Bulletin of Mathematical Biology (2009) 71: 750–769
DOI 10.1007/s11538-008-9380-8

ORIGINAL ARTICLE

Efficient Reconstruction of Metabolic Pathways by Bidirectional Chemical Search

Liliana Félix[a], Francesc Rosselló[b,c], Gabriel Valiente[a,c,*]

[a] *Algorithms, Bioinformatics, Complexity and Formal Methods Research Group, Technical University of Catalonia, 08034 Barcelona, Spain*
[b] *Department of Mathematics and Computer Science, University of the Balearic Islands, 07122 Palma de Mallorca, Spain*
[c] *Research Institute of Health Science (IUNICS), University of the Balearic Islands, 07122 Palma de Mallorca, Spain*

Received: 30 November 2007 / Accepted: 24 November 2008 / Published online: 20 December 2008

Abstract One of the main challenges in systems biology is the establishment of the metabolome: a catalogue of the metabolites and biochemical reactions present in a specific organism. Current knowledge of biochemical pathways as stored in public databases such as KEGG, is based on carefully curated genomic evidence for the presence of specific metabolites and enzymes that activate particular biochemical reactions. In this paper, we present an efficient method to build a substantial portion of the artificial chemistry defined by the metabolites and biochemical reactions in a given metabolic pathway, which is based on bidirectional chemical search. Computational results on the pathways stored in KEGG reveal novel biochemical pathways.

Keywords Artificial chemistry · Biochemical reaction · Metabolic pathway

1. Introduction

Metabolism can be regarded as a network of chemical reactions activated by enzymes and connected via their substrates and products, and a metabolic pathway can be regarded as a coordinated sequence of biochemical reactions (Deville et al., 2003). The definition of a metabolic pathway is not exact, and most pathways constitute indeed highly intertwined cyclic networks. In a cell, the substrates of a pathway are usually the products of another pathway, and there are junctions where pathways meet or cross (Karp and Mavrovouniotis, 1994).

*Corresponding author.
E-mail addresses: lfelix@lsi.upc.edu (Liliana Félix), cesc.rossello@uib.es (Francesc Rosselló), valiente@lsi.upc.edu (Gabriel Valiente).
This work has been partially supported by the Spanish CICYT project TIN2004-07925-C03-01 GRAMMARS, by Spanish DGI projects MTM2006-07773 COMGRIO and MTM2006-15038-C02-01, and by EU project INTAS IT 04-77-7178.

The analysis of metabolic pathways is motivated by the rapidly increasing quantity of available information on metabolic pathways for different organisms. One of the most comprehensive sources of metabolic pathway data is the Roche Applied Science Biochemical Pathways chart (Michal, 1999). There are also several databases on metabolic pathways, such as aMAZE (Lemer et al., 2004), BRENDA (Schomburg et al., 2002), MetaCyc (Caspi et al., 2006), KEGG (Kanehisa and Goto, 2000), and WIT (Overbeek et al., 2000). These databases contain hundreds of metabolic pathways and thousands of biochemical reactions, and even the metabolic pathway for a small organism constitutes a large network. For instance, the proposed metabolic pathway for the bacterium *E. coli* consists of 436 compounds (substrates, products, and intermediate compounds) linked by 720 reactions (Edwards and Palsson, 2000).

An artificial chemistry (Dittrich et al., 2001), on the other hand, is a computational model of a chemical system that consists of a set of objects (molecules), a set of reaction rules (that allow for the production of new molecules from already existing molecules), and a definition of the dynamics of the system (that is, application conditions for the reaction rules), aimed at answering qualitative questions about the chemical system. Thus, artificial chemistries model real chemistries, in which molecules represent chemical compounds and reaction rules represent chemical reactions and, in particular, artificial chemistries model organic chemistries (Benkö et al., 2003a, 2003b, 2004).

The chemical description of molecules in an artificial chemistry can be made at different levels of resolution, from simple molecular descriptors to structural formulas. One of these representations are *chemical graphs*, with nodes corresponding to the atoms of the molecules and edges indicating the bonds between them. Chemists have used chemical graphs to distinguish isomers since the second half of the nineteenth century, and in first course organic chemistry classes, chemical reactions are explained in terms of constitutional formulas and a handful of reaction mechanisms, which are nothing but chemical graphs and rules to modify them by means of breaking, forming, and changing the type of bonds. This leads in a natural way to artificial chemistries based on labeled graphs as molecules and graph transformation rules as reactions. Several such artificial chemistries have been proposed so far: see, for instance, (Benkö et al., 2003a, 2003b, 2004; McCaskill and Niemann, 2001; Rosselló and Valiente, 2005a).

Artificial chemistries can also be used to model biochemical systems such as metabolic pathways, in which molecules represent metabolites and reaction rules represent biochemical reactions (Rosselló and Valiente, 2005b), and they allow for answering qualitative questions about metabolism. In this paper, we present an efficient method to build a substantial portion of the artificial chemistry defined by the metabolites and biochemical reactions in a given metabolic pathway. Our method is based on bidirectional chemical search, and its implementation uses chemical graphs to represent sets of molecules. We report also on the results of some experiments applying this method to pathways stored in KEGG, which reveal novel biochemical pathways.

2. Modeling biochemical reactions as chemical graph transformations

Following (Rosselló and Valiente, 2005a), by a *chemical graph,* we understand a complete labeled weighted graph (V, E, ℓ, μ), with (V, E) an undirected graph (without multiple edges or self-loops), ℓ a labeling mapping that labels every node $v \in V$ with a chemical

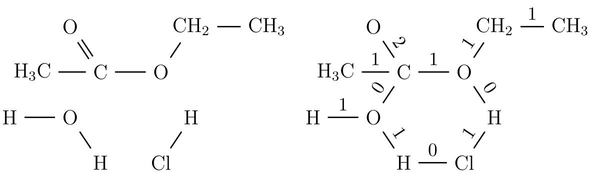


Fig. 1 A multi-molecule and a simplified representation of it as a chemical graph. Only some weight 0 edges that make the graph connected are shown for clarity.

element $\ell(v)$, and $\mu : E \to \mathbb{N}$ an edge weight function. We shall denote the weight of the edge joining nodes v and w by $\mu(v, w)$; notice that $\mu(v, w) = \mu(w, v)$ because the graph is undirected. A weight of 0 stands for a nonexisting bond, a weight of 1 for a single bond, a weight of 2 for a double bond, etc. The *valence* of a node in a chemical graph is the total weight of the edges incident to it.

To simplify the language, we shall call a *multi-molecule* to any set of molecules. Such a multi-molecule is described by the disjoint union of the chemical graphs representing the molecules and then adding weight 0 edges between atoms of different molecules. In this way, the molecules in the set are identified as maximal connected subgraphs with nonzero weight edges; see Fig. 1.

Given two chemical graphs $G_1 = (V_1, E_1, \ell_1, \mu_1)$ and $G_2 = (V_2, E_2, \ell_2, \mu_2)$, an *atom mapping* between them is a bijection $M : V_1 \to V_2$ such that, for every $v_1 \in V_1$:

- $\ell_1(v_1) = \ell_2(M(v_1))$.
- $\sum_{w_1 \in V_1} \mu_1(v_1, w_1) = \sum_{w_1 \in V_1} \mu_2(M(v_1), M(w_1))$.

When there exists an atom mapping between two chemical graphs G_1 and G_2, these chemical graphs (and the multi-molecules they represent) are said to be *compatible*: this means that they have the same number of nodes for each possible pair (label, valence). Notice that there is no stereochemical information in this simplified representation, and thus stereoisomers are represented by the same chemical graph. There is no electrical charge information either, and anions and cations are also represented by the same chemical graph.

A *chemical reaction graph* is a structure $R = (G_1, G_2, M)$, where $G_1 = (V_1, E_1, \ell_1, \mu_1)$ and $G_2 = (V_2, E_2, \ell_2, \mu_2)$ are compatible chemical graphs, called the *substrate* and the *product* chemical graphs, respectively, and $M : V_1 \to V_2$ is an atom mapping between them.

The application of a chemical reaction graph to a given chemical graph, consists of breaking, forming, and changing bonds in a subgraph of the chemical graph which is isomorphic to the substrate of the chemical reaction graph. Reversible chemical reaction graphs can also be applied in the opposite direction, by breaking, forming, and changing bonds in a subgraph of the chemical graph which is isomorphic to the product of the chemical reaction graph.

The *size* of an atom mapping M between two chemical graphs $G_1 = (V_1, E_1, \ell_1, \mu_1)$ and $G_2 = (V_2, E_2, \ell_2, \mu_2)$ is given by

$$\operatorname{size}(M) = \sum_{(v,w) \in E_1} \bigl|\mu_2\bigl(M(v), M(w)\bigr) - \mu_1(v, w)\bigr|.$$

Given two compatible chemical graphs $G_1 = (V_1, E_1, \ell_1, \mu_1)$ and $G_2 = (V_2, E_2, \ell_2 \mu_2)$, an *optimal* atom mapping between them is an atom mapping of minimal size, which always exists (but it needs not be unique). An optimal atom mapping models the classical principle of minimum structure change, by which a chemical reaction normally occurs through the redistribution of the minimum number of valence electrons, that is, the formation and breaking of the least number of covalent bonds (Temkin et al., 1996).

The *size* of a chemical reaction graph $R = (G_1, G_2, M)$ is simply the size of the corresponding atom mapping M.

3. Reconstructing metabolic pathways by bidirectional chemical search

Artificial chemistries (Dittrich et al., 2001) are computational models of chemical systems and, in particular, of biochemical systems such as metabolic pathways. An artificial chemistry consists of a set of *molecules*, a set of *reaction rules* that produce new molecules from already existing molecules, and the definition of the *dynamics* of the system, which specifies the application conditions of the rules, the preference in their application, etc. (Rosselló and Valiente, 2005b).

A metabolic pathway can be regarded as a coordinated sequence of biochemical reactions and is often described in symbolic terms, as a succession of transformations of one set of *substrate* molecules into another set of *product* molecules (Rosselló and Valiente, 2004). Substrate and product must be compatible chemical graphs for a pathway between them to exist (Rosselló and Valiente, 2004, 2005a, 2005b).

Metabolic pathways are often represented as directed hypergraphs, with substrate and product molecules as nodes and biochemical reactions as hyperarcs. Since a chemical graph can represent the disjoint union of a set of molecules, though, the equivalent representation of artificial chemistries and, in particular, metabolic pathways as directed graphs becomes more natural. An artificial chemistry defined by a set of chemical reaction graphs, is thus represented as a directed second-order graph with the chemical graphs that represent the sets of substrate and product molecules as vertices and applications of the chemical reaction graphs, including information on atom mapping, as arcs.

Unfortunately, the size of the artificial chemistry defined by a set M of chemical graphs and a set R of chemical reaction graphs is often exponential in the size of M and R, and thus artificial chemistries are known for very small instances only, involving a few dozens of molecules and biochemical reactions. Therefore, we consider in this paper the problem of obtaining a substantial portion of the artificial chemistry defined by a set of biochemical reactions while avoiding the complexity of reconstructing the whole artificial chemistry.

The constraints we impose on the reconstruction process are threefold:

(1) The *initial chemical graphs* represent all sets of at most m metabolites among those involved in the set R of reactions, for some fixed, but arbitrary, m (in examples and applications in this paper we shall always take $m = 2$).
(2) The reconstruction process is restricted to a fixed, but arbitrary, number k of derivation steps.
(3) The initial and final sets of metabolites of every metabolic pathway belong to the set of initial chemical graphs.

While the first two constraints (on the size of the initial chemical graphs and the lengths of the metabolic pathways under inspection) are motivated by complexity considerations alone, the third constraint allows for directing the search of new metabolic pathways *inside* the artificial chemistry. That is, instead of building the artificial chemistry by applying the biochemical reactions in every possible way to each of the initial chemical graphs, we perform a bidirectional search by constructing *forward* metabolic pathways of length at most k starting in initial chemical graphs and *backward* metabolic pathways of length at most k ending in initial chemical graphs, and then gluing them to obtain all metabolic pathways of length at most $2k$ *starting and ending* in initial chemical graphs.

Given a set R of biochemical reactions and a number k of derivation steps, the detailed procedure for reconstructing all metabolic pathways of length up to $2k$ using the metabolites and reactions in R and starting and ending in multi-molecules of at most m components, is the following:

- First, we extract the set M of all chemical graphs representing sets of at most m any metabolites appearing in substrates and products of the reactions in R. We call the elements of M the *initial chemical graphs*.
- Next, we identify all compatibility classes in M (maximal subsets of compatible initial chemical graphs). Biochemical reactions transform chemical graphs into compatible chemical graphs and, therefore, the origin and the end of a metabolic pathway will be compatible sets of metabolites. Thus, since we restrict ourselves to metabolic pathways starting and ending in initial chemical graphs, we can restrict ourselves to search for metabolic pathways starting and ending in each compatibility class of initial chemical graphs.
- Then each compatibility class C in M is considered as a set of potential substrates $C_F^{(0)}$ and a set of potential products $C_R^{(0)}$ for the reactions in R.
- For every $i = 1, \ldots, k$, the forward application of the reactions in R to the elements of $C_F^{(i-1)}$ produces a set of multi-molecules $C_F^{(i)}$, while the reverse application of these reactions to the molecules in $C_R^{(i-1)}$ produces a set of multi-molecules $C_R^{(i)}$.
- Any nonempty intersection of a set obtained by forward application and a set obtained by reverse application of reactions yields a new pathway between elements of C. To avoid repetitions, it is enough to check whether each $C_F^{(i)}$ intersects $C_R^{(i)}$ and $C_R^{(i-1)}$. More specifically:
  - For $i = 1$, the forward application of the reactions in R to the molecules in $C_F^{(0)}$ produces a set $C_F^{(1)}$ of new molecules, and the reverse application of the reactions in R to the molecules in $C_R^{(0)}$ produces a set $C_R^{(1)}$ of new molecules.

$$\begin{tikzcd} C_F^{(0)} \arrow[r] & C_F^{(1)} \arrow[r, equal] \arrow[rr, bend left, equal] & C_R^{(1)} & C_R^{(0)} \arrow[l] \end{tikzcd}$$

  Then
  - Every member of $C_F^{(1)} \cap C_R^{(0)}$ yields a new pathway $C_F^{(0)} \to C_F^{(1)} \cap C_R^{(0)}$ of length 1.
  - Every member of $C_F^{(1)} \cap C_R^{(1)}$ yields a new pathway $C_F^{(0)} \to C_F^{(1)} \cap C_R^{(1)} \to C_R^{(0)}$ of length 2.

- For $i = 2$, the forward application of the reactions in R to the molecules in $C_F^{(1)}$ produces a set $C_F^{(2)}$ of new molecules, and the reverse application of the reactions in R to the molecules in $C_R^{(1)}$ produces a set $C_R^{(2)}$ of new molecules.

$$\begin{tikzcd}
C_F^{(0)} \arrow[r] & C_F^{(1)} \arrow[r] & C_F^{(2)} \arrow[r, equal] \arrow[rr, bend left, equal] & C_R^{(2)} & C_R^{(1)} \arrow[l] & C_R^{(0)} \arrow[l]
\end{tikzcd}$$

  Then
  - Every member of $C_F^{(2)} \cap C_R^{(1)}$ yields a new pathway of length 3

$$C_F^{(0)} \to C_F^{(1)} \to C_F^{(2)} \cap C_R^{(1)} \to C_R^{(0)}.$$

  - Every member of $C_F^{(2)} \cap C_R^{(2)}$ yields a new pathway of length 4

$$C_F^{(0)} \to C_F^{(1)} \to C_F^{(2)} \cap C_R^{(2)} \to C_R^{(1)} \to C_R^{(0)}.$$

- And, recursively, the forward application of the reactions in R to the molecules in $I_F = C_F^{(i-1)}$ produces a set $C_F = C_F^{(i)}$ of new molecules, and the reverse application of the reactions in R to the molecules in $I_R = C_R^{(i-1)}$ produces a set $C_R = C_R^{(i)}$ of new molecules.

$$\begin{tikzcd}
C_F^{(0)} \arrow[r] & \cdots \arrow[r] & I_F \arrow[r] & C_F \arrow[r, equal] \arrow[rr, bend left, equal] & C_R & I_R \arrow[l] & \cdots \arrow[l] & C_R^{(0)} \arrow[l]
\end{tikzcd}$$

  Then
  - Every member of $C_F \cap I_R$ yields a new pathway of length $2i - 1$

$$C_F^{(0)} \to \cdots \to I_F \to C_F \cap I_R \to \cdots \to C_R^{(0)}.$$

  - Every member of $C_F \cap C_R$ yields a new pathway of length $2i$

$$C_F^{(0)} \to \cdots \to I_F \to C_F \cap C_R \to I_R \to \cdots \to C_R^{(0)}.$$

The following result shows that in this way we obtain all metabolic pathways of length at most $2k$ under constraints (1) and (3) above.

Lemma 1. *For every $i = 1, \ldots, k$, all metabolic pathways of length $2i - 1$ and $2i$ starting and ending in initial chemical graphs are obtained in the ith iterative step of the procedure explained above.*

Proof: If

$$m_0 \to m_1 \to \cdots \to m_i \to \cdots \to m_{2i-1}$$

is a pathway with m_0 and m_{2i-1} initial chemical graphs, then $m_j \in C_F^{(j)}$ for every $j = 0, \ldots, i$ and $m_{2i-1-l} \in C_R^{(l)}$ for every $l = 0, \ldots, i - 1$, and hence in particular, $m_i \in C_F^{(i)} \cap C_R^{(i-1)}$. Therefore, this path is obtained in the ith iterative step of the procedure explained above.

On the other hand, if

$$m_0 \to m_1 \to \cdots \to m_i \to \cdots \to m_{2i}$$

is a pathway with m_0 and m_{2i} initial chemical graphs, then $m_j \in C_F^{(j)}$ for every $j = 0, \ldots, i$ and $m_{2i-l} \in C_R^{(l)}$ for every $l = 0, \ldots, i$, and hence, in particular, $m_i \in C_F^{(i)} \cap C_R^{(i)}$. Therefore, this path is also obtained in the ith iterative step of that procedure. □

Example 1. Let a, b, c, d, e, f be metabolites such that b, d, e, f are compatible with each other, a is compatible with $b + b$ and c is compatible with $b + b + b$. Consider the toy artificial chemistry given by the following reactions (where only the first four reactions are reversible):

$$a + b \leftrightarrow c, \qquad a \leftrightarrow d + e, \qquad b + d \leftrightarrow b + e, \qquad b + b \leftrightarrow d + f,$$

$$c \to e + b + b, \qquad d + d \to a, \qquad a + f \to b + e + e$$

Let us look for metabolic pathways starting and ending with metabolites and pairs of metabolites $a, \ldots, f$ globally compatible with $b + b + b$. Then the set M of all initial chemical graphs can be identified with the set of monomials of total weight at most 2 over the alphabet $\{a, b, c, d, e, f\}$ and the class C of the initial chemical graphs compatible with bbb (we omit henceforth the $+$ sign for simplicity) is

$$C = \{c, ab, ad, ae, af\}.$$

So, we are looking for metabolic pathways starting and ending in elements of this set C. The intermediate multi-molecules of these pathways will belong to the set of all multi-molecules formed by metabolites a, b, c, d, e, f compatible with bbb: these are the multi-molecules in C plus any combination of three metabolites b, d, e, f.

Taking

$$C_F^{(0)} = C_R^{(0)} = C = \{c, ab, ad, ae, af\},$$

we obtain the following one step derivations:

$C_F^{(0)}$	$\to$	$C_F^{(1)}$
c	$\to$	(ab, bbe)
ab	$\to$	(c, bde)
ad	$\to$	dde
ae	$\to$	dee
af	$\to$	(def, bee)

$C_R^{(1)}$	$\to$	$C_R^{(0)}$
(def, ddf)	$\to$	af
(dde, dee)	$\to$	ae
(dde, ddd)	$\to$	ad
(c, bde, bdd)	$\to$	ab
ab	$\to$	c

Notice that some elements of $C_F^{(1)}$ and $C_R^{(1)}$ do no longer belong to M, as we warned. Then

$$C_F^{(1)} = \{c, ab, bbe, bde, bee, dde, dee, def\}$$

$$C_R^{(1)} = \{c, ab, bdd, bde, ddd, dde, ddf, dee, def\}$$

and hence

$$C_F^{(1)} \cap C_R^{(0)} = \{ab, c\}, \qquad C_F^{(1)} \cap C_R^{(1)} = \{ab, c, bde, dde, dee, def\}.$$

From these intersections, we deduce that all metabolic pathways of lengths 1 and 2 starting and ending in C are

$$c \to ab, \qquad ab \to c, \qquad c \to ab \to c, \qquad ab \to c \to ab, \qquad ab \to bde \to ab,$$

$$ad \to dde \to ad, \qquad ad \to dde \to ae, \qquad ae \to dee \to ae, \qquad af \to def \to af.$$

For $k = 2$, we obtain:

$C_F^{(0)}$	$\to$	$C_F^{(1)}$	$\to$	$C_F^{(2)}$	$C_R^{(2)}$	$\to$	$C_R^{(1)}$	$\to$	$C_R^{(0)}$
c	$\to$	(ab, bbe)	$\to$	$((c, bde), (bbd, def))$	$((af, bbe), bbd)$	$\to$	(def, ddf)	$\to$	af
ab	$\to$	(c, bde)	$\to$	$((ab, bbe), (ab, bdd, bee))$	(ad, ae)	$\to$	(dde, dee)	$\to$	ae
ad	$\to$	dde	$\to$	(ad, ae)	$(ad, \emptyset)$	$\to$	(dde, ddd)	$\to$	ad
ae	$\to$	dee	$\to$	ae	$(ab, (ab, bdd, bee), bde)$	$\to$	(c, bde, bdd)	$\to$	ab
af	$\to$	(def, bee)	$\to$	$((af, bbe), bde)$	(c, bde, bdd)	$\to$	ab	$\to$	c

Then

$$C_F^{(2)} = \{c, ab, ad, ae, af, bbd, bbe, bdd, bde, bee, def\}$$

$$C_R^{(2)} = \{c, ab, ad, ae, af, bbd, bbe, bdd, bde, bee\}$$

and hence

$$C_F^{(2)} \cap C_R^{(1)} = \{c, ab, bdd, bde, def\},$$

$$C_F^{(2)} \cap C_R^{(2)} = \{c, ab, ad, ae, af, bbd, bbe, bdd, bde, bee\}.$$

From these intersections, we deduce that all metabolic pathways of lengths 3 and 4 starting and ending in C are

$$c \to ab \to c \to ab, \qquad c \to ab \to bde \to ab,$$

$$ab \to c \to ab \to c, \qquad ab \to bde \to bdd \to ab,$$

$$af \to bee \to bde \to ab, \qquad c \to bbe \to def \to af,$$

$$ab \to bde \to ab \to c, \qquad c \to ab \to c \to ab \to c,$$

$$c \to bbe \to bbd \to ddf \to af, \qquad c \to ab \to bde \to ab \to c,$$

$$c \to ab \to bde \to bdd \to ab, \qquad ab \to c \to ab \to c \to ab,$$

$$ab \to c \to ab \to bde \to ab, \qquad ab \to c \to bbe \to def \to af,$$

$$ab \to bde \to ab \to c \to ab, \qquad ab \to bde \to ab \to bde \to ab,$$

$$ab \to bde \to bdd \to ab \to c, \qquad ab \to bde \to bdd \to bde \to ab,$$

$$ab \to bde \to bee \to bde \to ab, \qquad ad \to dde \to ad \to dde \to ad,$$

$$ad \to dde \to ad \to dde \to ae, \qquad ad \to dde \to ae \to dde \to ae,$$

$$ae \to dee \to ae \to dde \to ae, \qquad af \to def \to af \to def \to af,$$
$$af \to def \to bbe \to def \to af, \qquad af \to bee \to bde \to ab \to c.$$
$$af \to bee \to bde \to bdd \to ab,$$

As it can be seen in the previous example, the raw application of the procedure explained above generates all metabolic pathways of length up to $2k$ starting and ending in sets of at most m metabolites used by the reactions in R, but most of these metabolic pathways will be redundant, for instance because they are cyclic, or because they do not contain any new multi-molecule that has not appeared in shorter metabolic pathways. Therefore, several reconstruction problems may be addressed in this context. In this work, we consider only three of them:

(a) to produce all metabolic pathways of length up to $2k$
(b) to produce all shortest metabolic pathways of length up to $2k$
(c) to produce all minimal acyclic metabolic pathways of length up to $2k$

in all cases under restrictions (1) to (3) made explicit above.

Here, by a *shortest* metabolic pathway between metabolite sets I and F, we understand a metabolic pathway from I to F of shortest length among all metabolic pathways from I to F, and by a *minimal acyclic* metabolic pathway we understand a metabolic pathway that contain no directed cycles and no other, shorter metabolic pathways with intermediates in I or F. For instance, the shortest path derivation

$$ab \to c \to bbe \to def \to af$$

in Example 1 is acyclic but not minimal, because it contains the derivation $c \to bbe \to def \to af$, while the minimal acyclic derivation

$$c \to bbe \to bbd \to ddf \to af$$

is not shortest, because there is a shorter derivation $c \to bbe \to def \to af$ from c to af.

We give our reconstruction algorithms in full pseudocode next. Algorithm 1 one formalizes the procedure explained above.

The first three lines of this algorithm produce the different compatibility classes of initial chemical graphs. Then for each compatibility class C and for each $i = 1, \dots, k$:

- It receives the sets $I_F = C_F^{(i-1)}$ and $I_R = C_R^{(i-1)}$ of the results of all direct and reverse applications, respectively, of $i-1$ consecutive rules in R to multi-molecules in C (when $i = 1$, $C_F^{(0)} = C$ and $C_R^{(0)} = C$) and it produces the sets $N_F = C_F^{(i)}$ and $N_R = C_R^{(i)}$ of the results of all direct and reverse applications, respectively, of rules in R to multi-molecules in I_F and I_R, respectively. That is, the sets of the results of all direct and reverse applications, respectively, of i consecutive rules in R to multi-molecules in C.
- The lines starting with *output* call a procedure that outputs the list of all metabolic pathways of lengths $2i - 1$ and $2i$ obtained so far. When $i = 1$:
  - the first *output* line gives all length 1 pathways $m \to m_f^{(1)}$, with $m \in C$,
  - the second *output* line gives all length 2 pathways $m \to m_r^{(1)} \to m'$ with $m, m' \in C$.

  And when $i > 1$:

Algorithm 1. Given a set R of biochemical reactions and a number k of derivation steps, obtain the set of all metabolic pathways of length up to $2k$ using the metabolites and reactions in R starting and ending in sets of at most m metabolites among those involved in the reactions in R.

```
$M \leftarrow$ substrate and product metabolites of the reactions in $R$
$M \leftarrow \bigcup_{j=1}^{m} M^j$
$E \leftarrow M/_{\cong}$ where $m \cong m'$ if and only if $m$ and $m'$ are compatible
foreach $C \in E$ do
    $I_F \leftarrow I_R \leftarrow C$
    foreach $i \leftarrow 1$ to $k$ do
        $N_F \leftarrow \emptyset$
        foreach $m \in I_F$ do
            foreach $r \in R$ do
                foreach $n \leftarrow$ forward application of $r$ to $m$ do
                    $N_F \leftarrow N_F \cup \{n\}$
        $N_R \leftarrow \emptyset$
        foreach $m \in I_R$ do
            foreach $r \in R$ do
                foreach $n \leftarrow$ reverse application of $r$ to $m$ do
                    $N_R \leftarrow N_R \cup \{n\}$
        output $C \rightarrow \cdots \rightarrow I_F \rightarrow N_F \cap I_R \rightarrow \cdots \rightarrow C$
        output $C \rightarrow \cdots \rightarrow I_F \rightarrow N_F \cap N_R \rightarrow I_R \rightarrow \cdots \rightarrow C$
        $I_F \leftarrow N_F$
        $I_R \leftarrow N_R$
```

- The first *output* line gives all length $2i - 1$ pathways

$$m \rightarrow m_f^{(1)} \rightarrow \cdots \rightarrow m_f^{(i-1)} \rightarrow m_f^{(i)} = m_r^{(i-1)} \rightarrow m_r^{(i-2)} \rightarrow \cdots \rightarrow m_r^{(1)} \rightarrow m'$$

  with $m, m' \in C$.
- The second *output* line gives all length $2i$ pathways

$$m \rightarrow m_f^{(1)} \rightarrow \cdots \rightarrow m_f^{(i-1)} \rightarrow m_f^{(i)} = m_r^{(i)} \rightarrow m_r^{(i-1)} \rightarrow \cdots \rightarrow m_r^{(1)} \rightarrow m'$$

  with $m, m' \in C$.

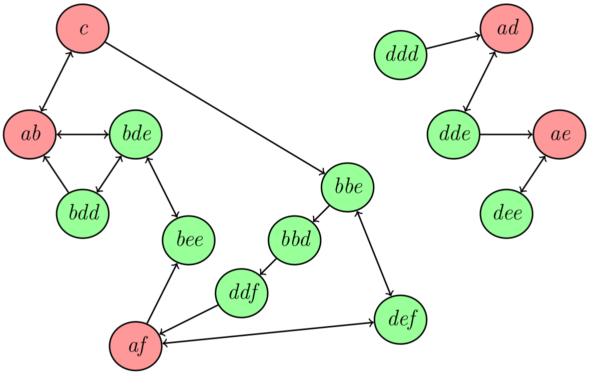
Algorithm 2 produces a metabolic network (X, Y) containing all metabolic pathways up to a given length, where the vertex set X contains the initial and final metabolite sets together with all those new metabolite sets produced by the forward and reverse application of the given biochemical reactions, and the arc set Y consists of all direct derivations thus obtained.

Now, upon the metabolic network (X, Y) obtained with the previous algorithm, the set of all shortest metabolic pathways of length up to $2k$, using the metabolites and reactions in R starting and ending in sets of at most m metabolites among those involved in the reactions in R, can be obtained by using an all-pairs shortest path algorithm (Dijkstra, 1959; Floyd, 1962; Johnson, 1977; Takaoka, 1998) upon each element of C as source vertex and each element of C as target vertex in turn.

Algorithm 2. Given a set *R* of biochemical reactions and a number *k* of derivation steps, obtain the metabolic network (X, Y) containing all metabolic pathways of length up to $2k$, using the metabolites and reactions in *R* starting and ending in sets of at most *m* metabolites among those involved in the reactions in *R*.

```
$M \leftarrow$ substrate and product metabolites of the reactions in $R$
$M \leftarrow \bigcup_{j=1}^{m} M^j$
$E \leftarrow M/_{\cong}$ where $m \cong m'$ if and only if $m$ and $m'$ are compatible
$X \leftarrow Y \leftarrow \emptyset$
foreach $C \in E$ do
    $I_F \leftarrow I_R \leftarrow C$
    foreach $i \leftarrow 1$ to $k$ do
        $N_F \leftarrow \emptyset$
        foreach $m \in I_F$ do
            foreach $r \in R$ do
                foreach $n \leftarrow$ forward application of $r$ to $m$ do
                    $N_F \leftarrow N_F \cup \{n\}$
                    $X \leftarrow X \cup \{m, n\}$
                    $Y \leftarrow Y \cup \{(m, n)\}$
        $N_R \leftarrow \emptyset$
        foreach $m \in I_R$ do
            foreach $r \in R$ do
                foreach $n \leftarrow$ reverse application of $r$ to $m$ do
                    $N_R \leftarrow N_R \cup \{n\}$
                    $X \leftarrow X \cup \{m, n\}$
                    $Y \leftarrow Y \cup \{(n, m)\}$
        $I_F \leftarrow N_F$
        $I_R \leftarrow N_R$
return $(X, Y)$
```

Example 2. The toy artificial chemistry of Example 1, obtained from the class $C = \{c, ab, ad, ae, af\}$ of the initial chemical graphs compatible with *bbb* by bidirectional search of metabolic pathways of length up to 4, is the following:



Then the enumeration of all-pairs shortest paths in (X,Y) starting and ending in the elements of $C = \{c, ab, ad, ae, af\}$ produces the following derivations:

$$
\begin{aligned}
&c \rightarrow ab,\\
&c \rightarrow bbe \rightarrow def \rightarrow af,\\
&ab \rightarrow c,\\
&ab \rightarrow c \rightarrow bbe \rightarrow def \rightarrow af,\\
&ad \rightarrow dde \rightarrow ae,\\
&af \rightarrow bee \rightarrow bde \rightarrow ab,\\
&af \rightarrow bee \rightarrow bde \rightarrow ab \rightarrow c.
\end{aligned}
$$

Algorithm 3 extracts the set of all minimal acyclic metabolic pathways of length up to $2k$, using the metabolites and reactions in R starting and ending in sets of at most m metabolites among those involved in the reactions in R, from the metabolic network (X,Y) produced by Algorithm 2.

In this algorithm, each path of the form $u \rightarrow \cdots \rightarrow v$ is extended in all possible ways by arcs in Y of the form $v \rightarrow w$ until reaching an element $w \in C$, where the test $w \notin p$ ensures the resulting paths are acyclic.

Algorithm 3. Given a metabolic network (X,Y) and a set C of initial and final metabolite sets, enumerate all minimal acyclic metabolic pathways contained in (X,Y) which start and end in metabolite sets from C.

```
foreach $v \in C$ do
    $p \leftarrow \{v\}$
    $acyclic(C, Y, v, p)$
```

where $acyclic(C, E, v, p)$ is defined as follows:

```
foreach $(v, w) \in E$ do
    if $w \notin p$ then
        if $w \in C$ then
            print $p \cup \{w\}$
        else
            $p \leftarrow p \cup \{w\}$
            $acyclic(C, E, w, p)$
            $p \leftarrow p \setminus \{w\}$
```

Example 3. In the metabolic network (X,Y) of Example 2, which corresponds to the toy artificial chemistry of Example 1, the enumeration of minimal acyclic paths starting and ending in the elements of $C = \{c, ab, ad, ae, af\}$ produces the following derivations:

$$c \rightarrow ab,$$

$$c \rightarrow bbe \rightarrow bbd \rightarrow ddf \rightarrow af,$$

$$c \rightarrow bbe \rightarrow def \rightarrow af,$$

$$ab \rightarrow c,$$

$$ad \rightarrow dde \rightarrow ae,$$

$$af \rightarrow bee \rightarrow bde \rightarrow ab,$$

$$af \rightarrow bee \rightarrow bde \rightarrow bdd \rightarrow ab.$$

Remark 1. Notice that the shortest path derivation $ab \rightarrow c \rightarrow bbe \rightarrow def \rightarrow af$ is not minimal, and the minimal acyclic derivation $c \rightarrow bbe \rightarrow bbd \rightarrow ddf \rightarrow af$ is not shortest.

4. Results and discussion

The size of an artificial chemistry is often exponential in the number of initial metabolites and biochemical reactions, and thus some method is needed for obtaining a significant portion of an artificial chemistry while avoiding the complexity of a complete reconstruction. The techniques we have introduced in this paper represent an important step in this direction, because they impose the only constraint on the reconstruction process that biochemical reactions be applied to combinations of at most m metabolites. Nevertheless, they allow for

(1) Obtaining all pathways of length up to $2k$ by bidirectional search,
(2) Storing them in a compact representation, and
(3) Extracting shortest pathways and minimal acyclic pathways from the compact representation,

where m and k are the only parameters of the reconstruction algorithms.

The metabolic reconstruction algorithm was implemented as a Perl script, using the `Chemistry::Reaction` module from the PerlMol collection of Perl modules for computational chemistry (Tubert-Brohman, 2004). The core of the methodology is embodied in the `Chemistry::Artificial` Perl module, which is available from the authors and will also be available from the PerlMol collection of Perl modules for computational chemistry (Tubert-Brohman, 2004). This module can be used to reconstruct the artificial chemistry defined by a given set of reaction equations written in reaction SMILES format (Weininger, 1988). For instance, the following Perl script first stores the artificial chemistry containing all derivations of length up to $2k = 4$ starting and ending in sets of at most $m = 2$ metabolites using the reaction equations in file `rctn.smi` (Algorithm 2) and then, extracts all shortest derivations and all minimal acyclic derivations (Algorithm 3).

```
use Chemistry::Artificial;
use strict;

my $m = 2;
my $k = 2;
my $c = Chemistry::Artificial->new("rctn.smi",$m,$k);
$c->bidirectional;
$c->shortest;
$c->minimal_acyclic;
```

We have performed a series of experiments in order to reconstruct metabolic pathways for all known reference pathway maps. The protocol we have used is as follows:

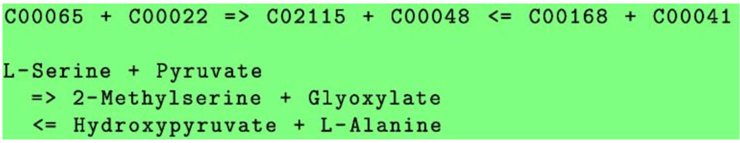
(1) Obtain reference pathway maps from the KEGG (Kanehisa et al., 2006) database. We have used KEGG release 42.0 in all our experiments.
(2) Solve the optimal atom mapping problem for all of the reactions in the reference pathways, using the optimal atom mapping by chemical substructure search algorithm and tool support (Félix and Valiente, 2007).
(3) Reconstruct metabolic pathways of length up to 8 for each reference pathway.
(4) Orient the reactions, according to the study of irreversibility of reactions in KEGG carried out in (Ma and Zeng, 2003).
(5) Filter out those metabolic pathways that involve irreversible reactions applied in the reverse direction.
(6) Identify the new metabolites thus obtained, by chemical structure search in CheBi (Brooksbank et al., 2005), MetaCyc (Caspi et al., 2006), KEGG (Kanehisa et al., 2006), and SciFinder Scholar (Wagner, 2006).
(7) Analyze the new metabolic pathways for coexistence of metabolites and enzymes in each particular organism.

Preliminary results obtained by following the aforementioned experimental protocol upon 13 of the 308 reference pathway maps in KEGG are summarized in Tables 1 and 2. For the reference pathway map β-Alanine metabolism (00410), for instance, during the bidirectional chemical search for $k = 1$, the number of new metabolites was $264 - 106 = 158$ and four new shortest pathways and also four new minimal acyclic metabolic pathways were obtained; for $k = 2$, the number of new metabolites was $293 - 158 = 135$ and two new minimal acyclic metabolic pathways were obtained; and for $k = 3$, the number of new metabolites was $316 - 293 = 23$, while no further new minimal acyclic pathway was found for $k = 3, 4$, and thus four new shortest pathways and six new minimal acyclic metabolic pathways were found while generating 7189 new metabolites.

The biological significance of these results can be assessed by examining the actual pathways found by bidirectional search, using the metabolites and reactions stored in KEGG for a particular reference pathway map. Besides obtaining again some of these reactions, an intermediate step is added in some metabolic pathways to one of the reactions stored in KEGG. For instance, using the metabolites and reactions stored in KEGG for glycine, serine, and threonine metabolism (reference pathway map 00260), we have obtained the following pathway:

Table 1 Number of vertices (n) and arcs (m) of the metabolic network containing all metabolic pathways of length up to $2k$ found by bidirectional chemical search upon the metabolites and reactions stored in KEGG for several reference maps (map), for $k = 1, 2, 3, 4$

map	$k = 0$	$k = 1$		$k = 2$		$k = 3$		$k = 4$	
	n	n	m	n	m	n	m	n	m
00010	529	870	690	931	854	931	854	931	854
00020	82	253	350	458	818	737	1712	785	1876
00030	314	1148	1678	2284	4788	2988	6770	3021	6836
00031	23	33	20	33	20	33	20	33	20
00040	330	707	756	870	1178	888	1214	915	1268
00051	702	913	422	943	488	943	488	943	488
00053	201	660	1108	1285	2982	1819	4618	2276	6046
00061	53	102	118	102	118	102	118	102	118
00062	290	2359	4188	5042	10884	5706	12212	6012	12824
00071	372	2550	4418	4977	10322	5314	10996	5314	10996
00072	8	8	0	8	0	8	0	8	0
00100	229	229	0	229	0	229	0	229	0
00120	292	1901	3254	3442	7680	3442	7680	3442	7680
00130	267	289	44	296	58	296	58	296	58
00150	290	290	0	290	0	290	0	290	0
00190	14	14	0	14	0	14	0	14	0
00220	238	399	326	437	422	439	426	439	426
00231	18	45	54	45	54	45	54	45	54
00251	24	44	44	52	60	52	60	52	60
00252	146	235	186	260	242	274	270	280	282
00260	604	841	482	915	632	929	676	929	676
00271	386	633	502	788	850	943	850	943	850
00272	95	110	36	111	38	111	38	111	38
00280	320	1206	1778	2595	5200	3129	6286	3134	6298
00290	161	350	390	350	390	350	390	350	390
00300	152	287	276	287	276	287	276	287	276
00310	188	380	394	381	396	381	396	381	396
00311	14	27	26	27	26	27	26	27	26
00330	289	376	180	383	194	383	194	383	194
00340	129	323	390	385	536	385	536	385	536
00360	157	244	178	246	182	246	182	246	182
00400	37	54	34	54	34	54	34	54	34
00410	106	264	320	293	382	316	428	316	428
00471	13	30	34	37	54	37	54	37	54
00590	870	3128	4672	5501	10278	7052	14456	7189	14824
00906	594	1181	1250	1345	1780	1357	1818	1357	1818



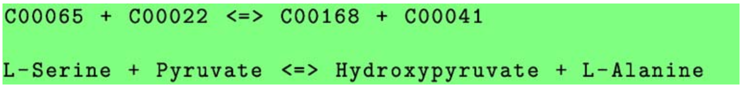
```
C00065 + C00022 => C02115 + C00048 <= C00168 + C00041

L-Serine + Pyruvate
  => 2-Methylserine + Glyoxylate
  <= Hydroxypyruvate + L-Alanine
```

While the methylation of L-Serine to 2-Methylserine and demethylation of Pyruvate to Glyoxylate followed by the methylation of Glyoxylate to L-Alanine and demethylation of 2-Methylserine to Hydroxypyruvate is chemically feasible, the Serine pyruvate aminotransferase enzyme (2.6.1.51) allows for the oxidative deamination of L-Serine into L-Alanine, as stated in KEGG reaction R00585:

Table 2 Number of shortest pathways (short) and the number of minimal acyclic pathways (min) of length up to $2k$ found by bidirectional chemical search upon the metabolites and reactions stored in KEGG for several reference maps (map), for $k = 1, 2, 3, 4$

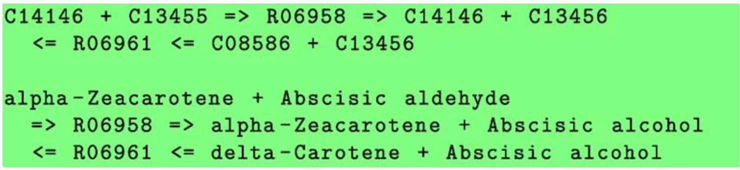
map	$k = 1$		$k = 2$		$k = 3$		$k = 4$	
	short	min	short	min	short	min	short	min
00010	8	8	8	8	8	8	8	8
00020	8	8	8	8	8	8	8	8
00030	6	10	6	44	6	326	6	1714
00040	2	2	2	2	2	2	2	2
00053	50	194	52	672	52	3250	52	17412
00061	20	20	20	20	20	20	20	20
00062	50	50	50	50	50	50	50	50
00071	62	62	62	62	62	62	62	62
00120	24	36	30	192	30	984	30	4716
00220	4	4	4	4	4	4	4	4
00251	2	4	2	4	2	4	2	4
00252	6	8	8	12	8	12	8	12
00260	8	8	8	8	8	8	8	8
00271	8	8	8	12	8	24	8	36
00272	6	6	6	6	6	6	6	6
00280	6	6	6	6	6	6	6	6
00290	12	12	12	12	12	12	12	12
00300	4	6	4	6	4	6	4	6
00310	10	10	10	10	10	10	10	10
00330	6	6	6	6	6	6	6	6
00340	2	2	2	2	2	2	2	2
00360	4	4	4	4	4	4	4	4
00410	4	4	4	6	4	6	4	6
00590	156	156	156	180	156	228	156	228
00906	76	76	76	78	76	78	76	78



```
C00065 + C00022 <=> C00168 + C00041

L-Serine + Pyruvate <=> Hydroxypyruvate + L-Alanine
```

Among the novel metabolic pathways found by bidirectional search, using the metabolites and reactions stored in KEGG for carotenoid biosynthesis (reference pathway map 00906), we have obtained the following metabolic pathway:



```
C14146 + C13455 => R06958 => C14146 + C13456
  <= R06961 <= C08586 + C13456

alpha-Zeacarotene + Abscisic aldehyde
  => R06958 => alpha-Zeacarotene + Abscisic alcohol
  <= R06961 <= delta-Carotene + Abscisic alcohol
```

A KEGG pathway reference map contains information for several organisms. Thus, it is important to find evidence that all four metabolites appearing in this pathway are present in a same organism, and also that the enzyme activating the reverse biochemical reaction R06961 (carotene 7,8-desaturase, 1.14.99.30) is indeed expressed in that particular organism.

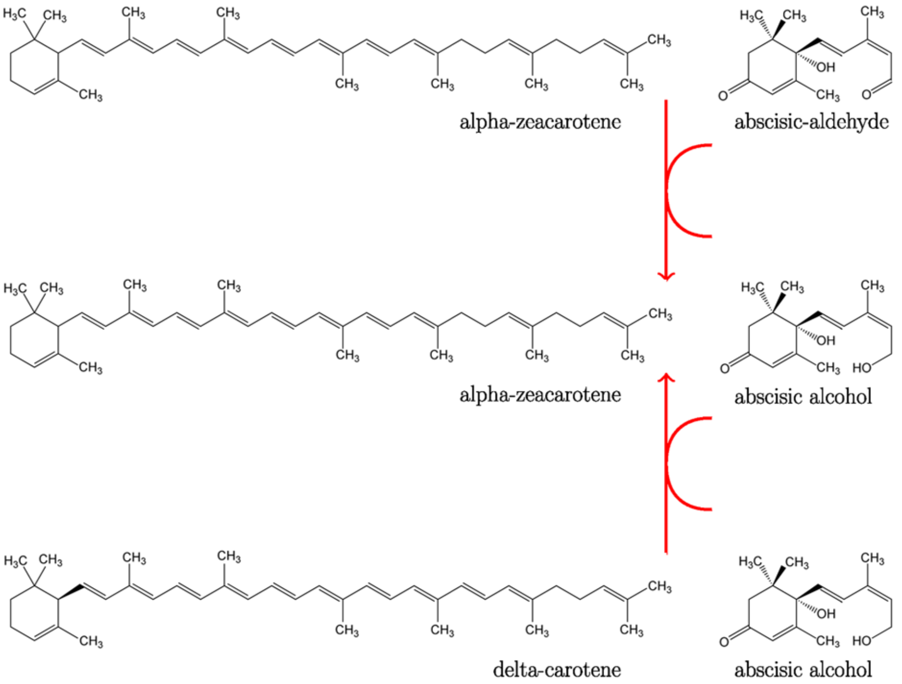


Fig. 2 A novel metabolic pathway found in the biosynthesis of steroids.

Carotenoid biosynthesis spans several related pathways: spheroidene, normal-spirilloxanthin, unusual-spirilloxanthin, abscisic acid biosynthesis, and astaxanthin biosynthesis. However, there are organisms whose metabolism does not include both carotenoid biosynthesis and abscisic acid biosynthesis. In fact, *Arabidopsis thaliana* (thale cress) is the only organism for which the four metabolites are annotated in KEGG to carotenoid biosynthesis, and the gene coding for carotene 7,8-desaturase, AT3G04870, is indeed expressed in *A. thaliana* (Bartley et al., 1999; Scolnik and Bartley, 1995).

On the other hand, there is a biosynthetic pathway, the plastidic 2C-methyl-D-erythritol 4-phosphate (MEP) pathway that involves the four metabolites and occurs in plastids, protozoa, most bacteria, and algae (Estévez et al., 2001). In the MEP pathway, carotenoid biosynthesis is a precursor of abscisic acid biosynthesis (Estévez et al., 2001, Fig. 1). In the novel metabolic pathway, alpha-Zeacarotene (C14146) and delta-Carotene (C08586) are involved in carotenoid biosynthesis whereas Abscisic aldehyde (C13455) and Abscisic alcohol (C13456) are involved in abscisic acid biosynthesis. Such a possible link between the early and later stages of the biosynthesis of steroids was established in (Estévez et al., 2001), where it is argued that only specific carotenoid intermediates (direct precursors of the abscisic acid biosynthesis) are increased or reduced, and further studied in (Seo and Koshiba, 2002) when regulating the early stages of abscisic acid biosynthesis in plants. The new metabolic pathway, shown in Fig. 2, is thus a novel pathway in the biosynthesis of carotenoid indeed.

Table 3 Number of potential biochemical reactions between sets of at most m metabolites among those involved in the reactions stored in KEGG for several reference maps. For each value of m, the first column gives the number of classes with two or more molecules (which indicates the possibility of a biochemical reaction among them) and the second column gives the total number of classes

map	$m = 1$		$m = 2$		$m = 3$		$m = 4$		$m = 5$	
00010	6	45	293	920	5552	11199	60502	94731	446942	601910
00020	1	37	97	665	2250	7068	24833	51429	170858	279264
00030	7	41	263	714	3696	6828	28860	42957	151233	198172
00031	2	17	40	160	365	955	2121	4254	9072	15268
00040	9	41	346	768	4986	7977	40053	53668	213052	258300
00051	11	34	301	482	2985	3794	17651	20210	75027	81554
00053	7	32	199	439	2086	3207	11631	14982	43952	51606
00061	0	45	113	814	2791	8237	28979	55956	183880	281947
00062	0	33	91	362	1118	2118	5855	8397	20396	25711
00071	1	60	271	1415	7218	18353	88066	156445	664502	965103
00072	0	14	7	112	99	560	642	2072	2675	6137
00100	14	70	825	2262	20756	40791	285566	454527	2439893	3381759
00120	10	50	396	1061	6685	12885	67353	106346	471120	654410
00130	5	46	246	1038	5395	13902	64431	120826	467334	719792
00150	9	41	294	744	4201	7645	35515	52648	205587	269069
00190	0	14	8	111	106	561	712	2146	3148	6706
00220	0	63	244	1771	8846	26869	125700	242959	921233	1387232
00231	0	22	12	263	248	2022	2564	11662	17577	53983
00251	3	48	209	1122	5512	15603	74831	143221	604301	911037
00252	5	51	301	1239	7186	17277	89453	154299	664930	941960
00260	5	84	624	3100	23630	63164	424027	790950	4252609	6407444
00271	2	69	336	2222	12816	41942	236574	518317	2567554	4365113
00272	2	44	138	944	3409	12221	45450	109489	389767	724507
00280	8	42	312	762	4409	7540	34765	49006	185510	233275
00290	8	37	274	645	3895	6601	33131	46469	196273	245176
00300	2	54	251	1318	6750	18193	85928	163830	670524	1047724
00310	4	69	393	2210	13161	40178	220249	464308	2165271	3607261
00311	1	38	62	749	1656	9189	23656	78677	209181	492812
00330	6	68	490	2088	14025	35627	203409	377970	1741156	2671199
00340	1	59	255	1551	7897	22521	104759	199413	763662	1153907
00360	7	51	364	1105	6230	12342	55100	85952	309536	418892
00400	2	53	191	1379	6345	20855	95031	198504	781143	1261828
00410	3	53	234	1329	6273	18727	81546	165437	610070	967461
00471	2	22	55	259	670	1939	5061	10810	27605	48249
00590	10	29	213	404	2178	3349	14417	19596	70620	88693
00906	18	67	786	1608	13100	20264	120642	159842	719545	869908

While these preliminary results already reveal a number of new biochemical pathways, the artificial chemistry reconstruction starting from all sets of at most m metabolites among those involved in the set of reactions (the third constraint imposed on the reconstruction process) might reveal the existence of a much larger number of new biochemical pathways for $m > 2$. As can be seen in Table 3, the number of potential biochemical reactions grows fast with m for the reference maps stored in KEGG.

Acknowledgements

The research described in this paper has been partially supported by the Spanish CICYT, project TIN 2004-07925-C03-01 GRAMMARS and project MTM2006-07773 COMGRIO, and by EU project INTAS IT 04-77-7178. A preliminary version of this paper has appeared in (Félix et al., 2007).

References

Bartley, G.E., Scolnik, P.A., Beyer, P., 1999. Two *Arabidopsis thaliana* carotene desaturases, phytoene desaturase and ζ-carotene desaturase, expressed in *Escherichia coli*, catalyze a poly-*cis* pathway to yield pro-lycopene. Eur. J. Biochem. 259(1–2), 396–403.

Benkö, G., Flamm, C., Stadler, P.F., 2003a. Generic properties of chemical networks: artificial chemistry based on graph rewriting. In: Proc. 7th European Conf. Advances in Artificial Life, Lect. Notes Comput. Sci., vol. 2801, pp. 10–19. Springer, Berlin.

Benkö, G., Flamm, C., Stadler, P.F., 2003b. A graph-based toy model of chemistry. J. Chem. Inf. Comput. Sci. 43(4), 1085–1093.

Benkö, G., Flamm, C., Stadler, P.F., 2004. Multi-phase artificial chemistry. In: Schaub, H., Detje, F., Brüggemann, U. (Eds.), The Logic of Artificial Life: Abstracting and Synthesizing the Principles of Living Systems, pp. 16–22. IOS Press, Amsterdam.

Brooksbank, C., Cameron, G., Thornton, J., 2005. The European Bioinformatics Institute's data resources: towards systems biology. Nucleic Acids Res. 33(D), D46–D53.

Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y., Tissier, C., Zhang, P., Karp, P.D., 2006. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. 34(D), D511–D516.

Deville, Y., Gilbert, D., van Helden, J., Wodak, S.J., 2003. An overview of data models for the analysis of biochemical pathways. Brief. Bioinform. 4(3), 246–259.

Dijkstra, E.W., 1959. A note on two problems in connexion with graphs. Numer. Math. 1(1), 269–271.

Dittrich, P., Ziegler, J., Banzhaff, W., 2001. Artificial chemistries—a review. Artif. Life 7(1), 225–275.

Edwards, J.S., Palsson, B.O., 2000. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. P. Natl. Acad. Sci. USA 97(10), 5528–5533.

Estévez, J.M., Cantero, A., Reindl, A., Reichler, S., León, P., 2001. 1-deoxy-D-xylulose-5-phosphate synthase, a limiting enzyme for plastidic isoprenoid biosynthesis in plants. J. Biol. Chem. 276(25), 22901–22909.

Félix, L., Rosselló, F., Valiente, G., 2007. Reconstructing metabolic pathways by bidirectional chemical search. In: Proc. 5th Int. Conf. Computational Methods in Systems Biology, Lect. Notes Bioinformatics, vol. 4695, pp. 217–232. Springer, Berlin.

Félix, L., Valiente, G., 2007. Validation of metabolic pathway databases based on chemical substructure search. Biomol. Eng. 24(3), 327–335.

Floyd, R.W., 1962. Algorithm 97: Shortest path. Commun. ACM 5(6), 345.

Johnson, D.B., 1977. Efficient algorithms for shortest paths in sparse networks. J. ACM 24(1), 1–13.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28(1), 27–30.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M., 2006. From genomics to chemical genomics: New developments in KEGG. Nucleic Acids Res. 34(D), D354–D357.

Karp, P.D., Mavrovouniotis, M.L., 1994. Representing, analyzing, and synthesizing biochemical pathways. IEEE Expert 9(2), 11–21.

Lemer, C., Antezana, E., Couche, F., Fays, F., Santolaria, X., Janky, R., Deville, Y., Richelle, J., Wodak, S.J., 2004. The aMAZE LightBench: a web interface to a relational database of cellular processes. Nucleic Acids Res. 32(D), 443–448.

Ma, H., Zeng, A.-P., 2003. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 19(2), 270–277.

McCaskill, J., Niemann, U., 2001. Graph replacement chemistry for DNA processing. In: DNA 2000, Lect. Notes Comput. Sci., vol. 2054, pp. 103–116. Springer, Berlin.

Michal, G. (Ed.), 1999. Biological Pathways: An Atlas of Biochemistry and Molecular Biology. Wiley, New York.

Overbeek, R., Larsen, N., Pusch, G.D., D'Souza, M., Selkov, E., Kyrpides, N., Fonstein, M., Maltsev, N., Selkov, E., 2000. WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res. 28(1), 123–125.

Rosselló, F., Valiente, G., 2004. Analysis of metabolic pathways by graph transformation. In: Proc. 2nd Int. Conf. Graph Transformation, Lect. Notes Comput. Sci., vol. 3256, pp. 70–82. Springer, Berlin.

Rosselló, F., Valiente, G., 2005a. Chemical graphs, chemical reaction graphs, and chemical graph transformation. Electron. Notes Theor. Comput. Sci. 127(1), 157–166.

Rosselló, F., Valiente, G., 2005b. Graph transformation in molecular biology. In: Formal Methods in Software and System Modeling, Lect. Notes Comput. Sci., vol. 3393, pp. 116–133. Springer, Berlin.

Schomburg, I., Chang, A., Schomburg, D., 2002. BRENDA, enzyme data and metabolic information. Nucleic Acids Res. 30(1), 47–49.

Scolnik, P.A., Bartley, G.E., 1995. Nucleotide sequence of zeta-carotene desaturase (accession no. U38550) from arabidopsis. Plant Physiol. 109(4), 1499.

Seo, M., Koshiba, T., 2002. Complex regulation of ABA biosynthesis in plants. Trends Plant Sci. 7(1), 41–48.

Takaoka, T., 1998. Subcubic cost algorithms for the all pairs shortest path problem. Algorithmica 20(3), 309–318.

Temkin, O.N., Zeigarnik, A.V., Bonchev, D., 1996. Chemical Reaction Networks: A Graph-Theoretical Approach. CRC Press, Boca Raton.

Tubert-Brohman, I., 2004. Perl and chemistry. Perl J. 8(6), 3–5. PerlMol is available at http://www.perlmol.org/.

Wagner, A.B., 2006. Scifinder scholar 2006: An empirical analysis of research topic query processing. J. Chem. Inf. Model. 46(2), 767–774.

Weininger, D., 1988. SMILES, a chemical language and information system, 1: introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28(1), 31–36. http://www.daylight.com/dayhtml/doc/theory/.