Autocatalytic Sets: From the Origin of Life to the Economy

WIM HORDIJK

The origin of life is one of the most important but also one of the most difficult problems in science. Autocatalytic sets are believed to have played an important role in the origin of life. An autocatalytic set *is a collection of molecules and the chemical reactions between them, such that the set as a whole forms a functionally closed and self-sustaining system. In this article, I present an overview of recent work on the theory of autocatalytic sets and on how this theory can be used to study the probability of emergence, the overall structure, and the further evolution of such systems, both in simple mathematical models and in real chemical systems. I also describe some (still speculative) ideas of how this theory can potentially be applied to living systems in general and perhaps even to social systems such as the economy.*

Keywords: complex systems, modeling, life history, autocatalytic sets

Thinking of biology means thinking about life. And thinking about life inevitably brings up the questions *What is life?* and *How did it start?* All life on Earth, from the simplest bacteria to human beings, is based on intricate molecular machinery—a complex interplay among RNA, DNA, and proteins. Within this network of molecules and chemical reactions, there are many mutual dependencies and interactions, such that the network as a whole forms a functionally closed and self-sustaining system. In other words, a living system is able to produce, from whatever resources are available in its environment, those molecules that are necessary for its own growth, maintenance, and reproduction. However, the probability that such a complex and functionally complete system emerged all at once from nonliving molecules and basic chemistry seems extremely small—too small for some to take the notion of a spontaneous origin of life very seriously.

Starting in the early 1970s, several theories were developed to try to explain the origin (and workings) of life on the basis of this idea of a functionally closed and self-sustaining system (Eigen and Schuster 1979, Maturana and Varela 1980, Dyson 1985, Rosen 1991, Kauffman 1993, Gánti 2003). Some of these theories remained mostly at a conceptual level. Others were worked out in great mathematical detail, showing how such a system, if it exists (a big *if*!), is indeed stable and can survive and maintain itself. However, none of them provided a real fundamental insight into the actual probability that such a fully functional system could emerge spontaneously or how it could be built up from smaller systems with similar properties or could grow (and evolve) into larger and more complex systems. As a consequence, these theories—although they capture an essential property of life—were not yet sufficient to explain life's origin.

The idea garnered additional interest, however, when progress was made in experimentally constructing examples of such functionally closed and self-sustaining chemical networks in the laboratory (Sievers and von Kiedrowski 1994, Ashkenasy et al. 2004, Lincoln and Joyce 2009, Vaidya et al. 2012). Although they are relatively small and artificially designed and produced, these experimental networks show the validity of the original idea. However, they still leave open the main questions about the probability of spontaneous emergence and these systems' ability to grow and evolve.

Taking a more mathematical and computational approach, my colleague Mike Steel, of the University of Canterbury, and I have developed a formal framework with which these remaining and important questions can be investigated in more detail. Interestingly, the results from our studies actually paint a very positive picture of the probability of and a possible mechanism for the spontaneous origin and further evolution of such functionally closed and self-sustaining systems. Moreover, the framework can be applied directly to investigate some of the mentioned experimental chemical networks, providing additional insights and predictions about their structure and behavior that would be difficult to obtain from the experiments alone. And, finally, we argue that this framework could possibly also be applicable beyond the origin of life to living systems in general, from individual

BioScience 63: 877–881. ISSN 0006-3568, electronic ISSN 1525-3244. © 2013 by American Institute of Biological Sciences. All rights reserved. Request permission to photocopy or reproduce article content at the University of California Press's Rights and Permissions Web site at *www.ucpressjournals.com/reprintinfo.asp*. doi:10.1525/bio.2013.63.11.6

cells to bacterial colonies and entire ecologies and perhaps even to social networks and the economy.

This framework therefore provides, for the first time, a solid mathematical foundation (and plausible justification) for the original idea of life emerging as a functionally closed and self-sustaining chemical reaction network. In this essay, I present a brief overview of the formal framework, its main results and application to an experimental chemical system, and how it might be generalized beyond chemistry: from the origin of life to the economy.

Autocatalytic sets of molecules and chemical reactions

The framework was originally developed in the context of a *chemical reaction system*, which can be described formally as a set (collection) of molecules; possible chemical reactions between these molecules; and, additionally, catalysts. A *catalyst* is a molecule that significantly increases the rate at which a chemical reaction happens, without being consumed in that reaction. In this context, catalysts can be viewed as providing functionality, because they determine which reactions happen at high enough rates to be relevant. In fact, without catalysts, life would most likely not be possible at all, because the chemical reactions vital for life would not happen fast enough, and they would not be synchronized with one another. Finally, we assume that there are small numbers of molecules, called the *food set*, that are assumed to be freely available from the environment. This reflects the notion that at least certain types of molecules (e.g., water, hydrogen, nitrogen, carbon dioxide, iron) would have been around on the early Earth, before the origin of life, and could be used freely as chemical building blocks.

Given such a chemical reaction system, a subset of its chemical reactions, together with the molecules involved in them, is called an *autocatalytic set* if (a) every reaction in the subset is catalyzed by at least one molecule from this subset and (b) every molecule in the subset can be produced from the food set by a series of reactions from this subset only. This two-part definition formally captures the idea of a functionally closed (part a) and self-sustaining (part b) system. The molecules mutually help (through catalysis) in each others' production, and the set as a whole can be built up and maintained (through these mutually catalyzed reactions) from a steady supply of food molecules.

Stuart Kauffman (1971) was one of the first scientists to introduce this notion of autocatalytic sets. He subsequently constructed a simple mathematical model of chemical reaction systems to argue that such autocatalytic sets will arise spontaneously (Kauffman 1986, 1993). In his model (known as the *binary polymer model*), molecules are represented by simple bit strings (sequences of zeros and ones) of maximum length *n*. The chemical reactions consist of either gluing two bit strings together into a larger one (e.g., $000 + 11 \rightarrow 00011$), or cutting one bit string into two smaller ones (e.g., $010101 \rightarrow 01 + 0101$). The molecules (bit strings) are then assigned randomly, with a given probability, *p*, as catalysts for the possible reactions. In other words, there is a probability, *p*, that an arbitrary molecule will catalyze an arbitrary reaction. By changing the values of the parameters *n* and *p* and randomly generating the catalysis assignments, different instances of the model can be created.

Kauffman then developed a mathematical argument to show that, in his binary polymer model, given a fixed value for the probability of catalysis, *p*, and a large enough value for the maximum molecule length, *n*, the existence of autocatalytic sets is basically inevitable. However, this argument was later criticized (Lifson 1997) because it implies an exponential increase in the (average) level of catalysis. In other words, every time the maximum length *n* of the molecules (bit strings) in the model is increased by one, each molecule will end up catalyzing about twice as many reactions as before. This will indeed eventually lead to the existence of autocatalytic sets (for large enough *n*), but at a chemically unrealistically high level of catalysis. Furthermore, this notion of autocatalytic sets was also criticized for lacking evolvability (Vasas et al. 2010). In Kauffman's argument, an autocatalytic set will appear as one "giant connected component" in the chemical reaction network. This, however, implies that there is no room for change, growth, or adaptation—in other words, no possibility for the autocatalytic set to evolve.

The probability and structure of autocatalytic sets

In our own work, Steel and I have developed and investigated Kauffman's original idea and its subsequent criticisms more formally. First, we formulated the notion of autocatalytic sets in a rigorous mathematical way, calling them *RAF sets* (for *reflexively autocatalytic and food-generated*; see box 1). Next, we developed an efficient computer algorithm to detect RAF sets in general chemical reaction systems. Then, we applied this RAF algorithm to many instances of the binary polymer model (with different values for the parameters *n* and *p*), and collected statistics on when and how often RAF sets were found. Finally, we derived useful mathematical theorems on the basis of the formal definition of RAF sets. With these results, we were actually able to counter the main criticisms of the original autocatalytic sets idea.

First of all, the results of our computer simulations indicate that there is indeed a high probability that RAF sets exist in the binary polymer model and that only a very moderate level of catalysis is required (Hordijk and Steel 2004). On average, each molecule needs to catalyze only between one and two reactions to get autocatalytic sets with high probability (even for *n* up to 50). This is chemically very plausible; it is well known that many molecules can indeed catalyze more than one chemical reaction. Moreover, the simulation results show that only a linear increase in this level of catalysis is required (with increasing *n*), as opposed to the exponential increase in Kauffman's original argument. In fact, this linear relationship was formally proven in a subsequent mathematical analysis (Mossel and Steel 2005).

Box 1. The mathematics of RAF sets.

Mathematically, a chemical reaction system can be defined as a tuple: $Q = \{X,R,C\}$, where $X = \{x_1,x_2,\ldots,x_n\}$ is a set of molecule types; $R = \{r_1,r_2,\ldots,r_m\}$ is a set of chemical reactions such that a reaction is of the form $r_i{:}A_i \rightarrow B_i$, converting a set of reactants $A_i \subset X$ into a set of products $B_i \subset X$; and $C = \{(x,r)|x \in X, r \in R\}$ is a set of molecule-reaction pairs specifying which molecules can catalyze which reactions. Finally, there is a food set $F \subset X$.

We define the *closure*, $cl_{R'}(F)$, of the food set, F, relative to a (sub)set of reactions, $R' \subseteq R$, as the set of molecules, $W \subseteq X$, that contains the food set, F, plus all molecules that can be produced starting from the food set and using only reactions from R'. In other words, W is the unique (minimal) subset of molecules such that $F \subseteq W$ and for each reaction $r{:}A \rightarrow B$ in $R'{:}A \subseteq W \Rightarrow B \subseteq W$.

Given a chemical reaction system, $Q = \{X,R,C\}$, and a food set, $F \subset X$, an autocatalytic (RAF) set is now formally defined as a subset, $R' \subseteq R$ (plus associated molecules) such that (a) for each reaction $r \in R'$, there exists a molecule $x \in cl_{R'}(F)$ such that $(x,r) \in C$, and (b) for each reaction $r{:}A \rightarrow B$ in R', $A \subseteq cl_{R'}(F)$.

In Hordijk and Steel (2004), we introduced a polynomial-time algorithm ($O(|R|^2\log|R|)$ worst-case running time) for finding RAF sets in any chemical reaction system. In Mossel and Steel (2005), it was proved mathematically that the average number, $f = p|R|$, of reactions catalyzed per molecule need only grow linearly with the maximum molecule length, n, to obtain a high probability, P_n, of finding RAF sets in instances of the binary polymer model. This provides a theoretical confirmation of an earlier conjecture (Steel 2000) and of the computational results obtained from applying the RAF algorithm to many instances of the binary polymer model (Hordijk and Steel 2004).

An RAF set found using our algorithm (if one exists) is what we refer to as the *maximal* RAF set (maxRAF) of a chemical reaction system—that is, the union of all RAF (sub)sets contained in R. An RAF set from which no reactions can be removed without losing the RAF property is called an *irreducible* RAF set (irrRAF). We have shown that a maxRAF can (potentially) contain an exponential number of irrRAFs (Hordijk et al. 2012) and that finding even the smallest irrRAF is, in general, an NP-complete (from *nondeterministic polynomial time*) problem (Steel et al. 2013). The RAF subsets (subRAFs) of a maxRAF actually form a *partially ordered set* (poset)—that is, a hierarchical structure of subRAFs with the maxRAF at the top and the irrRAFs at the bottom. This hierarchical network (called a *Hasse diagram*) elucidates all the different ways in which subRAFs in a chemical reaction system can be (re)combined and grow into larger RAF sets, providing a mechanism for the evolution and emergence of more-complex RAF sets (Hordijk et al. 2012).

Next, we looked at the structure of the autocatalytic sets that were found by our RAF algorithm in instances of the binary polymer model. In contrast to Kauffman's original argument, in which autocatalytic sets appear as giant connected components, RAF sets are actually often composed of smaller subsets that are RAF sets, themselves. Moreover, these RAF subsets, in turn, consist of yet smaller RAF subsets. In short, there generally exists an entire hierarchy of smaller RAF sets in slightly larger RAF sets, in yet larger ones, and so on. This hierarchical structure contains all the possible ways in which RAF (sub)sets can be split up and (re)combined and can grow into bigger and more complex RAF sets (Hordijk et al. 2012). In fact, some of our colleagues, led by the well-known evolutionary biologist Eörs Szathmáry, have convincingly shown that if such multiple autocatalytic subsets exist, this can indeed lead (under certain additional conditions) to an evolutionary process, including competition and selection (Vasas et al. 2012).

To illustrate these results, figure 1 shows an example of an RAF set found using our algorithm in an instance of the binary polymer model with $n = 5$ and $p = .0045$ and taking as the food set all bit strings of length one and two. The color-outlined polygons indicate the different RAF subsets of which the full RAF set is composed. These subsets can grow or be combined in various ways to get larger RAF sets. For example, the purple subset can grow into the red subset or be combined with the yellow subset, or the blue subset can be extended with the green one, and so on. Of course, this is just a small example (to keep it visually clear); however, in realistic networks of hundreds or even thousands of chemical reactions, these hierarchical RAF structures can indeed be extremely rich and diverse.

In short, the results from our formal RAF framework (combined with results from our colleagues) seem to resolve the earlier criticisms and therefore support the notion of autocatalytic sets as a plausible and useful formalism. One additional point of criticism, however, could be that the binary polymer model is too simplistic to be chemically realistic. To address this issue, we have investigated more-realistic extensions of the model, such as a version in which the potential catalyst and the reactants in a reaction have to match in some way—for example, as in the formation of complementary base pairs in RNA molecules. The main results from this template-based catalysis model version are largely the same as those from the original model (Hordijk et al. 2011), and, moreover, they can be predicted analytically (Hordijk and Steel 2012a).

Furthermore, we have applied the RAF framework to an experimental chemical reaction system of catalytic RNA molecules—so-called *ribozymes*—in which autocatalytic networks emerge spontaneously (Vaidya et al. 2012).

Note that Vaidya and colleagues (2012) used the term *cooperative networks*. The term *cooperation* is often used incorrectly or can cause confusion in a biological or evolutionary

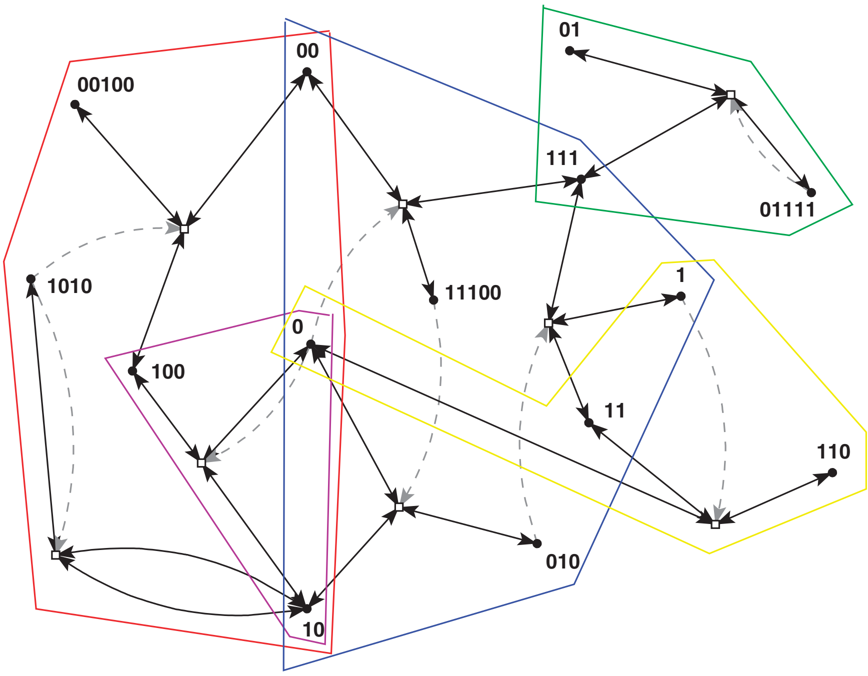

Figure 1. An example of an RAF (reflexively autocatalytic and food-generated) set in an instance of the binary polymer model. The black dots represent molecules (labeled by bit strings); the white boxes represent reactions. The solid arrows indicate molecules going in and out of a reaction (all reactions are bidirectional). The dashed arrows indicate catalysis. The food molecules are all bit strings of one or two characters. All of the reactions are catalyzed by a molecule from the RAF set, and all of the molecules can be produced from the food set by using reactions from the RAF set. The colored boxes indicate the various RAF subsets that exist within the full RAF set. Source: Reprinted with permission from Hordijk and Steel (2012b).

context (Szathmáry 2013); therefore, we prefer to use the term *autocatalytic networks*, which is less ambiguous.

Not only is our model capable of reproducing the main results from this experimental system, but it also provides additional results and predictions that would be difficult to obtain from the experiments alone (Hordijk and Steel 2013). For example, our model predicts the existence of multiple possible sequences of larger and larger RAF sets that the system can go through over time and that the autocatalytic networks of molecules appearing in this system are more robust against perturbations than an equivalent collection of "selfish" RNA molecules. These predictions can be tested directly with the actual experimental system. Our results, therefore, clearly show that the RAF framework is not restricted to the binary polymer model but can be applied directly and meaningfully to real chemical reaction systems. This provides an important first step toward merging experimental and theoretical lines of work on autocatalytic sets.

Note that the RAF framework is mostly motivated by and described in terms of chemical reactions and catalysis (in other words, *metabolism*). Of course, life as we know it depends on other aspects as well, such as information storage in the genetic system and spatial components, such as cell walls and structural proteins. These different aspects are all coupled together, as in Gánti's (2003) chemoton model. However, many of these additional aspects can also be encompassed by the RAF framework. For example, cell walls prevent molecules from diluting away and can therefore be considered catalysts for metabolic reactions to happen, whereas the molecules necessary to build and maintain the cell wall itself are produced by the metabolic reaction network. Similarly, the genetic system produces catalysts required in the metabolic network, which, in turn, produces the basic building blocks to maintain the genetic system. Even so, the RAF framework is clearly not yet complete. Other aspects that are still missing, for example, are those of inhibition, energy requirements, and reaction rates, although we are currently addressing some of these questions.

As a final note on the usefulness and applicability of the RAF framework, instead of providing a direct explanation for how life *did* happen on Earth, it addresses the more general question of how life *could* happen. In other words, it provides a model for the necessary (or minimal) conditions, from an organizational or process point of view, for a life-like system to come into existence, regardless of its actual chemical implementation. However, as was described above, the framework can, of course, be applied to very specific chemical systems, with the aim of explaining various aspects of the actual origin of life on Earth. Indeed, the RAF framework is already viewed, in some cases, as theoretical support for empirical observations (Martin and Russell 2007, Vaidya et al. 2012).

Beyond chemistry: From the origin of life to the economy

If life indeed started as an autocatalytic set of chemical reactions, which, subsequently, would have grown and evolved into larger and more-complex autocatalytic sets, eventually giving rise to the intricate RNA–DNA–protein molecular machinery on which all life is based, this immediately gives rise to the question, *Is life itself an autocatalytic set?* In other words, could a living cell (such as a bacterium) be considered an autocatalytic set in itself? This, of course, goes back to the original idea of life as a functionally closed and self-sustaining system on which the notion of autocatalytic sets is based. But what about entire colonies of bacteria, in which, for example, some bacteria live on the waste products of other bacteria or depend on the exchange of genetic material. Could such a bacterial ecosystem, with its complex network of mutual dependencies, be considered an autocatalytic set—or, rather, an autocatalytic *superset* (the colony as a whole) of autocatalytic *subsets* (the individual bacteria), similar to the hierarchical structure of RAF sets in the binary polymer model? If so, what about an entire ecology of mutually dependent species, with all of its trophic levels, symbioses, energy exchanges, and so on? These ideas, as was originally also proposed by Kauffman, do indeed sound quite attractive and powerful; with the RAF framework, we now seem to have the appropriate formalism and tools available to actually develop and test such ideas.

Finally, to take the analogy even one step further, what about the economy? Consider an economic production function, such as transforming inputs (e.g., raw materials such as wood, oil, ores) into products (e.g., chairs, plastic, cans). This can be compared with a chemical reaction transforming molecular reactants into products, and, as with chemical reactions, these economic production functions are often facilitated (catalyzed) by other items (e.g., hammers, mills, conveyor belts), which are not consumed by the production function but are, themselves, the products of yet other production functions. So, there is a complex network of dependencies, in which production functions create products, and some of these products, in turn, catalyze the production functions in a closed, self-sustaining manner. In other words, the economy can perhaps be viewed as an autocatalytic set (Kauffman 2011).

Furthermore, the economy grows and evolves, with new technologies giving rise to even more possible production functions and products, which, in turn, can trigger yet more innovations. For example, a company such as Google could only have come into existence once the Internet had been established, just as the green reaction subset in figure 1 can only exist (i.e., be viable) once the blue RAF subset exists. This is also similar to an ecosystem, in which the appearance of one species creates new niches (which did not exist before), enabling yet other species to come into existence and survive—RAF sets enabling the appearance of other RAF sets.

We hope that these new and exciting ideas will eventually lead to a generalized theory of autocatalytic sets (Hordijk and Steel 2012b, Hordijk et al. 2012) beyond chemistry and origin of life. Indeed, several ecologists, economists, and social and cognitive scientists are interested in exploring these ideas further, building on the encouraging results of the formal RAF framework. And, to return to where it all started, if living systems can truly be described in terms of hierarchies of autocatalytic sets, the RAF results seem to suggest that a spontaneous origin of life from pure chemistry is, after all, less improbable than we may have thought.

References cited

Ashkenasy G, Jegasia R, Yadav M, Ghadiri MR. 2004. Design of a directed molecular network. Proceedings of the National Academy of Sciences 101: 10872–10877.

Dyson FJ. 1985. Origins of Life. Cambridge University Press.

Eigen M, Schuster P. 1979. The Hypercycle. Springer.

Gánti T. 2003. The Principles of Life. Oxford University Press.

Hordijk W, Steel M. 2004. Detecting autocatalytic, self-sustaining sets in chemical reaction systems. Journal of Theoretical Biology 227: 451–461.

———. 2012a. Predicting template-based catalysis rates in a simple catalytic reaction model. Journal of Theoretical Biology 295: 132–138.

———. 2012b. Autocatalytic sets extended: Dynamics, inhibition, and a generalization. Journal of Systems Chemistry 3 (art. 5).

———. 2013. A formal model of autocatalytic sets emerging in an RNA replicator system. Journal of Systems Chemistry 4 (art. 3).

Hordijk W, Kauffman SA, Steel M. 2011. Required levels of catalysis for emergence of autocatalytic sets in models of chemical reaction systems. International Journal of Molecular Sciences 12: 3085–3101.

Hordijk W, Steel M, Kauffman SA. 2012. The structure of autocatalytic sets: Evolvability, enablement, and emergence. Acta Biotheoretica 60: 379–392.

Kauffman SA. 1971. Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. Journal of Cybernetics 1: 71–96.

———. 1986. Autocatalytic sets of proteins. Journal of Theoretical Biology 119: 1–24.

———. 1993. The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press.

———. 2011. Economics and the collectively autocatalytic structure of the real economy. 13.7 Cosmos and Culture, National Public Radio. (21 November 2011; *www.npr.org/blogs/13.7/2011/11/21/142594308/economics-and-the-collectively-autocatalytic-structure-of-the-real-economy*)

Lifson S. 1997. On the crucial stages in the origin of animate matter. Journal of Molecular Evolution 44: 1–8.

Lincoln TA, Joyce GF. 2009. Self-sustained replication of an RNA enzyme. Science 323: 1229–1232.

Martin W, Russell MJ. 2007. On the origin of biochemistry at an alkaline hydrothermal vent. Philosophical Transactions of the Royal Society B 362: 1887–1925.

Maturana HR, Varela FJ. 1980. Autopoiesis and Cognition: The Realization of the Living. Reidel.

Mossel E, Steel M. 2005. Random biochemical networks: The probability of self-sustaining autocatalysis. Journal of Theoretical Biology 233: 327–336.

Rosen R. 1991. Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life. Columbia University Press.

Sievers D, von Kiedrowski D. 1994. Self-replication of complementary nucleotide-based oligomers. Nature 369: 221–224.

Steel M. 2000. The emergence of a self-catalysing structure in abstract origin-of-life models. Applied Mathematics Letters 3: 91–95.

Steel M, Hordijk W, Smith J. 2013. Minimal autocatalytic networks. Journal of Theoretical Biology 332: 96–107.

Szathmáry E. 2013. On the propagation of a conceptual error concerning hypercycles and cooperation. Journal of Systems Chemistry 4 (art. 1).

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. Proceedings of the National Academy of Sciences 107: 1470–1475.

Vasas V, Fernando C, Santos M, Kauffman SA, Szathmáry E. 2012. Evolution before genes. Biology Direct 7 (art. 1).

Wim Hordijk (wim@worldwidewanderings.net) is the owner of SmartAnalytiX.com, in Lausanne, Switzerland.