The relevance of determining these metabolic pathways has been elementary flux modes (Schuster 1999), path finding approaches. Stoichiometric approaches, which underlie we classified these approaches into stoichiometric approaches and approaches have been proposed. In Planes and Beasley (2008), order to achieve this goal, different mathematical (computational) steps in the search to fully comprehend cellular organization. In Modelling these metabolic pathways represents one of the key noted in de Figueiredo (2009). This also restricts the scope of path finding approaches in studying more quantitative aspects of metabolism, e.g. maximal molar yield (Schuster et al., 1999). In addition, metabolic pathways do not necessarily constitute a linear sequence of biochemical reactions, e.g. pentose phosphate pathway or TCA cycle. However, path finding approaches have been shown to be a valid framework for analysing network properties (Jeong et al., 2000), gene expression (Kharchenko et al., 2005) and radioactive tracers (Arita, 2003). Different methods have been presented to obtain the whole set of elementary flux modes in a metabolic network (Schuster et al., 1999; Urbanczik and Wagner, 2005). However, as the metabolic network increases in size the number of elementary flux modes explodes in a combinatorial fashion (Klamt and Stelling, 2002). This makes a detailed analysis at the genome-scale impracticable.

Motivation: Different mathematical methods have emerged in the post-genomic era to determine metabolic pathways. These methods can be divided into stoichiometric methods and path finding methods. In this paper we detail a novel optimization model, based upon integer linear programming, to determine metabolic pathways. Our model links reaction stoichiometry with path finding in a single approach. We test the ability of our model to determine 40 annotated Escherichia coli metabolic pathways. We show that our model is able to determine 36 of these 40 pathways in a computationally effective manner.

ABSTRACT

Fig. 1. (A) An example metabolic network comprising eight reactions (labelled R1–R8, respectively) and eight compounds (labelled C1–C8, respectively). Reaction R3, for example, converts one molecule of C5 into two molecules of C4, one molecule of C6 and one molecule of C7. Each reaction has a specified direction so a reversible reaction contributes two different reactions. For example, R6 and R7 are the reverse of each other. (B) An elementary flux mode that converts C1 into C7. Note here that the elementary flux modes approach divides the full set of compounds into internal and external compounds (Schuster et al., 2000). We assume that (C2, C3, C5, C6, C8) are internal compounds and (C1, C4, C7) are external compounds. External compounds are shaded. The numbers in brackets after each reaction label are the reaction ticks (fluxes). (C) A pathway that contains a directed path (a metabolic path) from C1 to C7. Note here that path finding approaches need to define a source compound and a target compound, here C1 and C7, respectively. One metabolic path is C1→R1→C2→R2→C5→R3→C7, although there may be more than one metabolic path. Here there is another metabolic path C1→R1→C3→R2→C5→R3→C7.

1 INTRODUCTION

1.1 Background

It is well known that the vast repertoire of biochemical reactions occurring inside the cell is grouped into metabolic pathways. The relevance of determining these metabolic pathways has been extensively described in the literature (Schuster et al., 2006). Modelling these metabolic pathways represents one of the key steps in the search to fully comprehend cellular organization. In order to achieve this goal, different mathematical (computational) approaches have been proposed. In Planes and Beasley (2008), we classified these approaches into stoichiometric approaches and path finding approaches. Stoichiometric approaches, which underlie elementary flux modes (Schuster et al., 2000), define a metabolic pathway as a minimal set of biochemical reactions at pseudo steady state. Path finding approaches view a metabolic pathway as a set of biochemical reactions that contains a directed path from a given source compound to a given target compound (Arita, 2000). To illustrate the two different perspectives a simple example is shown in Figure 1.

As can be seen in Figure 1, metabolic (directed) paths, as opposed to elementary flux modes, typically omit stoichiometric information. Due to this fact, we may find misleading metabolic pathways, as noted in de Figueiredo et al. (2009). This also restricts the scope of path finding approaches in studying more quantitative aspects of metabolism, e.g. maximal molar yield (Schuster et al., 1999). In addition, metabolic pathways do not necessarily constitute a linear path (Arita, 2000). To whom correspondences should be addressed.
1988; Meléndez-Hevia reaction steps (Meléndez-Hevia, 1990; Meléndez-Hevia and Torres, 2000). However, path finding approaches have generally been concerned with finding a small number of paths (typically up to 10) as an optimization criterion. This is a well-known (and computationally tractable) problem in graph theory referred to as the k-shortest path problem. The hypothesis here is that relevant pathways can be found among the k-shortest paths. The main difficulty is how to provide biological significance to the sought k-shortest paths, since, as noted in Ma and Zeng (2003), not every possible path in the metabolic network is biologically meaningful, especially due to the bimolecular nature of biochemical reactions (Palsson, 2006). In order to meet this issue, path finding approaches generally consider two factors: distance metric and pruning of the metabolic network. Typically the distance metric utilized is the number of reaction steps (Arita, 2000; Jeong et al., 2000; Ma and Zeng, 2003; Rahmann et al., 2005; Wagner and Fell, 2001). Use of this metric is supported by more theoretical studies that have suggested that metabolic pathways have evolved to minimize the number of reaction steps (Meléndez-Hevia, 1990; Meléndez-Hevia and Torres, 1988; Meléndez-Hevia et al., 1994). A different distance metric can be found in Croes et al. (2005), where the connectivity of the intermediate compounds in the network is the metric to be minimized (thus avoiding paths through highly connected compounds such as ATP). We also conducted a similar analysis (Planes and Beasley, 2009), finding that, though not totally precise, this distance metric has an acceptable performance.

Pruning of the metabolic network rests on the fact that some arcs in the metabolic network are not biologically meaningful, as discussed in Figure 2. Different attempts have been made to properly prune the metabolic network. Ma and Zeng (2003) carried out this pruning manually. A more systematic procedure is found in the work of Arita (2000) and Rahmann et al. (2005), who account for atomic mapping and structural similarity, respectively. Other authors remove highly connected compounds from the metabolic network (Croes et al., 2005; Wagner and Fell, 2001).

1.2 Optimization model

Using stoichiometric approaches such as elementary flux modes to gain insight into the structure of a metabolic network is common. However, the number of reactions active in an elementary flux mode can be more than one hundred in a large metabolic network. For this reason, using elementary flux modes to represent actual and functional metabolic pathways (which typically involve only a small number of reactions) generates less insight. Building upon our previous work (Beasley and Planes, 2007; Planes and Beasley, 2008, 2009), we present in this article an optimization model for metabolic pathways that refines and integrates previous stoichiometric and path finding approaches.

Our model starts from the idea that metabolic pathways do contain at least one directed (metabolic) path from the source compound to the target compound. This is supported by the work of Ihnems et al. (2004), which suggests that metabolic flow is driven by linear pathways. Our model allows pathways to contain more than one directed path from the source compound to the target compound, as opposed to typical path finding approaches. This makes it suitable to determine (if necessary) metabolic pathways that are branched.

In addition, our model directly addresses pathway stoichiometry, namely by requiring that directed paths respect specified stoichiometric constraints. Stoichiometric approaches typically assume a steady-state condition by defining a subset of compounds (internal compounds) constrained to be balanced for every metabolic pathway. The requirement to balance all internal compounds is the reason why elementary flux modes involve a large number of reactions. As noted in Planes and Beasley (2008), the definition of this internal compound subset is ambiguous. Indeed, this classification is generally done manually, based on user supplied biological knowledge. In Beasley and Planes (2007), we systematically defined this subset of compounds (referred as to low presence compounds) based on compound connectivity in the metabolic network. The steady-state condition however does not apply for every annotated metabolic pathway. In other words, it is common to find co-substrates and by-products in metabolic pathways. In Beasley and Planes (2007), we manually excluded these compounds from satisfying a balancing constraint, and this is a major limitation of that work, since we need to know beforehand the low presence unbalanced compounds for each metabolic pathway.

In order to overcome this issue in this article, we require that intermediate compounds in metabolic pathways must be balanced, but do not explicitly state which compounds must be balanced. This ‘relaxed’ stoichiometric constraint (steady-state condition) allows the same compound to be balanced in some pathways, whilst being consumed or produced (i.e. unbalanced) in other pathways.

In essence, our model views a metabolic pathway as being made up of a finite set of metabolic pathways from the source compound to the target compound with the intermediate compounds on the metabolic pathway being balanced. Interestingly, our mathematical model provides a link between path finding and stoichiometric approaches. To the best of our knowledge, no approach in the literature has to date combined both perspectives.

In order to preserve relevant biochemical along the different reaction steps in the metabolic paths, we introduce a number of systematic pruning rules. These rules relate to highly connected compounds (referred to here as high presence compounds, e.g. ATP, ADP, h, nad, nadh), inorganic compounds (e.g. Fe, O2) and cofactors (e.g. fud-arh), which typically do not appear as intermediate compounds in annotated pathways. The direct removal of links containing these compound sets is not always a good strategy. Indeed, one may find biochemical reactions composed of only, for example, high presence compounds, e.g. nad + nadh → nadt + nadp. For this reason, we introduce a more sophisticated method for dealing with such links, as discussed in the ‘Methods’ section.
In our optimization model, the minimization objective adopted gives primary weight to a factor termed specificity, $\Psi$ (see ‘Methods’ section). The specificity of a pathway is a function related to the connectivity of intermediate compounds in the metabolic paths. Here we aim to avoid highly connected compounds appearing in section). The specificity of a pathway is a function related to the metabolic network and the parameters ($S, T, Q_S, Q_T, K$).

2 METHODS

In our model we have a metabolic network 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 that transforms $Q_T$ molecules of source compound $S$ into $Q_S$ molecules of target compound $T$ and contains a maximum of $K$ directed metabolic paths from the source compound to the target compound. So here the input to our model is the metabolic network and the parameters ($S, T, Q_S, Q_T, K$).

2.1 Variables and constraints related to metabolic pathway stoichiometry

The material in this section has been taken from Bradley and Plantas (2007). Refer to that paper for a more detailed description of the mathematical variables and constraints given below. The variables are:

- $z_r = 1$, if reaction $r$ is active in the pathway, 0 otherwise;
- $t_r$ the number of ticks of reaction $r$ in the pathway. This must be a non-negative integer variable with value 0 if the reaction is not active;
- $b_c = 1$, if for compound $c$ the number of molecules needed is equal to the number produced (so the compound is balanced), 0 otherwise;
- $e_c = 1$, if for compound $c$ the number of molecules needed is less than the number produced (so the compound is produced to excess), 0 otherwise;
- $f_c = 1$, if for compound $c$ the number of molecules needed is greater than the number produced (so the compound must be freely available), 0 otherwise.

Let $n_c$ be the number of molecules of compound $c$ needed as input for one tick of reaction $r$ and $p_r$ be the number of molecules of compound $c$ produced as output by one tick of reaction $r$. Let $M_1$ and $M_2$ be large positive constants. The constraints are:

\[
\begin{align*}
    z_r &\leq M_1 t_r & \forall r = 1, \ldots, R \\
    z_r &\leq t_r & \forall r = 1, \ldots, R \\
    b_c + e_c + f_c &\leq 1 \quad c = 1, \ldots, C \\
    e_c &\leq \sum_{r=1}^{R} p_r t_r - \sum_{r=1}^{R} n_r t_r & \forall c = 1, \ldots, C \\
    f_c &\leq \sum_{r=1}^{R} n_r t_r - \sum_{r=1}^{R} p_r t_r & \forall c = 1, \ldots, C \\
    f_c &\leq \sum_{r=1}^{R} t_r - \sum_{r=1}^{R} t_r & \forall c = 1, \ldots, C \\
    \sum_{r=1}^{R} n_r t_r &= Q_S \\
    \sum_{r=1}^{R} p_r t_r &= Q_T \\
    \sum_{r=1}^{R} t_r &= 0 & \text{if } S \neq T \quad \text{(9)}
\end{align*}
\]

Equation (1) ensures that if a reaction is not active ($z_r = 0$) then the tick value is also zero. Equation (2) ensures that if a reaction is active ($z_r = 1$) then the tick value is non-zero. Equation (3) is a logical constraint which says that a compound is either balanced, or produced to excess, or freely available. Equations (4) and (5) link the excess variable $e_c$ for compound $c$ to the reactions that involve compound $c$. Equations (6) and (7) perform the same role for the freely available variable $f_c$. Equation (8) ensures that we consume $Q_T$ molecules of source compound $S$ and produce $Q_S$ molecules of target compound $T$. Equation (9) ensures that we produce no source compound, nor consume any target compound.

2.2 Variables and constraints related to metabolic paths

The material in this section has been derived (in an obvious fashion) from that given in Planes and Bradley (2009). Refer to that paper for more insight into the mathematical variables and constraints given below. The variables are:

- $u_{Tk} = 1$, if the arc from reaction node $r$ to compound node $c$ is in metabolic path $k$, 0 otherwise;
- $v_{rk} = 1$, if the arc from reaction node $r$ to compound node $c$ is in metabolic path $k$, 0 otherwise.

If $p_r = 0$, i.e. compound $c$ is not an input compound for reaction $r$, the arc does not exist, so we set $u_{Tk} = 0 \forall k$. Similarly if $p_r = 0$ then $v_{rk} = 0 \forall k$.

The constraints are:

\[
\begin{align*}
    \sum_{r=1}^{R} u_{Tk} &= \sum_{r=1}^{R} v_{rk} = 1, & k = 1, \ldots, K \\
    \sum_{r=1}^{R} u_{T} &= \sum_{r=1}^{R} v_{T} = 0 \quad S \neq T \quad \text{(10)} \\
    \sum_{r=1}^{R} u_{rk} &= \sum_{r=1}^{R} v_{rk} = 0, & k = 1, \ldots, K \\
    c_{r} &= r, 1, \ldots, R; \quad k = 1, \ldots, K \quad \text{(11)} \\
    u_{Tk} &\leq 1, & r = 1, \ldots, R; \quad k = 1, \ldots, K \quad \text{(12)} \\
    v_{rk} &\leq 1, & r = 1, \ldots, R; \quad c = 1, \ldots, C \quad \text{if } S \neq T; \quad k = 1, \ldots, K \quad \text{(13)} \\
    c_{r} &= r, 1, \ldots, R; \quad k = 1, \ldots, K \quad \text{(14)} \\
    \sum_{r=1}^{R} u_{Tk} &\leq 1, & c = 1, \ldots, C \quad \text{if } S \neq T; \quad k = 1, \ldots, K \quad \text{(15)}
\end{align*}
\]

Equation (10) ensures that for each metabolic path $k$ we have one arc out of the source compound and one arc into the target compound. Equation (11) ensures that for each metabolic pathway we have no arc into the source compound, and no arc out of the target compound. Equation (12) ensures that the number of arcs (for each metabolic path) into a reaction node equals the number of arcs out. Equation (13) performs the same role for compound nodes. Equation (14) ensures that at most one arc enters a reaction node on each metabolic path. Equation (15) performs the same role for compound nodes.

We need to prevent cycles appearing in the solution. Consider metabolic path $k$ once we have solved our mathematical model. We may have a cycle for the non-zero variables ($u_{Tk}, v_{rk}$) associated with this path. If $S \neq T$, a cycle defines a path of successive arcs associated with non-zero variables ($u_{Tk}, v_{rk}$) in the directed network that starts and ends at the same compound, whilst if $S = T$ a cycle defines a path of successive arcs associated with non-zero variables ($u_{Tk}, v_{rk}$) in the directed network that starts and ends at the same compound and does not contain compound $S$.

The constraints to eliminate a cycle are: (sum of the $u_{Tk}$ and $v_{rk}$ variables for arcs appearing in the cycle) ≤ (number of arcs in the cycle − 1) $k = 1, \ldots, K$. Note that this cycle elimination constraint applies.
for all metabolic paths, irrespective of the metabolic path in which it was discovered. Technically this is because we are allowing K paths. If we detect a cycle in a solution associated with a specific metabolic path then unless we eliminate that cycle in all K paths simultaneously we may well find computationally that we solve K times, each time discovering what is essentially the same cycle, but with a different metabolic path, k label.

2.3 Linking constraints between metabolic pathway stoichiometry and the K metabolic paths

In this section, we show how to incorporate stoichiometric information into the K metabolic paths. Specifically, we need to link variables (υc, bc) related to metabolic pathway stoichiometry to variables (υr, bc) related to metabolic paths. Since the material in this (and all subsequent) sections has not been presented in the literature before, we describe our model in more detail.

First, we need to relate the appearance of a reaction node in a metabolic path to the variables signifying whether or not a reaction is present in the pathway. This is done by:

\[ \sum_{r=1}^{R} \sum_{c=1}^{C} \delta_{r,c} \geq \sum_{r=1}^{R} \kappa_{c} \quad c=1, \ldots, C \quad (16) \]

\[ \sum_{r=1}^{R} \sum_{c=1}^{C} \delta_{r,c} \geq \sum_{r=1}^{R} \kappa_{c} \quad c=1, \ldots, C \quad (17) \]

These constraints ensure that if υc is present in any of the K metabolic paths then υc is forced to be one. If an arc is not used then these constraints are inactive. Similarly if υr is zero then these constraints ensure that no arc associated with reaction r can be used in any metabolic path.

We impose the constraint that if a reaction is present in the pathway, then it must lie on one of the K metabolic paths. This constraint is:

\[ \sum_{r=1}^{R} \sum_{c=1}^{C} \delta_{r,c} \geq \sum_{r=1}^{R} \kappa_{c} \quad c=1, \ldots, C \quad (18) \]

which ensures that if a reaction r is active then we have at least one arc coming into that reaction associated with one of the K metabolic paths. Note that this constraint allows a reaction to be on more than one metabolic path.

With respect to the compound nodes in the metabolic paths, we impose the constraint that if an intermediate compound is on a metabolic path then the compound must be balanced. This constraint is:

\[ \sum_{r=1}^{R} \sum_{c=1}^{C} \kappa_{c} \leq \sum_{r=1}^{R} \kappa_{c} \quad r=1, \ldots, R \quad (19) \]

\[ \sum_{r=1}^{R} \sum_{c=1}^{C} \kappa_{c} \leq \sum_{r=1}^{R} \kappa_{c} \quad c=1, \ldots, C \quad (20) \]

which ensure that if an arc associated with an intermediate (c ≠ S, T) compound is used in any of the K metabolic paths then the compound must be balanced. If an arc is not used then these constraints are inactive. Similarly if υc is zero then these constraints ensure that any arc associated with compound c cannot be used in any metabolic path.

Equations (19) and (20) represent our 'relaxed' steady-state condition, since they force intermediate compounds on a metabolic path to be balanced, but do not explicitly state which compounds must be balanced. This allows the same compound to be balanced in some pathways, whilst being unbalanced in other pathways. Note here though that in our objective below we include a term related to minimizing the number of unbalanced compounds.

2.4 Constraints related to pruning of the metabolic network

In this section, we give some systematic rules to detect reactions/arc(s) in the metabolic network that cannot be involved in any of the metabolic paths in the pathway being considered.

It may be that for S and T there is a single reaction that has S as an input compound and T as an output compound. If such cases we may, perhaps, find a pathway that comprises just this reaction. If such a reaction exists then we exclude it from appearing in the pathway, i.e.

\[ \sum_{r=1}^{R} \sum_{c=1}^{C} \delta_{r,c} = 0, \quad \text{if} \quad \sum_{r=1}^{R} \kappa_{c} \geq 0 \quad r=1, \ldots, R \quad (21) \]

Table 1. Structural and (QS, QT) recovery for K = 1, ..., 5

<table>
<thead>
<tr>
<th>Number</th>
<th>Name</th>
<th>Structural recovery?</th>
<th>(QS, QT) recovery?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Glucose metabolism</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Glyoxylate cycle</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>Glycolysis</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Proline biosynthesis</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>Kynurenine metabolism</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>Pentose phosphate</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Salvage pathway</td>
<td>No</td>
<td>–</td>
</tr>
<tr>
<td>8</td>
<td>Deoxysynephrine phosphate</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>Tetracycline biosynthesis</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>10</td>
<td>Heme biosynthesis</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>11</td>
<td>Arginine biosynthesis</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>12</td>
<td>Anthraquinone biosynthesis</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>13</td>
<td>Xanthine biosynthesis</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>14</td>
<td>Adenine biosynthesis</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>15</td>
<td>Threonine biosynthesis</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>16</td>
<td>Glycolysis</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>17</td>
<td>De novo synthesis of pyrimidine ribonucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>18</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>19</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>20</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>21</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>22</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>23</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>24</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>25</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>26</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>27</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>28</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>29</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>30</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>31</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>32</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>33</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>34</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>35</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>36</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>37</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>38</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>39</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>40</td>
<td>De novo synthesis of pyrimidine nucleotides</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

We define the percentage presence \( k_{c} \) of a compound c to be \( k_{c} = \frac{100}{n_{c}} \) (number of reactions in which c appears)/R = 100\( \sum_{c=1}^{C} \min(\delta_{r,c}, \kappa_{c}) \)/R. Compounds for which \( k_{c} \geq \Delta \) (where Δ is an input parameter) we call low presence compounds. Compounds for which \( k_{c} < \Delta \) we call high presence compounds. The set of high presence compounds is \( D_{H} = \{ c | \kappa_{c} > \Delta \} \). In the results shown in Table 1 we used \( \Delta = 4\% \).

We denote \( D_{H} \) as the set of all (inorganic) compounds which do not involve carbon in their molecular composition. Hence \( D_{H} = \{ c | \text{compound } c \text{ does not include carbon in its molecular composition} \} \).
We impose the constraint that if a compound is in the set of high presence compounds $D_1$, or in the set of inorganic compounds $D_2$, then it cannot be on a metabolic path through a reaction provided it is possible for that reaction to have other compounds on such a path.

For a reaction $r$, other compounds associated with $r$ exist and can be on a metabolic path into $r$ if and only if there exists an input compound $c$ for reaction $r$ (i.e. $n_c \leq 1$) with $c \notin D_1 \cup D_2 - \{r\}$. For each reaction $r$ that satisfies this condition, we impose the constraint that no compound in $D_1 \cup D_2$ (but excluding source) can be on a metabolic path into $r$, i.e.

$$u_{kr} = 0, \forall c \in D_1 \cup D_2 - \{r\}, \quad k = 1, \ldots, K$$ (22)

For a reaction $r$, other compounds associated with $r$ exist and can be on a metabolic path out of $r$ if and only if there exists an output compound $c$ for reaction $r$ (i.e. $n_c \geq 1$) with $c \notin D_1 \cup D_2 - \{T\}$. For each reaction $r$ that satisfies this condition, we impose the constraint that no compound in $D_1 \cup D_2$ (but excluding target) can be on a metabolic path out of $r$, i.e.

$$v_{rak} = 0, \forall c \in D_1 \cup D_2 - \{T\}, \quad k = 1, \ldots, K$$ (23)

We denote $D_s$ as the set of cofactors, formally $D_s = \{(a,\beta) | \text{compounds } a \text{ and } \beta \text{ are a cofactor pair}\}$ so $D_s$ is a set of compound pairs, not simply a set of compounds (cf $D_1$ and $D_2$ above). Note here that the set of cofactors used in this article was obtained by a systematic analysis of the frequency with which pairs of compounds appear in the same reaction (see Supplementary Material).

For convenience in presenting our constraints in a mathematical form below we adopt the convention that each compound pair appears twice in $D_s$ (i.e. if we had a single cofactor pair composed of $q_8$ and $q_{62}$, then we would have $D_s = \{(q_8,q_{62}), (q_{62},q_8)\}$).

If two compounds ($a$ and $\beta$) are a cofactor pair then we impose the constraint that the compounds in this cofactor pair cannot be on the metabolic path through any reaction that involves them both provided it is possible for that reaction to have other compounds on such a path.

Consider each reaction $r$ in turn. Consider each cofactor pair $(a,\beta) \in D_s$ in turn. If $(a,\beta)$ satisfies:

• $a \neq S$ and $\beta \neq T$ (so the cofactor pair does not involve either the source compound or the target compound);
• $n_{ar} \geq 1$ and $p_{\alpha r} \geq 1$ (so the cofactor pair is involved with reaction $r$ with $a$ as the input compound and $\alpha$ as the output compound); and
• for reaction $r$ there exists an input compound $\lambda$ (i.e. $n_{r\lambda} \geq 1$) with $\lambda \neq a$, $\lambda \notin D_1 \cup D_2 - \{S\}$ and an output compound $\mu$ (i.e. $p_{\mu r} \geq 1$) with $\mu \neq \beta$, $\mu \notin D_1 \cup D_2 - \{T\}$ such that $(\lambda,\mu) \notin D_s$ (so the pair $(\lambda,\mu)$ is not itself a cofactor pair).

Then we impose the constraint that the cofactor pair $(a,\beta)$ cannot be on a metabolic path through $r$, i.e.

$$u_{\alpha r} = v_{\mu r} = 0, \quad k = 1, \ldots, K$$ (24)

Finally, for notational convenience, denote $D_{s0} = \{(c | c \notin D_1 \cup D_2 \text{ and } c = a, \beta, \lambda, \mu) \in D_s \text{ and } n_c = 1, \ldots, C\}$ as the set of main compounds. A main compound is a low presence, organic compound not involved in any cofactor pair.

### 2.5 Objective function

We define $N_c$ to be the relative presence for compound $c$ as input to reaction $r$, formally $N_{cr} = h_c \min\{a_i, p_{cr} > 0, a_1, \ldots, C\}$ if $p_{cr} > 0$, (else $N_{cr}$=0, if $p_{cr}$=0). If a compound $c$ has the lowest percentage presence over all input compounds for reaction $r$ then $N_{cr}$ will have the value one. For $N_{cr}$=1, then compound $c$ can be regarded as the most specific of the input compounds for reaction $r$, since it participates in less biochemical reactions in the metabolic network than any other input compound for that reaction.

We define $P_{cr}$ to be the relative presence for compound $c$ as output from reaction $r$, formally $P_{cr} = h_c \min\{a_i, p_{cr} > 0, a_1, \ldots, C\}$ when $p_{cr} > 0$; (else $P_{cr}$=0, if $p_{cr}$=0). If a compound $c$ has the lowest percentage presence over all output compounds for reaction $r$, then $P_{cr}$ will have the value one. Similar to input compounds if $P_{cr}$=1 then compound $c$ can be regarded as the most specific of the output compounds for reaction $r$.

We define the specificity of a pathway, $\Psi$, using

$$\Psi = \sum_{k=1}^{E} \sum_{c=1}^{C} D_{ncr,ucrk} + P_{cr} + \frac{1}{\lambda}$$ (25)

Equation (25) is a weighted sum over all metabolic paths of the arcs in the paths, each arc having a weight equal to its relative presence ($\Psi$ simply being a scaling factor). Specificity is bounded from below by the number of reaction steps in the pathway, since if the metabolic paths in the pathway consisted of the input/output compounds with the lowest percentage presence for the reactions used then the specificity of the pathway would equal the number of reaction steps.

We denote by $W$, the number of unbalanced main compounds (excluding the source and target compound), formally

$$W = \sum_{c=1; k \neq k_1}^{C} (e_c + f_c).$$ (26)

In our optimization model, we seek to minimize both specificity and $W$. We believe that minimizing specificity is appropriate as minimizing that is achieved by having the most specific compounds (those little used elsewhere in the metabolic network) on the metabolic paths, and compounds on such paths must be balanced (Equations (19) and (20)). Minimizing specificity will also typically ensure that high presence compounds are not included on a metabolic path (cf. Creso et al., 2005, 2006). We believe that minimizing $W$ is appropriate as minimizing that will minimize the number of unbalanced main compounds. In order to deal with this bi-objective problem we adopt a simple linear weighted objective, namely:

$$\min 10\Psi + W.$$ (27)

The relative weighing between $\Psi$ and $W$ ($\Psi:W=10:1$) was decided empirically based on computational results from a few pathways.

### 2.6 Constraints related to ATP and NADH production

Denoting (for simplicity) $ATP$ as compound 1 and $NADH$ as compound 2 we can force these compounds to be produced using the following constraints:

$$\sum_{r=1}^{R} p_{r1} - \sum_{r=1}^{R} n_{c1} \geq 1,$$ (28)

$$\sum_{r=1}^{R} p_{r2} - \sum_{r=1}^{R} n_{c2} \geq 1.$$ (29)

For a cyclic pathway, defined to be a pathway satisfying $S = T$ and $Q_S = Q_T$, then it is clear that the pathway must have a function beyond converting source compound to target compound. For the two cyclic pathways we examined, pathways 8 and 36, see ‘Results’ section, pathway 8 was recovered when Equations (28) and (29) were imposed; pathway 36 when Equation (29) was imposed (with $K=1$, which seems reasonable for cyclic pathways). The Glycolysis pathway (pathway 3) was recovered when Equation (28) was imposed. Equations (28) and (29) were not imposed for any of the other pathways.

### 2.7 Summary

Our optimization model, an integer linear program, is optimized (27) subject to (1)-(26), plus (28), (29) if appropriate. We need to test whether the model we have presented is capable of predicting known pathways.

Aside from knowledge of the metabolic network, our model requires as input parameters $(S, T, Q_S, Q_T, K)$, so the source and target compounds $(S, T)$, the number of molecules of source and target $(Q_S, Q_T)$ and the maximum number of directed metabolic paths $(K)$.
As will become apparent from the computational results given below, the parameter $K$ turns out to be relatively unimportant. Clearly $S$ and $T$ are vital. The role of $Q_S$ and $Q_T$ is more interesting. Logically one can ask why the number of molecules of source ($Q_S$) and target ($Q_T$) is needed as input information for the model—surely the model itself should be able to decide (output) appropriate values for $Q_S$ and $Q_T$. In fact, as discussed in the Results section below, our model can be used to decide values for $Q_S$ and $Q_T$ by considering as input a number of different ($Q_S$, $Q_T$) pairs and then picking the dominant pair.

3 RESULTS

3.1 Data

Our model was applied to the 40 *Escherichia coli* annotated pathways shown in Table 1. We used the *E.coli* metabolic network of Reed et al. (2003), which comprises 880 cytosolic reactions and over 600 compounds. The pathways used were taken from Keseler et al. (2005), Nelson and Cox (2005) and http://biocyc.org/ECOLI. We used Cplex (http://www.ibm.com/products/cplex) to solve our integer linear program optimization model. Details as to the source and target compounds for each pathway in Table 1 can be found in Supplementary Information.

3.2 Structural recovery

Results are shown in Table 1. In that table “structural recovery” means that, once our optimization model is solved, the solution is precisely the same as the annotated metabolic pathway, both in terms of the reactions/compounds involved in the pathway and in terms of its inherent stoichiometry (reaction ticks/fluxes). For structural recovery results we used input parameters ($S$, $T$, $Q_S$, $Q_T$) taken from the annotated pathway.

Our model also needs an input parameter $K$, which defines the maximum number of metabolic paths. The fundamental reason for considering $K$ metabolic paths was to meet the issue of branched metabolic pathways. We applied our model for $K=1, 2, 3, 4, 5$ so as to see performance for different $K$ values. Note here that allowing a maximum of $K$ metabolic paths does not imply that the solution will contain $K$ different metabolic paths (since the model allows the solution to contain $K$ copies of exactly the same metabolic path).

Table 1 indicates that we achieved structural recovery for 35 of our 40 annotated pathways when $K=1$, and 36 pathways when $K=2, 3, 4, 5$. Our results for structural recovery for $K=1$ differ from those for $K=2, 3, 4, 5$ only in that pathway 6 (pentose phosphate) needs $K\geq 2$ to be recovered. Although there is no single description of this pathway we treated it as a non-cyclic branched pathway containing two distinct metabolic paths (for a fuller discussion of this pathway see Supplementary Information). Structural recovery results for $K=2$ are identical to those for $K=3, 4, 5$; although average computation time over the forty pathways increases from 7.5 s for $K=2$ (1.86 GHz pc, 2 GB RAM) to 69.6 s for $K=5$. For this reason we henceforth fix $K=2$.

Our objective function gives primary weight to minimizing the number of molecules of source and target compounds ($Q_S$, $Q_T$) involved in the pathway. For structural recovery, as mentioned above, these values have been taken as equal to those associated with the annotated pathway.

In order that our model might be used to decide (i.e. output) appropriate ($Q_S$, $Q_T$) values we adopted the procedure of taking our model and solving it for different input ($Q_S$, $Q_T$) pairs ($Q_S$, $Q_T$ fixed) this leads to 36 different solutions, one for each ($Q_S$, $Q_T$) pair. By examining all of these 36 solutions, we identify a single dominant solution. For reasons of space we have not elaborated on how this is done in any detail here, but precise details are given as Supplementary Material. We refer to the ($Q_S$, $Q_T$) pair associated with this dominant solution as the dominant ($Q_S$, $Q_T$) pair. If the dominant pair is exactly the same as the number of molecules of source and target seen in the annotated pathway, then our model does recover the ($Q_S$, $Q_T$) pair observed in the annotated pathway.

The summary of our analysis with respect to ($Q_S$, $Q_T$) can be seen in Table 1. For 33 of the 36 structurally recovered pathways,
our model also recovers the \([Q_S, Q_T]\) pair observed in the annotated pathway.

4 SUMMARY

On the basis of results discussed above we have that, for 36 out of 40 pathways:

- given \(S, T\), \([Q_S, Q_T]\) recovery together, our results are highly significant (\(P\)-value \(1.64 \times 10^{-4}\), one-tailed test, see Supplementary Material).

Note that the phrase ‘optimization solution’ here simply means the solution from our model, which is obviously dependent on the reaction/compound database used. We are not claiming that the pathway we find is optimal in the sense of being the best possible pathway that can exist (such as for example as is addressed in the work of Meléndez-Hevia et al. (1994)).

5 DISCUSSION

The analysis of large metabolic networks from a pathway-oriented perspective constitutes a critical activity in systems biology. Early stoichiometric approaches, which underlie elementary flux modes, introduced an elegant mathematical formalism for the modelling of metabolic pathways. However, they encounter severe difficulties due to the combinatorial explosion. Path finding approaches have emerged as a (computationally) effective alternative for the analysis of pathways at the genome-scale. However, their inherent simplicity can lead to misleading conclusions and the results from such approaches are of a different nature from those derived from stoichiometric approaches. In this article, we have shown that a computationally effective mixed approach is possible via an integer linear programming optimization model.

Our model uses an objective function based on the connectivity of intermediate compounds in the metabolic paths. As shown in the Results section, this objective achieves good overall performance. Our results contrast with the work done by Jeong et al. (2000), which suggested the importance of hubs in metabolic pathways.

The fact that we have non-recovered pathways indicates that our model can be refined, perhaps by modifying the objective function and/or constraints presented in the Methods section. Other objectives are clearly possible, e.g. directly minimizing the number of reactions or minimizing the number of unbalanced main compounds. Work that we have carried out examining these objectives has not, however, yielded a better objective than that reported in this article.

In summary of this article, we presented a computationally effective mathematical optimization model for metabolic pathways. Our numeric results provide evidence as to the validity of our model. We believe that our model can serve as a reference for future work to build novel mathematical approaches that allow us to gain insight into the properties and functions of metabolic networks.

Conflict of Interest: none declared.

REFERENCES