Abstract

Metabolism can be regarded as a network of biochemical reactions, connected via their substrates and products. A metabolic pathway is thus a coordinated series of reactions, and it is often described in symbolic terms, as a succession of transformations of a set of substrate molecules into a set of product molecules. In this paper, a set of molecules is described by the disjoint union of the chemical graphs representing them, chemical reactions are described by chemical reaction graphs, and metabolic pathways are computationally modeled by artificial chemistries. An artificial chemistry is represented as a directed graph with the chemical graphs that represent the sets of substrate and product molecules as nodes and applications of the chemical reaction graphs as arcs. We present simple algorithms for building optimal artificial chemistries, which transform a substrate chemical graph to a product chemical graph using a given set of chemical reaction graphs such that the least possible total number of chemical bonds are created or broken. These algorithms are based on a classification of chemical reactions in metabolic pathways as either decomposition, pseudo-exchange, displacement, or isomerase reactions.

1. Introduction

Metabolism is one of the most complex cellular processes, cells function as organized chemical engines carrying out a large number of transformations, in a suited behavior, called bioreactions or biochemical reactions.
analysis of biocatalysis and biodegradation processes, and consistency checking of pathway databases [6].

The organization of the rest of the paper is as follows. In Section 2, we establish the notation used along this work, and previous work supporting this paper. In Section 3, we define the way to form the set of molecules to work. In Section 4, we construct the artificial chemistry this paper deals with. Finally, in Section 5, we establish some directions of future work.

All figures in this paper were obtained with the DEPICT algorithm and tool support [17]. In particular, Figure 1 was obtained with a tool that produces a pair of GIF files containing a diagram of the reaction, one for the substrate and the other one for the product, with appropriate atom coloring (omitted in this paper) to illustrate the correspondence among substrate and product atoms, furthermore, so that molecules can be represented as strings using the SMILES [16, 18] language, we are including the respective string in this language with each figure.

2 Preliminaries

2.1. Chemical Graphs

The notation for graphs used here is stated in [13] and is as follows. A **chemical graph** is a weighted graph \((V, E, \mu)\), with \((V, E)\) an undirected graph (without multiple edges or self-loops) all whose nodes are labeled by means of chemical elements, and \(\mu : E \to \mathbb{N}\) a weight function. A weight of 0 stands for a non-existing 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. The label of a node \(v \in V\) will be denoted by \(\ell(v)\).

A **chemical reaction graph** is a structure \((V, E, \sigma, \pi)\), with \((V, E, \sigma)\) and \((V, E, \pi)\) chemical graphs, called the **substrate** and the **product** chemical graphs respectively, satisfying the following conditions:

- There is no \(e \in E\) such that \(\sigma(e) = \pi(e) = 0\).
- For every \(v \in V\), if \(e_1, \ldots, e_k\) are the edges incident to it, then
  \[
  \sigma(e_1) + \cdots + \sigma(e_k) = \pi(e_1) + \cdots + \pi(e_k) \geq 1.
  \]

The size of a chemical reaction graph is the total number of bonds that are broken by the reaction, which must coincide with the total number of bonds that are created by the reaction:

\[
\sum_{e \in E} |\sigma(e) - \pi(e)|.
\]

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.

A set of molecules is consequently described by the disjoint union of the chemical graphs representing them, whereas chemical graphs themselves need not be connected.

2.2. Artificial Pathways

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 [12]. Substrate and product must be compatible chemical graphs for a pathway between them to exist [5].

Two chemical graphs \(G_1 = (V_1, E_1, \mu_1)\) and \(G_2 = (V_2, E_2, \mu_2)\) are said to be **compatible** if

- \(|V_1| = |V_2|\),
- there exists a bijection \(\phi : V_1 \to V_2\) such that \(\ell_1(v_1) = \ell_2(\phi(v_1))\) for all \(v_1 \in V_1\), and
- \(\sum_{e_1 \in E_1} \mu_1(e_1) = \sum_{e_2 \in E_2} \mu_2(e_2)\).

In such a case, \(\phi : V_1 \to V_2\) is said to be a compatible bijection.

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 graph with the chemical graphs that represent the sets of substrate and product molecules as nodes and applications of the chemical reaction graphs as arcs.

2.3. Optimal Artificial Pathways

A set of biochemical reactions applying the optimality criterium proposed in [5], which can be stated as the least possible total number of broken and created bonds through the various biochemical reactions, is defined as follows:

- Given two compatible chemical graphs \(G_1 = (V_1, E_1, \mu_1)\) and \(G_2 = (V_2, E_2, \mu_2)\), together with a compatible bijection \(\phi : V_1 \to V_2\), the difference between \(G_1\) and \(G_2\) over \(\phi\) is
  \[
  G_1 \ominus \phi G_2 = \sum_{e_1 \in E_1} |\mu_1(e_1) - \mu_2(\phi(e_1))| = \sum_{e_2 \in E_2} |\mu_2(e_2) - \mu_1(\phi^{-1}(e_2))|,
  \]
  where the latter equality holds because \(\phi\) is a compatible bijection.
• The least difference between two compatible chemical graphs \(G_1 = (V_1, E_1, \mu_1)\) and \(G_2 = (V_2, E_2, \mu_2)\), denoted by \(G_1 \ominus G_2\), is the minimum difference over all bijections between their vertex sets:

\[ G_1 \ominus G_2 = \min_{\phi:V_1 \rightarrow V_2} G_1 \ominus_\phi G_2. \]

• Let \(\mathcal{G}\) be a set of chemical graphs, and let \(\mathcal{R}\) be a set of chemical reaction graphs over \(\mathcal{G}\). An artificial pathway from a substrate chemical graph \(S \in \mathcal{G}\) to a product chemical graph \(P \in \mathcal{G}\), is a nonempty sequence \(R_1, \ldots, R_n\) of chemical reaction graphs, with \(R_i \in \mathcal{R}\) for \(1 \leq i \leq n\), such that

- \(T_0 = S\),
- \(R_i\) can be applied to \(T_{i-1}\) and results in a chemical graph \(T_i\), for \(1 \leq i \leq n\), and
- \(T_n = P\).

That is,

\[ S = T_0 \xrightarrow{R_1} T_1 \xrightarrow{R_2} T_2 \xrightarrow{R_3} \ldots \xrightarrow{R_n} T_n = P \]

In such a case, let \(\phi_i\) be the compatible bijection from the nodes of \(T_{i-1}\) to the nodes of \(T_i\) that results from the application of \(R_i\) to \(T_{i-1}\). An artificial pathway from \(S\) to \(P\) is said to be optimal if the total size of the chemical reaction graphs in the pathway is equal to \(S \ominus P\), that is, if

\[ \sum_{1 \leq i \leq n} T_{i-1} \ominus_{\phi_i} T_i = S \ominus P. \]

• For a given set \(\mathcal{G}\) of chemical graphs and a given set \(\mathcal{R}\) of chemical reaction graphs, an artificial chemistry is a directed graph, possibly with loops and multiple arcs, with chemical graphs (including those in \(\mathcal{G}\)) as nodes and applications of the chemical reaction graphs in \(\mathcal{R}\) as arcs.

3. Optimal Artificial Chemistries

Artificial chemistries are computational models of chemical systems and, in particular, of biochemical systems such as metabolic pathways. However, these pathways can range in size from involving a few enzymes and metabolites (a collection of substrates and products) to the complete pathway of an organism that can have thousands of them. Therefore, artificial chemistries are known for very small instances only, involving a few dozens of molecules and biochemical reactions [5]. Besides, not all the pathways in an artificial chemistry are biologically meaningful.
These schemes have proven to be very useful in practice because in general, the automatic mapping problem is NP-hard, even in a restricted form [1]. Previous work on this problem was either centered on heuristic maximum common subgraph algorithms [2, 9] or focused on particular cases [1]. Furthermore, as shown in [6], these groups represent two of the most frequent enzymatic reactions existing in the KEGG LIGAND [8] reaction database.

Secondly, once the least difference between a substrate and a product is known, the second stage of the process consists of actually constructing the optimal artificial chemistry, that is, an artificial chemistry in which the transformation of a substrate chemical graph $S$ into a product chemical graph $P$ using a given set of chemical reaction graphs, achieves the optimal value for $S \ominus P$.

Since the size of the artificial chemistry can be exponen-
tial in the size of $G$ and $R$, we restrict the size of $R$ by ap-
plying only well-known reactions, such as the Diels-Alder reaction [7], one of the most important reactions in organic chemistry (see Figure 1), or by applying only a given set of reactions that are known to belong in the metabolic pathway under study. To carry out this application, we are using the PerlMol implementation of chemical graph transformation [13], where each chemical reaction graph $R \in R$ is created from the substrate and product chemical graphs $S$ and $P$ of the reaction and the optimal mapping between them, $S \ominus P$, obtained by search over the space of all matchings, as explained above. Figure 2 shows the molecule for the chemical reaction graph $R$ corresponding to the Diels-Alder reaction.

In our approach, we are considering two different ways to construct an artificial chemistry, that is, two ways to col-
lect the chemical graphs of the new molecules, adding them to the set of chemical graphs $G$. The first construction con-
ists in collecting the new molecules in a given number $k$ of steps, and the second construction consists in the exhaustive collection of the new molecules, that is, to collect them until no new molecule is obtained. We are considering the first construction because of the resulting high cost of obtaining the whole meaningful artificial chemistry, therefore, we re-
strict the number of derivation steps to a given number $k$ in order to get only a slice of the whole artificial chemistry.

Furthermore, in order to ensure the creation of chemical graphs representing the new molecules that result from the application of a chemical reaction graph $R \in R$ to a chemi-
cal graph $G \subseteq G$, we need to include in $G$, at each derivation step, the chemical graphs representing the disjoint union of pairs, triples, etc. In this fashion, we create the chemical graphs of the new molecules in a unique way. While we are getting new molecules in this way, we continue with the ad-
dition of such chemical graphs at the next derivation step. However, as soon as no $R \in R$ applied to the chemical graphs $G \subseteq G$ gives a new molecule, the process must stop. Again, in order to restrict the size of the resulting artificial chemistry, we consider only pairs and triples of chemical graphs as candidate substrate or product molecules, because most biochemical reactions in metabolic pathways involve no more than two or three substrate compounds and two or three product compounds.

Figure 3 shows a hypothetical example for a metabolic pathway containing only two molecules. Figure 4 show some of the new molecules created from the application by breaking, forming and changing bonds in a subgraph of each chemical graph $G \subseteq G$ which is isomorphic to the sub-
strate of each chemical reaction graph $R \in R$ through the first step, with $k = 1$, while Figure 5 shows some of the new molecules created until the second step, with $k = 2$.

Notice that, since substrate and product are sets of molecules, and these are described by the disjoint union of
the chemical graphs representing them, our approach to the construction of an artificial pathway can be applied to any metabolic pathway database and, in particular, to the KEGG LIGAND [8] reaction database.

4. Constructing Optimal Artificial Chemistries

Recall that an artificial chemistry is a directed graph with chemical graphs $G \subseteq \mathcal{G}$ as nodes and applications of the chemical reaction graphs $R \subseteq \mathcal{R}$ as arcs. Then, for a given set $\mathcal{G}$ of chemical graphs and a given set $\mathcal{R}$ of chemical reaction graphs, let $AC = (V, A)$ be the graph representing the artificial chemistry, with $V \subseteq \mathcal{P}(\mathcal{G})$ as vertex set and $A \subseteq \mathcal{R}$ as arc set.

In order to construct such a graph, let $M$ be the set of chemical graphs representing the original molecules involved in the studied pathway, let $\mathcal{G}$ be the set of chemical graphs representing molecules, and let $\mathcal{R}$ be the set of chemical reaction graphs.

We follow the approach explained in Section 3. The first construction consists in the collection of the new molecules in a given number $k$ of steps. Thus, we construct the artificial chemistry graph $AC$ as follows: we apply each chemical reaction graph $R \in \mathcal{R}$ to each chemical graph $G \subseteq \mathcal{G}$ (including the chemical graphs representing the disjoint union of pairs and triples of chemical graphs). This application consists of breaking, forming, and changing bonds in each subgraph of the chemical graph $G$ which is isomorphic to the substrate of the chemical reaction graph $R$, and gives the chemical graph $N$ representing a new substrate. If $N$ is not contained in $\mathcal{G}$, we add $N$ to the set of chemical graphs $\mathcal{G}$. Besides, we need to compute the least difference between $G$ and $N$, $G \ominus N$, in order to establish the optimal path. To do so, we first need the bijection $\phi$. We thus assume there is a reaction $G \iff N$ in $\mathcal{R}$, where $G$ represents the substrate and $N$ represents the product of the reaction.

Now, if $G$ and $N$ are two compatible chemical graphs or the disjoint union of chemical graphs, we separate $G$ and $N$ into their connected components, the chemical graphs $gg_{G} \in G$ and $gn_{N} \in N$, such that

\[
\bigcap_{j=1}^{\ell_{G}} gg_{j} = \emptyset \quad \text{and} \quad \bigcup_{j=1}^{\ell_{G}} gg_{j} = G
\]

and

\[
\bigcap_{j=1}^{\ell_{N}} gn_{j} = \emptyset \quad \text{and} \quad \bigcup_{j=1}^{\ell_{N}} gn_{j} = N .
\]

This separation allows us to deal with the substrate chemical graphs $gg \in G$ and the product chemical graphs $gn \in N$ using the schemes explained in section 3:

- Decomposition reactions,
  \[G \iff gn_{1} + gn_{2} .\]
Pseudo-exchange and displacement reactions,
\[ gg_1 + gg_2 \Leftrightarrow gn_1 + gn_2 \]

Isomerase reactions,
\[ G \Leftrightarrow N \]

These schemes lead us to calculate the bijection \( \phi \) and, therefore, the least difference,
\[ G \ominus N = \min_{\phi: V_1 \rightarrow V_2} G \ominus_\phi N. \]

The detailed algorithm for computing the optimal bijection between two chemical graphs, is presented in pseudocode form in Fig. 6. Only the schemes for one substrate and two product compounds (decomposition reactions) and two substrate and two product compounds (pseudo-exchange and displacement reactions) are shown. Other schemes can be found in [6].

Finally, we add the chemical graphs \( G \) and \( N \) to the vertex set \( V \) of the artificial chemistry graph \( AC \), and the chemical reaction graph \( R \) (actually, the triplet formed by the substrate chemical graph \( G \), the product chemical graph \( N \), and the chemical reaction graph \( R \)) to the arc set \( A \) of the artificial chemistry graph. Every time the whole process ends, we get a derivation of \( i \) steps, for \( i = 1, \ldots, k \).

This algorithm returns the graph \( AC \) for the artificial chemistry and updates the set of chemical graphs \( G \) representing the original and new molecules of the studied pathway, both results for a derivation of \( k \) steps. The detailed algorithm for constructing a slice of an optimal artificial chemistry, is presented in pseudocode form in Fig. 7.

The second construction consists in the exhaustive collection of the new molecules, that is, to collect the new molecules until there is no new one. We thus construct the artificial chemistry graph \( AC \) in a similar way to the previous one, except that, we ignore the number of steps required for the whole artificial chemistry. Then, we control the existence of new molecules, that is, meanwhile new molecules are created the steps of derivation continue, at the moment there does not exist any new molecule, the process finishes, returning the artificial chemistry graph \( AC \) and with the set of chemical graphs \( G \) containing the original and the new molecules created during the derivation process.
The detailed algorithm for the exhaustive construction of an optimal artificial chemistry, is presented in pseudocode form in Fig. 8.

Let $n$ be the number of chemical graphs in $G$, let $m$ be the number of chemical reaction graphs in $R$, let $n'$ be the size of the largest chemical graph in $G \times G \times G$, and let $f(n')$ be the cost of subgraph isomorphism over the chemical graphs, pairs of chemical graphs, and triples of chemical graphs in $G$. The algorithm for building a slice of $k$ derivation steps of the optimal artificial chemistry for $G$ and $R$, takes $O(kmn^3f(n'))$ time.

5. Conclusions

The classification of chemical reactions in metabolic pathways as either decomposition, pseudo-exchange, displacement, or isomerase reactions, allows for the computation of slices of optimal artificial chemistries, which transform a substrate chemical graph to a product chemical graph using a given set of chemical reaction graphs such that the least possible total number of chemical bonds are created or broken.

We are actually implementing this framework on top of the PerlMol collection of Perl modules for computational chemistry [15]. This implementation strongly relies on pattern matching of the substrate of a chemical reaction graph and a larger chemical graph (molecule), but the pattern matching algorithm currently available in PerlMol is essentially an exhaustive enumeration algorithm. Since metabolic pathways contain macro-molecules, that is, chemical graphs with hundreds of nodes, the extension of PerlMol with a more efficient pattern matching algorithm is an interesting line of future research.

We are also interested in providing appropriate support for the visualization of artificial chemistries, in order for biologists and biochemists to be able to analyze them. A first such visualization tool has already been developed [11] and will be freely available for download from the PerlMol collection of Perl modules for computational chemistry [15], module Chemistry::Artificial::Graphics.

6. Acknowledgement

The research described in this paper was partially supported by the Spanish CICYT, project GRAMMARS (TIN 2004-07925-C03-01), and by the Japan Society for the Promotion of Science through Long-term Invitation Fellowship L05511 for visiting JAIST (Japan Advanced Institute of Science and Technology).
References


