Chapter 10
Network Theory in Prebiotic Evolution

Sara Imari Walker and Cole Mathis

Abstract One of the most challenging aspect of origins of life research is that we do not know precisely what life is. In recent years, the use of network theory has revolutionized our understanding of living systems by permitting a mathematical framework for understanding life as an emergent, collective property of many interacting entities. So far, complex systems science has seen little direct application to the origins of life, particularly in laboratory science. Yet, networks are important mathematical descriptors in cases where the structure of interactions matters more than counting individual component parts—precisely what we envision happens as matter transitions to life. Here, we review a few notable examples of the use of network theory in prebiotic evolution, and discuss the promise of systems approaches to origins of life. The end goal is to develop a statistical mechanics useful to origins of life—that is, one that deals with interactions of system components (rather than merely counting them) and is therefore equipped to model life as an emergent phenomena.

S. I. Walker (✉)
Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, AZ, USA

School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA

ASU-SFI Center for Biosocial Complex Systems, Arizona State University, Tempe, AZ, USA

Blue Marble Space Institute for Science, Seattle, WA, USA
e-mail: sara.i.walker@asu.edu

C. Mathis
Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, AZ, USA

Department of Physics, Arizona State University, Tempe, AZ, USA
e-mail: cole.mathis@asu.edu

C. Menor-Salván (ed.), *Prebiotic Chemistry and Chemical Evolution of Nucleic Acids*, Nucleic Acids and Molecular Biology 35,
https://doi.org/10.1007/978-3-319-93584-3_10

10.1 Introduction

One of the most challenging aspects of origin of life research is identifying those properties of life likely to be characteristic not only of life as it exists today, after >3.5 billion years of evolutionary refinement, but also at its origin. In reality, the problem is harder than even this, as we must identify properties of life that could have preceded its origin *and* could also be responsible for driving the transition to the living state. So far, the community of origin of life researchers has seen tremendous progress in synthesizing different molecular components of life, including lipids and amino and nucleic acids. This approach assumes the properties of life preceding it should include some of current life's basic molecular components (although it is still debated which ones).[1] However, to become life, these molecular components would necessarily have to interact to generate chemical systems exhibiting the emergence of increasingly "lifelike" properties. But, what are the emergent "lifelike" properties prebiotic chemists should focus on?

In biology, it is often the case we deal with highly complex interacting systems, with hundreds or thousands of components (see, e.g., Fig. 10.1). Understanding the fundamental processes driving the large-scale organization of living systems is therefore no small challenge (and it is more challenging still to distill and import relevant ideas to prebiotic chemistry). Network theory has become an indispensable tool for making sense of the mess of biology by reducing the study of complex interacting systems to the study of the statistical properties of their graphical representation. A graph (or network) is a set of nodes and edges, sometimes with additional attributes and structure. A simple example is shown in Fig. 10.2. In chemistry nodes could be molecules, where two molecules are connected by an edge if they participate in the same reaction. As a mathematical abstraction, networks have found utility in describing the large-scale statistical properties of living systems from the functioning of cells to the organization of cities. For example, studies of biochemical networks led to the discovery of the "scale-free" structure of metabolism (see Sect. 10.2), which describes a heterogeneity in the global organization of chemical reactions associated with bioenergetics, common to all three domains of life (Jeong et al. 2000). In addition to revealing organizational properties, once the structure of the graph is known, its generative and evolutionary mechanisms can be identified (Barabasi and Albert 1999) and its robustness and stability properties characterized (Larhlimi et al. 2011). Network theory therefore provides a set of mathematical tools, which could be utilized to understand not only the organization of living networks but also how this organization emerges in the first

[1] It is important to point out it is an assumption of our theories for the origin of life that the process started with molecules we would identify as biological. Alternative hypotheses, such as Cairns-Smith's "clay world" (Cairns-Smith 1986), make different assumptions. It is a reasonable assumption to make, but in the field of origins where we remain largely in the dark about exactly what happened, it is important to be aware of the starting points we adopt to make traction on the problem.

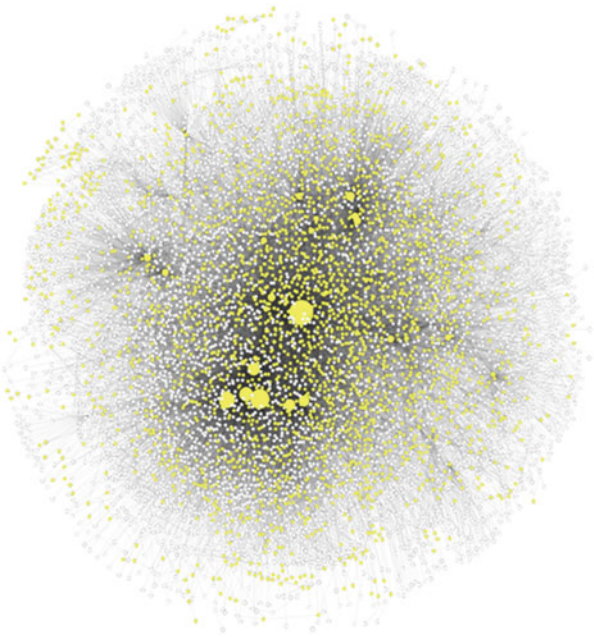

Fig. 10.1 A network representation of global biochemistry, containing thousands of compounds cataloged across organism on Earth. Highlighted in yellow are compounds (represented by nodes in the network) common to all three domains of life. Figure adopted from Kim et al. (2018)

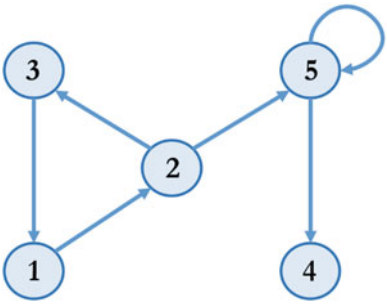

Fig. 10.2 A network with nodes (circles) and edges connecting them (arrows). Here, additional attributes and structure include labeling of nodes and arrows on the edges, respectively

place. Combining novel techniques from systems chemistry with the mathematical formalism of network science will enable prebiotic chemists to develop new concepts with testable consequences (Cronin and Walker 2016).

In this chapter, our goal is to introduce the concepts of network theory to prebiotic chemists as a mathematical formalism for making sense of prebiotic systems and as a tool to identify the processes driving the emergence of life. We first review the basics of network theory, discussing some of the successes in its application to chemical and biological systems. Our focus is on properties of biochemical networks of use to prebiotic chemists interested in studying the emergence of more "lifelike" chemical systems in the lab. We also discuss future directions for both how the study of biochemical networks might inform origin and how the study of chemical networks

relevant to the origin of life might also provide tighter constraints on what network properties could truly distinguish living from nonliving organization.

10.2 What Is a Network?

The phrase "World Wide Web" vividly captures the complex web of interactions connecting computers across the globe. Many other biological and technological systems share similar weblike structure, being comprised of many heterogeneous interconnected components (see, e.g., Fig. 10.1). Over the past several decades, it was realized a new statistical mechanics was necessary to describe such systems, which goes beyond the nineteenth-century statistical mechanics of idealized non-interacting particles to include the topology of interactions among system components and their resultant dynamics (Albert and Barabási 2002) The natural mathematical framework for developing such a theory is network theory, which projects the complex web of interactions in real systems onto an abstract representation as a graphical object (Barabási 2016). Networks are important mathematical descriptors in cases where the structure of interactions matters more than counting individual component parts—precisely what we envision happens as nonliving matter transitions to life. Due to its utility in concisely describing complex, interacting systems, network theory has been applied to an increasing number of systems in fields ranging from biology (Barabási and Oltvai 2004), to engineering, to the social sciences (Wasserman and Faust 1994).

Mathematically, networks are studied using the tools of graph theory, where entities are represented by *nodes* (also called vertices) and their interactions by *edges* (also called links), as in the simple network shown in Fig. 10.2 where nodes are depicted as circles and edges as arrows. Familiar examples include social networks, such as Facebook, where the nearly two billion individuals on Facebook could be mathematically represented by nodes and their friendships by edges (in practice it is difficult computationally to construct and analyze networks this large, but many networks of interest are smaller than Facebook or subnetworks can be studied). In a graph-theoretic representation of Facebook, an individual would be connected to every individual they "like," and network dynamics might include studying how the structure of interactions changes as individuals "like" and "unlike" one another or as new individuals are added to the network and others lost from it. Likewise, chemical species reacting with one another form networks within cells where nodes represent molecular species or reactions and edges represent connections of molecular species to reactions they participate in (see Fig. 10.3).

There are in fact many different ways to represent a network. For example, the Facebook network mentioned above could be represented as a *directed* network, meaning that the connections between nodes are not symmetric but instead reflect the directionality of "like" relations. Cole "liking" Sara's Facebook page does not imply Sara also "likes" Cole's page. The network in Fig. 10.2 is an example of a directed network, where edges are represented by arrows delineating the

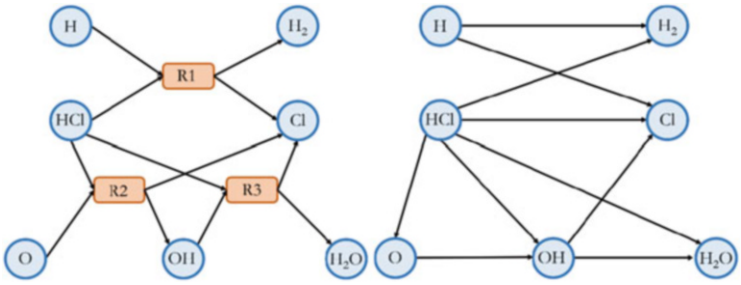

Fig. 10.3 Two different graphical representations of the same chemical reaction network. On the left is a bipartite substrate-reaction network, and on the right is shown the same system represented as a unipartite substrate-substrate network (see text for descriptions)

directionality of the relationship (in the Facebook example, a "like" relationship might point from the node labeled "Cole" to the node labeled "Sara"; if Sara also "likes" Cole's page, there would also be an arrow connecting the same two nodes but pointing in the opposite direction). Networks may also be *undirected*, where edges do not encode the directionality of the relationship. The choice of network is often motivated by the problem of interest and the measures one is interested in calculating (e.g., there is a richer literature of network measures for undirected networks as they are simpler, but the trade-off is they do not capture as much information as directed ones).

There are many different ways to graphically represent chemical systems, each permitting quantitative analysis of different aspects of global organization and in turn identification of the role of specific molecules in the robustness and function of living systems. Two examples of the most commonly implemented graphical representations for chemical systems are shown in Fig. 10.3, which both represent the following sequence of reactions:

$$\begin{aligned} \mathrm{H} + \mathrm{HCl} &\rightarrow \mathrm{H_2} + \mathrm{Cl} \\ \mathrm{HCl} + \mathrm{O} &\rightarrow \mathrm{Cl} + \mathrm{OH} \\ \mathrm{HCl} + \mathrm{OH} &\rightarrow \mathrm{Cl} + \mathrm{H_2O} \end{aligned}$$

The left panel of Fig. 10.3 shows a *bipartite* network, called a reaction-substrate graph, where substrates (reactants and products) and their reactions are both nodes and edges connect substrates to their relevant reactions. Bipartite networks are so-called because there exist two distinct types of nodes in the network: here, molecular species represent one type of node (circles), while chemical reactions represent the other (squares). The representation of the same network in the right panel of Fig. 10.3 is an example of a *unipartite*, substrate-substrate graph, where reactions are abstracted away and reactants are directly connected to products by an edge (if they are from the same reaction). A further refinement in the unipartite graph is representing edges as undirected (loosing directionality in the relationship of substrates and products as discussed above). In the substrate-substrate

representation, an edge between a pair of nodes can be thought of as a group of processes converting some molecular species to others. There are many other types of network descriptors including *weighted* networks (Newman 2004), where edges have a strength of weight associated with them, and *multilayer* networks, which contain different types of edges representing different connections (Boccaletti et al. 2014). Selection of which graphical representation to use depends on the question of interest and the relevant quantities to be measured. A review of many of the different representations of biochemical networks and their utility and shortcomings as applied to different scientific questions is discussed in Montañez et al. (2010).

10.2.1 Measuring Statistical Properties of Networks

In a seminal paper published in 2000, Jeong et al. reported metabolic networks of 43 distinct organisms—representing all three domains of life—are *scale-free* (Jeong et al. 2000), meaning their degree distributions roughly follow a power-law $P(k) \sim k^{-\alpha}$. An example of a scale-free network is shown in Fig. 10.4a and the corresponding power-law degree distribution in Fig. 10.4c. Here, $P(k)$ is the probability a given molecular species participates in k reactions. In graph theory k is called the *degree* of a node, corresponding to the number of edges connected to it. The *degree distribution,* or degree sequence, is the probability distribution of node degree taken over an entire network.[2] In the simple example network of Fig. 10.1, the degrees are 2, 2, 3, 1, and 3 for nodes 1, 2, 3, 4, and 5, respectively, yielding a degree distribution of $P(k) = 1/5$, 2/5, and 2/5 for $k = 1$, 2, and 3, respectively. This distribution has a *mean degree* $\langle k \rangle = 2.2$, and there are no outlier nodes with a significantly higher degree than the others. In this respect, the network is fairly homogeneous (the network in Fig. 10.2 is of course is too small to make statistically meaningful statements, but it serves for illustrative purposes). For a longtime it was thought most networks were homogeneous, but in the late 1990s and early 2000s, it was discovered most real-world biological and technological networks are in fact very *heterogeneous*, with *heavy-tailed* degree distributions consistent with power-law or lognormal fits [see, e.g., Barabasi (2009) for perspective]. In many real-world networks, most nodes have very few connections, but a few nodes called *hubs* have many connections and link less connected nodes together. The discovery of the power-law scaling in metabolic networks by Jeong et al. was part of this watershed moment in our understanding of the organization of biological and technological systems, but the significance of this property and its evolutionary origins still remain poorly understood.

In our Facebook example, a very small minority of Facebook's >2 billion users are hubs, such as Mark Zuckerburg with 98,885,179 "likes" (as of writing).

[2]The degree distribution is calculated by determining the frequency of the degree for each node, and is often normalized by dividing by the total number of edges in the graph, which can be interpreted as a probability of connection and the resulting distribution interpreted as probability distribution.

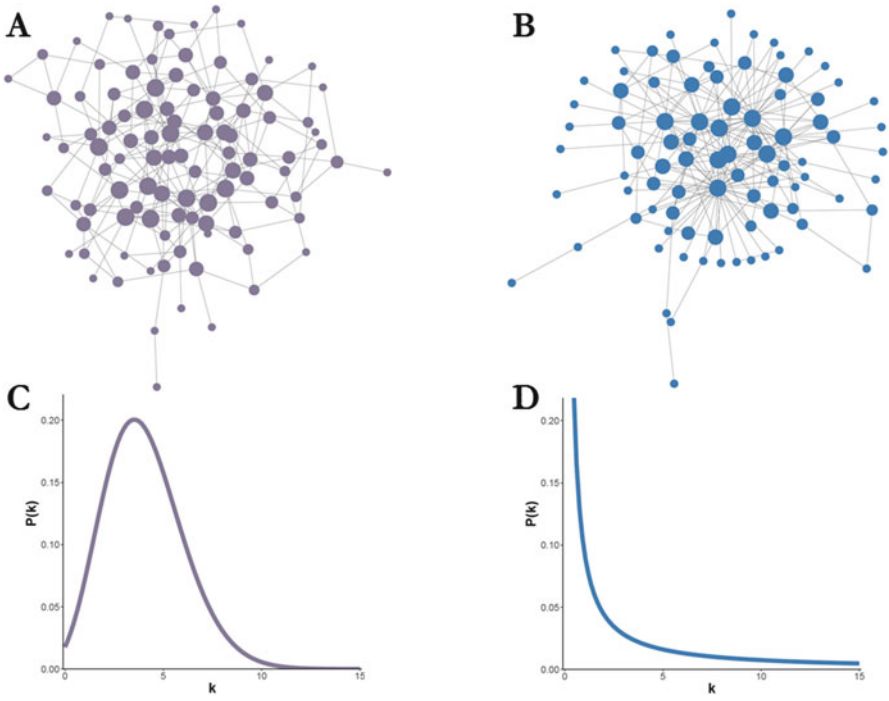

Fig. 10.4 Homogenous (left) and heterogeneous (right) networks. Shown are (**a**) a Erdös-Rényi (ER) random graph and (**b**) a scale-free network. Node sides correspond to degree in both images. Corresponding degree distributions are shown in (**c**) for the Poisson degree distribution characteristic of ER random graphs and (**d**) the power-law degree distribution indicative of scale-free network structure

Individuals with only a handful of connections are much more common but are also much less connected (e.g., by comparison the authors each have only a few hundred "likes"). As with social networks, metabolic networks also contain hubs, which include highly utilized molecules in biochemistry such as H_2O and ATP (Andreas Wagner 1998). These molecules participate in hundreds of reactions, with a comparably high node degree, whereas the mean degree of metabolic networks globally is in the range of just 2–5 connections (Jeong et al. 2000; Kim et al. 2018). Projecting metabolism onto a substrate-substrate network representation yields fits for the degree sequence that in general follow a power-law fit, indicative of scale-free structure. However, rigorously confirming a power-law fit for a given degree distribution is a challenging technical problem and an active area of research in the statistical inference community. In recent years, new tools have been developed for reliably determining cases of scaling consistent with a true power-law behavior, as opposed to other heavy-tailed degree distributions, such as lognormal (see, e.g., Clauset et al. 2009). Our recent analysis applying these tools to a dataset of >28,000 biochemical networks extracted from genomic and metagenomic data revealed a majority of biochemical networks can plausibly be fit to true power-law scaling, but

not all (Kim et al. 2018). A further complication is scale-free structure can depend on the network projection, leading to complications in interpreting results of fits to degree distributions without reference to other properties of the network or randomized controls.

Nonetheless, important structural differences between networks can often be seen directly from the degree distribution and other topological measures: in many cases these are indicative of properties that seem to be distinctive to living networks. For example, random networks, such as Erdös-Rényi (ER) networks, are characterized by degree distributions which are Poisson distributed, meaning that most nodes share roughly the same number of edges and the probability is exponentially suppressed for the highest degree nodes (the $P(k) \sim e^{-k}$ for $k \gg 1$) (Erdös and Rényi 1959). Because most nodes share similar degree, random networks are described as *homogenous* in their distribution of edges among nodes (like our small network in Fig. 10.1). An example of a homogenous network is shown in Fig. 10.4b and the corresponding Poisson degree distribution in Fig. 10.4d. The network structure and degree distribution are visually very different for homogeneous networks when compared to heavy-tailed or "scale-free" networks. Due to these structural differences, the systematic observations of heterogeneous networks in living systems provides a window into their large-scale organizational properties that distinguishes living networks from generic random ones. One key structural difference from an evolutionary standpoint is robustness to random mutation: random loss of nodes in heterogeneous networks will most often not affect overall topology so long hubs remain intact, whereas for homogenous networks, there are no hubs to maintain overall network connectedness. As such, heterogeneous networks are in general more robust to random failure or mutation, perhaps motivating their preferential selection in living organization.

In order to characterize a network, a number of different statistics about the network can be measured beyond degree distribution alone (Barabási 2016). For example, the mean degree of the network, as mentioned above, can be calculated as the mean value of the degree distribution and provides information about how connected each node in the network is on average. For directed networks this can be broken down into two contributing terms: the mean in-degree (number of edges pointing into a node) and mean out-degree (number of edges emanating from a node), which represent sinks and sources in chemical transformation space. Another important network statistic, the *average shortest path length,* measures the average number of steps it would take to get from one node to any other node in the network by taking steps along its edges. That is, it quantifies the minimal number of chemical transformations it takes, on average, to convert one molecular species to another.

Networks with a low average shortest path length are sometimes said to have a *small-world* property because it is relatively easy to get from one node to any other: one need only traverse a few steps. Readers may be familiar with the term "six degrees of separation" to describe this small-world property in human social networks. For a while it was thought metabolism too had the small-world property, meaning it only should take a handful of chemical reactions to transform any molecule to any other in a biochemical network (Wagner and Fell 2001). However,

subsequently, it was discovered metabolism is not small in a study performing detailed analysis of the network structure of *E. coli* (Arita 2004). In the context of prebiotic chemistry, we need an accurate picture of the structure of biochemical networks in order to identify how they can be generated in the absence of life—only then will we be able to map this to the appropriate chemical and physical properties driving prebiotic network evolution (not a small task!).

There are other network measures too that could aid in this effort. Another important statistic is *betweenness centrality*, which measures how often a particular node is on the shortest path between all other nodes in the network. Nodes with high betweenness centrality can sometimes be low degree but nonetheless essential to dynamics and function since they play a key structural role by connecting many otherwise disconnected or distant nodes. In many cases however, high betweenness centrality is correlated with high degree (hubs). For example, in a network representation of social media interactions, one might expect Barack Obama to have high betweenness centrality as he is a highly connected node (hub) through which many other individuals are connected. In biochemical networks, molecules like H_2O and ATP tend to have both very high degrees and high centrality, due to their fundamental roles in aqueous organic chemistry and metabolism, respectively. Hubs with high betweenness are therefore often among the most vulnerable nodes in a network for directed attack (rather than random loss): targeting removal of such nodes can lead to a network breaking apart into smaller isolated graphs. This is a chief vulnerability of the Internet (Cohen et al. 2001) and is often more technically discussed in terms of breaking apart the *largest connected component* of a network. A *connected component* is a subgraph of a network (e.g., a subset of nodes) where there exists a path between any two nodes in the subgraph. For understandable reasons, metabolic networks are dominated by a single large connected component [see, e.g., supplement of Kim et al. (2018) for size of largest connected component in biochemical networks]. The idea of connected components becomes important in discussions of graph-theoretic models of the origins of life, such as autocatalytic sets, which also form as connected components (discussed in Sect. 10.3). The early emergence of molecules with high betweenness centrality may have therefore been critical to rapid formation of connected components in prebiotic evolution. Identification of molecules fulfilling this role prebiotically could therefore provide insights the emergence of many key structural properties of living networks.

Many of the measures discussed so far track statistical properties of individual nodes, or paths between two nodes, but there are also many measures for higher-order properties of networks. For example, *clustering coefficient* tracks how many tightly knit communities exist within a given network, typically measured by counting the number of complete triangles connecting three nodes. Networks with high clustering coefficients have many clusters of nodes with above-average connections between them (relative the rest of the network). Complete triangles represent one example of a *network motif*. Network motifs are subgraphs which have specific connection patterns and which are overrepresented in biological systems with respect to randomized networks (Milo 2002). They were first uncovered in networks as diverse as those from biochemistry, ecology, neurobiology, and

engineering and have been proposed as a means to uncover the building blocks of functional networks (Alon 2003). From this perspective, they are an important concept for prebiotic evolution—identifying the network motifs which readily form under abiotic conditions and could combine to form more complex, lifelike systems would advance our understanding of key structural properties needed for assembly of living networks. As just one example, we recently constructed all possible three-member networks of cooperating RNA using reaction rate data from a real RNA system based on the *Azoarcus* ribozyme (Mathis et al. 2017b). The goal was to determine the types of cooperation possible when building prebiotic networks from their component parts. Here cooperation was defined in terms of the structure of the subgraph (see Mathis et al. 2017b). Our results demonstrate the triplet network interactions among genotypes (nodes) in the real *Azoarcus* ribozyme system were intrinsically biased to favor cooperation due to the particular distribution of catalytic rate constants in the real system, as compared to other possible distributions for the rate constants. This example demonstrates how coupling properties of chemistry with network structure can provide new insights into the emergent properties of prebiotic networks, such as whether we should expect them to be cooperative.

10.2.2 Generative and Evolutionary Models

Knowledge of topological properties, such as the small-world property or scale-free structure, can provide insights into how networks with those properties can arise in the first place. To get at the interesting properties, we must first identify what features are expected to arise randomly. There are many different models to generate random networks for comparison. These models are known as *random graphs*. We introduced above the first class of random graphs to be formalized, the ER graph, which was developed by Erdös and Rényi in 1959 (called Erdös-Rényi or ER random graphs) (Erdös and Rényi 1959). ER random graphs are defined by the number of nodes (n) and number of edges (v) they contain. A single instance of an ER random graph is generated by starting with n unconnected nodes and randomly assigning edges between nodes with *equal and independent* probability p, until v edges exist. For a very long time, it was assumed that ER random graphs represented ideal null models for network organization. However, as more empirical examples of networks were accrued through the 1970s and early 1980s, it became apparent that ER random graphs failed to produce statistical features common in real-world networks, such as high clustering coefficients, small-world topology, and heterogeneous degree distributions discussed above.

The degree distribution of ER random networks is always homogeneous, described by a Poisson distribution. These provided a stark contrast with real-world networks which are observed to have small-world properties and heterogeneous degree distributions. To address this, in 1998 Watts and Strogatz published a random graph model combining some properties of ER random graphs with regular

graphs (which are similar to lattice structures) in order to generate networks with high clustering coefficients and small-world properties, much like real-world systems (Watts and Strogatz 1998). However, these failed to produce the heterogeneous, heavy-tailed degree distributions characteristic of many real-world systems. In 1999 Barabási and Albert introduced a model using preferential attachment to generate networks with the desired scaling properties (Barabasi and Albert 1999). Preferential attachment models start with a small network of nodes and add nodes one at a time to the network by preferentially attaching new nodes to nodes with high degree. In the context of life's chemical networks, such a growth model would imply metabolic networks grow by adding new metabolites, with the most highly connected nodes also being the most likely candidates for being the oldest. Among these ancient, highly connected nodes in metabolism are intermediates of glycolysis and the tricarboxylic acid cycle (TCA); consistent with the hypothesis of Morowitz and later Smith and Morowitz, the evolution of biochemistry is recapitulated in intermediary metabolism (with TCA being among the most ancient components) (Smith and Morowitz 2016). However, modifications to the Barabási-Albert preferential attachment model are necessary to explain the network evolution of biochemistry: the model always produces exactly scale-free networks, whereas observed biochemical networks have heavy tails but are not precisely scale-free (see Kim et al. 2018; Clauset et al. 2009). A number of models have been developed to address this gap (e.g., Bianconi and Barabási 2000). Identifying prebiotically relevant random graph models will be an important step toward understanding the transition from nonliving to living matter.

10.3 Prebiotic Chemical Networks: Prospects and Promise

An important feature of the Erdös-Rényi (ER) model described in the previous section is the existence of a phase transition as the probability of two nodes being connected by an edge increases. At a critical connection probability, pc, corresponding to a critical mean degree, ER graphs transition from having many disconnected components to being dominated by one large connected component. Although this transition occurs within an abstract mathematical object, it has implications for the origins of life. Kauffman was the first to recognize this link (Kauffman 1993). In a stroke of insight, he realized a similar process should exist in chemistry: if enough reactions are possible in a given chemical system, one should end up with a large connected set of reactions. Chemical reaction networks should therefore exhibit a phase transition much like the ER transition, where increasing the number of possible reactions among a set of molecules induces a transition from many disconnected networks to a large connected one. To model this process, he considered abstract proteins represented as binary sequences of "0"s and "1"s, often referred to as *binary polymer models* in the artificial chemistry literature. An example autocatalytic network of binary polymers is represented in Fig. 10.5. Kauffman showed that if there is a small, independent, and identical probability

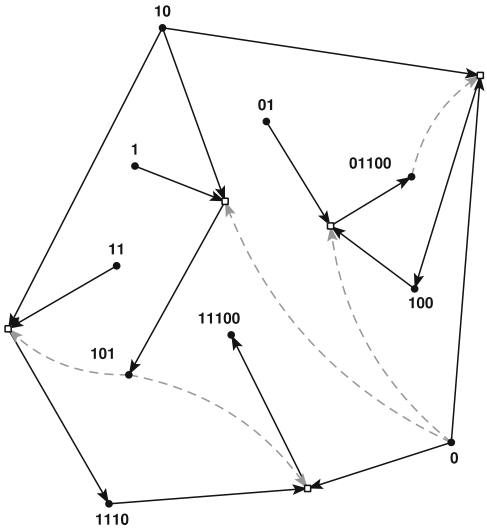

Fig. 10.5 Example of an autocatalytic network within the binary polymer model, where polymers consisted of two monomer species "0" and "1." In this multilayer graph, black nodes represent molecules, white nodes represent reactions, black arrows connect molecules to reactions they participate in, and gray-dashed arrows point from catalysts to reactions they catalyze. Figure adopted from Hordijk et al. (2012)

p that any protein up to a length L catalyzes any given reaction, then the probability P of finding a connected set, such as the one in Fig. 10.5, increases as the length L of the longest sequences increases. In fact, in the mathematical model, the probability approaches 100% as the max length L of the proteins grows, even if the probability of any given protein being catalytic active, quantified by the p, is made arbitrarily small. Placed within the broader context of network theory discussed in Sect. 10.2, Kauffman was not only looking for connected components but a specific motif. His interest was in collectively reproducing sets of molecules: these are network motifs composed of closed cycles of reactions. Analyzing chemical reaction graphs for the existence of these motifs forms the foundation of autocatalytic set theory, the first systematic application of network science to prebiotic chemistry.

Other elements of network science have been suggested in prebiotic chemistry over the years, although they are often not studied with the formalism of graph theory. As one example, in 1978, Eigen and Shuster introduced the hypercycle as a proposed solution to the error threshold problem in prebiotic evolution (Eigen and Schuster 1978). The error threshold sets a fundamental bound on the minimal amount of information necessary to transmit between successive generations for heredity to be possible (Eigen 1971). For a prebiotic replicator, such as a self-copying RNA, this bound is approximately a mutation rate $\mu = 1/L$,where μ is the mutation rate per monomer and L is the length of the sequence (in reality the critical mutation rate depends on the shape of the fitness landscape). In prebiotic systems, before the evolution of error-correction mechanisms of modern cells, error rates were high. The intrinsically high error rates of nonenzymatic templated replication place a strict limit on the amount of information an individual sequence can faithfully copy before error-correcting enzymes evolved. In the words of Szathmary and Maynard Smith, this encompasses a catch-22 for prebiotic evolution: "no enzymes before genes, no genes before enzymes" (Smith and Szathmáry 1995). The hypercycle, as

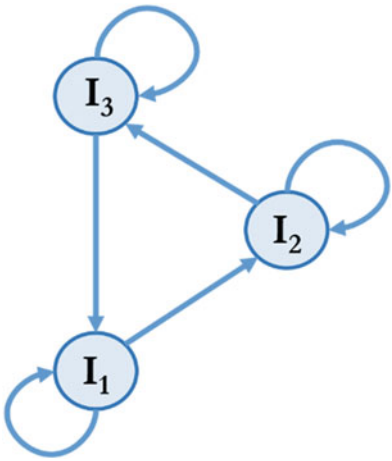

Fig. 10.6 A graphical representation of a three-node hypercycle composed of three replicators I_1, I_2, and I_3. Self-loops indicate self-replication and directed arrows between nodes indicate direction of catalysis

first imagined by Eigen, was proposed as a resolution to this paradox. His idea was to couple two or more replicating species, where each was capable of promoting the replicative efficiency of the other forming a cyclic graph with self-loops; see Fig. 10.6. This constitutes a simple network, where every chemical species can be represented as a node and catalytic connections by edges. Thinking graphically, a solution to the error threshold problem is to distribute information over a network of interacting molecules, rather than storing all of it within a single molecule. For a number of decades, Eigen's idea remained hypothetical, but recently hypercycle networks have been demonstrated in real chemical systems of interacting RNA molecules, providing an important empirical window into understanding potential early stages of evolution and cooperation in molecular systems (Vaidya et al. 2012).

10.3.1 Autocatalytic Set Theory

Chemically, autocatalytic sets are collections of molecules, where every molecule in the set is produced by a reaction catalyzed by another molecule in the set. Graphically, autocatalytic sets represent a class of subgraphs, or network motifs, with directed paths forming closed cycles. It is in this respect the hypercycle can be considered an example of an autocatalytic set. Since the early work of Kauffman, Eigen, and others exploring how closed cycles might lead to collectively reproducing systems, there have been a number of efforts to both develop better theoretical and experimental approaches to understanding these systems. Of note, Kauffman's original idea has been formalized within the context of RAF theory. RAF is short for *reflexively autocatalytic and food-generated.* RAF sets are graphical structures forming closed cycles with inputs for food to the cycle (Hordijk and Steel 2004; Hordijk et al. 2012). Within this formalism a variety of properties of

autocatalytic sets have been proven over the past few years, strengthening the potential prebiotic relevance of the theory. In particular, a major criticism of Kauffman's original model was the required level of catalysis, which was deemed too high to be realistic (Lifson 1997). Within the RAF formalism, Hordjik et al. have proven autocatalytic sets are guaranteed for realistic levels of catalysis and when more complicated constraints of real-world chemistries are imposed (such as base pairing) (Hordijk et al. 2011).

Computational and analytical results show autocatalytic sets are common in chemical reaction networks with random independent and identically distributed catalysts (Hordijk and Steel 2017) and also occur in more realistic scenarios with heterogeneously distributed rates of catalysis (Hordijk et al. 2014). However, while autocatalysis is a common network motif, multiple studies have shown that relatively few networks are capable of fixating dynamically when kinetics are taken into account. Wynveen et al. simulated the dynamics of a binary polymer system constructed using the same algorithm as Kauffman. They demonstrated that relatively few networks were able to depart from an expected maximum entropy state, meaning that networks composed of random and identically distributed catalysis rarely display "lifelike" dynamics (Wynveen et al. 2014). Similarly, Filisetti et al. found only a small fraction of autocatalytic sets which were able to increase the abundance of their constituents above a background level expected nonenzymatically (Filisetti et al. 2012).

As theory improves and makes closer contact with experiment, the challenge ahead will be to understand better the properties of real biochemical networks and how those arise prebiotically. It has already been confirmed RAFs exist in real biochemical networks, such as the metabolic network of *E. coli* (Sousa et al. 2015). Additionally, some work has been done to connect RAF theory to the structure of real biochemical networks. When Kauffman originally devised autocatalytic set theory, it was thought most real-world networks were homogenous, like the ER model. However, as discussed above, it was subsequently discovered heterogeneous networks are more common in real-world systems. Recent work has also shown RAFs are common in catalytic networks with power-law distributed catalysis (Hordijk et al. 2014), which more closely resembles the distribution of catalysis in metabolic networks. Future work should further the connections between these abstract models and the properties of real biochemical networks.

10.3.1.1 Evolvability of Autocatalytic Sets

A key transition in the origin of life on Earth was the emergence of Darwinian evolution via natural selection (Nowak and Ohtsuki 2008). Natural selection requires mechanisms to generate variation among individuals, which can then be selected. For single molecule replicators or single cells, these requirements are easy to satisfy as the "unit" of evolutionary selection is readily identifiable (a replicating sequence or cell, respectively). However, for collectively reproducing systems without a well-defined boundary of "self" and "other," the concepts of individuality and heredity

are poorly defined. It is not yet even clear well-defined units for selection exist in such systems. As such, the evolvability of catalytic networks has been a subject of intense debate in origin of life research. At stake is whether catalytic networks are indeed a viable alternative to genetic polymers as the first hereditary system capable of Darwinian evolution.

Among models proposed for catalytic network evolution is the "lipid world" scenario proposed by Segré, Lancet, and collaborators (Segré et al. 2001). The model system includes simulated random networks with lognormal distributed catalytic efficiencies, meant to capture aspects of the asymmetry of catalytic efficiency in real systems. The lognormal distribution of catalysis can be modeled by a strongly connected, weighted network. From this model, Segré et al. have shown in some situations these networks are capable of evolution by natural selection (Segre et al. 2000). However, using the same model, Vasas et al. have shown that, in general, these networks cannot be evolved to generate arbitrary steady states (Vasas et al. 2010). The problem arises because random networks generated using the lognormal catalytic distributions contain subtle motifs which prevent the maintenance of variation between competing networks. This led Vasas et al. to claim that autocatalytic networks, in general, are not evolvable. In subsequent work, Vasas et al. investigated the evolvability of autocatalytic sets similar to those first suggested by Kauffman (Vasas et al. 2012). These more recent results suggest autocatalytic sets can indeed evolve in a limited sense, as long as they contain multiple *viable cores.* Viable cores are a specific network motif composed of completely connected catalytic subgraphs. Taken together, these results indicate it may be possible for catalytic networks to evolve in the absence of genes, but the details are sensitive to network topology in ways genetic propagation of information isn't (perhaps one selective factor in the transition to genetic heredity during early evolution).

The jury is still out on whether general, evolvable models of catalytic networks are possible and what the key properties of such networks might be. In an attempt to summarize the current state of the field, and to project what network properties might emerge as those most essential to defining evolvability, Nghe et al. recently identified six key network parameters to focus research efforts (Nghe et al. 2015). Among these were the concepts of viable cores. Other parameters include familiar concepts in prebiotic chemistry, such as resource availability, and compartmentalization. Resource availability is essential to maintaining collective reproduction (e.g., the "food" in RAF sets), and compartmentalization is essential for forming selectable units (this could occur via localization on a surface and need not be physical compartmentalization). Other parameters may be less familiar and are more intrinsic to network organization, including its connectivity, controllability, and scalability. Connectivity, combined with the availability of resources determines how effectively molecular species outside of viable cores can be produced. Very sparse, poorly connected networks will have limited evolvability, as there are no paths for transitions between graphs. In the other extreme, networks that are too densely connected will generate non-specific tars. Controllability can be implemented in chemical system through dynamic feedback. These feedbacks stabilize network dynamics

against random perturbations, enhancing the robustness of chemical networks in fluctuating environments and can play a critical role in inheritance. As one example, Kaneko and collaborators have worked out a model catalytic network, where reproduction is controlled by a "minority" population of molecules regulating the behavior of the rest of the network. Their proposal is that these minority molecules played the role of primitive genes (Kaneko and Yomo 2002). The scalability of chemical networks refers to an ability to grow in size while maintaining functional modules. In order to scale efficiently, prebiotic chemical networks must be sparse, meaning most nodes have few connections, reducing the likelihood new functional modules will interfere with the rest of the network. This introduces a tension between scalability and evolvability as sparsity favors one but not the other. In order to understand the evolution of primitive chemical networks, future studies must constrain these six key parameters in real networks, and theory must be developed to better understand how each impacts evolutionary outcomes.

10.3.2 Autocatalytic Sets in the Lab

Identifying and exploring dynamic, complex chemical networks represents a major analytical challenge for organic chemists. The most interesting aspects of complex (bio)chemical networks are due to interactions between tens to thousands of dynamically coupled reactions, meaning any given network cannot be understood as the sum of many isolated reactions. Many of the standard techniques for characterizing reaction products depend on identifying single molecule products which can be compared to lab standards. Understanding the properties of chemical networks depends on first understanding how the topology of the networks is related to their dynamics. Luckily, there are many questions in this area that can be addressed using different chemical models, and the earliest investigations of chemical networks have come from polypeptide systems as well as RNA systems.

In 2003 Ashkenasy et al. predicted and constructed a complex network of peptide fragments (Ashkenasy et al. 2004). The authors had previously demonstrated a reaction between electrophilic (E) and nucleophilic (N) peptide fragments could be promoted by a template peptide (T). The peptides form a quaternary complex, and in isolated reactions they had shown that the efficiency of the reactions could be predicted by the stability of the complex, which could be in turn estimated from the template structure. Using nine different templates, Ashkenasy et al. constructed a weighted network with the templates as the nodes and while the edges represented the predicted catalytic pathways, a schematic representation of this network is shown in Fig. 10.7. When they implemented this network in the lab, they found that some of the predicted edges were not realized due to competition for shared substrates. This network represents one of the earliest physical instantiations of Kauffman's autocatalytic set theory. This work demonstrates real chemical networks are not simply the sum of all possible reaction pathways but include emergent properties arising due to the complex interplay between topology (catalytic efficiency) and dynamics

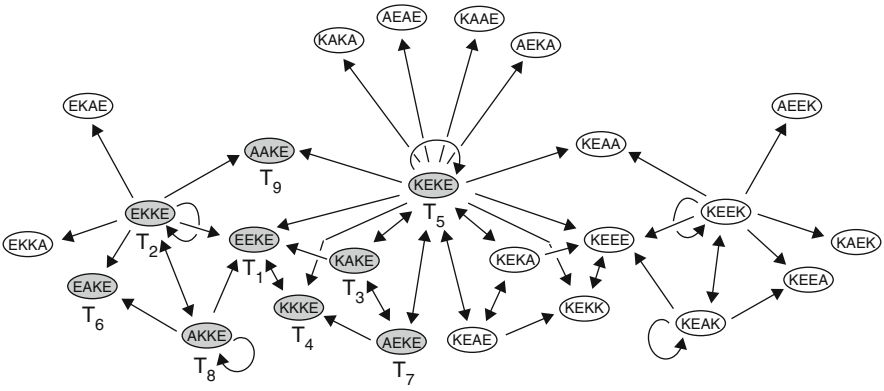

Fig. 10.7 Illustration of a self-organized peptide network composed of 25 nodes joined by 53 edges. Nodes are different peptide templates, while edges represent catalytic activity. Adapted from Ashkenasy et al. (2004)

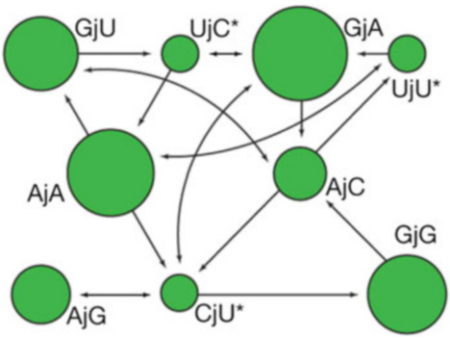

Fig. 10.8 Network structure of cooperative RNA hypercycle. Node labels correspond to different genotypes, while edges represent catalytic activity. Node sizes correspond to steady-state abundances. Adapted from Vaidya et al. (2012)

(resource availability) [see, e.g., also Vaidya et al. (2013) for a combined theory-experiment model of the role of limited resources in RNA systems] (Fig. 10.7).

In 2012 Vaidya et al. demonstrated hypercycle networks could form spontaneously in RNA networks, using the *Azoarcus* ribozyme system, one of which is shown in Fig. 10.8 (Vaidya et al. 2012). The *Azoarcus* ribozyme is a ~200 nt RNA sequence capable of self-assembly. By varying four different bases in the ribozyme, 48 different genotypes can be made. Each genotype can catalyze its own assembly as well as the assembly of other genotypes with different efficiencies. Lehman et al. demonstrated strongly cooperative triplet motifs could form, within the larger 48 node network. Subsequent studies have shown the dynamics of those motifs can be described using the tools of evolutionary game theory, suggesting the networks are evolvable. Our work on the degree of cooperation within triplet motifs

demonstrated that the catalytic rates observed in the lab, which are derived from the energetics of RNA base pairing, promoted the frequency these cooperative triplets relative to selfish alternatives (Mathis et al. 2017b). A key result of the 2012 study was the observation cooperative interactions among related RNA replicators can lead to the spontaneous formation of ordered dynamics (Vaidya et al. 2012). This highlighted the key role of network interactions in understanding the spontaneous organization of lifelike entities (Fig. 10.8).

10.3.3 Network Expansion

Autocatalytic sets were envisioned as self-generating networks, which collectively can act as selectable evolutionary units. A different approach to the application of network-theoretic ideas to understanding the early evolution of biochemistry is to instead consider the properties of ecosystem or biosphere-level models (without specific knowledge of individual evolutionary units). One motivation for this approach is the observation extant ecosystems display greater regularity in terms of their stability and function than individuals do (Dinsdale et al. 2008). If we are to uncover general, and perhaps even universal, principles of biological organization through a network-based approach, it is therefore at the level of ecosystems (or even the biosphere as a whole) where we may have the greatest success (Smith and Morowitz 2016). Under this view, compartment-free, ecosystem-level models could provide the most promising insights into the processes governing the emergence and evolution of life's biochemical networks. It is worth noting we also do not know the level of complexity where "individuals" first emerged in prebiotic evolution.

One approach, first developed by Handorf et al. (2005), implements network expansion algorithms to explore the temporal order of incorporation of metabolic pathways in global biochemistry (Handorf et al. 2005). Network expansion leverages the availability of databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata et al. 1999), which provide publicly accessible catalogs containing a large majority of all known biochemical reactions (the network in Fig. 10.1 was generated from all enzymatic reactions cataloged in KEGG). The algorithm proceeds by recursively determining the set of all possible molecules (the "scope"), which can be produced from an initial "seed set" of molecules. Each time the algorithm is iterated, the newly produced products of reactions are added to the graph, expanding the network. Prebiotically relevant seed sets might include simple molecular species proposed in different origin of life scenarios, such as H_2CO, CH_3SH, NH_3, and $P_2O_7^{4-}$, for example (see, e.g., Handorf et al. 2005). Starting from a given seed set, and iteratively expanding the network along all possible enzymatically catalyzed reactions, permits asking questions about the ordering and interdependency of the expansion of biochemical networks at a global scale. For example, Raymond and Segre utilized network expansion on a biosphere-level network representing global biochemistry to uncover a critical role for O_2 in

permitting the emergence of metabolic pathways associated with complex life (Raymond and Segre 2006). Network expansion has also been implemented to study the coevolution of enzymes and metabolic pathways, revealing enzymatic novelty emerging in punctuated clusters corresponding to enzyme classes (Schütte et al. 2010).

More recently, network expansion has been applied directly to a problem of relevance to the origin of life—could a primitive core metabolism exist in the absence of phosphate? The answer, according to Goldford et al., is "yes" (Goldford et al. 2017). Starting from a set of prebiotically plausible seed molecules, exclusive of phosphate, they identified a phosphate-independent core metabolism, which could in principle support synthesis of a broad array of bioessential compounds. This model lends support to the concept of a "thioester world," preceding the use of ATP as the major energy currency of life. Other origin of life theories could similarly be tested with network expansion algorithms to identify possible ancestral networks, which could then be leveraged to generate new hypotheses about different origin of life scenarios.

10.3.4 Graph Grammars and Generative Models

One challenge of the approach provided by network expansion is it is not predictive but can only retrodict the potential pathways by which metabolism could have expanded through evolutionary history. Ideally, we should be able to predict all possible chemical pathways and networks and then identify the possible paths traversed in transitioning from nonlife to life and in the subsequent evolution of life. With this knowledge in hand, it would be easier to ask questions about why life arose and what its characteristic properties are. To do this, in addition to knowing the biochemistry of life, we must have some knowledge of the chemical networks *not* selected by life. Predictive theory in chemistry is an area of intensive research, with much progress made but much further to go. One promising area is the development of graph grammars as applied to predicting transformations on chemical structure. In this approach, a molecule itself is mathematically represented as a graph, where atoms correspond to nodes (vertices) and bonds to edges in the graphical representation of a molecule. Reactions are then modeled as rewiring transformations that transform the graph into another graph (representative of a different molecule) (Andersen et al. 2017). As an example of the application of graph grammars to prebiotic chemistry, this formalism has been applied to HCN polymerization, demonstrating the combination of graph grammars with experimental data can lead to guide exploration of different chemical pathways and roots to open-ended evolution (Andersen et al. 2013).

As systems biologists, network theorists, physicists, biochemists, and others collaboratively illuminate the structure of biochemical networks and the generative mechanisms which produce them, prebiotic chemists are charged with the task of explaining the origins of that structure. In order to effectively explain the topological

properties of living networks, prebiotic network scientists will need to choose random network models to compare their networks against. The ubiquity of heterogeneous degree distributions suggest the Barabasi-Albert model might be a good place to start; however, the modularity described by Jeong et al. implies random hierarchical graphs might be better suited (Ravasz et al. 2002). Each random graph model contains within it implicit assumptions about the generative mechanisms involved in networks. For a given network and a set of questions about it, the appropriate random models will be different. For example, we recently compared the network structure of biochemical networks at the scale of individuals, ecosystems, and the biosphere as a whole (Kim et al. 2018) (see discussion below). For this work, the appropriate random graph for individual organismal biochemical networks involved constructing networks by randomly sampling biochemical reactions from the KEGG database, while the appropriate random graph model for ecosystems instead constructed networks by randomly sampling whole genome networks and merging them.

Prebiotic chemists, who straddle abiotic organic chemistry and biochemical networks, will need to develop appropriate random models to understand the transition from nonliving to living networks. Graph grammars provide a useful framework for understanding the generative mechanisms underlying organic chemistry, while thermodynamic calculations can generate networks of plausible geochemical reactions. An important next step in understanding the large-scale structure of biochemical reaction networks will involve deploying machine learning techniques to infer the generative mechanisms underlying those networks. Once those generative mechanisms are identified, comparing them to expected abiotic mechanisms will allow prebiotic chemists to separate the roles of chance and necessity in the evolution of biochemical networks.

10.4 Future Directions

So far, we have discussed general background in network theory and provided some applications where it has been successfully applied to modeling origin of life processes. Up to now, most research investigating the network structure of biochemical and chemical networks has focused purely on their graphical properties. To understand the physical and chemical principles underlying the origins of life, closer contact must be made with understanding—in terms of physics and chemistry—why particular network architectures are selected by life (Walker 2017). One approach is to identify universal structural properties of living networks. The “scale-free” property is one such candidate property. But, as we have discussed, fitting degree distributions is only one way to gain insights into the structure of a network and is challenging to interpret because of ambiguities in identifying the correct fit for a given distribution. Nonetheless, Jeong et al.’s work demonstrating a universal network structure for metabolism across all three domains of life does hint there exist organizational properties of biochemistry common to all life on Earth. Just as

astrobiologists discuss the "universal" nature of biochemical components—e.g., all known life is composed of DNA, RNA, proteins, etc.—in informing models for origins of life, we must also consider the "universal" nature of biochemical organization, e.g., the network structure of biochemical reactions.

10.4.1 Universal Properties of Biochemical Networks

One major hurdle for making claims about universality is the common ancestry of all life on Earth. When we talk about universal properties of life, we mostly mean universal properties of life on Earth, and not necessarily properties truly universal to life, characteristic of any life in our universe. Such principles, if they exist, would form the foundation of a new research field in universal biology (Sterelny 2015; Goldenfeld et al. 2017). Discovery of alien life would obviously enable us to identify such properties, if they exist. But, in the absence of discovering alien life, is there any way we might confidently make claims of universality? This is a question the origins of life field is primed to address. We must understand life and its universal properties in order to definitively say how such systems can arise in the first place. One advantage of network thinking is we need not think of life as a level-specific phenomenon in the same way we must if we focus on chemistry alone as the defining feature of life. We are accustomed to thinking of life as a chemical phenomenon, i.e., defined by the "right" chemical building blocks. But, shifting our thinking to organizational properties of networks, and their informational properties, allows studying recurring properties of life across different scales of organization. If common patterns are found in how living matter organizes across scales within the biosphere, it increases our confidence those patterns of derivative of universal laws, rather than shared common ancestry.

We recently analyzed the structure of biochemical networks across multiple levels of organization in the biosphere ranging from the chemical reaction networks within cells, to ecosystems, to the biosphere as a whole (Kim et al. 2018). Biochemical reaction networks were constructed using annotated genomic data from 21,637 bacteria taxa, 845 archaea taxa, 77 eukaryotic taxa, and 5587 metagenomes, using methods developed by Jeong et al. (2000). A biosphere-level network was constructed from all enzymatically catalyzed reactions cataloged in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, which is the network shown in Fig. 10.1. Analyzing the topological structure of these networks as a function of network size (number of compounds) and level or organization reveals universal structural properties across all biochemical networks on Earth. These are described by universal scaling laws; scaling behavior is shown in Fig. 10.9. Scaling laws are often cited as a candidate for universal biology as they unify trends across different biological organisms and scales of organization (Gisiger 2001; West 1999a). Familiar examples of scaling behavior from physics include critical phenomena near phase transitions, where physical properties such as heat capacity, correlation length, and susceptibility all follow power-law behavior. The scaling of

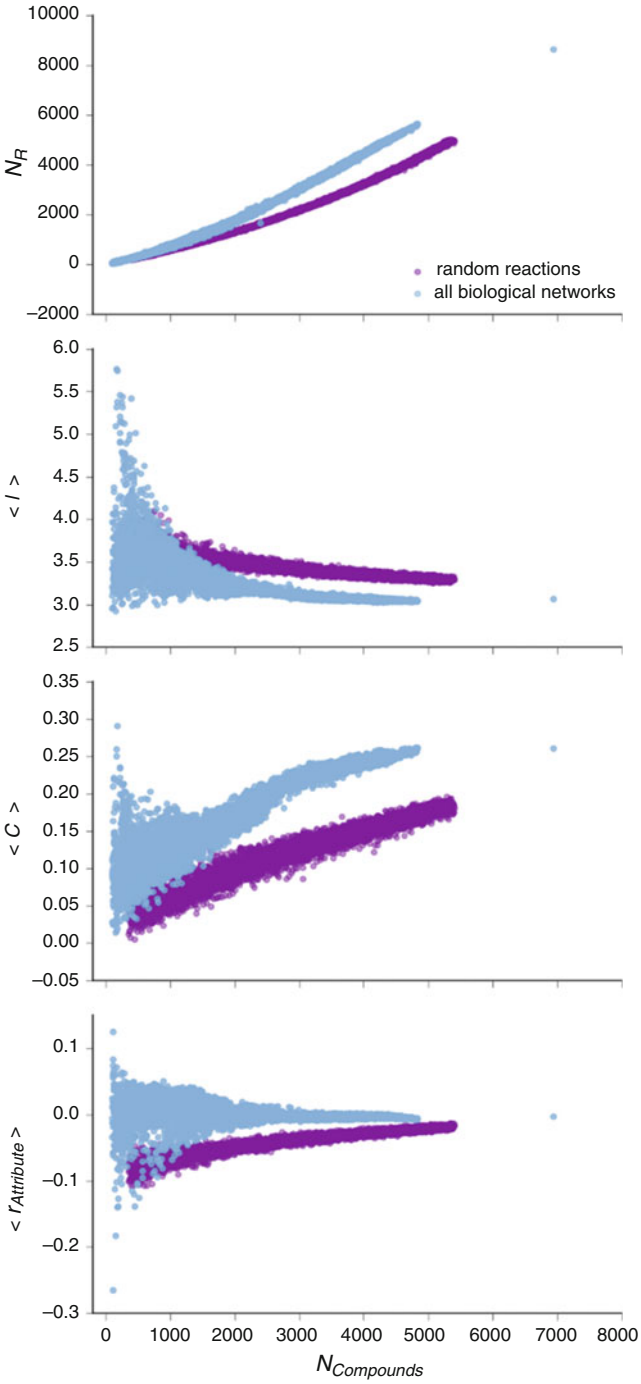

Fig. 10.9 Scaling of network attributes with network size. Biological networks (blue) scale differently from random collections of biochemical reactions (purple). Figure adapted from Kim et al. (2018)

network structure across levels of organization is different than the power-law relationship for degree distribution of scale-free networks described in the previous section because it applies *across* networks and not just *within* networks of individual organisms.

Randomly sampling reactions from known biochemistry to construct networks of similar size to organismal and ecosystem-level biochemical networks does not reproduce the scaling observed for living networks (Kim et al. 2018). This suggests it is the particular manner in which reactions are organized in living matter, and not the compounds or set of reactions alone, which yield the distinctive properties of living systems. Network growth models, such as preferential attachment, can reproduce some aspects of the architecture of life, but are not physically or chemically motivated and do not explain the constraints given rise to an observed network architecture. Scaling relations, due to their ability to "predict" the values of system parameters based on other measured quantities, represent one of the closest approaches so far to a predictive theoretical biology, akin to theoretical physics. Using the observation that cells and organisms are constrained in their growth by resource distribution networks, predictive models can be generated that accurately provide values for the scaling exponents observed in a number of diverse biological systems (West 1999b). Similar predictive models should be generated for biochemical network scaling (a work in progress) and would provide insights into universal constraints on biochemical architecture, which likely played a role in shaping the earliest networks in the transition from nonlife to life.

10.4.2 Information and Controllability

One of the most widely discussed, distinctive characteristics of life is its "informational" properties (Walker and Davies 2013; Yockey 2005; Kuppers 1990). The study of biology is replete with informational analogies, such as coding, signaling, sensing, interpretation, etc. As such, information theory is increasingly being utilized to characterize living systems across all scales (Davies and Walker 2016). An open question is how useful the concept of information is for origin of life research. A first step is to identify in what sense information could distinguish living networks from nonliving ones. The static network picture discussed throughout much of this chapter is not readily amenable to analysis from an informational perspective as many measures from information theory rely on knowledge of the dynamic properties of a system. One class of models for biological networks where information theory is readily applied is so-called random Boolean network (RBN) models (Wang et al. 2012), which are most commonly used to model gene regulatory networks where genes can be represented in one of two states "1" (activated) or "0" (inhibited). While this may seem an abstract representation, such models have been successful in systems biology and are widely applied. They may also have some utility in prebiotic evolution as peptide networks have been shown to capable to execute simple Boolean logic (Ashkenasy and Ghadiri 2004). In a recent study by one of us, it was shown Boolean models for biological gene regulatory networks (GRN) do in

fact display different patterns in how information is processed when compared to random networks with similar topological properties, e.g., random networks with the same degree distribution (Kim et al. 2015; Walker et al. 2016). This suggests at least some of life's biochemical networks might be optimized for information processing, and these optimization properties may go above and beyond the topological structure of the network alone for networks encoding function.

A relevant question for origins is why biology is optimized in this manner. In the case of the GRN models, the distinctive informational properties are associated with their controllability. The models under study were the GRNs describing the state of genes responsible for regulating the cell cycle in the fission yeast *Schizosaccharomyces pombe* and the budding yeast *Saccharomyces cerevisiae*. The Boolean models correctly reproduce the sequence of gene expression patterns observed in dividing fission and budding yeast, respectively. Regulating a small subset of nodes, called the control kernel, drives both networks toward their resting phenotype. That is, by regulating just a few nodes, one can control the function of the entire network. Intriguingly, these nodes also dominate the distinctive informational properties of these networks, suggesting a relationship between information processing and controllability in the biological networks, which is not generally present in random graphs.

In Nghe et al., information control was recognized as among the six key network parameters we must understand better in order to build an evolutionary theory of catalytic networks (Nghe et al. 2015). Most origins of life research so far has focused on the role of genes in information storage and propagation and not their role in information processing or as regulators of biological function in a dynamic system. However, control is essential to biological function, for example, in maintaining homeostasis. As with the toy model of the cell cycle networks, there likely is a connection between the function of early genes as control elements in early cells and their role in information processing and storage. Hints of this are apparent in models exploring these concepts. For example, Kaneko and colleagues discovered a key role for "minority molecules" in regulating reproduction of catalytic networks in a protocell model (Kamimura and Kaneko 2010). The minority molecules are kinetically slower components in an autocatalytic network and were suggested to play the role of primitive genes, regulating reproduction of the entire system. More models and more empirical work studying how networks might evolve control nodes are necessary to understand how very primitive biochemical networks first evolved regulatory feedback and may provide insights into the early evolution of genetic function.

10.4.3 A Network Theory of Planetary Biospheres

In the previous sections, we have talked about two distinct layers of biochemical networks—metabolic networks describing all of the catalyzed (programmed) chemical reactions transforming molecular compounds within cells and gene regulatory

networks which regulate cellular function (e.g., do the programming). While we often study these systems separately, in reality, they are tightly coupled. The biochemical network organization of the biosphere emerges due to the structure of reactions which are *enzymatically catalyzed*—that is, the subset of the Earth's chemistry life controls. That control is itself implemented through gene regulatory networks, which represent a "higher level" in life's hierarchical organization. As we go up in the hierarchy of structure and function in biological systems, we see similar motifs of interacting networks, where some biological networks regulate the function of others. The phenomena of life itself may be thought of as a hierarchy of interacting networks. One critical question for origin of life research is to uncover how such a hierarchy emerges in the first place.

In order to answer this question, we must consider the coupling of Earth's biological networks to their geochemical and atmospheric context. At some level, terrestrial biochemistry should be continuous with terrestrial geochemistry, implying geochemistry should represent the bottom level of life's hierarchy (Shock and Boyd 2015). In a similar vein, the biosphere's coupling to atmospheric chemistry has driven the most dramatic planetary scale changes in Earth's history (Sessions et al. 2009). If this strong coupling between life and its planetary environment is considered fundamental to living processes, the emergence of feedbacks between "life" and environment must be an important process even prior to life's emergence, perhaps even driving it (Mathis et al. 2017a). To develop quantitative frameworks for understanding the emergence of life as a planetary process, a network theory of biogeochemistry is necessary. One possible mathematical framework for formalizing such a theory is multiplex networks. In a multiplex (or multilayer) network, nodes are connected by different types of edges (to be contrasted with bipartite networks where different types of nodes are connected by edges) (Boccaletti et al. 2014). Multiplex networks are often visualized as several networks layered on top of one another. A multiplex network of planetary chemistry would involve several scales of organization, connecting a network of geochemical reactions (at the "bottom") to metabolic processes, to atmospheric chemistry (at the "top"). While biochemical networks are well characterized, geochemical networks and atmospheric networks remain relatively unexplored. Some work on the network structure of planetary atmospheres has revealed topological differences between Earth's atmospheric reaction network and that of other planetary bodies (with atmospheres) in our solar system, such as Mars, Venus, and Titan (Sole and Munteanu 2004; Gleiss et al. 2001); suggestive network theory can distinguish properties of living worlds from those of nonliving worlds. Developing a network theory of planetary (biogeo)chemistry would also allow astrobiologists to incorporate information about exoplanets, such as atmospheric spectra and planetary composition, into a unified framework that will be essential for characterizing alien biosignatures.

10.5 Conclusions

In the last century, prebiotic chemists have focused on identifying the molecular aspects of biochemistry which may have played prominent roles in the origin of life on Earth. However, while there is much debate about whether or not all life will share a common chemistry, it is less debated that life will display some form of organization (Schrodinger 1944). The organizational principles of life likely transcend levels of organization within the biosphere: we see evidence of the role of information in organizing living matter within cells, in intracellular signaling in living tissues, in ecosystems, and in societies. Adopting a view whereby it is the structure of interactions and transformations which defines the living state, and how that is mediated by information, life becomes a property not just of the chemistry within our cells but of the organization of that chemistry. Network science provides a natural quantitative framework to study this. By shifting to a system-level perspective, and embracing the tools of provided by network science, prebiotic chemistry will be able to not only understand the synthesis of molecules relevant to life on the primitive Earth but, perhaps more importantly, how those molecules collectively drove the emergence of the first living systems.

References

Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74 (1):47–97. https://doi.org/10.1103/RevModPhys.74.47

Alon U (2003) Biological networks: the tinkerer as an engineer. Science 301(5641):1866–1867. https://doi.org/10.1126/science.1089072

Andersen JL, Andersen T, Flamm C, Hanczyc MM, Merkle D, Stadler PF (2013) Navigating the chemical space of HCN polymerization and hydrolysis: guiding graph grammars by mass spectrometry data. Entropy 15(10):4066–4083. https://doi.org/10.3390/e15104066

Andersen JL, Flamm C, Merkle D, Stadler PF (2017) An intermediate level of abstraction for computational systems chemistry. Philos Trans R Soc A Math Phys Eng Sci 375 (2109):20160354. https://doi.org/10.1098/rsta.2016.0354

Arita M (2004) The metabolic world of *Escherichia coli* is not small. Proc Natl Acad Sci U S A 101 (6):1543–1547. https://doi.org/10.1073/pnas.0306458101

Ashkenasy G, Ghadiri MR (2004) Boolean logic functions of a synthetic peptide network. J Am Chem Soc 126(36):11140–11141. https://doi.org/10.1021/ja046745c

Ashkenasy G, Jagasia R, Yadav M, Ghadiri MR (2004) Design of a directed molecular network. Proc Natl Acad Sci U S A 101(30):10872–10877. https://doi.org/10.1073/pnas.0402674101

Barabasi A-L (2009) Scale-free networks: a decade and beyond. Science 325(5939):412–413. https://doi.org/10.1126/science.1173299

Barabási A-L (2016) Network science. Cambridge University Press, Cambridge

Barabasi L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–513. https://doi.org/10.1126/science.286.5439.509

Barabási A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113. https://doi.org/10.1038/nrg1272

Bianconi G, Barabási A-L (2000) Competition and multiscaling in evolving networks. Europhys Lett 54(May):436–442. https://doi.org/10.1209/epl/i2001-00260-6

Boccaletti S, Bianconi G, Criado R, del Genio CI, Gómez-Gardeñes J, Romance M, Sendiña-Nadal I, Wang Z, Zanin M (2014) The structure and dynamics of multilayer networks. Phys Rep 544(1):1–122. https://doi.org/10.1016/j.physrep.2014.07.001

Cairns-Smith AG (1986) Clay minerals and the origin of life. Cambridge University Press, New York

Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661–703. https://doi.org/10.1137/070710111

Cohen R, Erez K, Ben-Avraham D, Havlin S (2001) Breakdown of the Internet under intentional attack. Phys Rev Lett 86(16):3682–3685. https://doi.org/10.1103/PhysRevLett.86.3682

Cronin L, Walker SI (2016) Beyond prebiotic chemistry: what dynamic network properties allow the emergence of life? Science 352(6290):1174–1175. https://doi.org/10.1126/science.aaf6310

Davies PCW, Walker SI (2016) The hidden simplicity of biology. Rep Prog Phys 79(10):102601. https://doi.org/10.1088/0034-4885/79/10/102601

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M et al (2008) Functional metagenomic profiling of nine biomes. Nature 452(7187):629–632. https://doi.org/10.1038/nature06810

Eigen M (1971) Self-organisation of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465–523

Eigen M, Schuster P (1978) The hypercycle. a principle of natural self-organisation. Part C: the realistic hypercycle. Naturwissenschaften 65(2):341–369. https://doi.org/10.1007/BF00420631

Erdös P, Rényi A (1959) On random graphs. Publ Math 6:290–297. https://doi.org/10.2307/1999405

Filisetti A, Graudenzi A, Serra R, Villani M, Füchslin RM, Packard N, Kauffman SA, Poli I (2012) A stochastic model of autocatalytic reaction networks. Theory Biosci 131(2):85–93. https://doi.org/10.1007/s12064-011-0136-x

Gisiger T (2001) Scale invariance in biology: coincidence or footprint of a universal mechanism? Biol Rev Camb Philos Soc 76(2). Arizona State University Libraries:S1464793101005607. doi: https://doi.org/10.1017/S1464793101005607

Gleiss PM, Stadler PF, Wagner A, Fell D a (2001) Relevant cycles in chemical reaction networks. Adv Complex Syst 4(02n03):207–226. https://doi.org/10.1142/S0219525901000140

Goldenfeld N, Biancalani T, Jafarpour F (2017) Universal biology and the statistical mechanics of early life. Philos Trans R Soc A 375(2109):20160341. https://doi.org/10.1098/RSTA.2016.0341

Goldford JE, Hartman H, Smith TF, Segrè D (2017) Remnants of an ancient metabolism without phosphate. Cell 168(6):1126–1134.e9. https://doi.org/10.1016/j.cell.2017.02.001

Handorf T, Ebenhöh O, Heinrich R (2005) Expanding metabolic networks: scopes of compounds, robustness, and evolution. J Mol Evol 61(4):498–512. https://doi.org/10.1007/s00239-005-0027-1

Hordijk W, Steel M (2004) Detecting autocatalytic, self-sustaining sets in chemical reaction systems. J Theor Biol 227(4):451–461. https://doi.org/10.1016/j.jtbi.2003.11.020

Hordijk W, Steel M (2017) Chasing the tail: the emergence of autocatalytic networks. BioSystems 152:1–10. https://doi.org/10.1016/j.biosystems.2016.12.002

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. https://doi.org/10.3390/ijms12053085

Hordijk W, Steel M, Kauffman S (2012) The structure of autocatalytic sets: evolvability, enablement, and emergence. Acta Biotheor 60(4):379–392. https://doi.org/10.1007/s10441-012-9165-1

Hordijk W, Hasenclever L, Gao J, Mincheva D, Hein J (2014) An investigation into irreducible autocatalytic sets and power law distributed catalysis. Nat Comput 13(3):287–296. https://doi.org/10.1007/s11047-014-9429-6

Jeong H, Albert R, Ottval ZN, Barabási AL (2000) The large scale organization of metabolic networks. Nature 407(6804):651–654

Kamimura A, Kaneko K (2010) Reproduction of a protocell by replication of a minority molecule in a catalytic reaction network. Phys Rev Lett 105(26):1–4. https://doi.org/10.1103/PhysRevLett.105.268103

Kaneko K, Yomo T (2002) On a kinetic origin of heredity: minority control in a replicating system with mutually catalytic molecules. J Theor Biol 214(4):563–576. https://doi.org/10.1006/jtbi.2001.2481

Kauffman S (1993) The origins of order: self-organization and selection in evolution. Oxford University Press, New York

Kim H, Davies P, Walker SI (2015) New scaling relation for information transfer in biological networks. J R Soc Interface 12(113):20150944. https://doi.org/10.1098/rsif.2015.0944

Kim H, Smith H, Mathis C, Raymond J, Walker SI (2018) Universal scaling across biochemical networks on earth. bioRxiv. https://doi.org/10.1101/212118

Kuppers B-O (1990) Information and the origin of life. MIT Press, Cambridge, MA

Larhlimi A, Blachon S, Selbig J, Nikoloski Z (2011) Robustness of metabolic networks: a review of existing definitions. BioSystems 106(1):1–8. https://doi.org/10.1016/j.biosystems.2011.06.002

Lifson S (1997) On the crucial stages in the origin of animate matter. J Mol Evol 44(1):1–8. https://doi.org/10.1007/PL00006115

Mathis C, Bhattacharya T, Walker SI (2017a) The emergence of life as a first-order phase transition. Astrobiology 17(3):266–276. https://doi.org/10.1089/ast.2016.1481

Mathis C, Ramprasad S, Walker S, Lehman N (2017b) Prebiotic RNA network formation: a taxonomy of molecular cooperation. Life 7(4):38. https://doi.org/10.3390/life7040038

Maynard Smith J, Szathmáry E (1995) The major transitions in evolution. Oxford University Press, Oxford

Milo R (2002) Network motifs: simple building blocks of complex networks. Science 298 (5594):824–827. https://doi.org/10.1126/science.298.5594.824

Montañez R, Medina MA, Solé RV, Rodríguez-Caso C (2010) When metabolism meets topology: reconciling metabolite and reaction networks. BioEssays 32(3):246–256. https://doi.org/10.1002/bies.200900145

Newman MEJ (2004) Analysis of weighted networks. Phys Rev E Stat Nonlinear Soft Matter Phys 70(5 2):1–9. https://doi.org/10.1103/PhysRevE.70.056131

Nghe P, Hordijk W, Kauffman SA, Walker SI, Schmidt FJ, Kemble H, Yeates JAM, Lehman N (2015) Prebiotic network evolution: six key parameters. Mol BioSyst 11(12):3206–3217. https://doi.org/10.1039/c5mb00593k

Nowak MA, Ohtsuki H (2008) Prevolutionary dynamics and the origin of evolution. Proc Natl Acad Sci U S A 105(39):14924–14927. https://doi.org/10.1073/pnas.0806714105

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27(1):29–34. https://doi.org/10.1093/nar/27.1.29

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi A-L (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–1555. https://doi.org/10.1126/science.1073374

Raymond J, Segre D (2006) The effect of oxygen on biochemical networks and the evolution of complex life. Science 311(5768):1764–1767. https://doi.org/10.1126/science.1118439

Schrodinger E (1944) What is life? Cambridge University Press

Schütte M, Skupin A, Segrè D, Ebenhöh O (2010) Modeling the complex dynamics of enzyme-pathway coevolution. Chaos 20(4):1–12. https://doi.org/10.1063/1.3530440

Segre D, Ben-Eli D, Lancet D (2000) Compositional genomes: prebiotic information transfer in mutually catalytic noncovalent assemblies. Proc Natl Acad Sci U S A 97(8):4112–4117. https://doi.org/10.1073/pnas.97.8.4112

Segré D, Ben-Eli D, Deamer DW, Lancet D (2001) The lipid world. Orig Life Evol Biosph 31 (1–2):119–145. https://doi.org/10.1023/A:1006746807104

Sessions AL, Doughty DM, Welander PV, Summons RE, Newman DK (2009) The continuing puzzle of the great oxidation event. Curr Biol 19(14):R567–R574. https://doi.org/10.1016/j.cub.2009.05.054

Shock EL, Boyd ES (2015) Principles of geobiochemistry. Elements 11(6):395–401
Smith E, Morowitz H (2016) The origin and nature of life on earth: the emergence of the fourth geosphere. Cambridge University Press, Cambridge
Sole RV, Munteanu A (2004) The large-scale organization of chemical reaction networks in astrophysics. EPL (Europhys Lett) 68(2):170
Sousa FL, Hordijk W, Steel M, Martin WF (2015) Autocatalytic sets in *E. coli* metabolism. J Syst Chem 6(1):4. https://doi.org/10.1186/s13322-015-0009-7
Sterelny K (2015) Universal biology. Br Soc Philos Sci 48(4):587–601
Vaidya N, Manapat ML, Chen IA, Xulvi-Brunet R, Hayden EJ, Lehman N (2012) Spontaneous network formation among cooperative RNA replicators. Nature 491(7422):72–77. https://doi.org/10.1038/nature11549
Vaidya N, Walker SI, Lehman N (2013) Recycling of informational units leads to selection of replicators in a prebiotic soup. Chem Biol 20(2):241–252. https://doi.org/10.1016/j.chembiol.2013.01.007
Vasas V, Szathmary E, Santos M (2010) Lack of evolvability in self-sustaining autocatalytic networks constraints metabolism-first scenarios for the origin of life. Proc Natl Acad Sci U S A 107(4):1470–1475. https://doi.org/10.1073/pnas.0912628107
Vasas V, Fernando C, Santos M, Kauffman S, Szathmáry E (2012) Evolution before genes. Biol Direct 7(1):1. https://doi.org/10.1186/1745-6150-7-1
Wagner A (1998) The large-scale structure of metabolic networks: a glimpse of life's origin? Wiley Periodicals 8(1): 15–19
Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc R Soc B Biol Sci 268(1478):1803–1810. https://doi.org/10.1098/rspb.2001.1711
Walker SI (2017) Origins of life: a problem for physics, a key issues review. Rep Prog Phys 80 (9):092601. https://doi.org/10.1088/1361-6633/aa7804
Walker SI, Davies PCW (2013) The algorithmic origins of life. J R Soc Interface 10(79):20120869. https://doi.org/10.1098/rsif.2012.0869
Walker SI, Kim H, Davies PCW (2016) The informational architecture of the cell. Philos Trans R Soc A Math Phys Eng Sci 374(2063). https://doi.org/10.1098/rsta.2015.0057
Wang RS, Saadatpour A, Albert R (2012) Boolean modeling in systems biology: an overview of methodology and applications. Phys Biol 9(5):055001. https://doi.org/10.1088/1478-3975/9/5/055001
Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
West GB (1999a) The fourth dimension of life: fractal geometry and allometric scaling of organisms. Science 284(5420):1677–1679. https://doi.org/10.1126/science.284.5420.1677
West GB (1999b) The origin of universal scaling laws in biology. Phys A Stat Mech Appl 263 (1–4):104–113. https://doi.org/10.1016/S0378-4371(98)00639-6
Wynveen A, Fedorov I, Halley JW (2014) Nonequilibrium steady states in a model for prebiotic evolution. Phys Rev E 89(2):22725. https://doi.org/10.1103/PhysRevE.89.022725
Yockey H (2005) Information theory, evolution and the origin of life. Cambridge University Press, Cambridge