Graph Transformation in Molecular Biology*

Francesc Rosselló[1] and Gabriel Valiente[2]

[1] Department of Mathematics and Computer Science, Research Institute of Health Science (IUNICS), University of the Balearic Islands, E-07122 Palma de Mallorca
[2] Department of Software, Technical University of Catalonia, E-08034 Barcelona

Abstract. In the beginning, one of the main fields of application of graph transformation was biology, and more specifically morphology. Later, however, it was like if the biological applications had been left aside by the graph transformation community, just to be moved back into the mainstream these very last years with a new interest in molecular biology. In this paper, we review several fields of application of graph grammars in molecular biology, including: the modeling higher-dimensional structures of biomolecules, the description of biochemical reactions, the analysis of metabolic pathways, and their potential use in computational systems biology.

1 Introduction

Once upon a time, biology was one of the main fields of application of graph transformation, as it is proved by the maiden name (back in 1978) "Workshop on Graph Grammars and Their Application to Computer Science and Biology" of the current "International Conference on Graph Transformation." Those early applications of graph rewriting in biology mostly belonged to the field of morphogenesis.

It is common knowledge that graphs describe structures in a simplified but explicit way. In such descriptions, nodes correspond to substructures and arcs represent relations among substructures. These arcs can be directed if the relation is so, labelled if one wants to record the kind of relation they stand for, and so on. On their turn, nodes may be labelled to make explicit what they symbolize, whith labels that may be not only raw names, but also graphs themselves, or other higher-order objects that can be used to abstract the details of the substructure represented by the node in hierarchical structures. In any case, the actual meaning of the nodes and the arcs will depend on the actual application. Under this graphical representation of structures, the evolution of the latter can be described by graph rewriting mechanisms, where one or several subgraphs are replaced by other graphs in a way determined by evolution rules specified in a graph grammar.

* This work has been partially supported by the Spanish CICYT, project MAVERISH (TIC2001-2476-C03-01) and by the Spanish DGES and the EU program FEDER, project BFM2003-00771 ALBIOM.

It was soon noticed that the development states of an organism can be described as graphs in this way, with nodes representing for instance cells, body segments, or tissues, and arcs representing spatial or biological relations among nodes. The nodes' labels may be used to denote their type and the arcs' labels the type of interaction they stand for. The rules governing some aspect of the development of such an organism can be described in this framework as graph rewriting rules and gathered in a graph grammar. In a given application, these rules can be fired simultaneously, in a synchronized way, or following some priority order. It was precisely the possibility of modelling the development of organisms where changes and segmentations take place simultaneously at different places that lead to the notion of parallel graph grammars, also called *graph L-systems*, as a generalization of string L-systems. They were introduced about thirty years ago by K. Culik and A. Lindemayer [1], previously hinted by B. Mayoh [2], and they have been used since then in many applications of graph rewriting in morphogenesis.

This was the first kind of applications of graph rewriting in biology, and actually the use of graph grammars as models of the development of organisms is still alive. For instance, Beck, Benkö et al [3] have proposed recently the use of graph transformation as an alternative to standard morphospace representations and geometric morphometrics in the field of theoretical morphology.

The success of graph grammars in the description of development pathways can be seen as a simple instance of their pattern handling power. According to D. Gernert [4], as soon as patterns are represented as graphs, graph grammars are a natural tool to describe the fundamental operations related to patterns: pattern generation, pattern transfer (the duplication of a certain subpattern and its insertion in a different location), pattern recognition, pattern interpretation (the influence of certain subpatterns on the behavior of whole system) and pattern application (the transmission of a certain pattern to another location). A type of graph grammars specifically tailored to handle patterns was proposed in [5].

Patterns that are conveniently modelled as graphs are found everywhere in biology, and not only in morphology. Molecular biology is no exception: the inner structure of chemical compounds [6], the tridimensional structure of nucleic acids and proteins [7], the chemical reactions [8], the biochemical and metabolical pathways [9], most formal components of molecular biology can be represented as graphs. This fact must be added to what is called in sociology of science "the phenomenon of the earlier tool" [4]: when some branch of mathematics reaches a high standard or it becomes fashionable, then it will be surely used in many other sciences.[1] Therefore, it should not be a surprise that, with the recent

[1] Historians of science put more emphasis on the converse phenomenon, when a problem in some science gives rise to new a branch of mathematics or gives new life to an already existing branch; for instance, the theory of Abstract Data Types gave a boost to universal algebra... and H. Ehrig [10] has his share of guilt, in this connection. Graph grammars can also be seen as an example of this phenomenon, as they were born to solve the problem of specifying the transformation of non linear structures in software systems.

thriving of computational molecular biology and computational systems biology, graph grammars have initiated what will probably become a second silver age of applications in biology.

The goal of this paper is to overview some applications of graph rewriting in molecular biology. In the next section we shall write about the modeling of tridimensional structures of nucleic acids and proteins. In Section 3 we will cover the modeling of chemical compounds and chemical reactions in artificial chemistries and then, in Section 4, the application of the latter in the analysis of biochemical pathways and, in Section 5, in computational systems biology. With this short survey we want to call the attention of graph-grammarists mostly oriented to software systems specification and to invite them to catch a glimpse of a completely different world of possible applications.

2 Higher-dimensional structures of biomolecules

A biomolecule can always be viewed as an oriented chain of monomers, which in turn can be mathematically described as a string over a suitable alphabet. This string is called the *primary structure* of the molecule. For instance, a DNA or an RNA molecule is a chain of nucleotides, each one of them characterized by the base attached to it: adenine, A, cytosine, C, guanine, G, or thymine, T, (in RNA, thymine is replaced by uracil, U). Thus, the primary structure of a DNA molecule is a string over $\{A, C, G, T\}$, while the primary structure of an RNA molecule is a string over $\{A, C, G, U\}$. In a similar way, proteins are chains of aminoacids, and hence the primary structure of a protein is a string over a 20-letter alphabet, for instance

$$\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\},$$

each letter representing an aminoacid: A for Alanine, C for Cysteine, D for Aspartic acid, etcetera.

In the cell and *in vitro*, each RNA molecule and protein folds into a tridimensional structure, and this is structure what determines its biochemical function. The understanding of the folding process of these biomolecules and the prediction of their tridimensional structure from their primary structure are two of the main open questions in molecular biology.

As different levels of graining are suitable for different problems [11], we can sometimes forget about the detailed description of these tridimensional structures and consider only a simplified model of them, like for instance their contact structures. The *contact structure* [12] of a biomolecular tridimensional structure is the set of all pairs of monomers that are either consecutive in the chain or, in some specific sense, neighbors in the structure. Such a contact structure can be mathematically described as an undirected graph without multiple edges or self-loops, with sets of nodes representing the monomers numbered according to their position along the chain and with edges of two types: those that join pairs of consecutively numbered monomers, which are said to form the *backbone* of the contact structure, and the other ones, which are called *contacts*.

The secondary structures of RNA molecules form a special class of contact structures. In them, contacts represent the *hydrogen bonds* between pairs of nonconsecutive bases[2] that hold together the tridimensional structure. A restriction, called the *unique bonds condition*, is added to the definition of RNA secondary structure [7]: a base can only pair with at most another base. It is usual to impose a final restriction on *RNA secondary structures*, by forbidding the existence of *knots*, i.e., of contacts that "cross" each other. This restriction has its origin in the first dynamic programming methods to predict RNA secondary structures [7, 13], but real RNA structures can contain knots, which are moreover important structural elements of them [14]. Contact structures with unique bonds and knots can also be used to represent the local basic building blocks of protein structures, like *α-helixes* or *β-sheets*, often called *protein secondary structures*.

Beyond secondary structures, the representation of the neighborhood in tridimensional structures of RNA molecules and proteins needs contact structures without unique bonds. The full contact structure of an RNA molecule may contain sets of contacts that violate the unique bonds condition, like base triplets and guanine platforms [15, 16], and in the contact structure of a protein, one aminoacid is usually an spatial neighbor of several aminoacids [17, 18].

Although the theory of formal languages was born in the 1950s, and then almost simultaneously to modern molecular biology (recall that F. Crick and J. Watson discovered DNA's double helix in 1953 and N. Chomsky published *Syntactic structures* in 1957), it was not until the 1980s that formal grammars methods started to be applied to biomolecular sequences [19]. A little later it was also noticed that string grammars could also be used to model and study not only the primary structure of biomolecules, but also certain aspects of their contact structures, as for instance secondary structures of RNA molecules [20, 21]. In these approaches, an RNA secondary structure is represented by a derivation tree of a certain context free grammar, while RNA contact structures with unique bonds and knots must be generated by new types of string grammars [22]. Many more works have focused on secondary structures, or, more in general, contact structures with unique bonds, which can be easily described as strings over a complemented language. There have also been a few attempts to model and study simple aspects of the secondary structures of proteins using string grammar methods. Two typical examples are the SMART [23] and the TOPS [24] systems.

The goal of the representation of contact structures of biomolecules by means of grammars is to contribute to both main questions about contact structures of biomolecules mentioned above. From the theoretical point of view, one expects to deduce properties of their folding mechanisms from the performance of these grammars and the accuracy with which they generate real structures. In this way, these grammars would yield to a better understanding of the folding process of nucleic acids and proteins. From the practical point of view, stochastic versions of these grammars can be used to predict contact structures. Recall that a *stochastic grammar* specifies a probability for each production, and in this way it assigns

[2] Actually, a hydrogen bond can only form between bases that are at least four positions apart in the chain.

a probability to every derivation. Once a grammar is *trained*, i.e., its probability parameters are tuned on a set of training examples, it can be used to predict the contact structure corresponding to a given primary structure as the most likely derivation of a structure with this primary structure. In the case of string regular grammars, this last step can also done using the very popular, and equivalent, formalism of Hidden Markov Models [25], while in the case of stochastic context-free grammars ad-hoc parsing methods are used [26].

Nevertheless, it is clear that it will be difficult to go beyond these results using string grammars in the study of protein structures, because of their high complexity [27]. As contact structures of proteins are graphs, the clear candidate to generate them are graph grammars.

There have been several important advances in the theoretical study of the protein folding problem using graph grammars in a hidden way. In these approaches, "rules take the form of local structure generators, from which structure evolves via iterative application of elementary steps" [29, p. 409]. Actually, the first rule-based approach to protein folding [30] dates of 1977, and consists of several explicitly described composition rules for the formation, growth and coalescence of β-sheets that could perfectly be formalized as graph rewriting rules. Another description of the formation of protein domains[3] and their relative position as the result of the hierarchical application of explicit rules that are reminiscent of graph transformation rules is due to Lesk [31]. Rule-based descriptions of folding processes of RNA molecules have also been proposed [32, 33].

A paradigmatic, and very interesting, work in this line of research is Przytycka et al's rule-based description of a certain class of protein contact structures, the so-called *all-β proteins*, that admit a high variety of topologies and are difficult to predict from their primary structure. These researchers use a grammar consisting on four composition rules, or rather four families of composition rules, motivated by biophysical considerations that make them conjecture that their rules have physical correlates in the actual mechanism of protein folding [29]. Contrary to all previous rule-based approaches to protein contact structures, their rules are explicitly presented as a graph grammar. For the purpose of this grammar, all-β proteins are represented by graphs with nodes corresponding to β-sheets and two types of edges: there are *domain edges*, that are generated by the application of the folding rules and combine the β-sheets to generate more complex folds, and *neighbor edges*, that represents the spatial juxtaposition of non-consecutive β-sheets after the application of a rule by means of a closure operation that can also be represented by a graph rewriting rule. The start graph has only neighbor edges between consecutive β-sheets and no domain edge, and successive applications of the rules group β-sheets by means of domain edges into more complex domains and connect β-sheets that are distant in the sequence but that become juxtaposed in space due to spatial restrictions. Figure 1 displays a derivation of this graph grammar, extracted from [29].

[3] A *domain* of a protein is a piece that folds into a stable higher order contact structure.

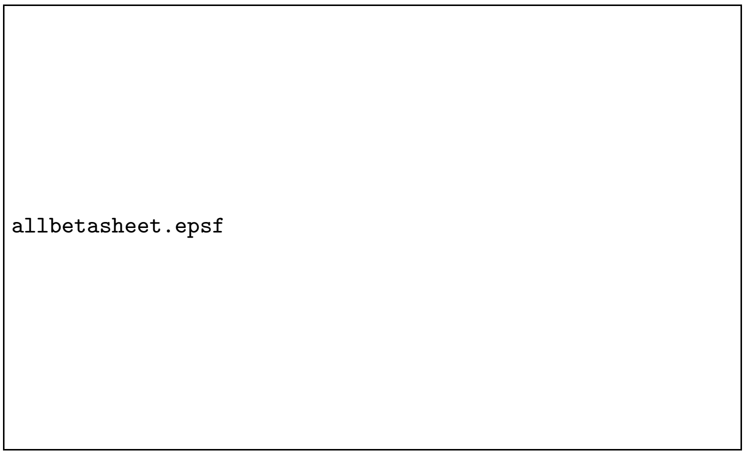


Fig. 1. A derivation of Przytycka et al's grammar

What can the graph grammar community bring to this line of research? To our opinion, it is the biologists' task to propose new rules of formation of contact structures of RNA molecules and proteins, but graph-grammarians could and should collaborate, among other tasks, in formalizing these rules and analyzing the redundancies that appear in the grammars; in determining the properties, for instance related to parallelism and concurrency, of the rewriting systems they define, which might lead to uncovering properties of the real folding process; and in developing general methods to characterize the sets of structures generated by any given set of biomolecules' folding rules, i.e., to determine which structures are possible and which ones are not under any set of folding mechanisms, which might give new insights not only on folding processes but also on the evolution mechanisms.

As far as the prediction of protein contact structures goes, Abe and Mamitsuka [34] proposed in 1997 a stochastic tree grammar to predict β-sheets in a way similar to those developed for RNA secondary structures using string grammars that we recalled above. Stochastic versions of more general graph grammars will be necessary to predict contact structures of RNA molecules and proteins beyond secondary structures using this kind of methods. To do that, one should find a set of rules, perhaps in the spirit of those discussed in the previous paragraphs, that capture the formation of the different components of the contact structures of the target biomolecules as well as their relation; one should develop an efficient technique to estimate the grammar's parameters from a set of training graphs; and one should devise an efficient method to find the most probable structure of a protein given the grammar.

3 Artificial chemistry

Roughly speaking, an *artificial chemistry* [35, 36] is a computational model of a chemical system. It consists of a set (a *soup*) of objects, called *molecules*, a set of *reaction rules* that produce new molecules from already existing molecules, and the definition of the *dynamics* of the system, that specifies the application conditions of the rules, the preference in their application, etc. Against other types of computational models, the goal of an artificial chemistry is to answer qualitative, rather than quantitative, questions: the existence of steady, or closed and self-maintaining, states, the size and diversity of the soup at some moment, etc.

The nature of the molecules, the reactions, and the dynamics of an artificial chemistry can be quite diverse. For instance, in one of the first artificial chemistries, Walter Fontana's *AlChemy* [37], objects were λ-terms, a reaction consisted of the application of the first λ-term to the second one, and the dynamics followed a combination of randomness (in the selection of the pair of molecules) and an explicit algorithm (to decide whether the reaction took place or not).

Now, although artificial chemistries can be, and have been, used to model many kinds of systems, their primary targets are 'real' chemistries, in which case molecules should be representations of chemical compounds, and reaction rules of chemical reactions. Now, chemical descriptions of 'real' molecules can be made at different levels of resolution:

- A *molecular descriptor* uniquely identifies a molecule in a biochemical database. For instance, beta-D-Glucose is entry number C00221 in the KEGG database [38].
- A *molecular formula* indicates the number of each type of atom in a molecule. For instance, beta-D-Glucose has the molecular formula $C_6H_{12}O_6$.
- A *constitutional formula* or *chemical graph* indicates which pairs of these atoms are bonded. For instance, beta-D-Glucose has the following chemical graph displayed in Figure 2.

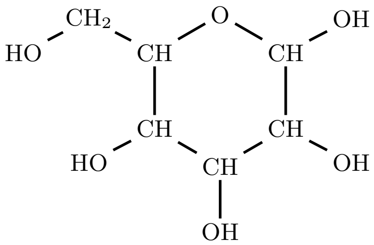


Fig. 2. Beta-D-Glucose's chemical graph.

- A *structural formula* refines a chemical graph by indicating those stereochemical distinctions that are required to uniquely identify a molecule. For instance, Figure 3 downloaded from the KEGG database, displays the structural formula of beta-D-Glucose; in it, plain lines depict bonds approximately in the plane of the drawing, bonds to atoms above the plane are shown with a bold wedge, and bonds to atoms below the plane are shown with short parallel lines.

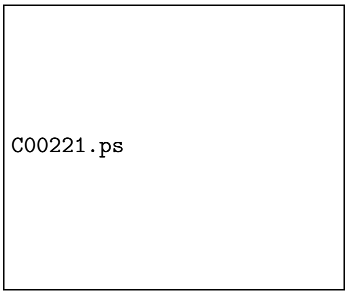


Fig. 3. Beta-D-glucose's structural formula.

This representation allows to distinguish beta-D-Glucose from other chemical compounds with the same chemical graph. D-glucose and L-glucose are mirror images and therefore they share the same chemical graph. Further, there are two possible orientations for the upper-right OH group, which is linked to the CH group number 7 in the ring structure: below the plane of the drawing (alpha-D-glucose) and above the plane of the drawing (beta-D-glucose).

A chemical description at the level of molecular descriptors and molecular formulas is useful for database retrieval purposes, and they can be used in an artificial chemistry when the knowledge of the structure of the chemical compounds is not necessary. In this case, molecular descriptors and formulas play the roles of simple labels, but then chemical reactions cannot be defined by means of local interactions of the atoms of the substrate's molecules.

Chemical graphs are probably *the* natural and the most familiar representation of molecules [39]. In first course Organic Chemistry classes, chemical reactions are explained in terms of constitutional formulas and a handful of reaction mechanisms, which corresponds to (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 labelled graphs as molecules and graph transformation rules as reactions.

Several such artificial chemistries have been proposed so far. J. McCaskill and U. Niemann [40] proposed in 2000 a artificial chemistry for DNA and RNA processing based on graph transformation. In it, molecules are labelled graphs of a specific type, called *variable graphs*, that can represent nucleotides, nucleic acid single or double strands, or sets of all the latter. The reaction rules represent several types of chemical reactions: unimolecular (only one molecule is involved), bimolecular (two molecules react together) and enzymatic (a special type of unimolecular reaction that represents the attachment or the removal of an enzyme in a specific position of a molecule). These reaction rules are graph transformation rules that act in the usual matching-replacement-embedding way, in a way reminiscent of single-pushout approach [41]: when a node is removed, all nodes adjacent to it are also removed. All other reactions, including complex enzymatic reactions, can be decomposed into a series of applications of these reaction rules. The dynamics of the system simply consists of performing all possible reactions through a branching process to obtain all possible derivation paths. The final goal is to predict all libraries of nucleic acids arising from a given set of strands by means of a given set of enzymatic reactions. The authors have implemented their artificial chemistry in a computer program called MOLGRAPH.

More recently, an artificial chemistry for organic chemistry called the *Toy Model* has been developed by G. Benkö, C. Flamm and P. Stadler [39, 42, 43]. In it, and following [44], molecules are *orbital graphs*: undirected graphs with nodes representing outer atom orbitals, labelled by the atomic element and the hybridization type of the orbital, and edges representing overlaps of adjacent orbitals. These orbital graphs represent sets of chemical compounds and they are uniquely determined by the chemical graphs of the chemical compounds, but they moreover incorporate chemically meaningful energy functions that allow the computation of reaction energies.

The reactions rules translate to this level of abstraction the basic organic reaction mechanisms as graph transformation rules that preserve the vertex labels and the total degrees of corresponding nodes, to capture the conservation of atoms and valences in organic reactions. These graph rewriting rules are actually double pushout production rules [45] over orbital graphs: the left-hand side, context, and right-hand side are orbital graphs with the same labelled nodes; the left-hand side graph represents the substrate, the right-hand side graph represents the product and the context graph has as edges those appearing in both the substrate and the product with the same type.

Consider,[4] for example, the Diels-Alder reaction [46], one of the most important reactions in organic chemistry. The substrate of the reaction, 1,3-butadiene (C_4H_6) and ethylene (C_2H_4), is combined to form cyclohexene (C_6H_{10}), as described by the double-pushout transformation rule displayed in Figure 4.

A forward application of the previous double-pushout transformation rule to 1,3-butadiene (C_4H_6) and dihydro-2,5-furandione ($C_4H_4O_3$) to form 1,3-isobenzofurandione ($C_8H_8O_3$), corresponds to the double-pushout transformation in Figure 5.

In this artificial chemistry, the rules can be applied randomly, or according to the reactivity index of the matching step computed using suitable formulas, that

[4] Usually, hydrogen atoms and the corresponding bonds are not represented explicitly in constitutional formulas.

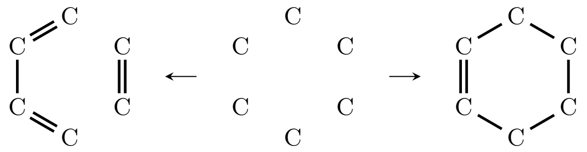


Fig. 4. The Toy model double-pushout rule for the Diels-Adler reaction

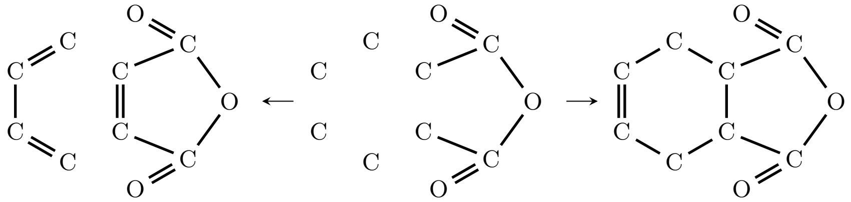


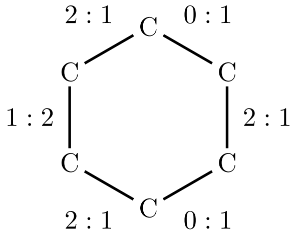
Fig. 5. A reaction in the Toy model

can for instance be translated into a reaction rate constant. This graph rewriting system has been implemented in *Maude* as a client/server application. The final goal is again to compute extensively all possible results of any specific instance of this artificial chemistry, in this case under the form of large chemical reaction networks defined by an initial set of molecules and the set of allowed reactions. We shall talk more on chemical networks, which are graphs themselves, in the next section.

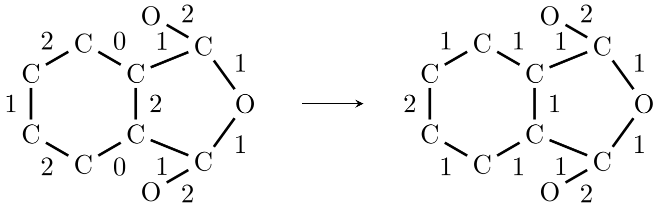
We have recently started to develop an artificial network based on graph grammars[47, 48], also with the final aim of studying biochemical networks. In our approach, and following Fujita's *imaginary transition structures*[49, 50, 8] to model chemical reactions, molecules are *generalized chemical graphs*: chemical graphs with possibly some extra edges labelled 0 and reactions are described as edge relabeling graph transformation rules. These reactions can be *explicit* and *implicit.*

An *explicit chemical reaction* is an undirected graph (without multiple edges or self-loops) whose nodes are labeled by means of chemical elements and whose edges are labeled by the combination of two natural numbers: a *substrate weight* and a *product weight.* No edge can have both substrate and product weights equal to zero and, for all nodes, the total substrate weight cannot be equal to zero and must coincide with the total product weight, over all edges incident with the node.

For instance, in the aforementioned Diels-Alder reaction, there are six carbon atoms involved, one single bond (which is turned into a double bond), and three double bonds (which are turned into single bonds). Further, two new bonds are created. Thus, the following explicit chemical reaction models the Diels-Alder reaction, where a label of the form $x : y$ next to an edge means that the edge has substrate weight x and product weight y:



An application of a generalized chemical reaction replaces the substrate weights by the product weights in the matching subgraph. For instance, an application of the previous explicit chemical reaction to 1,3-butadiene (C_4H_6) and dihydro-2,5-furandione ($C_4H_4O_3$) to form 1,3-isobenzofurandione ($C_8H_8O_3$), corresponds to the following edge relabeling graph transformation:



On its turn, an *implicit chemical reaction* is a compact representation of an explicit chemical reaction by means of a finite set of elementary edge relabeling operations that, when applied to a graph taking into account that the total degree of each node must remain constant and that no arc labeled 0 can still be labeled 0 after the application, determine uniquely the product chemical graph. Since the undirected graph underlying the substrate chemical graph is finite, such a minimal set of edge relabeling operations will always exist, although it need not be unique. Our conjecture is that any implicit chemical reaction, at least for reactions with molecules only involving hydrogen, oxygen, nitrogen and carbon, is given by any one of the relabeling operations in an explicit chemical reaction, but we still have not been able to check it.

We have implemented our generalized chemical graphs and reactions on top of PerlMol, but we still have not defined the dynamics of our artificial chemistry.

4 Analysis of Metabolic Pathways

Metabolism is one of the most complex cellular processes. Cells function as organized chemical engines carrying out a large number of transformations, called bioreactions or biochemical reactions, in a suited behavior. One of the interacting molecules in a reaction might act only as a catalyst, that is, it facilitates the association of several molecules to form a compound, and it decreases the energy barrier required for the bond rearrangements that establish a reaction. But in the end the molecule itself separates from the compound and returns to its original state, the law of conservation of mass states: Atoms are neither created, nor destroyed, during any chemical reaction.

Enzymes are proteins, that is, macromolecules consisting of a long sequence of compounds called amino acids. The structure of each enzyme is encoded in the cell's genome by a gene. In order for an enzyme to catalyze a reaction, the molecules on which it acts, called substrates, must have just the right structures and orientations to interact with the intricate 3-dimensional shape of the enzyme [6].

Four major factors influence the rate at which enzymes work:

1. Enzyme concentration: Influencing the rate of a biochemical reaction by changing the enzyme compound is expensive in terms of energy. The cell has to make complex enzyme molecules from simpler ones; this process involves the consumption of energy and time and may, therefore, be unsatisfactory for urgent response.
2. pH and Temperature variation: External (outside the enzyme) conditions are likely to be most inuential within the body or cell of an organism. The wide range of pH and temperature variation in a laboratory test tube may have little relevance to an organism in which body temperature and pH are closely controlled.
3. Substrate concentration: A high substrate concentration may increase the rate of enzyme action to ensure its rapid breakdown. A high product concentration may also inhibit enzyme action so that less product is formed.

Biochemical pathways such as metabolic, regulatory, and signal transduction pathways, are often described in symbolic terms, as a succession of transformations of one set of molecules (called reactants) into another set (called products); reactants and products are collectively referred to as metabolites [7, 9].

Metabolites are classified as internal or external according to whether or not they are to satisfy the quasi-steady-state condition, that is, the total rate of production of each internal metabolite equals the total rate of its consumption. In contrast, external metabolites do not satisfy this condition because they participate in additional reactions that are not involved in the system under study [51].

The analysis of metabolic pathways is a fertile field because of, among other reasons:

1. The completion of genomes has made the comparison of complete metabolic pathways possible, and its analysis across species [1] applies to:
   - Understanding the evolutional relationships between species.
   - Development of species-specific drug targets.
   - Identification of previously unknown parts of pathways in a species.

2. It represents a natural step up in modeling of biological systems relative to the study of biological macromolecules [2].
3. Modeling is important in guiding the biotechnological engineering of cells to maximize the industrial output of specified products [2].

The mathematical analysis of metabolic pathways has been approached through a large sort of techniques contemplating distinctive objectives and hope. Metabolic pathways can range in size from involving a few enzymes and metabolites to the complete pathway of an organism that can have thousands of them [2].

Some of these approaches are: (...)

Metabolism is the general term for all the chemical reactions in the body.

Metabolic pathways are represented in a natural way as directed graphs, with the substrates, products, and enzymes as nodes and the chemical reactions catalyzed by the enzymes as arcs.

5 Computational systems biology

Term rewriting systems have also been used to define biochemically inspired computational systems [52].

Pathway logic,

[53] [54]

ciliats

6 Conclusion

Everybody has in his or her mind some case where a problem in some science has given rise to new a branch of mathematics, or at least has given new life to an already existing branch: for instance, the theory of Abstract Data Types gave a boost to universal algebra,[5] which otherwise perhaps would not have survived as an active field of research. But the reciprocal phenomenon has also happened, usually at a lower level, and it is called "the phenomenon of the earlier tool" in sociology of science [4]. When some branch of mathematics reaches a high standard then it will be surely used in many other sciences, specially in those that are fashionable a that moment, be it because of pure scientific reasons, or because of funding reasons; many methods of algebraic geometry have entered physics in this way in the 1980s, but also the surge of applications of physics in biology after the World War II has a certain component of this effect.

Graph rewriting was born more than 30 years ago with an eye on its applications, but since then it has become a very popular mathematical field of research in theoretical computer science. The obvious fact that graphs can be used to model many kind of structures, and hence graph rewriting can be used to model the transformation of these structures, with some contribution of the phenomenon of the earlier tool, have made graph rewriting to find many applications, including in molecular biology.

[5] And H. Ehrig [10] has his share of guilt, here!

Regarding the description of biochemical reactions, an interesting open problem consists in the automatic detection of the difference between candidate substrate and product molecules, which would lead to the automatic construction of chemical reaction graphs from a large database of chemical graphs.

It could be interesting to study application and uniqueness conditions for implicit chemical reactions.

References

1. Culik II, K., Lindenmayer, A.: Parallel rewriting on graphs and multidimensional development. Int. Journ. of General Systems **3** (1976) 53–66
2. Mayoh, B.: Multidimensional Lindemayer organisms. In: L-systems. Volume 15 of Lecture Notes in Computer Science., Springer-Verlag (1974) 302–326
3. Beck, M., Benkö, G., G. Eble, C.F., Müller, S., Stadler, P.: Graph grammars as models for the evolution of developmental pathways. In: The Logic of Artificial Life: Abstracting and Synthesizing the Principles of Living Systems (Proceedings of GWAL 2004), IOS Press (2004) 8–15
4. Gernert, D.: Graph grammars as an analytical tool in physics and biology. Biosystems **43** (1997) 179–187
5. Mayoh, B.: On patterns and graphs. Preprint (1995)
6. Cayley, A.: On the mathematical theory of isomers. Philosophical Magazine **47** (1874) 444–446
7. Waterman, M.S., Smith, T.F.: RNA secondary structure: a complete mathematical analysis. Math. Biosci. **42** (1978) 257–266
8. Fujita, S.: Computer-Oriented Representation of Organic Reactions. Yoshioka Shoten, Kyoto (2001)
9. Michal, G., ed.: Biological Pathways: An Atlas of Biochemistry and Molecular Biology. John Wiley & Sons, New York (1999)
10. Ehrig, H., Mahr, B.: Fundamentals of algebraic specification I: Equations and initial semantics. Springer Verlag (1985)
11. Reidys, C., Stadler, P.F.: Bio-molecular shapes and algebraic structures. Computers & Chemistry **20** (1996) 85–94
12. Chan, H.S., Dill, K.A.: Compact polymers. Macromolecules **22** (1989) 4559–4573
13. Zuker, M., Sankoff, D.: RNA secondary structures and their prediction. Bull. Math. Biol. **46** (1984) 591–621
14. Westhof, E., L. Jaeger, L.: RNA pseudoknots. Curr. Opinion Struct. Biol. **2** (1992) 327–333
15. Batey, R.T., Rambo, R.P., Doudna, J.A.: Tertiary motifs and folding of RNA. Angew. Chem. Int. Ed. **38** (1999) 2326–2343
16. Moore, P.B.: Structural motifs in RNA. Annu. Rev. Biochem. **68** (1999) 287–300
17. Chan, H.S., Dill, K.A.: Sequence space soup of proteins and copolymers. J. Chem. Phys. **95** (1991) 3775–3779
18. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yeo, D.P., Thomas, P.D., Chan, H.S.: Principles of protein folding: A perspective from simple exact models. Protein Science **4** (1995) 561–602
19. Brendel, V., Busse, H.G.: Genome structure described by formal languages. Nucleic Acid Research **12** (1984) 2561–2568
20. Searls, D.: Formal language and biological macromolecules. In: Mathematical Support for Molecular Biology. Volume 47 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science., AMS (1999) 128–141

21. Searls, D.: The computational linguistics of biological sequences. In: Artificial Intelligence and Molecular Biology, AAAI Press (1993) 47–120
22. Rivas, E., Eddy, S.R.: The language of RNA: a formal grammar that includes pseudoknots. Bioinformatics **16** (2000) 334–340
23. Schultz, J., Milpetz, F., Bork, P., Ponting, C.: SMART, a simple molecular architecture research tool. PNAS **95** (1998) 5857–5864
24. Westhead, D., Slidel, T., Flores, T., Thornton, J.: Protein structural topology: automated analysis and diagrammatic representation. Protein Science **8** (1999) 897–904
25. Durbin, R., Krogh., A., Mitchison, G., Eddy, S.: Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge Univ. Press, Cambridge (1998)
26. Sakakibara, Y., Brown, M., Hughey, R., Mian, I., Sjolander, K., Underwood, R., Haussler, D.: Stochastic context-free grammars for tRNA modeling. Nucleic Acids Research **22** (1994) 5112–5128
27. Searls, D.: The language of genes. Nature **420** (2002) 211–217
28. Mayoh, B.: DNA pattern multigrammars. Preprint (1995)
29. Przytycka, T., Srinivasan, T., Rose, G.: Recursive domains in proteins. Protein Science **11** (2002) 409–417
30. Richardson, J.: β-sheet topology and the relatedness of proteins. Nature **268** (1977) 495–500
31. Lesk, A.M.: Systematic representation of protein folding patterns. J. Mol. Graph. **13** (1995) 159–164
32. Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M., Schuster, P.: Fast folding and comparison of RNA secondary structures. Monatsh. Chem. **125** (1994) 167–188
33. Kister, A., Magarshak, Y., Malinsky, J.: The theoretical analysis of the process of RNA molecule self-assembly. BioSystems **30** (1993) 31–48
34. Abe, N., Mamitsuka, H.: Predicting protein secondary structure using stochastic tree grammars. Machine learning **29** (1997) 275–301
35. Dittrich, P., Ziegler, J., Banzhaff, W.: Artificial chemistries—a review. Artificial life **7** (2001) 225–275
36. Speroni, P.: Artificial chemistries. Bulletin EATCS **76** (2002) 128–141
37. Fontana, W.: Algorithmic chemistry. In: Artificial life II. Volume 47 of Santa Fe Institute Studies in the Sciences of Complexity., Addison-Wesley (1992) 159–210
38. Kanehisa, M., Goto, S.: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research **28** (2000) 27–30
39. Benkö, G., Flamm, C., Stadler, P.F.: A graph-based toy model of chemistry. Journal of Chemical Information and Computer Sciences **43** (2003) 1085–1093
40. McCaskill, J., Niemann, U.: Graph replacement chemistry for DNA processing. In: DNA 2000. Volume 2054 of Lecture Notes in Computer Science., Springer-Verlag (2001) 103–116
41. Ehrig, H., Heckel, R., Korff, M., Löwe, M., Ribeiro, L., Wagner, A., Corradini, A.: Algebraic approaches to graph transformation. part II: Single pushout approach and comparison with double pushout approach. In Rozenberg, G., ed.: Handbook of Graph Grammars and Computing by Graph Transformation. Volume 1: Foundations. World Scientific (1997) 247–312
42. Benkö, G., Flamm, C., Stadler, P.F.: Generic properties of chemical networks: Artificial chemistry based on graph rewriting. In: Proc. 7th European Conf. Advances in Artificial Life. Volume 2801 of Lecture Notes in Computer Science., Springer-Verlag (2003) 10–19

43. Benkö, G., Flamm, C., Stadler, P.F.: Multi-pase artificial chemistry. Submitted (2004)
44. Polanski, O.: Graphs in quantum chemistry. MATCH **1** (1975) 183–195
45. Corradini, A., Montanari, U., Rossi, F., Ehrig, H., Heckel, R., Löwe, M.: Algebraic approaches to graph transformation. Part I: Basic concepts and double pushout approach. In Rozenberg, G., ed.: Handbook of Graph Grammars and Computing by Graph Transformation, Volume 1: Foundations. World Scientific (1997) 163–246
46. Fringuelli, F., Taticchi, A.: The Diels-Alder Reaction: Selected Practical Methods. John Wiley & Sons, Chichester, England (2002)
47. Rosselló, F., Valiente, G.: Analysis of metabolic pathways by graph transformation. In: Proc. 2nd Int. Conf. Graph Transformation. Lecture Notes in Computer Science, Springer-Verlag (2004) to appear
48. Rosselló, F., Valiente, G.: Chemical graphs, chemical reaction graphs, and chemical graph transformation. Electronic Notes in Theoretical Computer Science (2004) to appear
49. Fujita, S.: Description of organic reactions based on imaginary transition structures. Part 1–5. Journal of Chemical Information and Computer Sciences **26** (1986) 205–242
50. Fujita, S.: Description of organic reactions based on imaginary transition structures. Part 6–9. Journal of Chemical Information and Computer Sciences **27** (1987) 99–120
51. Schuster, S., Fell, D.A., Dandekar, T.A.: A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nature Biotechnology **18** (2000) 326–332
52. Berry, G.: The chemical abstract machine. Theoretical Computer Science **96** (1992) 217–248
53. Danos, V., Laneve, C.: Graphs for core molecular biology. In: Proc. 1st Int. Workshop Computational Methods in Systems Biology. Volume 2602 of Lecture Notes in Computer Science., Springer-Verlag (2003) 34–46
54. Danos, V., Laneve, C.: Formal molecular biology. Theoretical Computer Science (2004) in press