Autocatalytic networks and the origin of life

A summer project at the University of Canterbury, funded by the
Allan Wilson Centre for Molecular Ecology and Evolution

Joshua Smith

During the summer of 2012/2013, I completed a summer research project under the supervision of Prof. Mike Steel at the University of Canterbury (UC). The project involved contributing to the ongoing development of "RAF theory", a framework for mathematical investigation into the origin of life problem. The development of RAF theory began in 2000 with Mike's extension of Stuart Kauffman's informal mathematical work (Kauffman 1986, Steel 2000), and has continued as a collaboration with Wim Hordijk and others (Hordijk & Steel et al, 2004, 2005, 2010, 2012).

RAF theory examines the structure of networks of interdependent chemical reactions which grow from an initial collection of simple molecules known as the *food set*. While no restriction is placed on the type of molecules involved, we often study the *binary polymer model*, in which all molecules are modelled as bit strings: polymers consisting of the monomers 0 and 1 (see Figure 1). In this context reactions are either ligation of two molecules into one, or cleavage of a single molecule into two. Specifically, we are interested in subsets of such sets of reactions that satisfy two important mathematical properties. Here I will define these properties only informally, as space precludes the presentation of the formal set-theoretic definitions.

- A set of reactions is *reflexively autocatalytic* if and only if every reaction is catalysed, and the reactions collectively produce all of their own catalysts from the food set.
- A set of reactions is *Food-generated* (F-generated) if and only if the set is structured such that all the involved molecules could be built up in steps over time, beginning with only the food set.

A set of reactions satisfying both of these properties is known as a RAF (reflexively autocatalytic and F-generated) set, or just a RAF for short. The subset of reactions $\{r_1, r_2, r_3\}$ in Figure 1 is an example of a RAF. The mathematical properties of RAF sets are designed to echo properties of real chemical networks that might be essential in an origin of life scenario. One of the goals of studying RAF sets is to generate formal results and use them to make predictions about real experimental systems, therefore accelerating progress in this multidisciplinary area of research.

In the early stages of my summer project, I read existing academic papers on RAF theory, familiarising myself with the concepts and the set-theoretic notation required to follow the mathematics involved. I also read about *chemical organisation theory* (COT), an alternative formalism of chemical networks. As I became comfortable with these ideas, I was able make a small contribution to Mike and Wim's latest paper (Steel, Hordjik & Smith 2013, arXiv:

1212.4450). One section of the paper introduces the concept of co-RAFs, which, informally, are defined as sets of reactions that when combined with an existing RAF form a larger RAF. In Figure 1, the set $\{r_3, r_4\}$ is not itself a RAF, but it is a co-RAF: when it is combined with the RAF $\{r_1, r_2\}$ the result is a larger RAF. My main contribution to the paper was the proof of a result which provides four alternative mathematical definitions of what it means for a set to be a co-RAF.

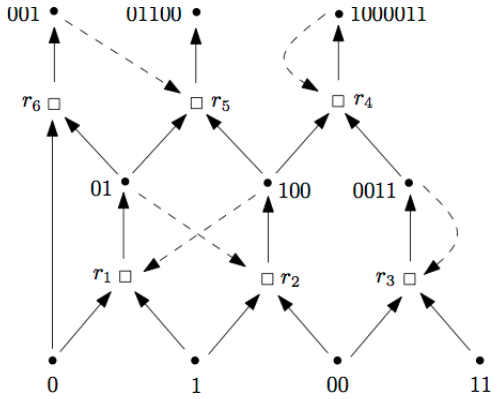

Figure 1: an example of a small set of reactions within the binary polymer model. Dashed arrows are between molecules and the reactions they catalyse.

The example shown in Figure 1 is a very small set of reactions: we are usually interested in studying much larger systems, involving perhaps tens of thousands of reactions, which may be generated on a computer. Due to the size of such systems, searching for RAF subsets becomes computationally demanding. In 2004 Hordijk and Steel presented a search algorithm which can discover, in polynomial time, a RAF subset of any given set of reactions (if one exists). During the project I worked with Wim to learn about implementing this algorithm, and with Mike to understand basic computational complexity theory (not all algorithms are guaranteed to run in a practical amount of time: roughly, those that run in *polynomial time* are practically useful, while those that require more time are not). This was very valuable, since prior to this project I had no knowledge of the difference classes of problem complexity. A consequence of the computationally demanding nature of searching through these large systems is that a desktop computer quickly becomes ineffective. Towards the end of my project, I was able to open a user account on BlueFern (the UC supercomputer). We hope to implement BlueFern during my Honours project in 2013, which will build on the work done during this project.

Another exciting aspect of the project was being involved in writing an academic paper and submitting it to a journal for publication. This was new to me, and it was great to learn about the process.

In summary, I have thoroughly enjoyed the project, and it has given me a head start on my Honours project at UC. I would like to thank Mike and Wim for their excellent teaching and encouragement during the project, as well as Mike Steel and the Allan Wilson Centre for their generous funding.