Molecularisation of the RAF [MoRAF]

Autocatalytic sets in a Toy Chemistry

10.12.2015

Motivation and Background

The origin of life (OoL) on earth is still in search of a satisfactory explanation. The field is dominated by many facets and partial explanations. Facets include frequency of planetary systems, climatology of early earth, chirality in naturally occurring compounds, abiotic production of molecules, etc. Partial explanations include the naturally occurring self-reproducing molecules, quasispecies and hypercycles, an RNA world, natural formation of micelles, etc. However, most of these explanations are incomplete or based on speculation.

Many other fields in the biosciences have benefited from the introduction of formal models, which forced researchers to be explicit about assumptions made, and allowed mathematical reasoning to be applied and computational experiments to be performed. Such models have been introduced in the context of the OoL, but research in them has not been very dominant so far. However, as OoL research gains pace, they will be given more attention. Examples of formal models related to OoL are Conway [Game of Life], von Neumann (1967), Ganti (1997), Kauffman (1986), and Steel (2000). For formal models to be useful they should capture some essence of the empirical problem and as time passes they should be forced towards increasingly realistic descriptions of the phenomena. The formalisation of catalytic reaction systems by Steel (2000), based on an initial idea by Kauffman, consists of

- a set of molecule types;
- a set of reactions where each reaction converts one set of molecules (reactants) into another set (products);
- a set of catalysations: molecules that accelerate a reaction (or set of reactions);
- a food set: a small set of molecules assumed to be freely available and constantly replenished.

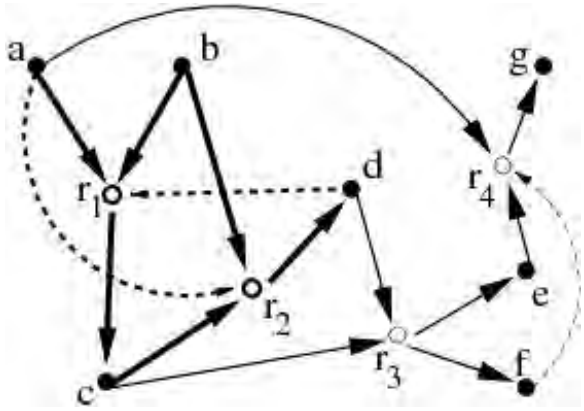


Figure 1: An example of a catalytic reaction system (CRS). The subset of reactions {r1,r2} (shown in bold) is an RAF set.

The questions of interest in these models are conditions for the appearance of (sub)sets of molecules/reactions that are self-sustainable: each reaction in the set is catalysed by at least one molecule from the set, and each molecule can be created, starting from the food set, by repeated reactions from the same set. This idea of *autocatalytic sets* was introduced in Kauffman (1986), and formalised as *RAF sets* and subsequently studied more extensively in Steel (2000), Hordijk and Steel (2004), and Mossel and Steel (2005). Investigations into these models represent significant progress relative to less precise models, but the Steel model needs elaboration to be more realistic in addressing the probability of spontaneous occurrences of RAF sets.

The concept of RAF is not the final formal framework for studying the Origins of Life computationally: It lacks individuality (no cell), it doesn't evolve, it is presently Boolean (no concentrations) and it is an abstract theory with no reference to real molecules. But it is major step forward relative to earlier lose speculations.

In describing the behavior of a set of molecules, know which reactions can take place and which molecules within the set would accelerate the reactions are clearly central. Typically, in systems biochemistry, metabolic control analysis and kinetic analysis, these are assumed known. However, in reality they are not known and answering them is highly relevant as having a set of molecules and wanting to predict the changes in concentrations is central to both systems chemistry, origin of life research and beyond – the equations of systems biology are often taken from "established biological knowledge" that is bound to contain many errors.

The motivation for this problems comes from "systems chemistry" (Hunt and Otto, 2011) and formal models of origin of life [Steel, 2000], where it would be a major advantage if observed molecules, reactions and catalysis could be computed as these in real systems will consist of multiply interacting molecules that is not pure or controlled by an experimenter.

The OL-problem is at the at the sub-25 atom level!! The nucleotide adenine [with phosphate and sugar] has 26 atoms and 19 if excluding the hydrogens. Small molecules have been exhaustively enumerated up to 17 non-hydrogen atoms (and there were 166 [American] Billion) by Reymond (2012). Enumeration is not so much in itself but illustrates that one can soon have a list of ALL basic building blocks. It also illustrates that you only get to the nucleotides when you have listed many American trillions of smaller molecules, so assuming life started in a soup of suitable monomers is a tall assumptions. There is a strong inclination in computational OoL research to assume that we can start in a pure monomer soup. Kauffman (1986) and several of the more recent Steel and Hordijk papers assume this, but that is taking one huge convenient step that skips most of the hard bit.

Any description of the behaviour of many different molecules in a solution is bound to make strong approximation and then add realism as predictions fail.

What must be done to molecularise the RAF?

The abstract models of von Neumann [Universal Replicator], Ganti [Chemoton] and Steel [RAF] has the advantage that it forces you to exact reasoning, but so far only the Chemoton is cell and molecularisation of the Chemoton would certainly be worthwhile, but I think considerably more difficult than molecularisation of the RAF, since one would have to define how molecules made a cell and as it is defined presently, the Chemoton doesn't replicate it only perpetuates its own existence. So starting with the RAF is the natural starting point.

The three problems – molecule enumeration [or defining a universe of molecules], reaction prediction and catalysis prediction – are very different. The first has a more than hundred year history, the second decades, while that last is only started to be studied.

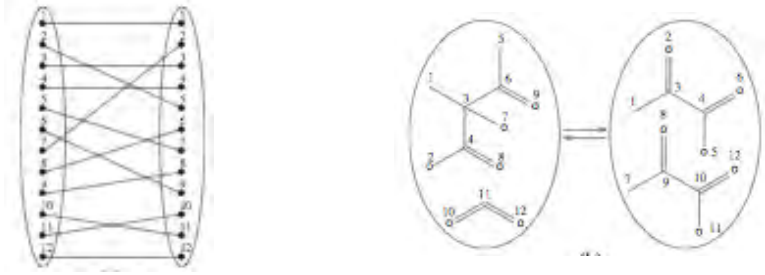
A. Much of modern combinatorics was initiated by attempts to ***count molecules***. Cayley counted saturated hydrocarbons [alkanes] in mid-19th century, which happens also to be equivalent to counting the shapes of phylogenies in biology. Polya (1937) extended the ability to count molecules considerably by the use of group theory to be able to take account of symmetries. Researchers like Leslie Ann Goldberg (Oxford) and Jean-Loup Faulon published on computational aspects of Polya-enumeration with a view toward molecules. Exhaustive listing becomes unrealistic already for 15-20 atoms and a way of investigating molecular systems with more atoms is to use algorithms that samples in this much larger set.

In the last decades molecule enumeration has been used to screen virtual drug targets computationally (Blum, van Deursen and Reymond, 2011). Molecule enumeration views a molecule as a node labeled graph. Viewing a molecule as a 3D entity will introduce chirality of certain atoms, when a molecule is different from its mirror image. Additionally, a large set of molecule graphs are hard to realize as physical structures due to steric collision, making them energetically unfavourable. Since, a molecule with a sub-molecule that energetically unfavourable, is itself energetically unfavourable, this lends itself to a branch-and-bound traversal of the set of molecules.

Many problems remains open in this line of research and numbers become very large as a function of number of atoms, so imposing restrictions on the set of relevant molecules is important to become a realistic tool. Andronico et al. (2011) discussed a variety of methods to predict the 3D structure from the molecular graph, where using databases of known small molecule structures to extract constraints on fragments of the molecules seemed to be the most most promising approach.

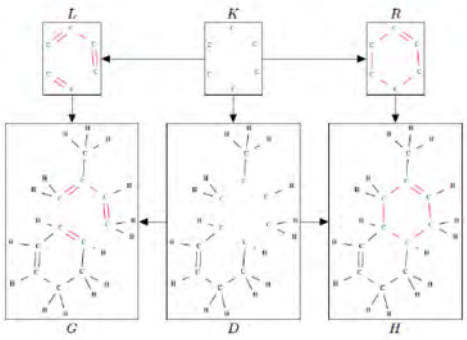
B. Most ***reactions*** are either fission/fusion of molecules or swapping of groups. This is so for the simple reason that more that two molecules colliding in space is unlikely and we will restrict ourselves to these case. Higher order reactions with more molecules reacting simultaneously can be explained as a string of reactions only involving pairs of molecules. To investigate if A+B→C or A+B→C+D is possible suggests a natural approach: Are **A** and **B** sub-graphs of ***C*** [or ***C*** and ***D***] such that they don't overlap and that they cover ***C*** [or ***C*** and ***D***]?

Kayala et al. (2011) discus machine learning approaches to predict reactions incorporating various levels of chemical realism in the form of molecular orbital information.



To the left which atoms in A, B could map to which atoms in C,D. The two covering disjoint sub-sets from A,B is found that maps into two covering disjoint sub-sets in C,D that preserves subgraphs of the sub-sets (From Boyer and Viari)

Graph Grammars is an extremely convenient framework for representing molecules and reactions as graphs and graph edit operations. Thus 3D and energies are ignored. The Double Pushout Automaton seems the best formalism for doing this. A graph [read molecule] (K) is used for embedding or identifying atom by atom a molecules in a framework and modifications to the embedded molecule can be then be defined to the right [R] and left [L]. The approach will also need a list of reactions (below Diels-Alder) that is allowed to be applied to the molecules. Given a Food set molecules, one can define 1st generation, 2nd generation etc of products obtainable from this formalism. Below is illustrated the molecules obtained up to 3rd generation.

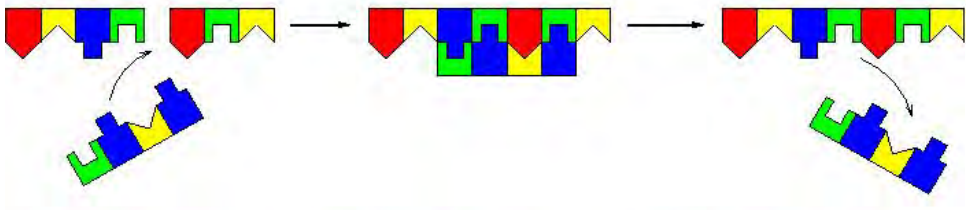


This was implemented by Federico Paoletti based on existing software [Flamm, Stadler, Lykke Jakobsen], but we ran short of time before we could get to the real intersing questions:

1. How big is the set of molecules for a given Food and Reaction Set at generation within the total set of molecules of the same size? As one let k grow the size seems overwhelming, but is probably vanishing within the set of possible molecules.
2. Given a set of target molecules like amino acids, nucleotides and sugars, which Food and Reaction Set could eventually generate these and in how many generations? In actual enumeration schemes, these molecules are assigned extremely high numbers so one could fear that they are only reached when a VERY large molecule universe have been created. Is there a rational and chemically credible way of getting interesting molecules?? It would be easy to stop declare certain molecules unreactive in generation k and thus not lead to new molecules in generation k+1.
3. As k grows when does the first RAFs appear? Again there will be a parameter tuning problem since we don't want RAFs everywhere, but only to start appearing for medium size molecules with say 15+ atoms.

C. ***Catalysis*** is harder to predict as it doesn't lend itself as easily to be reduced to a simple computational problem such as sub-graph isomorphism with constraints. However, there are special cases where catalysis is well established and that is in the case of ligation and hydrolysis of RNA/DNA. ***A+B→C*** [concatenation of ***A*** and ***B***] is catalysis by a molecule ***M*** if ***M*** is complementary to the end of ***A*** and beginning of ***B***.
Abstracting a little in the hope of applying it generally, the double complementarity of ***M*** to ***A*** and ***B*** can be viewed as double docking. Docking [Brooijmans and Kuntz, 2003] is when two molecules "lock" into each other. Docking is hard to predict as it involves real molecules that should be represented in full and described by quantum mechanics. This has led to a series of simplified approaches. These approaches have not been reliable in being able to predict docking with high confidence, but it still represent progress relative to uniform random guessing. Docking comes in many flavours dependent on the problem at hand and the computational resources available. Complementary is always a goal, but the shapes can be refined with charge and flexibility. One

approach to docking would be that ***M*** should dock to ***A*** and to ***B***, so that ***A*** and ***B*** was close, but not physically overlapping.



Two polymers of length 4 and 3, respectively are brought together by a polymer of length 4 that is complementary to the end of the first polymer and the beginning of the next polymer.

Since catalysis prediction seems a new topic and this should be doable by a student in 2-6 months, we will here emphasize simplicity both in the method and in the application. It will be challenging enough.

The method of KATCHALSKI-KATZIR[1] et al. (1992) is simple to program and generalize. The fundamental idea is to represent a molecule by a region called INNER, OUTER and SURFACE. An function a(x,y.z) will take different values dependent in which region (x,y,z) is. Typically 0 outside, 1 inside and some parameter on the surface. Given two molecules [A and B], the docking problem becomes to create as much overlap between the surfaces while not having much overlap between the INNER of A and B.

In this framework our catalysis prediction can be formulated as C has to be able to dock to A and B simultaneously, while not creating overlaps between the INNERs of A and B. A and B need not dock. If we let T be translation, the function to optimize would be $D_{A,B} = \min_{T,R} f(A)*f(R(B+T))$, where * is dot product. This is an optimization with 5 parameters: 3 for translation and 2 angles for rotation. Additionally a series of further parameters will have to be set: How big is the box that contains the two molecules, how fine is the grid, how thick is the surface and what is the penalty for overlap of the two molecuels and reward for overlap of surfaces.

If we have 3 molecules (A,B,C) there will two translations and two rotations to optimize over. The object function will be $D_{C,A+}D_{C,B} + D'_{A,B.}$ D' is the function where there is no reward for overlap between two molecules.

This will not be a fantastic catalysis predictor since catalysis prediction is probably genuinely hard, but it will be pionering. Docking of small molecule versus small molecule is a non-standard problem for which methods have not been developed, but the problems would probably be analogous to ligand-protein and protein-protein docking: Surfaces are flexible and attraction between molecules are complicated by charges and hydrophobic bonds. Undoubtedly, improved methods could be quickly devised by using the methods developed over decades for ligand-protein and protein-protein docking.

D. A ***Food*** Molecule Set must be chosen for the RAF to be defined and decent sets of these can easily be fund in the literature with the most famous example is Miller-Urey from 1953, which included H_2, CH_4, NH_3 and CO_2, but other sets have been explored remoring H_2 and/or adding phospate. List of molecules under priomordial conditions or at hot vents are obvious starting points.

Mini Project proposals

1. Continue the Grammar Reaction work of Federico Paoletti.
2. Continue the Catalysis Prediction work of Ewan Dwyer. Example a: Make idealized complementary shapes - like 2 half moons and test the algorithm on those. Example b: Find atomic coordinates of RNA 6-mer: UCGAUA and two 3-mers: AGC and UAU and see if they will join in one complex. Optimize this method so it is as fast as possible can see if C can bring A and B together.

Scaling to Dphil

A Dphil should should bring together RAF[Food + Molecules + Reactions + Catalysis] !!

It is of course unlikely that a single Dphil will solve OoL, but it would bring forward MoRAF and contribute significantly to computational approaches to OoL.

References

Ganti. Biogenesis itself. *Journal of Theoretical Biology* 187, pp. 583-593, 1997.

[1] Only considerably time after having read this method, did we realise that the first author had been Prime Minister of Israel!! And will refer to the algorithm as PMI.

Hordijk and Steel. Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *Journal of Theoretical Biology* 227(4), pp. 451-461, 2004.
Kauffman. Autocatalytic sets of proteins. *Journal of Theoretical Biology* 119, pp. 1-24, 1986.
Mossel and Steel. Random biochemical networks: the probability of self-sustaining autocatalysis. *Journal of Theoretical Biology* 233(3), pp. 327-336, 2005.
Steel. The emergence of a self-catalysing structure in abstract origin-of-life models. *Applied Mathematics Letters* 3, pp.91-95, 2000.
Von Neumann. *Self-reproducing automata*. 1967.
Gil Benko¨,† Christoph Flamm,*,† and Peter F. Stadler (2003) "A Graph-Based Toy Model of Chemistry" *J. Chem. Inf. Comput. Sci. 43,* 1085-1093
Brooijmans N and I Kuntz (2003) "Moleculear Recognition and Docking Algorithms" Annu.Rev.Biophys.Biomol.Struct. 32.335-73.
JChen, Baldi. (2009)"No Electron Left-Behind: a Rule-Based Expert System to Predict Chemical Reactions and Reaction Mechanisms". *J. of Chem Inf Mod.* 49, 9, 2034-2043
Diettrich, Ziegler and Banzhaf (2001) "Artificial Chemistries—A Review" Artificial Life 7: 225–275
Halperin et al. (2002) "Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions" Proteins 47:409–443
Hunt and Otto (2011) "Dynamic combinatorial libraries: new opportunities in systems chemistry" Chem. Commun., 2011, 47, 847–858
KATCHALSKI-KATZIR et al. (1992) "Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques" PNAS 89, pp. 2195-2199
Kayala et al. (2011) "Learning to Predict Chemical Reactions" *Journal of Chemical Information and Modeling.* 2011
Kerber,„Laue, Meringer,§ and Rucker (2007) "Molecules in Silico: A Graph Description of Chemical Reactions" *J. Chem. Inf. Model. 47,* 805-817
Kuntz, ID; Blaney, JM; Oatley, SJ; Langridge, R; Ferrin, TE (1982). "A geometric approach to macromolecule-ligand interactions". *Journal of molecular biology* **161** (2): 269–88
Rosello and Valiente (2005) "Chemical Graphs, Chemical Reaction Graphs, and Chemical Graph Transformation" Electronic Notes in Theoretical Computer Science 127. 157–166
Lorenz C. Blum • Ruud van Deursen • Jean-Louis Reymond (2011) "Visualisation and subsets of the chemical universe database GDB-13 for virtual screening"J Comput Aided Mol Des (2011) 25:637–647
Reymond, van Deursen, Blum and Ruddigkeit (2010) Chemical space as a source for new drugs Med. Chem. Commun., 2010, 1, 30–38

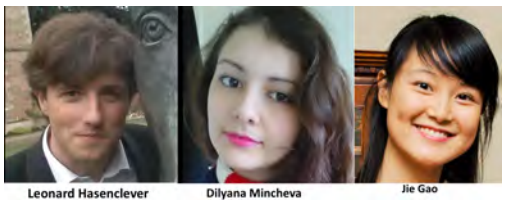
Acknowledgements: JH has received advice on literature from Tack Kuntz, Rene Thomsen, Martin Simonsen, Charlie Carter, Sijbren Otto, Wim Hordijk, Pierre Baldi and probably more. JH is surprised that not more work has been done on catalysis prediction as there clearly are a series of unexplored approaches.

Oxford Work This means supervised or intiated by Jotun Hein. A lot of interesting work is going on in Oxford that I will not mention here. Might be in future editions of enlarged project description.


Wim Hordijk

I taught molecular evolution since early 90s at Aarhus University and decided to include the Origins of Life in my courses, which led me to read a lot on this topic which is extremely diverse [planetary science, climatology, ancestral reconstructions in phylogenetics, molecular biology, emergence [ufff bad word] of chirality in non-chiral systems,...] and very few people have an overview. It was clear that this field was rife with unfounded speculations and was not much respected or well-funded. Clearly, one of the most interesting questions conceivable, but what was its applications? It is my sense that that this has seriously changed in the last 2 decades. Origins of Life has been rebranded as Astrobiology with at least 2 journals with this in their titles and several regular workshops/conferences dedicated to the topic. The rise of of Synthetic Biology and Systems Chemistry with real applications have also helped OoL research. And finally the pervasive use of simulations in biology where the role of computations are permanently expanded. At Århus University I had 8 60 minutes lectures on the Origin of Life. I was not very satisfied with them – there wasn't much of a unifying theme in this field, but it was fund to read about Planetary Systems, the distribution of the elements, the age of the Earth, dating first life on Earth, chirality, designed self-replicating systems, alternative suggestions for possible life (especially Freeman Dyson had an arsenal of crazy ideas), meteor/coment storms, comparative climatology, the distribution of stars types in the galaxies. I tried to find algorithmic papers on what they did in SETI, but I still haven't read a single interesting paper on this topic, which is one of of the most boring areas of science with nothing to show for 55 years of research – it is basically a series of patient people sitting and listening to white noise. In Århus we celebrated end of term and my course with going and watching "Armagedon" with Bruce Willis saving us all from an approaching meteor. In Oxford my lecture series was reduced to 90 minutes introductory lecture at DTC. I worry it erased any confidence students would have in me.


Ina Trolle Andersen Maiken Ina Siegismund Kjærsgaard Nan Lin


Leonard Hasenclever Dilyana Mincheva Jie Gao

In 2011 I was a Miller Fellow in Berkeley, which was an escellent and well funded program and a very large part of the Fellows were working on extrasolar planetary systems, so there I had a chote to discuss it with the experts. It was exciting with the planetary zoo out there and an increasing number of Earth like systems. The very large number of planets has clearly helped increase the status of OoL research. So has the discovery of Water on Mars.

Being face-to-face with the enormous literature with little concrete progress made me a fan of exact models. I have not had a full Dphil student or postdoc in this area, but I have supervised at least 6 projects involving 12 student that have made serious progress. This has led two publications, each time thanks to Wim Hordijk who co-supervised 3 of the above projects at Oxford Summer School in Computational Biology [OSSCB]. OSSCB was great, but also frustrating since we often did real progress, but then it was over and nobody had the energy to pursue writing a paper or putting it on archive. 6-8 papers were published but we had 50-60 projects and 3 times as many students.


Edward Rolls Laurence Hutton-Smith Eli Bingham

Project 1 OSSCB10 [Maiken Ina Siegismund Kjærsgaard, Ina Trolle Andersen, Lin Nan] was focussed on investigating self-replicating systems of RNA motivated by experimental work of Günther von Kiedrowski and Julius Rebek in the 90s and the work was very successful and the report can be found on the project www-pages.

Project 2 OSSCB13 [Jie Gao, Leonard Hasenclever & Dilyana Mincheva] investigated the consequences of giving the probability of catalysis of a given reaction a power law distribution. They fully completed the project and it lead to a publication (An investigation into irreducible autocatalytic sets and power law distributed catalysis)


Lukas Hutter

Project 3 OSSCB13 [Eli Bingham, Laurence Hutton-Smith & Eddie Rolls] was our first serious attempt to say "Just say no to polymers!!" and do molecularisation of the RAF. The team was extremely strong but ran into problems: it wasn't planned well enough and one of the teachers had to leave most of the time due to serious illness in the family. It was depressing.

DTC mini-project 1 2013 [Lukas Hutter] wanted to find RAFs in real molecular reactions and Lukas focussed on a paper on fire!! Fire is clearly self-replicating!! Lukas made very

important observations and we had a worrying discussion about "catalysis" (which is key to RAF): i. There are no catalysis in fire - the reactions are all thermodynamically feasible. ii. Catalysis is often a kind/set of reaction(s), so we considered defining a catalyst as a molecule that was necessary for a set of reactions, but entered and exited in the same amount. A catalyst could be a set of molecules. If this was the case it would pull the carpet away under much RAF analysis and would pose the new problem of algorithmically recognising catalysts or catalyst sets in a reaction graph. I loved this fire-systems and had dreams of a set of reactions that would burn either green or orange dependent on which initial conditions they were given.


Frederico Paoletti

DTC mini-project 2 2015 [Federico Paoletti] Normally we don't do the same project twice, but in case of molecularisation of the RAF we decided to redo the OSSCB13 project 3 but now much more focussed and well-defined. Thus only focus on Graph Grammars as means of defining reactions, I also discussed with potential co-supervisors around in the world [Flamm, Peter Stadler, Jakob Lykke Andersen, Hordijk] and had an explicit reading list we had to get through. We had to read 6-10 real hard papers on RAF theory and Double Pushout Automata, that we finally understood properly. This clearly paid off and Federico did a very large amount of progress and implemented algorithms and made simulations. Given a set of Food molecules and a set of reactions, one could get an expanding set of molecules [1^{st}, $2^{nd,...}$ k^{th} generation] and we ready to start to ask question like "How set the k'th generation in the possible set of molecules of that size?" Vanishing of course, but exactly how small. An important missing incredient was catalysis predicting, without which the real molecules might just as well be abstract nodes in a graph. The project clearly added one of the key ingredients in any further progress on MoRAF.

A Single Summer Student 2015 [Ewan Dwyer] was here in 5 weeks and implemented the Israeli Prime Minister [IPM] Algorithm, but did not have time for testing or writing a real report. Alex Popinga [American doing her PhD in Auckland, NZ under the supervision of Chris Wills] was here for the same period and also did RAFs but not with focus on OoL.


Ewan Dwyer

In my promotion of a computational approach to OoL, I discussed with several people with some relevant expertise. In Oxford you invite them to lunch at your College. Frederick Taylor had written at least 4 relevant books: Mars, Venus, Titan, Planetary Atmospheres – I have only read the last, but would have read the first 3 if a Dphil student showed up so we could have defined the relevant food sets. Graham Richards had promoted investigating the interaction of small molecules by harvesting unused computational power of PCs and I was interested in how he implemented docking. The other famous project doing the same is SETI, but I still don't know what they are calculating.


Alex Popinga

There is an awfull lot of very bad books on OoL, but 3 relevant I enjoyed are: i. Maynard Smith and Szhathmary Major Evolutionary Transitions, which has the virtue that it tries to extract the basic IDEAS of different scenarios, not just throw a lot of chemistry and geology in your head, ii. How to Find a Habitable Planet by Kastings [now I think of it: That is easy, you just close your eyes, count to 50, open your eyes and there it is!! I must email to Kasting and he will put this observation in an appendix] and iii. Stephen Mason Chemical Evolution. And there is a surprising good literary/historical book on the history of SET by Steven Dick: Life on Other Worlds. The ideas about life elsewhere has a surprisingly long history.

If somebody knows of a good recent book, please contact me.

And take a look at the photos of these students and note how happy people who work on OoL all look !!!