**Acta Biotheoretica manuscript No.**
(will be inserted by the editor)

# The Structure of Autocatalytic Sets: Evolvability, Enablement, and Emergence

**Wim Hordijk · Mike Steel · Stuart Kauffman**

Received: date / Accepted: date

**Abstract** This paper presents new results from a detailed study of the structure of autocatalytic sets. We show how autocatalytic sets can be decomposed into smaller autocatalytic subsets, and how these subsets can be identified and classified. We then argue how this has important consequences for the evolvability, enablement, and emergence of autocatalytic sets. We end with some speculation on how all this might lead to a generalized theory of autocatalytic sets, which could possibly be applied to entire ecologies or even economies.

**Keywords** Origin of life · autocatalytic sets · evolvability · emergence · functional organization

## 1 Introduction

Origin of life research seems divided between two paradigms: *genetics-first* and *metabolism-first*. However, a common theme in both is the idea of *autocatalysis*. In fact, a third alternative, that of *collectively autocatalytic sets*, was introduced more or less independently several times [13, 6, 5], and used in later origin of life models [27, 7, 21, 16]. The idea of autocatalytic sets has not been without its criticism

Wim Hordijk
SmartAnalytiX.com
E-mail: wim@SmartAnalytiX.com

Mike Steel
Biomathematics Research Centre
University of Canterbury
Christchurch, New Zealand
E-mail: m.steel@math.canterbury.ac.nz

Stuart Kauffman
University of Vermont
Burlington, VT, USA
and
Tampere University of Technology
Tampere, Finland
E-mail: stukauffman@gmail.com

[17,20,26], but recent experimental advances in creating such sets in a laboratory setting [22,1,8,24] have sparked a renewed interest in autocatalytic sets. Moreover, there is growing evidence that simple autocatalytic cycles may indeed have been at the core of the origin of life [2].

In this paper, we continue our own (theoretical) studies of autocatalytic sets. In previous work [13–15,23,11,18,9,10,12], we used a mathematical framework to investigate the probability of existence of autocatalytic sets under various conditions (model parameters), to answer questions about the required level of catalysis needed for autocatalytic sets to emerge, and we considered various model variants. We also compared and contrasted our own work with other, related models and methods (which we will not repeat here, but see, e.g., [9]), and presented results that complement, but also go beyond what has been reported elsewhere so far.

Here, we present new results from a detailed study of the actual *structure* (as opposed to merely the existence) of autocatalytic sets. In particular, we show empirically, theoretically, and through an illustrative example, that autocatalytic sets can often be decomposed into smaller subsets which themselves are autocatalytic, and how these subsets can be identified and classified. We then argue that this structural decomposition of autocatalytic sets has important consequences for their potential *evolvability*, how they can *enable* their own growth and also the coming into existence of other autocatalytic (sub)sets, and how this can possibly give rise to higher-level, *emergent* structures. Finally, we end the paper with some provoking but plausible ideas of how the theory of autocatalytic sets, having started in the context of the origin of life, might be generalized to a theory of functional organization and emergence. Such a generalized theory of autocatalytic sets could, as we hope and envision, perhaps even be applicable to ecology and economics.

## 2 Chemical reaction systems and autocatalytic sets

In previous work we introduced and studied a formal model of chemical reaction systems and autocatalytic sets in the context of the origin of life problem [13–15, 23,11,18,9,10,12]. Here, we will only briefly review the relevant definitions and results.

A *chemical reaction system* (CRS) is defined as a tuple $Q = \{X, \mathcal{R}, C\}$ consisting of a set of molecule types $X$, a set of reactions $\mathcal{R}$ (where each reaction transforms a set of reactants into a set of products), 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. In one particular model of a CRS, known as the binary polymer model [13–15], molecule types are represented as bit strings up to a certain length $n$, reactions are simply ligation ("gluing" two bit strings together into one longer one) and cleavage (splitting one bit string into two shorter ones), and catalysis is assigned at random according to some parameter $p$ (the probability that a molecule type catalyzes a reaction). The food set consists of all molecule types up to a certain length $t \ll n$.

Informally, an *autocatalytic set* (or RAF set, in our terminology) 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.

A more formal definition of RAF sets is provided in [11,10], where we also introduced a polynomial-time (in the size of the reaction set $\mathcal{R}$) algorithm for finding RAF sets in a general CRS $Q = \{X, \mathcal{R}, C\}$. Figure 1 shows an example of an RAF set that was found by our algorithm in an instance of the binary polymer model with parameter values $n = 5$, $t = 2$, and $p = 0.0045$.

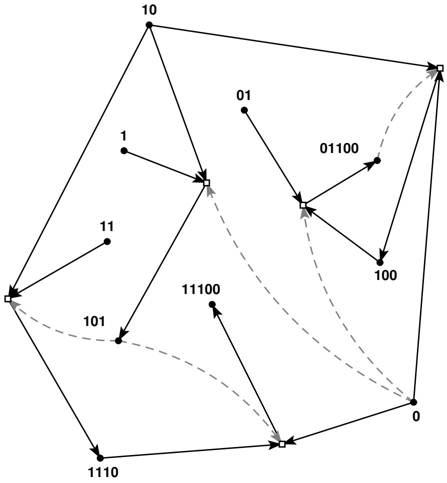

**Fig. 1** An example of an RAF set that was found by the RAF algorithm in an instance of the binary polymer model. Molecule types are represented by black dots and reactions by white boxes. Solid arrows indicate reactants and products coming in and out of a reaction, while dashed arrows indicate catalysis. The food set is $F = \{0, 1, 01, 10, 11\}$.

In [11] we showed computationally, and then proved theoretically in [18], that only a linear growth rate in the level of catalysis $f$ (i.e., the average number of reactions catalyzed per molecule type) with the size of the largest molecules $n$, suffices to get RAF sets with high probability in instances of the binary polymer model. Furthermore, the level of catalysis only needs to be roughly $1 < f < 2$, i.e., between one and two reactions catalyzed per molecule (for $n$ at least up to 20), which is a chemically realistic number. In [10,12] we studied a variant of the binary polymer model where catalysis is based on a more realistic template-matching constraint. However, the main results from the basic model remain the same and, moreover, can be mathematically predicted [12].

Finally, we note that the RAF sets that are found by our algorithm are what we refer to as *maximal* RAF sets (maxRAFs). However, a maxRAF could possibly consist of several smaller (independent or overlapping) subsets which themselves are RAF sets (subRAFs). If such a subRAF cannot be reduced any further without losing the RAF property, we refer to it as an *irreducible* RAF (irrRAF). In [11] we presented an extension of the RAF algorithm to find one (arbitrary) irrRAF within a given maxRAF. These notions of subRAFs are relevant for what follows below.

## 3 The structure of autocatalytic sets

In the original argument for the existence of autocatalytic sets in the binary polymer model, it was assumed that when the probability of catalysis $p$ is slowly increased (for a given $n$), at some point an autocatalytic set will occur as a "giant component", i.e., containing all (or most of) the reactions in the reaction set $\mathcal{R}$ [13–15]. This is similar, it was claimed, to the sharp phase transition observed in percolation networks or random graphs. Furthermore, the proof that only a linear growth rate in the level of catalysis $f$ is sufficient to get RAF sets with arbitrary high probability assumes that RAF sets contain the entire molecule set $X$ [18]. However, as was already shown in [11, 10], in practice these assumptions are stronger than necessary. Here, we investigate this issue of the size and structure of RAF sets in more detail.

### 3.1 Empirical results

First, we look at the *relative* size of RAF sets. For maxRAFs, we calculate the relative size as the size (in number of reactions) of a maxRAF divided by the size of the full reaction set $\mathcal{R}$ it was found in. For irrRAFs, we calculate the relative size as the size of an irrRAF divided by the size of the maxRAF it is part of. For various values of $n$ in the binary polymer model, we measured these relative sizes of maxRAFs and irrRAFs at a level of catalysis for which the probability of finding (max)RAF sets (using our RAF algorithm) is around $P_n = 0.5$. We then averaged this over a sample of instances of the model (given the practical computational limitations, as the size of the reaction set $\mathcal{R}$ increases exponentially with $n$). Figure 2 shows the results.

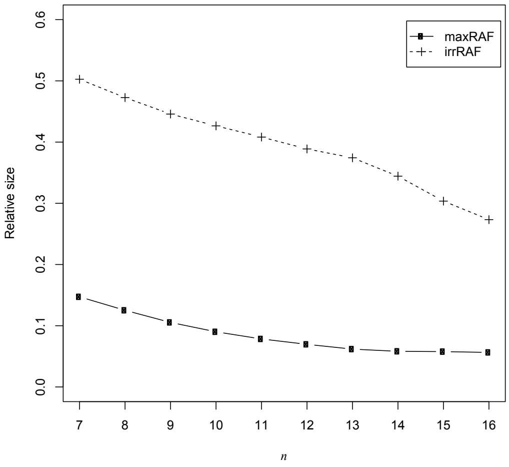

**Fig. 2** The (average) relative size of maxRAFs and irrRAFs for increasing $n$ and $P_n \approx 0.5$.

This plot clearly shows that maxRAFs are not necessarily giant components. In fact, for smaller values of $n$ the maxRAFs are only about 15% of the size of the full reaction set, decreasing to about 6% for $n = 16$. Furthermore, irrRAFs tend to be about half the size of a maxRAF for smaller values of $n$, quickly decreasing to close to one quarter for $n = 16$ (with a continuing decreasing trend). This seems to imply that maxRAFs indeed consist of multiple (possibly overlapping) irrRAFs, plus perhaps a number of reactions that are part of the maxRAF but not necessarily of any irrRAF.

### 3.2 Theoretical results

Next, we state several mathematical results (see Theorem 1 below) concerning the structure of RAF sets, which confirm and formalize the implications that follow from the empirical results. The first result shows that an RAF set may indeed contain many irreducible RAF sets, in fact, possibly exponentially many (see Fig. 3 for a simple example). One immediate consequence of this is that there is no hope of devising a polynomial-time algorithm that can guarantee to find *all* irrRAFs in an arbitrary RAF set (though an interesting, but still open question is whether irrRAFs can be merely counted in polynomial time, or can be listed by an algorithm which runs in polynomial time in the number of irrRAFs and the size of the RAF set). We will return to this issue with another, but more positive consequence (from an evolutionary point of view) with the example in Section 3.3.

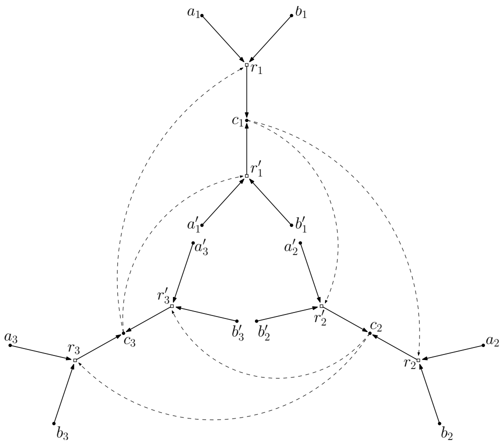

**Fig. 3** An RAF set containing eight (overlapping) irreducible RAF sets, which illustrates the type of construction used to create $2^k$ irrRAFs (for $k = 3$) in the proof of Theorem 1(1).

For the second part we first need to introduce one more definition. A *maximal proper subRAF* $\mathcal{R}''$ of an RAF $\mathcal{R}'$ is a proper subRAF $\mathcal{R}'' \subset \mathcal{R}'$ (i.e., $\mathcal{R}''$ is an RAF and is contained in but not equal to $\mathcal{R}'$) such that there is no other proper subRAF $\mathcal{R}^* \subset \mathcal{R}'$ with $\mathcal{R}'' \subset \mathcal{R}^* \subset \mathcal{R}'$. Part 2 of Theorem 1 shows that the number of maximal proper subRAFs of an RAF can only grow linearly with the size of the RAF. Thus, although an RAF can have many *minimal* (i.e., irreducible) proper subRAFs, it cannot have too many *maximal* proper subRAFs.

The third part of Theorem 1 includes three related points. The first point will be used (in Section 3.4) to show how we can represent all the subRAFs of an RAF by constructing a (Hasse) diagram; this provides a convenient way to visualize how the subRAFs sit inside each other. Next, we show there is an efficient (polynomial-time) method to determine whether any given RAF set can be decomposed into two (overlapping or disjoint) subRAFs. Together these two results provide a way to describe the possible pathways through which any particular RAF set might be built up by allowing subRAFs to combine in pairs, or to co-opt other reactions (see also the example in Section 3.3). As a third point, we show that it is possible to efficiently determine whether or not any particular reaction or, more generally, any non-empty set of reactions, is "essential" within any given RAF, in the sense that *any* subRAF of the original RAF must contain this given set of reactions.

The results described above are stated formally in the following theorem, the proof of which is provided in the Appendix.

**Theorem 1**

1. *There exist RAF sets $\mathcal{R}'$ for which the number of irreducible RAF subsets is exponential in the number of molecules and reactions in $\mathcal{R}'$.*
2. *For any RAF set $\mathcal{R}'$, the number of maximal proper subRAFs of $\mathcal{R}'$ can never exceed $|\mathcal{R}'|$.*
3. *Given a catalytic reaction system, $Q = (X, \mathcal{R}, C)$, a food set $F \subset X$, and an RAF $\mathcal{R}' \subseteq \mathcal{R}$, there exist polynomial time (in $|Q|$) algorithms that solve the following problems:*
   *(i) generate a list of all the maximal proper subRAFs of $\mathcal{R}'$;*
   *(ii) determine whether or not $\mathcal{R}'$ is the union of two proper subRAFs, and if so find all such pairs of subRAFs;*
   *(iii) for any given non-empty subset of $\mathcal{R}'$, determine whether that subset is contained in every subRAF of $\mathcal{R}'$.*

### 3.3 An illustrative example

To illustrate this idea of the decomposition of an RAF set into smaller subsets, we provide a simple but illustrative example. Recall the maxRAF shown in Figure 1, which consists of the following five reactions (the catalysts are shown above the reaction arrows):

$$\begin{aligned}
r_1 &: 10 + 0 \xrightarrow{01100} 100\\
r_2 &: 01 + 100 \xrightarrow{0} 01100\\
r_3 &: 10 + 1 \xrightarrow{0} 101\\
r_4 &: 11 + 10 \xrightarrow{101} 1110\\
r_5 &: 1110 + 0 \xrightarrow{101} 11100
\end{aligned}$$

The food set is $F = \{0, 1, 00, 01, 10, 11\}$, i.e., all molecules up to length $t = 2$. This maxRAF can actually be decomposed into two independent subsets, both of which are RAF sets themselves. The first subRAF is $\mathcal{R}_1 = \{r_1, r_2\}$ and the second subRAF is $\mathcal{R}_2 = \{r_3, r_4, r_5\}$. These two subRAFs are shown in Figure 4.

The subRAFs $\mathcal{R}_1$ and $\mathcal{R}_2$ are independent in the sense that they do not share any reactions. The only overlap is in the two food molecules 0 and 10 which are used

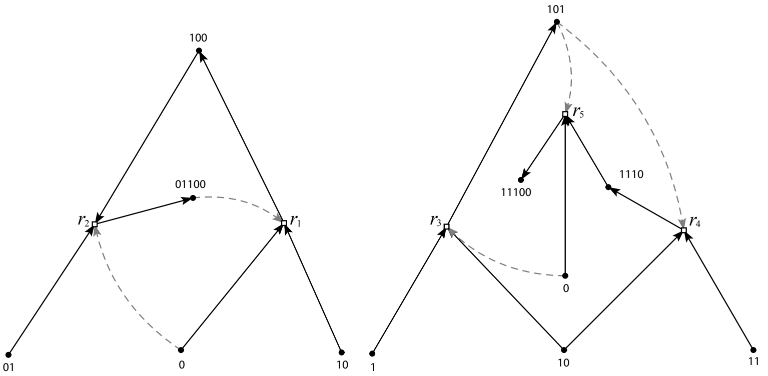

**Fig. 4** The two independent subRAFs $\mathcal{R}_1$ (left) and $\mathcal{R}_2$ (right) of which the maxRAF shown in Figure 1 is composed.

in both subRAFs as reactants and a catalyst. Note, furthermore, that subRAF $\mathcal{R}_1$ is an irreducible RAF. None of the two reactions $r_1$ or $r_2$ can be removed without losing the RAF property. However, subRAF $\mathcal{R}_2$ is not an irrRAF, as first $r_5$ and then $r_4$ can be removed without losing the RAF property, leaving an irrRAF consisting of just one reaction ($r_3$, in which both reactants and the catalyst are food molecules). We discuss further details and implications of this example in section 4 below.

### 3.4 Classifying autocatalytic subsets

The collection of subRAFs of any RAF $\mathcal{R}$ form a partially ordered set (i.e., a *poset*) under inclusion, and it is convenient to visualize this poset by its associated *Hasse diagram* [3]. This is a directed graph whose nodes are labeled by subRAFs of $\mathcal{R}$, with the top node (root) labeled $\mathcal{R}$ and the bottom nodes (leaves) labeled by the irrRAFs of $\mathcal{R}$. In this diagram we place an (upward oriented) edge from node $\mathcal{R}''$ to node $\mathcal{R}'$ precisely if $\mathcal{R}''$ is a maximal proper subRAF of $\mathcal{R}'$. For example, the Hasse diagram of the subRAFs of the 5-reaction maxRAF (from Figure 1) is shown in Figure 5. This Hasse diagram displays all the possible ways that the maxRAF can be built up from simpler subRAFs, starting from one or more irrRAFs.

In general, the poset of subRAFs of an RAF set can be very large. For example, as we have shown in Theorem 1(1), an RAF set may have exponentially many irrRAFs, and each will appear as a separate node at the bottom of the Hasse diagram. Thus, there is no algorithm for constructing the Hasse diagram of the subRAFs that is guaranteed to run in polynomial time in $|Q|$ for all cases. However, Theorem 1(3(i)) provides us with the next best thing – namely, an algorithm that will be fast when the Hasse diagram is not too large. We formalize this as follows; a proof is provided in the Appendix.

**Corollary 1** *Given a catalytic reaction system, $Q = (X, \mathcal{R}, C)$, a food set $F \subset X$, and an RAF $\mathcal{R}' \subseteq \mathcal{R}$, there is an algorithm for constructing the Hasse diagram of the poset $P$ of subRAFs of $\mathcal{R}'$ whose running time is polynomial in the size of $P$ and $\mathcal{R}'$.*

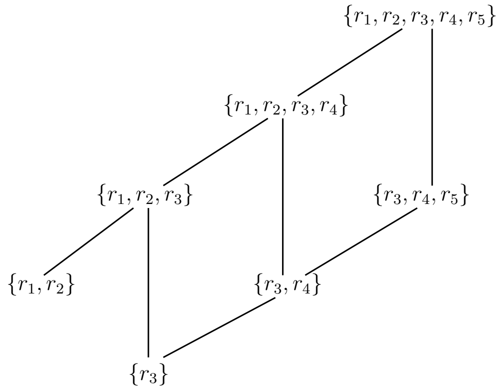

**Fig. 5** The Hasse diagram of the poset of subRAFs of the maxRAF shown in Figure 1.

## 4 Evolvability, enablement, and emergence

The "decomposability" of RAF sets, as detailed in the previous section, actually gives rise to a possible mechanism for evolution to happen in autocatalytic sets. As was convincingly shown only very recently [25], the process of combining, splitting, and recombining different subRAFs can give rise to inheritance, mutation, and competition, i.e., indeed *evolvability.* For example, the first subset $\mathcal{R}_1 = \{r_1, r_2\}$ is what is called a "viable core" in [25]. This subset actually needs to be "seeded" by either the molecule type 100 or 01100, but once that happens, it will be able to sustain itself. This seeding can happen, for example, by one of the two reactions $\{r_1, r_2\}$ happening spontaneously (i.e., uncatalyzed, which can always happen but at a much lower rate), or one of the two required molecules being produced by perhaps another subRAF (assuming that the 5-reaction RAF of Figure 1 is itself a subset of a larger RAF set).

The second subset $\mathcal{R}_2 = \{r_3, r_4, r_5\}$ depends only on food molecules, and so will always exist (given that food molecules are always available). However, imagine that molecule type 11 is not a food molecule, but is a product of some other subRAF. Then, once this molecule is available, the subset $\mathcal{R}_2$ forms another viable core (once $r_3$ and $r_4$ can happen, $r_5$ will automatically happen as well). Now, these subRAFs (viable cores) can be combined and recombined in various combinations. In this case, we can have $\{r_3\}$, $\{r_1, r_2, r_3\}$, $\{r_3, r_4, r_5\}$, and $\{r_1, r_2, r_3, r_4, r_5\}$ as possible combinations. This, as argued and shown in [25], is exactly what allows adaptation and evolution to happen. And it is in this context that a possibly exponential number of irrRAFs that can exist within a given maxRAF has an important (and positive, in terms of evolvability) consequence.

Next to evolvability, the example above also illustrates how RAF (sub)sets can *enable* their own growth or, more importantly, each others coming into existence. Returning to subRAF $\mathcal{R}_2$ in Figure 4, once reactions $r_3$ and $r_4$ can happen (e.g., seeded by molecule type 11), reaction $r_5$ is automatically added to the set as well, given that it depends on one food molecule and the products of $r_3$ and $r_4$. So, existence of the subset $\{r_3, r_4\}$ automatically enables its own growth to the 3-reaction subRAF $\mathcal{R}_2$. Furthermore, as the example implies, subRAFs can be (possibly mutually) dependent on each other, and the existence of one subRAF can create the required conditions (act as a "seeder") for another to come into existence.

Finally, taking this one step further, one could imagine a collection of mutually dependent RAF sets forming a meta-RAF set: one set enabling (catalyzing) the existence of another, in mutually beneficial ways. In other words, self-sustaining, functionally closed structures can arise at a higher level (an autocatalytic set of autocatalytic sets), i.e., true *emergence*. And this, in turn, opens up the possibility of *open-ended* evolution.

## 5 A generalized theory of autocatalytic sets

We end this paper by discussing an as yet speculative, but potentially powerful and far-reaching idea. By definition, RAF sets are self-sustaining entities (supported by a food set) that exhibit catalytic closure. As we have argued here (largely inspired by [25]), they can give rise to evolvability, by virtue of their property of being decomposable into (possibly exponentially many) subsets. As such, the Hasse diagram of a (maximal) RAF set, as described above, can be interpreted as containing possible "paths" that evolution could potentially follow by combining, splitting, and recombining RAF subsets into different variants. Furthermore, as also argued here, RAF sets can enable their own growth, or even the coming into existence of other RAF sets, creating mutually dependent collections of RAF sets, and possibly emergent meta-RAF sets, thus giving rise to open-ended evolution.

So, if collections of molecules and chemical reactions between them, "facilitated" by catalysis, can form autocatalytic sets, perhaps even at several emergent levels, then this begs the question: "Could we consider a complete cell as an (emergent) autocatalytic set?" We argue that this may indeed be the case, particularly for autotrophs. And once autotrophs existed, they enabled the coming into existence of heterotrophs, i.e., collections of mutually dependent cells that feed on each other's waste or side products (or simply on each other). Perhaps it is not too far-fetched to think, for example, of the collection of bacterial species in your gut (several hundreds of them) as one big autocatalytic set.

Taking this a step further, why not consider any ecology of mutually dependent organisms as an emergent autocatalytic set, with one (group of) species enabling the evolution of (i.e., *creating niches for*) other, new, species. And what about the economy? If we view the process of transforming raw materials (reactants) into products as a "production function" (the equivalent of a chemical reaction), with objects like hammers, conveyor belts, and factory machines as "facilitators" (the equivalent of catalysts, which themselves are products of other production functions), perhaps we can also view the economy as an (emergent) autocatalytic set, exhibiting some sort of functional closure. And, as with an ecology, the existence of one or more autocatalytic subsets (economic agents such as companies or government organizations) enables the coming into existence of new ones.

We admit, once more, that all this is perhaps rather speculative, but at the same time we believe that these ideas are worth pursuing and developing further. In fact, the theory of autocatalytic sets, as we have only begun to formulate in the context of the origin of life, could perhaps be generalized into a theory of functional organization, and possibly also of emergence (collections of autocatalytic sets forming meta-autocatalytic sets). The notion that certain aspects of networks are universal to life is quite common (see e.g. [4,19]), but the specific ideas presented here would seem to provide a new and potentially powerful approach. As such, a "generalized theory of autocatalytic sets" might well play an important role in understanding living systems and their organization and evolution.

**Acknowledgements** This paper was finalized while WH and SK were visiting the Computational Systems Biology Research Group of the Tampere University of Technology, Finland. MS thanks the Royal Society of New Zealand for funding support. We also thank Vera Vasas for helpful and stimulating discussions.

## References

1. Ashkenasy, G., Jegasia, R., Yadav, M., Ghadiri, M.R.: Design of a directed molecular network. PNAS **101**(30), 10,872–10,877 (2004)
2. Braakman, R., Smith, E.: The emergence and early evolution of biological carbon-fixation. PLoS Computational Biology **8**(4), e1002,455 (2012)
3. Cameron, P.J.: Combinatorics: Topics, Techniques, Algorithms. Cambridge University Press (1995)
4. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press (2003)
5. Dyson, F.J.: A model for the origin of life. Journal of Molecular Evolution **18**, 344–350 (1982)
6. Eigen, M., Schuster, P.: The hypercycle: a principle of natural self-organization. Part A: Emergence of the hypercycle. Naturwissenschaften **64**, 541–565 (1977)
7. Gánti, T.: Biogenesis itself. Journal of Theoretical Biology **187**, 583–593 (1997)
8. Hayden, E.J., von Kiedrowski, G., Lehman, N.: Systems chemistry on ribozyme self-construction: Evidence for anabolic autocatalysis in a recombination network. Angewandte Chemie International Edition **120**, 8552–8556 (2008)
9. Hordijk, W., Hein, J., Steel, M.: Autocatalytic sets and the origin of life. Entropy **12**(7), 1733–1742 (2010)
10. Hordijk, W., Kauffman, S.A., Steel, M.: Required levels of catalysis for emergence of autocatalytic sets in models of chemical reaction systems. International Journal of Molecular Sciences **12**(5), 3085–3101 (2011)
11. Hordijk, W., Steel, M.: Detecting autocatalytic, self-sustaining sets in chemical reaction systems. Journal of Theoretical Biology **227**(4), 451–461 (2004)
12. Hordijk, W., Steel, M.: Predicting template-based catalysis rates in a simple catalytic reaction model. Journal of Theoretical Biology **295**, 132–138 (2012)
13. Kauffman, S.A.: Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. Journal of Cybernetics **1**(1), 71–96 (1971)
14. Kauffman, S.A.: Autocatalytic sets of proteins. Journal of Theoretical Biology **119**, 1–24 (1986)
15. Kauffman, S.A.: The Origins of Order. Oxford University Press (1993)
16. Letelier, J.C., Soto-Andrade, J., Abarzúa, F.G., Cornish-Bowden, A., Cárdenas, M.L.: Organizational invariance and metabolic closure: Analysis in terms of (M;R) systems. Journal of Theoretical Biology **238**, 949–961 (2006)
17. Lifson, S.: On the crucial stages in the origin of animate matter. Journal of Molecular Evolution **44**, 1–8 (1997)
18. Mossel, E., Steel, M.: Random biochemical networks: The probability of self-sustaining autocatalysis. Journal of Theoretical Biology **233**(3), 327–336 (2005)
19. Newman, M.E.J.: Networks: An Introduction. Oxford University Press (2010)
20. Orgel, L.E.: The implausibility of metabolic cycles on the prebiotic earth. PLoS Biology **6**(1), 5–13 (2008)
21. Rosen, R.: Life Itself. Columbia University Press (1991)
22. Sievers, D., von Kiedrowski, G.: Self-replication of complementary nucleotide-based oligomers. Nature **369**, 221–224 (1994)
23. Steel, M.: The emergence of a self-catalysing structure in abstract origin-of-life models. Applied Mathematics Letters **3**, 91–95 (2000)

24. Taran, O., Thoennessen, O., Achilles, K., von Kiedrowski, G.: Synthesis of information-carrying polymers of mixed sequences from double stranded short deoxynucleotides. Journal of Systems Chemistry **1**(9) (2010)
25. Vasas, V., Fernando, C., Santos, M., Kauffman, S., Sathmáry, E.: Evolution before genes. Biology Direct **7**, 1 (2012)
26. Vasas, V., Szathmáry, E., Santos, M.: Lack of evolvability in self-sustaining autocatalytic networks constraints metabolism-first scenarios for the origin of life. PNAS **107**(4), 1470–1475 (2010)
27. Wächterhäuser, G.: Evolution of the first metabolic cycles. PNAS **87**, 200–204 (1990)

## Appendix

*Proof of Theorem 1:*

*Part 1:* First, consider a directed graph $G$ that has $2k$ vertices $r_1, r_2, ..., r_k$, and $r'_1, r'_2, ..., r'_k$. For each $i = 1, 2, \ldots, k-1$, place a directed edge from $r_i$ to $r_{i+1}$ and also one from $r_i$ to $r'_{i+1}$. Next, for each $i = 1, 2, ..., k-1$, place a directed edge from $r'_i$ to $r_{i+1}$ and also one from $r'_i$ to $r'_{i+1}$. Finally place directed edges from $r_k$ back to $r_1$ and to $r'_1$; similarly place directed edges from $r'_k$ back to $r_1$ and to $r'_1$.

Notice that the number of minimal directed cycles in this digraph is $2^k$, since we have complete freedom to select $r_i$ or $r'_i$ at each step in the cycle, and we must select one of them (to get a cycle) but not more than one (to get a minimal cycle).

We now use this graph to construct an RAF set that has exponentially many irrRAFs as follows. Associate with $r_i$ the reaction $a_i + b_i \rightarrow c_i$ and with $r'_i$ the reaction $a'_i + b'_i \rightarrow c_i$, where:

(i) the $a_i, b_i, a'_i, b'_i$ and $c_i$ are all distinct from each other (and across different choices of $i$ there is no repetition), and
(ii) the $a_i, b_i, a'_i, b'_i$ are all in the food set $F$ (for all $i$).

For the catalysis set $C$, we let $c_i$ catalyze $r_{i+1}$ and $r'_{i+1}$ (for $i = 1, 2, \ldots, k-1$). In addition, let $c_k$ catalyze $r_1$ and $r'_1$. Figure 3 illustrates this RAF set for the case $k = 3$.

The irrRAFs in this resulting RAF set are now in one-to-one correspondence with the minimal directed cycles of the graph $G$ described above, and there are $2^k$ such minimal cycles, but only $2k$ reactions and $5k$ molecules. So, the number of irrRAFs is exponential in the size of the RAF set. Notice that this construction can be carried out within the binary polymer model.

*Part 2:* For an arbitrary subset $\mathcal{R}'' \subseteq \mathcal{R}$, let $s(\mathcal{R}'')$ denote the (possibly empty) subset of $\mathcal{R}$ obtained by applying the RAF algorithm to $\mathcal{R}''$ and $F$, and let $\mathcal{R}''_{\neq\emptyset}$ be the set of reactions $r$ in $\mathcal{R}''$ for which $s(\mathcal{R}'' - \{r\}) \neq \emptyset$. We first establish the following result:

*Claim 1:* If $\mathcal{R}'$ is any RAF, then $\mathcal{R}''$ is a maximal proper subRAF of $\mathcal{R}'$ if and only if

(a) $\mathcal{R}'' = s(\mathcal{R}' - \{r\})$ for some reaction $r \in \mathcal{R}'_{\neq\emptyset}$, and
(b) $\mathcal{R}''$ is not strictly contained within any other set of type (a).

To verify this claim, suppose that $A$ is a maximal proper subRAF of $\mathcal{R}'$. Then there is at least one reaction $r \in \mathcal{R}' - A$. Notice that, since $A \subseteq \mathcal{R}' - \{r\}$, $s(A) = A$ is a non-empty subset of $s(\mathcal{R}' - \{r\})$; moreover $s(\mathcal{R}' - \{r\})$ is a strict subRAF of $\mathcal{R}'$ since $s(\mathcal{R}' - \{r\})$ does not include $r$ while $\mathcal{R}'$ does. Thus, since $A$ is a maximal proper subRAF of $\mathcal{R}$ we have

$$A = s(A) = s(\mathcal{R}' - \{r\}),$$

and so (a) holds. Property (b) now follows by the maximality assumption.

Conversely, suppose that (a) and (b) hold for $\mathcal{R}''$. Then $\mathcal{R}'' = s(\mathcal{R}' - \{r\})$ is nonempty and so $s(\mathcal{R}' - \{r\})$ is a proper subRAF of $\mathcal{R}'$, and if it were not a maximal proper subRAF of $\mathcal{R}'$ then, from the first part of the proof $s(\mathcal{R}' - \{r\})$ would need to be strictly contained within $s(\mathcal{R}' - \{r'\})$ for some reaction $r' \in \mathcal{R}'_{\neq\emptyset}$, and this is impossible since we are assuming that (b) holds.

From Claim 1, the number of maximal proper subRAFs is at most the number of sets of the form $s(\mathcal{R}' - \{r\})$ for $r \in \mathcal{R}'$, and there are at most $|\mathcal{R}'|$ such sets across the possible choices of $r$ from $\mathcal{R}'$.

*Part 3:* Part (i) follows directly from Claim 1, since the collection of RAF sets $\{s(\mathcal{R}' - \{r\}) : r \in \mathcal{R}'_{\neq\emptyset}\}$ can be computed in polynomial time, and property (b) in Claim 1 can then also be checked in polynomial time.

Part (ii) also follows from Claim 1, since this shows that $\mathcal{R}'$ is the union of two proper subRAFs if and only if

$$\mathcal{R}' = s(\mathcal{R}' - \{r_1\}) \cup s(\mathcal{R}' - \{r_2\}) \tag{1}$$

for some pair of distinct elements $r_1, r_2$ of $\mathcal{R}'_{\neq\emptyset}$.

From this, it is clear how to obtain a polynomial time algorithm: first construct the set $\mathcal{R}'_{\neq\emptyset}$, and, provided this set is non-empty, search for all pairs $r_1, r_2 \in \mathcal{R}'_{\neq\emptyset}$ for which Eqn. (1) holds; for each such pair we can set $\mathcal{R}_i := s(\mathcal{R}' - \{r_i\})$, for $i = 1, 2$ so that $\mathcal{R}' = \mathcal{R}_1 \cup \mathcal{R}_2$. If no such pair $r_1, r_2$ exists (or if $\mathcal{R}'_{\neq\emptyset}$ is empty), then report that $\mathcal{R}'$ cannot be decomposed further. This completes the proof of the part (ii).

For part (iii), it suffices to verify the following:

*Claim 2:* If $\mathcal{R}'$ is any RAF set and $\mathcal{R}_0$ is any non-empty subset of $\mathcal{R}'$ then $\mathcal{R}_0$ is contained within every subRAF of $\mathcal{R}'$ if and only if $s(\mathcal{R}' - \{r\}) = \emptyset$ for all $r \in \mathcal{R}_0$.

To verify this claim, first suppose there exists $r \in \mathcal{R}_0$ with $s(\mathcal{R}' - \{r\}) \neq \emptyset$. Then $s(\mathcal{R}' - \{r\})$ is a subRAF of $\mathcal{R}'$ and yet RAF $s(\mathcal{R}' - \{r\})$ does not contain $\mathcal{R}_0$, since $s(\mathcal{R}' - \{r\})$ is a subset of $\mathcal{R}' - r$ and so does not contain $r \in \mathcal{R}_0$. Conversely, suppose there exists a subRAF $\mathcal{R}''$ of $\mathcal{R}'$ which does not contain $\mathcal{R}_0$. Select any reaction $r \in \mathcal{R}_0 - \mathcal{R}''$. Then $\mathcal{R}'' \subseteq s(\mathcal{R}' - \{r\})$ and so $s(\mathcal{R}' - \{r\}) \neq \emptyset$. This establishes Claim 2, as required, and completes the proof.

*Proof of Corollary 1:* The algorithm constructs the Hasse diagram from the top down, starting from the single node $\mathcal{R}'$. We apply Part 3(i) of Theorem 1 to list all the maximal proper subRAFs of $\mathcal{R}'$, and then place edges from each of these to $\mathcal{R}'$ (if $\mathcal{R}'$ has no maximal proper subRAFs then $\mathcal{R}'$ is irreducible and we leave the node as it is). Now we repeat this step recursively on these subRAFs, introducing edges as before, and also identifying any two (or more) nodes labeled by the same subRAF. We continue in this way until the network can be extended no further, in which case all the nodes with no children comprise the set of irrRAFs of $\mathcal{R}'$.

The resulting network $N$ that we have constructed contains all the nodes of the Hasse diagram of the poset (i.e. it contains all the subRAFs of $\mathcal{R}'$); moreover, the edge set is a subset of the edges in the Hasse diagram. This last claim needs a short proof: if we have constructed an edge in $N$ from $\mathcal{R}_1$ to $\mathcal{R}_2$, where $\mathcal{R}_1 \subset \mathcal{R}_2$ we need to show that there is no other path in $N$ from $\mathcal{R}_1$ to $\mathcal{R}_2$ via a sequence of increasing subRAFs (which would make the edge $(\mathcal{R}_1, \mathcal{R}_2)$ redundant). Suppose there were such a second path, and let $(\mathcal{R}_3, \mathcal{R}_2)$ be the last edge on this path. Then, referring to Claim 1 (in the proof of Part 2 of Theorem 1), $\mathcal{R}_1 = s(\mathcal{R}_2 - \{r\})$ would be strictly contained in $\mathcal{R}_3 = s(\mathcal{R}_2 - \{r'\})$ for some reactions $r, r'$ and this is forbidden in allowing $\mathcal{R}_1$ to be selected as a maximal proper subRAF of $\mathcal{R}_2$.

Thus, each edge in $N$ will be present as an edge in the Hasse diagram. Moreover, all edges in the Hasse diagram are present in $N$, for suppose that in the Hasse diagram there is an edge from $\mathcal{R}_1$ to $\mathcal{R}_2$, where $\mathcal{R}_1 \subset \mathcal{R}_2$. Then $\mathcal{R}_1$ must be a maximal subRAF of $\mathcal{R}_2$ and so, by construction, the algorithm inserts an edge from $\mathcal{R}_1$ to $\mathcal{R}_2$ during the step at which the subRAF $\mathcal{R}_2$ and its maximal subRAFs are considered.

In summary, we have verified that the algorithm described constructs exactly the Hasse diagram of subRAFs of $\mathcal{R}'$.