Motif Search in Graphs:

Application to Metabolic Networks

Vincent Lacroix [1,2,◇], Cristina G. Fernandes [3], Marie-France Sagot [1,2,4]

[1] Équipe BAOBAB, Laboratoire de Biométrie et Biologie Evolutive (UMR 5558);CNRS; Univ. Lyon 1, 43 bd 11 nov, 69622, Villeurbanne Cedex, France.

[2] Projet Helix, INRIA Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot Saint-Martin, France

[3] Instituto de Matemática e Estatística, Universidade de São Paulo,Rua do Matão, 1010, Cidade Universitária, CEP 05508-090, São Paulo, Brazil

[4] Department of Computer Science, King's College London, Strand, London WC2R 2LS, England

[◇] Corresponding author (lacroix@biomserv.univ-lyon1.fr)

Abstract

The classic view of metabolism as a collection of metabolic pathways is being questioned with the currently available possibility of studying whole networks. Novel ways of decomposing the network into modules and motifs that could be considered as the building blocks of a network are being suggested. In this work, we introduce a new definition of motif in the context of metabolic networks. Unlike in previous works on (other) biochemical networks, this definition is not based only on topological features. We propose instead to use an alternative definition based on the functional nature of the components that form the motif, which we call a *reaction motif*. After introducing a formal framework motivated by biological considerations, we present complexity results on the problem of searching for all occurrences of a reaction motif in a network, and introduce an algorithm that is fast in practice in most situations. We then show an initial application to the study of pathway evolution. Finally, we give some general features of the observed number of occurrences in order to highlight some structural features of metabolic networks.

1 Introduction

Network biology is a general term for an emerging field that concerns the study of interactions between biological elements [1]. The term *molecular interaction networks* may designate several types of networks depending on the kind of molecules involved. Classically, one distinguishes between gene regulatory networks, signal transduction networks and metabolic networks. Protein-protein interaction networks represent yet another type of network, but this term is rather linked to the techniques (such as Yeast-2-hybrid) used to produce the data and covers possibly several biological processes (including, for example, the formation of complexes and phosphorylation cascades) [15].

One of the declared objectives of network biology (or systems biology in general) is whole cell simulation [8]. However, studying the dynamics of a network requires knowledge on reaction mechanisms such as the kinetic parameters describing a Michaelis-Menten equation. Besides the fact that such knowledge is often unavailable or unreliable, the study of the static set of reactions that constitute a biochemical network is equally important, both as a first step towards introducing dynamics, and in itself. Indeed, such static set represents not what is happening at a given time in a given cell but instead the capabilities of the cell, including capabilities the cell does not use. A careful analysis of this set of reactions for a given organism, alone or in comparison with the set of other organisms, may also help to arrive at a better understanding of how metabolism evolves. It is this set we propose to study in this paper. More precisely, in the following sections, the term "metabolism" should be understood as the static set of reactions involved in the synthesis and degradation of small molecules. Regulation information is not taken into consideration for now. It may be added in a later step, as the "software" running on the "hardware" of a metabolic network [14].

A major issue concerning the study of biochemical networks is the problem of their organisation. Several attempts have been made to decompose complex networks into parts. These "parts" have been called modules or motifs, but no definition of such terms seems to be completely satisfying.

Modules have first been mentioned by Hartwell *et al.* [5] who outline the general features a module should have but provide no clear definition for it. In the context of metabolic networks, a natural definition of modules could be based on the decomposition of a metabolic network into the metabolic pathways one can find in databases: modules would thus be the pathways as those have been established. The advantage of this definition of module is that it reflects the way metabolism has been discovered experimentally (starting from key metabolites and studying the ability of an organism to synthesize or degrade them). The drawback is that it is not based on objective criteria and therefore is not universal (indeed, the number of metabolic pathways and the frontiers between them vary from one database to the other).

Several attempts to give systematic and practical definitions have been made using graph formalisms [4, 9, 13] and constraint-based approaches [10]. Graph-based methods range from a simple study of the local connectivity of metabolites in the network [13] to the maximisation of a criterion expressing modularity (number of links within modules) [4]. The only information used in these methods is the topology of the network. In the case of constraint-based approaches, the idea is quite different. First, a decomposition of the network into functional sets of reactions is performed (by analysis of the stoichiometric matrix [11]) and then modules are defined from the analysis of these functional states. The result is not a partition in the sense that a single reaction might belong to several modules.

Unlike the definition of module, the notion of motif has not been studied in the context of metabolic networks. In general, depending on what definition is adopted for modules and motifs, there is no clear limit between the two notions besides the difference in size. In the context of regulatory networks, motifs have been defined as small, repeated and perhaps evolutionary conserved subnetworks. In contrast with modules, motifs do not function in isolation. Furthermore, they may be nested and overlapping [22]. This definition refers to general features that regulatory motifs are believed to share but it provides no practical way to find them. A more practical definition has been proposed, still in the context of gene regulatory networks (and other types of non-biological networks such as the web or social networks). These are "network motifs" and represent patterns of interconnections that recur in many different parts of a network at frequencies much higher than those found in randomized networks [16]. This definition is purely topological and disregards the nature of the components in a motif. It assumes that the local topology of the network is sufficient to model function (which is understood here as the dynamic behaviour of the motif). This assumption seems acceptable when studying the topology of the internet and may also hold when analysing gene regulatory networks, but it appears not adapted to metabolic networks. In a static context, a topological definition of motif seems indeed inappropriate as similar topologies can give rise to very different functions.

In the definition of motif we introduce, the components of the network play the central part and the topology can be added as a further constraint only. This is the main biological contribution of this paper.

Its main algorithmical contribution comes from the fact that the definition of motif we adopt leads to new questions. Indeed, if searching for “purely” topological motifs may be formally modelled as a subgraph isomorphism problem, this no longer applies when searching for motifs where the features describing the components are the important elements and topology is initially indifferent (connectivity only is taken into account). Besides, the problem we address is also different from pathway alignment because we wish to go beyond the notion of pathway in order to study the network as a whole therefore leading to noticeable differences in the modelling decisions. Indeed, in [18] and [12], the pathways are modelled as, respectively, chains and trees to simplify the problem. This simplification may seem reasonable when modelling pathways but it is no longer so in the case of general networks.

The paper addresses complexity issues related to this new definition of a graph motif, providing hardness results on the problem, and then presents an exact algorithm that is fast in practice for searching for such motifs in networks representing the whole metabolism of an organism. The paper ends with two applications of the algorithm: one concerning the formulation of hypotheses on the evolution of pathways, the other being related to the study of some structural features of metabolic networks.

This paper is an extended version of a previous paper presented at WABI in 2005.

2 Preliminaries

2.1 Data

The metabolic network analysed in this work was obtained from the PATHWAY database from KEGG [7]. Data describing reactions, compounds and enzymes were downloaded and stored locally using a relational database management system (postgreSQL). The KEGG database contains metabolic data concerning 209 sequenced organisms. The network we built from such data is therefore a consensus of our current knowledge on the metabolisms of all those organisms. As a consequence, sequences of reactions present in the network may have been observed in no organism. To avoid this configuration, one can "filter" the consensus network by an organism of interest, keeping only in the dataset reactions catalysed by enzymes the organism is considered to be able to synthesize. We adopt a different strategy by choosing to perform our motif search on the consensus network and to possibly filter the results in a second step, allowing for easier comparative analysis among organisms.

Moreover, we use an additional information present in KEGG: the notion of primary/secondary metabolites. Indeed, in the KEGG reference pathway diagrams (maps), only primary metabolites are represented and connect reactions together, whereas secondary metabolites are not drawn (even though they participate in the reaction). A typical example of a secondary metabolite is the ATP molecule in an ATP-consuming reaction. (Observe that, unlike the notion of ubiquitous compound [13], the notion of primary/secondary metabolite is relative to a reaction.) Keeping all metabolites in the network leads to the creation of artefactual links between reactions and the bias introduced can lead to inaccurate results such as considering metabolic networks as small-world networks as shown in [2]. Withdrawing secondary metabolites may not be the best strategy to adopt, but it represents a simple way of avoiding this bias.

2.2 Graph Models

Several formal models have been used to study metabolic networks. The choice of a formal model seems to depend mainly on the nature of the hypotheses one wishes to test (qualitative or quantitative, static or dynamic) and on the size of the network under study. Differential equations seem well adapted to study the dynamic aspects of very small networks whereas graphs enable the static study of very large networks.

Between these two ends of the spectrum, semi-quantitative models have been proposed. For example, Petri nets allow for the simulation and dynamical analysis of small networks [20], while constraint-based models provide a mathematical framework enabling to decompose the network into functional states starting only from information on stoichiometry and making the assumption that the network is at steady-state [11].

As our goal is to deal with large networks and work with the least possible *a priori*, graph models seem appropriate. In previous genome-scale studies [6], graphs have been used mainly for topological analyses regardless of the nature of their components (reactions, compounds and enzymes). We propose to enrich the graph models and take into consideration some of the features of such components.

Formally, a graph G is defined as a pair (V, E), with V a set of *vertices* and $E \subseteq V \times V$ a set of *edges*. The edges represent the relations between the vertices and may be directed or undirected. The vertices and edges of the graph can be labelled.

The most intuitive graph representation of a metabolic network is provided by a bipartite graph. A bipartite graph has two types of vertices which in the context of metabolic networks represent, respectively, reactions and chemical compounds, and edges only between these compounds. The compound graph is a compact version of the bipartite graph where only compound vertices are kept and information on the reactions is stored as edge labels. The reaction graph is the symmetric representation of a compound graph (*i.e.*, reaction vertices are kept and information on the compounds is stored as edge labels). Directed versions of these graphs can be drawn expressing the irreversibility of some reactions. The information concerning the reversibility of reactions is generally not well-known. Indeed, contradictions may be found within a same database. We therefore consider this information as uncertain and, in an initial step, assume that all reactions are reversible. This apparently strong hypothesis seems preferable than considering a reaction as irreversible when it actually is reversible (leading to a loss of information).

For practical reasons, the algorithm developed uses the reaction graph as model for the metabolic network. In the following sections, the graph $G(V, E)$ we consider will be undirected and labelled at the vertices. Vertices of G may be labelled by more than one element from a finite set C whose elements we refer as colours. In the context of metabolic networks, these colours correspond to reaction types. Details will be given in Section 2.5.

2.3 Motif Definition

We define a motif using the nature of the components it contains.

Definition of a motif

A motif *is a multiset of elements from the set C of colours.*

As mentioned earlier, we choose in this definition not to introduce any constraint on the order of the reactions nor on topology. This choice is motivated by the wish to explore the network with the least possible *a priori* information on what we are searching for. Topology and order of the reactions can be used later as further constraints. The advantage of this strategy is that the impact of each additional constraint can then be measured.

2.4 Occurrence Definition

Intuitively, an occurrence is a connected set of vertices labelled by the colours of the motif. For a precise definition, let R be a set of vertices of G and let M be a motif of the same size as R. Let $H(R, M)$ denote the bipartite graph whose set of vertices is $R \cup M$ and where there is an edge between a vertex v of R and a vertex c of M if and only if v has c as one of its colours.

Definition of an exact occurrence of a motif

An exact occurrence *of a motif M is a set R of vertices of G such that $H(R, M)$ has a perfect matching and R induces a connected subgraph of G.*

If one searches for exact occurrences of a motif, the risk is to find a single occurrence, or none [2]. Moreover, since studying the evolution of what the graph G

represents is one of our main objectives, it seems relevant to allow for flexibility in the search.

With this in mind, we introduce a function S (detailed later) that assigns, to each pair c_i, c_j in $C \times C$, a score which measures the similarity between c_i and c_j. Two colours are considered similar if this score is superior to a threshold s. We then adapt our definition of exact occurrence by modifying $H(R, M)$ in the following way. There will be an edge between a vertex v in R and a colour c in M if and only if there exists a colour c' of v such that the value of $S(c', c) \geq s$. Further, we generalise this to the case where the threshold s is different for every element c in M. The latter is motivated by the idea that some elements in the motif we are searching for may be more crucial than others.

Another type of flexibility can then be added, that allows for gaps in the occurrences. By this we mean, roughly, allowing the occurrence to have more vertices just to achieve the connectivity requirement. These extra vertices are not matched to the elements of the motif. Two types of control on the number of gaps are considered: local and global. Intuitively, a local gap control policy bounds the maximum number of consecutive gaps allowed between a pair of matched vertices of R. A global control policy bounds the total number of gaps in an occurrence.

This leads to the following definition of an approximate occurrence of a motif, where we denote by G_R the subgraph of G induced by a set R of vertices of G.

Definition of an approximate occurrence of a motif

Let lb and gb be the local and global gap control bounds and let M be a motif. For each c in M, let s_c be a number. An approximate *occurrence of M (with respect to lb, gb and the thresholds s_c) is a set R of vertices which is contained in a set R' of vertices of G that satisfies the following conditions:*

1. *the bipartite graph $H(M \cup R, E_H)$ with $E_H = \{\{c, v\} \in M \times R \mid$ there exists a colour c' of v such that $S(c', c) \geq s_c\}$ contains a perfect matching;*

2. *for each subset B of R such that $B \neq \emptyset$ and $R \setminus B \neq \emptyset$, the length of a shortest path in $G_{R'}$ between B and $R \setminus B$ is at most lb;*

3. *$|R'| - |R| \leq gb$.*

Observe that when no gaps are allowed then $R = R'$ and condition 2 means simply that G_R is connected. An example is given in Figure 1.

2.5 Colours and Similarity

We now discuss the choice of a set C and a function S for the problem of metabolic networks and reaction motifs in such networks. A major obstacle in this task is that no systematic classification of reactions is available. Instead, enzyme classifications can be used, but a problem one has to face is that several enzymes can catalyse the same reaction (and a single enzyme can catalyse several reactions).

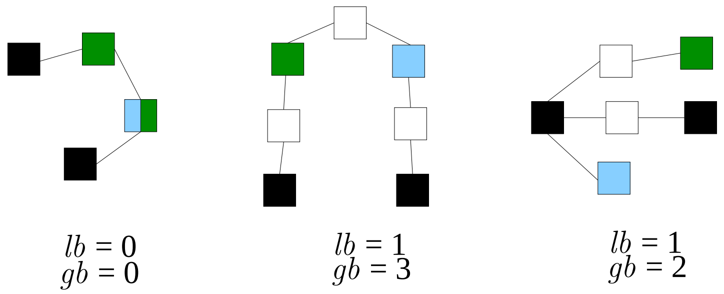

Figure 1: Subgraphs induced by occurrences for the motif {black, black, dark grey, light grey}.

This observation explains the necessity to allow for multiple labelling of nodes in the reaction graph.

We will describe here two possible ways of defining C and S. The first one is based on alignment. Indeed, in order to compare reactions, which is what function S is used for, one can compare the enzymes that catalyse these reactions by performing an alignment of their sequences. An element of C would then be a protein sequence. The function S assigns a sequence alignment score and s is a user-defined threshold that has to be met to consider the sequences similar. The same method can be applied using protein structures instead of sequences but in the case of whole networks, sequences are preferable since many structures are not known.

The second example is the one we adopt in this paper. It is based on a hierarchical classification of enzymes developed by the International Union of Biochemistry and Molecular Biology (IUBMB) [21]. It consists in assigning to each enzyme a code with 4 numbers expressing the chemistry of the reaction it catalyses. This code is known as the enzyme's EC number (for Enzyme Commission Number). The first number of the EC number can take values in $[1\,..\,6]$, each number symbolising the 6 broad classes of enzymatic activity. (1. Oxidoreductase, 2. Transferase, 3. Hydrolase, 4. Lyase, 5. Isomerase, 6. Ligase.) Then each of the three remaining numbers of the EC number provides additional levels of detail. For example, the EC number 1.1.1.1 refers to an oxidoreductase (1) with CH-OH as donor group and NAD+ as acceptor group.

An element of C is in this case an EC number. The function S then assigns a similarity score between two EC numbers that corresponds to the index of the deepest level down to which they remain identical. For example, $S(1.1.1.2, 1.1.1.3) = 3$. Two EC numbers are considered similar if their similarity score is above a user-defined cut-off value s in $[0\,..\,4]$. The advantage of this definition of similarity between colours, *i.e.*, reaction labels, is that it is more directly linked to the notion of function. Reactions compared with this measure are likely to be functionally related (and possibly evolutionarily related also).

3 Algorithmics

3.1 Hardness Results

The formal problem we address is the following:

Search Problem. Given a motif M and a labelled undirected graph G, find all occurrences of M in G.

As mentioned earlier, this problem is different from subgraph isomorphism because the topology is not specified for the motif.

In the following section, we may assume the graph is connected and all vertices have colours that appear in the motif. Otherwise, we preprocess the graph withdrawing all the vertices having no colour appearing in the motif and solve the problem in each component of the resulting graph.

This assumption only holds when no gaps are allowed. Some indications on how to deal with gaps will be given in Section 3.2.2, but the complexity analysis will be done on the exact case.

A natural variant of the Search Problem consists in, given a motif and a labelled graph, deciding whether the motif occurs in the graph or not. As before, we may assume the graph is connected, all vertices are labelled with colours and all colours appear in the motif. It is easy to see that the decision version of the Search Problem is in NP. We show next that it is NP-complete even if G is a tree, which implies that the Search Problem is NP-hard for trees.

3.1.1 NP-Complete for Trees.

We have the following proposition.

Proposition 1. *The decision version of the Search Problem is NP-complete even if G is a tree.*

Proof. We present a reduction from EXACT COVER BY 3-SETS (X3C):

INSTANCE: Set X with $|X| = 3q$ and a collection $\mathcal{C}$ of 3-element subsets of X.

QUESTION: Does $\mathcal{C}$ contain an exact cover for X, *i.e.*, a subcollection $\mathcal{C}' \subseteq \mathcal{C}$ such that every element of X occurs in exactly one member of $\mathcal{C}'$?

Let $X = \{1, \ldots, 3q\}$ and $\mathcal{C} = \{C_1, \ldots, C_n\}$ be an instance of X3C. The instance for the decision version of the Search Problem consists of a motif $M = \{Y, B, \ldots, B, 1, \ldots, 3q\}$, where B appears q times in M, and a tree T as follows. (See Figure 2 for an example.) There are four vertices in T for each i, $1 \leq i \leq n$, three of them are leaves in T, each one labelled by one of the elements of C_i. The fourth vertex, named r_i, is adjacent to the three leaves and has colour B. Besides these $4n$ vertices, there is only one more vertex in T, which is labelled Y and is adjacent to each r_i. This completes the description of the instance. Clearly it has size polynomial in the size of X and $\mathcal{C}$.

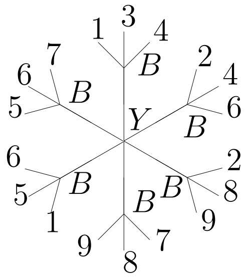

Figure 2: Tree T and its labels for $X = \{1, \ldots, 9\}$ and $\mathcal{C} = \{\{1,3,4\}, \{2,4,6\}, \{2,8,9\}, \{7,8,9\}, \{1,5,6\}, \{5,6,7\}\}$. For this example, $M = \{Y, B, B, B, 1, \ldots, 9\}$.

To complete the reduction, we need to argue that the motif M occurs in T if and only if there is a subcollection $\mathcal{C}'$ of $\mathcal{C}$ such that each element of X occurs exactly in one member of $\mathcal{C}'$.

Suppose there is such a $\mathcal{C}'$. Clearly $|\mathcal{C}'| = q$. Let R be the set of vertices of T consisting of the vertex labelled Y and the four vertices of each C in $\mathcal{C}'$. The subgraph of T induced by R is connected. Also, in R, there is a vertex labelled Y, q vertices labelled B (one for each C in $\mathcal{C}'$) and one labelled by each element in X (because of the property of $\mathcal{C}'$). That is, R is an occurrence of M in T.

Now, suppose there is an occurrence of M in T, that is, there is a set R of $1 + 4q$ vertices of T that induces a connected subgraph of T and has a vertex labelled by each of the colours in M. Let $\mathcal{C}'$ consist of the sets C_i in $\mathcal{C}$ whose vertex r_i in T is in R. Let us prove that each element of X appears in exactly one of the sets in $\mathcal{C}'$. First, note that the vertex labelled Y is necessarily in R, because it is the only one labelled Y and there is a Y in M. Then, as R induces a connected graph, a leaf from a set C_i is in R if and only if r_i is also in R. But R must contain exactly q vertices labelled B. Consequently, $|\mathcal{C}'| = q$ and, as R must contain $1 + 4q$ vertices, all three leaves of each C in $\mathcal{C}'$ must be in R, and these are all vertices in R. As R must contain a vertex labelled after each element in X, there must be exactly one set in $\mathcal{C}'$ containing each element in X. $\Box$

3.1.2 Fixed Parameter Tractability.

Finding all occurrences of a motif M in an undirected labelled tree T (Search Problem in trees) is fixed-parameter tractable with parameter the size of the motif (denoted by k). Indeed, a naive fixed-parameter algorithm consists in generating all possible topologies for the input motif M, and then searching for each topology by using a subtree isomorphism algorithm. Since it is enough to generate all possible tree topologies for M, the number of topologies to consider depends (exponentially) on k only, and subtree isomorphism is polynomial in the size of both the motif M and the tree T where M is sought. This reasoning is not valid anymore when the motif must be searched in a general graph G as subgraph isomorphism is NP-complete even when the motif is a tree [3].

3.1.3 General Complexity Results.

In this paper, we define a motif as a multiset of colours. Other works [16] have defined a motif as a subgraph. In both cases, deciding if a motif occurs in a graph is NP-complete. Nevertheless, using a combination of constraints on the topology and the colours of the motif may, in some cases, lead to polynomial algorithms. In this section, we compare the complexity of several variants of those problems in order to determine which constraint is responsible for NP-completeness.

In the following discussion, we shall first concentrate on the general case where the network is a graph and then comment on the cases where it is a tree.

If we only consider the topology as a constraint (TOPOLOGICAL MOTIFS), all considered variant of the problem are NP-complete. Indeed, subgraph isomorphism is NP-complete even if the motif is a path. This can be shown by a reduction from Hamiltonian Path. The reduction is based on an instance in which the motif is a path of size n, with n being the number of vertices of the graph in which the motif is being sought.

If we consider both the topology and the colours as constraints (TOPOLOGICAL COLOURED MOTIFS), the problem is NP-complete in all cases (reduction from Hamiltonian Path using the same technique as before), except when the colours are fixed (each node has a colour assigned to it) and no repetition of colours is allowed. This last statement is still a conjecture for which we have strong evidence in the case where the topology is a path or a tree, and partial results for the case of a general graph. The proof is not trivial and will be the subject of a forthcoming paper.

In the case where the network is a tree, every variant of the problem is polynomial (subtree isomorphism can be solved in polynomial time) except the variants where the topology is not known (Coloured Motifs). This result holds even when we do not allow for repetitions of the colours. The latter can be shown using the same reduction as in Section 3.1.1. The only difference is in the building of the instance of the Search Problem. In this case, the motif is defined as $M = \{Y, B_1, B_2, \ldots, B_q, 1, \ldots, 3q\}$, and the tree is the same as in Section 3.1.1 except that each of the n subtrees of four nodes connected to the central node (colour Y) appears now in q copies. The copies vary only in the colour of their root node which has colour B_j, $1 \leq j \leq q$.

Table 1: Complexity results for the motif Search Problem.

Input Graph Motif		Tree	Arbitrary
Topological motifs		polynomial	NP-complete
Coloured topological motifs	General Case	polynomial	NP-complete
	fixed colours and no repetition	polynomial	polynomial (conjecture)
Coloured motifs (this paper)		NP-complete, FPT in k	NP-complete

Table 1 summarises the complexity of the decision version of the Search Problem for the variants mentioned above. Reaction motifs, which are the subject of this paper, fall into the larger category of COLOURED MOTIFS, which also include the case where no colour is repeated.

In practice, metabolic networks are graphs and not trees. But fortunately they are relatively small (3184 vertices and 17642 edges for the network built from the KEGG Pathway database). Even though our initial problem is NP-complete, it is nevertheless conceivable to choose to solve it exactly, provided some efficient pruning is applied. This is described in the next section.

3.2 Exact Algorithm

3.2.1 Version with no Gaps.

We now present an exact algorithm which solves the Search Problem. We first explain it for the simple case where the gap parameters lb and gb are set to 0 and then we show how it can be extended to the general case.

Let M be the motif we seek. A very naive algorithm would consist in systematically testing all sets R of k vertices as candidates for being an occurrence, where $k = |M|$. For R to be considered an occurrence of M, the subgraph induced by R must be connected and there must be a perfect matching in the bipartite graph $H(R, M)$ that has an edge between $r \in R$ and $c \in M$ if and only if c is similar to one of the colours at vertex r. The search space of all combinations of k vertices among the n vertices in G is huge. We therefore show two major pruning ideas arising from the two conditions that R has to fulfill to be validated as an occurrence of M.

The connectivity condition can be checked by using a standard method for graph traversal, such as breadth first search (BFS). In our case, a BFS mixed with a backtracking strategy is performed starting from each vertex in the graph. At each step of the search, a subset of the vertices in the BFS queue is marked as part of the candidate set R. The queue, at each step, contains only marked vertices and neighbours in G of marked vertices. Also, there is a pointer p to the last considered vertex in the queue.

At each step, there are two cases to be analysed: either there are k vertices marked or not. If there are k vertices marked, we have a candidate set R at hand. We submit R to the test of the colouring condition, described below, and we backtrack to find the next candidate set. If there are less than k vertices marked, then there are two possible cases to be analysed: either p is pointing to the last vertex in the queue or not. If p is not pointing to the last vertex in the queue, we move p one position ahead in the queue, mark the next vertex and queue its neighbours that are not yet in the queue (checking the latter can be done in constant time by adding a flag to each vertex in the original graph). Then we repeat, that is, start a new step. If, on the other hand, p is pointing to the last vertex in the queue, then we move p to the last marked vertex (if no such vertex exists, the search is finished) and backtrack. The backtracking consists of (1) unmarking the vertex pointed to by p, (2) unqueueing its neighbours that were added when it was marked, and (3) starting a new step. Next we describe the test of the colouring condition.

Given a candidate set R, one can verify the colouring condition by building the graph H and checking whether it has a perfect matching or not. In fact, we can apply a variation of this checking to a partial set R, that is, we can, while constructing a candidate set R, check whether the corresponding graph H has or not a complete matching. The latter is a matching that completely covers the partial candidate set R. If there is no such matching, we can move the search ahead to the next candidate set. This verification can be done in constant time using additional data structures that are a constant times the size of the motif.

Extra optimisations can also be added. For instance, instead of using every vertex as a seed for the BFS, we can use only a subset of the vertices: those coloured by one of the colours from the motif, preferably the less frequent in the graph.

3.2.2 Allowing for Gaps.

Allowing for local but not global gaps (*i.e.*, setting $lb > 0$ and $gb = \infty$) can easily be done by performing the $lb-$transitive closure of the initial graph G and applying the same algorithm as before to the graph with augmented edge set. The

$p-$transitive closure of a graph G for a positive integer p is the graph obtained from G by adding an edge between any two vertices u and v such that the length l of a shortest path from u to v in the original graph satisfies $1 < l \leq p$. The $p-$transitive closure can be done at the beginning of the algorithm or on the fly. In the latter case, when a next vertex is added to the queue, instead of queueing its neighbours only, all vertices at distance at most $p + 1$ from it are queued (if they are not already in the queue) where by distance between any two vertices we mean the number of edges in a shortest path between them.

If we allow for global gaps only (that is, setting $lb = \infty$ and $gb > 0$), the problem can be reformulated as the one of finding a minimum Steiner tree for a set of vertices of a graph. The set of vertices would be any set R of vertices of G that match the colours in the motif, that is, that verify the colouring condition. If a minimum Steiner tree for R has size no more than $|R| + gb - 1$, then any such Steiner tree for R is a minimal occurrence of the motif (the minimality being in relation to the total number of gaps in the occurrence). The Steiner tree problem in general graphs is NP-hard, but there are several exact solver for it.

Here again, a less naive strategy than testing all possible sets R can be adopted. This strategy can follow the lines of the graph traversal with backtracking method described for the case of exact matches. The situation in this case is however more tricky. The reason is the following. Each vertex v may be considered for incorporation into the queue because it is at a distance d at most gb of different marked vertices. The distance d depends of course on the marked vertex. The number of gaps v may add to a candidate set R for a motif thus depends on which marked vertices responsible for the addition of v to the queue are already in R. We must therefore keep track of this information as we add and subtract vertices from the queue (and from R). This information can be kept using a balanced tree of size proportional to $k = |M|$ associated with each queued unmarked vertex v. Each node in the tree corresponds to a marked vertex u that is responsible for the queueing of v and is labelled by the distance from v to u (this distance is at most gb). Keeping, updating and using the extra information adds a multiplicative term in $O(k \log k)$ to the time complexity of the algorithm, which seems reasonable.

With this system, it now becomes straightforward to take into account both a local and a global gap parameter (that is, to have $lb = \infty$ and $gb = \infty$). In this case indeed, one needs only to further check that the distance above between v (the unmarked vertex being added to the queue) and u (the marked vertex that led to the queueing of v) is at most $\min\{lb, gb\}$. Another way of proceeding, would consist in (1) finding motifs that verify the local gap parameter and, (2) solving a minimum Steiner tree problem in G with the sets R given by the solutions to step 1.

3.2.3 Time Results.

On average, searching for all occurrences of a motif of size 4 with no gaps and threshold $s = 3$ takes 8 microseconds of CPU time on a Pentium 4 (CPU 1.70 GHz) with 512 Mb of memory.

4 Application

The approach we propose, and have described in the previous sections, should enable both to generate some hypotheses on the evolution of metabolic pathways, and to analyse some global features of the whole network.

We start by presenting a case study motivated by trying to understand how metabolic pathways evolve. We do not directly answer this question, which is complex and would be out of the scope of this paper. Instead, we give a first example of the type of evolutionary question people have been asking already and have addressed in different, often semi-manual ways in the past [19], and that the algorithm we propose in this paper might help treat in a more systematic fashion.

As in [19], one is often interested in a specific pathway, and, for instance, in finding whether this pathway can be considered similar to other pathways in the whole metabolic network thus suggesting a common evolutionary history. The metabolic pathway we chose as example is valine biosynthesis. Focusing on the last five steps of the pathway, we derived a motif

$M = \{1.1.1.86, 1.1.1.86, 4.2.1.9, 2.6.1.42, 6.1.1.9\}$ and performed the search for this motif using initially a cut-off value s of 4 for the similarity score between two EC numbers (that is, between two reaction labels). With this cut-off value, the motif was found to occur only once (see Figure 3.)

From this strictly defined motif, we relaxed constraints by first lowering the cut-off value s from 4 to 3 and then setting the gap parameters to 1 (motif denoted by M'). Additional occurrences were found. Three of them particularly drew our attention (see Figure 3.)

The first one corresponds to the five last steps of the isoleucine biosynthesis. The second one corresponds to the five last steps of the leucine biosynthesis. Together, they suggest a common evolutionary history for the biosynthesis pathways of valine, leucine and isoleucine.

An interesting point concerning the second occurrence is the fact that the order of the reactions is not the same as in the other pathways. This occurrence would not have been found if we had used a definition of motif where the order was specified.

Finally, the third occurrence that drew our attention was formed by reactions from both the biosynthesis of valine and a distinct metabolic pathway, namely the biosynthesis of Panthotenate and CoA. This latter case illustrates a limit of our current general way of thinking about metabolism: frontiers between metabolic pathways as defined in databases are not tight. If we had taken such frontiers

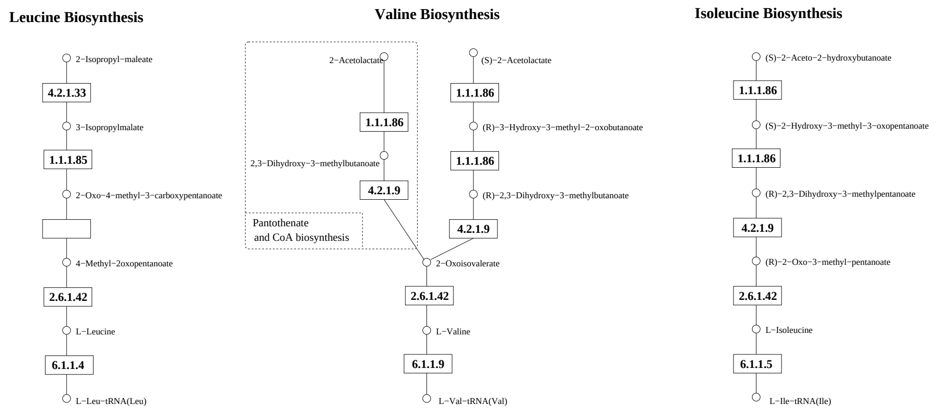

Figure 3: Bipartite representation of the results obtained when searching for the following motif: $M' = \{1.1.1, 4.2.1, 2.6.1.42, 6.1.1\}$ with local and global gap bounds set to 1. The empty box in the leucine biosynthesis represents a spontaneous reaction.

into account, we would not have found this occurrence that overlaps two different pathways. Yet such occurrence can be given a biological meaning: it can be seen as a putative alternative path for the biosynthesis of valine.

To complement this analysis, one should add that the results presented in this section hold for 125 organisms in KEGG among which *S. cerevisiae* and *E. coli*.

4.1 Preliminary results on some global features of the network

We now report a few preliminary results concerning some global features of the network. These results are based on the analysis of the number of occurrences found for all possible motifs of fixed size k (we considered $k = 3, 4$) and fixed cut-off value s (we considered $s = 2, 3$). We also report some general properties of the occurrences of the motifs that appear in the network with such parameters.

The number of motifs to be tested is the number of multi-choices of k elements among $|C_s|$, where $|C_s|$ is the number of EC numbers with a cut-off value s. For $s = 3$, $|C_s| = 153$ and the number of motifs of size 3 (respectively 4) is 608685 (respectively 23738715).

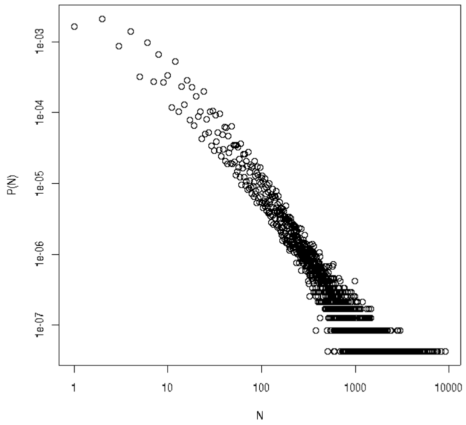

Figure 4: Log-log representation of the observed number of occurrences for motifs of size $k = 4$ with cut-off value $s = 3$. $P(N)$ is the probability that a motif occurs N times. This probability decays as a power law: $P(N) = N^{-1.6}$.

Surprisingly, the distribution of the number of occurrences of motifs of size k is found to be well modelled by a power law (see Figure 4 for $k = 4$ and $s = 3$). Similar results are obtained for $k = 4$, $s = 2$ and for $k = 3$, $s = 2, 3$ (data not shown). This type of distribution has already been shown to represent a good model for the distribution of the degrees of the vertices in metabolic networks, indicating the existence of a few hubs (highly connected vertices) and a vast majority of poorly connected vertices [6]. What had not been reported before (to the best of our knowledge) is that a similar observation can be made for motifs. Indeed, the distribution of the number of occurrences of motifs of a given size indicates that some motifs are highly repeated in the metabolic network while many have only a few occurrences. Furthermore, many motifs never occur in the network. Such motifs are naturally absent from a log-log representation but they correspond to the vast majority of the tested motifs (95% in the case of motifs of size 3 and 98% for motifs of size 4, even with a cut-off value in both cases of 3).

Observe that in this paper we do not discuss over-representation of motifs, which is a difficult theoretical issue in the case of labelled graphs such as metabolic networks. We stress only the fact that, of course, highly represented motifs do not necessarily correspond to over-represented motifs.

Another interesting result that comes out of this systematic study is the number of occurrences that are not included in a single pathway. We call them "inter-pathway occurrences". As we have shown in the previous application, such occurrences exist. In most cases, they are hard to interpret but they could correspond to alternative reaction paths that are not visible with the classical representation of metabolism as a collection of disjoint metabolic pathways. We found that, on average, a motif of size 3 (respectively 4) has 74% (respectively 92%) of its occurrences that are inter-pathway occurrences (always with a cut-off value of 3).

All inter-pathway occurrences may not represent biologically meaningful chemical paths but the proportions above suggest that a lot of information may be lost when we study pathways and not networks.

5 Conclusion

In this paper, we presented a novel definition of motif, called a "reaction motif", in the context of metabolic networks. Unlike previous works, the definition of motif is focused on reaction labels while the topology is not specified. Such a novel definition raises original algorithmic issues of which we discuss the complexity in the case of the problem of searching for such motifs in a network. To demonstrate the utility of our definition, we show an example of application to the comparative analysis of different amino-acid biosynthesis pathways. This work represents a first step in the process of exploring the building blocks of metabolic networks. It seems promising in the sense that, with a simple definition of motif, biologically meaningful results are found.

We are currently working on an enriched definition of motif that will take into account information on input and output compounds. The current definition already enables to discover regularities in the network. Enriched definitions should enable to test more precise hypotheses.

In this paper, we used a particular formalism for analysing a metabolic network through the identification of motifs. Other formalisms have been employed or could be considered. As Stelling indicated in his review of 2004 [17], each formalism gives a different perspective and confronting them seems to be a promising way of getting at a deeper understanding of such complex networks.

Acknowledgements The authors would like to thank Anne Morgat, Alain Viari and Eric Tannier for very fruitful discussions. The work presented in this paper was funded in part by the ACI Nouvelles Interfaces des Mathématiques (project π-*vert*) of the French Ministry of Research, by the ARC (project *IBN*) from the INRIA and by the ANR (project *REGLIS*).

References

[1] E. Alm and A. P Arkin. Biological networks. *Curr. Opin. Struct. Biol.*, 13(2):193–202, Apr 2003.

[2] M. Arita. The metabolic world of Escherichia coli is not small. *Proc. Natl. Acad. Sci. USA*, 101(6):1543–1547, Feb 2004.

[3] M.R. Garey and D.S. Johnson. *Computers and Intractability. A Guide to the Theory of NP-Completeness*. Freeman, 1979.

[4] R. Guimerà and L.A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, Feb 2005.

[5] L. Hartwell, J. Hopfield, A. Leibler, and A. Murray. From molecular to modular cell biology. *Nature*, 402:c47–c52, 1999.

[6] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[7] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32:277–280, 2004.

[8] H. Kitano. Systems biology: A brief overview. *Science*, 295:1662–1664, 2002.

[9] H.-W. Ma, X.-M. Zhao, Y.-J. Yuan, and A.-P. Zeng. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics*, 20(12):1870–1876, Aug 2004.

[10] J.A. Papin, J.L. Reed, and B.O. Palsson. Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. *Trends Biochem. Sci.*, 29(12):641–647, Dec 2004.

[11] J.A. Papin, J. Stelling, N.D. Price, S. Klamt, S. Schuster, and B.O. Palsson. Comparison of network-based pathway analysis methods. *Trends Biotechnol.*, 22(8):400–405, Aug 2004.

[12] R.Y. Pinter, O. Rokhlenko, D. Tsur, and M. Ziv-Ukelson. Approximate labelled subtree homeomorphism. In *Proceedings of the 15th Annual Symposium on Combinatorial Pattern Matching (CPM)*, volume 3109 of *LNCS*, pages 59–73, 2004.

[13] S. Schuster, T. Pfeiffer, F. Moldenhauer, I. Koch, and T. Dandekar. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. *Bioinformatics*, 18(2):351–361, Feb 2002.

[14] D. Segrè. The regulatory software of cellular metabolism. *Trends Biotechnol.*, 22(6):261–265, Jun 2004.

[15] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–504, 2003.

[16] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.*, 31(1):64–68, May 2002.

[17] J. Stelling. Mathematical models in microbial systems biology. *Curr. Opin. Microbiol.*, 7(5):513–518, Oct 2004.

[18] Y. Tohsato, H. Matsuda, and A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:376–383, 2000.

[19] A. M. Velasco, J. I. Leguina, and A. Lazcano. Molecular evolution of the lysine biosynthetic pathways. *J. Mol. Evol.*, 55:445–459, 2002.

[20] K. Voss, M. Heiner, and I. Koch. Steady state analysis of metabolic pathways using Petri nets. *In Silico Biol.*, 3(3):367–387, 2003.

[21] E.C. Webb. *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Oxford University Press, 1992.

[22] D.M. Wolf and A.P. Arkin. Motifs, modules and games in bacteria. *Curr. Opin. Microbiol.*, 6(2):125–134, Apr 2003.