An Inherited Efficiencies Model of Non-genomic Evolution

Michael H. New and Andrew Pohorille
Exobiology Branch and Evolutionary Cell Computing Group
NASA Ames Research Center
Mailstop 239-4
Moffett Field, CA 94035

ABSTRACT

A model for the evolution of biological systems in the absence of a nucleic acid-like genome is proposed and applied to model the earliest living organisms — protocells composed of membrane encapsulated peptides. Assuming that the peptides can make and break bonds between amino acids, and bonds in non-functional peptides are more likely to be destroyed than in functional peptides, it is demonstrated that the catalytic capabilities of the system as a whole can increase. This increase is defined to be *non-genomic evolution*. The relationship between the proposed mechanism for evolution and recent experiments on self-replicating peptides is discussed.

1. INTRODUCTION

The ability of organic molecules to self-organize into self-sustaining, reproducing and evolving structures governed the transformation of matter from inanimate to animate on the early earth. Probably the earliest such structures were protocells — membrane-enclosed, cell-like structures capable of supporting essential life functions, such as the capture and utilization of energy and synthesis of proteins. [5] In modern organisms, most of these life functions are performed by proteins which are, in turn, synthesized on an RNA template. It is, however, unlikely that both proteins and RNA arose simultaneously and immediately became interconnected. The discovery of catalytic properties of RNA led to a suggetion that the present world of nucleic acids and proteins was preceded by the "RNA World," wherein RNA molecules alone acted as both catalysts and information storage systems. [2; 1] This concept, however, encounters considerable difficulties. RNA is fragile and no efficient prebiotic syntheses of its building blocks have been found. Furthermore, RNA cannot be readily incorporated into membranes to perform functions which, in modern cells, include energy transduction and transport. Finally, since there is no relationship between the function of a catalytic RNA and the function, if any, of the protein for which it can code, there is no clear path from the RNA World to today's world of protein catalysis and nucleic acid information storage. We therefore hypothesize that initially protocells evolved in the absence of a nucleic acid-based genome and only later did coded information storage emerge. While peptides do not suffer from similar problems as RNA, amino acids cannot base-pair like nucleic acids, so it is not clear how peptides, alone, could transfer information between generations. Thus a new conception of "evolution" is necessary that does not require a nucleic acid–based, or similar, genome.

Central to this new concept of *non-genomic* evolution is the emergence of peptide-bond forming protoenzymes (ligases). In all likelihood, they were initially very weak, non-specific catalysts, joining amino acids to form peptides of various lengths and sequences. A few of the peptides so generated could have been better catalysts of peptide bond formation than the protoenzymes which formed them. These better protoenzymes would, in turn, generate even more peptides, increasing the rate at which a protocell "searched" the space of all peptides for functional ones. Some of the peptides generated in this search would undoubtedly function as proteases, cutting peptide bonds. Since proteases cleave unstructured peptides more rapidly than structured ones, and since functional peptides have to have some degree of ordered structure, the proteases would preferentially destroy non-functional peptides. Occasionally, the newly produced peptides would be capable of performing novel functions. If they integrated into the protocellular metabolism, they could increase its capabilities. This process would eventually lead to the emergence (or utilization) of nucleic acids and their coupling with peptides to yield a genomic system.

For this process to be effective, it is required that protocells grow and divide either by acquiring amphiphilic material from the environment or by producing it internally. The contents of the two "offspring" protocells would not be identical and some would not contain the proper suite of components for self-maintenance. Nevertheless, over time, the catalytic efficiency of a community of protocells might increase. This increase in overall efficiency is *non-genomic evolution.*

Recent breakthroughs in experimental protein chemistry open the gates for systematic experimental and theoretical tests of the ideas undelying non-genomic evolution. Szostak and Roberts [6] have modified the methods of *in vitro* evolution, previously only applicable to nucleic acids, to select peptides with specific properties. This work will provide needed

information on the distribution of catalytic abilities among small peptides. In a series of elegant papers, Ghadiri and co-workers [4; 3; 7] have produced a self-replicating peptide system with an inherent error-correction mechanism and have demonstrated the evolution of populations of peptides. Most recently, Chmielewski, *et al.* [8] have constructed another peptide system capable of auto- and cross-catalysis and generating self-replicating peptides that were not present in the original mixture.

2. THE INHERITED EFFICIENCIES MODEL

To examine the evolutionary potential of a non-genomic system, we have employed a simple, computationally tractable model which is still capable of capturing the essential biochemical features of the real system. In this model, protocellular walls are permeable to amino acids but not to oligopeptides of any length. Within the protocell, the formation and destruction (also called hydrolysis) of bonds between consecutive amino acids in oligopeptides (peptide bonds) occur through catalyzed, albeit possibly very inefficient, pathways. A peptide of any length can act in a double role as a substrate for polymerization or hydrolysis, or as a catalyst of chemical reactions. Since only two reactions are considered in the present model, all peptides are characterized by three traits: their length and their efficiencies as catalysts of ligation and hydrolysis of peptide bonds. These efficiencies can be interpreted as the inverse of turnover rates and are currently assumed to be independent of each other.

In a system composed of different types of amino acids, peptides of the same length but different composition vary in their catalytic ability. In a detailed model, this can be accounted for by providing microscopic rules that relate the peptide sequence to its catalytic efficiency. Since these rules, however, are not known at present, we adopt a stochastic model, in which the specific identities of amino acids are not considered. Instead, the dependence of the catalytic efficiency on the sequence, ϵ, is captured by assuming that the efficiencies of peptides of length n for catalyzing ligation and hydrolysis reactions are distributed with probabilities $p_n^L(\epsilon)$ and $p_n^H(\epsilon)$, respectively. In the current implementation, these probability distributions are Gaussian:

$$p_n^L(\epsilon) = \frac{1}{\sigma^L(n)\sqrt{2\pi}} \exp\left(-\frac{\left(\epsilon-\epsilon_0^L(n)\right)^2}{2\sigma^L(n)^2}\right) \quad (1)$$

$$p_n^H(\epsilon) = \frac{1}{\sigma^H(n)\sqrt{2\pi}} \exp\left(-\frac{\left(\epsilon-\epsilon_0^H(n)\right)^2}{2\sigma^H(n)^2}\right) \quad (2)$$

The position of the maximum of each distribution function increases, in a sigmoidal fashion, with the length of the polymer:

$$\epsilon_0^L(n) = \frac{1}{2}\left(\epsilon_{\max}^L + \epsilon_{\min}^L\right) + \frac{1}{2}\left(\epsilon_{\max}^L - \epsilon_{\min}^L\right)\tanh\left(r_L\left(n-n_L\right)\right) \quad (3)$$

$$\epsilon_0^H(n) = \frac{1}{2}\left(\epsilon_{\max}^H + \epsilon_{\min}^H\right) + \frac{1}{2}\left(\epsilon_{\max}^H - \epsilon_{\min}^H\right)\tanh\left(r_H\left(n-n_H\right)\right) \quad (4)$$

The parameter r_L (r_H) sets the rate at which the mean efficiencies vary between their minimum value of $\epsilon_{\min}^L$ ($\epsilon_{\min}^H$) and their maximum value of $\epsilon_{\max}^L$ ($\epsilon_{\max}^H$) and n_L (n_H) is the length at which the mean efficiency is halfway between its maximum and minimum. This relationship captures the biochemically plausible property that initially the efficiencies increase, on average, only slightly with the length of the polymer. Only when peptides reach lengths sufficient for them to be able to adopt an ordered three-dimensional structures do the average efficiencies start increasing markedly. Then, for even longer polymers, the average efficiencies again stabilize, since gaining additional length no longer produces significant improvement in catalytic properties. The widths of the distributions $\sigma^L(n)$ ($\sigma^H(n)$) are chosen such that probabilities of sampling negative efficiencies are quite small. If such instances occur, the efficiencies are reflected across the origin.

When two peptides are joined together, the catalytic efficiencies of the product of this reaction are related to the efficiencies of the reactants. For example, the product of the addition of a small peptide to a much longer peptide has efficiencies which closely resemble the efficiencies of the longer "parent". To underscore this relationship the model is called an Inherited Efficiencies Model. Statistically, catalytic efficiencies of the product of a ligation reaction are chosen from a conditional probability, $\mathcal{P}_{n,k,l}^L(\epsilon|\epsilon',\epsilon'')$, which gives the probability of creating a peptide of length $n = k + l$ with efficiency ϵ, given peptides of length k and l with efficiencies ϵ' and ϵ'', respectively. Since this probability is a property of the ligation process, the same form is used to assign efficiencies of ligation and hydrolysis to the product. In the present implementation, this probability has the form of a multivariate Gaussian:

$$\mathcal{P}_{n,k,l}^L(\epsilon|\epsilon',\epsilon'') = \frac{n}{\sigma_n\sqrt{4\pi kl}} \times \exp\left[-\frac{n^2}{4kl}\left(\frac{\epsilon-\epsilon_n^0}{\sigma_n} - \frac{k}{n}\frac{\epsilon'-\epsilon_k^0}{\sigma_k} - \frac{l}{n}\frac{\epsilon''-\epsilon_l^0}{\sigma_l}\right)^2\right] \quad (5)$$

Here, to simplify the notation, $\sigma_j = \sigma^L(j)$ ($\sigma^H(j)$) and $\epsilon_j^0 = \epsilon_0^L(j)$ for the ligation (hydrolysis) properties of the substrates or the product.

A similar approach is taken to define a conditional probability for the products of hydrolysis reactions.

However, since hydrolytic enzymes act more efficiently on disordered peptides than on ordered peptides, not all peptide bonds are equally likely to be hydrolyzed. Although our model does not explicitly include the degree of ordering of different polymers, we exploit the relationship between structure and function: without a stable three-dimensional structure, high efficiency protein catalysis is impossible. In the current implementation of the model, the degree of structure of a peptide, s, is computed using:

$$s = \max[\epsilon^L, \epsilon^H]. \tag{6}$$

Clearly, other mappings between efficiency and structure are possible. The bias of hydrolytic enzymes towards disordered peptides is modelled by a decreasing sigmoidal function of structure, $\beta(s)$. As stipulated by the model, this implies that efficient catalysts are less likely to be hydrolyzed than inefficient, presumably disordered, peptides. The maximum value of the bias, almost always equal to unity, will be denoted by $\beta_{\max}$ and the mininum value by $\beta_{\min}$. The degree of structure for which the bias is halfway between its maximum and minimum values will be denoted s_0 and the rate of decrease of the bias will be controlled by a parameter denoted as r_b. When a peptide is hydrolyzed to form two new peptides, the catalytic efficiencies of the "offspring" are, once again, chosen from a conditional probability, $\mathcal{P}^H_{k,l,n}(\epsilon', \epsilon''|\epsilon)$, of creating peptides of lengths k and l, with catalytic efficiencies ϵ' and ϵ'', respectively, from a peptide of length $n = k + l$ with efficiency ϵ. To find the form of this conditional probability we note that the making and breaking of peptide bonds are, in a way, inverses of each other and, therefore, the conditional probabilities describing the properties of the products of ligation and hydrolysis reactions are related by Bayes's Theorem. Considering that in a peptide of length n, $n-1$ bonds can be hydrolized and including the structural bias function $\beta(s)$ we obtain:

$$\mathcal{P}^H_{k,l,n}(\epsilon', \epsilon''|\epsilon) = \frac{\mathcal{P}^L_{n,k,l}(\epsilon|\epsilon', \epsilon'')p_k(\epsilon')p_l(\epsilon'')}{p_n(\epsilon)(n-1)\beta(s)} \tag{7}$$

Evaluating this expression for the specific forms of the probabilities, we obtain:

$$\begin{aligned}\mathcal{P}^H_{k,l,n}(\epsilon', \epsilon''|\epsilon) = {} & \frac{1}{(n-1)\beta(s)} \frac{n}{2\pi\sigma_k\sigma_l\sqrt{2kl}} \\ & \times \exp\Bigg[-\frac{n^2}{4kl}\Bigg(\frac{k(k+2l)}{n^2}\Big(\frac{\epsilon' - \epsilon^0_k}{\sigma_k}\Big)^2 \\ & + \frac{l(2k+l)}{n^2}\Big(\frac{\epsilon'' - \epsilon^0_l}{\sigma_l}\Big)^2 + \frac{k^2+l^2}{n^2}\Big(\frac{\epsilon - \epsilon^0_n}{\sigma_n}\Big)^2 \\ & + \frac{2kl}{n^2}\Big(\frac{\epsilon' - \epsilon^0_k}{\sigma_k}\Big)\Big(\frac{\epsilon'' - \epsilon^0_l}{\sigma_l}\Big) \\ & - 2\Big(\frac{k}{n}\frac{\epsilon' - \epsilon^0_k}{\sigma_k} + \frac{l}{n}\frac{\epsilon'' - \epsilon^0_l}{\sigma_l}\Big)\Big(\frac{\epsilon - \epsilon^0_n}{\sigma_n}\Big)\Bigg)\Bigg]\end{aligned} \tag{8}$$

where s is the degree of structure of the "parent" n-mer.

Simulations of the Inherited Efficiencies Model are carried out using a Monte Carlo method. Each Monte Carlo cycle consists of three stages: (1) the reaction to be performed (ligation or hydrolysis) is chosen, (2) the substrate or substrates are chosen from the list of peptides present in the system, (3) the properties of the product or products of the reaction are sampled from the appropriate distributions and the list of polymers is updated. The number of monomers in the protocell is held fixed to reflect the equilibrium between the concentrations of amino acids inside the protocell and in the environment, facilitated by the permeation properties of the protocellular boundary. The probabilities for the two reaction types are computed from the corresponding total catalytic capabilities of the peptides within the protocell. Once the reaction type is chosen, the probabilities of individual reactions are used to choose the substrate(s) of the reaction. Finally, the properties of the products of the reactions were chosen from the conditional probabilities described above.

3. RESULTS AND DISCUSSION

Several properties of the inherited efficiencies model have been explored *via* Monte Carlo simulation. For a range of parameters, the increase in the catalytic capabilities of the protocell that defines non-genomic evolution has been observed. Here we describe the results of simulations aimed at assessing the role played by the details of the hydrolysis bias and the balance between the ease of creation of efficient ligases and proteases.

The bias in the action of the hydrolytic enzymes towards the destruction of less efficient peptides is expected to play an important role in non-genomic evolution. Several simulations were performed to explore this. In all cases, the number of monomers within the protocell was fixed at 1000, the maximum efficiency means ($\epsilon^L_{\max}$ and $\epsilon^H_{\max}$) were 1000.0, the minimum efficiency means ($\epsilon^L_{\min}$ and $\epsilon^H_{\min}$) were 1.0 and the maximum of the bias ($\beta_{\max}$) was 1.0. The simulations were performed for 2×10^6 Monte Carlo cycles. Variations in the location of the midpoint of the bias (s_0) and of the rate of decrease of the bias (r_b) were not observed to have a qualitative effect on the behavior of the model (data not shown). Changes to the relative depth of the bias had a marked effect, however. The results of three representative simulations, for $\beta_{\min} = 0.05$, 0.025, and 0.01, are shown in Figure 1. In these simulations, the functions governing the means of the efficiency distributions were adjusted to make the formation of ligases slightly easier than the formation of proteases ($n_L = 20$, $r_L = 0.235$, $n_H = 21$, $r_H = 0.230$). As the minimum value of the hydrolysis bias was decreased from 0.1 to 0.001, the behavior of the protocell changed: systems with $\beta_{\min} < 0.01$ exhibited

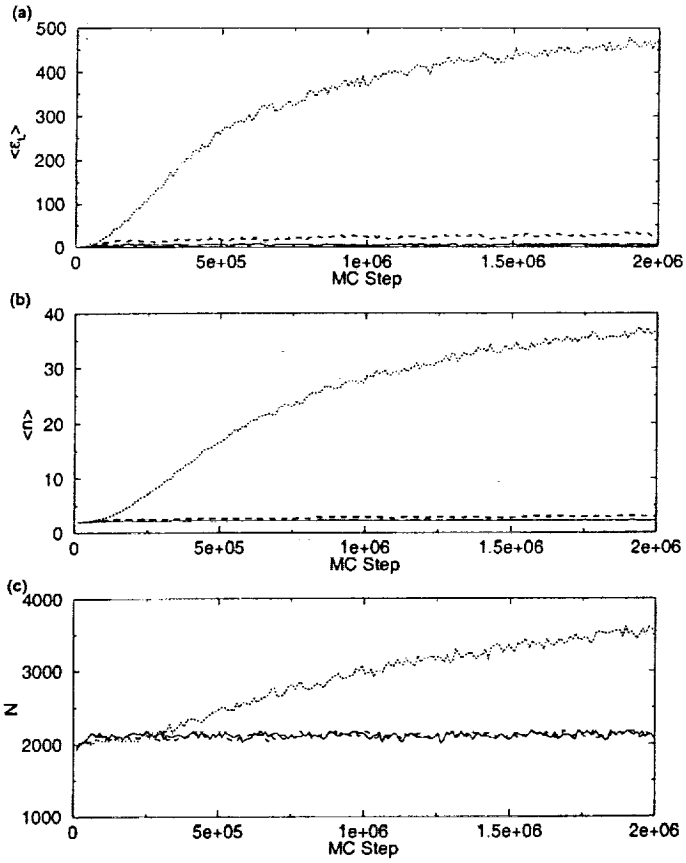

Fig. 1. Results for $\beta_{\min} = 0.05$ (solid lines), 0.025 (dashed lines), and 0.01 (dotted lines). (a) The average ligation efficiency of the polymers in the protocell. (b) The average length of the polymers within the protocell. (c) The number of polymers within the protocell.

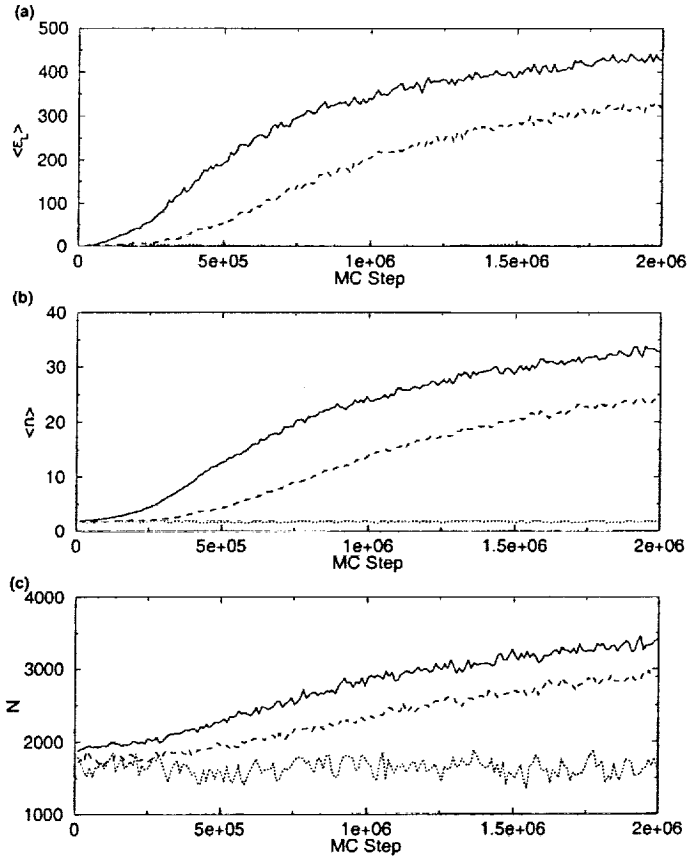

Fig. 2. Results for $n_L = 21$ (solid lines), 25 (dashed lines), and 30 (dotted lines) for $n_H = 20$. (a) The average ligation efficiency of the polymers in the protocell. (b) Average length of the polymers within the protocell. (c) The number of polymers within the protocell.

a large and sustained increase in both the average length and average catalytic efficiencies of the peptides within the protocell. In contrast, systems with $\beta_{\min} > 0.05$ showed little increase in either the average length or the average catalytic efficiencies of their peptides. A single system was simulated with $\beta_{\min} = 0.025$ and it exhibited very slight growth in the length and catalytic efficiencies of its peptides.

Since $\beta_{\min}$ is the value of the hydrolysis bias for highly structured, and therefore highly efficient, peptides, its value determines the "lifespan" of highly efficient peptides. Large values of $\beta_{\min}$ mean that the probability that a highly efficient peptide will be hydrolyzed is not much reduced over the probability that a peptide of average efficiency will be hydrolyzed. Thus, when highly efficient peptides are generated in a system with a large $\beta_{\min}$, they are hydrolyzed before their actions greatly affect the population of peptides within the protocell and the rate with which the protocell explores the space of all peptides is not changed. In contrast, for small values of $\beta_{\min}$, the probability that a highly efficient peptide will be hydrolyzed is much smaller than the probability that a peptide of average efficiency will be hydrolyzed. Thus, highly efficient peptides are long-lived and their presence can increase the rate at which the protocell generates new peptides. The protocell can then evolve non-genomically.

The rate at which novel peptides are generated within a protocell not only depends on the depth of the hydrolysis bias but is also sensitive to the balance between the rates of creation of small, efficient ligases and small, efficient proteases. Clearly, if highly efficient ligases are much more easily formed than efficient proteases, the protocell will fill with a diverse array of long peptides. Eventually, the protocell will burst. At the other extreme, if small, efficient proteases are much more easily formed than small, efficient ligases, the formation of long peptides will proceed slowly and any peptides formed will be hydrolyzed rapidly; the overall catalytic efficiency of the protocell will therefore remain small. A series of simulations were performed to examine the sensitivity of non-genomic evolution to slight imbalances in the ease of creation of ligases and proteases. Particular attention was paid to cases where proteases were slightly easier to produce than ligases. As before, the number of monomers within the protocell was fixed at 1000, the maximum efficiency means ($\epsilon^L_{\max}$ and $\epsilon^H_{\max}$) were 1000.0, the minimum efficiency means ($\epsilon^L_{\min}$ and $\epsilon^H_{\min}$) were 1.0 and the maximum of the bias ($\beta_{\max}$) was 1.0. The minimum of the bias

($\beta_{\min}$) was set to 0.01, the rate of decrease of the bias (r_b) to 0.065 and the midpoint of the bias decrease (s_0) to 58.0. The parameters governing the means of the hydrolysis efficiency distributions were fixed to $n_H = 20$ and $r_H = 0.235$. The parameter governing the rate of change of the means of the ligation efficiency distributions, r_L, was fixed at 0.230 and three values for n_L were considered: 21, 25, and 30. The simulations were performed for 2×10^6 Monte Carlo cycles.

The results of these simulations are displayed in Figure 2. Clearly shown is the sensitive dependence of the rate of evolution on the ease of creation of ligases: as n_L increases, the rate of improvement in the average efficiency and length of the polymers in the protocell decreases. No real improvement is seen when $n_L = 30$. These data demonstrate that the rates with which ligases and proteases are formed must be in a close balance for non-genomic evolution to occur.

4. SUMMARY

The results presented here demonstrate the possibility of a novel mechanism of early protocellular evolution. This mechanism does not require the presence of a genome, nor does it rely on any form of sequence complementarity or the exact replication of protein sequences. In fact, the sloppy replication of protein sequences is an advantage in the earliest phase of evolution because it allows for the rapid exploration of the space of proteins and the discovery of new functions. It is the preservation of these functions and their interrelationships which must be maintained during this early stage of evolution, not the identity of the actors performing those functions. Further, evolution progresses through improvements of the whole community rather than the most fit individuals.

The proposed model makes truly minimal assumptions — the existence of polymers capable of performing constructive and destructive processes and some preference for the destruction of non-functional polymers. This preference, well-motivated by the known biochemistry of protein enzymes, drives the evolution of protocells.

Although specific interactions between peptides are not included here, they can be readily incorporated into the proposed concept of evolution. In fact, there is no conflict between this concept and the work of Ghadiri and Chmielewski. Since non-genomic evolution is necessarily limited by its inability to transfer information sufficiently precisely, specificity of peptide interactions would improve the fidelity of information transfer, hence increasing evolutionary potential of the system. Ultimately, however, a truly advanced protocell would have to find a better method of transferring information to its offsprings.

The model can be naturally extended to include the possibility of producing peptides capable of performing new protocellular functions and to describe growth and division of protocells. Perhaps more importantly, recent advancements that allow the *in vitro* evolution of catalytic peptides [6] provide firm ground for improving the model and testing its predictions experimentally.

REFERENCES

[1] A.J. Hager, J.D. Pollard, and J.W. Szostak: "Ribozymes: Aiming at RNA replication and peptide synthesis." Chemistry and Biology, Vol. 3, pp 717–725, 1996.

[2] G.F. Joyce: "Ribozymes — building the RNA world." Current Biology, Vol. 6, pp 965–967, 1996.

[3] D. H. Lee, K. Severin, Y. Yokobayashi, and M.R. Ghadiri: "Emergence of symbiosis in peptide self-replication through a hypercyclic network." Nature, Vol. 390, pp 591–594, 1997.

[4] D.H. Lee, J.R. Granja, J.A. Martinez, K Severin, and M.R. Ghadiri: "A self-replicating peptide." Nature, Vol. 382, pp 525–528, 1996.

[5] H. J. Morowitz, B. Heinz, and D. W. Deamer: "The chemical logic of a minimum protocell." Origins of Life and Evolution of the Biosphere, Vol. 18, pp 281–287, 1988.

[6] R.W. Roberts and J.W. Szostak: "RNA-peptide fusions for the in vitro selection of peptides and proteins." Proceedings of the National Academy of Science USA, Vol. 94, pp 12297–12302, 1997.

[7] K. Severin, D.H. Lee, J.A. Martinez, M. Vieth, and M.R. Ghadiri: "Dynamic error correction in autocatalytic peptide networks." Angewandte Chemie International Edition, Vol. 37, pp 126–128, 1998.

[8] S. Yao, I. Ghosh, R. Zutshi, and J. Chmielewski: "Selective amplification by auto- and cross-catalysis in a replicating peptide system." Nature, Vol. 396, pp 447–450, 1998.