



## The Molecular Roots of Compositional Inheritance

DANIEL SEGRÉ\*, BARAK SHENHAV, RON KAFRI AND DORON LANCET†

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

(Received on 14 February 2001, Accepted in revised form on 31 August 2001)

Non-covalent compositional assemblies, made of monomeric mutually catalytic molecules, constitute an alternative to alphabet-based informational biopolymers as a mechanism of primordial inheritance. Such assemblies appear implicitly in many “Metabolism First” origin of life scenarios, and more explicitly in the Graded Autocatalysis Replication Domain (GARD) model [Segré *et al.* (2000). *Proc. Natl Acad. Sci. U.S.A.* **97**, 4112–4117]. In the present work, we provide a detailed analysis of the quantitative molecular roots of such behavior. It is demonstrated that the fidelity of reproduction provided by a newly defined heritability measure  $\eta_s^*$ , strongly depends on the values of molecular recognition parameters and on assembly size. We find that if the catalytic rate acceleration coefficients are distributed normally, transfer of compositional information becomes impossible, due to frequent “compositional error catastrophes”. In contrast, if the catalytic acceleration rates obey a lognormal distribution, as actually predicted by a statistical formalism for molecular repertoires, high reproduction fidelity is obtained. There is also a clear dependence on assembly size  $N$ , whereby maximal  $\eta$  is seen in a narrow range around  $N \sim 3.5N_G/\lambda$ , where  $N_G$  is the size of the primordial molecular repertoire and  $\lambda$  is a molecular interaction statistical parameter. Such relationships help define the physicochemical conditions that could underlie the early steps in pre-biotic evolution.

© 2001 Academic Press

### Introduction

Biological inheritance is embodied in a capacity to store and transfer information. Two major schools address the origin of this process in radically different ways. Proponents of the “Genome First” scenario suggest the primordial “digital” copying of alphabet-based covalent biopolymer, performed through template-catalysed replication of sequences (Cech, 1993; Eigen, 1971; Orgel, 1992; Swetina & Schuster, 1982). In the alternative, “Metabolism First” or “Composition First” school, early information is contained in the counts of molecules within non-covalent assem-

blies, and inheritance is effected through the action of metabolic cycles, that generate additional copies of the relevant components (Oparin, 1957; Bagley & Farmer, 1991; Dyson, 1999, 1982; Kauffman, 1993; Morowitz, 2000; New & Pohorille, 2000; Wächtershäuser, 1988). One example of a metabolism-first scenario without a nucleic-acid-based genetic apparatus invokes the formation of small aggregates, or coacervation, using the terminology coined by one of the pioneers in this field (Bungenberg de Jong, 1936), and widely used by Oparin (Oparin, 1957).‡

‡ Since criticism has been expressed (cf. Yockey, 1992; Fry, 2000) on Oparin’s philosophical positions as seen in his later work (Oparin, 1957), it is stressed that we refer here solely to the mechanistic-reductionist aspects of his life-long contributions (Oparin, 1957, 1967).

\*Present address: Department of Genetics, Harvard Medical School, 200 Longwood Ave, Boston, MA 02115, U.S.A.

† Author to whom correspondence should be addressed.  
E-mail: [doron.lancet@weizmann.ac.il](mailto:doron.lancet@weizmann.ac.il)

This last view overcomes some debated aspects related to the spontaneous abiotic formation of information-coding biopolymers (Shapiro, 1984, 2000).

The fidelity of information transfer in polymer-based mechanisms has been characterized in detail (Alves & Fontanari, 1998; Eigen, 2000; Swetina & Schuster, 1982). In contrast, fewer attempts have been made to formally delineate the mechanisms of information transfer in metabolism- and composition-based self-reproducing systems (Morowitz, 1967; Szathmáry, 1999). The paucity of quantitative analyses is considered a weak point of such collective replication models, and mutually catalytic sets are often regarded as too imprecise to represent units of heredity. The present work addresses this point, showing that while compositional entities may succumb to error catastrophes under certain conditions, high fidelity of compositional information transfer may be attained under other, biochemically sound conditions.

The Graded Autocatalysis Replication Domain (GARD) model (Segré *et al.*, 1998a, 2000a) provides a quantitative tool for detailed analyses of inheritance without biopolymers. GARD has its departure point with a random collection of organic molecules. It does not concern itself with the explicit mechanisms of their formation, or with their exact identity. Controversies exist regarding the nature and substance concentrations of the primordial chemical mixture (“primeval soup”) (Gesteland, 1999; Yockey, 1992). The relative insensitivity of the GARD model to the exact composition of such mixtures has been previously discussed (Segré *et al.*, 2001). As in the realms of random and combinatorial chemistry, it is assumed that sufficient molecular diversity has formed by a multitude of pre-biotic chemical pathways (Bernstein *et al.*, 1999; Chyba & Sagan, 1992; Deamer, 1997; Kauffman, 1993; McCollom *et al.*, 1999; Schwartz, 1996). GARD, similar to previous metabolism-based pre-biotic models, depicts subsets of such molecules that may undergo self-reproduction, mediated by a complex web of mutually catalytic interactions. A set of differential equations is then used to simulate the chemical kinetics of processes within such a molecular assembly, showing that composition is homeostatically preserved upon growth (Segré

*et al.*, 1997, 1998b). It is shown that when a GARD system undergoes a physical separation process (“splitting”) and is kept far from equilibrium, multiple quasi-stationary states (“composomes”) appear, which display an evolution-like behavior (Segré *et al.*, 2000a).

We have now explored the universe of physicochemical parameters that determines whether such evolutionary phenomena arise. It is shown that only certain ranges of statistical molecular interaction parameters and assembly size lead to high fidelity self-reproduction.

### Compositional Space

Consider a collection of  $N$  molecules constrained in physical space. In an aqueous environment, this could be through the formation of non-covalent molecular assemblies, e.g. micelles or colloidal particles (Bachmann *et al.*, 1992; Deamer, 1997; Luisi *et al.*, 1999; Bungenberg de Jong, 1936; Oparin, 1957; Ourisson & Nakatani, 1994; Segré *et al.*, 2001). If the constituent molecules are drawn from a repertoire of  $N_G$  different kinds, then the assembly may be represented by an  $N_G$ -dimensional vector  $\mathbf{n}$ , whose  $i$ -th component  $n_i$  indicates the count of molecular type  $i$ .

The assembly size is then given by

$$N = \sum_{i=1}^{N_G} n_i. \quad (1)$$

Given two such assemblies  $\mathbf{n}$  and  $\mathbf{m}$ , the compositional similarity can be computed as the scalar product

$$H(\mathbf{n}, \mathbf{m}) = \frac{\mathbf{n} \cdot \mathbf{m}}{|\mathbf{n}| |\mathbf{m}|}, \quad (2)$$

where  $|\mathbf{n}|$  and  $|\mathbf{m}|$  are Euclidean norms.  $H$  is an analog of the Hamming distance, often applied to biopolymer sequence comparisons (Swetina & Schuster, 1982), and may similarly be used for defining the fidelity of reproduction (Segré *et al.*, 2000a).

An analogous similarity measure could be defined in terms of mutual information or entropy of random variables (Shannon, 1948; Yockey, 1992). Figure 1(a) demonstrates that the two measures are related to each other. Thus, for

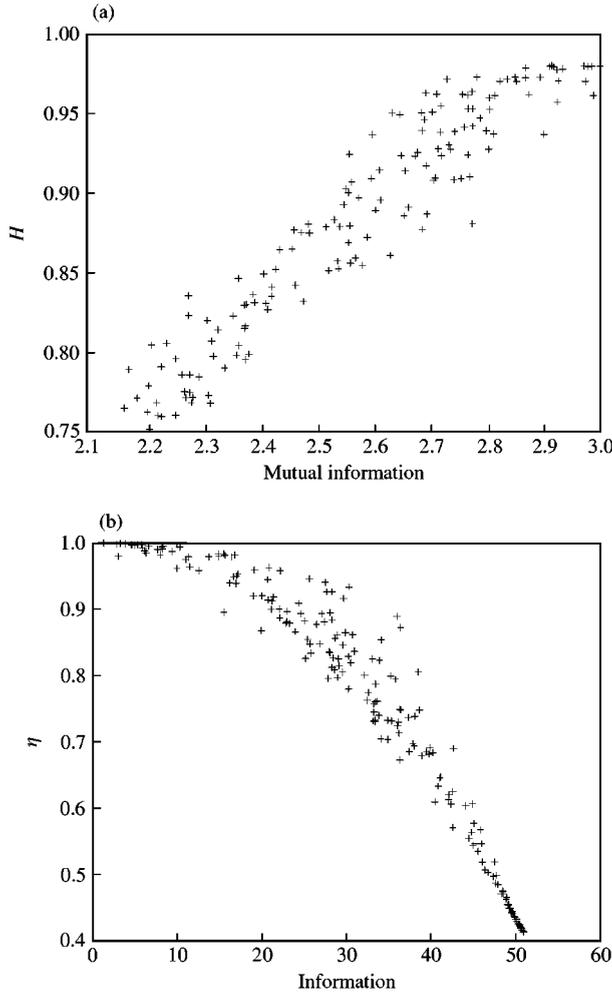


FIG 1. (a) A graph showing the correlation between compositional similarity and mutual entropy. For pairs of random vectors  $\mathbf{n}$  and  $\mathbf{n}'$  the compositional similarity  $H$  [eqn (2)] is plotted vs. mutual entropy, defined as follows:

$$I(\mathbf{n}, \mathbf{n}') = \sum_{i,j} p(n_i, n'_j) \log \left( \frac{p(n_i, n'_j)}{p(n_i)p(n'_j)} \right)$$

(Reza, 1994; Yockey, 1992). All the components of vector  $\mathbf{n}$  are selected from the same uniform distribution. The component  $i$  of vector  $\mathbf{n}'$  is obtained through

$$n'_i = wn_i + (1 - w)u,$$

where  $u$  is a uniformly distributed number, and  $w$  is a weighting parameter that can vary between 0 and 1. In this way, the degree of correlation between the two vectors can be arbitrarily assigned. Vectors of length 1000, and a uniform distribution for  $n_i$  and  $u$  between 0 and 60 were used. Calculations here and elsewhere were performed with Matlab 5.3 (Mathworks, 1999) on a Pentium III processor. (b) A correlation plot relating heritability,  $\eta$ , to the conditional Shannon information for the splitting process  $\mathbf{n} \rightarrow \mathbf{n}'$ . The conditional Shannon information was computed through eqn (7) using the Monte Carlo approximation. For each of 200 points, 1000 random splitting events (i.e. samplings) were computed. Here we use  $k_f = 10^{-3}$ ,  $k_b = 0$ ,  $N_G = 100$ ,  $N = 20$ , and  $\lambda$  uniformly distributed between 1 and 50.

the simple case of compositional vectors dealt with here, eqn (2) seems to be adequate.

For a sample  $\mathbf{n}'$  of the assembly  $\mathbf{n}$ , with size  $N' = fN$  ( $f < 1$ ), we define a heritability measure  $\eta$ , as a mean similarity between an assembly  $\mathbf{n}$  and all its samples of size  $N'$ , as follows:

$$\eta(\mathbf{n}, f) = \langle H(\mathbf{n}, \mathbf{n}') \rangle_{\mathbf{n}'} = \sum_{\mathbf{n}'} P_s(\mathbf{n}'|\mathbf{n}) \cdot H(\mathbf{n}, \mathbf{n}'), \quad (3)$$

where  $P_s$  is the hypergeometric multivariate for sampling without replacements defined as

$$P_s(\mathbf{n}'|\mathbf{n}) = \frac{1}{\binom{N}{N'}} \prod_{i=1}^{N_G} \binom{n_i}{n'_i}. \quad (4)$$

Under the Stirling approximation,  $P_s$  also fulfills the equation

$$\log P_s(\mathbf{n}'|\mathbf{n}) = I(\mathbf{n}') + I(\mathbf{n} - \mathbf{n}') - I(\mathbf{n}), \quad (5)$$

where  $I$ , in analogy to the Shannon information content (Shannon, 1948; Reza, 1994) is defined as

$$I(\mathbf{n}) = -N \sum_{i=1}^{N_G} \frac{n_i}{N} \log \left( \frac{n_i}{N} \right). \quad (6)$$

Thus,  $\log P_s$  is equal to the difference between the information content of the original assembly,  $\mathbf{n}$ , and the sum of the information contents for the two progeny assemblies, the sample  $\mathbf{n}'$  and its complement  $\mathbf{n} - \mathbf{n}'$ . We asked how the heritability  $\eta$  is correlated to the conditional Shannon information (entropy) (Shannon, 1948; Yockey, 1992). For this we utilized a modified expression [cf. Reza, 1994, eqns (3–86)], in which the conditional entropy

$$I(\mathbf{n}'|\mathbf{n}) = - \sum_{\mathbf{n}'} P_s(\mathbf{n}'|\mathbf{n}) \log P_s(\mathbf{n}'|\mathbf{n}) \quad (7)$$

is not averaged on the entire probability distribution, but relates to a specific value of  $\mathbf{n}$ . Figure 1(b) suggests that  $\log P_s(\mathbf{n}'|\mathbf{n})$  is statistically correlated to  $H(\mathbf{n}, \mathbf{n}')$ . Thus, through a comparison to eqn (3), the heritability  $\eta$  may be interpreted as roughly equivalent to information transfer.

Since a direct calculation of  $\eta$  according to eqn (3) would demand an excessive computing

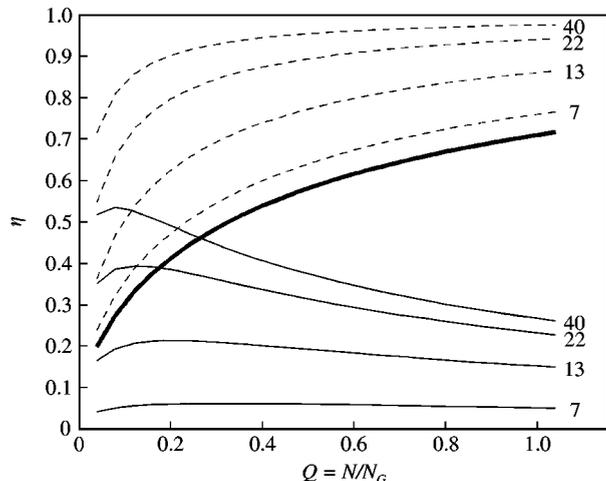


FIG. 2. The behavior of heritability measures  $\eta^*$  (dashed lines),  $\eta_0$  (bold line) and  $\eta_s^*$  (continuous lines) as a function of  $Q = N/N_G$ , for different values of the parameter  $\lambda$ , as marked. For low  $\lambda$  values,  $\eta^*$  is rather similar to the baseline  $\eta_0$ , and therefore the specific heritability  $\eta_s^*$  is quite small. In contrast, for high  $\lambda$  values, and in particular for low  $Q$ , the value of the difference  $\eta_s^*$  becomes appreciable. The curves are computed with  $N_G = 100$ , a lognormal distribution ( $c = 0$ ), and reaction parameters  $k_f = 10^{-3}$ ,  $k_b = 0$ ,  $\rho = 1/N_G$  and  $\alpha/RT = 0.7$ . The heritability measures were computed using a Monte Carlo method, assuming that the frequency of random samples of an initial vector  $\mathbf{n}$  automatically reflect their probability  $P_s$  [eqn (4)].

time, we use two estimations: a Monte Carlo approximation (cf. Fig. 2), and a heuristic approximation

$$\eta \approx 1 - \frac{1}{J(aN + b)}, \quad (8)$$

where  $a$  and  $b$  are positive constants,  $J = |\mathbf{n}|/N$  and  $f = 0.5$  is assumed.  $J$  is another information-related parameter, which measures the deviation from compositional uniformity, having a minimum  $J = N_G^{-0.5}$  for an equimolar (diagonal) composition  $\mathbf{n}_0$ , and a maximum  $J = 1$  if only one species is present. In both estimates, we simplify the computational problem through the assumption of sampling with replacements.

Equation (8) provides a quantitative embodiment of a previous statement (Segré & Lancet, 2000) that assemblies which have a high degree of bias, and/or a large size are more likely to generate progeny with a composition similar to their own. Indeed, in the realm of  $Q > 1$  (where  $Q = N/N_G$ ), even randomly composed assemblies

will display effective reproduction (cf. (Morowitz, 1967; Segré & Lancet, 2000)), but practically no compositional variability will be generated. To correct for this trivial effect, we further define  $\eta_s$ , the specific heritability, as

$$\eta_s(\mathbf{n}) = \eta(\mathbf{n}) - \eta_0, \quad (9)$$

where  $\eta_0 = \eta(\mathbf{n}_0)$  represents the baseline heritability of an assembly with an unbiased (equimolar) composition  $\mathbf{n}_0$  and with the same size  $N$  (Fig. 2).  $\eta_s$  is always positive, since  $\mathbf{n}_0$  is the worst replicator of size  $N$ . A rough estimator of  $\eta_0$  is obtained by substituting in eqn (8) the value  $J = N_G^{-1/2}$ , leading to  $\eta_0 = 1 - (N_G)^{1/2}/(aN + b)$ .

### Compositional Dynamics

The foregoing definitions are in the domain of “statics”, and can be applied to assemblies with specific compositions and sizes, irrespective of their chemical dynamics. They may, however, also be used to shed quantitative light on the time-dependent trajectories of compositional assemblies, as exemplified by those embodied in the GARD model. The growth process of a GARD compositional assembly is described by the differential equations (Segré *et al.*, 2000a)

$$\frac{dn_i}{dt} = (k_f \rho_i N - k_b n_i) \left( 1 + \frac{1}{N} \sum_{j=1}^{N_G} \beta_{ij} n_j \right), \quad (10)$$

where  $k_f$  and  $k_b$  are, respectively, the uncatalysed forward and backward rates for monomer accretion,  $\rho_i$  is the external concentration of molecular species  $i$ , and  $\beta_{ij}$  is an element of the  $N_G \times N_G$  positive matrix that defines a network of mutually catalytic interactions, governed by a statistical formalism (Lancet *et al.*, 1994b; Segré & Lancet, 1999; see also Appendix A).

As previously described (Segré *et al.*, 2000a), assemblies are further assumed to split upon reaching a limiting size, and a constant population constraint (Küppers, 1983) is utilized to keep the number of assemblies fixed. These constraints maintain the system far from equilibrium, as the low free energy state represented by large assemblies is continuously disrupted. Under these conditions, eqn (10) may be approximated by the

linear matrix equation

$$\frac{d\mathbf{n}}{dt} = B\mathbf{n} \quad (11)$$

with  $B_{ij} = k_f \rho (1 + \beta_{ij})$  ( $\rho$  is assumed to be equal for all species). A modified form of eqn (10) in terms of molar fractions  $n_i/N$ , rather than molecular counts,  $n_i$ , is formally equivalent to the quasi-species equation (Jain & Krishna, 1998).

A system involving assembly splitting, and governed by eqn (10) or (11) may manifest multiple quasi-stationary states (Segré *et al.*, 2000a). In the analyses below, we refer specifically to the solution corresponding to the eigenvector with highest eigenvalue, whose components are guaranteed to be real and positive by the Perron–Frobenius theorem (Jain & Krishna, 1998). This solution,  $\mathbf{n}^*$ , represents the canonical steady state, or canonical composome, reached by the molar fractions of GARD components upon assembly exponential growth without disruptions.

In a typical computer experiment we simulate the behavior of an assembly of size  $N$  in a universe of  $N_G$  chemical species. The elements of a catalytic matrix  $\beta$  are sampled from a distribution  $\Phi(\beta_{ij})$  (see below), and the canonical composome  $\mathbf{n}^*$  is numerically computed. The heritability parameter  $\eta_s^* = \eta_s(\mathbf{n}^*)$  (Fig. 2) is then used to represent the kinetic and static effect of parameter values on the simulated reproduction behavior.

### The $\Phi$ Distribution of Rate Parameters

The catalytic interactions in a pre-biotic scenario involving random repertoires of small organic molecules were likely to be relatively ineffective and non-specific, antipodally different from those seen in present day enzymes and substrates. In the absence of detailed knowledge about the pre-biotic components and the interactions among them, we have proposed to utilize a statistical formalism as a means for computing molecular interactions (Lancet *et al.*, 1993, 1994a). This is based on similar principles to those used in the realm of enzyme design (Marks *et al.*, 1992; Tawfik & Griffiths, 1998), as well as random chemistry and combinatorial libraries

(Altreuter & Clark, 1999; Cousins *et al.*, 2000; Hoogenboom, 1997).

To carry out the computations in the present study, we use the explicit definitions of a modified Receptor Affinity Distribution (RAD) model (Lancet *et al.*, 1993, 1994a, b; Rosenwald & Lancet, submitted) in a Poisson approximation for the distribution of rate enhancement values (see Appendix A). This distribution contains a molecular interaction parameter  $\lambda$ , which may be interpreted as the average number of successful intermolecular subsites recognition events. In the continuous approximation we use here, the rate enhancement parameters obey a lognormal probability distribution  $\phi(\beta_{ij})$  (see Appendix A). In order to test the relevance of this specific shape in the realm of mutually catalytic assemblies, we further define a class of distributions of rate enhancement factors that can gradually transform from normal to lognormal, as follows:

$$\beta_{ij} = t^{(1-c)} e^{ct}, \quad (12)$$

where the variable parameter  $t$  is assumed to be distributed normally with parameters  $\mu$  and  $\sigma$  and  $c \in (0, 1)$  effects the graded shift from a normal ( $c = 0$ ) to a lognormal ( $c = 1$ ) distribution.

### Conditions for Faithful Compositional Inheritance

Figure 3 shows the joint dependence of  $\eta_s^*$  on the catalytic interaction parameter  $\lambda$  and the normal to lognormal transition variable  $c$ . It can be seen that both a high proportion of lognormal shape, as well as a high  $\lambda$  value are necessary in order to attain high heritability values near  $\eta_s^* = 0.5$ . Specifically, a normal distribution ( $c \approx 0$ ) cannot lead to high heritability, even for very large  $\lambda$ . The dependence on both  $\lambda$  and  $c$  has a sigmoidal increase, resembling a slow phase transition. Above the transition,  $\eta_s^*$  reaches high values, that represent assemblies with a significant level of homeostasis, and therefore of faithful transmission of compositional information in a process akin to reproduction.

Figure 4 provides an indication that the dependence of the heritability on  $\lambda$  resembles the classical quasi-species error threshold. When  $\lambda$  is large, most assemblies belong to a population

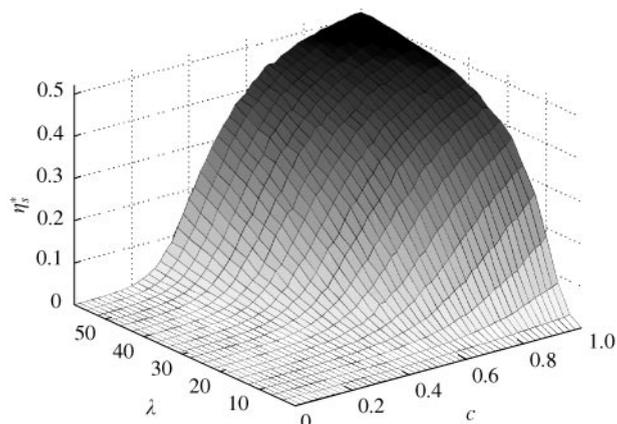


FIG. 3. A plot of the heritability  $\eta_s^*$  as a function of the parameters  $\lambda = \sigma^2$  and  $c$ , the lognormality parameter [eqn (12)], which govern the distribution of catalytic rate enhancement factors. Lognormal distributions (high  $c$  values) with large enough  $\lambda$  are required for effective compositional inheritance. The parameters used were as in Fig. 2, with  $N = 30$  and  $\mu = \lambda$ . This was also verified for a wider range of independently selected  $\mu$  values. For each point in the  $(c, \lambda)$  grid, 1000 different  $\beta$  matrices were sampled and diagonalized for the calculation of the canonical composome  $\mathbf{n}^*$ . Heritability values were then computed by using eqn (8). This was shown to be a good heuristic approximation of the Monte Carlo estimate. The correlation coefficient between these two estimates was found to be 0.997 for a set of over 13000 points. The parameters  $a$  and  $b$  of eqn (8) were estimated with a linear fit ( $a = 0.177$ ;  $b = 10.5$ ). The dependence of the approximated formula on  $N_G$  was found to be negligible in the analysed region (data not shown).

with high heritability. For decreasing  $\lambda$ , intermediate  $\eta_s^*$  classes appear, and finally, at  $\lambda \sim 5$  an abrupt transition occurs to a state where most assemblies belong to a class with heritability values characteristic of a state of randomness in a kind of “error catastrophe”. This may be seen as analogous to the threshold of information transfer in Shannon’s (1948) Channel Capacity Theorem.

We next explored the way by which heritability depends on assembly size  $N$ . Figure 5 shows the dependence of  $\eta_s^*$  on  $Q = N/N_G$  for different values of  $\lambda$ .  $\eta_s^*$  is found to have a maximum ( $Q_{max} = N_{max}/N_G$ ) generated by the interplay of the dependence of  $\eta^*$  and  $\eta_0$  on  $Q$ . Assemblies with  $Q$  near to  $Q_{max}$  are the best candidates for propagating compositional information without significant losses. Interestingly, when this behavior is compared with different  $\lambda$  values, a simple linear relationship is found to occur between

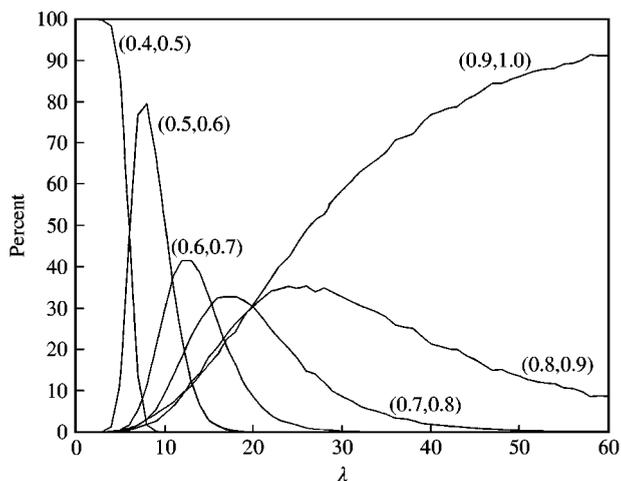


FIG. 4. The dependence of the heritability measure  $\eta_s^*$  on  $\lambda$ , in a form that resembles the quasi-species error threshold analysis (Swetina & Schuster, 1982). Each curve represents the dependence on  $\lambda$  of the percent of assemblies with  $\eta_s^*$  in a range  $(\eta_s^*, \eta_s^*)$  as marked. For a fixed value of  $\lambda$ , the different curves indicate the percent ratios of assemblies belonging to different categories of heritability levels. For example, when  $\lambda$  approaches very small values, 100% of the assemblies have heritability in the lowest range. Lognormal distributions with different  $\lambda$  values may lead to drastically different behaviors, ranging from an extreme in which practically all assemblies display high heritability, to another in which most assemblies undergo random split without information transfer (low  $\eta_s^*$ ). Computations as in Fig. 2. A value  $\mu = \lambda$  was used for the lognormal distribution.

$1/Q_{max}$  and  $\lambda$  (Fig. 5, inset) leading to a relationship  $N_{max} \approx 3.5N_G/\lambda$ . This implies that in mixtures with high  $\lambda$  values, i.e. with a large mean and variance of the distribution  $\Phi(\beta_{ij})$ , smaller compositional assemblies can faithfully reproduce.

Not all values of  $\lambda$  are equally probable physicochemically. As the pre-biotic organic compounds would typically have low molecular mass, high  $\lambda$  values should be rendered unlikely (Appendix A). To embody this, a normal probability distribution  $P(\lambda)$  is used as a weighting factor (Fig. 6). The plot displays a maximum, which corresponds to the point at which heritability is high, but not trivial, and the  $\lambda$  is still physically probable. In other words, this is the region in the  $(Q, \lambda)$  parameter space in which non-covalent aggregates were most likely to initiate a progression towards a compositionally based protocellular system.

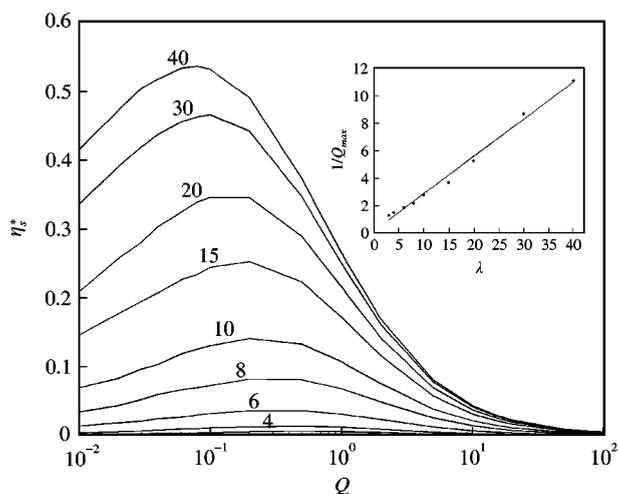


FIG. 5. The dependence of  $\eta_s^*$  on  $Q = N/N_G$ , for different values of  $\lambda$ , as indicated. A  $\lambda$ -dependent maximum is observed for each curve at a value of  $Q_{max}(\lambda)$ . The  $\eta_s^*$  values were calculated as in Fig. 2. Inset: The dependence of  $1/Q_{max}$  on  $\lambda$ . A parameter fit gave a linear relationship with a slope of 0.28.

### Discussion

Compositional inheritance in assemblies with rudimentary metabolic networks may have constituted a form of self-reproduction during the earliest period of the evolution of life. The present study attempts to form a quantitative basis for such reproduction capacity, in parallel to studies of the replication fidelity of nucleotide sequences. The analysis presented here may be important for further exploration of the transition from compositional to sequence-based systems, as described by Segré *et al.* (2000b, c).

Here, as well as in previous reports (Segré *et al.*, 2000a) we have demonstrated that small non-covalent assemblies of mutually catalytic molecules may display a robust behavior typified by homeostatic growth and effective self-reproduction capacity. The emerging metabolism-like networks of catalytic interactions are shown through explicit computation to be able to be resistant to physical splitting of the assemblies, i.e. to manifest self-reproduction characteristics.

The present description for the reproduction of compositional assemblies involves two separable domains of reference. The first is an explicit kinetic mechanism for assembly growth, whereby additional copies of the non-covalently bound molecular components are recruited or gener-

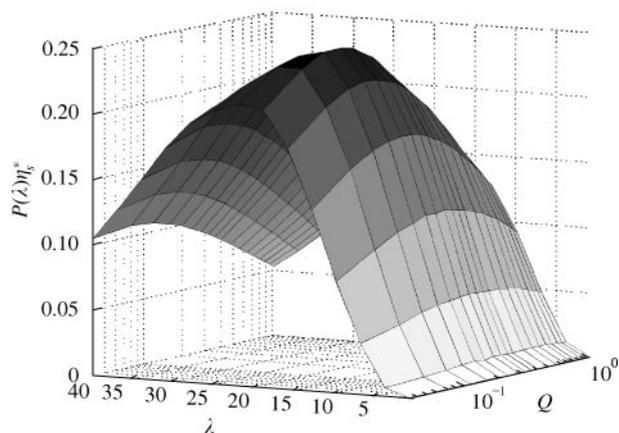


FIG. 6. A three-dimensional diagram of the dependence of  $P(\lambda)\eta_s^*$  on  $Q$  and  $\lambda$ .  $P(\lambda)$  is assumed to obey a Gaussian distribution (See Appendix A), centered around  $\lambda = 2$  with standard deviation of 4. The region of maximal probability for the spontaneous emergence of faithfully reproducing compositional assemblies corresponds to dark colored peak. Computations were carried out with the Monte Carlo method, as in Fig. 2.

ated. This realm is governed by dynamic parameters such as  $\lambda$  or  $c$ , which in turn determine the distribution of the mutual catalysis parameters  $\beta_{ij}$ . The second domain of relevance is related to assembly splitting, a crucial step which embodies progeny generation. We use here a newly defined graded heritability measure,  $\eta_s^*$ , to quantify an assembly's capacity to give rise to new assemblies whose compositions resemble its own. This definition is based on static parameters ( $\mathbf{n}$ ,  $N$ ,  $N_G$ ,  $H$ ), that quantify properties of assemblies and their mutual relationships, irrespective of the dynamic processes that may have underlain their appearance and change.

The essence of the present treatise is the integration of static and dynamic concepts to allow a mathematical description of compositional reproduction for the capacity to transfer information to progeny. A cornerstone of this analysis is the use of a statistical formalism for pair-wise molecular interactions in a random chemistry setting. This provides a means for tracing the molecular roots of compositional inheritance.

We realize that  $\eta_s^*$  does not manifest the entire gamut of dynamic properties of a compositional assembly. First, they are based only on the canonical composome  $\mathbf{n}^*$ , while a complete description would require looking at multiple steady states. Also, relying on  $\eta_s^*$  does not allow one to

fully capture the complex universe of kinetic fates available to an assembly, whereby if low heritability obtains after splitting, the progeny could trace a dynamic trajectory that would lead it back to the canonical quasi-stationary state. However, the broader significance of the heritability measures lies in the fact that assemblies displaying high heritability are less susceptible to compositional mutation-like events that might bring them to other domains in compositional space.

An important result of the present study is a clear definition of the conditions for attaining a high heritability value. We show that homeostatic compositional replicators emerge if the catalytic rate enhancement factors are distributed lognormally. In contrast, if a normal distribution prevails, assemblies capable of generating faithful progeny are not observed in the computer simulations. This is true even for high  $\lambda$  values that guarantee appreciable mutual catalysis. The long tail of the lognormal distribution appears to provide an appropriate diversity of rate enhancement values, thus affording optimal mutually catalytic networks. It appears significant that lognormal distributions also govern the interactions of present-day biomolecules, in particular when combinatorial libraries are involved (Goldstein, 1975; Inman, 1978; Lancet *et al.*, 1993, 1994a; Macken & Perelson, 1991; Rosenwald & Lancet, submitted).

Barabasi and colleagues (Albert *et al.*, 2000; Barabasi & Albert, 1999; Jeong *et al.*, 2000) have recently suggested that large-scale cellular metabolic networks in numerous present-day organisms derive their stability from specific global properties. In such “scale-free” networks, the probability distribution of the number of links for each node decays as a long-tailed power law. The stability of the networks described in the present study may similarly stem from the long-tail distribution of mutual interactions embodied within them. Likewise, our formalisms may be generalized to any system with agent-like units whose mutual interactions are distributed lognormally. Future studies should further explore this intriguing analogy.

In studies that address the fidelity of replication for sequence-based biopolymers, a copying machinery (polymerase) is usually assumed to be

at work, which generates similes of an original sequence with a particular fidelity. It was demonstrated that if the accuracy of replication per monomer position drops below a threshold, an “error catastrophe” results, whereby most of the generated sequence progeny are rather distant from the original one. Significantly, the present analysis of compositional dynamics in a GARD system suggests the occurrence of a similar error threshold phenomenon. We demonstrate that this will happen in GARD assemblies when either a global molecular interaction parameter  $\lambda$  drops below a minimal value, or the size of the assembly is outside a specific range.

The results of the present work go beyond the specific GARD embodiment. This is because the heritability  $\eta_s^*$  depends more on the statistical behavior of compositional vectors than on the dynamic behavior as manifested in the kinetic equations. We believe that the formalisms used here may serve as a platform for future research aiming at a general mathematical description of the emergence of molecular self-reproduction and evolution. More detailed knowledge of the statistical properties of a pre-biotic mixture, as well as more explicit models for non-covalent assembly growth and splitting would eventually lead to more realistic predictions of constraints for the origin of life.

This paper is dedicated to the memory of Shneior Lifson, friend and teacher, a pioneer of the origin of life research, who passed away on January 23, 2001. Doron Lancet holds the Ralph and Lois Silver Chair in Human Genomics. This work was supported by the Crown Human Genome Center, the Krupp foundation, and the Weizmann Institute Glasberg, Levy, Nathan Brunschwig and Levine funds. We thank Dr Luca Peliti and Dafna Ben-Eli for useful discussions and Orna Man for her comments on the manuscript.

## REFERENCES

- ALBERT, R., JEONG, H. & BARABASI, A. L. (2000). Error and attack tolerance of complex networks. *Nature* **406**, 378–382.
- ALTREUTER, D. H. & CLARK, D. S. (1999). Combinatorial biocatalysis: taking the lead from nature. *Curr. Opin. Biotechnol.* **10**, 130–136.
- ALVES, D. & FONTANARI, J. F. (1998). Error threshold in finite populations. *Phys. Rev. E* **57**, 7008–7013.
- BACHMANN, P., LUISI, P. & LANG, J. (1992). Autocatalytic self-replicating micelles as models for prebiotic structures. *Nature* **357**, 57–59.

- BAGLEY, R. J. & FARMER, J. D. (1991). Artificial life II. In: *Spontaneous Emergence of a Metabolism* (Langton, C. G., Taylor, C., Farmer, J. D. & Rasmussen, S., eds), pp. 93–140. Reading, MA: Addison-Wesley.
- BARABASI, A. L. & ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509–512.
- BERNSTEIN, M. P., SANDFORD, S. A., ALLAMANDOLA, L. J., GILLETTE, J. S., CLEMETT, S. J. & ZARE, R. N. (1999). UV irradiation of polycyclic aromatic hydrocarbons in ices: production of alcohols, quinones, & ethers. *Science* **283**, 1135–1138.
- BUNGENBERG DE JONG, H. G. (1936). La coacervation, les coacervats et leur importance en biologie. *Actualites Sci. Indust.* (Paris) **398**.
- CECH, T. R. (1993). The efficiency and versatility of catalytic RNA: implications for an RNA world. *Gene* **135**, 33–36.
- CHYBA, C. F. & SAGAN, C. (1992). Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: an inventory for the origin of life. *Nature* **355**, 125–132.
- COUSINS, G. R., POULSEN, S. A. & SANDERS, J. K. (2000). Molecular evolution: dynamic combinatorial libraries, autocatalytic networks and the quest for molecular function. *Curr. Opin. Chem. Biol.* **4**, 270–279.
- DEAMER, D. (1997). The first living systems: a bioenergetic perspective. *Microbiol. Mol. Biol. Rev.* **61**, 239.
- DYSON, F. (1999). *Origins of Life*. Cambridge: Cambridge University Press.
- DYSON, F. J. (1982). A model for the origin of life. *J. Mol. Evol.* **18**, 344–350.
- EIGEN, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465–523.
- EIGEN, M. (2000). Natural selection: a phase transition? *Biophys. Chem.* **25**, 101–123.
- FRY, I. (2000). *The Emergence of Life on Earth*. NJ: Rutgers University Press.
- GESTELAND, R. F., CECH, T. R. & ATKINS, J. F. (eds) (1999). *The RNA World*, 2nd Edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
- GOLDSTEIN, B. (1975). Theory of hapten binding to IgM: the question of repulsive interactions between binding sites. *Biophys. Chem.* **3**, 363–367.
- HOOGENBOOM, H. R. (1997). Designing and optimizing library selection strategies for generating high-affinity antibodies. *Trends Biotechnol.* **15**, 62–70.
- INMAN, J. K. (1978). The antibody combining region: speculations on the hypothesis of general multispecificity. In: *Theoretical Immunology* (Bell, G. I., Perelson, A. S., and Pimbley, G. H. Jr, eds). New York: Marcell Dekker.
- JAIN, S. & KRISHNA, S. (1998). Autocatalytic sets and the growth of complexity in an evolutionary model. *Phys. Rev. Lett.* **81**, 5684–5687.
- JANIN, J. (1996). Quantifying biological specificity: the statistical mechanisms of molecular recognition. *Proteins: Structure, Function Genetics* **25**, 438–445.
- JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z. N. & BARABASI, A.-L. (2000). The large-scale organization of metabolic networks. *Nature* **407**, 651–654.
- KAUFFMAN, S. A. (1993). *The origins of order: self-organization and selection in evolution*. NY: Oxford University Press.
- KRAUT, J. (1988). How do enzymes work? *Science* **242**, 533–540.
- KÜPPERS, B.-O. (1983). *Molecular Theory of Evolution*. Berlin-Heidelberg: Springer-Verlag.
- LANCET, D., SADOVSKY, E. & SEIDEMANN, E. (1993). Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system. *Proc. Natl Acad. Sci. U.S.A.* **90**, 3715–3719.
- LANCET, D., HOROVITZ, A. & KATCHALSKI-KATZIR, E. (1994a). Molecular recognition in biology: models for analysis of protein/ligand interactions. In: “*Perspectives in Supramolecular Chemistry*” (Behr, J.-P., ed.). New York: John Wiley.
- LANCET, D., KEDEM, O. & PILPEL, Y. (1994b). Emergence of order in small autocatalytic sets maintained far from equilibrium: application of receptor affinity distribution (RAD) model. *Ber. Bunsenges. Phys. Chem.* **98**, 1166–1169.
- LUISI, P. L., WALDE, P. & OBERHOLZER, T. (1999). Lipid vesicles as possible intermediates in the origin of life. *Curr. Opin. Colloid & Interface Sci.* **4**, 33–39.
- MACKEN, C. A. & PERELSON, A. S. (1991). Affinity maturation on rugged landscapes. In: *Molecular Evolution on Rugged Landscapes: Santa Fe Institute Studies in the Sciences of Complexity* (Perelson, A. S. & Kauffman, S. A., eds). Reading, MA: Addison-Wesley.
- MARKS, J. D., HOOGENBOOM, H. R., GRIFFITHS, A. D. & WINTER, G. (1992). Molecular evolution of proteins on filamentous phage. *J. Biol. Chem.* **267**, 16007–16010.
- MATHWORKS, T. (1999). Matlab 5.3.
- MCCOLLOM, T. W., RITTER, G. & SIMONEIT, B. R. T. (1999). Lipid synthesis under hydrothermal conditions by Fisher-Tropsch-type reactions. *Origins Life Evol. Biosphere* **29**, 153–166.
- MOROWITZ, H. J. (1967). Biological self-replicating systems. In: *Progress in Theoretical Biology* (Snell, F. M., ed.), pp. 35–58. New York: Academic Press.
- MOROWITZ, H. J. (2000). The origin of intermediary metabolism. *Proc. Natl Acad. Sci. U.S.A.* **97**, 7704–7708.
- NEW, M. H. & POHORILLE, A. (2000). An inherited efficiencies model of non-genomic evolution. *Simulation Practice Theory* **8**, 99–108.
- OPARIN, A. I. (1957). *The Origin of Life on the Earth*. London: Oliver and Boyd.
- OPARIN, A. I. (1967). The origin of life, Trans. A. Sygne. In: J. D. Bernal, *The origin of life*, pp. 199–234. London: Weidenfeld and Nicolson (First Russian edition: Moscow 1924).
- ORGEL, L. E. (1992). Molecular replication. *Nature* **358**, 203–209.
- OURISSON, G. & NAKATANI, Y. (1994). The terpenoid theory of the origin of cellular life: the evolution of terpenoids to cholesterol. *Chem. Biol.* **1**, 11–23.
- REZA, F. M. (1994). *An Introduction to Information Theory*. New York: Dover.
- ROSENWALD, S. & LANCET, D. Test of a statistical model for molecular recognition in biological repertoires. *J. theor. Biol.* (under revision).
- SCHWARTZ, A. W. (1996). Did minerals perform prebiotic combinatorial chemistry? *Chem. Biol.* **3**, 515–518.
- SEGRÉ, D. & LANCET, D. (1999). A statistical chemistry approach to the origin of life. *Chemtracts—Biochem. Mol. Biol.* **12**, 382–397.
- SEGRÉ, D. & LANCET, D. (2000). Composing life. *EMBO Reports* **1**, 217–222.
- SEGRÉ, D., PILPEL, Y., GLUSMAN, G. & LANCET, D. (1997). Self-replication and evolution in primordial mutually

- catalytic sets. In: *Astronomical and Biochemical Origins and the Search for Life in the Universe, Proceedings of the fifth International Conference on Bioastronomy, IAU Colloquium N. 161* (Cosmovici, C. B., Bowyer, S. & Werthimer, D., eds), pp. 469–476. Bologna: Editrice Compositori.
- SEGRÉ, D., LANCET, D., KEDEM, O. & PILPEL, Y. (1998a). Graded Autocatalysis Replication Domain (GARD): kinetic analysis of self-replication in mutually catalytic sets. *Origins Life Evol. Biosphere* **28**, 501–514.
- SEGRÉ, D., PILPEL, Y. & LANCET, D. (1998b). Mutual catalysis in sets of prebiotic organic molecules: evolution through computer simulated chemical kinetics. *Physica A* **249**, 558–564.
- SEGRÉ, D., BEN-ELI, D. & LANCET, D. (2000a). Compositional genomes: prebiotic information transfer in mutually catalytic non-covalent assemblies. *Proc. Natl Acad. Sci. U.S.A.* **97**, 4112–4117.
- SEGRÉ, D., BEN-ELI, D. & LANCET, D. (2000b). Prebiotic evolution of amphiphilic assemblies far from equilibrium: from compositional information to sequence-based biopolymers. In: *A New Era in Bioastronomy, ASP Conference Series*, Vol. 213 (Lemarchand, G. A. & Meech, K. J., eds), pp. 373–378. Michigan: Sheridan Books.
- SEGRÉ, D., BEN-ELI, D. & LANCET, D. (2000c). The prebiotic transition from compositional to sequence-based information. *Origins Life Evol. Biosphere* **30**, 174–194.
- SEGRÉ, D., BEN-ELI, D., DEAMER, D. & LANCET, D. (2001). The lipid world. *Origins Life Evol. Biosphere* **31**, 119–145.
- SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423; 623–656.
- SHAPIRO, R. (1984). The improbability of prebiotic nucleic acid synthesis. *Origins Life Evol. Biosphere* **14**, 565–570.
- SHAPIRO, R. (2000). A replicator was not involved in the origin of life. *IUBMB Life* **49**, 173–176.
- STEWART, J. D. & BENKOVIC, J. S. (1995). Transition-state stabilization as a measure of the efficiency of antibody catalysis. *Nature* **375**, 388–391.
- SWETINA, J. & SCHUSTER, P. (1982). Self-replication with errors—A model for polynucleotide replication. *Biophys. Chem.* **16**, 329–345.
- SZATHMÁRY, E. (1999). Chemes, Genes, Memes: a revised classification of replicators. *Lect. Math. the life sci.* **26**, 1–10.
- TAWFIK, D. S. & GRIFFITHS, A. D. (1998). Man-made cell-like compartments for molecular evolution. *Nature Biotechnol.* **16**, 652–656.
- WÄCHTERSCHÄUSER, G. (1988). Before enzymes and templates: theory of surface metabolism. *Microbiol. Rev.* **52**, 452–484.
- YOCKEY, H. P. (1992). *Information Theory and Molecular Biology*. Cambridge: Cambridge University Press.

## APPENDIX A

Consider the following unimolecular conversion reaction that may proceed with catalysis ( $k_{catal}$ ) or without it ( $k_{uncat}$ ):

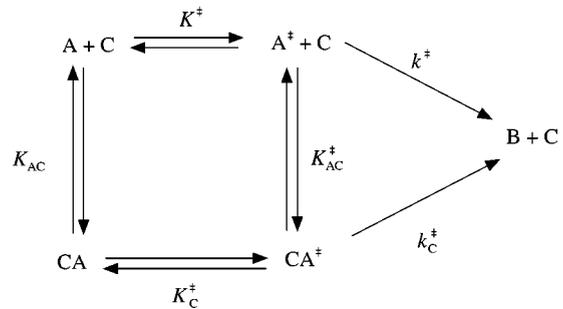
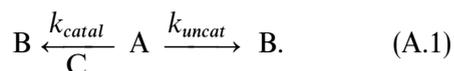


FIG. A1. Pseudo-thermodynamic cycle, showing a kinetic scheme for catalysed and uncatalysed reactions: C, catalyst; A, substrate; B, product; the superscript double dagger ( $\ddagger$ ) denotes the transition state species.

The catalytic rate enhancement coefficient is defined by

$$\beta = k_{catal}/k_{uncat}. \quad (\text{A.2})$$

Applying a transition state theory formalism for enzymes or catalytic antibodies (Kraut, 1988; Stewart & Benkovic, 1995) to this simple catalytic reaction, leads to the scheme shown in Fig. A1, with  $k_{uncat} = K^\ddagger k^\ddagger$  and  $k_{catal} = K_{AC} K_C^\ddagger k_C^\ddagger = K^\ddagger K_{AC}^\ddagger k_C^\ddagger$ . Assuming the catalyst does not affect the vibrational modes of the substrate,  $k^\ddagger = k_C^\ddagger$  obtains, and therefore

$$\beta = K_{AC}^\ddagger. \quad (\text{A.3})$$

Hence, the catalysis enhancement coefficient  $\beta$  is the measure of a binding pseudo-equilibrium constant between the transition state species A and the catalyst C.

This equality enables us to apply existing models for the distribution of receptor binding affinities (Goldstein, 1975; Janin, 1996; Lancet *et al.*, 1993; Rosenwald & Lancet, submitted) to the context of prebiotic catalytic networks. One of these statistical models, the Receptor Affinity Distribution (RAD) model (Lancet *et al.*, 1993) was first introduced for studying recognition properties in multireceptor repertoires, as found in the immune and in the olfactory systems. The RAD model suggests that the number of interactions  $L$  contributing to the energy of binding between a receptor and an arbitrary set of ligands

is distributed binomially:

$$P(L) = \frac{B!}{L!(B-L)!} \left(\frac{1}{S}\right)^L \left(1 - \frac{1}{S}\right)^{B-L}, \quad (\text{A.4})$$

where  $B$  is the number of subsites within the binding site and  $S$  is the subsite diversity.

The affinity of binding  $K$  is related to  $L$  through the following relation:  $L = (RT/\alpha) \log(K)$  where  $\alpha$  is the average energy contribution for a single subsite interaction,  $R$  is the gas constant and  $T$  is the absolute temperature. Based on the transition state logic, we propose that the distribution of the  $\beta_{ij}$  values behaves similarly and therefore write

$$L = (RT/\alpha) \log(\beta_{ij}). \quad (\text{A.5})$$

In a Poisson approximation the probability distribution can be written as (Rosenwald & Lancet, submitted)

$$P(L) = \frac{\lambda^L}{L!} e^{-\lambda} \quad (\text{A.6})$$

with  $\lambda = B/S$ .  $L$  may further be interpreted as a continuous variable in a Gaussian approximation

$$P(L) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(L-\mu)^2}{2\sigma^2}\right], \quad (\text{A.7})$$

where  $\mu = \lambda$  and  $\sigma = \lambda^{1/2}$ . By using the standard rule for transforming probability distributions

$$\Phi(\beta_{ij}) = P(L) \frac{dL}{d\beta_{ij}} \quad (\text{A.8})$$

and based on eqns (17) and (18), an explicit expression is obtained for  $\Phi(\beta_{ij})$  as a log-normal distribution:

$$\Phi(\beta_{ij}) = \frac{1}{\beta_{ij}(\sigma\alpha/RT)\sqrt{2\pi}} \times \exp\left[-\frac{(\log \beta_{ij} - (\mu\alpha/RT))^2}{2(\sigma\alpha/RT)^2}\right]. \quad (\text{A.9})$$

The Poisson RAD model implies a linear relationship between the average size of the binding determinants (number of subsites,  $B$ ) and the molecular interaction parameter  $\lambda$ .  $B$  is related to the average size of the smaller reactant in a binary association. Therefore, we deduce that  $P(\lambda)$ , the probability distribution for the parameter  $\lambda$ , is similar to that of molecular sizes in a primordial mixture. For simplicity, we assume that  $P(\lambda)$  is Gaussian (Fig. 6), manifesting the notion that extreme values of  $\lambda$  are improbable.