Astronomical Society of the Pacific Conference Series
6th Bioastronomy meeting, Kohala Coast, Hawaii, August 1999
(in press)

Prebiotic Evolution Of Amphiphilic Assemblies Far From Equilibrium: From Compositional Information To Sequence-Based Biopolymers

Daniel Segré, Dafna Ben-Eli and Doron Lancet

Dept. of Molecular Genetics and the Crown Human Genome Center, The Weizmann Institute of Science, 76100 Rehovot, Israel.
E-mail: bmsegre@wicc.weizmann.ac.il.

Abstract. The primordial emergence of biopolymers, agents of the genetic machinery in modern cells, is not less enigmatic than the emergence of the genetic code itself. Here we discuss how potential early replicating protocellular systems based on a rudimentary form of inheritance, a "compositional genome", could evolve towards the emergence of "alphabetic" polymers, predating the genetic code. A computer simulated evolutionary process based on our previously proposed kinetic model may help understand the appearance of chemical combinatorics through early natural selection.

1. Introduction

Biological information has often been identified with the "digital" genetic machinery: four letters for the language of the genes, twenty for the language of proteins. Considerable effort has been invested in trying to trace the history of the genetic code (Trifonov and Bettecken 1997) and to study its formal properties (Niesert-Struwe and Wills 1997) and its optimization (Knight et al. 1999, Di Giulio and Medugno 1999). Yet, much less analysis has been centered on the emergence and the evolution of the combinatorial system underlying it[1]. A necessary condition for the rise of molecular combinatorics in biopolymers is the availability of their monomeric constituents. Increasing evidence is being collected that the prebiotic chemical inventory on early Earth comprised thousands of different organic compounds (Bernstein et al. 1999, Irvine 1998). This repertoire very likely included many of the organic compounds constitut-

[1]Interestingly, a similar twofold problem is faced in the study of the emergence of human language: besides investigating the evolution of words and grammars in different languages, one might ask, at a more elementary level, what could be the fundamental factors inducing, within a primitive society, a transition from a communication based on isolated simple sounds to one involving the combination of such sounds into words (Nowak and Krakauer 1999).

ing the fundamental building blocks for present day cellular life. But no clear explanations seem to have been offered for how the amino acid and nucleic acid monomers abiotically segregated from the original complex mixtures and interacted only among themselves to form higher level oligomers. The emergence of an RNA world, for example, would require a relatively high localized abundance of the four RNA building blocks, and the absence of any other monomer that might interfere with "smooth" RNA polymerization (Shapiro 1999).

A possible alternative to an abiotic segregation process is one based on natural selection (cf. Lifson and Lifson 1999). Based on former notions (Oparin 1957, Dyson 1985, Morowitz 1992, Bachmann et al. 1992), our previously explored Amphiphilic Graded Autocatalysis Replication Model (A-GARD, cf. Segré et al. 1999b) analyzes the kinetics of protocellular amphiphilic aggregates and their self-reproduction behavior in absence of biopolymers. Such aggregates could form spontaneously under prebiotic condition, due to hydrophobic forces, giving rise to micelles and vesicles (Deamer 1997, Luisi et al. 1999). In A-GARD, a probabilistic formalism (Lancet et al. 1993, 1994) similar to that involved in combinatorial chemistry is used to describe the catalytic interactions within random collections of prebiotic lipid-like molecules, as suggested also by a statistical reexamination of experimental values for membrane mimetic compounds (Fendler 1982). The model is analyzed through computer simulations with stochastic chemical kinetics rules (Segré et al. 1999b). This results in homeostatic preservation of the molecular composition and assembly growth. With one additional assumption, whereby an amphiphilic assembly may split and generate two distinct new aggregates, it is possible to observe the propagation of compositional information to daughter micelles. In other words information is stored and propagated as a "compositional genome" (Segré and Lancet 1999a), rather than in the sequences of biopolymers. A small fraction of all the possible compositions behave as metastable states for the system, due to the efficient catalytic networks connecting their components. The reduced number of molecular types within such compositional states (or "composomes" (Segré et al. 1999b)) increases the probability of high parent-offspring similarity after division (Morowitz 1967).

In the present work, based on the results summarized above, we explore a possible stepwise mechanism for the appearance of the first polymers within a system of compositional assemblies, and for the emergence of sequence-based information, encoded in oligomeric chemical "words".

2. From compositions to sequences

If biopolymers are the outcome of natural selection, it should be possible to identify their evolutionary advantages, within compositional assemblies, at the earliest steps of emergence. One potential advantage of longer molecular chains over single monomers is the greater chance of significant catalytic activity. As in a modern combinatorial chemistry experiment, the large number of possible covalent associations of different units induces a large variety of folding shapes, binding pockets, catalytic activities (MacBeath et al. 1998).

Another more subtle advantage is related to the splitting of a micelle or vesicle. In contrast to the situation depicted in Fig.1A, in which the division

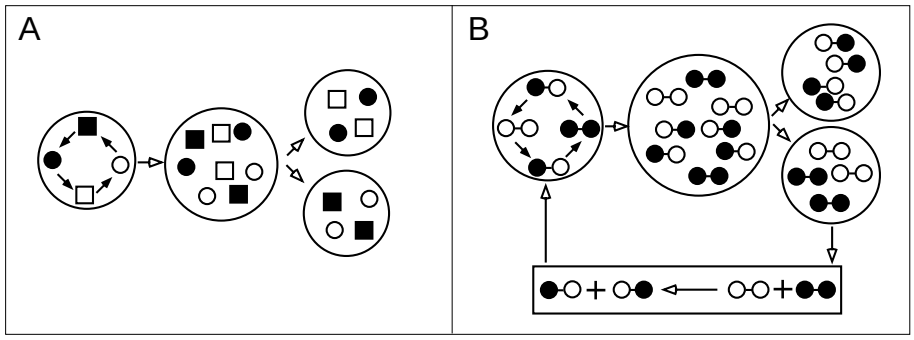

Figure 1. Molecular aggregates that grow and divide. In the two examples illustrated here, the division processes fail to generate offspring similar to the parent assemblies. The arrows within the circles represent catalytic enhancement effects. **A** For monomers the situation is really unrecoverable. **B** If the molecules are dimers formed by two different monomers covalently linked, then a rearrangement of the monomers might still generate an assembly similar to the original one.

fails to generate an offspring identical to the parent assembly, consider a case in which the molecules involved in the catalytic network are dimers formed by combining in all possible ways two different monomers (Fig.1B). Even after a "bad splitting", rearrangement of the monomers could bring an assembly back to its original successful dimer composition. In general, in a compositional assembly, the initial necessary condition of keeping high the count of species that are members of the catalytic network, in order to guarantee successful splittings, might be gradually substituted by more sophisticated mechanisms of rearrangements of molecular building blocks. Further insight on these processes may be gained through computer simulations, as delineated below.

3. Emergence of molecular alphabets in a computer model

We performed computer simulations for a modified A-GARD system, in which the monomers within an assembly can bind covalently to each other to form oligomers. The Dimer GARD (Segré et al. 1998b) is a special case with oligomers not longer than 2. The present approach to a polymeric A-GARD involves an assembly whose composition is allowed to move on a defined fitness landscape (cf. Segré et al. 1998a). The elementary compositional changes can be of two different kinds: (a) monomers recruitment/leaving, (b) polymerization, i.e. joining of any two species (polymers or monomers) to form a new oligomeric molecule. In the simulations, the tendency of a composition to maximize the fitness is opposed by the externally imposed break up of polymers at an arbitrary bond, with a fixed rate. The fitness function is defined as follows:

$$f = \sum_{\sigma,\tau} \rho_\sigma \beta_{\sigma\tau} n_\sigma n_\tau \qquad (1)$$

where ρ_σ weights the contribution of each catalyzed polymer with the fraction of its monomeric constituents actually available within the assembly and n_σ is

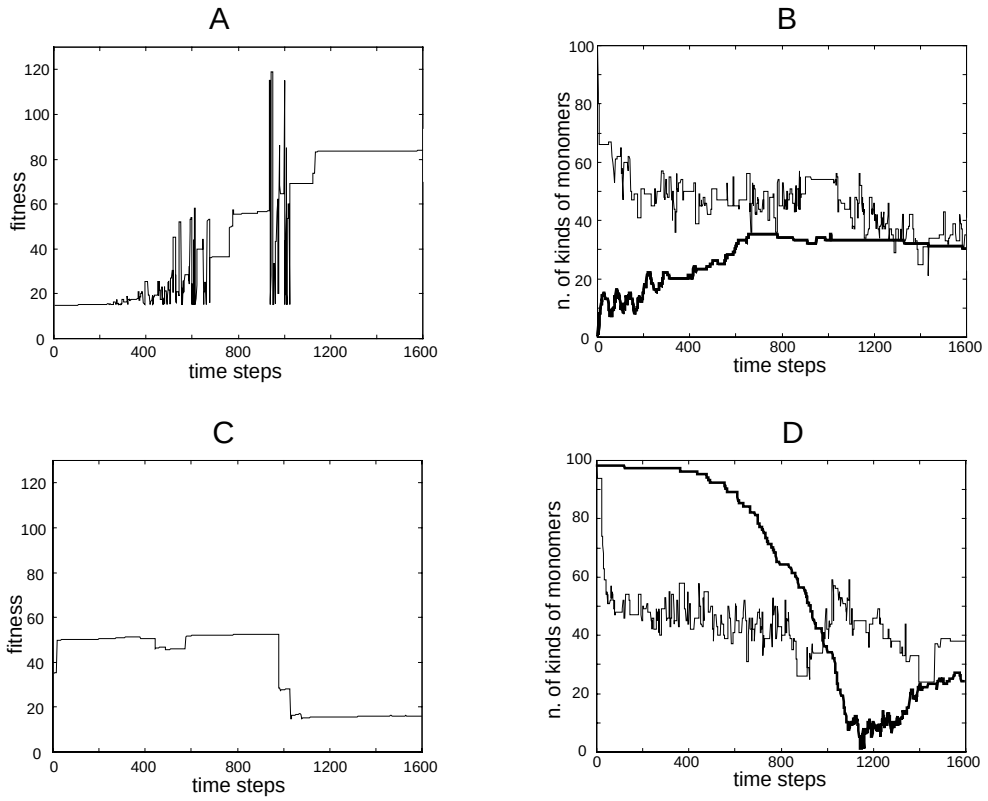

Figure 2. Computer simulated evolutionary process for a polymeric A-GARD system with $N_G = 100$ (see text). Initial conditions are: only monomers in **A** and **B**; 10 random polymers of length 30 in **C** and **D**. **A** and **C** show the time dependence of the fitness function. **B** and **D** depict the time course of the number of different kinds of monomer species, as isolated units in the aggregate (thin line), and as components of polymers (thick line).

the number of polymers with a given sequence σ. A sequence is represented by a string $\sigma = (s_1, s_2, \ldots, s_{L(\sigma)})$ of numbers $s_i \in \{1, \ldots, N_G\}$. The N_G sequences with L=1 are the monomeric building blocks, whose amounts constituted the **n** vector in previous monomeric A-GARD embodiments (Segré et al.1999b). The matrix element $\beta_{\sigma\tau}$ represents the degree of catalysis exerted by a polymer with sequence τ on a polymer with sequence σ. In the kinetic version of A-GARD these values would multiply the rates of the joining/leaving processes (Segré et al. 1999b). The $\beta_{\sigma\tau}$ values for monomers are sampled, at the beginning of a simulation, from a probability distribution $\Phi(\beta_{\sigma\tau})$ of catalytic rate enhancement factors (Segré et al. 1999b), based on the statistics of molecular recognition (Lancet et al. 1993, 1994). An analogous lookup table for the catalysis among polymers up to length L_{max}, would require dealing with approximately $N_G^{2(L_{max}+1)}$ numbers. It is therefore preferable to use an algorithm that computes the $\beta_{\sigma\tau}$ values only for the polymers actually present in the assembly. The polymer to polymer rate enhancement algorithm is based on a string matching rule, whereby the value of $\beta_{\sigma\tau}$ is calculated as the product of the catalytic factors among all the pairs of monomers facing each other (Lancet et al. 1994, cf. Bagley and Farmer 1991).

The fitness function is designed for rewarding assemblies which are homeostatic (i.e. such that they tend to conserve their composition upon growth). The fitness function is therefore defined as the scalar product between the compositional change and the current composition. The absence of normalization in the definition of f (Eq. 1) implies that, for assemblies with the same degree of homeostasis, a higher value will be found for a faster growing assembly.

Some results obtained with the algorithms described above are shown in Fig.2. In the first experiment (Figs.2A,B), an assembly was randomly seeded only with monomers of $N_G = 100$ different kinds. A rough plateau was reached when approximately 40 monomers types were selected (Fig.2B, thin line). This "reduction of possibilities" is compatible with analogous results previously observed in kinetic simulations (Segré and Lancet 1998c, Segré et al. 1999b). Polymers growing in the assembly would tend to utilize all the available monomeric repertoire; however, the reduction of possibilities which takes place simultaneously, limits the number of accessible monomers species. Hence, at the plateau, the size of the molecular alphabet of the polymers (Fig.2B, thick line) is approximately equal to the number of available isolated monomers kinds. Note that the fitness function displays regions of abrupt jumps, and long plateaus (Fig.2A). Jumps occur when new highly catalytically efficient polymers form. Initially, the small number of copies of such polymers makes their presence very vulnerable to decomposition. Only when a large amount of efficient polymers is reached, the fitness function becomes more stable. In the second experiment (Figs.2C,D), the assembly was seeded with 10 random polymers of length 30. Despite the larger catalytic effects among them, these polymers, that were not obtained through evolution, have no potential for further fitness improvement, and gradually decompose until only monomers remain.

4. Discussion

The results presented here constitute an initial attempt of uncovering possible paths for the emergence of biopolymers, under the hypothesis that a primitive

form of life, based on compositional information transfer, preceded biopolymers. Some ideas and results stemming from this preliminary analysis might be rather general, and pave the road for more rigorous investigations. Experimental tests of the proposed scenario may be possible in the near future, thanks to newly arisen interest in polymerization phenomena on monolayer lipid surfaces (Oliver and Singh 1997).

In the simulated stepwise evolutionary process, the reduction of the number of monomer types in an assembly, observed also in previous A-GARD simulations (Segré et al. 1999b), is coupled to the rise of polymers. This emergence of polymers is envisaged here as a consequence of evolutionary pressure that calls for more efficient members in an already established catalytic network. This is quite different from other scenarios, in which polymers are assumed to be present from the beginning, and replicators have a certain probability to spontaneously emerge. It may be possible to test these two scenarios by comparing the characteristics of sequences obtained in different simulated processes. Sequences that evolve during a reduction of possibilities period may develop multilevel combinatorial structures (cf. Karlin and Brendel 1993), as observed in preliminary analyses. If, for example, a short sequence happened to be catalytically successful at an early stage of the evolutionary process, it will likely be found incorporated in longer sequences at later stages. Among such complex segments, only some will survive further, giving rise to a hierarchy of "words". Future research could investigate methods for comparing such structures with the patterns observed in real ancient nucleotide or peptide sequences.

Acknowledgments. Thanks to Drs. Ora Kedem, Shneior Lifson and Yitzhak Pilpel for stimulating discussions, and to Dr.Helmut Zepik and Yoav Gilad for critical reading of the manuscript.

References

Bachmann, P., Luisi, P. L. and Lang, J. (1992) Nature 357, 57-59.

Bagley, R. J., Farmer, J.D. 1991, In Artificial Life II, Langton, C. G. et al. , ed. Addison-Wesley, pp 93-140.

Bernstein, M. P., Sandford, S. A., Allamandola, L. J. et al. 1999, Science 283, 1135-1138.

Deamer, D. W. 1997, Microbiol. Molecular Biology Reviews 61(2), 239-261.

Di Giulio, M. and Medugno, M. 1999, J. Mol. Evol. 49(1), 1-10.

Dyson, F. 1985, Origins of Life; Cambridge University Press: Cambridge.

Fendler, J. H. 1982, Membrane Mimetic chemistry, Wiley.

Irvine, W. M. 1998, Orig. Life Evol. Biosph. 28, 365-383.

Karlin, S. and Brendel V. 1993, Science 259, 677-679.

Knight, R. D., Freeland, S. J. and Landweber, L. F. 1999, Trends Biochem. Sci. 24(6), 241-247.

Lancet, D., Sadovsky, E. and Seidemann, E. 1993, Proc. Natl. Acad. USA. 90, 3715-3719.

Lancet, D., Kedem, O. and Pilpel, Y. 1994, Ber Bunsenges. Phys. Chem. 98, 1166-1169.

Lifson, S. and Lifson H. 1999, J. Theor. Biol. 199, 425-433.

Luisi, P.L., Walde, P. and Oberholzer, T. 1999, Current Opinion in Colloid & Interface Sci., 4, 33.

MacBeath G., Kast, P., Hilvert, D. 1998, Protein Sci. 7(2), 325-335.

Morowitz, H.J. 1967, In Progress in Theor. Biol., Snell, F.M., ed. Academic Press, pp 35-58.

Morowitz, H. J. 1992, Beginnings of cellular life, Yale University Press, London.

Niesert-Struwe, K. and Wills, P. R. 1997, J. theor. Biol. 187, 1-14.

Nowak, M. and Krakauer, D. C. 1999, Proc. Natl. Acad. USA. 96, 8028-8033.

Oliver, J. S. and Singh, J. 1997, J. Org. Chem., 62, 6436-6438.

Oparin, A. I. 1957, The origin of life on the earth, Oliver and Boyd, London.

Segré, D., Pilpel, Y. and Lancet, D. 1998a, Physica A 249, 558-564.

Segré, D., Lancet, D., Kedem, O. and Pilpel, Y. 1998b, Orig. Life Evol. Biosph. 28, 501-514.

Segré, D. and Lancet, D. 1998c, in Exobiology, Chela-Flores, J. and Raulin F., eds., Kluwer, The Netherlands, pp. 123-131.

Segré, D. and Lancet, D. 1999a, Chemtracts - Biochemistry and Molecular Biology 12, 382-397.

Segré, D., Ben-Eli, D. and Lancet, D. 1999b, submitted.

Shapiro, R. 1999, Proc. Natl. Acad. USA. 96, 4396-4401.

Trifonov, E. N. and Bettecken, T. 1997, Gene, 205, 1-6.