BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 1 2007, pages 92–98
doi:10.1093/bioinformatics/btl554

Systems biology

# Recovering metabolic pathways via optimization

John E. Beasley[1,*] and Francisco J. Planes[1,2]

[1]Mathematical Sciences, Brunel University, Uxbridge, UB8 3PH, UK and [2]CEIT and TECNUN,
University of Navarra, Manuel de Lardizabal 15, 20018 San Sebastian, Spain

Received on June 6, 2006; revised on September 7, 2006; accepted on October 20, 2006
Advance Access publication October 26, 2006
Associate Editor: Alvis Brazma

**ABSTRACT**

**Motivation:** A metabolic pathway is a coherent set of enzyme catalysed biochemical reactions by which a living organism transforms an initial (source) compound into a final (target) compound. Some of the different metabolic pathways adopted within organisms have been experimentally determined. In this paper, we show that a number of experimentally determined metabolic pathways can be recovered by a mathematical optimization model.

**Contact:** john.beasley@brunel.ac.uk

## 1 INTRODUCTION

Figure 1 shows a simple example of a metabolic pathway, converting one compound (the source compound) into another (the target compound), as a directed graph. In this paper, we are concerned with the problem of, given a database of reactions/compounds, recovering via mathematics the specific set of reactions/compounds that have been experimentally determined to be active in a metabolic pathway.

Previous approaches to this problem have typically been based on utilizing a database of reactions/compounds to enumerate possible paths, satisfying various constraints, from the source compound to the target compound (Croes *et al*., 2005, 2006; Dooms *et al*., 2005, http://www2.info.ucl.ac.be/people/YDE/Papers/wcb05.pdf; King *et al*., 2005; Küffner *et al*., 2000; Mavrovouniotis *et al*., 1990; Mavrovouniotis, 1992; McShan *et al*., 2003; Seressiotis and Bailey, 1986, 1988). This is based on the fact that one of these paths must, by definition, correspond to the set of reactions/compounds involved in the experimentally determined (observed) pathway. Such approaches do not directly address pathway stoichiometry, instead that must (somehow) be deduced once the reactions/compounds involved are known, i.e. knowledge of the (ordered) set of reactions/compounds involved in the path from the source compound to the target compound does not uniquely define the pathway. To illustrate this, Figure 2 involves precisely the same source compound to target compound path as Figure 1, but is a non-trivial alternative pathway. In addition, enumeration approaches do not directly address the issue of the number of molecules of source compound converted into target compound; this is illustrated in Figure 3.

A fundamental difficulty with enumeration approaches, aside from the issues referred to above is that there are a large number of possible paths. Küffner *et al*., 2000, for example, found that there were some 500 000 paths from glucose to pyruvate. Because of this large number of possible paths all of the work reported to date has had limited and mixed success, frequently failing to unambiguously recover the experimentally determined pathway.

Our approach to recovering metabolic pathways is fundamentally different to previous approaches. We develop below a mathematical optimization model, which directly addresses pathway stoichiometry, to decide which reactions should be active in a pathway. When we solve our optimization model for a number of different pathways we do recover the experimentally determined pathway in many cases.

## 2 METHODS

### 2.1 Reaction variables and constraints

In our approach we have a database of R reactions (where each reaction has a specified direction so a reversible reaction contributes two different reactions to the total number R), which collectively involve C different compounds. Suppose we are seeking a pathway (coherent set of reactions) that transforms $Q_S$ molecules of source compound S into $Q_T$ molecules of target compound T. A reaction may, or may not, be active in the pathway. So we have the binary (zero-one) variable:

$z_r = 1$ if reaction r is active in the pathway, 0 otherwise ($r = 1, ..., R$) and the associated tick variable:

$t_r$ the number of ticks of reaction r in the pathway (this must be an integer variable ($\geq 0$) with value 0 if the reaction not active).

We need a constraint relating the number of ticks of a reaction to the zero-one variable signifying whether the reaction is active or not, this is:

$$t_r \leq M_1 z_r \quad r = 1, \ldots, \pm R, \tag{1}$$

where $M_1$ is a large positive constant that represents the maximum number of ticks of any reaction (since $z_r = 1$ implies $t_r \leq M_1$). If the reaction does not tick then it must be inactive, so we have the constraint:

$$z_r \leq t_r \quad r = 1, \ldots, R \tag{2}$$

### 2.2 Compound variables and constraints

Our approach involves deciding whether compounds are balanced (or not). A balanced compound is one where the number of molecules needed (consumed) is equal to the number produced. A compound which is balanced can either be active (number of molecules needed = number produced > 0) or inactive (number of molecules needed = number produced = 0) in the pathway. Considering Figure 1, for example, the active balanced compounds are C2, C5 and C6.

Let $n_{cr}$ be the number of molecules of compound c needed as input for one tick of reaction $r$ and $p_{cr}$ be the number of molecules of compound c produced as output by one tick of reaction $r$. For each compound $c$ ($c = 1, \ldots, C$) define:

$b_c = 1$ if for compound c the number of molecules needed is equal to the number produced (i.e. if $\sum_{r=1}^{R} n_{cr} t_r = \sum_{r=1}^{R} p_{cr} t_r$), 0 otherwise. If $b_c = 1$ compound c is balanced.

[*]To whom correspondence should be addressed.

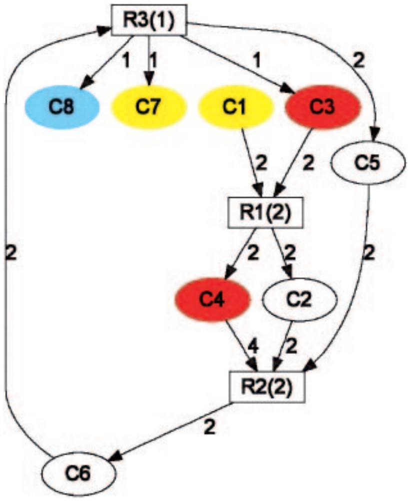

**Fig. 1.** An example metabolic pathway. The reactions and the compounds (labelled R and C, respectively) are the nodes in the above directed graph. The numbers associated with each arc are the number of molecules of each compound. For example, reaction R3 takes two molecules of C6 and transforms them into one molecule of C3, C7 and C8 and two molecules of C5. The source and target compounds (C1 and C7, respectively) are coloured yellow and two molecules of C1 are transformed into one molecule of C7. The numbers in brackets after each reaction label are the number of ticks, so for example reaction R1 ticks twice, each time converting one molecule of C1 and C3 into one molecule of C2 and C4. Compounds coloured blue are produced to excess (number of molecules needed is less than the number produced) whilst compounds coloured red are freely available (number of molecules needed is greater than the number produced). Compounds shown in white are balanced (number of molecules needed is equal to the number produced).

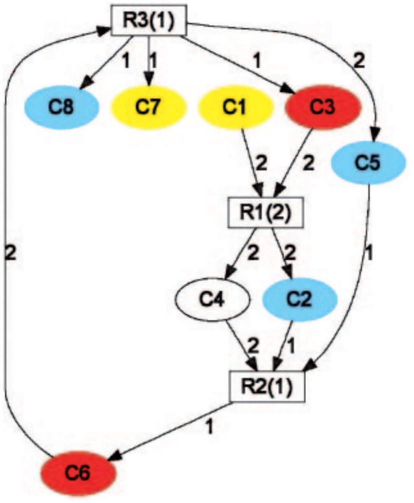

**Fig. 2.** A second pathway involving the same set of reactions/compounds. Using the path enumeration approach both Figures 1 and 2 correspond to a path solution of the form C1→R1→C2→R2→C6→R3→C7, from the source compound C1 to the target compound C7 (equivalently R1→R2→R3). Although, we have the same set of reactions/compounds here as in Figure 1 reaction ticks are not a common multiple of those shown in Figure 1. R1 and R3 have the same tick value as in Figure 1, R2 ticks once as opposed to twice in Figure 1. Hence, although we have the same path, we have a different pathway.

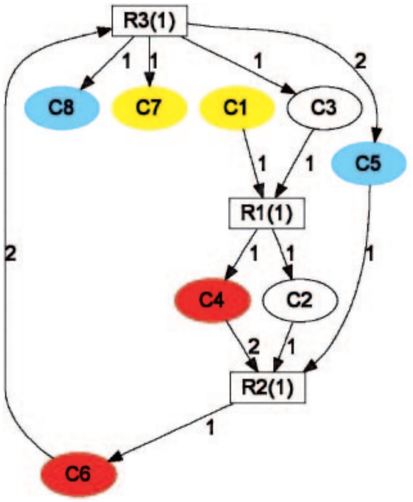

**Fig. 3.** A third pathway involving the same set of reactions/compounds. As for Figures 1 and 2 we have a path solution of the form C1→R1→C2→R2→C6→R3→C7. Here one molecule of C1 is transformed to one molecule of C7. In contrast, Figures 1 and 2 both transform two molecules of C1 into one molecule of C7.

$e_c = 1$ if for compound c the number of molecules needed is less than the number produced (i.e. if $\sum_{r=1}^{R} n_{cr}t_r < \sum_{r=1}^{R} p_{cr}t_r$), 0 otherwise. If $e_c = 1$ compound c is produced to excess, since we have 'spare' molecules of the compound to be disposed of (in other pathways).

$f_c = 1$ if for compound c the number of molecules needed is greater than the number produced (i.e. if $\sum_{r=1}^{R} n_{cr}t_r > \sum_{r=1}^{R} p_{cr}t_r$), 0 otherwise. If $f_c = 1$ compound c must be freely available, since we need 'spare' molecules of the compound that have come from other pathways.

Considering Figure 1, for example, compound C8 is produced to excess (denoted by the blue colouring) and compounds C3 and C4 are freely available (denoted by the red colouring). We have the constraint:

$$b_c + e_c + f_c = 1 \quad c = 1, \ldots, C. \tag{3}$$

In order to link the variables $e_c$ and $f_c$ to the number of molecules of each compound produced we need the constraints.

$$e_c \geq \left( \sum_{r=1}^{R} p_{cr}t_r - \sum_{r=1}^{R} n_{cr}t_r \right) / M_2 \quad c = 1, \ldots, C \tag{4}$$

$$e_c \leq 1 + \left( \sum_{r=1}^{R} p_{cr}t_r - \sum_{r=1}^{R} n_{cr}t_r - 1 \right) / M_2 \quad c = 1, \ldots, C \tag{5}$$

$$f_c \geq \left( \sum_{r=1}^{R} n_{cr}t_r - \sum_{r=1}^{R} p_{cr}t_r \right) / M_2 \quad c = 1, \ldots, C \tag{6}$$

$$f_c \leq 1 + \left( \sum_{r=1}^{R} n_{cr}t_r - \sum_{r=1}^{R} p_{cr}t_r - 1 \right) / M_2 \quad c = 1, \ldots, C, \tag{7}$$

where $M_2$ is a large positive constant. Equation (4) forces the zero-one variable $e_c$ to be one if $\sum_{r=1}^{R} n_{cr}t_r < \sum_{r=1}^{R} p_{cr}t_r$ whilst Equation (5) forces $e_c$ to be zero if $\sum_{r=1}^{R} n_{cr}t_r \geq \sum_{r=1}^{R} p_{cr}t_r$. Equations (6) and (7) are as Equations (4) and (5) but with $n_{cr}$ and $p_{cr}$ interchanged.

## 2.3 Metabolic constraints

The above has defined the variables that we need and the constraints that logically (mathematically) must be satisfied given these definitions. We now present the metabolic constraints that we included in our optimization approach.

We need constraints specifying that the required number of molecules of the source compound S ($Q_S$) and target compound T ($Q_T$) are involved—these are:

$$\sum_{r=1}^{R} n_{Sr} t_r = Q_S \text{ and } \sum_{r=1}^{R} p_{Tr} t_r = Q_T. \tag{8}$$

If the source compound and target compound are different then we produce none of the source compound and consume none of the target compound, i.e.

$$\sum_{r=1}^{R} p_{Sr} t_r = \sum_{r=1}^{R} n_{Tr} t_r = 0 \quad \text{if } S \neq T. \tag{9}$$

We have found it necessary in our approach to distinguish between compounds that appear in a significant number of different reactions and compounds that appear in just a few reactions. We define the percentage presence ($\delta_c$) of a compound c to be $\delta_c = 100$ (number of reactions in which c appears)/R $= 100 \sum_{r=1}^{R} \min[\max(p_{cr}, n_{cr}),1]/R$. Note that $\delta_c$ is defined purely with respect to the set of reactions that are considered and hence will vary as that set of reactions changes. Compounds for which $\delta_c \leq \Delta$, (where $\Delta$ is an input parameter) we call low presence compounds. Compounds for which $\delta_c > \Delta$ we call high presence compounds. Other authors (Croes *et al*., 2006; Horne *et al*., 2004; Jeong *et al*., 2000; Ma and Zeng, 2003; Wagner and Fell, 2001) have also found it necessary to distinguish compounds that commonly appear from those that appear less often when considering metabolic networks. In the computational results reported later we used $\Delta = 4\%$. Although this might seem a small value, for our relatively large database (R = 880 cytosolic reactions, involving C = 605 compounds) there were only 16 compounds (shown in Table 1) that had $\delta_c > \Delta$ and so were considered high presence compounds.

**Table 1.** High presence compounds

| Compound | Percentage presence |
|---|---|
| Hydrogen ion | 43.86 |
| Water | 28.98 |
| ATP | 18.98 |
| Adenosine diphosphate | 14.89 |
| Phosphate | 14.32 |
| Nicotinamide adenine dinucleotide | 9.77 |
| Nicotinamide adenine dinucleotide—reduced | 9.32 |
| Diphosphate | 8.98 |
| Nicotinamide adenine dinucleotide phosphate | 7.16 |
| Carbon dioxide | 7.05 |
| Nicotinamide adenine dinucleotide phosphate—reduced | 6.93 |
| L-Glutamate | 5.91 |
| Coenzyme A | 5.23 |
| Pyruvate | 4.77 |
| Ammonium | 4.43 |
| Adenosine monophosphate | 4.43 |

The logic behind this distinction is that high presence compounds appear in so many reactions that we can reasonably assume that if the metabolic pathway we are seeking either needs to obtain molecules of a high presence compound (produced by other pathways); or produces molecules of a high presence compound that have to be disposed of (in other pathways); then this can be achieved. High presence compounds can therefore be regarded as being 'freely available' or being 'produced to excess' if necessary. Another way to view high presence compounds is that they represent the interaction/interface between the pathway we are considering, (which is unknown, but is to be found) and all the other pathways that exist, (which are unknown, and remain unknown in terms of our mathematical model).

Low presence compounds, in contrast, cannot be reasonably assumed to be so easily obtained from, or disposed of in, other pathways and so must be balanced, i.e. any molecules involved must be internally produced/disposed of in the pathway chosen from S to T. Hence we have the constraint:

$$b_c = 1 \quad \text{if } \delta_c \leq \Delta; c \neq S, T; c = 1, \ldots, C, \tag{10}$$

which forces low presence compounds (excluding S and T) to be balanced. Note here that this constraint does not force compounds to be active in the pathway, merely to be balanced. Equation (10) links our approach to flux balance analysis (FBA), which underlies extreme pathways and elementary flux modes, as the requirement that compounds are balanced is the fundamental constraint applied there (Klamt and Stelling, 2003; Papin *et al*., 2003; Schilling *et al*., 1999, 2000; Schuster *et al*., 2000; Stelling *et al*., 2002).

A similarity between our approach and FBA is that low presence compounds correspond to internal and high presence compounds to external, compounds in FBA. However one difference is that in our approach external compounds can be in any state (freely available, produced to excess or balanced) and their state is decided as a result of solving our optimization model. A further difference between our approach and FBA is our focus on a single metabolic pathway.

Our approach is essentially different from either extreme pathways or elementary flux modes. From a linear optimization viewpoint these approaches take the entire feasible (continuous) solution space and identify a special set of solutions within it. For example, the extreme pathway set is such that all feasible solutions can be written as a non-negative linear combination of solutions in this set. Because we adopt (see below) an optimization objective we are seeking just a single solution, not a set of solutions. Moreover as our optimization model involves integer valued variables we no longer have a continuous solution space, rather a discrete disconnected solution space.

In our approach each reaction active in the pathway has at least one active balanced compound as an output, except any reaction producing the target compound T. Should a reaction be in the pathway and not satisfy this condition it can only be producing high presence compounds, which by definition are freely available anyway. Hence we impose the constraint:

$$\sum_{c=1,\, p_{cr}\geq 1}^{C} b_c \geq z_r \quad p_{Tr} = 0;\ r = 1, \ldots, R \tag{11}$$

We need to consider the issues of cycles (a closed path in the directed graph representation, e.g. C3-R1-C2-R2-C6-R3-C3 in Fig. 1) in the pathway. Cycles do exist in metabolic pathways, but in our approach some types of cycles are allowed, others are disallowed.

Each reaction in our database of R reactions has a specified direction associated with it. Define the set B $=\{(\alpha,\beta) \mid$ reaction $\alpha$ and reaction $\beta$ are the reverse of each other, $\alpha < \beta\}$. In order to disallow a cycle around a reaction and its reverse we impose the constraint:

$$z_\alpha + z_\beta \leq 1 \quad \forall(\alpha,\beta) \in B. \tag{12}$$

Considering a pathway as a directed graph we define a c-cycle to be an alternating sequence of c active balanced compounds and c active reactions that starts and ends at the same compound and within which no compound/reaction is repeated except at the start/end of the sequence. An example 2-cycle (C5-R2-C6-R3-C5) can be seen in Figure 1. In our approach we regard a c-cycle in a metabolic pathway as allowable if and only if:

- the source compound and the target compound are the same (S = T) and the c-cycle involves that compound or
- the c-cycle involves exactly one high presence balanced compound.

If the above conditions are not met then the c-cycle is disallowed.

The first of these conditions is a logical one. If S = T then the pathway must be a cycle by definition and so must be allowed. The second of these conditions is based on examination of known pathways. In a random sample of 25 pathways (taken from http://biocyc.org/ECOLI/, but excluding the ten pathways dealt with here) we found five pathways where there was an allowable c-cycle, but only one pathway where there was a disallowed c-cycle.

To illustrate this second condition if and only if exactly one of the two balanced compounds (C5 and C6) in the 2-cycle (C5-R2-C6-R3-C5) in Figure 1 is a high presence compound would the 2-cycle be allowed, otherwise it would be disallowed.

If a c-cycle is disallowed a constraint must be imposed to prevent it appearing in the pathway. To illustrate this consider the case c = 2. A 2-cycle involves two balanced compounds and two reactions. Any two reactions $\alpha$, $\beta$ for which there exist two compounds d, e for which: d is an input for $\alpha$ ($n_{d\alpha} > 0$); and e is an output from $\alpha$ ($p_{e\alpha} > 0$) and e is an input for $\beta$ ($n_{e\beta} > 0$) and d is an output from $\beta$ ($p_{d\beta} > 0$); gives rise to a potential 2-cycle (d-$\alpha$-e-$\beta$-d). If this 2-cycle is disallowed (it does not satisfy the conditions given above) then the constraint $b_d + z_\alpha + b_e + z_\beta \leq 3$ prevents it from appearing. In general the constraint required to prevent a c-cycle from appearing is: sum of the b variables for the compounds in the c-cycle plus sum of the z variables for the reactions in the c-cycle $\leq 2c - 1$.

## 2.4 Objective

Above we have set out a series of variables and constraints (a mathematical model) that we believe can be used to recover a metabolic pathway. It is likely that there is more than one feasible solution to the above mathematical model and so to arrive at a pathway we propose an objective that is to be optimized. Our computational results (reported below) indicate that two factors are of importance in terms of an optimization objective: the total number of reactions involved in the pathway and the number of excess molecules of Adenosine Triphosphate (ATP).

The total number of reactions involved in the pathway $\left(\sum_{r=1}^{R} z_r\right)$ should be minimized. This makes biological and evolutionary sense as minimising the number of reactions involved reduces the 'complexity' of the pathway. Broadly speaking we would expect that the fewer the reactions involved in a pathway the fewer the enzymes that will be needed by an organism to catalyse the reactions in the pathway. Moreover we would expect that the more reactions involved in a pathway the greater the chance that it may be disrupted, for example should an enzyme not be present due to a genetic defect. Other authors (Ebenhöh and Heinrich, 2003; Meléndez-Hevia, 1990; Meléndez-Hevia and Isidoro, 1985; Meléndez-Hevia and Torres, 1988; Meléndez-Hevia *et al*., 1994,1996; Mittenthal *et al*., 1998) have also emphasized minimization of the number of reactions involved in a metabolic pathway.

Denoting ATP as compound 1 for simplicity the number of excess molecules of ATP $\left(\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r\right)$ should be maximized.

This makes biological and evolutionary sense as ATP is a key metabolic compound. ATP has been termed the cell's energy currency and is the universal carrier of chemical energy in the cells of all living organisms from bacteria and fungi to plants and animals including humans. It captures the chemical energy released by the combustion of nutrients and transfers it to reactions that require energy. Previous work examining optimality criteria associated with the structure of metabolic pathways (Heinrich and Ebenhöh, 2001; Heinrich *et al*., 1997; Meléndez-Hevia *et al*., 1996,1997; Stephani and Heinrich, 1998; Stephani *et al*., 1999) has also focussed on the optimization of (net) ATP production.

Maximising excess ATP:

- if $\left(\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r\right) > 0$ produces as many 'spare' molecules of ATP as possible for use in other pathways
- if $\left(\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r\right) < 0$ uses as few 'spare' molecules of ATP (generated in other pathways) as possible in the pathway from S to T.

Attempting to minimize one factor (total number of reactions) whilst simultaneously maximising another (excess ATP) involves a tradeoff. Whilst this tradeoff can be treated in a number of ways (e.g. see Heinrich *et al*., 1991) in this paper we examine the two extreme cases of this tradeoff:

$$\text{minimize } M_3\left(\sum_{r=1}^{R} z_r\right) - \left(\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r\right) \qquad (13)$$

$$\text{maximize } M_3\left(\sum_{r=1}^{R} p_{1r}t_r - \sum_{r=1}^{R} n_{1r}t_r\right) - \left(\sum_{r=1}^{R} z_r\right), \qquad (14)$$

where $M_3$ is a large positive constant. Objective (13) gives primary weight to minimising the total number of reactions and secondary weight to maximising excess ATP, whilst objective (14) gives primary weight to maximising excess ATP and secondary weight to minimising the total number of reactions.

## 2.5 Overview

Our mathematical optimization model given above for recovering a metabolic pathway [optimize (13) or (14) subject to (1–12) plus c-cycle constraints] is a linear integer program. Algorithmically such programs are solved by linear programming based tree search. Modern software packages to perform this task, such as (ILOG CPLEX, 2005, http://www.ilog.com/products/cplex/news/whatsnew.cfm#cplex90), which we used, are well developed and highly sophisticated.

One computational point here deals with our treatment of c-cycles. We imposed constraints to prevent all disallowed 2-cycles directly and solved the integer program as given above. The solution obtained was then checked to see whether it contained any disallowed c-cycles (for any c>2). Finding a cycle in the directed graph composed of balanced compounds and active reactions is (algorithmically) an easy task, and checking to see whether a c-cycle is allowed or not is trivial. If any disallowed c-cycles were found then constraints to eliminate them (as discussed above) were added and the process repeated until a solution without any disallowed c-cycles was found.

Our model neglects three issues: bioenergetics (Gibbs free energy), enzymes and cofactors/coenzymes. Mathematically all of these can be easily incorporated into our model, but details as to this have been omitted here as the data available for the 880 reactions considered was not sufficient to enable any of these issues to be implemented computationally. Extending our model to deal with pathways with multiple source/target compounds is also easily done.

# 3 RESULTS

## 3.1 Recovery of known pathways

We applied our optimization model to the 10 pathways shown in Table 2, which includes a number of well-known pathways frequently encountered in biochemistry texts (e.g. Nelson and Cox, 2005). The reaction/compound database used was drawn from (Reed *et al*., 2003, http://systemsbiology.ucsd.edu/organisms/ecoli/ecoli_reactions.html) and the pathways from (Keseler *et al*., 2005; Nelson and Cox, 2005, http://biocyc.org/ECOLI/). One complication arises with pathways 5–10 in Table 2 in that they contain some low presence unbalanced compounds, [which our approach would force to be balanced, Equation (10)]. Hence for these pathways we did not force these compounds to be balanced [i.e. we excluded them from Equation (10)]. Table 2 indicates that for 9 of our 10 pathways one (or both) of our objectives results in the solution to our mathematical optimization model being precisely the same as the experimentally determined pathway. In other words we recover not only the reactions and compounds involved in the experimentally determined pathway but also its stoichiometry (reaction ticks). Statistically this is a highly significant result (significant at the 0.005% level).

**Table 2.** Metabolic pathways considered

| Pathway number | Pathway name | Pathway recovered? Objective (13) | Objective (14) |
|---|---|---|---|
| 1 | Gluconeogenesis | Yes | No |
| 2 | Glycogen | Yes | No |
| 3 | Glycolysis | Yes | Yes |
| 4 | Proline biosynthesis | Yes | No |
| 5 | Ketogluconate metabolism | No | No |
| 6 | Pentose phosphate | Yes | No |
| 7 | Salvage pathway deoxythymidine phosphate | Yes | No |
| 8 | Tricarboxylic acid (citric acid, citrate, TCA, Krebs) cycle | No | Yes |
| 9 | NAD biosynthesis | Yes | No |
| 10 | Arginine biosynthesis | Yes | No |
| Number of 'yes' entries | | 8 | 2 |

Note in particular that the TCA pathway is recovered by objective (14). Since this pathway is a cycle of reactions from one compound to itself (i.e. S = T, $Q_S = Q_T$) it is perhaps not surprising that this pathway is one that gives secondary weight to the total number of reactions (since we might expect that a cycle from S back to itself that involves just a few reactions can readily be found).

Based on Table 2 it appears clear that a pathway is best recovered by objective (13) if the source compound and target compound are different, but by objective (14) if the source compound and target compound are the same.

With respect to computation time the average computation time over the 20 cases shown in Table 2 was 11 s, no case requiring more than 85 s (3 GHz pc, 1 Mb RAM). For 9 of the 10 pathways in Table 2 optimizing using objective (14) took longer than optimizing using objective (13), on average five times longer.

In 9 of the 10 'yes' cases in Table 2 there is a unique pathway providing the optimal objective function value and in only one case is there an alternative pathway providing the same optimal objective function value.

As we have a significant number of constraints in our optimization model the question arises as to the relevance of the objective adopted. In the limit for example there may be only one unique solution satisfying the constraints, and if so the objective adopted becomes irrelevant. We have investigated this issue and have found that in all 10 'yes' cases in Table 2 we have more than one solution satisfying the constraints.

Equation (9) explicitly excludes solutions in which reactions in the pathway produce any of the source compound (or consume any of the target compound). If we amend our optimization model, (which is trivially done) to allow such solutions then, with respect to Table 2, we degrade the results slightly, failing to recover pathway 9, NAD biosynthesis. In a random sample of 25 pathways (taken from http://biocyc.org/ECOLI/, but excluding the 10 pathways dealt with here) we found only one pathway in which Equation (9) was violated (and that was for a pathway where the source compound was itself a high presence compound).

If we do not impose the constraint on allowable c-cycles then, with respect to Table 2, we degrade the results significantly. Objective (14) now fails to recover any pathway and objective (13) now only recovers 6 pathways.

**Table 3.** Sensitivity analysis relating to Δ

| Value of Δ (%) | Number of 'yes' entries Objective (13) | Objective (14) | Total |
|---|---|---|---|
| 2.5 | 4 | 0 | 4 |
| 3 | 7 | 2 | 9 |
| 3.5 | 7 | 2 | 9 |
| 4 | 8 | 2 | 10 |
| 4.5 | 8 | 5 | 13 |
| 5 | 8 | 6 | 14 |
| 5.5 | 8 | 6 | 14 |

As our approach is linked to FBA the question arises as to the results we would obtain were we to apply a FBA based approach. Here such an approach would be to optimize (13) or (14) subject to (1–10), i.e. excluding Equations (11) and (12) and c-cycle constraints. If we do this then, with respect to Table 2, we degrade the results significantly. Objective (13) now only recovers 5 pathways and objective (14) fails to recover the TCA pathway (only recovering one pathway).

## 3.2 Discussion

In our optimization model it is necessary to specify the user defined input parameter Δ, which determines whether a compound is a low presence, or a high presence compound. We conducted a sensitivity analysis as to how the results change as Δ changes. This can be seen in Table 3, where we have summarized the number of 'yes' entries that we obtained in the equivalent of Table 2 for varying Δ values. It is clear from this table that over a fairly wide range of Δ values a significant number of 'yes' entries are obtained.

Note here that the value of Δ = 4% associated with the results presented in Table 2 was originally chosen based on limited computational experience with a number of pathways. It was not chosen via systematic enumeration of results for all pathways for a range of Δ values and then selection of the best Δ value. As can be seen from Table 3 we could improve the results presented in Table 2 were we to use Δ = 5% for example.

In our optimization model it is necessary to specify the number of molecules of the source and target compounds ($Q_S$, $Q_T$) involved in the pathway. For the results shown in Table 2 these values have (obviously) been taken as equal to those associated with the experimentally determined pathway. Our optimization model can also recover the particular ($Q_S$, $Q_T$) values observed.

As an illustration of this the Gluconeogenesis pathway in Table 2 has $(Q_S, Q_T) = (2, 1)$, requiring nine reactions and consuming four molecules of ATP. For this pathway Table 4 shows for a number of different ($Q_S$, $Q_T$) pairs ($Q_S, Q_T \leq 6$) the number of reactions and the excess ATP when our optimization model is solved using objective (13). For this pathway our optimization model indicates that the pair $(Q_S, Q_T) = (2, 1)$ dominates all other cases (since it involves an equal number of reactions but uses less ATP). Hence in this case our optimization model recovers the $(Q_S, Q_T) = (2, 1)$ pair observed in the experimentally determined pathway.

**Table 4.** Optimization solution, expressed as (number of reactions, excess ATP), for varying $Q_S$ and $Q_T$ for gluconeogenesis

| | | Number of molecules $Q_T$ of target compound | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Number of molecules $Q_S$ of source compound | 1 | X[a] | X | X | X | X | X |
| | 2 | (9, −4) | X | X | X | X | X |
| | 3 | X | X | X | X | X | X |
| | 4 | X | (9, −8) | X | X | X | X |
| | 5 | X | X | X | X | X | X |
| | 6 | X | X | (9, −12) | X | X | X |

[a] Situations where our optimization model indicated that no feasible solution exists are indicated by X. In these cases no (integer) values for the decision variables exist which satisfy all the constraints [(1–12) plus c-cycle constraints] for the particular ($Q_S$, $Q_T$) pair examined.

We have repeated the analysis shown in Table 4 for the other pathways. Our judgment in that for 8 of the 10 pathways our optimization model recovers the ($Q_S$, $Q_T$) pair observed in the experimentally determined pathway. Statistically this is a highly significant result (significant at the 0.001% level).

## 4 CONCLUSIONS

In essence the approach given above hypothesises that metabolic pathways have evolved so as to be the optimal solution of the mathematical optimization model we have proposed (balancing minimising the number of reactions with maximising excess ATP, whilst subject to a variety of constraints). Our success (though not complete) at using our optimization model to recover the experimentally determined pathways, including their stoichiometry, we have examined supports this hypothesis.

Although for reasons of space we will not explore it here, having a mathematical model for metabolic pathway recovery advances our ability:

- to predict pathways, e.g. those resulting should a reaction not function due to a genetic defect meaning a catalysing enzyme is not available,
- to investigate how to disrupt pathways, e.g. by finding those reactions/enzymes that should be disabled so as to prevent efficient pathway functioning.

Being able to accomplish prediction and disruption via a mathematical model has clear and significant advantages over any other means.

## ACKNOWLEDGEMENT

*Conflicts of Interest:* none declared.

## REFERENCES

Croes,D. *et al*. (2005) Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Res*., **33**, W326–W330.

Croes,D. *et al*. (2006) Inferring meaningful pathways in weighted biochemical networks. *J. Mol. Biol*., **356**, 222–236.

Dooms,G. *et al*. (2005) Constrained metabolic network analysis: discovering pathways using CP(Graph).

Ebenhöh,O. and Heinrich,R. (2003) Stoichiometric design of metabolic networks: multifunctionality, clusters, optimization, weak and strong robustness. *Bull. Math. Bio*., **65**, 323–357.

Heinrich,R. *et al*. (1991) Mathematical analysis of enzymic reaction systems using optimization principles. *Eur. J. Biochem*., **201**, 1–21.

Heinrich,R. *et al*. (1997) Theoretical approaches to the evolutionary optimization of glycolysis. Thermodynamic and kinetic constraints. *Eur. J. Biochem*., **243**, 191–201.

Heinrich,R. and Ebenhöh,O. (2001) Evolutionary optimization of metabolic pathways. Theoretical reconstruction of the stoichiometry of ATP and NADH producing systems. *Bull. Math. Bio*., **63**, 21–55.

Horne,A.B. *et al*. (2004) Constructing an enzyme-centric view of metabolism. *Bioinformatics*, **20**, 2050–2055.

ILOG CPLEX (2005).

Jeong,H. *et al*. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Keseler,I.M. *et al*. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res*., **33**, D334–D337.

King,R.D. *et al*. (2005) On the use of qualitative reasoning to simulate and identify metabolic pathways. *Bioinformatics*, **21**, 2017–2026.

Klamt,S. and Stelling,J. (2003) Two approaches for metabolic pathway analysis? *Trends Biotech*., **21**, 64–69.

Küffner,R. *et al*. (2000) Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, **16**, 825–836.

Ma,H. and Zeng,A.-P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, **19**, 270–277.

Mavrovouniotis,M.L. *et al*. (1990) Computer-aided synthesis of biochemical pathways. *Biotechnol. Bioeng*., **36**, 1119–1132.

Mavrovouniotis,M.L. (1992) Synthesis of reaction-mechanisms consisting of reversible and irreversible steps. 2. Formalization and analysis of the synthesis algorithm. *Ind. Eng. Chem. Res*., **31**, 1637–1653.

McShan,D.C. *et al*. (2003) PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics*, **19**, 1692–1698.

Meléndez-Hevia,E. (1990) The game of the pentose phosphate cycle—a mathematical approach to study the optimization in design of metabolic pathways during evolution. *Biomedica Biochimica Acta*, **49**, 903–916.

Meléndez-Hevia,E. and Isidoro,A. (1985) The game of the pentose phosphate cycle. *J. Theor. Biol*., **117**, 251–263.

Meléndez-Hevia,E. and Torres,N.V. (1988) Economy of design in metabolic pathways—further remarks on the game of the pentose phosphate cycle. *J. Theor. Biol*., **132**, 97–111.

Meléndez-Hevia,E. *et al*. (1994) Optimization of metabolism: the evolution of metabolic pathways toward simplicity through the game of the pentose phosphate cycle. *J. Theor. Biol*., **166**, 201–220.

Meléndez-Hevia,E. *et al*. (1996) The puzzle of the Krebs citric acid cycle: assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. *J. Mol. Evol*., **43**, 293–303.

Meléndez-Hevia,E. *et al*. (1997) Theoretical approaches to the evolutionary optimization of glycolysis. Chemical analysis. *Eur. J. Biochem*., **244**, 527–543.

Mittenthal,J.E. *et al*. (1998) Designing metabolism: alternative connectivities for the pentose phosphate pathway. *Bull. Math. Bio*., **60**, 815–856.

Nelson,D.L. and Cox,M.M. (2005) *Lehninger Principles of Biochemistry, 4th edn*. Worth Publishers, NY.

Papin,J.A. *et al*. (2003) Metabolic pathways in the post-genome era. *Trends in Biochem. Sci*., **28**, 250–258.

Reed,J.L. *et al*. (2003) An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Gen. Biol*., **4**, R54.

Schilling,C.H. *et al*. (1999) Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog*., **15**, 296–303.

Schilling,C.H. *et al*. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol*., **203**, 229–248.

Schuster,S. *et al*. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotech*., **18**, 326–332.

Seressiotis,A. and Bailey,J.E. (1986) MPS: an algorithm and data base for metabolic pathway synthesis. *Biotech. Lett.*, **8**, 837–842.

Seressiotis,A. and Bailey,J.E. (1988) MPS—an artificially intelligent software system for the analysis and synthesis of metabolic pathways. *Biotech. & Bioeng.*, **31**, 587–602.

Stelling,J. *et al.* (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190–193.

Stephani,A. and Heinrich,R. (1998) Kinetic and thermodynamic principles determining the structural design of ATP-producing systems. *Bull. Math. Bio.*, **60**, 505–543.

Stephani,A. *et al.* (1999) Optimal stoichiometric designs of ATP-producing systems as determined by an evolutionary algorithm. *J. Theor. Biol.*, **199**, 45–61.

Wagner,A. and Fell,D.A. (2001) The small world inside large metabolic networks. *Proc. R. Soc. Lond. Ser. B*, **268**, 1803–1810.