[a, b]→ C
C →[c]→ D
D →[-c]→ C
C + 2 D →[d, e, f]→ 2 E + F
2 E + F →[-f, -e, -d]→ C + 2 D
A →[g]→ E
E →[-g, a, b]→ C
E →[h, -k]→ 2 F
E →[h, m]→ F + L
F →[k, m]→ L

The set of metabolites becomes:

C: [a, b], [c], [-c], [d, e, f], [-f, -e, -d], [-g, a, b]
D: [c], [-c], [d, e, f], [-f, -e, -d]
E: [d, e, f], [-f, -e, -d], [g], [-g, a, b], [h, m], [h, -k]
F: [d, e, f], [-f, -e, -d], [h, m], [k, m], [h, -k]

Since D involves now only 4 pathways, it is processed next. The fact that the coefficient of D in [d, e, f] and [-f, -e, -d] is 2 must be reflected in the construction of the combinations. The new pathways are constructed as 2[c]+[d,e,f]=[2c, d, e, f] and [-f,-e,-d]+2[-c]=[-f, -e, -d, 2 -c], and all four pathways that involved D are deleted. The set of active pathways is now significantly smaller:

A →[a, b]→ C

3 C →[2c, d, e, f]→ 2 E + F
2 E + F →[-f, -e, -d, 2 -c]→ 3 C
A →[g]→ E
E →[-g, a, b]→ C
E →[h, -k]→ 2 F
E →[h, m]→ F + L
F →[k, m]→ L

Only three metabolites remain:

C: [a, b], [2c, d, e, f], [-f, -e, -d, 2 -c], [-g, a, b]
E: [2c, d, e, f], [-f, -e, -d, 2 -c], [g], [-g, a, b], [h, m]
F: [2c, d, e, f], [-f, -e, -d, 2 -c], [h, m], [k, m], [h, -k]

C is processed next and leads to two combinations, 3[a,b]+[2c,d,e,f]=[3a, 3b, 2c, d, e, f] and 3[-g, a,b]+[2c,d,e,f]=[3 -g, 3a, 3b, 2c, d, e, f]. Then the active pathways are:

3 A →[3a, 3b, 2c, d, e, f]→ 2 E + F
A →[g]→ E
E →[3 -g, 3a, 3b, 2c, d, e, f]→ F
E →[h, -k]→ 2 F
E →[h, m]→ F + L
F →[k, m]→ L

The two metabolites remaining are:

E: [3a, 3b, 2c, d, e, f], [g], [3 -g, 3a, 3b, 2c, d, e, f], [h, m], [h, -k]
F: [3a, 3b, 2c, d, e, f], [3 -g, 3a, 3b, 2c, d, e, f], [h, m], [k, m], [h, -k]

The two metabolites can be processed in either order to yield the final results. Processing F leads to three new combinations of pathways: [3a, 3b, 2c, d, e, f]+[k, m]=[3a, 3b, 2c, d, e, f, k, m]; [h, m] + [k, m] = [h, k, 2 m]; and finally [3 -g, 3a, 3b, 2c, d, e, f]+[k, m]=[3 -g, 3a, 3b, 2c, d, e, f, k, m]. After the original 5 pathways in which F participated are deleted, the remaining pathways are:

3 A →[3a, 3b, 2c, d, e, f, k, m]→ 2 E + L
A →[g]→ E
E →[3 -g, 3a, 3b, 2c, d, e, f, k, m]→ L
E →[h, k, 2m]→ 2 L

Processing E (and omitting pathways that include the same reaction in opposing directions) leads to the combinations: 1/3[3a, 3b, 2c, d, e, f, k, m] + 2/3[3 -g, 3a, 3b, 2c, d, e, f, k, m] = [2 -g, 3a, 3b, 2c, d, e, f, k, m][3]; [3a, 3b, 2c, d, e, f, k, m] + 2[h, k, 2m] = [3a, 3b, 2c, d, e, f, 3k, 5m, 2h]; and the much simpler [g]+[h, k, 2m]=[g, h, k, 2m]. Thus, the final pathways are:

A →[2 -g, 3a, 3b, 2c, d, e, f, k, m]→ L
3 A →[3a, 3b, 2c, d, e, f, 3k, 5m, 2h]→ 5 L
A →[g, h, k, 2m]→ 2 L

These three pathways are feasible solutions to the original synthesis problem. All other feasible pathways are linear combinations of pathways from

this set, with positive coefficients.

When the algorithm is not permitted to run to completion (because of limited computational resources), it can provide useful *partial* results. Specifically, it will return a list of pathways that satisfy only *some* of the constraints involved; it will also return the list of *unprocessed constraints*. In the detailed example discussed in this section, if the algorithm must stop before the last step, it returns a list of four pathways that satisfy all constraints *except for the constraint designating E as an excluded reactant and excluded product;* the algorithm also indicates that the constraint on E has not been satisfied.

5 A Case Study: Lysine Pathways

We are going to perform, in this chapter, a case study on the synthesis and evaluation of biochemical pathways for the production of lysine from glucose and ammonia [Mavrovouniotis, *et al* 1990a].

It should be emphasized right from the start that the analysis we perform here is not exhaustive; our aim is merely to demonstrated the concerted application and utility of our methods in a real system and not to arrive to definitive answers on the synthesis of lysine.

The basic procedure we will follow in this case study is as follows:

- We synthesize a pathway as close as possible to the pathway believed to prevail
- We identify bottlenecks in the pathway by performing a maximum-rate analysis for each reaction
- We synthesize pathways that bypass bottlenecks
- We synthesize other pathways to explore alternatives that omit key enzymes
- We try to identify fundamental constraints on the structure and yield of the pathways

Note that this is a procedure suggested from the point of view of the goals of the analysis. The exact application of the methods may take place following a number of different structures. For example, one can generate *all* pathways producing the desired product from the substrates *a priori*, carry out all the maximum rate calculations for all pathways, and then perform all the tasks by appropriate search through this (potentially very big) set of pathways.

Table 5 shows the abbreviations that we will use for metabolic intermediates throughout this chapter. The core of the bioreaction network with which we will work is shown in Figure 6. It includes:

ABBREVIATION	METABOLITE
2PG	2-phosphoglycerate
3P-OH-Pyr	3-Phosphohydroxypyruvate
3P-Ser or P-Ser	3-Phospho-serine
3PG	3-phosphoglycerate
AcCoA (or Acetyl-CoA)	Acetyl-Coenzyme-A
αkG	α-ketoglutarate
Ala	Alanine
ASA	Aspartate-semialdehyde
Asp	Aspartate
Cit	Citrate
DHAP	Dihydroxyacetone-phosphate
Fru6P	Fructose-6-phosphate
FruDP	Fructose-1,6-diphosphate
Fum	Fumarate
GAP	Glyceraldehyde-3-phosphate
Glc	Glucose
Glc6P	Glucose-6-phosphate
Gln	Glutamine
Glt or Glu	Glutamate
Gly	Glycine
Glyox	Glyoxylate
i-Cit	Isocitrate
Lys	Lysine
Mal	Malate
OxAc	Oxaloacetate
PEP	Phosphoenolpyruvate
Pyr	Pyruvate
Suc	Succinate
SucCoA	Succinyl-Coenzyme-A

Table 14: Abbreviations of the names of metabolites

- Glycolysis
- *Lactate dehydrogenase*, converting pyruvate to lactate (a common anaerobic fate for pyruvate)
- The usual citric acid cycle (or tricarboxylic acid cycle, which will be referred to as TCA), with the exception of the bioreaction *α-ketoglutarate dehydrogenase* which will be assumed to be absent or non-functional
- The glyoxylate shunt to complement TCA and make up for the absence of

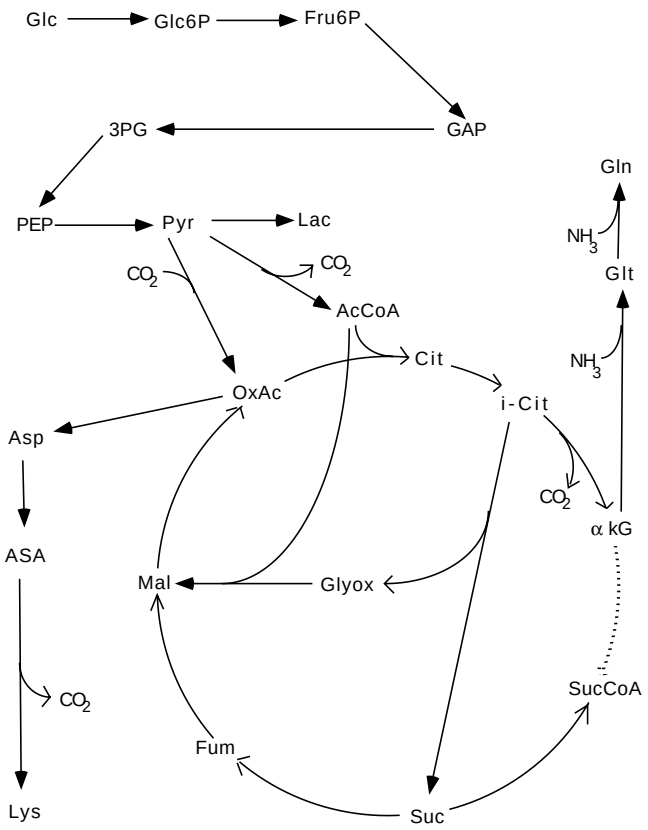


Figure 6: The basic bioreaction network for the synthesis of lysine

α-ketoglutarate dehydrogenase

- The bacterial pathway that leads from oxaloacetate to aspartate and on to lysine.
- *Glutamate dehydrogenase* and *glutamine synthetase* for the synthesis of glutamate and glutamine

Figure 6 was constructed to conform to the bioreaction network used in the analysis of experiments of lysine production. Note that the figure is substantially simplified, as:

- Many side-reactants and side-products are not shown.
- Many reactions are lumped together. In particular, the arrow drawn from aspartate-semialdehyde (ASA) to lysine represents 6 individual bioreac-

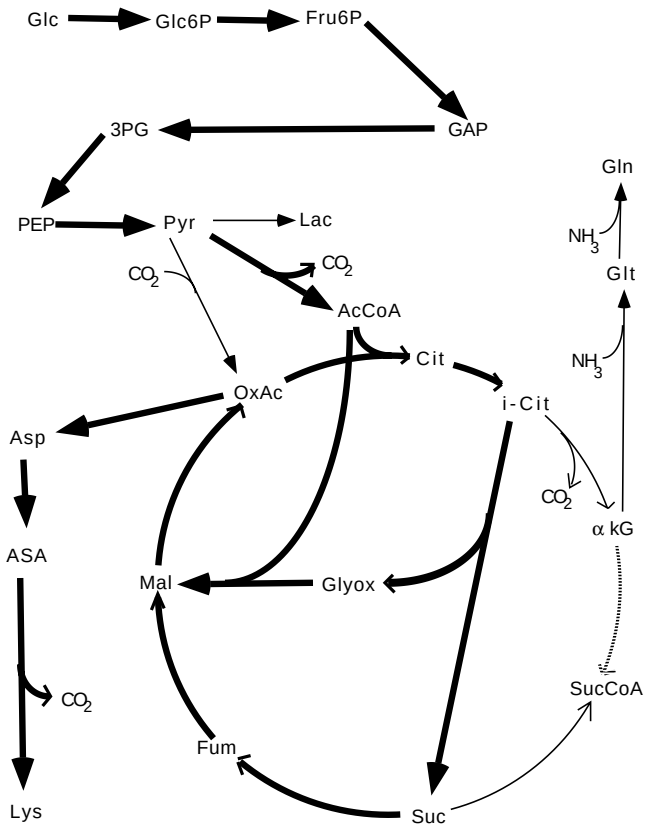


Figure 7: Partial pathway of the normal route of lysine production. Bold arrows indicate the reactions used in the pathway. Dotted arrow indicates a prohibited reactions.

tions.

One of the basic pathways believed to function in bacteria (such as *Brevibacterium Flavum*) for the conversion of glucose to lysine is shown in Figure 7. The pathway uses the glyoxylate shunt to bypass *α-ketoglutarate dehydrogenase*, which has not been included in the network.

The pathway of Figure 7 is actually only a partial one, because the pathway leading from aspartate to lysine requires succinyl-CoA and glutamate (at the same time producing succinate and α-ketoglutarate).

We can complete the pathway by looking for the smallest possible pathway which contains all the reactions already marked and additional reactions to balance the stoichiometries for succinyl-CoA, glutamate, succinate, and α-ketoglutarate. The stoichiometries of other metabolites (such as ATP or

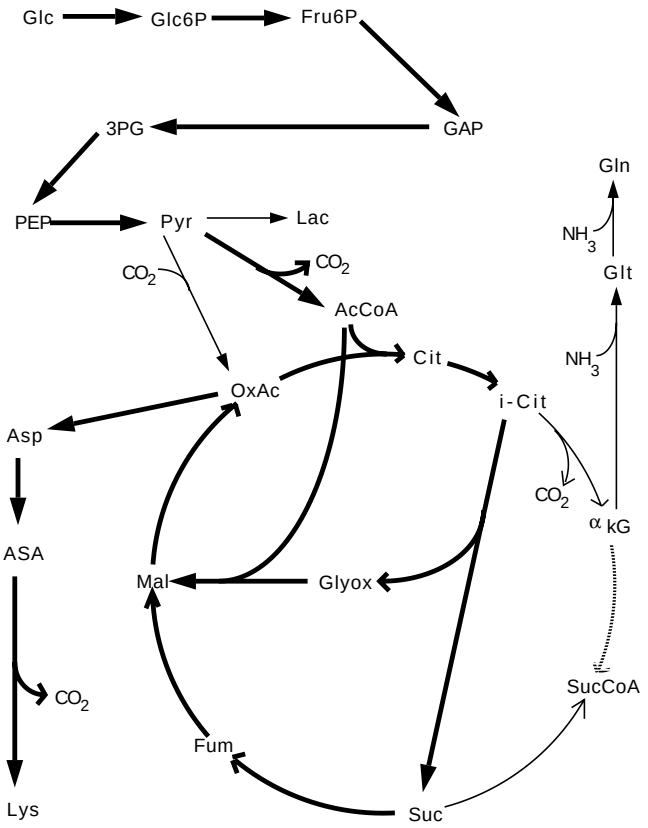


Figure 8: Completion of the basic pathway for the synthesis of lysine.

NAD) are also unbalanced, but we will consider them to be allowed reactants and allowed products.

The completed pathway is shown in Figure 8. The reactions added are *succinate kinase* and *glutamate dehydrogenase*, which, as might be expected, comprise the simplest alternative.

Maximum-rate analysis. We will calculate here the minimum enzyme requirement, which is a parameter related to the maximum rate of each bioreaction [Mavrovouniotis *et al.*, 1990b]. By comparing the rates of the bioreactions in the basic pathway of Figure 8 we will uncover kinetic bottlenecks. For the estimation of the maximum rate of a bioreaction, we need the equilibrium constant of the reaction, the concentrations of the reactants and products of the reaction, and the concentrations of enzymes:

- We would like to be as conservative as possible in choosing a value for the

equilibrium constant, because it has a very strong effect on the maximum rate. We take the equilibrium constant of a bioreaction to be equal to the maximum value among:

◊ Any available data from the literature (residing in the database)

◊ The value estimated by the group-contribution method

- The concentrations of all metabolites are assumed to be in the default range we normally use for physiologically acceptable conditions. Thus, the concentrations of the products of each bioreaction are set to $5x10^{-6}$ and the concentrations of reactants are set to $5x10^{-3}$.
- The concentration of the enzyme is not assumed to have any particular value. Since the maximum rate is proportional to the concentration of the enzyme, we can estimate [Mavrovouniotis, *et al* 1990b] the quantity r/E, i.e., maximum rate divided by the enzyme concentration, leaving the enzyme concentration unspecified.

Instead of using the ratio r/E, where the quantity E (in mol/l) refers to intracellular concentration and r (in mol/s l) refers to rate per unit cell volume, we can equivalently estimate the inverse of that ratio, i.e., E/r, which denotes the *minimum enzyme requirement* (per unit rate) for the bioreaction. The actual (i.e., experimental) E/r of a reaction must be higher than our estimate; since actual enzymes are less efficient it takes a higher (than ideally estimated) enzyme concentration to achieve a given rate. The minimum enzyme requirement, E/r, is a particularly convenient quantity because the minimum enzyme requirement of the whole pathway can be obtained simply by adding together the requirements of all the reactions.

In this context, it is convenient to take r not as the rate of the bioreaction examined, but rather as the rate of production of the final product. To achieve this transformation of reference-rate, we only need to multiply the initial enzyme requirement of each reaction by the corresponding coefficient of the reaction-stoichiometry of the pathway.

Note that, since a pathway involves many enzymes, the enzyme requirement of the pathway denotes the sum of the concentrations of different enzymes. This is not unreasonable considering that the different enzymes have to coexist and function in the same cell, and compete, in their synthesis, for same limited resources of the cell. Similarly, the pathway as a whole competes for resources for all of its enzymes, because it is functional only when sufficient quantities of all enzymes are present. Thus, in evaluating a pathway as a whole and comparing it to other pathways, it is useful to lump the concentrations of all the enzymes in the pathway and estimate the minimum enzyme requirement of the pathway.

The minimum enzyme requirement for each bioreaction in the basic pathway of Figure 8 is shown in Figure 9. For each reaction in Figure 9, the num-

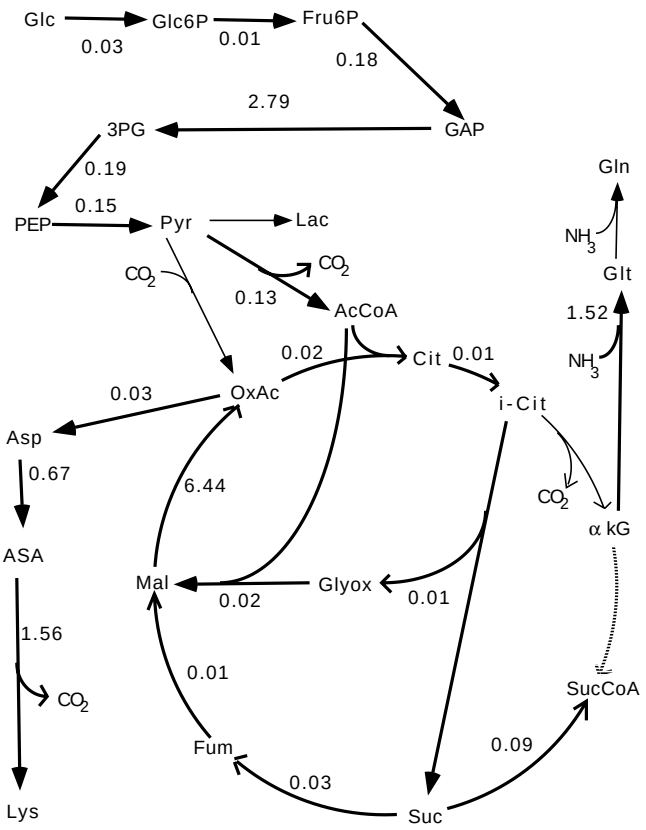


Figure 9 Calculation of minimum enzyme requirements for the basic pathway for lysine production

ber shown is the enzyme requirement of that reaction, in milliseconds. The total enzyme requirement for the whole pathway of Figure 9 is approximately 14 ms.

About half of the enzyme requirements of the pathway come from the bioreaction *malate dehydrogenase*, which has an enzyme requirement of 6.44 ms. The next larger contribution, equal to 2.7 ms, comes from *glyceraldehyde-phosphate dehydrogenase*. However, we have very little control over that enzyme since it belongs to glycolysis. Thus, *malate dehydrogenase* remains the main kinetic bottleneck of the pathway.

Bypassing the Potential Kinetic Bottleneck. We seek now new pathways that eliminate the kinetic bottleneck of malate dehydrogenase. In Figure 10 we show a first possibility, which has been already determined (experimentally) to function under certain conditions. This pathway involves the

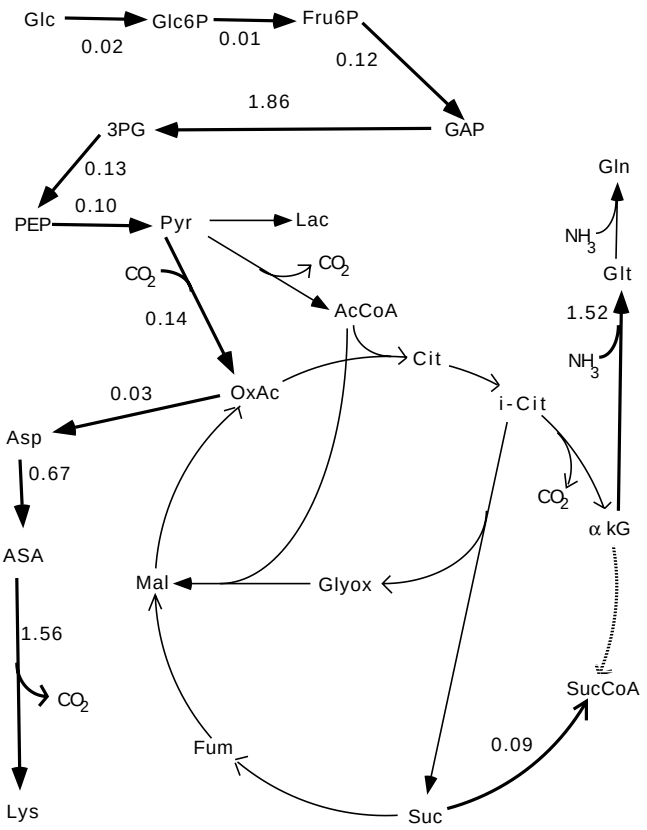


Figure 10: Minimum enzyme requirements for a lysine pathway involving carboxylation of pyruvate

carboxylation of pyruvate, bypassing the whole TCA cycle. This direct conversion of pyruvate to oxaloacetate can be achieved by two distinct bioreactions:

- Pyruvate carboxylase
- Oxaloacetate decarboxylase

The pathway of Figure 10 successfully bypasses the kinetic bottlenecks because its minimum enzyme requirement is only 6.4 ms, roughly equal to one half the requirement of the initial pathway. This pathway also has a higher maximum molar yield. Its yield is 100%, i.e., the pathway yields one mole of lysine per mole of glucose, as compared to a molar yield of 67% for the initial pathway of Figure 9.

If the original pathway has some good traits, we might prefer to bypass only the immediate vicinity of the bottleneck and retain much of the structure

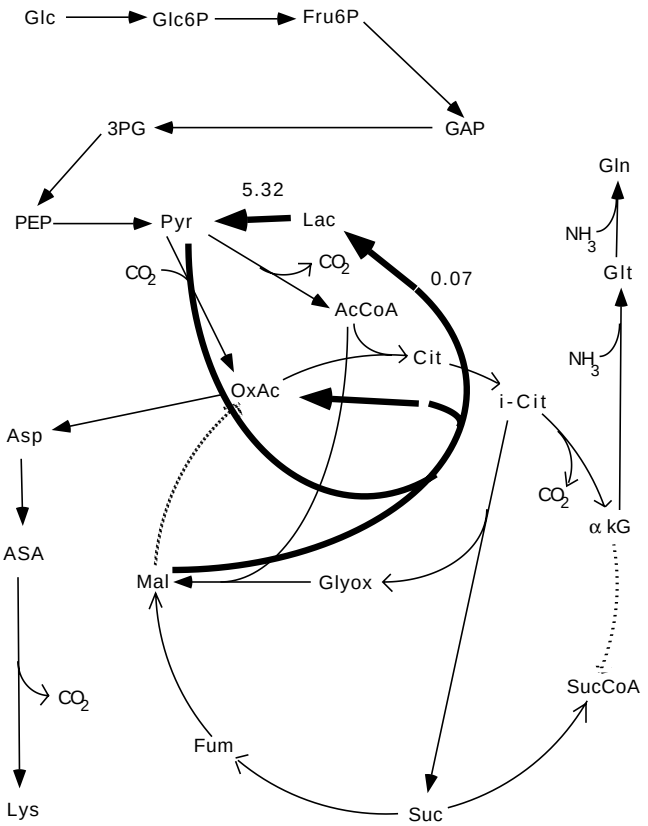


Figure 11 Pathway converting malate to oxaloacetate, with lactate and pyruvate as intermediates

of the original pathway intact, including the TCA cycle. A first alternative, shown in Figure 11, involves bypassing *malate dehydrogenase* with a set of just two reactions:

• *Lactate-Malate transhydrogenase* achieves the conversion:

MALATE + PYRUVATE → OXALOACETATE + LACTATE

• *Lactate dehydrogenase* achieves the conversion:

LACTATE → PYRUVATE

The combination of the two reactions converts malate to oxaloacetate. Unfortunately, the enzyme requirement of this bypass is approximately the same as that of *malate dehydrogenase*. Specifically, *lactate dehydrogenase* has a requirement of 5.32 ms (compared to 6.44 for *malate dehydrogenase*).

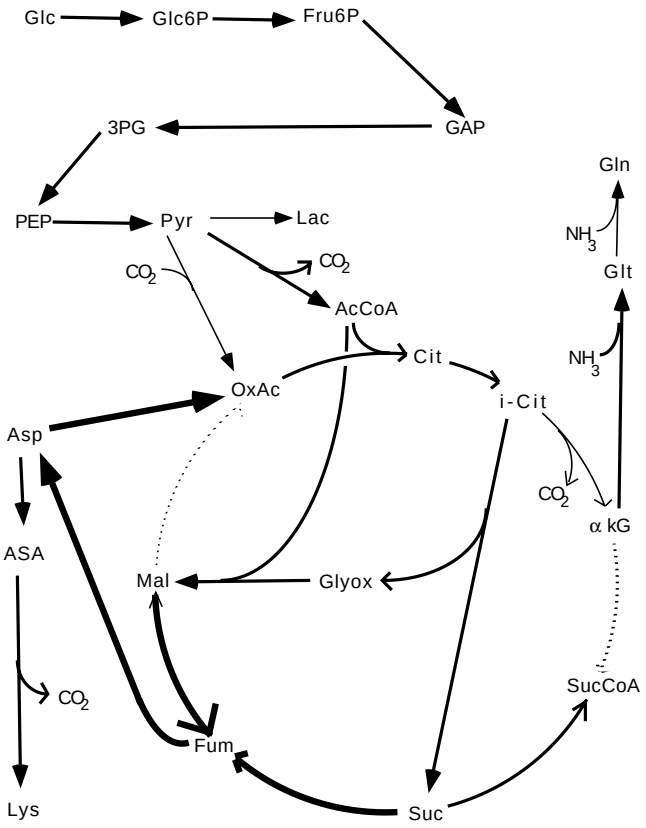


Figure 12.. The simplest of the pathways bypassing malate dehydrogenase by converting fumarate to aspartate

Thus, this particular pathway offers little improvement over the original one. It is interesting to note that this pathway uses *lactate dehydrogenase* in the direction opposite to that originally drawn in Figure 6

Two more interesting alternatives are shown in Figures 12 and 13. They both involve:

- Conversion of malate to fumarate by using *Fumarase* in the direction opposite to that initially assumed in Figure 6
- Conversion of succinate to fumarate by *Succinate dehydrogenase* as in the original pathway
- Conversion of fumarate into aspartate through *Aspartate aminolyase*

Since oxaloacetate is used in order to form citrate, half of the aspartate

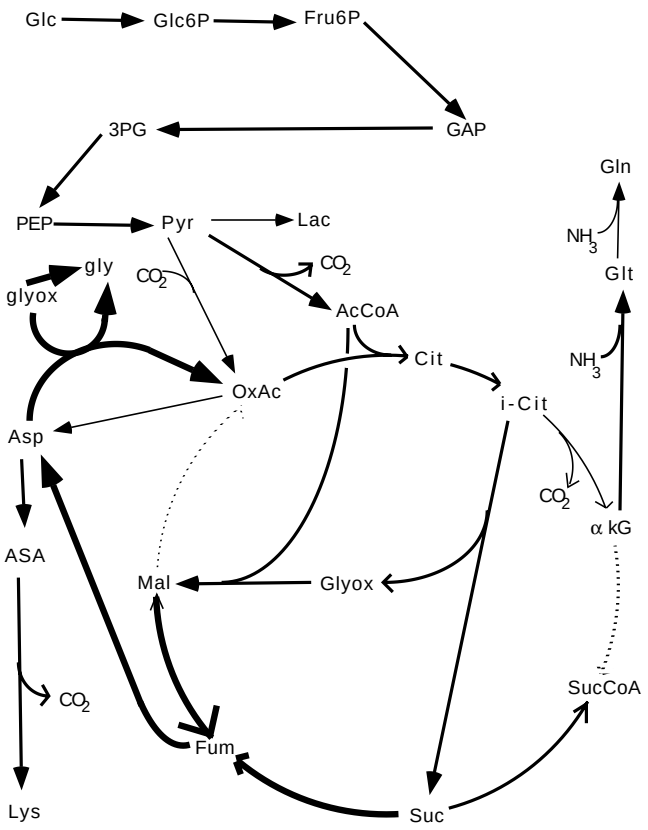


Figure 13. The kinetically most efficient of the pathways bypassing malate dehydrogenase by converting fumarate to aspartate

must be recycled back into oxaloacetate to close the TCA loop. The two pathways use different ways to achieve this:

- In the pathway of Figure 12 the reaction *aspartate glutamate transaminase* converts oxaloacetate to aspartate, by operating in the direction reverse to that assumed in the original bioreaction network (Figures 8 to 11).
- The pathway of Figure 13 uses a set of two reactions, *Glycine dehydrogenase* and *Glycine-oxaloacetate aminotransferase*, involving interconversion of glycine and glyoxylate.

The pathway of Figure 13 is longer, but it is actually the most efficient (kinetically) of all the pathways sharing the TCA structure of the original pathway. Its minimum enzyme requirement is 8 ms, i.e., almost half of the requirement of the original pathway.

Persistent Intermediates. In the two pathways discussed above, oxaloacetate is partly bypassed, in that it is needed only for the synthesis of citrate, and not directly for the synthesis of aspartate and lysine. An interesting question is whether we can bypass oxaloacetate *altogether* and produce aspartate directly from pyruvate or glucose.

With the reactions in our database, this turns out to be impossible. Thus, it appears that oxaloacetate is a key intermediate in the production of aspartate and lysine. The pathways we discuss in this chapter (and other pathways which were constructed but will not be discussed) indicate, in fact, that, the only *persistent* intermediates, i.e., intermediates that occur in all pathways are:

- The intermediates of glycolysis from glucose to phosphoenolpyruvate, with that section of the pathway fixed
- The intermediates of the pathway from aspartate to lysine, a pathway that is also fixed
- Oxaloacetate, for which no surrounding reaction is fixed, but the intermediate itself is always present (participating in different reactions)

One might argue that the conclusion that oxaloacetate is a necessary intermediate is obvious, because standard biochemistry textbooks classify lysine in the aspartate family [Mandelstam *et al.*, 1982, Rawn, 1983, Snyder *et al.*, 1985], and aspartate is commonly synthesized from oxaloacetate. However, the pathways discussed here involve several different bioreactions consuming or producing oxaloacetate; thus, the metabolism in the region of this intermediate can hardly be characterized as fixed. Our conclusion states that any lysine-producing pathway involves at least 2 of these reactions (hence there is no pathway that can avoid the intermediate altogether).

In the pathway of Figure 12 (and its variation suggested by Figure 13) aspartate and lysine are *not* directly derived from oxaloacetate, because fumarate is converted to aspartate by a single enzyme. In fact, aspartate is converted *into* oxaloacetate (rather than the reverse). Thus, the metabolism in the neighborhood of aspartate, fumarate, malate and oxaloacetate is quite different from what one would find in a standard biochemistry textbook. This portion of the metabolism suggests that it *is* possible to derive aspartate without the intervention of oxaloacetate. It turns out however that, within the enzyme database used here, the necessary TCA intermediates (malate or succinate) cannot be produced from glucose without the intervention of oxaloacetate; this constraint necessitates the presence of oxaloacetate in any pathway leading from glucose to lysine. In effect, the real obstacle is that production of fumarate from glucose requires the TCA cycle and hence oxaloacetate.

To illustrate this point better, assume that (in addition to glucose) we

could use succinate as an allowed reactant. *A priori* biosynthetic classifications would still entail oxaloacetate as a required intermediate. Inspection of Figure 13 reveals, however, that succinate can be converted to fumarate and on to aspartate (by *aspartate aminolyase*), without the intervention of malate or oxaloacetate. Thus, with succinate as an additional substrate, it is entirely possible to synthesize lysine with a pathway that does not entail oxaloacetate.

If one rests with the preconceived pathways of biochemistry textbooks, one would draw a variety of conclusions about essential enzymes and intermediates. It would, for example, appear safe to assume that the carboxylation of pyruvate to oxaloacetate must involve either *pyruvate carboxylase* or *oxaloacetate decarboxylase*. This assumption would not be correct, because there are non-obvious alternatives, such as the pathway of Figure 12. Other pathways discussed here (e.g., Figures 11 or 13) contain other non-obvious possibilities for different biotransformations.

Fundamental Constraints. Some of the most interesting results of applying the synthesis algorithm involve not particular pathways found, but rather demonstrations that no pathways exist to meet certain sets of specifications.

We discussed already the fact that there is no pathway that will reach aspartate (and consequently lysine) from glucose without going through oxaloacetate. A second interesting constraint that was uncovered by the algorithm refers to the maximum yield of the pathway:

- The yield can exceed 67% only if carbon dioxide is recovered by some bioreaction.

In effect, if we eliminate reactions that consume carbon dioxide, the yield is restricted to be 67% or less. A point to keep in mind is that these constraints only hold for the set of reactions present in our database. It is entirely possible that inclusion of additional reactions will change these results.

5 Concluding Remarks

The problem of synthesizing qualitatively feasible biochemical pathways was discussed in this chapter. With respect to thermodynamic feasibility, a group contribution technique that allows the estimation of equilibrium constants of bioreactions was described. With respect to stoichiometric requirements, an algorithm for pathway synthesis was presented, based on the iterative satisfaction of constraints, and the transformation of the initial set of reactions (which can be thought of as one-step pathways) into a final set of pathways which satisfy all constraints. The algorithm generates all biochemical production routes that satisfy a set of linear stoichiometric constraints; these constraints designate bioreactions and metabolites (in their role as reactants, products, or intermediates of the pathways) as required, allowed, or

prohibited. For the task of synthesis of biochemical pathways, this is the first algorithm that is formal and well-defined, with proven properties like completeness and correctness.

The algorithm is of significant value in the investigation of alternative biochemical pathways to achieve a given biotransformation (which is defined by a set of stoichiometric specifications). It can also produce pathways that bypass bottlenecks of a given pathway. A variety of alternative non-obvious routes for the synthesis of lysine demonstrates the utility of computer-based, systematic construction of pathways. Furthermore, the algorithm can identify fundamental limitations that govern the biochemical pathways and the process. In the case of lysine-producing pathways, it was shown that oxaloacetate is always present as an intermediate, and that in the absence of recovery of carbon dioxide by some bioreaction the yield of lysine over glucose is restricted to be 0.67 or less.

If the database of bioreactions is expanded to include a much larger number of bioreactions (and ultimately all known bioreactions), the computational performance of the algorithm would have to be drastically improved, in terms of both conceptual structure and actual implementation.

Notes

[1] The constraints are assumed to be consistent. For example, if S is a required product, it cannot be a required reactant.

[2] On Symbolics computers there are 36 bits in each word (32 bits for data and 4 bits for data-type). In LISP implementations on general-purpose hardware, one word might actually correspond to ~6 bytes. The distinction between list-words and structure-words is only important if one is recycling objects (and hence structure-words).

[3] To obtain smaller integer coefficients for the combination pathway, the fractions 1/3 and 2/3 were used instead of 1 and 2 in the construction of the combination. This has the same effect as dividing the resulting pathway by 3; clearly, the essence of the transformation and the overall significance of the pathway are not affected by multiplicative constants. Only the molar *proportions* of metabolites and reactions matter.

References

Barman, T. E. *Enzyme Handbook*, Supplement 1 (Springer-Verlag, New York, 1974).

Barman, T. E. *Enzyme Handbook*, Volume 1 (Springer-Verlag, New York, 1969).

Barman, T. E. *Enzyme Handbook*, Volume 2 (Springer-Verlag, New York, 1969).

Benson, S. W. *Thermochemical Kinetics* (Wiley, New York, 1968).

Benson, S. W., Cruickshank, F. R., Golden, D. M., Haugen, G. R., O'Neal, H. E., Rodgers, A. S., Shaw, R., and Walsh, R. *Chemical Rev.*, **69**, 279 (1969).

Domalski, E. S., and Hearing, E. D. "Estimation of the Thermodynamic Properties of Hydrocarbons at 298.15 K." *Journal of Physics and Chemistry Ref. Data*, **14**, 1637 (1988).

Edsall, J. T., and Gutfreund, H. *Biothermodynamics* (Wiley, New York, 1983).Hinz, H.-J. *Thermodynamic Data for Biochemistry and Biotechnology* (Springer-Verlag, New York, 1986).

Joback, K. G., and Reid, R. C. Estimation of Pure-Component Properties from Group Contributions. *Chem. Eng. Comm.*, **57**, 233 (1987).

Lehninger, A.E. *Biochemistry*, 2nd ed. (Worth, New York, 1975).

Lehninger, A.E. *Principles of Biochemistry* (Worth, New York, 1986).

Mandelstam, J., McQuillen, K., and Dawes, I. *Biochemistry of Bacterial Growth, 3rd edition*, pp. 163-165. Wiley, New York, 1982.

Mavrovouniotis, M. L, *Symbolic Computing in the Prediction of Properties of Organic Compounds*, Technical Report SRC TR 89-95 (Systems Research Center, University of Maryland, College Park, MD, 1989).

Mavrovouniotis, M. L. *Computer-Aided Design of Biochemical Pathways*. Ph.D. Thesis, Dept. of Chemical Engineering, Massachusetts Institute of Technology, 1989.

Mavrovouniotis, M. L. "Estimation of Properties from Conjugate Forms of Molecular Structures: The ABC Approach", 29: 1943-1953 *Industrial and Engineering Chemistry Research*, 1990a.

Mavrovouniotis, M. L. Group Contributions to the Gibbs Energy of Formation of Biochemical Compounds in Aqueous Solution. *Biotechnology and Bioengineering*, 36, 1070-1082, 1990b.

Mavrovouniotis, M. L. Estimation of Standard Gibbs Energy Changes of Biotransformations. *Journal of Biological Chemistry*, 266, 14440-14445, 1991.

Mavrovouniotis, M. L., Bayol, P., Lam, T.-K. M., Stephanopoulos, G., and Stephanopoulos, G. *Biotechnology Techniques*, **2**, 23 (1988).

Mavrovouniotis, M. L., Stephanopoulos, G., and Stephanopoulos, G. Computer-Aided Synthesis of Biochemical Pathways. *Biotechnology and Bioengineering*, 36, 1119-1132, 1990a.

Mavrovouniotis, M. L., Stephanopoulos, G., and Stephanopoulos, G. Estimation of Upper Bounds for the Rates of Enzymatic Reactions. *Chemical Engineering Communications*, 93, 211-236, 1990b.

Morrison, R. T., and Boyd, R. N. *Organic Chemistry, 3rd edition* (Allyn and Bacon, Boston 1973).

Mostow, J. "Rutgers Workshop on Knowledge-Based Design" *SIGART Newsletter* (90):19-32, October, 1984.

Mostow, J. "Toward Better Models of the Design Process" *AI Magazine* 6(1):44-56, Spring, 1985.

Old, R.W., and Primrose, S.B. *Principles of Gene Manipulation, 3rd edition*. Blackwell Scientific Publications, London, 1985.

Rawn, J. D. *Biochemistry*, pp. 883-888. Harper and Row, New York, 1983.

Reid, R. C., Prausnitz, J. M., and Poling, B. E. *The Properties of Gases and Liquids, 4th edition* (McGraw-Hill, New York, 1987).

Reid, R. C., Prausnitz, J. M., and Sherwood, T. K. *The Properties of Gases and Liquids, 3rd edition* (McGraw-Hill, New York, 1977).

Snyder, L.A., Freifelder, D., and Hartl, D.L. *General Genetics.* Jones and Bartlett Publishers, Boston, 1985.

Sober , H.A. (ed) *Handbook of Biochemistry* (CRC, Cleveland, Ohio, 1970).

Thauer, R. K., Jungermann, K., and Decker, K. *Bacteriological Reviews,* 41: 148 (1977).