Advances in Complex Systems, Vol. 6, No. 1 (2003) 15–35

MESOBIOTIC EMERGENCE: MOLECULAR AND ENSEMBLE COMPLEXITY IN EARLY EVOLUTION

BARAK SHENHAV

Department of Molecular Genetics and Crown Human Genome Center,
Weizmann Institute of Science, Rehovot 76100, Israel

DANIEL SEGRÈ

Lipper Center for Computational Genetics and Department of Genetics,
Harvard Medical School, Boston, MA 02115, USA

DORON LANCET

Department of Molecular Genetics and Crown Human Genome Center,
Weizmann Institute of Science, Rehovot 76100, Israel
doron.lancet@weizmann.ac.il

Received 9 November 2002
Accepted 4 December 2002

In addition to the visible complexity expressed in the morphogenesis of multicellular organisms, two levels of microscopic complexity may be discerned within every living cell. The first level is related to covalently bonded structures, namely molecules. The second level has to do with the generation of non-covalent molecular assemblies. Origin of life research has largely focused on the first complexity level, i.e. the appearance of covalent biopolymers. We present a life emergence scenario based mainly on the second complexity level. We argue that homeostatic molecular ensembles, for which we have coined the term "mesobiotic," have assumed a half-way position between prebiotic organic synthesis and full-fledged cellular (biotic) life.

Keywords: Mesobiotic; origin of life; metabolism first; chemical complexity; ensemble complexity.

1. The Complexity of Life

Individual molecule complexity. Because of the peculiarities of carbon chemistry, organic molecules are composed of atoms in configurations that span a vast range of sizes. The simpler ones have only about a dozen atoms, for instance, the amino acid alanine. More elaborate molecules contain 100–200 atoms, as exemplified by peptides, phospholipids or oligosaccharides. Next are proteins and RNAs, with a few thousand atoms. Heading the list are chromosomes, containing single DNA molecules with millions or even billions of atoms.

Such molecular complexity is the most obvious chemical attribute of life. It has been the subject of nearly two centuries of biological chemistry research. By now, chemical bond configurations are known for practically all the molecules in living cells, and there is also an accurate three-dimensional structure for thousands of them. The most celebrated achievement of the previous century in this respect is the elucidation of the exact molecular configuration — the DNA sequence — of a large number of chromosomes, constituting entire genomes [37, 77].

But it is important to note that this first level of complexity is also abundantly present in the inanimate world. Interstellar dust particles [43] and carbonaceous meteorites [12, 57] have been shown to contain hundreds of organic substances. Prebiotic experimental simulations of electric sparks and mineral surface chemistry have shown evidence for the formation of numerous organic compounds [47, 79]. All the above appear to also contain relatively large "tar-like" organic polymers [45, 48]. In fact, the word "organic," often used in this context, is judged by some to be misleading. Simply interpreted, it means "derived from life," but ever since Wöhler's synthesis of urea [33] it is no longer taken to mean "present only in life." Thus, molecular complexity, as manifested in the chemistry of carbon, hydrogen, nitrogen, oxygen, phosphorous and sulfur (CHNOPS) may be considered a universal phenomenon, not restricted to living cells.

One relevant property of large molecules is combinatorial complexity. This term usually implies that if one examines molecules belonging to a given class, say hexapeptides, there exists a very large set of analogous structures, obtained as different combinations of the same or similar building blocks. In this example, if only the 20 natural L-amino acids are allowed, there are $20^6 = 64{,}000{,}000$ different possible hexapeptides. If additional amino acid types are considered, such as could well have been present under prebiotic conditions, the numbers grow far larger, increasing as X^6, where X is the number of different amino acid types.

Molecular ensemble complexity. Living cells are collections of organic substances. Their constituent molecules are held together by a variety of non-covalent forces. These include hydrophobic interactions that accrete lipids in biological membranes, hydrogen bonds with which DNA strands adhere to each other and dipole and ionic interactions that stabilize the quaternary structure of multi-subunit proteins. Non-covalent interactions make the emergence of structures much larger than even the largest of molecules possible, hence generate a new level of complexity. Even the simplest living cell is a "coalition" of thousands of different types of molecules, mostly attached to each other by variations on the "lock and key" theme [41].

In the final account, many of the properties of a living entities may be regarded as stemming from the cell's molecular composition, i.e. which molecules are present within it and in how many copies. The more elaborate properties often arise from specific interactions among such molecules to form supramolecular structures such as ribosomes, chromosomes, membranes, centrioles and enzyme complexes. Thus, the second level of complexity in living organisms resides in bringing together in a rather orderly manner ensembles of organic molecules, large and small. The configuration of such assemblies may be regarded as the most unique aspect of life.

Forming clumps of organic molecules is of course rather trivial. What distinguishes life is the generation of many similar copies of such assemblies, with well-defined, idiosyncratic, molecular composition. This idiosyncrasy resides in the counts of molecular types present, and even more significantly, in those types that are not present. Life entails a "reduction of possibilities," whereby only a minute subset of all possible covalent structures is actually present within any living cell [66].

It is here that combinatorial chemistry and ensemble complexity interrelate. A living cell is built in such a way that only very few members of any relevant combinatorial repertoire are actually present. This is true all across the spectrum, from the level of monomers (sugars, lipids, amino acids, nitrogen bases), all the way to biopolymers. Usually, when pondering a cell's protein inventory, one tends to focus on the *molecular* complexity of each individual polypeptide. But what is more awe-inspiring is that cells are equipped with kinetic bias mechanisms, which account for the fact that only a very small subset of all the possible strings of amino acids in the appropriate range of sizes and compositions are present in them. For every protein that is present, there is an astronomically large number of amino acid strings that is absent. Thus, a crucial facet of cellular ensemble complexity is manifested in idiosyncratic chemical composition.

Prebiotic complexity. Several generations of origin-of-life scientists have investigated possible mechanisms by which complexity of the first kind has emerged. They have asked about pathways for prebiotic synthesis of amino acids and nucleotides, and studied the means by which such monomers could assemble into longer covalent polymers, RNA and proteins, as a preamble for life's emergence. It is still widely assumed that if we understood how such individual molecules formed, an understanding of life's origin would readily follow. The RNA world view is so strongly rooted, that even when discussing pre-RNA scenarios, these are usually taken to imply just different chemistries of RNA-like polymers [17, 39].

A dissenting view, expounded in the present paper, is that the processes leading to life would be much better understood if ensemble complexity were given a more thorough consideration. The molecules that harbor this level of complexity may themselves be relatively simple. They must be able to accrete together spontaneously, even from dilute "soup" solutions, to form assemblies, whose dynamic behavior manifests life-like attributes. In the last few years, several researchers, including our own group, have demonstrated the validity of this approach. It was shown that some non-covalent assemblies are capable of propagating their compositional complexity, or compositional information, without the involvement of long biopolymers such as RNAs or protein enzymes ([64], see below). Accordingly, "pre-RNA" is taken to imply chemistries totally different from that of RNA.

This unorthodox view encounters considerable opposition, based, among others, on the notion that no life is possible without biopolymers (see for example [39, 43]). A general response to this criticism might be that if one wishes to understand the seemingly intractable process that led from inanimate chemistry to cellular life, it is necessary to keep one's mind open to the possibility that early life forms may have utilized functional components very different from those of their present day counterparts. An extreme form of such proposition is that life began with inorganic clays [8]. Here, we subscribe to the more conservative notion that life has emerged based on organic compounds. A redefinition of terms may be helpful for discussing the possibility of life without biopolymers.

Defining mesobiotic entities. Perhaps the most crucial issue for discussing prebiotic evolution is embodied in the following paradox. When studying early events in the emergence of life, one is forced to discuss a transition from what is unanimously defined as inanimate, to entities that practically all researchers will consider alive, e.g. a primitive bacterium. But, because of the broad agreement that living entities could not have emerged without selection and evolution, and because most researchers would consider evolution an attribute of life, the term "prebiotic evolution" harbors an intrinsic inconsistency.

It would therefore be helpful to regard life's emergence as a graded series of steps rather than as an abrupt transition (Fig. 1). In the definition favored here,

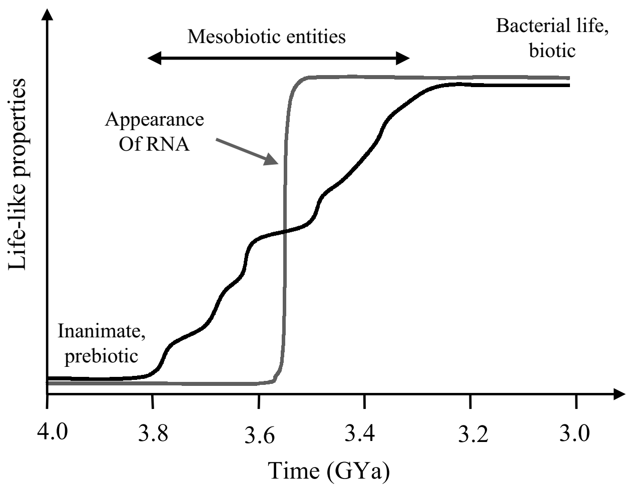

Fig. 1. Two conceptual views of the origin of life. In the first (gray line) a relatively abrupt transition is depicted where upon the appearance of RNA, entities begin to be considered alive. The second (black line) holds that the process which led from inanimate to living was a long and very graded one. The departure from baseline indicates a capacity to store and propagate information, which in the first period may have occurred without the involvement of templating RNA-like biopolymers. The time axis is provided for illustration only. But it highlights the idea that a graded origin might have taken hundreds of millions of years, rendering some of the arguments about the exact dating of the first cellular life more intractable.

"prebiotic" would relate to organic molecules, even with high covalent complexity, generated by inanimate chemical processes. It is customary to reserve the term "biotic" to cells that contains elements of the standard life machinery, i.e. RNA-like and/or protein-like molecules. To deal with the graded transition between these two dichotomous states, it would be useful to employ a new term, *mesobiotic entities*, to describe intermediate stages. This is also suitable because it relates to the notion that the early molecular assemblies belonged to the realm of mesoscopic entities, on the border between the microscopic and macroscopic [24, 42, 53].

A mesobiotic entity is envisioned as being endowed with both molecular and ensemble complexity. It should be capable of rudimentary catalysis, as well as of information storage and transfer, but still lack many of the standard components of living cells. In particular, mesobiotic entities may be thought to be largely devoid of long catalytic or informational biopolymers. The definition of such entities, half way along the transition from inanimate to living, is analogous to some extent to previously used definitions, e.g. that of *protocells* [51], and is expected to generate a well-defined arena for discussing the mechanisms of the origin of life.

2. Random Chemistry and Life's Emergence

Planetary random chemistry. Prebiotic earth has probably seen a plethora of chemical processes which occurred randomly, with relatively few specific guiding principles. Kinetic and thermodynamic parameters have surely resulted in biases towards certain molecular structures. External agents such as mineral catalysts could also exert their biasing action [82]. But if one considers the realm of random chemistry, where millions or billions of compounds may form by alternative arrangements of relatively few types of building blocks, such differentiating parameters could not play a decisive role. By and large, it may be assumed that early terrestrial chemistry displayed a very large diversity of chemical compounds.

A major question related to mesobiotic evolution is where, along the time axis, entities appeared which manifested increasing degrees of deviation from randomness. This emergence process is at the heart of the unsolved question of life's beginning. We propose that this emergence occurred mainly along the lines of generation of assemblies (complexity of the second kind), and that it may be possible to look for relevant guidance in the realm of cellular network analysis, and in systems biology [34]. This is in contrast to the prevailing trend which centers on the field of organic chemistry and the rules of formation of organic polymers (complexity of the first kind). In our view, covalent organic chemistry provides the building blocks, which then cross the elusive boundary between inanimate and living, based on complexity of molecular assemblies.

Metabolism first. The notion that the crux of life's emergence rests in molecular ensembles rather than in the properties of individual molecules is not really new. In fact, the very first systematic scientific endeavor in this domain, by Alexander

Oparin [55, 56], related to molecular ensembles (coacervates). Later, proponents of related views preached a "metabolism first" scenario [2, 16, 18, 30]. They argued that metabolic networks may get established through the appearance of numerous molecular species, which get formed from each other and from "foodstuff" precursors. This happens in a similar way to how dozens of compounds are generated by molecular interconversions in a cellular pathway, such as glycolysis. Specific chemistries have also been invoked to support this notion, for example molecules akin to those found in the citric acid cycle [52], stemming from thioester chemistry [10], or related to those that form by the catalytic and energetic impact of the sulfur iron mineral, pyrite [78]. Proponents of the metabolism first view were, however, disadvantaged by the question of how information could be stored and propagated along generations in a system composed only of low molecular weight metabolites.

Important progress in this domain was made through the idea that replication could, after all, be embodied in a metabolic cycle [49, 50]. A careful examination of the citric acid cycle reveals that in going around a full circle of biosynthetic reactions, one begins with a single copy of a molecule like citric acid, and ends up with two copies. The second copy gets synthesized in a way that involves the reaction of inorganic precursors with intermediates in the circle. Thus, conceptually, the metabolic pathway may be viewed as capable of making more copies of itself. A more general conceptual view in this vein was made by invoking a large network of interactions that results in a self-reproduction behavior [16, 31, 66], as described below.

In the metabolism first models, genetic information is assumed to be formed as a result of metabolic reactions. In other words, the diverse reactions of metabolism lead to the synthesis of informational biopolymers — RNA and DNA. Proponents of this alternative view to the RNA world do not negate the central role that RNA-type molecules may have played. But, they dismiss the notion that an abiotic set of reactions could have led directly to a full-fledged self-templating polymer. This argument is partially a semantic one, since all agree that chemistry had to generate the monomers necessary for information-storing polymers, and that catalytic events of the type found in metabolism were also needed for polymerization. The argument's core is whether the events preceding the first RNA molecule were just a set of inanimate reactions, perhaps driven by external mineral catalysts, or were reactions taking place within an entity describable as having attributes of life. Here, again, the proposed terminology becomes useful: organic synthesis may be defined as prebiotic, rudimentary proto-metabolism as mesobiotic, and RNA-containing entities as biotic.

Replication and metabolism. Practically all students of life's emergence agree that an entity can be considered to have crossed the line between inanimate and living, if it is capable of self replication and of undergoing natural selection. The incredulity with which the idea of metabolism first is often viewed by proponents of RNA world relates to disbelief in a metabolic system's capacity to store and propagate information. In other words, many cannot imagine how a "naked metabolism" could transfer useful properties from one generation to another [38]. In a conceptual framework governed by modern cell biology, any information transfer mechanism inferior to base-pair templating is considered implausible. Yet, we believe that transcending this dogma is essential for any progress in the quest for life's beginning.

There is thus a sore need to invoke a mesobiotic mechanism for information storage and transfer, independent of polymer strings whose sequence constitutes the biologically relevant information. One has to carefully define information in a more general way than related to a sequence of symbols or monomers. This is done in Sec. 5 below.

3. The Chicken-and-Egg Riddle

Chicken and egg. Numerous researchers have pointed out the inherent paradox of the origin of life: metabolism cannot form without genetic material and *vice versa.* This paradox is accentuated if one adheres to the idea that life must have begun with one "pure" type of entity, either catalysts capable of performing metabolic tasks, or polymeric replicators that can store and copy information. In present-day life, these two categories are irrevocably intertwined. No information-storing nucleic acids can perform even the most rudimentary self replication and coding without the help of protein catalysts, but no new copies of a protein may be made unless coded for by a nucleic acid. Why, then, should one insist on the dichotomous separation of metabolism and templating at the very early stages of life's emergence? Could there not have been a mesobiotic entity embodying primitive attributes of both? This idea has been explored by Freeman Dyson [16], although through most of his book he favors an alternative, dual origin scenario.

A lipid world. One of the innovative concepts developed recently in the origin of life community is that lipids may have played a key role in early evolution [11, 44, 63]. To many readers, the term "lipids" signifies rather dull molecules that make up cellular membranes. However, if one views lipid chemistry in a more open-minded way, a different picture may emerge. In this broader perspective, any molecule with one or more of a hydrophobic tail and a hydrophilic head is a lipid. Early lipids may have possessed a diversity of head groups, including an imidazole group or other functionalities that could perform catalytic functions. Indeed, catalytic capacities of lipids and other amphiphiles has been demonstrated experimentally [1, 7]. Catalytic lipid assemblies or "lipozymes" may participate in primitive metabolic networks [63]. Thus, lipids may have played a triple role, as catalysts, information carriers and also as compartment-forming molecules. It is however difficult to imagine how lipid assemblies would generate progeny. The answer is provided below. But in order to prepare for this, some additional clarifications of terminology are needed.

Replication and reproduction. In population biology and ecology, the term "reproduction" is often used to signify growth of a population due to a process in which individual members generate progeny. An argument has been made that when speaking about an assembly of molecules "reproduction" should be used in lieu of "replication" [6]. The latter, it was stated, should only be used to address to a process in which a single molecule makes a copy of itself in an autocatalytic or template related fashion. Yet, it is generally considered legitimate to use the word "replication" for a bacterial cell [23]. We would like to contend that if an entire molecular ensemble is the reference entity, the ensemble may be described as undergoing *replication* in analogy to an entire complex bacterium. This is provided that a process is envisioned in which, as for bacteria, an entire ensemble undergoes a complex sequence of steps that result in the generation of its replica.

Nominal portrayal of cell division. Based on the Central Dogma of biology, a typical text would describe cell division by the following sequence of events:

(i) DNA molecules undergo replication by a templating mechanism;
(ii) RNA molecules are generated by another template-based mechanism — transcription;
(iii) Ribosomes make additional copies of all proteins;
(iv) Self assembly processes allow for production of new copies of organelles, such as ribosomes and mitochondria;
(v) Protein-catalyzed metabolism synthesizes more of all the low molecular weight components, such as sugars, lipids, amino acids and nucleotides;
(vi) When the counts of all cellular components have doubled, thus increasing the cell's volume by a factor of two without appreciable change in concentrations, fission generates two similar daughter cells.

It may be readily realized that steps (i)–(v) describe a set of highly specific mechanisms found in present day cells. In contrast, step (vi) is very general, and summarizes, in a mechanism-independent fashion, the process of replication of a non-covalent molecular assembly, be it a bacterium or a coacervate. A crucial aspect of the models we have developed is an attempt to discern step (vi), without being biased by any of the details of steps (i)–(v). The crucial relevant question is how a mesobiotic entity much more primitive than today's cellular life, e.g. an aggregated molecular ensemble, may undergo a rough doubling of the amount of each and every one of its constituents. This is equivalent to asking what specific chemically-based mechanisms could lead to *homeostatic growth.* Of course early assemblies could not achieve the concerted processes prevailing in cells today, including exact doubling of the DNA molecule, but they might crudely obey similar principles.

4. Mutually Catalytic Assemblies

Thermodynamic idiosyncrasy. For an assembly of molecules to generate "more of itself," a set of mutual interactions must prevail within it. Consider a non-covalent molecular assembly in which N_E types of molecules are represented with different counts n_i. This idiosyncratic sub-repertoire is drawn from a larger repertoire of N_G types, all present in the external environment. For homeostatic growth to prevail within an assembly, it must preferentially absorb in the right ratios molecules belonging to its sub-repertoire, and largely reject other molecular types. Certain molecules may also be synthesized by covalent bond formation within the assembly, contributing to its idiosyncratic composition as compared to the environment. But, for simplicity of argument, and without loss of generality, we consider in this section only joining and leaving reactions. The question at hand is what general physicochemical mechanism would afford biased joining.

The simplest model is thermodynamically-based, drawn from the conceptual world of *self-assembly* reactions. Ribosomes and bacteriophages are known to be formed as highly defined and idiosyncratic molecular assemblies by absorbing certain molecular species from the environment. This process is guided by stereospecific molecular recognition events among the components, driven by thermodynamic free energy gradients. A similar mechanism may lead to the formation of complex supramolecular structures from a unique subset of organic molecules [40]. Any crystallization process also belongs in the same realm.

A similar type of self assembly could also occur in more dynamic, less organized soft-matter systems [61]. A well-known example is the formation of rafts and caveolae within lipid bilayers, whereby certain lipid types segregate in one domain of a two-dimensional structure, and other types are excluded, but are found in other domains [32]. Based on the same principles, it may be envisioned that a lipid assembly would preferentially absorb certain molecular species types, and undergo homeostatic growth, limited by external concentrations. Such forces may have indeed contributed to prebiotic events. However, what is common to all of the phenomena described here is that they are governed by thermodynamics, and occur in a system as it approaches chemical equilibrium.

The model described here provides a complementary route, which harbors some clear advantages related to life's emergence. Life is a non-equilibrium phenomenon, as a living cell is by no means the lowest energy state of its components. A cell maintains its homeostasis or homeostatic growth through a vast network of catalytic events that pertain to chemical reactions far from equilibrium. This is attained with the help of an external free energy supply, through the absorption of high-energy molecules or via the exploitation of light energy. It would be highly useful to explore a model for homeostasis in simpler molecular assemblies, based on similar principles.

Kinetic homeostasis. The general concepts on which *kinetic* homeostasis and homeostatic growth are based have been delineated by previous authors. They relate to the notion of mutually catalytic sets [2, 16, 18, 20, 30]. In this context, an assembly of molecules is envisioned, whereby each member can potentially catalyze the formation of one or more of the other molecule types in the same assembly. These molecules could be peptides of different lengths [31], but the specific nature of the molecules is unimportant. A system governed by mutual catalysis must, by definition, largely reside away from equilibrium.

One conceptual hurdle to accepting the notion of mutually catalytic sets is that the assumption of *pervasive catalysis* is required. In present day living cells, catalytic effects are always associated with macromolecules, proteins and RNAs. This is because a modern cell depends on highly efficient and specific catalysts, with a defined three-dimensional structure and a well-orchestrated multiple-contact active-site configuration. The price paid is the need for an elaborate mechanism — transcription and translation — that ensures the capacity to generate such polymeric catalysts. However, when the early rudiments of life appeared, it is likely that much weaker and non-specific catalysts were sufficient.

In principle, catalysis is a graded phenomenon. Every uncatalyzed chemical reaction has a basal rate, dictated by its free energy of activation. A decrement of this free energy barrier is normally considered as catalysis. While typical biopolymer enzymes induce rate enhancements ranging from 1,000 to more than 1,000,000 fold, smaller molecules may exert more modest rate increases, as low as $\times 1.1$, perhaps topping at $\times 100$. There is a broad literature on such weak catalysts, sometime called enzyme mimetic compounds. These include individual amino acids such as histidine [73] and other amino acids [35], small peptides [4, 62, 76], lipids [19], carbohydrates [9, 75, 76], organic metal chelates [27] and many more. Thus, there should be no reason to reject the notion that within an arbitrary set of molecules, some catalytic effects would exist. The strength of the mutually catalytic network will of course depend on the exact selection of the components. Some sets will show a much tighter catalytic network, others a rather loose one.

One approach to assessing the strength of mutual catalysis is based on defining a catalytic threshold. In this view, when examining pairwise catalytic interactions within a molecular set, a dichotomous probability distribution is envisioned, whereby each pair is classified as either catalytic (with a probability p) or not (with a probability $1 - p$) [31]. A central result of these studies is that for a defined value of p, there is a repertoire size N_G beyond which every molecule is catalytically related to at least one other, a state defined as *catalytic closure* [30, 31, 46]. A more general formulation proposed by us invokes a continuous distribution of rate enhancement values, whereby different catalytic strengths β, large or small, are described by a probability $P(\beta)$ [36, 64, 66]. In this formalism every molecular assembly is catalytically closed, but the effectiveness of closure may vary widely.

Whatever model is preferred, the conceptual outcome is highly important. Simulated chemical kinetic behavior of such systems demonstrates the potential existence of molecular assemblies which undergo dynamic exchange of matter with the environment, are energy-dependent, grow homeostatically, and have a potential to generate progeny [64, 69]. These systems harbor a capacity for *kinetic homeostasis*. They may assume stationary states far from equilibrium and display properties, similar, in a rudimentary way, to those of present-day living cells.

5. Compositional Information

Sequence versus composition. In the previous sections we alluded to the question of what kind of information could be propagated in an early molecular assembly, devoid of long biopolymers. The notion of sequence-based biological information is so deeply rooted that it is difficult to conceive an alternative. Yet, any proponent of a "metabolism first" scenario is obliged to come up with an alternative definition for information. To this end, we have developed the concept of compositional information [64, 69] (Fig. 2).

Both sequential and compositional information relate to manipulations carried out with a repertoire of monomers. In the case of sequential information, the monomers are threaded into strings, and the exact order matters. Of course, polymeric strings also have compositional information (e.g. amino acid compositions of proteins), but this type of information is rightly considered inferior. Notably, though, all the proteins in a large genome may be uniquely identified solely by specifying their amino acid composition [81].

For compositional information, only the tally of different molecular species is considered. To realize intuitively that compositions may carry a significant amount of information, consider the following example. A protein composed of a string of 20 different monomers and having a length of 23 can be constructed in 20^{23} different ways, hence has $\log_2(20^{23}) \approx 100$ bits of information, as defined by Shannon [80]. A

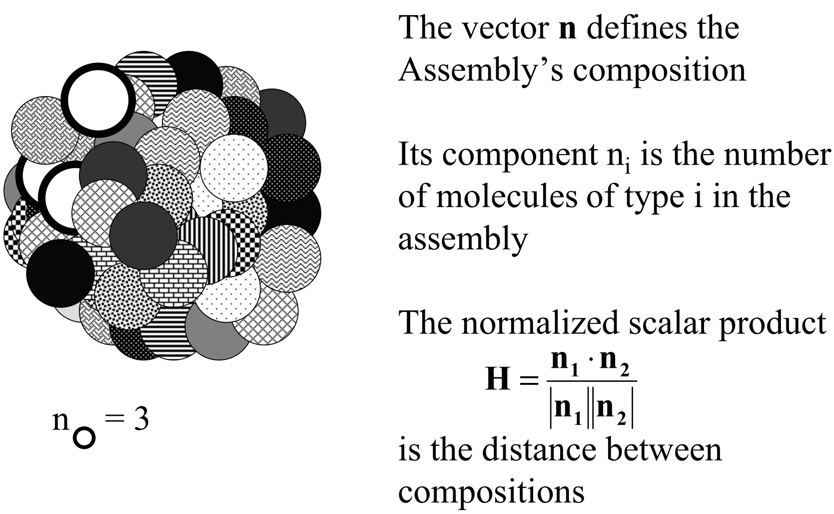

Fig. 2. Non-covalent assemblies have compositional information, as defined in this figure. The counts of different molecular kinds in the assembly constitute the compositional vector **n**. In this schematic representation of an assembly, the component correspondent to the count of "gray" molecular components is highlighted in the bottom left corner of the figure. The similarity measure H between different assemblies in composition space, which can be computed using their composition vectors, can serve as an indicator of homeostatic preservation along GARD assembly growth and evolution.

compositional assembly that contains between 1 and 20 copies of each of 23 types of monomers has exactly the same information content, because it is represented by a 23-long compositional vector n (see Fig. 2), with each component assuming values between 1 and 20. The price paid is that such an assembly has to contain, on average, 230 monomers, as opposed to only 23 for a sequential polymer. But this is not a vast difference in size. A more general illustration of such relationships for different entities containing 100 bits of information is shown in Fig. 3. It is clearly seen that for small N_G values, the compositional assemblies required are much larger than the polymer that carries an equivalent amount of information. But when N_G grows and reaches a still reasonable range of a few thousands, the required sizes become asymptotically similar.

Maintaining compositional information. Just as templating is an ideal mechanism for propagating sequential information, mutual catalysis is well suited to propagate compositional information. In the case of a molecular assembly, as for a living cell,

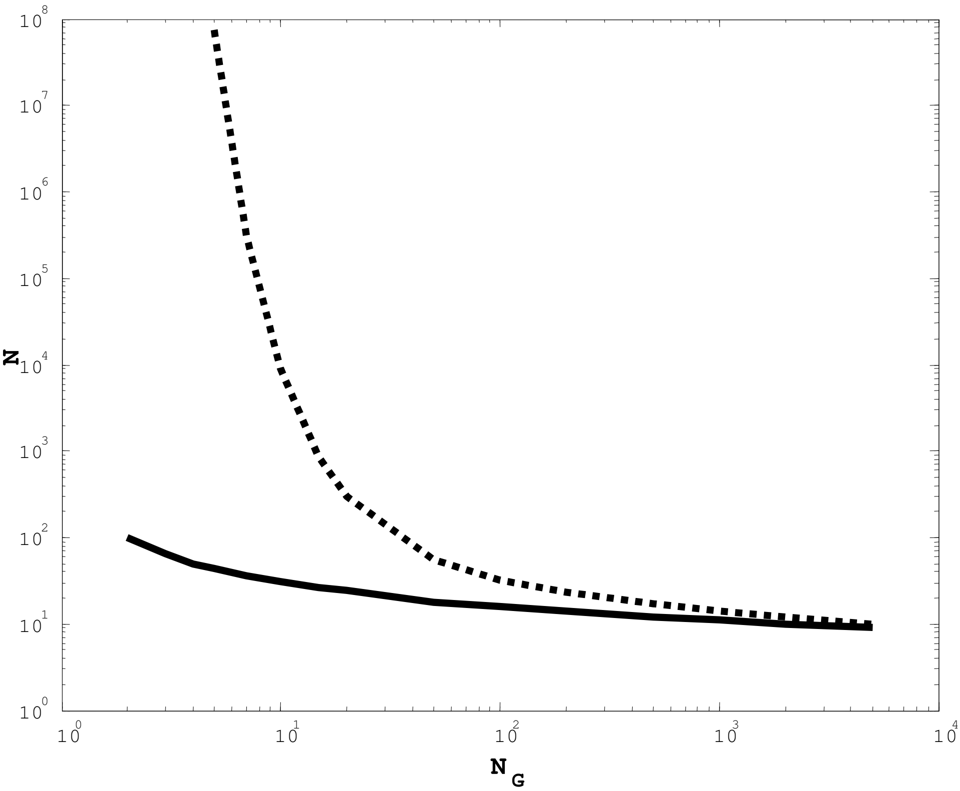

Fig. 3. Required N (sequence length or assembly size) for encoding 100 binary bits by a sequence (full line) or assembly (dashed line) as function of the size of the molecular repertoire (N_G). The number of bits encoded by a sequence of length N from a repertoire of size N_G is $N \log_2(N_G)$. The number of bits encoded by an assembly of size N from a repertoire of size N_G is $\log_2 \binom{N_G+N-1}{N}$.

the first step in generating compositional progeny is doubling the counts of all molecules. It is easy to realize that such an objective is entirely equivalent to a two fold homeostatic growth. This is because if all concentrations $c_i = n_i/V$ should remain unchanged when the volume V grows two-fold, the molecular counts n_i also have to increase by the same factor. A major claim of the work done in our laboratory is that mutually catalytic networks may undergo homeostatic growth [65, 67, 68]. The formal description of such behavior is briefly provided in the next section.

6. Propagating Compositional Information

The GARD model. The Graded Autocatalysis Replication Domain (GARD) model (Fig. 4) is so named because it describes a behavior of a spatial domain constituting a molecular assembly. Molecules in the assembly manifest mutual catalysis, resulting

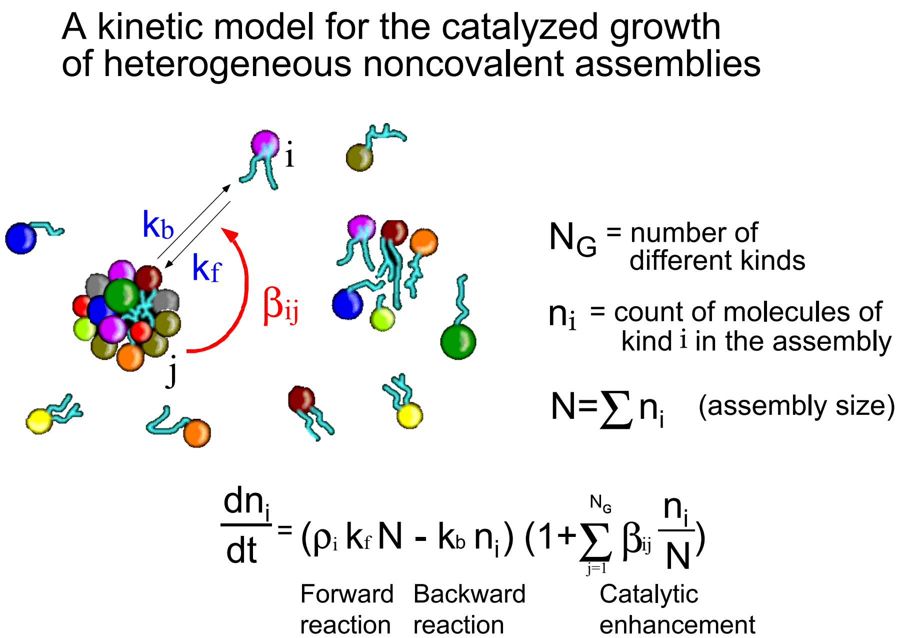

Fig. 4. **GARD: the Graded Autocatalysis Replication Domain model**. The model is based on computer simulations of faithful chemical behavior. It is thus distinct from Artificial Life approaches [5]. The model has been described in detail in several published papers [64, 67–69]. It involves a stochastic chemistry simulation based on a set of differential equations as shown. The main reaction step, in the simplest amphiphile GARD formulation, is the transfer of an amphiphilic molecule A_i between the environment and an assembly (straight arrows). A key aspect in reaching a kinetic homeostasis is the dependence of the reaction rates on the current composition of the assembly. Here, k_f and k_b are forward and backward basal rate constants, ρ_i are exernal concentrations of the different chemical species. The matrix β_{ij} signifies the mutual rate enhancement parameters for the catalysis exerted by species A_j on the joining and leaving reactions of A_i (curved arrow). The β matrix elements are drawn from a probability distribution generated through the Receptor Affinity Distribution model [36, 59].

in a global self-propagating behavior of the entire assembly, in a way reminiscent of autocatalysis. Under a certain set of constraints, this behavior resembles replication. Graded behavior enters in two different ways:

(i) Rate enhancement values related to mutual catalysis assume graded values;
(ii) Replication is graded both in its extent and its fidelity, because it is based on two graded variables, size and composition.

In its simplest embodiment, GARD is used to simulate the behavior of molecules that join and leave, and undergo only non-covalent reactions with other molecules [64, 65, 69]. In this embodiment no biosynthesis takes place, the only compounds present within the assembly are those supplied from the outside (complete heterotrophy). The free energy source that "fuels" such a "joining GARD" is the negative free energy gradient for lipid-like molecules transferred from an unfavorable aqueous medium to the inside of an amphiphilic micelle (Fig. 4). Mutual catalysis events occur on these downhill reactions, in ways analogous to catalyzed amphiphile flipping within membranes [13, 25] or to lipid-catalyzed ligand-receptor interactions [60]. It should be stressed, however, that the formalism explored here is not restricted to lipid micelles or vesicles, and is applicable also to other modes of molecular enclosures. One such example is aerosol droplets that constitute inverted micelles with aqueous enclosure [14, 15]. Unequal fission may be favored in such systems, a constraint which may be adapted by the GARD model (T. Shay and B. Shenhav, unpublished results).

GARD's time dependence. At time t, a GARD assembly contains a sub-repertoire of N_E molecule types out of a total repertoire of N_G different types. The GARD kinetic equations (Fig. 4) describe a process that resembles replicator dynamics [74]. Their solution signify the time-dependent trajectory of the composition vector $\mathbf{n}(t)$ in N_G dimensional space. The molecules present at a given time point are described by positive integer components in the compositional vector, while those absent are represented by zeros. This time trajectory is far from random, and depends on the mutual catalytic interaction parameters (defined by a matrix β_{ij}), on initial conditions and on statistical noise. The latter is related to the fact that small assemblies of discrete molecules are considered. This non-trivial behavior lasts as long as the assembly is away from equilibrium. At $t = \infty$ equilibrium sets in, and in a simplified case in which all molecules are thermodynamically equivalent, $\mathbf{n}(t)$ of the assembly will reflect the composition of the external milieu.

A much more interesting, non-trivial behavior, obtains when a GARD assembly is maintained in perpetuity away from equilibrium. This is achieved by adding a representation of an external free energy source to the simulation. Curiously, what is needed is *destructive* energy, i.e. the occasional disruption of the assembly [64]. One possible route for this is assembly splitting, e.g. fission. This simulated process would mimic natural ones, which occur when micelles grow beyond a certain limit. The relevant physical forces could be turbulence, temperature changes, surface tension and the like. The argument for the high probability of fission can be reformulated as a statement that under primitive earth conditions the chances of a molecular assembly to grow and never split are small.

Bequeathing compositional information. Rewardingly, the splitting assumption also serves to give the GARD model a capacity to emulate inheritance. If a perfect equal split occurs after perfect homeostatic growth, the two progeny assemblies will be identical. This might never happen in reality. Even in the incredibly orchestrated modern cell division, DNA miscopying and inequality of sharing of large molecules and organelles between progeny cells often occur. This imperfection may be deleterious in which case a daughter assembly might lose its mutually catalytic network and perish.

On the other hand, just as in the case of present-day cells, certain mutation-like compositional changes, resulting from imperfections of growth and fission, might actually be advantageous. For example, a more effective mutually catalytic network may be established through the fortuitous entry of a compound A_x not previously included in the assembly [68]. If this network is more efficient than the previous one, then it will now begin to acquire A_x from the environment, and the newly established composition will be propagated along many growth-split cycles. This amounts to a mutation being passed along to future generations. A similar instance of how one or very few molecules in a mutually catalytic set may determine the fate of an entire assembly, thus serving the rule of "genetic information" has been described [29]. The complex dynamics displayed by the GARD model (Fig. 5) is a result of this class of phenomena.

We have developed a simple formulation for assessing the success of progeny generation. A compositional similarity measure H, computed as a normalized scalar product of two compositional vectors, assists in this computation. It is possible to quantify accurately the similarity of two progeny assemblies to each other, and to the parent assembly. The dark squares in the autocorrelation diagram ("GARD carpet," Fig. 5), signify time periods during which, despite repeated growth and splitting events, compositional vectors for different time points remain clustered in N_G-dimensional space. In other words, the assembly undergoes consecutive periods of homeostatic growth.

Had different runs of the program (with different catalytic parameters) always resulted in just one large dark square, this would indicate success of assembly replication, but lack of any evolution-type change. In fact, only a few cases show such a "boring" behavior. In some of these the randomly sampled mutual catalysis matrix happens to have a single relatively large value on its diagonal, signifying that one of the N_G substances is a strong autocatalyst that dominates the dynamics of the GARD assembly. Yet, mesobiotic phenomena become significant when they involve ensemble complexity. Thus, a GARD composed of single autocatalyst must be regarded as degenerate.

Importantly, a majority of the simulations are *not* governed by a single autocatalyst, suggesting that under conditions of spontaneous formation, assemblies would tend to display mutually catalytic behavior. A typical such case is shown in Fig. 5. In this dynamics, every so often a homeostatic assembly gives way to a completely different one, which is also homeostatic. This is akin to multiple attractors, or set points, or stationary states, observed in other dynamic systems [3, 18, 26, 28, 74]. Because each of these states is characterized by a different molecular composition, we named them *composomes* [64]. The transitions between composomes likely arise from the accumulation of changes that occur due to imperfect mutual catalysis and imperfect fission. In other words, one may discern here a intriguing analogy to the way by which the accumulation of genomic mutations lead to speciation. In the

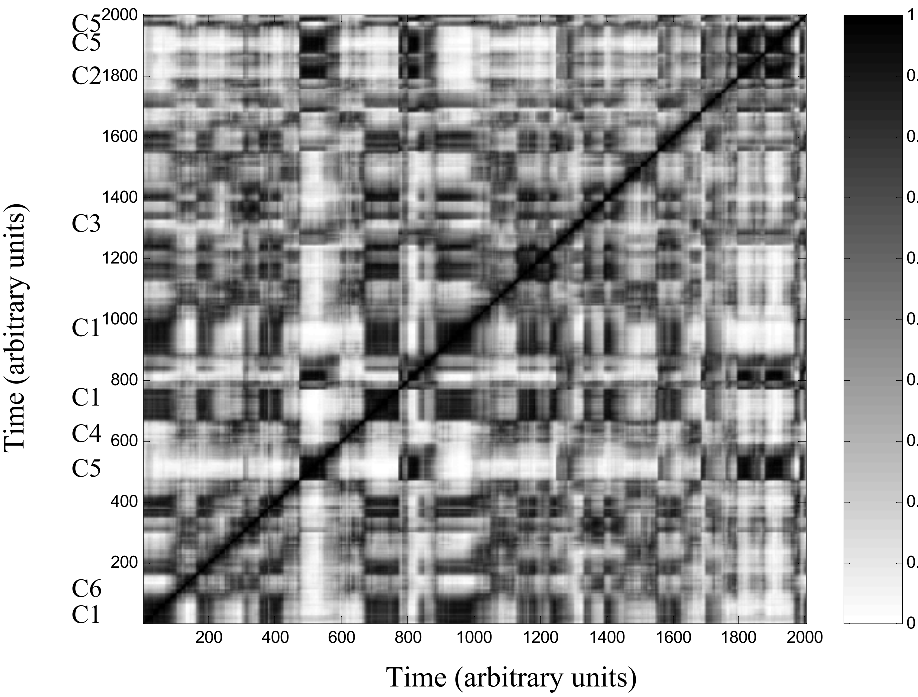

Fig. 5. (a) **GARD carpet.** "Compositional correlation carpet" of a GARD system, as previously described [64] with 100 different molecular species ($N_G = 100$). The drawing depicts a time correlation matrix, where both the ordinate and the abscissa represent the same time scale for the evolution of a particular GARD assembly for 20,000 time steps (about 2,000 growth split cycles). Each point in the two-dimensional graph is colored by its correlation measure, $H(\mathbf{n}_t, \mathbf{n}_{t'})$, for the compositions at times t and t' (cf. Fig. 2). The gray scale, signifying the range of H values, is shown on the right. The matrix displays dark rectangles, representing quasi-stationary compositions (composomes marked Ci on left) and abrupt phase transitions between them. The parameters of the simulation as described in [64] are $\Delta t = 0.03$, $k_f = 10^{-2}$, $k_b = 10^{-4}$, $\mu = -4$, $\sigma = 4$, $n_{\mathrm{TOT}} = 1000$. The simulation was conducted on a Pentium III 800MHz processor using MATLAB version 5.3 (Mathworks, Natick, MA), using a stochastic chemistry algorithm of Ref. 64 which is elaborated in Ref. 22.

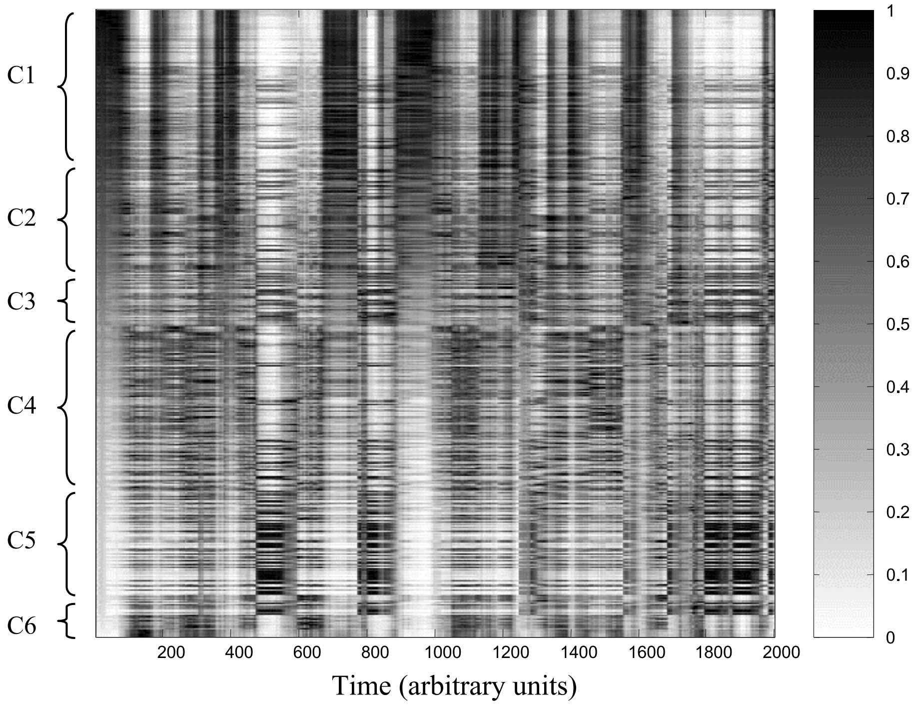

Fig. 5. (b) The same as Fig. 5(a), with rows sorted according to their value at $t = 20$. This affords a more facile visualization of the different composomes, as indicated on the left. This sorting was also used to label the composomes from 1 to 6 in Fig. 5(a). It should be noted that the main division is to three composomal types, C1, C5, C6, with the other three being respectively related to them ($C_2 \approx C_1$, $C_3 \approx C_5$ and $C_4 \approx C_6$). A thorough analysis of such relationships is underway (B. Shenhav, R. Kafri and D. Lancet, submitted).

spirit of the continuity principle, it should be also noted that the persistence of composomes associated with faster growing sets of mutally catalytic molecules has an interesting correspondence in the optimal growth performance observed in some present-day unicellular organisms [70].

Compositional genomes. In a GARD assembly compositional information plays a role analogous to that of sequence information in present day DNA- or RNA-based genomes. In these mesobiotic entities, the same molecules whose counts constitute a very primitive "compositional genome," also play the role of constituent building blocks, members of metabolic pathways and catalysts that make these reactions more efficient. A conceptually similar reasoning has been explored by invoking the control by a molecular minority in a mutually catalytic replicating system [29]. We believe that division of labor between genome-specific molecules, catalysts and compartment building blocks is the hallmark of modern (biotic) life. The riddle of how such function-specific molecules have emerged, and why they are so strongly inter-dependent, may be addressed by a scenario in which, early on, each molecular constituent contributed to some extent to all the above mentioned functions.

A seminal thought experiment related to the function of GARD assemblies as entities that harbor a compositional genome is as follows. Consider a highly effective GARD, i.e. one that possesses a well-connected catalytic network for catalyzed joining. The composition of such an assembly may be predicted using molecular dynamics algorithms, or by measurements of amphiphile dynamics and pairwise interactions in lipid bilayers. This micellar assembly is embedded in a medium containing a large variety of amphiphiles, including all those present in the "seed" assembly, and numerous others. Our simulations show that the assembly will grow by absorbing mainly the types of molecules already present within it. Furthermore, if agitation is employed, the growing assembly undergoes occasional fission, and the entire reaction volume will gradually be "infected" with more or less similar copies of the original assembly. Importantly, it is expected that the molecular composition of the assemblies will be different from that of the external environment. All of these attributes are highly reminiscent of the behavior of present-day free-living microorganisms.

7. Conclusion: Life as an Emergent Phenomenon

The true challenge of origin of life studies is to regard life as an emergent phenomenon. One should not assume what is sought, but allow the laws of nature to lead from what was very likely present in a prebiotic molecular mixture, to what is minimally needed for self-replication to take place. As dictated by Occam's razor, high levels of complexity of the first kind (covalent complexity) should be avoided, if elements of much lower complexity of the second kind (ensemble complexity) could serve as substitute.

The GARD model provides a route for simulated sequential improvement and consecutive addition of new model features. Such graded changes should be capable of delineating a smooth transition from a simple, ineffective assembly, capable only of non-covalent exchange with the environment, to more and more complex entities. The latter could gradually incorporate capacities such as making longer and longer oligomers [71], processes akin to polymer folding and conformational stability, increasingly better catalysis and so on. Hence, the transfer of complexity from second to first kind could constitute a novel way of explaining the emergence of biopolymers. Regarding compartmentalization, we plan to implement simulated parameters that would lead from simple micelles to lipid vesicles with an aqueous core, including passive and active transport. The simulation can also be made to include free energy input from chromophores that harvest light energy. All of this diversity of function comes naturally to a simulated heterogeneous molecular assembly, but not to simulated naked RNA replicators.

Time will tell whether the origin scenario delineated here is nearer to the true sequence of events than alternative ones. But using simple rules of biophysical chemistry, as well the increasing power of digital computers [72], it should be possible soon to find out how *plausible* this scenario is. The mutual catalysis mechanisms lend themselves to detailed scrutiny in terms of chemical kinetics simulations just as well as some of the other scenarios [21, 54, 58, 83]. They also provide a platform for computing probabilities and time scales at a planetary level. This should allow students of complexity and emergence to contribute in new ways to the tackling of one of the most important open questions in science.

Acknowledgments

We thank Drs. Avshalom Elitzur and Emanuel Lottem for critical comments on the manuscript.

References

[1] Bachmann, P., Luisi, P. and Lang, J., *Nature* **357**, 57 (1992).
[2] Bagley, R. J. and Farmer, J. D., in *Artificial Life II*, eds. Langton, C. G., Taylor, C., Farmer, J. D. and Rasmussen, S. (Addison-Wesley, 1991).
[3] Bagley, R. J. and Glass, L., *J. Theor. Biol.* **183**(3), 269 (1996).
[4] Barbier, B. and Brack, A., *Orig. Life Evol. Biosph.* **17**(3–4), 381 (1987).
[5] Bedau, M. A., McCaskill, J. S., Packard, N. H., Rasmussen, S., Adami, C., Green, D. G., Ikegami, T., Kaneko, K. and Ray, T. S., *Artif. Life* **6**(4), 363 (2000).
[6] Blomberg, C., *J. Theor. Biol.* **187**(4), 541 (1997).
[7] Bohler, C., Hill, A. R. Jr. and Orgel, L. E., *Orig. Life Evol. Biosph.* **26**(1), 1 (1996).
[8] Cairns-Smith, G., *Genetic Takeover and the Mineral Origins of Life* (Cambridge University Press, Cambridge, UK, 1982).
[9] Ceska, M., *Experientia* **28**(2), 146 (1972).
[10] De Duve, C., *Vital Dust*: *Life As a Cosmic Imperative* (Basic Books, New York, 1995).
[11] Deamer, D. W. and Oro, J., *BioSystems* **12**, 167 (1980).
[12] Deamer, D. W. and Pashley, R. M., *Orig. Life Evol. Biosph.* **19**(1), 21 (1989).
[13] Devaux, P. F., *Annu. Rev. Biophys. Biomol. Struct.* **21**, 417 (1992).
[14] Dobson, C. M., Ellison, G. B., Tuck, A. F. and Vaida, V., *PNAS* **97**(22), 11864 (2000).
[15] Donaldson, D. J., Tuck, A. F. and Vaida, V., *Orig. Life Evol. Biosph.* **32**(3), 237 (2002).
[16] Dyson, F., *Origins of Life* (Cambridge University Press, Cambridge, 1999).
[17] Eschenmoser, A., *Science* **284**(5423), 2118 (1999).
[18] Farmer, J. D., Kauffman, S. A. and Packard, N. H., *Physica* **22D**, 50 (1986).
[19] Fendler, H. J. and Fendler, E. J., *Catalysis in Micellar and Macromolecular Systems* (Academic Press, New York, 1975).
[20] Fontana, W. and Buss, L. W., *Proc. Natl. Acad. Sci. USA* **91**(2), 757 (1994).
[21] Fontana, W. and Schuster, P., *Science* **280**(5368), 1451 (1998).
[22] Gillespie, D. T., *J. Chem. Phys.* **115**(4), 1716 (2001).
[23] Goebel, W. and Kuhn, M., *Curr. Opin. Microbiol.* **3**(1), 49 (2000).
[24] Groot, R. D. and Rabone, K. L., *Biophys. J.* **81**(2), 725 (2001).
[25] Higgins, C. F. and Gottesman, M. M., *Trends. Biochem. Sci.* **17**(1), 18 (1992).
[26] Hogeweg, P., in *Artif Life*, eds. Adami, C., Belew, R. K., Kitano, H. and Taylor, C. E. (MIT Press, Cambridge, 1998).
[27] Hüber, C. and Wächtershäuser, G., *Science* **276**(5310), 245 (1997).
[28] Jain, S. and Krishna, S., *Proc. Natl. Acad. Sci. USA* **99**(4), 2055 (2002).

[29] Kaneko, K. and Yomo, T., *J. Theor. Biol.* **214**, 563 (2002).
[30] Kauffman, S. A., *J. Theor. Biol.* **119**, 1 (1986).
[31] Kauffman, S. A., in *The Origins of Order; Self-Organization and Selection in Evolution* (Oxford University Press, New York, 1993).
[32] Kenworthy, A., *Trends Biochem. Sci.* **27**(9), 435 (2002).
[33] Kinne-Saffran, E. and Kinne, R. K., *Am. J. Nephrol.* **19**(2), 290 (1999).
[34] Kitano, H., *Nature* **420**(6912), 206 (2002).
[35] Kochavi, E., Bar-Nun, A. and Fleminger, G., *J. Mol. Evol.* **45**, 342 (1997).
[36] Lancet, D., Sadovsky, E. and Seidemann, E., *Proc. Natl. Acad. Sci. USA* **90**(8), 3715 (1993).
[37] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K. *et al.*, *Nature* **409**(6822), 860 (2001).
[38] Lazcano, A. and Bada, J. L., *Science* **298**(5594), 749 (2002).
[39] Lazcano, A. and Miller, S. L., *Cell* **85**(6), 793 (1996).
[40] Lehn, J. M., *Science* **295**(5564), 2400 (2002).
[41] Lemieux, R. U. and Spohr, U., *Adv. Carbohydr. Chem. Biochem.* **50**, 1 (1994).
[42] Lindahl, E. and Edholm, O., *Biophys. J.* **79**(1), 426 (2000).
[43] Luisi, P. L., *Orig. Life Evol. Biosph.* **28**(4–6), 613 (1998).
[44] Luisi, P. L., Walde, P. and Oberholzer, T., *Curr. Opin. Colloid & Inter. Sci.* **4**(1), 33 (1999).
[45] McDonald, G. D., Whited, L. J., DeRuiter, C., Khare, B. N., Patnaik, A. and Sagan, C., *Icarus* **122**(1), 107 (1996).
[46] McMullin, B., *Annals of the New York Academy of Sciences* **901**, 163 (2000).
[47] Miller, S. L., Urey, H. C. and Oro, J., *J. Mol. Evol.* **9**(1), 59 (1976).
[48] Minard, R. D., Hatcher, P. G. and Gourley, R. C., *Orig. Life Evol. Biosph.* **28**(4–6), 461 (1998).
[49] Morowitz, H. J., in *Progress in Theoretical Biology*, Snell, F. M. ed. (Academic Press, 1967).
[50] Morowitz, H. J., Santa Fe Institute, 96-04-014 (1996).
[51] Morowitz, H. J., Heinz, B. and Deamer, D. W., *Orig. Life Evol. Biosph.* **18**(3), 281 (1988).
[52] Morowitz, H. J., Kostelnik, J. D., Yang, J. and Cody, G. D., *Proc. Natl. Acad. Sci. USA* **97**(14), 7704 (2000).
[53] Mouritsen, O. G. and Jorgensen, K., *Bioessays* **14**(2), 129 (1992).
[54] Nir, S. and Lahav, N., *Orig. Life Evol. Biosph.* **27**(5–6), 567 (1997).
[55] Oparin, A. I., *The Origin of Life*, 2nd edn. (Dover Pub., New York, 1953).
[56] Oparin, A. I., in *The Origin of Life*, Bernal, J. D. ed. (Weidenfeld and Nicolson, London, 1967).
[57] Pizzarello, S., Huang, Y., Becker, L., Poreda, R. J., Nieman, R. A., Cooper, G. and Williams, M., *Science* **293**(5538), 2236 (2001).
[58] Reidys, C., Forst, C. V. and Schuster, P., *Bull. Math. Biol.* **63**(1), 57 (2001).
[59] Rosenwald, S., Kafri, R. and Lancet, D., *J. Theor. Biol.* **216**(3), 327 (2002).
[60] Sargent, D. F. and Schwyzer, R., *Proc. Natl. Acad. Sci. USA* **83**(16), 5774 (1986).
[61] Saxena, A. and Lookman, T., *Physica* **D133**(1–4), 416 (1999).
[62] Scarso, A., Scheffer, U., Gobel, M., Broxterman, Q. B., Kaptein, B., Formaggio, F., Toniolo, C. and Scrimin, P., *Proc. Natl. Acad. Sci. USA* **99**(8), 5144 (2002).
[63] Segré, D., Ben-Eli, D., Deamer, D. W. and Lancet, D., *Orig. Life Evol. Biosph.* **31**(1–2), 119 (2001).

[64] Segré, D., Ben-Eli, D. and Lancet, D., *Proc. Natl. Acad. Sci. USA* **97**(8), 4112 (2000).
[65] Segré, D. and Lancet, D., *EMBO Rep.* **1**(3), 217 (2000).
[66] Segré, D. and Lancet, D., *Chemtracts — Biochemistry and Molecular Biology* **12**(6), 382 (1999).
[67] Segré, D., Lancet, D., Kedem, O. and Pilpel, Y., *Origins Life Evol. Biosphere* **28**, 501 (1998).
[68] Segré, D., Pilpel, Y. and Lancet, D., *Physica* **A249**(1–4), 558 (1998).
[69] Segré, D., Shenhav, B., Kafri, R. and Lancet, D., *J. Theor. Biol.* **213**(3), 481 (2001).
[70] Segré, D., Vitkup, D. and Church, G. M., *Proc. Natl. Acad. Sci. USA* (2002).
[71] Shenhav, B., Kafri, R. and Lancet, D., *Origins Life Evol. Biosphere* (in press).
[72] Shenhav, B. and Lancet, D., *Orig. Life Evol. Biosph.* (in press).
[73] Suwannachot, Y. and Rode, B. M., *Orig. Life Evol. Biosph.* **29**(5), 463 (1999).
[74] Szathmáry, E., *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **355**(1403), 1669 (2000).
[75] Takahashi, K., *Chem. Rev.* **98**(5), 2013 (1998).
[76] Tsutsumi, H., Hamasaki, K., Mihara, H. and Ueno, A., *Bioorg. Med. Chem. Lett.* **10**(8), 741 (2000).
[77] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. Q. H., Chen, L., Skupski, M. *et al.*, *Science* **291**(5507), 1304 (2001).
[78] Wächtershäuser, G., *Microbiol. Rev.* 452 (1988).
[79] Wächtershäuser, G., *Proc. Natl. Acad. Sci. USA* **87**(1), 200 (1990).
[80] Weiss, O., Jimenez-Montano, M. A. and Herzel, H., *J. Theor. Biol.* **206**(3), 379 (2000).
[81] Wilkins, in *2-D Proteome Analysis Protocols*, Link, A. J. ed. (Humana Press, New Jersey, 1998).
[82] Wills, C. and Bada, J., *The Spark of Life*: *Darwin and the Primeval Soup* (Perseus Publishing, 2001).
[83] Ycas, M., *Orig. Life Evol. Biosph.* **29**(1), 95 (1999).