Self-organization in macromolecular systems: The notion of adaptive value

(prebiotic evolution/statistical mechanics)

LLOYD DEMETRIUS
Max-Planck-Institut für Biophysikalische Chemie, 3400 Göttingen, Federal Republic of Germany

Communicated by P. W. Anderson, May 21, 1984

ABSTRACT Self-organization in macromolecular systems refers to the transition from a random assembly of interacting oligomers to a system of stable heteropolymers. The concept of adaptive value describes the correlation between environmental variability and the variability in replication and mutation rates of the interacting oligomers. This paper describes a model of self-organization in macromolecular systems based on the concept of adaptive value. The equilibrium states of a set of interacting polymers are described by states that maximize the adaptive value. The analytic basis for this notion of equilibrium, which is called the adaptive value principle, is given and this principle is invoked to explain two examples of macromolecular self-assembly.

The problem of self-organization in macromolecular systems centers around an understanding of the transition from a random assembly of oligomers to a stable self-replicating entity. Two examples of macromolecular assembly that have been investigated intensively are the experiments on template-free synthesis of macromolecules reported in refs. 1 and 2. Kornberg et al. (1) in their study of the DNA polymerase system have observed that a few small macromolecules can serve as template primers to produce long sequences of homopolymers. Moreover, although the nature and structure of these sequences vary with slight changes in the environmental conditions, sequences of similar length and structures are produced when the experiments are performed under the same external conditions. Biebricher et al. (2) in their experiments have observed that non-repetitive RNA sequences are produced in the absence of template. Moreover, when the experiments are carried out under the same environmental conditions, sequences of different lengths and different electrophoretic patterns are observed.

The DNA polymerase and $Q_{2}$ replicase are products of gene transcription by means of ribosomes and transfer RNA aided by several enzymes. Hence, both systems are much too complex to serve as a model for prebiotic evolution. These processes, however, share with prebiotic systems the property of a transition from a random assembly of oligomers to a stable array of heteropolymers. Hence, a quantitative understanding of these two systems can provide a basis for a theory of prebiotic evolution.

In this paper the basis for such a theory is proposed by giving a mathematical model of the process of self-organization, which incorporates the essential features of the two examples of macromolecular assembly described.

The mechanisms of self-organization invoked in this paper are those of random drift and natural selection, ideas whose analytic formulation in the context of population biology goes back to the works of Wright (3) and Fisher (4). The theory developed by these authors neglected the effect of environmental factors in the fecundity–mortality schedule of a population. The central concept that emerged from the work of these authors is the Malthusian parameter. This describes the intrinsic rate of increase of the population under constant environmental conditions. One of the central principles in the theory described in refs. 3 and 4 is summarized by the fundamental theorem of natural selection (4). This states that under constant environmental conditions, natural selection tends to increase the mean Malthusian parameter. The model described in this paper, in contrast to the work in refs. 3 and 4, takes into account the effect of environmental factors as an organizing force in natural selection. The notion of adaptive value is used to characterize the effect of random environmental factors on macromolecular evolution. This notion was introduced in a population biology context in Demetrius (5, 6). In the context of macromolecular evolution, the concept of “adaptive value” describes the correlation between the variability in the environment and the variability in the replication and mutation rates of the macromolecules. The model of self-organization proposed in this paper is based on the principle described in ref. 6, that in a population in interaction with environment, natural selection tends to increase the mean adaptive value. This principle is a generalization of the Wright–Fisher principle (4), which has been shown to be valid for constant environments.

The model exploits the connections between statistical mechanics and population biology described in Demetrius (7). In ref. 7 precise analogies between thermodynamic variables and the macroscopic parameters in population biology were established. Here it was shown that the free energy corresponds to the Malthusian parameter and the temperature is analogous to the reciprocal of the generation time of the population. In addition, it was shown that the fundamental principle of equilibrium statistical mechanics, which states that equilibrium structures are characterized by minimization of the free energy, is equivalent to the fundamental principle of natural selection, which states that equilibrium states are given by maximizing the Malthusian parameter. These two principles are valid for systems that are not in interaction with the environment. The new concept of adaptive value allows us to generalize the fundamental principle of natural selection in order to account for the effect of environmental factors on population dynamics. I postulate a second type of structure organized not as a consequence of maximizing the Malthusian parameter but as a result of environmental noise, which forces the system into a new organized state characterized by maximizing the adaptive value.

I refer to ref. 7 for the connection between population biology, statistical mechanics, and information theory, which forms the mathematical basis of this work. The concepts in information theory invoked are given in Billingsley (8). The works of Ruelle (9) and Dobrushin (10) are good sources for the statistical mechanics ideas used in this article.

I. A Model of Macromolecular Organization

Consider a population consisting of a family of oligomers. Each oligomer consists of a sequence of nucleotides from the...
Evolution: Demetrius

set A, G, C, U. I assume throughout that I am dealing with an RNA system. In the case of the DNA system, the nucleotide U is replaced by T. Species of oligomers are generated by the processes of slippage, elongation, point mutations, and replication. The generic term aggregation is used to describe the first three processes.

The processes are of the form:

\[ A \cdot G \cdot C \rightarrow A \cdot G \cdot C \cdot U \]

\[ (A \cdot G \cdot C \cdot U) + (G \cdot C \cdot A) \rightarrow A \cdot G \cdot C \cdot U \cdot G \cdot C \cdot A \]

\[ A \cdot G \cdot C \rightarrow A \cdot U \cdot C \]

Let \( S \) denote the set of different types of oligomers that can be generated by these processes. Each oligomer is made of elements from the nucleotides A, G, C, U. Assume that there are \( m \) distinct oligomers, which are denoted by \( I_1, I_2, \ldots, I_m \).

The generation of new oligomers and polymers from the initial set of oligomers can be represented by:

\[ \alpha_1 \]

\[ \alpha_2 \]

\[ \alpha_3 \]

\[ \cdots \cdots \cdots \]

\[ \alpha_m \]

Fig. 1.0

In this graph, each node \( i \) corresponds to an oligomer. The transitions \( i \rightarrow i \) describe the process of replication, whereas the transitions \( i \rightarrow j \) represent the process of aggregation in which new species of oligomers are formed as a result of the processes of mutation, insertion, deletion, and bonding. The weights attached to the transitions \( i \rightarrow j \) depend precisely on the reaction kinetics. The precise values of these kinetics will not concern us, as this model aims only to explain the qualitative behavior of the results of the processes of aggregation and replication. In the case of a process initiated by a set of oligomers of type \( I_1 \), for instance, the dynamical process in discrete time can be represented by:

\[ \begin{array}{c}
I_1 \\
I_2 \\
I_3 \\
\cdots \cdots \cdots \\
I_m \\
\end{array} \]

Fig. 1.1

The transitions \( I_1 \rightarrow I_1 \) represent replication and the transitions \( I_j \rightarrow I_i \) represent the processes of aggregation.

I shall call each path in the diagram given by Fig. 1.1 a genealogy. This is denoted by the sequence:

\[ x = (x_0, x_1, x_2, \ldots) \]

Here \( x_i \) corresponds to an oligomer \( I_k \), for instance, at generation \( i \).

Let \( \Omega \) denote the set of all genealogies generated by the process and let \( T \) denote the shift operator on \( \Omega \). Let \( \mu \) denote a shift invariant probability measure on the set \( \Omega \) of genealogies. In characterizing the equilibrium distribution on the set of genealogies, consider two cases:

(a) The constant environment model. In this case the processes of aggregation and replication are determined completely by the chemistry of the macromolecules. Defects due to noise will not significantly affect the aggregation or the attachment of the template to the enzyme.

(b) The variable environment model. Here I assume that replication and aggregation are influenced by environmental noise. This can occur as a result of variations in the temperature or enzyme concentration. Temperature fluctuations influence aggregation by periodically breaking apart and reforming conjugated pairs of macromolecules. Variations in enzyme concentration influence replication by its effect on the binding of the enzyme–template complex.

Constant Environment Models. The dynamics of the population of oligomers can be described by a potential function on the space of genealogies. This function ascribes to each genealogy a number that depends on the replication and mutation rates of the oligomers that constitute the genealogy. To describe this function, consider a finite set \( A = \{i_1, i_2, \ldots, i_m\} \) and let

\[ x_A = (x_{i_1}, x_{i_2}, \ldots, x_{i_m}) \]

denote a finite path \( \Omega_A \) in the graph given in Fig. 1.0. Interactions are defined as functions \( \phi \) on the space \( \cup \Omega_A \), where \( A \) runs over all nonempty finite subsets of the integers \( Z \). Assume that \( \phi \) is invariant with respect to the shift operator \( T \). The interaction \( \phi \) can be associated to a function \( \Phi_\phi \) on \( \Omega \) given by:

\[ \Phi_\phi(x) = \sum_{A \neq \emptyset} \frac{1}{|A|} \phi(x_A). \]

Henceforth \( \Phi \) is simply written to denote \( \Phi_\phi \).

Consider a set of oligomers of type \( I_1 \) and let \( \Omega' \) denote the set of genealogies generated by the type \( I_1 \). Let \( x = (x_0, x_1, \ldots, x_t) \) denote a genealogy of length \( t \) generated by the oligomers \( I_1 \) and let \( \{x\} = \{y \in \Omega, y_i = x_i, 1 \leq i \leq t\} \).

For \( z \in \{x\} \), write:

\[ S_z[\Phi(z)] = \sum_{i=1}^{t} \Phi(T^i(z)). \]

where \( T \) denotes the translation operator, and write

\[ N_z = \sum_{\sigma \in S_{I_1} \cdots S_{I_t}} \exp S_z[\Phi(z)]. \]

The number \( N_z \) describes the population size of the oligomers at time \( t \).

This interpretation is evident in the case in which the polymer growth has exponential kinetics. In this case, the dynamics can be described by a matrix \( A = (a_{ij}) \neq 0 \), where the diagonal elements describe the replication rates of the oligomers and the off-diagonal elements describe the mutation rates from oligomer \( I_i \) to \( I_j \). For this class of dynamical system (see ref. 7), the potential \( \Phi \) in the space of genealogies \( \Omega \) is given by a pair potential:

\[ \Phi(x) = \log a_{x_0 x_t}. \]

For this potential, by using Eq. 1.0,

\[ S_z[\Phi(z)] = \log a_{x_0 x_1} a_{x_1 x_2} \ldots a_{x_{t-1} x_t}. \]
Hence
\[ \exp S_t(\Phi(z)) = \prod_{i} a_{x_i, x_{i+1}} \cdots a_{x_{t-1}, x_t}. \tag{1.4} \]

This expression represents at time \( t \), the total number of oligomers with ancestry \( x_0, x_1, x_2, \ldots, x_t \).

The number \( N_t \) is given by
\[ \sum_{x_0, x_1, \ldots, x_t} a_{x_0, x_1} a_{x_1, x_2} \cdots a_{x_{t-1}, x_t}. \tag{1.5} \]

On account of the interpretation in Eq. 1.4, the expression in Eq. 1.5 represents the total number of oligomers in the population at time \( t \).

Now, consider Eq. 1.1. It can be shown [7] that \( N_t \) is subadditive; hence, we have
\[ \lim_{t \to \infty} \frac{1}{t} \log N_t = r(\Phi). \tag{1.6} \]

This asserts that \( N_t \) increases asymptotically at a rate denoted by \( r(\Phi) \) to denote the dependence on \( \Phi \).

When \( \Phi \) is given by Eq. 1.2, \( r(\Phi) \) is precisely \( \log \lambda \), where \( \lambda \) is the dominant eigenvalue of the matrix \( A \). The number \( r(\Phi) \) satisfies a variational principle (see ref. 7):
\[ r(\Phi) = \sup_{\mu} \left( H(\mu) + \int \Phi d\mu \right). \tag{1.7} \]

where \( H(\mu) \) is the entropy of the dynamical system \((\Omega, \mu, T)\).

The relation in Eq. 1.7 is analogous to the Gibbs variational principle and is the basis for the connection between statistical mechanics and population biology established in ref. 7.

On account of Eq. 1.7, \( \mu \) is said to be an equilibrium state for the population of oligomers described by the interaction \( \Phi \), provided
\[ r(\Phi) = H(\mu) + \int \Phi d\mu. \tag{1.8} \]

The choice of the potential function, and hence the number of equilibrium states that can arise in a macromolecular population whose evolution is independent of environmental influences, will depend on the complex chemistry of the macromolecules.

For example, in RNA synthesis, replication is determined by the secondary and tertiary structure of the macromolecule. Now sequences that contain strands with the structure G-G-C-C-G-C-... for instance, are inviable owing to their self-conjugating property. Exclusion of such self-conjugating sequences corresponds to a long-range interaction. These interactions can lead to multiple equilibrium states.

In template-free DNA replication, on the other hand, replication is determined by the primary structure, and, hence, self-conjugating sequences are viable. Since these self-conjugating sequences are not excluded, it follows that the potential function arising in DNA replication can be characterized by short-range interactions. This class of interactions is given by:
\[ \Phi(x_A) = \begin{cases} h(x_i); & x_A = (x_i) \quad A = \{i\} \subset Z \\
J_{i,j}(x_i, x_j); & x_A = (x_i, x_j) \quad A = \{i, j\} \subset Z \\
0; & \text{otherwise}, \quad |i - j| > r. \end{cases} \]

Here the functions \( h \) and \( J_{i,j} \) are real-valued functions on the space \( S \) and \( S \times S \), respectively, where \( S = \{1, 2, \ldots, m\} \) the set of macromolecular types, and \( r \) is the range of interactions.

It is well known that these classes of interactions give rise to unique equilibrium states.

**Variable Environment Models.** The environment is considered as an interference or noise that affects both the aggregation and the replication process. Environments can have both activating and inhibiting influences. Their effect will be to induce changes in the replication and aggregation of the oligomers and hence alterations in the sequence of genealogies produced.

To formalize the model, let \( \Omega \) denote the set of genealogies generated and \( \Omega' \) denote the set of genealogies that result from the action of the environment and let \( F \) and \( F' \) denote the Borel field generated by cylinder sets on \( \Omega \) and \( \Omega' \). The environment is described by a probability distribution \( \nu(x, \cdot) \) defined over \( F' \). This represents the probability that when a genealogy \( x \) is generated, the resulting genealogy belongs to the set \( F' \).

Consider a probability distribution \( \mu \) on the genealogies \( \Omega \). The effect of the noise will induce a new measure \( \mu' \) on \( \Omega' \). To describe \( \mu' \), consider \( A \subset \Omega \); \( B \subset \Omega' \). The probability \( \omega(A \times B) \) is the probability of the joint event \( x \in A, x' \in B \). This is given by:
\[ \omega(A \times B) = \int_A \nu(x, B) d\mu(x). \tag{1.9} \]

The expression for \( \mu' \) indicates that the distribution on the genealogies is altered by noise. We are interested in the set of genealogies that has the property that any alteration by noise will not transform the genealogies in the set closer to another reasonable genealogy than the original. This set can be characterized as follows:

Let \( \Omega^{(0)} \) denote the set of genealogies generated up to time \( t \) and let \( \Omega^{(0)} \) denote the resulting genealogies. Given \( S \subset \Omega^{(0)} \), let \( \nu(S, x) \) denote the probability that when a genealogy \( x \) is generated, the resulting genealogy belongs to the set \( S \).

Let \( L \) denote the set of mappings from the set \( \Omega^{(0)} \) to the set \( \Omega' \) and let \( W_f \) denote the set of elements \( x \in \Omega^{(0)} \) such that
\[ \nu(x, f^{-1}(x)) > 1 - \epsilon. \tag{1.10} \]

The set \( W_f \) has the property that I set out to characterize. Let \( N_f(x, \epsilon) \) denote the number of elements in this set. Write \( N_f(x, \epsilon) = \max N_f(x, \epsilon) \), where the maximum is taken over all \( f \in L \).

The number \( N_f(x, \epsilon) \) represents the number of genealogies that contribute to the growth of the population of oligomers. \( N_f(x, \epsilon) \) is called the effective population size.

We have (7)
\[ \lim_{t \to \infty} \frac{1}{t} \log N_f(x, \epsilon) = K(\nu). \tag{1.11} \]

This limit is denoted \( K(\nu) \) to indicate the dependence of the effective size on the environmental noise \( \nu \).

The number \( K \) satisfies a variational principle analogous to Eq. 1.7. We have (7)
\[ K(\nu) = \sup_{\mu} \left[ H(\mu) + H(\mu') - H(\omega) \right]. \tag{1.12} \]

where \( H(\mu), H(\mu'), \) and \( H(\omega) \) represent the entropy associated with the probability measures \( \mu, \mu', \) and \( \omega \), respectively.

Write \( \psi(\mu) = H(\mu) + H(\mu') - H(\omega) \) and \( \psi(\mu) \) is called the adaptive value of the population. The function \( \psi(\mu) \) repre-
sents the correlation between the variability in the environment and the variability in replication and mutation rates of the oligomers.

On account of Eq. 1.12, one says that the population (Ω, μ) is in equilibrium with the environment when:

\[ K(\nu) = \psi(\mu). \]  

The structure of the equilibrium state clearly depends on the environmental noise and the detailed chemistry of the macromolecules. The statistical properties of the noise are determined by the variations in the binding of the enzyme-template complex. The chemistry of the macromolecules is expressed through its secondary and tertiary structures.

II. Template-Free Synthesis of Macromolecules

The contrasting effects of environmental randomness and environmental constancy in determining the diversity in heteropolymer distributions that occur in template-free syntheses can be understood by distinguishing in detail between the DNA polymerase and the Q8 replication system.

(i) In de novo DNA synthesis, the primer is considered to be a short oligomer A-T-A-T, which generates longer polymers due to the process of slippage and replication. The following steps have been posited. (a) Replication of template with a new strand starting from the 3'-hydroxyl end of the template. (b) Melting of the newly formed helix and its annealing to expose a segment of template for further replication. In this step, reunion of the separated strands is correct in base-base pairing but displaced by one A-T notch from perfect alignment.

These two processes of replication and slippage depend only on the primary structure of the DNA. The rate at which these processes occur is fixed when environmental conditions such as pH and temperature are held constant. This implies that the potential functions on the space of genealogies can be described as a function of the form given in Eq. 1.2.

Since the secondary and tertiary structures of the DNA are not implicated in replication, environmental noise that exerts local effects on the DNA chain will not influence replication. Replication can be considered as occurring in a constant noise-free environment.

The processes of reiterative replication can be described by:

Fig. 2.1

Here 1 corresponds to the primer A-T-A-T, 2 corresponds to the oligomer A-T-A-T-A, and so on. The generation of genealogies can be described:

\[ A = (a_\nu) \geq 0 \text{ denote the matrix whose diagonal terms give the replication rates and whose non-zero off-diagonal terms give the rates of slippage. The potential function for this model is given by Eq. 1.2. Let } \lambda \text{ denote the dominant eigenvalue of } A. \text{ Then equilibrium states are probability measures } \mu \text{ such that} \]

\[ \log \lambda = H(\mu) + \int \Phi d\mu. \]  

For the potential given by Eq. 1.2, there exists a unique equilibrium state. This state is characterized by a population of homopolymers A-T-A-T-A-...-T of length k, for instance, together with a collection of polymers of a shorter length. The distribution of the homopolymers at equilibrium clearly depends on the rates of slippage and replication.

Now in the experiment described in ref. 1, sequences of similar length and structure are produced when the experiment is performed under the same environmental conditions. These findings are in accord with the uniqueness of the equilibrium state resulting from short-range potential under constant environmental conditions.

(ii) In the Q8 system studied in ref. 2, the primers consist of sequences of the form A-G-C, A-G-C-G, and so on. Synthesis is presumed to include (a) Replication of the oligomers with a new strand, starting from the 3'-hydroxyl end of the template. (b) Displacement of the replica from perfect alignment with the template. (c) Point mutations and insertions in the oligomer sequence.

Now in the Q8 system, replication depends on both the secondary and tertiary structures of the oligomer. On account of this, the enzyme-template complex determines the class of sequences that are replicated. This implies that variations in enzyme concentration will influence the rate at which replication occurs and that local defects in the template will affect its viability. The first remark indicates that the replication rates are time-dependent and are therefore described by long-range potentials. The second remark indicates that the replication process will be affected by noise.

The model for the Q8 system, therefore, involves two processes: (i) a process in which new oligomers are generated by the mechanisms of aggregation and replication and (ii) a process in which the noise influences the types of oligomers generated.

These processes can be formalized by the model described in Section I. We consider the oligomer A-G, for instance, as generating genealogies as a result of the process of aggregation and replication. The processes of aggregation and replication have the graphical representation given by Fig. 1.0. The generation of genealogies can be described as:

\[ \text{Fig. 2.2} \]

\[ \text{Fig. 2.3} \]
The processes of aggregation and replication can be described by a potential $\Phi$ on the space of genealogies. Let $\mu$ denote an equilibrium state for the potential $\Phi$.

Now let $\nu$ denote the environmental noise and assume that the probability characteristics of the noise can be described by a matrix $P = (p_{ij})$. Here $p_{ij}$ represents the probability that an oligomer $I_j$ results when $I_i$ is generated. The expression $K(\nu)$ given by Eq. 1.12 now becomes:

$$K(\nu) = \sup_{\nu} \sum_i \sum_j q_i p_{ij} \log \left( \frac{p_{ij}}{q_i q_j} \right). \quad [2.1]$$

Equilibrium states for the population are given by distributions $q = q_i$, which satisfies Eq. 1.12, where $K(\nu)$ is given by Eq. 2.1. In general, there are several such distributions $q = q_i$. This implies that there are several distributions $\mu$ and $\Omega$ for which the maximal adaptive value $K(\nu)$ is attained. Each equilibrium state $\mu$ corresponds to a distribution of polymers whose replication and mutation rates are correlated with the environment.

Now environmental noise may occur even though efforts are made to keep external environmental conditions constant. Thus, the multiplicity of the equilibrium states that arises from the effect of environmental noise suggests a mechanism, distinct from long-range interactions, whereby sequences of different lengths and different electrophoretic patterns, as seen in ref. 2, may be obtained.


**DISCUSSION**

The self-assembly of oligomers to form stable self-replicating heteropolymers is now a well-established experimental fact. The theoretical issue of understanding this self-organization and its implications for prebiotic evolution are central themes in the theory of macromolecular evolution.

On account of the connection between statistical mechanics and evolutionary biology, ideas from statistical mechanics have been invoked to study the problems of self-organization. Two recent contributions in this direction are the works of Dyson (11) and Anderson (12).

Dyson’s model deals with a primitive phase of macromolecular organization in which self-replication does not occur. In this work the mechanism of random drift is invoked to explain the transition from disorder to order in a population of mutually catalytic molecules. The analysis in this model revolves around a potential function that relates the catalytic activity of a newly placed monomer to the catalytic activity of its parent population. This potential function is assumed to depend on a single parameter; the analysis corresponds to a mean-field theory in statistical mechanics.

The model of Anderson (12) deals with both mutation and replication, and natural selection is invoked to explain the transitions to stable heteropolymers. The analysis in this work is based on a potential function described by random interactions. Although a death function is introduced into the potential function in order to describe the effect of environmental factors, the relationship between the different stable states that can arise in a noisy and noise-free environment is not considered. The analysis in ref. 12 is based on the spin-glass formalism in statistical mechanics.

The model proposed in this paper deals with both mutation and selection and invokes random drift and selection as mechanisms of organization. It differs from those in refs. 11 and 12 in that it considers environmental factors as a crucial element in organizing stable states. In noise-free environments, long-range interactions are required to generate multiple stable states. In noisy environments, multiple stable states can arise even when the environmental action is short-range. This distinction emphasizes the significance of environmental noise in generating new stable states. We have introduced the notion of adaptive value to formalize the effect of environmental factors on macromolecular variations. This notion represents the correlation between environmental variability and the variability in the rates of aggregation and replication of the different oligomers. Thus, owing to the effects of noise, natural selection will drive the system from an equilibrium given by maximizing the Malthusian parameter to a new equilibrium in which the adaptive value is maximized. This principle is shown to explain the structures observed in template-free syntheses.