Catalytic Reaction Sets, Decay, and the Preservation of Information

Wim Hordijk
Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand, wim@santafe.edu

José F. Fontanari
Instituto de Física, Universidade de São Paulo, São Carlos, Brazil, fontanari@if.sc.usp.br

Abstract --- *We study the ability to maintain information in a population of protocells that enclose sets of reacting polymers or metabolisms, under the influence of decay, i.e., spontaneous breakdown of large polymers. At a certain decay rate, it becomes impossible to maintain a significant concentration of large polymers, while it is still possible to maintain sets of smaller polymers that contain the same amount of information. We use a genetic algorithm to evolve reaction sets to generate specific polymer distributions under the influence of decay. In these evolved reaction sets, the beginnings of hypercycle-type structures can be observed, which are believed to have been an important step toward the evolution of the first living cells.*

1. INTRODUCTION

The current paradigm for the origin of life is the nucleic acids-first hypothesis – so called RNA world [1] – which speculates on the appearance, via spontaneous generation, of RNA-like polymers capable of self-replication, which later evolved the capacity to encode proteins. The theoretical consequences of this hypothesis were fully investigated by Eigen and co-workers in the 1970's (see, e.g., [2] and [3]). In this framework the information is carried by the self-replicating polymer itself, as in today's nucleic acids, with the amount of information being limited by the accuracy of the replication mechanism.

The overwhelming numerical improbability of the random assembly of the first replicator, however, points to a second, comparatively much less explored alternative - the proteins-first assumption [4], [5]. This is the realm of the protein-made metabolisms or autocatalytic sets of proteins, the emergence of which, out of a sufficiently large ensemble of randomly interacting peptides, is almost certain [6]. Here, the information is encoded in the types of molecules that comprise the reaction set as well as in the way they interact with each other. A metabolism takes in raw material from its environment (food), and transforms it into the forming units, thus perpetuating itself [7]. Despite claims that a metabolism may have the ability to "self-reproduce", in the sense that it can seed future generations, this capacity has not been implemented in the models proposed so far. In fact, the seeding or budding ability should be allowed only as an emergent property of the metabolism, and not as an ad hoc postulate. Metabolisms seem incapable of evolving under natural selection since they do not replicate, and overcoming this difficulty is a major challenge the proponents of the proteins-first assumption must face.

We do not attempt to tackle any of the above-mentioned long standing puzzles on the origin of life; rather, in this contribution we consider a likely subsequent scenario in which metabolisms and information-bearing polymers coexist, the latter being outcomes of the metabolism. Such a scenario is compatible with Dyson's double-origin theory [5], which assumes a genetic take-over of polynucleotides after a period of parasitic or symbiotic coexistence with metabolisms. Thus, in this sophisticated setting we address the very same issues of preservation of information investigated by Eigen and co-workers in the much simpler context of template replication.

The amount of information that can be stored and preserved in a population of reacting polymers depends crucially on the reaction efficiencies. For example, larger polymers are more likely to break down into smaller parts than shorter polymers. So, to maintain a significant concentration of a certain large polymer, there have to exist highly efficient reactions building up these large polymers from smaller ones. In fact, the efficiencies (or reaction rates) of these reactions have to be larger than the rate of polymer breakdown. This is somewhat equivalent to the error threshold phenomenon in self-replicating polynucleotides [2]. One proposed solution to circumvent the error catastrophe is the *hypercycle* [3], a catalytic feedback set where each polymer increases the efficiency of the creation of the next polymer in a (closed) reaction loop. This way, parts of the information can be stored in smaller polymers that help each other in maintaining a large enough concentration of each of them. The set as a whole can preserve the complete information, whereas one large polymer could not [8].

Here, we use a genetic algorithm (GA) to evolve catalytic reaction sets to generate a certain target distribution of polymers under the absence or existence of polymer decay (i.e., breakdown of large polymers). Our approach draws

heavily on the paper of Lohn et al [9], but has some additional features and a more realistic method of simulating polymer reactions. We then look at the differences between a target consisting of one large polymer and a target of three smaller ones, the lengths of which add up to that of the larger one, and if (or how) each target can be reached and maintained under both the absence and the existence of decay. We note that each individual in the GA implementation can be viewed as a protocell or vesicle that encloses the reaction set; the fact that the protocell is automatically endowed with an ad hoc replication ability is an unjustifiable weak point of our model (see [10] for a similar approach).

In the next section, a method for simulating simple chemical reactions on a computer is reviewed. In section 3 the model for evolving chemical reaction sets is explained. Section 4 then presents the results of this model comparing different target polymer distributions under the presence or absence of decay. Finally, section 5 summarizes the main conclusions.

2. SIMULATING CHEMICAL REACTIONS

The model we use considers simple polymers made up of only one type of molecule, and the types of interactions that are possible are bonding and breaking. The main characteristic of a polymer is its length, or the number of molecules in the polymer chain. Polymers of length i are denoted P_i. We restrict the length of polymers to a maximum of 35. The bonding reaction simply "glues" two polymers of lengths i and j together into one polymer of length $i+j$ (provided that $i+j \le 35$). The breaking reaction takes a polymer of length k and splits it into two polymers of lengths i and j where $i+j=k$. However, only catalytic reactions are considered. In other words, a reaction can only happen under the influence of an additional polymer that catalyzes the reaction but which is not involved in the reaction otherwise. A catalyzed bonding reaction is written as $P_i + P_j + P_k \rightarrow P_{i+j} + P_k$, where $i+j \le 35$. The catalyst P_k is not involved in the reaction itself, so it appears again on the right side as a reaction product. A catalyzed breaking reaction is written as $P_{i+j} + P_k \rightarrow P_i + P_j + P_k$, again with $i+j \le 35$ and P_k being the catalyst.

Suppose we have a mass conserving, well-stirred reactor that contains a large number of polymers. Reactions between polymers happen in this reactor based on the concentrations of the reactants (and catalysts) in the reactor. Usually, such a system is modeled with a set of coupled ordinary differential equations (ODE's), one equation for each type of polymer. However, such a system of ODE's quickly becomes analytically unsolvable or numerically cumbersome. An efficient method for numerically simulating such chemical reactions using a stochastic algorithm was proposed by Gillespie [11]. Instead of calculating changes in polymer concentrations over very small time steps (the ODE approach), Gillespie's algorithm is based on deriving a reaction probability density function (pdf) $P(\tau, m)d\tau$ that yields the probability at time t that the *next* reaction in the reactor will occur in the time interval $(t+\tau, t+\tau+d\tau)$ *and* will be of type m (given a certain number M of possible reactions). This pdf has certain parameters, the values of which depend on the current polymer concentrations in the reactor. The method then uses a Monte Carlo procedure to generate a stream of random numbers that are interpreted as reaction times and types, and the parameter values of the pdf are updated after every reaction to reflect the new polymer concentrations. In this simulation method, there is also a parameter c_i, the reaction efficiency, for each of the M reactions. In our model, we use the same value for c_i for each reaction (i.e., there is no difference in efficiencies for the different reaction types).

So, to summarize, we have a set of $N=35$ polymer types P_i, and a set of M possible reactions R_i where each R_i is a catalyzed bonding or breaking reaction. The algorithm for simulating this polymer reaction system is then:

1. Set the current time $t=0$, generate an initial polymer type distribution, and calculate the reaction pdf parameters based on this initial distribution. Set a "stopping" time T.
2. Generate a random pair (τ, m) from the reaction pdf $P(\tau, m)$.
3. Set the time $t \rightarrow t+\tau$ and perform reaction R_m. Update the polymer type concentrations and the reaction pdf parameters according to R_m.
4. If $t \ge T$, stop. Otherwise, go to step 2.

3. EVOLVING CHEMICAL REACTION SETS

Following Lohn et al [9] we use a genetic algorithm [12] to evolve a population of reaction sets towards a prespecified polymer distribution given some initial distribution. This approach is in stark contrast with previous work on autocatalytic sets that focused on the spontaneous emergence of metabolisms, where no targets were available [6], [7].

The population in our GA consists of reaction sets $\Re$, where each reaction set contains 100 reactions R_i, which are catalyzed bonding or breaking reactions. So, each R_i is of the form $P_i + P_j + P_k \rightarrow P_{i+j} + P_k$ (bonding), or $P_{i+j} + P_k \rightarrow P_i + P_j + P_k$ (breaking). We use a population size $S=100$, and the initial population is created at random, where the fraction of breaking reactions in each

reaction set is a parameter of the algorithm (usually set to 0.2).

The genetic operators are implemented as follows. In crossover, two "parent" reaction sets $\Re_{p1}$ and $\Re_{p2}$ are taken from the mating pool, and a random number c (the crossover point) between 1 and 100 is drawn from the uniform distribution. The first child, $\Re_{c1}$, is then formed by combining the first c reactions R_i from the first parent, $\Re_{p1}$, with the last $100-c$ reactions from the second parent, $\Re_{p2}$. The second child, $\Re_{c2}$, is formed in a similar way but with the opposite parts of the parents. The mutation operator simply replaces a reaction in a reaction set with a randomly chosen new reaction R_i (independent of the reaction being replaced). For selection, the standard roulette wheel selection method is used [12].

The fitness function of the GA is implemented as follows. Given an individual $\Re$ from the GA population and an initial polymer distribution, use the stochastic simulation method as described in the previous section to iterate this reaction set for T time units (in most runs we used $T=100$, and the values for the reaction efficiencies were set so that, at least initially, there are about 100 reactions performed in one time unit). Continue iterating the stochastic simulation method for another T time units, but after each time unit calculate a "target value" v_t. At the end, take the average of all target values and return that as the fitness value, i.e., the fitness of a reaction set $\Re$ is $f_{\Re} = \sum_{t=T+1}^{2T} v_t / T$.

In our experiments, we used two different ways of calculating the target values v_t. The first one is $v_t = n_{35}$, which simply means the number of polymers of length 35 (the maximum length) in the polymer population at time step t. So, for this target, the fitness of an individual is the number of polymers of maximum length, averaged over the second set of T time units. The second way of measuring the target values is

$$v_t = n_{10} + n_{12} + n_{13} - |n_{10} - n_{12}| - |n_{12} - n_{13}|.$$

So, for this target we try to get as many polymers of lengths 10, 12, and 13, but in roughly equal numbers (and again averaged over the second set of time steps). Note that the lengths of these polymers add up to 35, and indeed the main idea behind this target is to try to get the same "information" as in the first target, but split up in smaller pieces.

Finally, an element of spontaneous polymer breakdown, or decay, is added. In the stochastic simulation method, next to the set $\Re$ of reactions that forms an individual in the GA population, there is an independent set of decay reactions $P_{i+j} \rightarrow P_i + P_j$, which are not catalyzed. The reaction efficiencies of these reactions depend on the length of the polymer that is being broken down. In our simulations, the efficiency of a decay reaction is some constant d times the square of the length of the polymer ($i+j$). The constant d is another parameter in the GA, and can be set to 0 to turn decay off completely, or increased in value for increasing decay rates.

With this model setup, we can now study the differences between the two different targets (polymers of maximum length 35, or polymers of lengths 10, 12, and 13) under the influence (or absence) of decay.

4. RESULTS

Several GA runs were performed using the two different targets, both with and without decay. In this section, the main results of these different runs are presented. In the fitness calculations, an initial polymer distribution (the food set) as shown in figure 1 was used. In this initial distribution, there are 195 polymers each for the polymer types (lengths) 1 to 9, and 0 polymers of any other length. So, for both targets, there do not yet exist any target polymers in the initial polymer population.

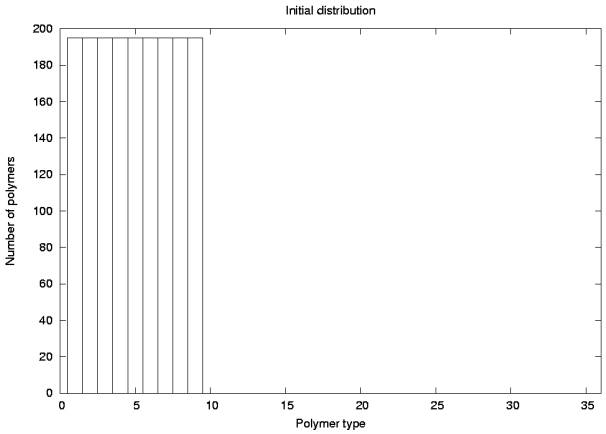

Figure 1 – The initial polymer distribution

First, the GA was run on the first target (polymers of length 35) without any decay (i.e., the decay parameter d was set to 0.0). In every run the GA was able to find reaction sets that produce around 200 polymers of this length. Figure 2 shows a typical result. Even though in the GA runs the reaction simulations were run up to $T=200$, all the results shown in this section are for $T=1000$. For example, in figure 2, one of the best individuals found by the GA was taken, and this reaction set was then re-iterated for $T=1000$ time steps (starting with the same initial distribution as shown in figure 1) to make sure that some equilibrium has been reached. In fact, this is a major progress with respect to the results of Ref. [9], since the target distribution in their case did not coincide with the steady-state polymer distribution. As the figure shows,

this reaction set produces slightly more than 200 polymers of length 35.

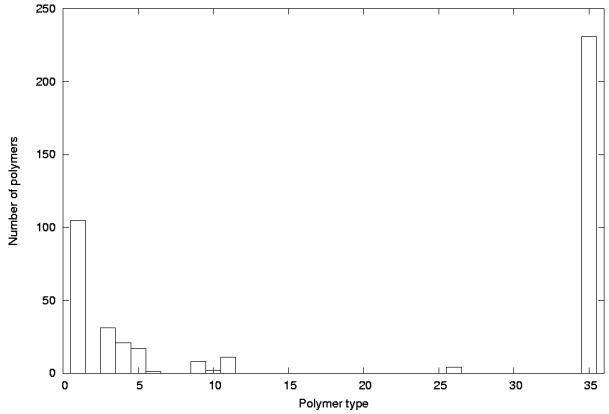

Figure 2 – First target, no decay

Although this reaction set was evolved without using decay, we can ask how it performs when iterated with decay turned on. Figure 3 shows the equilibrium distribution (again at $T = 1000$) of the same reaction set, but with the decay parameter set to $d = 0.0001$. In this case, it produces less than 20 polymers of maximum length, more than one order of magnitude less compared to the no-decay case.

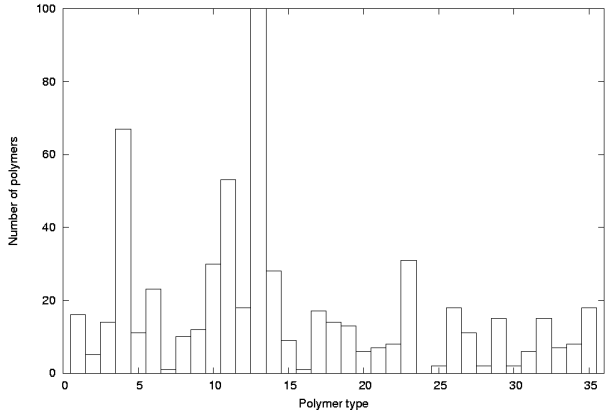

Figure 3 – First target, no decay, but iterated with decay

Of course this reaction set was not evolved to deal with decay, and so it is expected to perform poorly under the influence of decay. Next, the GA was run on the same target, but this time with decay (again with $d = 0.0001$). Figure 4 shows the equilibrium distribution of one of the best reaction sets found by the GA in this case. As the figure shows, even though the reaction set was evolved under the influence of decay, it still only manages to produce around 35 polymers of maximum length. This is slightly more than for the reaction set that was evolved for no-decay, but still significantly less than the more than 200 that can be reached without decay at all. So, apparently the decay in this case is too high to maintain a large enough number of maximum-length polymers, and the relevant "information" is lost or at least significantly reduced.

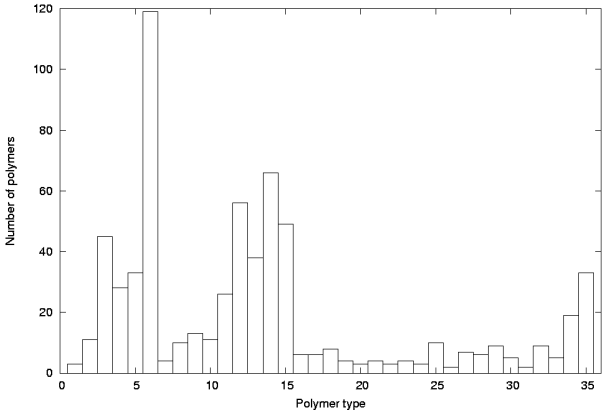

Figure 4 – First target, with decay

Next, the second target, polymers of lengths 10, 12, and 13, is investigated. The GA was run several times on this target, with the decay parameter at $d = 0.0001$. Figure 5 shows one of the best reaction sets evolved for this target. As the plot shows, it manages to produce around 200 polymers of each length, roughly equal to the amount of polymers of maximum length that can be produced without decay. So, even though the relevant "information" cannot be maintained in one long polymer under the influence of decay, it can be maintained by dividing the information up over smaller polymers. The information can be maintained at a similar level (around 200 polymers) using these smaller polymers.

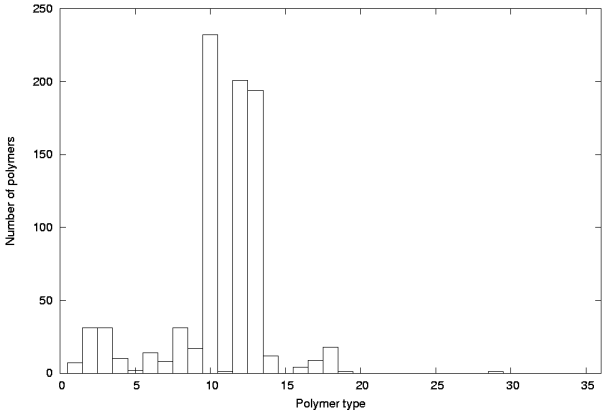

Figure 5 – Second target, with decay

It turns out that the performance of this particular reaction set is slightly less when iterated without decay. Figure 6 shows the equilibrium distribution in this case. This particular reaction set actually relies on the decay to break down longer polymers into smaller ones, which it can then use to create the target polymers. Without this breakdown, there are fewer smaller polymers available to create the targets ones, resulting in a somewhat lower production of target polymers.

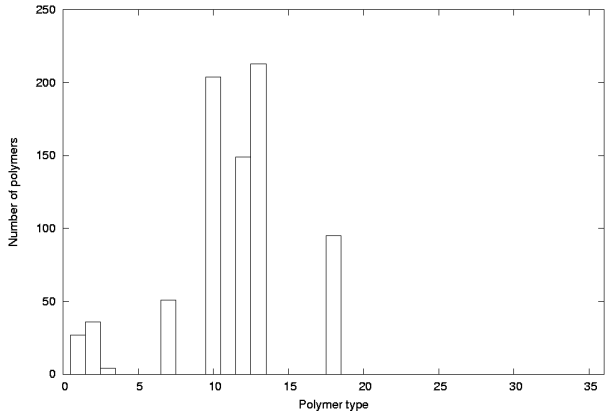

Figure 6 – Second target, with decay, but iterated without decay

On investigating the evolved reaction set, it turns out that there is a core set of only 13 reactions (out of the 100) that are mainly responsible for its performance. When isolating these 13 reactions, and iterating this core set on the same initial polymer distribution (from figure 1) and with the same decay rate ($d = 0.0001$), the equilibrium distribution is as shown in figure 7. The total number of target polymers produced is slightly less than with the complete set of 100 reactions, but is still around 200 each. So, the other 87 reactions only slightly increase the performance of this core set.

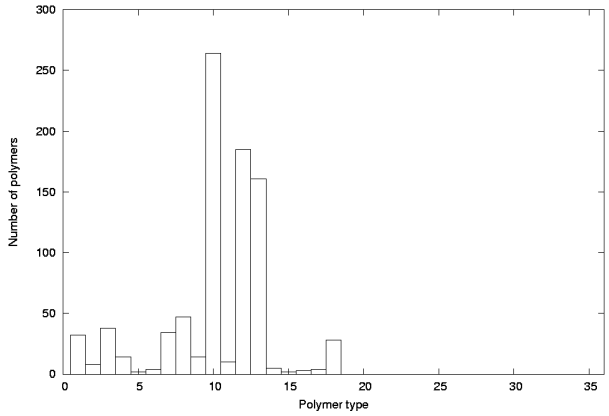

Figure 7 – Equilibrium distribution of the core set

Figure 8 shows the reaction graph of this core set of 13 reactions. The numbers indicate the polymer types (or lengths) and the black dots indicate reactions. The arrows going from polymers to reactions indicate the reactants going into the reaction, and the arrows going from reactions to polymers indicate the products coming out of the reaction. The gray arrows indicate the catalyst of a reaction. Table 1 lists the 13 core reactions.

As can be seen in the reaction graph, there are various "hypercycle-like" structures. For example, polymer type 13, one of the targets, serves as a catalyst in 6 different reactions, 3 of which produce target polymers. There are also several autocatalytic reactions, where the reaction product catalyzes its own creation (such as in 6+7+13→13+13, and 9+9+18→18+18). Furthermore, there are several closed loops in the graph, where the polymer types in this loop act alternately as reactants or catalysts and products. For example, 4+4+13→8+13, 8+13→2+6+13, and 6+4→2+4+4 is such a loop, and there are several more.

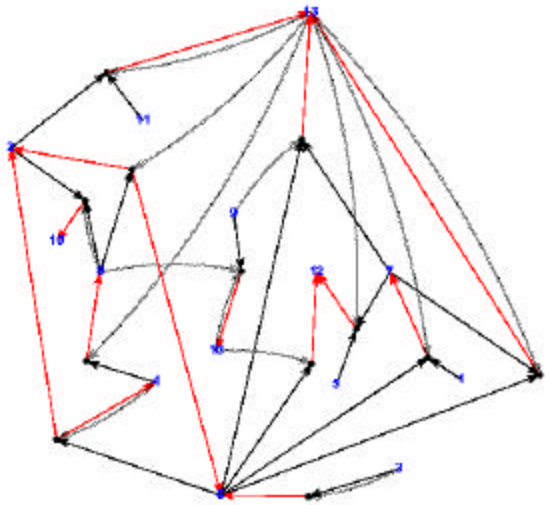

Figure 8 – The reaction graph of the core set

Table 1 – The 13 core reactions of the evolved set.

2+8+8 → 10+8	8+13 → 2+6+13
5+7+13 → 12+13	3+3+3 → 6+3
6+6+18 → 12+18	1+6+13 → 7+13
6+7+9 → 13+9	6+4 → 2+4+4
6+7+13 → 13+13	9+9+8 → 18+8
2+11+13 → 13+13	9+9+18 → 18+18
4+4+13 → 8+13	

One thing to note is that some polymer types in this reaction graph are not directly produced by any of the 13 reactions in the core set. For example, polymer types 11, 9, 4, 2 and some others are only used as reactants or catalysts. However, the core set relies on decay to produce these polymer types, by for example breaking down a polymer of length 13 into polymers of lengths 11 and 2, or 9 and 4, etc. So, instead of being hindered by decay, this reaction set has adapted to actually make good use of the existence of decay!

Results on other GA runs were similar, but often with slightly lower performances of the evolved reactions sets, or somewhat larger core sets. The result shown here was the best one found among the different runs.

5. CONCLUSIONS

The amount of information that can be maintained in a population of reacting polymers depends on the reaction efficiencies and the decay rate. For example, above a certain decay rate, it seems not possible anymore to maintain a significant number of large polymers. However, as shown here, it is possible to evolve reaction sets that are able to maintain the relevant information by using a set of smaller polymers, each of which holds only part of the information (in our case, the lengths of the smaller polymers add up to the length of the large one, but one can imagine encoding information in different and more sophisticated ways in polymers of different types and lengths).

So, whereas maintaining a certain amount of information in one large polymer breaks down at a certain decay rate, splitting the information up over several smaller polymers makes it possible to maintain the same amount of information (around 200 polymers of each type, in our case). In fact, the evolved reaction sets actually learn to make use of the decay by eliminating the use of reactions that create smaller polymers that can be used in building up the target ones. These evolved reaction sets rely on the decay to create these smaller polymers. This gives rise to relatively small core sets of reactions that are highly efficient and sufficient to reach the desired target polymer distribution.

Moreover, in these core sets the beginnings of hypercycle-type structures can be observed in the form of target polymers acting as catalysts, the existence of autocatalytic reactions, and several closed loops in the reaction graph. These results can also bear on other, more general questions relating to, e.g., the origin of life, where it is believed that hypercycle-type structures were an important step in achieving the complexity necessary to support living cells. The results presented here clearly show that it is indeed possible to evolve hypercycle-type structures to maintain a certain amount of information under the influence of decay, or polymer breakdown. This paper mainly presents work in progress, but our results so far are very encouraging, and demand further investigation into this phenomenon.

ACKNOWLEDGMENTS

This research was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo under project number 99/09644-9. WH also thanks the Allan Wilson Centre for Molecular Ecology and Evolution for partial funding while finishing this paper.

REFERENCES

[1] G. F. Joyce, " RNA evolution and the origins of life," *Nature* **338**, 217-224, 1989.

[2] M. Eigen, "Selforganisation of matter and the evolution of biological macromolecules," *Naturwissenchaften* **58**, 465-522, 1971.

[3] M. Eigen and P. Schuster, *The Hypercycle*, Berlin: Springer-Verlag, 1979.

[4] R. Shapiro, *Origins: A Skeptic's Guide to the Creation of Life on Earth*, New York: Bantam, 1987.

[5] F. Dyson, *Origins of Life*, Cambrigde: Cambridge University Press, 1985.

[6] S. A. Kauffman, "Autocatalytic Sets of Proteins," *Journal of Theoretical Biology* **119**, 1-24, 1986.

[7] R. J. Bagley and J. D. Farmer, "Spontaneous Emergence of a Metabolism," *Artificial Life II, SFI Studies in the Sciences of Complexity, vol. X, 93-140, Addison-Wesley, 1991.*

[8] J. Maynard Smith, "Hypercycles and the origin of life," *Nature* **280**, 445-446, 1979.

[9] J. D. Lohn, S. P. Colombano, J. Scargle, D. Stassinopoulos and G. L. Haith, "Evolving Catalytic Reaction Sets Using Genetic Algorithms," *IEEE International Conference on Evolutionary Computation*, 487-492, 1998.

[10] D. Segré, Y. Pilpel and D. Lancet, "Mutual catalysis in sets of prebiotic organic molecules: Evolution through computer simulated chemical kinetics," *Physica A* **249**, 558-564, 1998.

[11] D. T. Gillespie, "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions," *Journal of Computational Physics* **22**, 403-434, 1976.

[12] M. Mitchell, *An Introduction to Genetic Algorithms*, Cambridge: MIT Press, 1996.