Orig Life Evol Biosph (2014) 44:111–124
DOI 10.1007/s11084-014-9374-5

THEORETICAL MODELING

Conditions for Evolvability of Autocatalytic Sets: A Formal Example and Analysis

Wim Hordijk · Mike Steel

Received: 14 July 2014 / Accepted: 1 October 2014 /
Published online: 5 December 2014

Abstract We provide a formal but visually clear example of how a set of minimal necessary conditions for evolvability of autocatalytic sets is satisfied in a simple model of chemical reaction systems. Furthermore, we show how these conditions can be captured and analyzed with RAF theory, and how the results can be generalized with a somewhat more elaborate example. Finally, we argue that our results clearly support the hypothesis that autocatalytic sets can be evolvable, and that this might even be an expected property of such sets.

Keywords Origin of life · Autocatalytic sets · Evolvability

Introduction

The idea of collectively autocatalytic sets was introduced as a "metabolism-first" scenario in the context of the origin of life (Kauffman 1971, 1986, 1993; Dyson 1982, 1985). In this scenario, life supposedly started as a functionally closed, self-sustaining reaction network in which several molecules collectively support each other's production from basic nutrients through mutually catalyzed chemical reactions. This idea has not been without criticism (Lifson 1997; Orgel 2008; Vasas et al. 2010), but recent progress in constructing autocatalytic sets in the laboratory supports its plausibility (Sievers and von Kiedrowski 1994; Ashkenasy et al. 2004; Taran et al. 2010; Vaidya et al. 2012). Over the past decade or so, the

W. Hordijk (✉)
SmartAnalytiX.com, Lausanne, Switzerland
e-mail: wim@WorldWideWanderings.net

M. Steel
Allan Wilson Centre, University of Canterbury, Christchurch, New Zealand
e-mail: mike.steel@canterbury.ac.nz

concept of autocatalytic sets has been formalized mathematically and studied extensively in the form of RAF theory (Steel 2000; Hordijk and Steel 2004; Hordijk 2013), leading to the refutation of some of the earlier criticisms.

One of the main criticisms against the concept of autocatalytic sets was their apparent lack of evolvability (Vasas et al. 2010). In the original formulation, and its supporting arguments, autocatalytic sets appear as "giant connected components" within a chemical reaction network (Kauffman 1986, 1993), indeed leaving little room for change and adaptation. However, in a follow-up study the same authors of Vasas et al. (2010) showed that autocatalytic sets *can* actually be evolvable given a minimal set of necessary conditions (Vasas et al. 2012). Whether this set of conditions is also sufficient for true *open-ended* evolution is still unclear, and probably depends on whether new functionality can emerge in the evolving system (which is unlikely in the polymer model used to generate the results, even though the number and maximum length of polymers is, in principle, infinite). But these results clearly show that autocatalytic sets can, at the least, show a form of pre-template Darwinian evolution.

The necessary conditions for evolvability of autocatalytic sets are based on the notion of a viable core (Vasas et al. 2012). Informally, a *viable core* is a "minimal" autocatalytic set. The set of necessary conditions is then as follows:

1. The availability of compartments that can grow and divide;
2. The existence of multiple viable cores within the (complete) underlying reaction network that can potentially co-exist in various combinations inside a compartment;
3. A mechanism for the spontaneous gain or loss of viable cores during the actual dynamical "execution" of the reaction network.

Having multiple such viable cores that can co-exist in various combinations within compartments which can grow and divide, allows for inheritance, variation, competition, and thus evolvability. A "mutation" would be the (spontaneous) gain or loss of a viable core within a compartment. A viable core can, for example, be gained by a rare, spontaneous reaction that produces a "seed" molecule that is required to make a viable core come into existence (in a dynamical sense). On the other hand, a viable core could be lost during a compartment division with an unequal distribution of molecules (due to stochastic fluctuations) between the offspring compartments (Vasas et al. 2012).

In this paper, we show how RAF theory can be used to formally capture, exemplify, and analyze these necessary conditions, and how results obtained with the formal RAF framework support the possibility for evolvability of autocatalytic sets. In particular, we provide a small but visually clear example of how the above necessary conditions are satisfied within a simple model of a chemical reaction system, and how it can be analyzed and understood using RAF theory. We then provide a more general example using a well-studied polymer model to show that the results from the minimal example can be generalized. This, combined with earlier results from RAF theory, suggests that evolvability is a general, and perhaps even expected property of autocatalytic sets.

Autocatalytic Sets and RAF Theory

First, we define a *chemical reaction system* (CRS) as a tuple $Q = \{X, \mathcal{R}, C\}$ consisting of a set X of molecule types, a set $\mathcal{R}$ of chemical reactions, and a catalysis set C indicating

which molecule types catalyze which reactions. We also consider the notion of a food set $F \subset X$, which is a subset of molecule types that are assumed to be freely available from the environment. A graphical example of a CRS is provided in Fig. 1.

Informally, an *autocatalytic set* (or RAF set) is now defined as a subset $\mathcal{R}' \subseteq \mathcal{R}$ of reactions (and associated molecule types) which is:

1. *Reflexively Autocatalytic* (RA): each reaction $r \in \mathcal{R}'$ is catalyzed by at least one molecule type involved in $\mathcal{R}'$, and
2. *Food-generated* (F): all reactants in $\mathcal{R}'$ can be created from the food set F by using a series of reactions only from $\mathcal{R}'$ itself.

This definition captures the idea of a functionally closed (RA) and self-sustaining (F) reaction network. A more formal definition of RAF sets is provided in Steel (2000), Hordijk and Steel (2004), and Hordijk et al. (2011), including an efficient (polynomial-time) algorithm for finding RAF sets in a general CRS.

If an RAF set does exist, the RAF algorithm returns the unique *maximal* RAF set (maxRAF), which is the union of all RAF sets within a given CRS. If no RAF set exists within a CRS, the RAF algorithm returns an empty set. A maximal RAF set can often be decomposed into several smaller subsets which themselves are RAF sets, i.e., subRAFs (Hordijk et al. 2012). If such a subRAF cannot be reduced any further without losing the RAF property, it is referred to as an *irreducible* RAF, or irrRAF (Hordijk and Steel 2004).

Some of the main findings of RAF theory are that autocatalytic sets are highly likely to exist in random (polymer-based) models of reaction networks once a critical level of

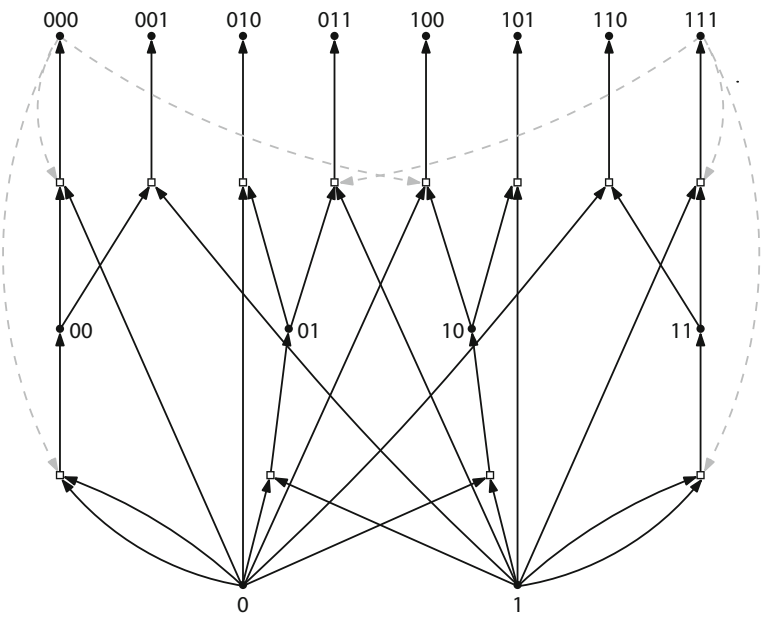

Fig. 1 A graphical representation of the example CRS. *Black dots* represent molecule types, white boxes reactions. *Solid arrows* are reactants going into and products coming out of a reaction. *Dashed arrows* indicate catalysis. The food set consists of the monomers 0 and 1

catalysis is exceeded (Hordijk and Steel 2004; Mossel and Steel 2005). This critical transition point already occurs at very modest levels of catalysis: between one and two reactions catalyzed per molecule type for moderate sized networks (Hordijk and Steel 2004; Hordijk et al. 2010). Moreover, only a linear growth rate in this critical level of catalysis is required to ensure that RAF sets exist with high probability for increasing polymer lengths (Hordijk and Steel 2004; Mossel and Steel 2005). These results hold up under a variety of more realistic model extensions, and even for non-polymer systems (Hordijk et al. 2011; Hordijk et al. 2014c; Smith et al. 2014; Hordijk et al. 2014a). Generally, there exist many hierarchical levels of subRAFs (Hordijk et al. 2012) (as opposed to having one "giant connected component"). Finally, the formal RAF framework can be directly and successfully applied to real chemical and biological systems to analyze the emergence and structure of autocatalytic sets (Hordijk and Steel 2013; Sousa et al. 2014).

An irrRAF is roughly the equivalent of a viable core as described in Vasas et al. (2012). Formally the equivalence is not exact: although every irrRAF can be considered a viable core, a viable core is either an RAF set or a slightly more general set that we have called a "pseudo-RAF" (Hordijk et al. 2014b). A pseudo-RAF would need at least one of its non-food molecules to come from somewhere else before the set as a whole can come into existence (in a dynamical sense). Such a "seed" molecule could come, by chance, from the environment, or be produced by some other (rare) reaction that is not part of the set itself. However, for most purposes, and in the examples below, viable cores and irrRAFs can be considered equivalent.

Conditions for Evolvability: A Formal Example

Here, we provide a simple but formal example to show how the conditions for evolvability of autocatalytic sets are satisfied in a particular instance of a simple polymer-type model of chemical reaction systems, and how they can be captured and analyzed with RAF theory.

The CRS

We consider a particular instance of the Wills-Henderson (W-H) model (Wills and Henderson 2000; Hordijk et al. 2014c). The W-H model consists of binary polymers (bit strings) that can grow by the ligation of a monomer (a 0 or a 1) to the end of a left-to-right oriented polymer. The molecule set X consists of all bit strings up to a maximum length n. In our example, we use $n = 3$. This value is small enough that the results can be easily understood and verified by eye, yet large enough to already show relevant behavior regarding the conditions for evolvability.

The set of reactions $\mathcal{R}$ in the W-H model consists of four categories: ($\mathcal{R}_1$) ligation of a 0 to a bit string ending with a 0, ($\mathcal{R}_2$) ligation of a 1 to a 0, ($\mathcal{R}_3$) ligation of a 0 to a 1, and ($\mathcal{R}_4$) ligation of a 1 to a 1. So, for example, a reaction in category $\mathcal{R}_1$ looks like $b0 + 0 \rightarrow b00$, where b is any (possibly empty) bit string of length at most $n - 2$ (Hordijk et al. 2014c). Note that in the example used here, we only consider reactions that create ligation products no longer than $n = 3$ bits.

Next, we assign two particular bit strings as catalysts: the bit string 000 catalyzes reactions in the first category $\mathcal{R}_1$, and the bit string 111 catalyzes reactions in the fourth category $\mathcal{R}_4$ (later on, in a slightly extended example, we also consider additional catalysts). Finally, the food set F consists of the monomers 0 and 1.

The complete example CRS is thus as follows:

$$
\begin{aligned}
X &= \{0, 1\}^{\leq 3} \\
\mathcal{R} &= \mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3 \cup \mathcal{R}_4 \\
C &= \{(000, r) | r \in \mathcal{R}_1\} \cup \{(111, r) | r \in \mathcal{R}_4\} \\
F &= \{0, 1\}
\end{aligned}
$$

This CRS is represented graphically in Fig. 1.

RAF Sets

Applying the RAF algorithm to this CRS results in a maximal RAF set $\mathcal{R}'$ of four reactions (this can be easily verified by hand):

$$\mathcal{R}' = \{0 + 0 \to 00, 00 + 0 \to 000, 1 + 1 \to 11, 11 + 1 \to 111\}.$$

Furthermore, this maxRAF can be decomposed into two independent irreducible RAF sets (or viable cores):

$$
\begin{aligned}
\mathcal{R}^0 &= \{0 + 0 \to 00, 00 + 0 \to 000\}, \\
\mathcal{R}^1 &= \{1 + 1 \to 11, 11 + 1 \to 111\}.
\end{aligned}
$$

The Simulation

To show how these irrRAFs can exist in various combinations within a compartment, we perform dynamical simulations of the example CRS using the Gillespie algorithm (Gillespie 1976, 1977). We assume that all reactions (in each category) can happen spontaneously (uncatalyzed) with a kinetic constant of $k_s = 0.1$. Catalyzed reactions (as determined by the catalysis set C) have a kinetic constant of $k_c = 1.0$. In the simulation, this k_c is multiplied by the number of catalysts currently present in the system to determine the actual reaction rates at each step.

We start the system with an initial concentration of 10 for each of the food molecules (the monomers 0 and 1). During the simulation, the food molecule concentrations are kept at this level, i.e., each time a monomer is used in a reaction, it is replaced by a new monomer of the same kind (to simulate influx of food molecules into the compartment).

In addition, we assume that trimers (bit strings of length three) flow out of the compartment at a given rate. These outflow reactions have a kinetic constant of $k_o = 0.8$, but these reactions are "catalyzed" by the trimer itself. In other words, the actual outflow reaction rate of a trimer depends on its own concentration. This can be interpreted as a kind of "osmotic pressure" condition, where the outflow of molecules is larger the higher their concentration inside the compartment is.

Finally, we set the volume of the reaction vessel (compartment) to one, and run the simulation for one time unit.

Basic Results

In general, there are four possible outcomes of a simulation: (1) none of the irreducible RAF sets have come into existence, (2) only irrRAF $\mathcal{R}^0$ exists, (3) only irrRAF $\mathcal{R}^1$ exists, and (4) both $\mathcal{R}^0$ and $\mathcal{R}^1$ exist. Outcomes (2) and (4) are illustrated in Fig. 2. Note that once an irrRAF comes into existence, the concentration of the trimer it produces initially increases at an exponential rate (due to the autocatalytic nature of the irrRAF), but then levels off

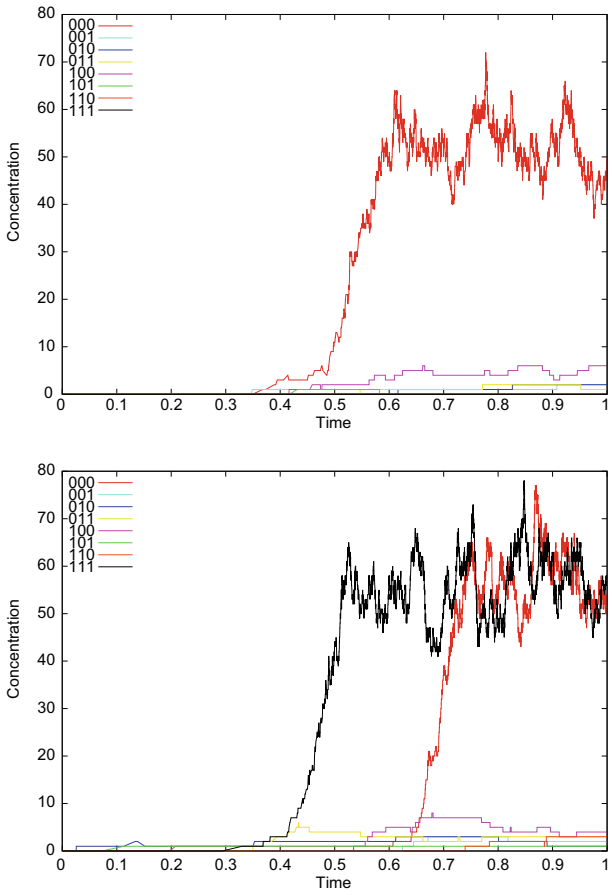

Fig. 2 Two of the four possible simulation results in terms of existence of irreducible RAF sets (*viable cores*). *Top*: only $\mathcal{R}^0$ exists. *Bottom*: both $\mathcal{R}^0$ and $\mathcal{R}^1$ exist

(due to the increasing rate of outflow of the trimer) until it reaches a (dynamic) equilibrium concentration.

All simulations start with the exact same initial conditions, but due to the stochastic nature of the Gillespie algorithm (as in real chemistry), different outcomes are possible depending on which spontaneous reactions happen first and when. Both irrRAFs ($\mathcal{R}^0$ and $\mathcal{R}^1$) initially require their reactions to happen spontaneously, i.e., uncatalyzed, at least once before the full irrRAF can actually come into existence. From the RAF analysis we know that they are present in the underlying reaction network (i.e., at a static level), but whether they are actually realized (i.e., at a dynamic level) depends on the occurrence of (rare) random reaction events.

Note that in this simple example, if the simulation is run for long enough (several time units), eventually both irrRAFs will come into existence and then co-exist "forever". This is due to the fact that they do not compete for the same resources (food molecules). IrrRAF $\mathcal{R}^0$ only requires the monomer 0, and irrRAF $\mathcal{R}^1$ only monomer 1. So, on shorter time scales there can be four different outcomes, but on longer time scales there really is only one "attractor".

Extended Results

To provide an example with more than one long-term attractor, consider the additional catalysts 010 and 101, where 010 catalyzes reaction category $\mathcal{R}_2$ and 101 catalyzes reaction category $\mathcal{R}_3$. So, we use the same CRS as before, but with the extended catalysis set

$$C = \{(000, r)|r \in \mathcal{R}_1\} \cup \{(010, r)|r \in \mathcal{R}_2\} \cup \{(101, r)|r \in \mathcal{R}_3\} \cup \{(111, r)|r \in \mathcal{R}_4\}.$$

Note that in Fig. 1 this would mean adding six additional catalysis (dashed) arrows. For this CRS, the entire reaction network becomes a maxRAF, and there is one additional irrRAF to make a total of three: $\mathcal{R}^0$ and $\mathcal{R}^1$ as before, plus

$$\mathcal{R}^{01} = \{0 + 1 \to 01, 1 + 0 \to 10, 01 + 0 \to 010, 10 + 1 \to 101\}.$$

In this slightly extended example, there are two distinct long-term attractors: (1) coexistence of $\mathcal{R}^0$ and $\mathcal{R}^1$, as illustrated in Fig. 2 (bottom), and (2) existence of $\mathcal{R}^{01}$ only, illustrated in Fig. 3 (top). Which of these two attractors is reached depends again on (rare) stochastic reaction events, but they are mutually exclusive due to competition for the same resources (food molecules). In other words, once one of these two attractors is reached, it "uses up" the available food molecules and intermediate products (dimers) at such a high rate, that the other attractor cannot get a foothold anymore to also come into existence.

On short time scales, however, there can still be any combination of the three irrRAFs. But there is an additional short-term dynamic as well. If only one of $\mathcal{R}^0$ or $\mathcal{R}^1$ exists (i.e., before they have both come into existence and reached the first stable attractor), it can be "taken over" by the appearance of $\mathcal{R}^{01}$, illustrated in Fig. 3 (bottom). Once $\mathcal{R}^{01}$ comes into existence, it "outcompetes" $\mathcal{R}^0$, which will eventually disappear (this simulation is run for 2 time units to clearly show this effect).

A More General Example

The examples above are carefully constructed by hand to clearly show how the necessary conditions for evolvability are satisfied in a simple binary polymer model. However, an obvious question is how many (different) irrRAFs and possible attractors can be expected to exist for a given (arbitrary) reaction network. This would give an indication of the possible "diversity" one can expect to see.

In Hordijk et al. (2012) we showed that, in principle, the number of irrRAFs within a given (max)RAF can grow exponentially with the size of the RAF, so in general it is not possible to efficiently enumerate all of them. However, using an extension of the basic RAF algorithm, it is possible to randomly sample irrRAFs within a given RAF set (Steel et al. 2013). Using such a sampling method, and a standard statistical test, we have shown elsewhere that also in practice there seem to be very large numbers of irrRAFs (possibly even hundreds of thousands) in random instances of a different, but related, binary polymer model (Hordijk et al. 2014b). With this many irrRAFs in one single reaction network there is obviously a certain amount of overlap between them, but on average about half of the reactions between an arbitrary pair of irrRAFs are different (in moderate-sized networks and at a level of catalysis where RAFs are likely to exist). In other words, a large number, and significant diversity, of irrRAFs can be expected in (random) reaction networks.

So, the next question is whether this large diversity of irrRAFs can indeed give rise to different dynamical behaviors and attractors. A full investigation of this is beyond the scope of the current paper (where the focus is mostly on the formal principles behind evolvability

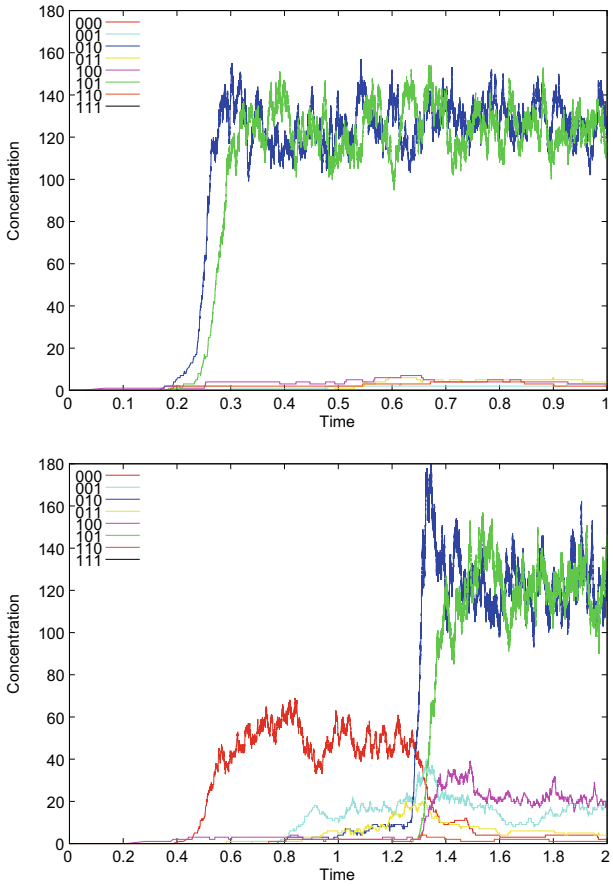

Fig. 3 Two new possible simulation results with the extended example CRS. *Top*: only $\mathcal{R}^{01}$ exists. *Bottom*: $\mathcal{R}^{0}$ comes into existence first, but is then outcompeted by $\mathcal{R}^{01}$

of autocatalytic sets), but the following analysis of one particular example suggests that this question can most likely be answered positively.

The CRS

In previous work we have often used a well-known binary polymer model related to the Wills-Henderson model, but which allows a more general set of polymer reactions and catalysis assignments. In this more general model, introduced by Kauffman (1986, 1993), molecules are also represented by bit strings up to a given maximum length n. The food set consists of all bit string up to a given small length $t < n$. The possible reactions are *ligation*, in which two bit strings are concatenated into a longer one (taking the maximum bit string constraint into account), and *cleavage*, where a bit string is split into two smaller ones. Finally, catalysts (bit strings) are assigned to reactions (ligations and cleavages) at random according to a probability of catalysis p, which is the probability that a given bit string x catalyzes a given reaction r (or, in other words, that a given molecule-reaction pair (x, r) will be included in the catalysis set C).

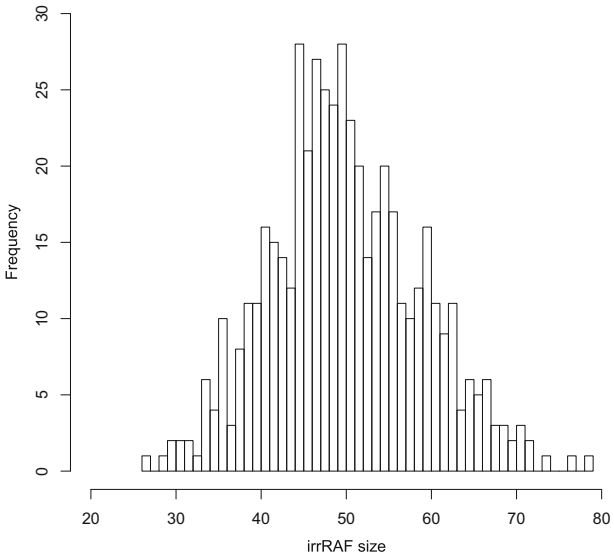

Fig. 4 A histogram of the sizes of $S = 500$ randomly sampled irrRAFs within the maxRAF in the used binary polymer model instance with maximum molecule length $n = 6$

We have taken a random instance of this binary polymer model with $n = 6$, $t = 2$, and $p = 0.003$, on which we applied the RAF algorithm and irrRAF sampling method.

RAF sets

The particular model instance used here consists of 1032 reactions, and contains a maxRAF of 288 reactions (i.e., 144 ligation reactions and their corresponding (reverse) cleavage reactions). Figure 4 shows a histogram of the sizes of irrRAFs within this maxRAF from a random sample of $S = 500$ irrRAFs. These sizes range from 26 to 79 reactions, with an average of 50.

It turns out that all these $S = 500$ irrRAFs are different, even though some of them are of the same size. There is an average overlap of $O = 0.32$, which means that, on average, an arbitrary irrRAF shares about one third of its reactions with another arbitrary irrRAF from the sample (Hordijk et al. 2014b). Repeating this sampling method, again with $S = 500$, results in 500 different irrRAF almost every time. Only occasionally will there exists a pair of equal irrRAFs within the random sample. This means that we can expect, with high probability, at least $\frac{500^2}{10} = 25,000$ irrRAFs to exist within the given maxRAF (Hordijk et al. 2014b).

Simulation Results

Next, we applied the Gillespie algorithm to the reaction network as defined by the maxRAF in the given model instance. As in the previous examples, catalyzed reactions have a kinetic constant of $k_c = 1.0$, but with a possibility of these same reactions happening spontaneously (uncatalyzed) with a kinetic constant of $k_s = 0.3$. Each simulation starts with a concentration of 10 of each of the six food molecule types (i.e., all bit strings up to length $t = 2$), which are continuously replenished whenever they are consumed in any reaction, as

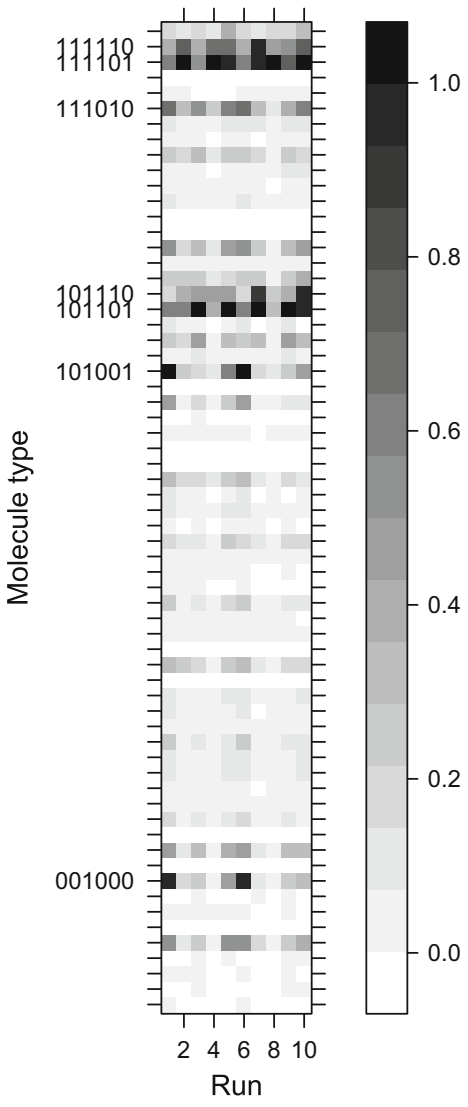

Fig. 5 Ten different simulation results, showing the distribution of relative concentrations (as a grey-scale) of maximum length ($n = 6$) molecules after two time units. Only the most dominant molecule types are labeled, but they are ordered lexicographically, with 000000 at the bottom and 111111 at the top

before. In this example there is no outflow of molecules, and the total volume (one) remains constant. The simulation is then run for a total of two time units.

Figure 5 shows the result of 10 different simulation runs. In this figure, the relative concentrations of maximum-length molecules (i.e., bit strings of length $n = 6$) after two time units are shown for each simulation. The molecule concentrations are normalized by the largest concentration in each run. In other words, the molecule type that is present in the largest amount (at the end of a given simulation) has a relative concentration of 1.0, a molecule type that is present in only half that amount has a relative concentration of 0.5, and molecule types that are not present at all have a relative concentration of 0.0. These relative concentrations are represented by a grey-scale (white is 0.0 and black is 1.0).

As Fig. 5 shows, different simulations can have different outcomes in terms of the most dominant molecule types. For example, in simulation runs 1 and 6, molecule types 001000 and 101001 are produced in the highest quantities (darkest dots), while the others exist in relatively low concentrations. In simulation run 3, molecule type 101101 is clearly dominant, while the others exist in much lower concentrations (including 001000 and 101001, which were dominant in runs 1 and 6). In run 7, molecule types 101101, 101110, 111101, and 111110 are by far the dominant types, whereas most other runs mostly produced some subset of these four types.

Of course this is only a very rough representation, but it does illustrate the possibility for different dynamical behaviors. We actually performed more than 10 simulation runs, many of which show similar outcomes, but Fig. 5 shows at least some of the more dissimilar ones. There are clearly not as many "attractors" as there are irrRAFs in this example, due to the partial overlap between the irrRAFs, and probably also due to a different number of spontaneous reactions required to allow these irrRAFs to come into existence (i.e., some of them are more likely to exist than others). However, there does appear to be significant variability in the dynamical behavior of the system over different runs, indicating that the results from the previous (more artificial) examples can be generalized.

Discussion

The above examples serve as a formal and visually clear illustration of how the necessary conditions for evolvability of autocatalytic sets, as stated in Vasas et al. (2012), are satisfied in simple polymer-type models of a chemical reaction system, and how they can be captured and analyzed with RAF theory. This allows for drawing more general conclusions about the possible evolvability of autocatalytic sets, also based on earlier results from RAF theory, as discussed next.

Compartments

The first condition for evolvability of autocatalytic sets is the availability of compartments. In our simple example, a "compartment" is simulated by having an inflow of food molecules (i.e., a constant concentration of monomers) and a continuous outflow of trimers from the system. The notion of a compartment, or more generally that of a boundary, is actually not explicitly included in the RAF formalism. However, we recently showed how this notion can be included implicitly (by considering the boundary as another "catalyst" for the reactions that happen within its enclosure), and how this gives rise to a mechanism for the emergence of "higher-level" RAF sets, i.e., a RAF (super)set of RAF (sub)sets (Hordijk and Steel 2014). This actually provides another mechanism for evolvability of autocatalytic sets, through the emergence of higher-level structures and functionality (which, as mentioned in the introduction, is most likely required for true open-ended evolution).

However, one element of the first condition as stated in the introduction, that of growth and division of compartments, is not included in our examples. Yet it seems that the basic requirements for evolvability are met (due to a different possible mechanism for loss of viable cores; see the further discussion below). This would imply that an even simpler set of necessary conditions can be formulated for the evolvability of autocatalytic sets.

The issue of compartmentalization is a general problem in the origin of life, regardless of whether one considers a metabolism-first or a genetics-first scenario. However, there appear to be possible solutions to this problem, e.g., the spontaneous formation, growth, and

division of lipid layers (Segré et al. 2001). Furthermore, there are also plausible origin of life scenarios that do not require the formation of an explicit boundary, at least not initially (Martin and Russel 2007; Wächtershäuser 2007).

Multiple irrRAFs

The second condition is the existence of multiple viable cores, or irreducible RAF sets, that can exist in various combinations within a compartment. Our simple but formal example clearly illustrates the basic principles of this condition. On short time scales, there can be various combinations of the three possible irrRAFs within a compartment, and on longer time scales there are two stable attractors. Using the RAF algorithm, we can detect the possible irrRAFs that exist in a given reaction network (at a static level), and with dynamical simulations we can see which ones actually come into existence, and in which combinations.

In principle the number of irrRAFs within a given (max)RAF can grow exponentially with the size of the RAF (Hordijk et al. 2012). Here, and in related work (Hordijk et al. 2014b), we have shown that also in practice there are large numbers of irrRAFs (possibly tens or even hundreds of thousands) in random instances of a simple binary polymer model. The example provided here gives a first indication that this can indeed give rise to different "attractors" (although the number of attractors is not necessarily of the same order as the number of irrRAFs).

Spontaneous Gain or Loss

The third condition is a mechanism for spontaneous gain or loss of viable cores, or irrRAFs. As was already shown in Vasas et al. (2012), viable cores can come into existence through rare spontaneous reactions that produce a required "seed" molecule. Indeed, as illustrated in our simple example, irrRAFs generally require one or more of their reactions to happen spontaneously (uncatalyzed) at least once before the full set can come into existence (in a dynamical sense). Due to the stochastic nature of these rare events, different simulations have different outcomes, even when starting from the same initial conditions. So, this need for one or more (initially) uncatalyzed reactions actually satisfies part of the third condition (spontaneous gain of viable cores). However, in general it is a difficult (NP-complete) problem to determine the smallest number of uncatalyzed reactions required to make an arbitrary (sub)RAF come into existence (Hordijk et al. 2014c).

This is in sharp contrast to so-called "constructible" autocatalytic sets, or CAFs (Mossel and Steel 2005). A CAF is an RAF that is immediately constructible from the food set without the need for any (initially) uncatalyzed reactions. This may sound like a desirable property, but there are two important reasons why this is not the case.

First, CAFs are much less likely to exist in (arbitrary) reaction networks than RAFs. Indeed, they require an exponential growth rate in the level of catalysis with increasing system sizes (Mossel and Steel 2005), whereas RAFs only require a linear growth rate in the level of catalysis (Hordijk and Steel 2004; Mossel and Steel 2005). Intuitively, this can be seen as follows. Given a number of available catalysts and reactions that need to be catalyzed, for an RAF to exist these catalysts can be assigned to the given reactions in any random order. However, there are only very few of those (random) assignments that will give rise to a CAF. As a consequence, with increasing system sizes, CAFs are exponentially less likely to exist than RAFs.

Second, CAFs do not provide the desired diversity and possibility for spontaneous gain of viable cores as required by the third condition for evolvability of autocatalytic sets. This

is because in any given CAF there can be only a small number of irreducible CAFs (and this number does not increase with increasing system size), and they will always (by definition) be of size one. Thus, in the context of the origin and early evolution of life, RAFs are of much more interest than CAFs.

In Vasas et al. (2012), the mechanism given for spontaneous loss of viable cores is through compartment division with an unequal distribution of molecules between the offspring compartments. In our simple example, compartment growth and division is not simulated, so this mechanism does not occur here. However, an alternative (and perhaps more plausible) mechanism for the loss of irrRAFs is clearly illustrated in Fig. 3 (bottom). An irrRAF can be lost because it is out-competed by another irrRAF that has (spontaneously) come into existence later on. When irrRAFs do not compete for the same resources, they can co-exist inside a compartment, as in Fig. 2 (bottom). However, when they compete for the same resources, it is likely that one of them will eventually outcompete the other and make it disappear.

A third possible mechanism for spontaneous loss of viable cores could be inhibition. For example, if one of the molecules in a newly formed irrRAF actually *inhibits* one or more reactions in another irrRAF, this other irrRAF may be lost. Inhibition is currently not explicitly included in the RAF formalism. In fact, it is known that including inhibition makes the problem of finding RAF sets in arbitrary reaction networks NP-complete (Mossel and Steel 2005). However, recent work shows that if the total number of inhibitors is limited, the problem can still be tractable, and that RAF sets still exist with a significant probability (Hordijk et al. 2014b) (see also Vasas et al. (2012)). Further work on RAF theory with inhibition included is currently in progress.

Conclusions

We have provided a formal but visually clear example of how a set of minimal necessary conditions for evolvability of autocatalytic sets is satisfied in a simple polymer model of a chemical reaction system. Furthermore, we have shown how these conditions can be captured and analyzed within RAF theory. Moreover, given the results of our more general example, combined with previous results, one of these conditions (existence of multiple viable cores, or irrRAFs) is an *expected* property of a (random) reaction network once a certain average level of catalysis is reached. Other conditions (mechanisms for spontaneous gain or loss of viable cores) are also clearly satisfied, or can be incorporated within RAF theory (such as boundaries). These results clearly support the hypothesis that autocatalytic sets can be evolvable. Additional mechanisms that can provide regulation and evolvability, such as inhibition and the emergence of higher-level structures, can also be included in RAF theory, and are a topic of ongoing work.

Acknowledgments We thank the Allan Wilson Centre, New Zealand, for helping fund this work, and two anonymous reviewers for helpful suggestions to improve the manuscript.

References

Ashkenasy G, Jegasia R, Yadav M, Ghadiri MR (2004) Design of a directed molecular network. PNAS 101(30):10872–10877

Dyson FJ (1982) A model for the origin of life. J Mol Evol 18:344–350

Dyson FJ (1985) Origins of Life. Cambridge University Press
Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys 22:403–434
Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81(25): 2340–2361
Hordijk W (2013) Autocatalytic sets. From the origin of life to the economy. BioScience 63(11):877–881
Hordijk W, Steel M (2004) Detecting autocatalytic, self-sustaining sets in chemical reaction systems. J Theor Biol 227(4):451–461
Hordijk W, Steel M (2013) A formal model of autocatalytic sets emerging in an RNA replicator system. J Syst Chem 4:3
Hordijk W, Steel M (2014) Autocatalytic sets and boundaries. J Syst Chem. In revision
Hordijk W, Hein J, Steel M (2010) Autocatalytic sets and the origin of life. Entropy 12(7):1733–1742
Hordijk W, Kauffman SA, Steel M (2011) Required levels of catalysis for emergence of autocatalytic sets in models of chemical reaction systems. Int J Mol Sci 12(5):3085–3101
Hordijk W, Steel M, Kauffman S (2012) The structure of autocatalytic sets: Evolvability, enablement, and emergence. Acta Biotheoretica 60(4):379–392
Hordijk W, Hasenclever L, Gao J, Mincheva D, Hein J (2014a) An investigation into irreducible autocatalytic sets and power law distributed catalysis. Natural Computing 13(3):287–296
Hordijk W, Smith JI, Steel M (2014b) Algorithms for detecting and analysing autocatalytic sets. Algorithms for Molecular Biology. In revision
Hordijk W, Wills P, Steel M (2014c) Autocatalytic sets and biological specificity. Bull Math Biol 761: 201–224
Kauffman SA (1971) Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. J Cybern 1(1):71–96
Kauffman SA (1986) Autocatalytic sets of proteins. J Theoretical Biol 119:1–24
Kauffman SA (1993) The Origins of Order. Oxford University Press
Lifson S (1997) On the crucial stages in the origin of animate matter. J Mol Evol 44:1–8
Martin W, Russel MJ (2007) On the origin of biochemistry at an alkaline hydrothermal vent. Philos Trans R Soc B 362:1887–1925
Mossel E, Steel M (2005) Random biochemical networks. The probability of self-sustaining autocatalysis. J Theor Biol 233(3):327–336
Orgel LE (2008) The implausibility of metabolic cycles on the prebiotic earth. PLoS Biol 6(1):5–13
Segré D, Ben-Eli D, Deamer DW, Lancet D (2001) The lipid world. Orig Life Evol Biosph 31(1-2):119–145
Sievers D, von Kiedrowski G (1994) Self-replication of complementary nucleotide-based oligomers. Nature 369:221–224
Smith J, Steel M, Hordijk W (2014) Autocatalytic sets in a partitioned biochemical network. J Syst Chem 5:2
Sousa FL, Hordijk W, Steel M, Martin W (2014) Autocatalytic sets in the metabolic network of E. coli. J Syst Chem. In revision
Steel M (2000) The emergence of a self-catalysing structure in abstract origin-of-life models. Appl Math Lett 3:91–95
Steel M, Hordijk W, Smith J (2013) Minimal autocatalytic networks. J Theor Biol 332:96–107
Taran O, Thoennessen O, Achilles K, von Kiedrowski G (2010) Synthesis of information-carrying polymers of mixed sequences from double stranded short deoxynucleotides. J Syst Chem 1(9)
Vaidya N, Manapat ML, Chen IA, Xulvi-Brunet R, Hayden EJ, Lehman N (2012) Spontaneous network formation among cooperative RNA replicators. Nature 491:72–77
Vasas V, Szathmáry E, Santos M (2010) Lack of evolvability in self-sustaining autocatalytic networks constraints metabolism-first scenarios for the origin of life. PNAS 107(4):1470–1475
Vasas V, Fernando C, Santos M, Kauffman S, Sathmáry E (2012) Evolution before genes. Biol Direct 7(1)
Wächtershäuser G (2007) On the chemistry and evolution of the pioneer organism. Chem Biodivers 4: 584–602
Wills PR, Henderson L (2000) Self-organisation and information-carrying capacity of collectively autocatalytic sets of polymers: ligation systems. In: Bar-Yam Y (ed) In Unifying Themes in Complex Systems: Proceedings of the First International Conference on Complex Systems. Perseus Books, pp 613–623