Generic Properties of Chemical Networks: Artificial Chemistry Based on Graph Rewriting

Gil Benkö^{1,2}, Christoph Flamm², and Peter F. Stadler^{1,2,3}

¹ Lehrstuhl für Bioinformatik, Institut für Informatik, Universität Leipzig, Kreuzstrasse 7b, D-04103 Leipzig, Germany.

peter.stadler@bioinf.uni-leipzig.de, http://www.bioinf.uni-leipzig.de/

² Insitut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria

³ Santa Fe Institute, Santa Fe, New Mexico

Abstract. We use a Toy Model of chemistry that represents molecules in terms of usual structural formulae to generate large chemical reaction networks. An extremely simplified quantum mechanical energy calculation and a straightforward implementation of reactions as graph rewritings ensure both transparency and closeness to chemical reality, both conditions that are necessary for the analysis of generic properties of large reaction networks. We show that some chemical networks graphs, e.g., repetitive Diels-Alder reactions, have the small-world property and exhibit a scale-free degree distribution. On the other hand, the Formose reaction does not fit well to this paradigm.

1 Introduction

Large-scale chemical reaction networks (CRN) appear as the metabolic networks of living cells [1], they describe the chemical processes in planetary atmospheres [2] as well as in combustion and in combinatorial chemistry [3]. Surprisingly, their generic features have not been studied much in the past. For instance, it is unknown whether the small-world properties of metabolic networks [4–6] are characteristic for living systems, or whether they are common to most or all large chemical networks, as suggested by data reported in [7].

The systematic study of these questions requires a computational model for their generation. In a recent study we have introduced such a Toy Model in which generic properties of extensive chemical reaction networks can be explored in detail and that at the same time preserves the "look-and-feel" of chemistry.

In the following three section we briefly review the main ingredient of a computational implementation of a purely graph-based Toy Model [8] of artificial chemistry: molecules and the evaluation of their physical properties, reaction mechanisms and reactivities, and the formation of large-scale networks. In section 5 we investigate a few reaction networks in detail and provide first results towards identifying generic properties of chemical reaction networks.

2 Molecules

Much of theoretical chemistry is concerned with solving the time-independent Schrödinger equation

$$\widehat{H}\Psi = E\Psi,\tag{1}$$

in the Born-Oppenheimer approximation that describes the electrons in a fixed arrangement of nuclei [9]. For our purposes, however, this is too demanding in terms of computational resources. We therefore resort to an Artificial Chemistry that it derived from equ.(1) by a series of approximations detailed below. In [8] we have shown that such a model retains the "look-and-feel" of organic chemistry; it is hence much more suitable for exploring the generic features of large chemical networks than more abstract systems such as Walter Fontana's AlChemy [10], see also [11] for a recent review on AC models.

We use here a simplified version of Extended Hückel Theory (EHT). EHT [12] is obtained from the wave function Ψ , equ.(1), of the complete molecule by first using the Born-Oppenheimer approximation to separate electron motion from nuclear motion. The orbital approximation then splits the electronic part into orthogonal functions Ψ_{α} called molecular orbitals (MOs). This brings the problem into a form tractable by the variational principle. The linear combination of atomic orbitals (AOs), the *LCAO approximation*, now assumes that one can write $\Psi_{\alpha} = \sum_{i} c_{\alpha,i}\chi_{j}$, where the AOs χ_{j} describe the spatial distribution of the individual electrons of an atom. The problem is thus reduced to starting with a basis set of atomic orbitals χ_{i} and a corresponding overlap matrix $S_{ij} = \int \chi_{i}\chi_{j} d\tau$ and solving the generalized eigenvalue problem

$$\mathbf{H}\boldsymbol{c}_{\alpha} = E_{\alpha}\mathbf{S}\boldsymbol{c}_{\alpha} \,. \tag{2}$$

Here **H** is the one-electron Hamilton matrix, c_{α} denotes the vector of the AO coefficients belonging to the molecular orbital Ψ_{α} with orbital energy E_{α} . The EHT model is now obtained by parametrizing **H** in terms of the overlap integrals S_{ij} , and the *atomic valence state ionization potentials* I_i .

We further simplify the EHT approximation by parametrizing the values S_{ij} of the overlap matrix based only on the type of the AOs χ_i and χ_j at adjacent atoms instead of computing S_{ij} from orbital functions in 3D space. As a consequence we disregard the spatial embedding of the molecule and instead represent it uniquely by its orbital graph [13] which has the outer atom orbitals as its vertices while edges represent overlaps. The orbital graph in turn is uniquely determined by the chemical structure formula, the chemical graph of the molecule, by virtue of the VSEPR rules [14]. Our Toy Chemistry is thus implemented at the level at which (organic) chemistry is usually taught and described: the level of chemical structural formulae. By definition, any property of the molecule can now be computed at least in principle from its wave function, i.e., from the solutions c_{α} of equ.(2). A detailed description of the model and its parametrization can be found in [8].

The main advantage of the Toy Model is that it incorporates a chemically meaningful energy function that is consistent with the framework of the simpli-

 $\mathbf{2}$



Fig. 1. Graph rewriting rule (top) and graph rewriting steps (bottom).

fied EHT theory: The total energy of atomization (TAE) of a molecule is given by

$$E = \sum_{\substack{\text{occupied}\\ \text{orbitals } \alpha}} n_{\alpha} E_{\alpha} \,. \tag{3}$$

From this quantity it is straightforward to compute the reaction energies as the difference of the TAEs of products and educts. The Toy Model hence has built-in chemical thermodynamics.

3 Reactions

With molecules represented as graphs it is natural to view chemical reactions as rewriting rules applied to molecular graphs. A graph grammar is a finite set of rules operating on edge and vertex labeled graphs. It typically consists of a left graph, a right graph, and a context graph defined as the parts of the graph that are removed, persist, and are newly introduced during a rewriting operation, Fig. 1 [15]. This graph rewriting formalism is very flexible and can be used to represent chemical reactions as well as chemically impossible yet strategically interesting reactions. A chemical reaction is the breaking, forming and changing of bonds. Thus the number and type of atoms must remain constant, which can be implemented by conservation of vertex labels. In analogy, the conservation of the number of valence electrons can be imposed on rewrite rules by ensuring conservation of total bond order. Both principles stem from the fact that chemical reactions are stoichiometric [16, 17].

The simulation of unimolecular reactions is a straightforward application of rewrite rules to a molecule. A bimolecular reaction or any other similar rule is split by into one half reaction rule for each educt molecule, and a final reaction rule. The two half reaction rules do not modify existing bonds and atoms in the molecules, they just add flag nodes to the atoms that will be joined during the total reaction. They serve to identify those reaction sites for the reactivity evaluation described in the following. The evaluation determines which pairs of reaction sites from each of the two reactants are joined by a temporary edge.



Fig. 2. One-mode projection from a bipartite graph to a substrate graph.

This temporary construct is then again submitted to the final reaction rule and transformed to the product.

The reaction rate for a specified reaction mechanism is at least in principle determined by the laws of quantum mechanics and can be expressed in terms of the wave function. It is natural therefore to use the Frontier Molecular Orbital (FMO) Theory [18] for this purpose. In the simplest version of FMO theory the reactivity is inversely proportional to the difference of the HOMO and LUMO energies E_{η} and E_{λ} of the reactants. The regioselectivity is determined by the MO coefficients at the reactive sites *i*, such that $\sum_{i} c_{\eta,i} c_{\lambda,i}$ is maximal. Again, we refer to [8] for a more detailed description.

4 Networks

The starting point for CRN generation is an initial set of molecules \mathcal{L}_0 . From this "seed set" we exhaustively generate the network in such a way that a reaction between the same molecules is computed only once, see [8]. While the generation of a large reaction network is rather straightforward, its representation is much less trivial. A plethora of graph-theoretical approaches have been discussed in the literature, see e.g. [19] for an overview. Following [20] a chemical reaction

$$aA + bB + \dots \rightarrow vV + wW + \dots$$
 (4)

can be described as a weighted directed hyperedge in a directed hypergraph $\mathcal{H}(V, E)$ with vertex set V. A hyperedge $\rho \in E$ is a pair of lists of vertices $(\{u_i\}, \{v_i\})$ where the u_i are the initial and the v_i are the terminal vertices of ρ . In chemical terms, the u_i are the educts and the v_i are the products of the reaction ρ . In addition we may have the *stoichiometric coefficients* (prefactors in the above reaction scheme) attached as weights to each u_i and v_i . The *stoichiometric matrix* \mathbf{S} , where $S_{x\rho}$ is the coefficient of the chemical species x in reaction ρ , is therefore the appropriate algebraic representation of a chemical reaction network.

Unfortunately, a well developed theory for the structural analysis of directed hypergraphs is not available. As an alternative, however, one can represent $\mathcal{H}(V, E)$ by an equivalent bipartite weighted digraph in which each hyperedge is "interrupted" by a special vertex representing the reaction itself. Incoming arcs connect this vertex with the educts, out-going arcs point to the products [21, 22,

Table 1. Characteristics of the two example CRNs in [8] and the substrate graph of the *E. coli* energy and biosynthesis metabolism [23].

Network	n	$\langle k \rangle$	$\langle L \rangle$	$\langle L_{\rm rand} \rangle$	$\langle C \rangle$	$\langle C_{\rm rand} \rangle$
Formose	48	3.25	3.55	3.28	0.15	0.068
Diels-Alder	40	4.65	2.15	2.40	0.72	0.110
E.coli	282	7.35	2.9	3.04	0.32	0.026

sect. 2.3.2], Fig. 2. This bipartite network graph is still a faithful representation of the entire reaction network.

Reduced representations of chemical networks are useful in particular when studying connectivity and clustering properties where the distinction of two classes of nodes with completely different interpretations may lead to artifacts in the data. In [23], for example, the *substrate graph* is defined as having the chemical species as vertices and an edge connecting two species x and y if they take part in the same reaction. This amounts to replacing each directed hyperedge ρ by a clique. Alternatively, it is obtained from the bipartite representation by one-mode projection, Fig. 2, [24].

Let us denote the number of vertices by n, and let m be the number of edges of the substrate graph of a chemical reaction network. In the context of analyzing large networks it has turned out the following measures yield useful characteristics:

The average node degree $\langle k \rangle = 2 \frac{m}{n}$.

The average length of the shortest path between to nodes, $\langle L \rangle$, see e.g. [25]. Let d_x be the degree of a vertex x and let q_x be the number of edges that connect neighbors of a vertex x. The clustering at x can then be measured by the fraction of possible triangles, at x, i.e., $C_x = 2q_x/((d_x(d_x - 1)))$. The clustering coefficient $\langle C \rangle$ then measures the overall cliquishness of the graph [26].

The distribution of detour length can be quantified in terms of the minimal cycle basis [7].

In the following section we will compare the characteristic quantitites of chemical networks with two classes of randomly generated graphs: Erdös-Renyi (ER) graphs which feature statistically independent edges. For later reference we note the expected mean path length and clustering coefficients for the ER model in terms of the average vertex degree k = 2m/n:

$$\langle L_{\rm rand} \rangle \approx \frac{\ln n}{\ln \langle k \rangle} \qquad \langle C_{\rm rand} \rangle = \frac{\langle k \rangle}{(n-1)}$$
(5)

Starting with the seminal paper by [26], it has been recognized that these real life network differ qualitatively from the classical ER random graph models by the so-called *small-world* property: while the graphs are very sparse on average, the mutual distances between their vertices of both graph types are



Fig. 3. Evolution of the network of a repetitive Diels-Alder network for different reaction rate thresholds. The CRN grows as the reactivity threshold $\ln k$ is lowered (from left to right: 108, 91.2, and 55.4). The simulation is started with $\sim \gamma \circ \Box \neg \Im$.

approximately of the same size. The major difference lies in the degree of local clustering which is much higher in the real life networks. The characteristic parameters of scale-free networks scale such as those obtained by the preferential attachment procedure [27] (AB model) behave as follows [28]:

$$\langle L_{\rm AB} \rangle \sim n^{-0.75} \qquad \langle C_{\rm AB} \rangle \sim \frac{\ln n}{\ln \ln n} \,.$$
 (6)

The cycles and edge-disjoint unions of cycles of a graph from a vector space (w.r.t. symmetric difference as addition) with dimension $\mu(G) = m - n + c(G)$, where c(G) is the number of connected components. A minimum cycle basis (MCB) is a cycle basis that minimized the total number of edges in the basis cycles. It can be shown that all MCBs have the same numbers n_{ℓ} of cycles of length ℓ [29]. We can therefore use the distribution of MCB cycle lengths to characterize the cycle structure of G, see also [7].

5 Results

The Diels-Alder reaction [30] has been extensively studied thanks to its importance in natural products synthesis and because it is easily tractable by simple semi-empirical methods. It is the typical test reaction for a semi-empirical quantum calculation methods such as ours, and furthermore for the numerous approaches of reaction description [31]. It involves the reaction between two linear π -systems of length 2 and 4, called *dienophiles* and *dienes*, and is thus called a [2+4]-cycloaddition. The product is again a dienophile and may react again in a Diels-Alder reaction. The reaction is used for the synthesis of polymers [32]. The generation of a repetitive Diels-Alder (DA) reaction network and a further network, the *Formose reaction*, is discussed in detail in [8]. Here their properties and those of DA networks with different reactivity thresholds are studied. The reactivity threshold determines which reactions are allowed, i.e. those above it.



Fig. 4. Evolution of network properties for a repetitive Diels-Alder CRN with different reaction rate thresholds (solid) and the AB (slashed) and ER (dotted lines) model. The AB graphs were only generated in the region of $n \in [89, 246]$, where $\langle k \rangle \approx 5$. (a) Reaction rate threshold (grey) and $\langle k \rangle$ (black); (b) Cycle statistics of triangles and cycles of size 4, 5, and 6; (c) $\langle L \rangle$; and (d) $\langle C \rangle$ vs. networks size.

The series of CRN in Fig. 3 is obtained by repetitive Diels-Alder reactions of a simple initial mixture of dienes and dienophiles with a decreasing threshold value.

Tab. 1 compares the network characteristics of the Diels-Alder, the Formose, and the *E. coli* metabolic network. They are all sparse graphs, i.e. they have much fewer edges than complete graphs, reflected by $m \ll \frac{n(n-1)}{2}$ or $\langle k \rangle \ll n$. Sparse networks are very common, ranging from the network of acquaintances to a neural network. In both cases, there are only few connections at each node. From the networks of Tab. 1, only Diels-Alder and *E. coli* fulfill the conditions $\langle C \rangle \gg \langle C_{rand} \rangle$ and $\langle L \rangle \leq \langle L_{rand} \rangle$ and thus are small-world networks in the strict sense.

The change of network characteristics with DA network size is shown in Fig. 4. Interestingly, both $\langle C \rangle$ and $\langle L \rangle$ stay constant for $n \in [89, 246]$. This property seems to be robust in chemical networks. This is not the case for ER and AB graphs as both $\langle C \rangle$ and $\langle L \rangle$ vary. Triangles dominate the minimum cycle basis of the substrate graph representation of a CRNs, see Fig. 4. For comparison we show the cycle distribution of scalefree graphs (AB model) with the same size



Fig. 5. Rank statistics of repetitive Diels-Alder (l.h.s.) and the Formose reaction network (r.h.s.). Datapoints are ranked by decreasing degree. The power-law regressions (solid lines) have slopes of -1.19 in both cases. The Poisson distribution expected for ER random graphs is shown for comparison as a dotted line.

and average degree. The number of triangles is $\Delta \approx \mu = m - n + 1 \approx n(\langle k \rangle - 2)/2$ for large *n*. So a linear rise of Δ with the network size is not surprising. As shown in Fig. 4, the fraction Δ/μ , however, also seems to be constant for increasing network size.

Finally, the degree distributions have been calculated (fig. 5). Both networks are scale-free, i.e. their degree distributions follow a power law. The cumulative representation of fig. 5 is equivalent to $\int_k^\infty P(x) dx$ vs. k. The regression $\int_{k}^{\infty} P(x) dx \sim k^{-1.19}$ is consistent with the values reported in [33]. The explanation of the origin of scale-freeness therein can be applied to the present examples. It relies on two generic mechanisms. First, the networks grows from an initial set of nodes by continuous addition of new nodes. Indeed, in the present case, there is an initial list of molecules, and the networks is built by adding molecules at every iteration (sect. 4). Second, the networks grows by *preferential* attachment, i.e. new nodes are preferably attached to nodes with high degree, or in the case of molecules, species who have already spawned many new other species are especially reactive and more likely to produce new molecules at each iteration. The power-law regression for the Formose reaction network fails for high k. The theoretical Poisson distribution for ER random graphs with high n, $P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$, also fails in this range. A degree distribution following a truncated power law has been referred to as broad-scale [34]. Again, the exponent of the power law is constantly around 1.18 ± 0.05 for the repetitive Diels-Alder reaction network for sizes $n \in [39, 246]$.

The Formose reaction network, in contrast, does not fit the predictions of from scale-free models very well, Fig. 5. It includes many non-reactive species with respect to keto-enol condensation, which leads to small cliquishness and longer paths. Furthermore, it has particularly reactive species (formaldehyde) whose neighbors in the substrate graph do not react with each other: a keto-enol isomerization must toggle the "state" of a network species from non-reactive to reactive first. This leads in consequence to an elongation of paths and a reduction of cliquishness of the Formose reaction network.

6 Discussion

The study of Artificial Chemistries (AC) is closely related to the field of Artificial Life. The many models of AC can be categorized according to abstraction level and intended application. Very abstract models simulate AC in which strings or logical elements interact, such as in Fontana's the λ -calculus. There, networks are built in a bottom-up approach to be studied phenomenologically, as in Artificial Life, for example for the emergence of structures or sustainability. On the other hand, models for practical applications tend to be top-down. They concentrate on describing interactions and try to predict, for example, the time and/or space evolution of concentrations.

The graph based Toy Chemistry Model possesses qualities from both approaches. The use of graph grammars to handle reactions makes our approach more bottom-up like, while the use of a simplified quantum mechanical energy calculation is clearely top-down. The graph based Toy Chemistry Model provides a very flexible and transparent framework for the exploration of the properties of chemical reaction networks. The built-in chemical thermodynamics makes the model self-consistent and helps to keep the resulting reaction networks close to chemical "reality". Since the used graph grammar (set of allowed reactions) can be chosen freely, a wide variety of reaction networks with different "chemical flavors" can easily be generated, which is an important prerequisite for the study of their network properties.

We found that the Diels-Alder reaction network exhibits small world properties similar to known cases e.g. the *E. coli* metabolic network. The Formose reaction network, on the other hand, does not fit the prediction for a small world network very well. We conclude that chemical networks do not fall into a single class of the small-world network classification scheme by Amaral *et al.* [34]. More detailed investigations will therefore be necessary.

References

- Fell, D.: Understanding the Control of Metabolism. Number 2 in Frontiers in Metabolism. Portland Press, London (1997)
- Yung, Y.L., DeMore, W.B.: Photochemistry of Planetary Athmospheres. Oxford University Press, New York (1999)
- Höllering, R., Gasteiger, J., Steinhauer, L., Schulz, K., Herwig, A.: The simulation of organic reactions: From the degradation of chemicals to combinatorial synthesis. J. Chem. Inf. Comput. Sci. 40 (2000) 482–494
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.: The large-scale organization of metabolic networks. Nature 407 (2000) 651–654

- 5. Wagner, A., Fell, D.A.: The small world inside large metabolic networks. Technical Report 00-07-041, Santa Fe Institute (2000)
- Fell, D.A., Wagner, A.: Structural properties of metabolic networks: implications for evolution and modelling of metabolism. In Hofmeyr, J.H.S., Rohwer, J.M., Snoep, J.L., eds.: Animating the cellular map, Stellenbosch, ZA, Stellenbosch University Press (2000) 79–85
- Gleiss, P.M., Stadler, P.F., Wagner, A., Fell, D.A.: Relevant cycles in chemical reaction network. Adv. Complex Syst. 4 (2001) 207–226
- Benkö, G., Flamm, C., Stadler, P.F.: A graph-based toy model of chemistry. J. Chem. Inf. Comput. Sci. (2003) in press; presented at *MCC 2002*, Dubrovnik CRO, June 2002; SFI # 02-09-045.
- Atkins, P., Friedman, R.: Molecular Quantum Mechanics. 3rd ed. edn. Oxford University Press (1997)
- Fontana, W., Buss, L.W.: What would be conserved if 'the tape were played twice'? Proc. Natl. Acad. Sci. USA 91 (1994) 757–761
- Dittrich, P., Ziegler, J., Banzhaf, W.: Artificial chemistries a review. Artificial Life 7 (2001) 225–275
- Hoffmann, R.: An Extended Hückel Theory. I. Hydrocarbons. J. Chem. Phys. 39 (1963) 1397–1412
- 13. Polansky, O.E.: Graphs in quantum chemistry. MATCH 1 (1975) 183-195
- Gillespie, R.J., Nyholm, R.S.: Inorganic Stereochemistry. Quart. Rev. Chem. Soc. 11 (1957) 339–380
- Nagl, M.: Graph-Grammatiken, Theorie, Implementierung, Anwendung. Vieweg, Braunschweig (1979)
- Dugundji, J., Ugi, I.: Theory of the be- and r-matrices. Top. Curr. Chem. 39 (1973) 19–29
- Ugi, I., Stein, N., Knauer, M., Gruber, B., Bley, K., Weidinger, R.: New elements in the representation of the logical structure of chemistry by qualitative mathematical models and corresponding data structures. Top. Curr. Chem. 166 (1993) 199–233
- Fleming, I.: Frontier Orbitals and Organic Chemical Reactions. Wiley: New York (1976)
- Bonchev, D., Mekenyan, O., eds.: Graph-Theoretical Approaches to Chemical Reactivity. Kluwer, Dordrecht, NL (1994)
- Zeigarnik, A.V.: On hypercycles and hypercircuits in hypergraphs. In Hansen, P., Fowler, P.W., Zheng, M., eds.: Discrete Mathematical Chemistry. Volume 51 of DIMACS series in discrete mathematics and theoretical computer science., Providence, RI, American Mathematical Society (2000) 377–383
- Balandin, A.A.: Multiplet theory of catalysis: Theory of hydrogenation. Classification of organic catalytic reactions. Algebra applied to strucutral chemistry. Volume 3. Moscow State Univ. (1970) in Russian.
- Zeigarnik, A.V., Temkin, O.N.: A graph-theoretic model of complex reaction mechanisms: bipartite graphs and the stoichiometry of complex reactions. Kinet. Catal. Engl. Transl. 35 (1994) 691–701
- Fell, D., Wagner, A.: The small world inside metabolic networks. Proc. R. Soc. Lond. Ser. B 268 (2001) 1803–1810
- 24. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press (1994)
- Newman, M.E.J., Moore, C., Watts, D.J.: Mean-field solution of the small-world network model. Phys. Rev. Lett. 84 (2000) 3201–3204
- Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393 (1998) 440–442

10

- Barabási, A.L., Albert, R., Jeong, H.: Mean-field theory for scale-free random networks. Physica A 173-187 (1999) 272
- 28. Bollobás, B., Riordan, O.: The diameter of a scalefree random graph. Technical report (2002) Preprint.
- Chickering, D.M., Geiger, D., Heckerman, D.: On finding a cycle basis of with a shortest maximal cycle. Inform. Processing Let. 54 (1994) 55–58
- Diels, O., Alder, K.: Synthesen in der hydroaromatischen Reihe. Liebigs Ann. Chem. 460 (1928) 98–122
- 31. Temkin, O.N., Zeigarnik, A.V., Bonchev, D.: Chemical Reaction Networks : a graph-theoretical approach. CRC Press Boca Raton, FL USA (1996)
- Morgenroth, F., Müllen, K.: Dendritic and hyperbranched polyphenylenes via a simple Diels-Alder route. Tetrahedron 53 (1997) 15349–15366
- Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science 286 (1999) 509–512
- Amaral, L.A.N., Scala, A., Barthelemy, M., Stanley, H.E.: Classes of small-world networks. Proc. Nat. Acad. Sci. 97 (2000) 11149–11152