Evolution and dynamics of chemical reaction networks

Group project report, CSSS 2006

Authors: Rogier Braakman[1, *], Jess Mar[2, *], Anirban Banerjee[3], Shahram Nosrati[4]

[1] *California Institute of Technology, Div. of Chemistry, Pasadena, CA, USA* [2] *Harvard University, Dept. of Biostatistics, Boston, MA, USA* [3] *Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany* [4] *Tehran Polytechnic University, Tehran, Iran*

[*]Corresponding authors: rogier@caltech.edu, jmar@hsph.harvard.edu

Abstract

Complexity underpins the very nature of known chemical interactions. In this paper, we offer a simulation-based model that mimics the behavior of a generic chemical reaction network. Using ideas from complex systems, we seed and evolve this reaction network with the view to study its dynamics and changing topology over time. Our simulation model allows for two sets of scenarios - the most minimal assumption of additive reactions between molecules only and the more complicated, more realistic assumption that allows for branching reactions to take place. Our results show behavior that we would expect – given a constant inflow of generic nutrients, after time eventually a steady-state system is realized. In the scenario where branching is permitted, the time to steady-state is delayed slightly due to the presence of apparent slower dynamics. This model can now be used to some of the basic assumptions of the model as well the effects of outside perturbation and other physical to the system or its environment.

Introduction

When a set of reactions share common components, this can inherently be described as a network. Looking at chemical reaction networks (CRN) in nature, the biochemistry of living

TABLE I – *Network characteristics: V = vertices; E = edges; $\langle K\rangle$ = mean degree; $\langle L\rangle$ = averaged shortest path; $\langle C\rangle$ = mean clustering; r = assortativity. See text for more details. ISM: Interstellar medium; HC: hydrocarbon chemical network of the giant planets.*

	V	E	$\langle K\rangle$	$\langle L\rangle$	$\langle L_{\rm rand}\rangle$	$\langle C\rangle$	$\langle C_{\rm rand}\rangle$	r	$r_{\rm rand}$	Modular
Earth	248	778	6.27	2.75	3.20	0.31	0.025	−0.31	−0.006	YES
Mars	31	144	9.29	1.89	1.73	0.61	0.31	−0.10	−0.007	NO
Titan	71	396	11.16	2.08	1.98	0.55	0.16	−0.17	−0.03	NO
Venus	42	175	8.33	2.07	1.94	0.59	0.20	−0.14	−0.06	NO
HC	39	270	13.85	1.65	1.64	0.68	0.37	−0.26	−0.06	NO
ISM	400	6102	30.51	1.99	2.01	0.52	0.07	−0.24	−0.006	NO
E. coli	741	2310	6.24	3.02	3.82	0.183	0.008	−0.17	0.004	YES

organisms, the atmospheric processes of a planetary atmosphere and the gas-dust pathways in interstellar chemistry, these all have very different characteristics and topologies. Solé and Munteanu (2004, henceforth SM) studied the topology of CRNs in a variety of these regimes, focusing on the clustering coefficient, average shortest path, degree distribution, assortativity and modularity, and compared these to a randomly generated Erdos-Renyi (ER) graph (1953). Their results are shown in table 1 (taken directly from SM). Several things can be seen from the data in this table. One of the most interesting results is the similarity between the overall CRN topologies of the Earth's atmosphere and the biochemical networks in *E.-Coli* bacteria. Both have a very similar network density, indicated by mean degree, $\langle K\rangle$, and averaged shortest path, $\langle L\rangle$, and are significantly sparser than all other networks. This is a strong indication of a coupling between the two CRNs (as suggested by Lovelock, 1988, 2003). Further comparisons shows that all CRN have small world characteristics, i.e. $\langle L\rangle \approx \langle L_{\rm rand}\rangle$ and $\langle C\rangle >> \langle C_{\rm rand}\rangle$, but the ratio of $\langle C\rangle$ / $\langle C_{\rm rand}\rangle$ is highest for the *E. Coli* network followed by Earths atmosphere. In all, these and other patterns in the data suggest that the topology of a CRN evolves over time, from dense highly connected graphs, more similar to an ER random graph in the interstellar medium and planetary atmosphere to sparser and more clustered networks with stronger small world characteristics in living organisms and Earths atmosphere. Second, the presence of life seems to play a crucial role in

driving the CRN away from an equilibrium set of reactions and towards a more hierachical topology (SM).

Much remains to be understood, however, in the evolution of reaction networks and metabolic networks. The patterns discussed above beckon some interesting questions. What is the driving force behind the change in topology; is it is simply the change in density and temperatures, and hence external? If this were the case it should be possible to develop an abstract, purely mathematical model to study the dynamics and evolution of such networks. If however, the appearance of new nodes with unexpected properties (for example as the result of the appearance of life) plays a crucial role, it might be unavoidable to add chemical reality into any such models. A somewhat deeper question then arises: Under what circumstances do new properties emerge in the network, and how does this relate to the topology? For example the appearance of highly connected nodes (hubs) and transitions between different degree distribution regimes, or the appearance of components with catalytic properties or that act as enzymes. To study questions like these a model was developed of a chemical reaction network as well network analysis tools to study its topology and dynamics

Group Project

The goal was to develop a simple, general chemical reaction network with minimal assumptions, and to study the effects of perturbations and other dynamics in the network. The behavior of the network can be followed in time, allowing the study of dynamics in the time domain. This is important to get a better understanding of how network switch between different regimes. Due to time constraints, the study of such dynamics in detail was not possible, but they will be discussed

in some detail under indication of future work. First, the model and the initial results obtained will be discussed.

Model

Reaction components X, Y and Z are the starting point of the model. These components can react with each other and with all products that are formed, according to the following reaction,

$$I + J \rightarrow IJ,$$

The rate of this reaction is given by the simple relationship:

$$R = r_{i,,j} = k_{i,,j} \cdot [\, i \,] \cdot [\, j \,]$$

Here, $r_{i,,j}$ is the reaction rate between *i* and *j* and gives the fraction of both components reacting away or the amount of product *ij* being formed. $k_{i,,j}$ is the reaction rate constant, or the probability of reaction once *i* and *j* meet. This meeting probability is in turn directly the product of the concentrations of *i* and *j*. All calculations can be done in a purely probabilistic manner, while both *R* and *K* can be written as a matrix. The values for $k_{i,j}$ are generated randomly by sampling from a uniform distribution between 0 and 1. Chemically this is not the most realistic method, but it is simple and for construction purposes will suffice. There is a constant inflow of 'nutrients' (molecules X, Y and Z), driving the dynamics of the basic network, and outflow is treated a simplistic manner, in which only molecules below size N are allowed to react. This can simply be done by setting rate constants *k* between those molecules and any others to zero. Thus, those molecules can be considered to be flowing out of the system (e.g. by condensation to solid phase,

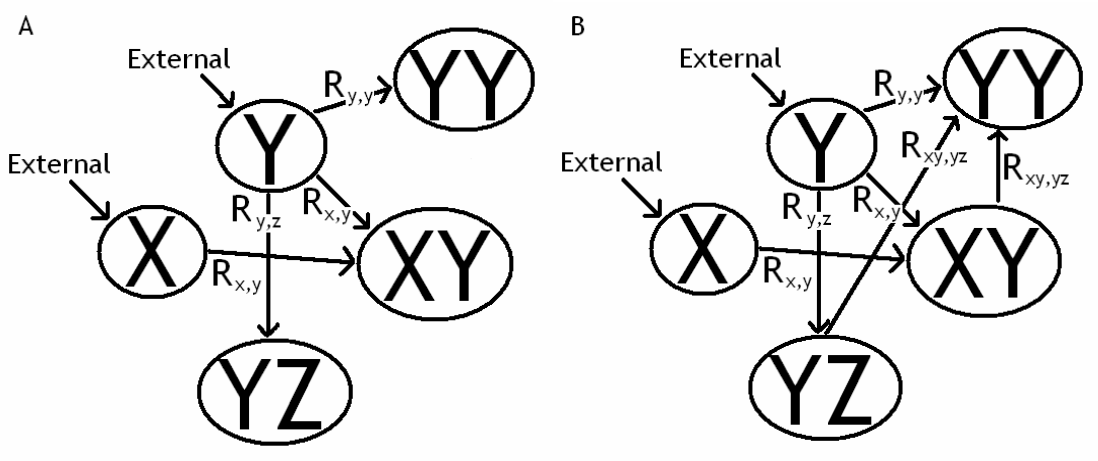

Figure 1. The two generations of the model. (A) Addition reactions only. (B) Including transfer reactions.

growth of biomass, etc.). With this construction, the systems should reach steady state once total inflow = total outflow. The final assumption in the basic model is that isomerism is ignored (i.e. XYX = XXY), which is simply to save computational cost.

To analyze the results of the model in terms of complex network theory, an additional matrix needs to be constructed, a connectivity matrix that gives information on the edges between different nodes. The nodes in this case are simply the different molecules in the reaction system and the edges are the flow of matter from one node to the other. The flow of matter from one node to another is given by the reaction rate, r, for the reaction involving the first node that leads to the second node. Summing all reaction rates between two nodes gives the total flow of matter, $f_{i \to ij}$, between the nodes. This gives us the flow matrix $F = f_{i \to ij}$, which is also the connectivity matrix. However, since all molecules can in principle react with each other, this gives a highly connected graph, whose connectivity does also not change over time. Only the weight of the edges, due to variations in the concentrations of reactants, varies over time. Thus to study dynamics on this network it needs to be treated as a weighted graph.

Two generations of the model were developed. In the first, only addition reactions are allowed in which the reactants are combined into one reaction product (ie. XX + YZ → XXYZ). This means that there is only one reaction contributing to the flow from one node to another, and flow is always unidirectional along an edge. However, in a study of the evolution of connectivity in metabolic networks Pfeiffer et al. (2006) found that group transfer reactions are essential for the emergence of hubs. Their study was a more complex model, involving enzymes, genes coding for enzymes and selection pressures. Nonetheless it shows the importance of including group transfer reactions. It is also more chemically realistic that reaction pathways are allowed other then simply adding molecules together. Therefore a second generation of the model was developed in which ALL reactions are in principle possible (i.e. XX + YZ → XXYZ, XXY + Z, XXZ +Y, XY +XZ, etc.). Realistically, however, only a few pathways actually participate in any reaction. Thus, the number of possible reactions was capped at 3, with half of reactions branching according to a 70:20:10 ratio and the other half according to a 40:30:30 ratio in cases where 3 or more paths are possible. In the case where only two paths are possible, the branching ratio is 60:40. Which pathways contribute as well as the ratio of the paths is chosen randomly. In this version of the model, several reactions can contribute to a single edge, and edges can now also be bi-directional.

Results

Initial results of both generation 1 (G1) and generation 2 (G2) models are shown in figure 2. The two models show different dynamics when evolving over time. Both reach steady state, but the dynamics of G2 appear to be somewhat slower and it reaches steady state at later times than G1. Further test also indicate that a lower total input into the model delays reaching steady state in both models. Both results are in accordance with expectations. One of the main reasons that G2

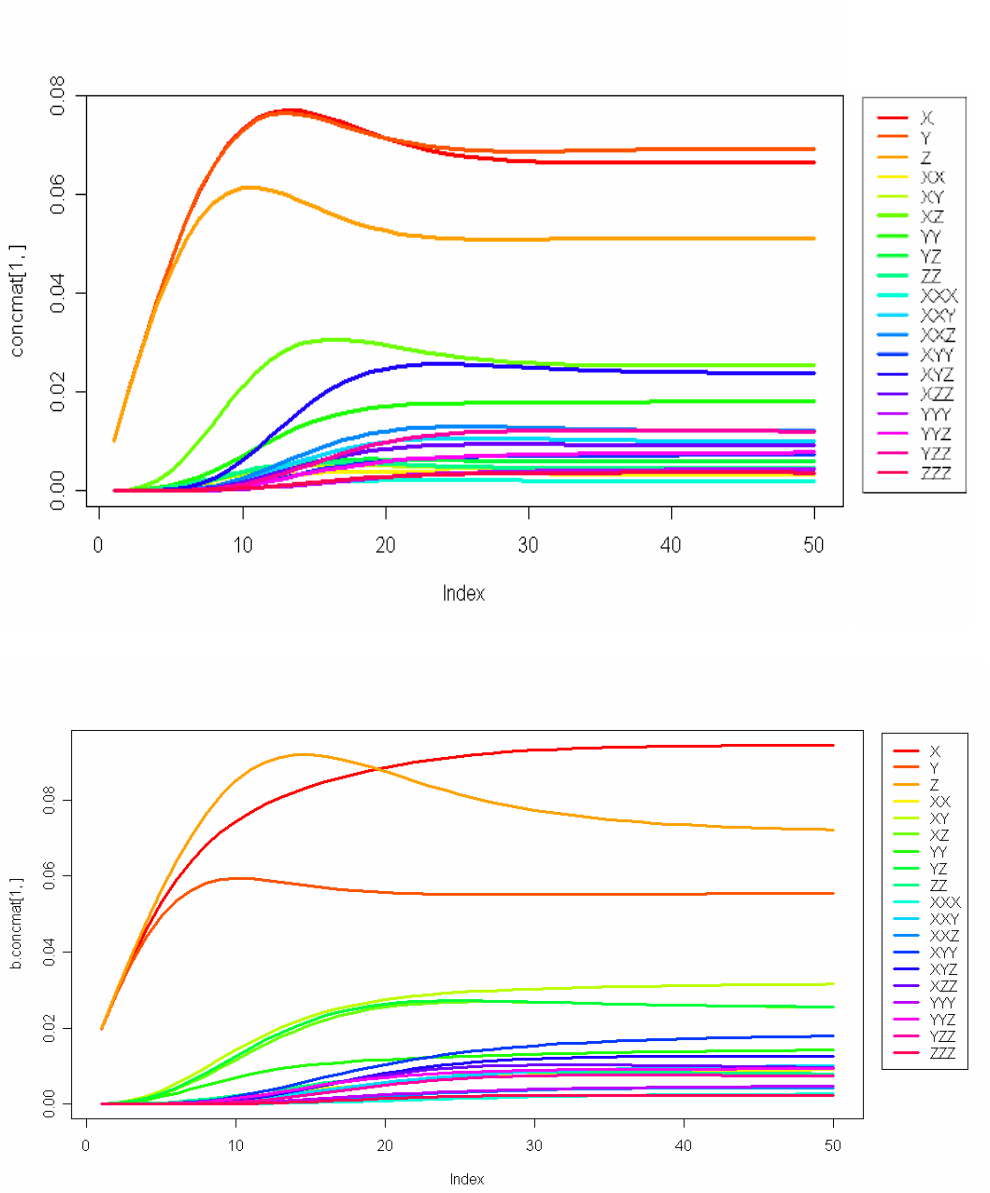

Figure 2. The concentrations of the different reactants over time. Top: Generation 1 of the model. Bottom: Generation 2 of the model

reaches steady state slower than does G1 is that the branching pathways create the presence of feedback loops in the model, which should result in more accumulation of matter in the network. This indeed appears to be the case.

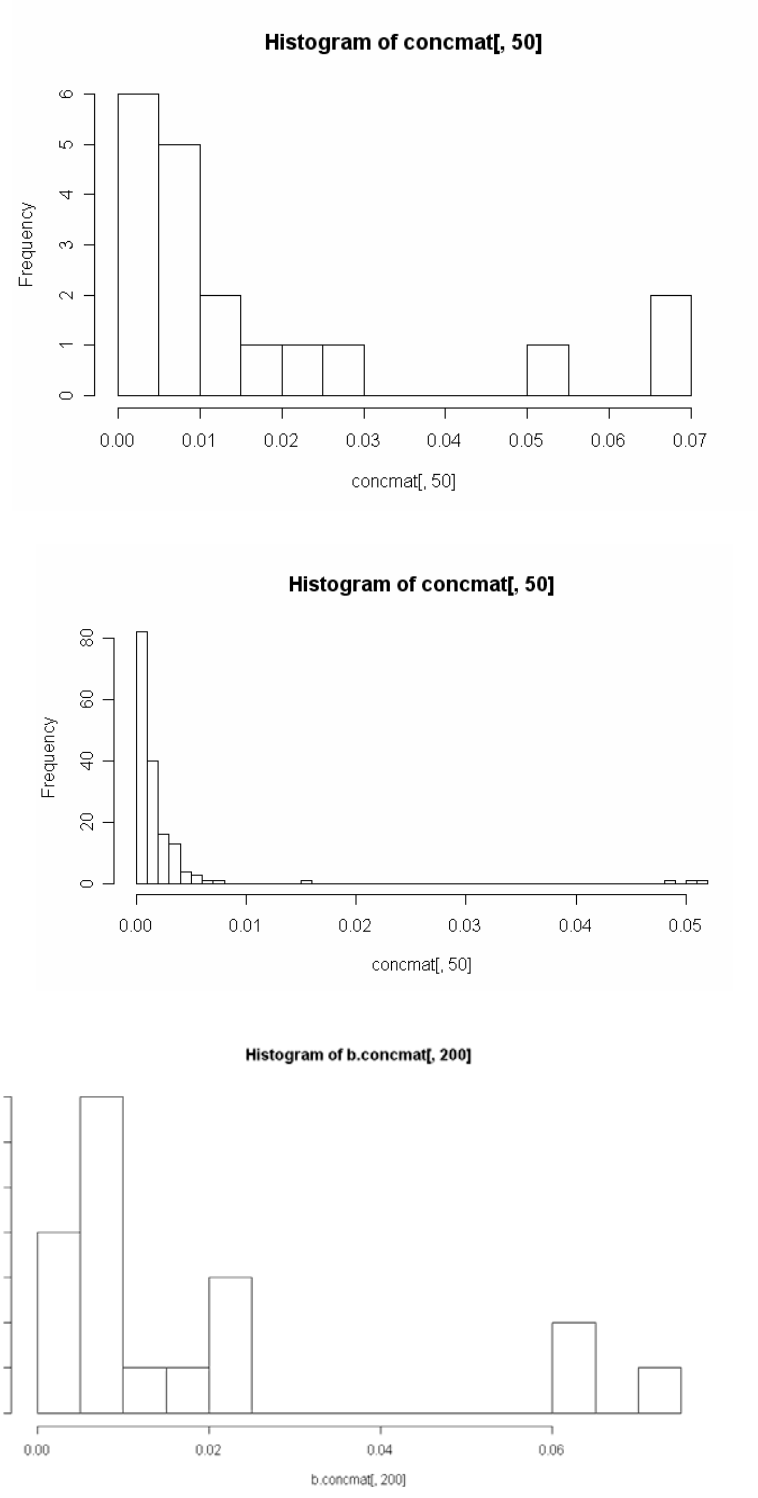

Figure 3. The distribution of reactant concentrations for both models, different sizes. (Top) G1, N = 3. (Center) G1, N = 8. (Bottom) G2, N = 3

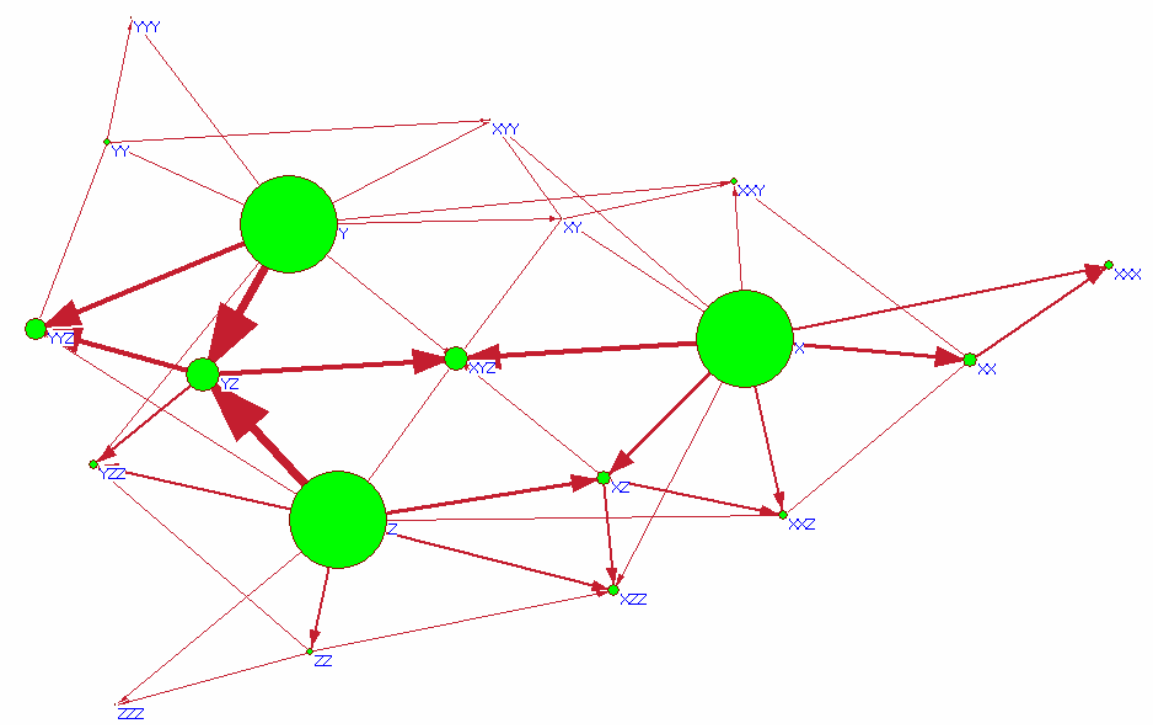

Figure 4. Graph representation of reaction network, G1. Size of nodes indicates concentrations of that molecules, the intensity of the edge indicates the amount of matter flowing between two nodes.

Figure 3 shows the concentration distribution for the different reactants in both models. The result for G2 shows more molecules with higher concentrations than G1. This point could be made conclusive simply by going to a larger system size. This is indicated by the concentration distribution for N = 8 in G1, also shown in figure 3. The increase from N = 3 to N = 8 results in a much smoother distribution, with the hint of an exponential shape. Thus it would be useful to run G2 at this size as well and compare the two results.

Finally, in figure 4, a graph representation of the G1 network is shown. It is expected that the concentrations of X, Y and Z should be highest because of the constant inflow. The flow of matter in the network can be followed very clearly along the different paths and it is obvious which edges contribute more significantly to the matter flow-through. A time series of such

graphs can easily be constructed, which gives a visual perspective on network dynamics over time.

Discussions and future work

A basic model for a generic chemical reaction network was constructed and tested for both generations of the model and both behave as expected. Due to time constraints further analysis and experiments were not conducted, but this basic model should allow for such tests in the future. Several things should be considered before further testing and experiments. One is the graphical representation used to study the network. The directed graph representation in the current model was chosen because it allows direct study of the flow of matter between nodes and thus the dynamics on the network. When picking any representation though, some information is inherently lost. Another option, for example, would be to use to a bipartite graph (e.g. Newman 2003), in which there are two types of nodes: molecules and reactions. The downside to using this method in our initial model is that the number of reactions is very large compared to the number of molecules, which makes it unclear if useful knowledge will be gained, but it might be useful for later generations with some of the alterations discussed below. Another representation commonly used in studying the topology of chemical reaction networks is the *hypergraph* (e.g. Flamm & Stadler, 2006).

Because the connectivity of the graph also does not change over time, some thought needs to be given into how to use traditional complex network analysis tools, such as average shortest path, clustering coefficient and mean degree and degree distribution of nodes. One option would be to simply choose a cutoff below which the weight of an edge is considered to be zero (not present). This can be done in a physically realistic manner by choosing a finite number of total molecules

and rounding down if the flow along an edge is less than a molecule. Another option is to simply take at the total flow into or out of a node and look at their 'binned' distribution.

Next, the chemical and physical reality of the model should be considered in more detail. In particular the method for generating the rate constants, k, should be reconsidered. Currently the values are randomly generated between 0 and 1, which gives an easy and reasonable first model, but it is not very realistic that this distribution should be uniform. Not only is it unlikely that all reaction rates should be equally represented, but there is also no conservation of reactivity for different species. It should be possible for certain species to be inherently reactive regardless of its reaction partner. And it should also be possible for unexpected properties to emerge when new molecules are formed. Perhaps this could be achieved by sampling from a bell shaped distribution, while simultaneously having a distribution around an initially picked value for similar species, as well as an increasing probability for the appearance of new properties as the size increases. Alternatively, basing the model on actual chemical data is another alternative.

Furthermore, the current outflow mechanism should be improved as well. Simply having all molecules above a certain size be non-reactive is both crude and unrealistic. A better way would be to create a more closed network, in which less large molecules are formed (biasing towards transfer reactions for larger molecules), and the outflow is more clearly specified. If the total number of molecules, n, is chosen to be a finite number, it would also be possible to truncate probabilities to zero at some value at which less than a single molecule is formed. Combined these different steps should result in a model that is more chemically and physically realistic.

At this point several interesting dynamical effects on the network could be tested, starting with the effect of external perturbations. The effect of fluctuations in nutrients can be monitored as it spreads through the network: sudden increase or decrease in concentration of nutrients,

continuous variation of input, appearance of new nutrients, or toxins binding specific molecules or inhibiting certain reactions, all these effects would drive a change in the topology of the network.

Other changes in the physical environment would be interesting to study as well. In most discussions about living organisms or the emergence of life, the presence of a membrane in which the biochemistry can be isolated from its surroundings is considered a crucial prerequisite (Ref…) With the model in this paper it would be possible to test how large the impact of constraining the molecules on the topology of the network is. Diffusion can be added to the model and the rate can be varied to mimic appearance of or damage to a membrane or simply being constrained on the surface of some mineral or some other non-biological source. The effect of temperature can also easily be added into the model, specifically by adding temperature dependence to the rate constants k

The final thing to consider in the model, the final step in going from interstellar non-biological networks to those of living organisms, would be to add in some information storage mechanism (genes) into the model that partially preserves the chemical information in the system between different generations. If these genes are then allowed to vary, or mutate, and certain molecules in the network are selected as biomass, it would be possible to study the effects of biological evolution on the topology of the network. Pfeiffer et al (2005) used a similar method.

Acknowledgements

The authors would like to acknowledge the Santa Fe Institute, the Chinese Academy of Sciences, Institute of Theoretical and the CSSS 2006 staff for organizing the summer school and giving the

opportunity to attend. We would also like to thank the other participants and lecturers, in particular Andreas, Dave, Cosma and Raissa for fruitful discussions.

References

1. Solé R. & Munteanu A., **2004**, *Europhys. Lett. 68 (2)*, p. 170-176
2. Watts D.J. & Strogratz S.H., **1998**, *Nature 393*, p. 440
3. Erdős P. & Rényi A., **1959**, *Publ. Math. (Debrecen) 6*, p. 290
4. Lovelock J., **1988**, *The Ages of Gaia* (Norton and Co. New York)
5. Lovelock J. **2003**, *Nature 426*, p. 769
6. Pfeiffer T., Soyer O.S., Bonhoeffer S., **2005**, *PLoS Biol. 3 (7)*, e228
7. Newman M. E. J., **2003**, *SIAM Review 45 (2)*, p. 167-256
8. Flamm C., stadler P.F., **2006**, *Manuscript.*.