J. theor. Biol. (2002) **217,** 167–181
doi:10.1006/yjtbi.3026, available online at http://www.idealibrary.com on IDEAL®

"Living" Under the Challenge of Information Decay: The Stochastic Corrector Model vs. Hypercycles

ELIAS ZINTZARAS†‡, MAURO SANTOS†§ AND EÖRS SZATHMÁRY*†‖

†*Collegium Budapest*, *Institute for Advanced Study*, *Szentháromság u.* 2, *H*-1014, *Budapest*, *Hungary*, ‡*National Agricultural Research Foundation* (*N.AG.RE.F.*), 19 *Egialias & Chalepa*, 15125 *Maroussi*, *Greece*, §*Departament de Genètica i de Microbiologia*, *Grup de Biologia Evolutiva*, *Universitat Autònoma de Barcelona*, 08193 *Bellaterra*, *Barcelona*, *Spain and* ‖*Department of Plant Taxonomy and Ecology*, *Eötvös University*, 2 *Ludovika tér*, *H*-1083, *Budapest*, *Hungary*

(*Received on* 6 *August* 2001, *Accepted in revised form on* 11 *February* 2002)

The combined problem of having a large genome size when the accuracy of replication was a limiting factor is probably the most difficult transition to explain at the late stages of RNA world. One solution has been to suggest the existence of a cyclically coupled system of autocatalytic and cross-catalytic molecular mutualists, where each member helps the following member and receives help from the preceding one (i.e., a "hypercycle"). However, such a system is evolutionarily unstable when mutations are taken into account because it lacks individuality. In time, the cooperating networks of genes should have been encapsulated in a cell-like structure. But once the cell was invented, it closely aligned genes' common interests and helped to reduce gene selfishness, so there was no need for hypercycles. A simple package of competing genes, described by the "stochastic corrector model" (SCM), could have provided the solution. Until now, there is no clear demonstration that the proposed mechanisms (compartmentalized hypercycles and the stochastic corrector model) do in fact solve the error threshold problem. Here, we present a Monte Carlo model to test the viability of protocell populations that enclose a hypercyclic (HPC) or a non-hypercyclic (SCM) system when faced with realistic mutation rates before the evolution of efficient enzymic machinery for replication. The numerical results indicate that both systems are efficient information integrators and are able to overcome the danger of information decay in the absence of accurate replication. However, a population of SCM protocells can tolerate higher deleterious mutation rates and reaches an equilibrium mutational load lower than that in a population of HPC protocells.

Introduction

The quasispecies concept of Eigen & Schuster (1979; see also Szathmáry, 1989a; Eigen, 1992; Kauffman, 1993; Maynard Smith & Szathmáry, 1995), which came up after modelling Darwinian evolution in a population of replicating molecules, shows how the accuracy of replication in

*Corresponding author. Collegium Budapest, Institute for Advanced Study, Szentháromság u. 2, H-1014, Budapest, Hungary. Tel.: +36-1-224-83-31; fax: +36-1-224-83-10.

E-mail addresses: zintza@nagref.gr (E. Zintzaras), mauro.santos@uab.es (M. Santos), szathmary@colbud.hu (E. Szathmáry).

0022-5193/02/$35.00/0

the early RNA world (Bartel & Unrau, 1999) placed limits on the size of the genome that could be maintained by selection. It is assumed that at that stage RNA-like polymers (templates) were replicated by a primitive mechanism involving complementary base pairing. Because in the absence of enzymes the average copying fidelity per nucleotide was probably between 0.99 and 0.90 (Friedberg *et al.*, 1995; see Johnston *et al.*, 2001), the size limit might well have been less than 100 bases. Therefore, the danger of information decay due to a relatively fast mutation accumulation was probably a major problem faced by early adaptive systems in order to avoid an "error catastrophe" (Eigen, 1992; Kauffman, 1993; Maynard Smith & Szathmáry, 1995). Since the minimum size of the genome of the last ribo-organism could have been at least 10 000–15 000 bases or bp (Jeffares *et al.*, 1998), elucidating how such a putative creature could have arisen by natural selection constitutes a major scientific challenge.

As a way out to the error catastrophe problem, Eigen & Schuster (1979) offered the "hypercycle" (HPC) as a model of information integration in which an arbitrary number of replicators are linked together in a catalytic loop. It is important to realize that the HPC is a doubly autocatalytic system. First, each member serves as a template that can catalyse its own replication and, second, each member receives help, through a replicase activity, from the preceding member of the cycle. Replication of each member thus depends on the product of its own concentration and that of the preceding one. Problems with the HPC arise, however, when mutations are taken into account. In spatially homogeneous settings HPCs would be eventually destroyed if we allow for selfish mutants (parasites) that are better targets for replication (Maynard Smith, 1979; Bresch *et al.*, 1980). Furthermore, if a mutant arises that happens to be a better replicase, HPCs will not evolve into more efficient systems because they are not evolutionary units (Bresch *et al.*, 1980; Szathmáry, 1989a). In other words, adaptive evolution requires the package of transmissible information for advantageous mutations in order not to aid less-efficient copies of the gene. Encapsulation of HPCs into compartments or cell-like structures (protocells) is thus a requisite for the evolution to continue, as well as a partial solution for parasites to be selected against at the compartment level (Michod, 1983; Eigen *et al.*, 1981). But the question naturally arises: was the package of a truly hypercyclic system into compartments a necessary intermediate stage of evolution? The answer is that we do not know with certainty, but more economic alternative systems such as the "stochastic corrector model" (SCM) (Szathmáry & Demeter, 1987; Szathmáry, 1989b; Grey *et al.*, 1995), which describes the dynamics of genes encapsulated in a reproductive protocell, could have fulfilled the same role.

A major difference between the SCM and HPC is that in the former system there is no hypercyclic coupling among the templates because they are replicated by a non-specific replicase. Within each compartment, the templates are free to compete, because they can reap the benefits of a common metabolism differently. The SCM assumes that there is an optimal template composition that ensures the fastest growth and division of protocells. The behavior of the system depends on two types of stochasticity: (i) replication of templates within protocells, and (ii) random assortment of templates into offspring protocells. Even though templates compete within compartments, selection on stochastically produced offspring variants ("between-protocell selection") can rescue the population from extinction, which reaches equilibrium with a constant frequency of the optimal protocell. The model obviously assumes that the number of different template types per protocell must be small; otherwise there is an increasing risk for a daughter cell to lack any essential gene and not to be viable. Although the SCM could efficiently integrate genetic information there is no sound comparative study showing its competitive superiority vs. that of compartmentalized HPCs (Szathmáry, 1989a).

Additionally, even though cellularization offers the most natural and efficient resolution to Eigen's error catastrophe, replication of RNA-like templates was not yet accurate and protocell populations could be liable to collapse due to a high input of deleterious mutations. The unit of selection here is the entire package of genes (i.e.,

the protocell), and to test the relative merits of HPC- and SCM protocells we should also consider whether or not the conditions under which they may survive are quantitatively consistent with reasonable assumptions on the mutation rates of early replication.

Compartmentalized Hypercycles vs. Stochastic Corrector Model under Deleterious Mutational Pressure

We developed a Monte Carlo model that examines the viability of ''conceptually analogous'' HPC and SCM systems. The simulation programs were independently written in FTN77 (1997) and MATLAB (1999), and the routines were repeatedly checked against any possible discrepancy in the outcomes. All numerical results in the present work are from the MATLAB version.

The basic ''biological'' unit in our model is a protocell with two different templates that stand for metabolic genes (M; *sensu* Gánti, 1975) essential for survivorship and are replicated by genes (R) with a replicase function (Fig. 1). These templates may differ in their rate of replication within a cell (within-compartment selection), and in their contribution to cell growth and division (between-compartment selection). The general features of the model are as follows. At generation t_0, a population is started with K (set to 150 for computational efficiency) identical protocells with n templates (genes). Each template is assumed to consist of three mutable sites (nucleotides). A protocell is randomly chosen according to its relative fitness (see below) for template replication with a deleterious mutation rate per nucleotide per gene equal to u. At this stage, we have defined two main versions of the simulation program: the ''discrete case'' and the ''continuous case''. In the discrete version the cell templates are randomly replicated according to their probabilities of

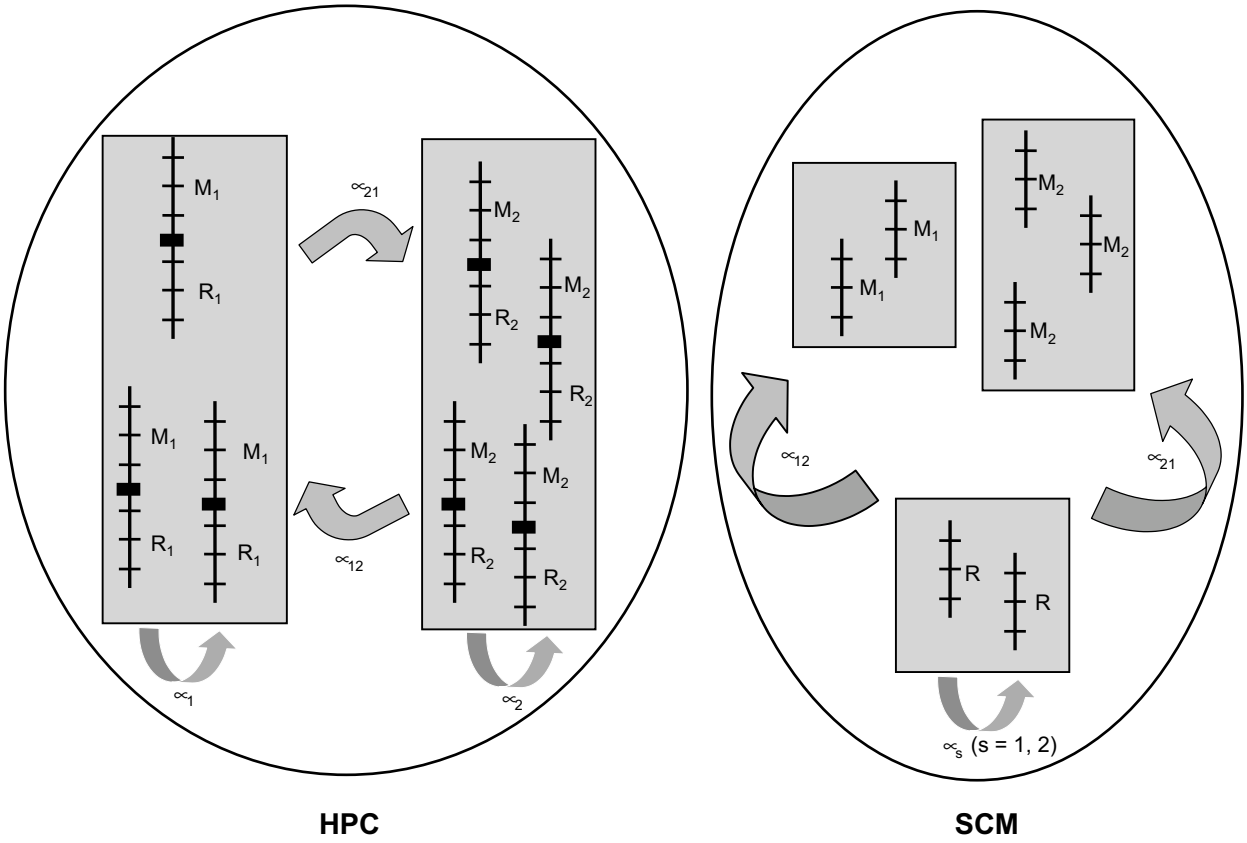


FIG. 1. Schematic representation of compartments (protocells) that integrate genetic information via the hypercycle (HPC) or the stochastic corrector model (SCM). M stands for metabolic templates (genes) that make a direct contribution to protocell fitness, R stands for replicase templates, and μ's for replication rates (see text for details).

replication (see below) until their number reaches $2n$, then the protocell randomly assorts the $2n$ genes into two offspring protocells. In the continuous version one randomly chosen template is replicated according to its probability of replication and the protocell is then turned back to the population whenever its template number is less than $2n$, otherwise it divides as before. In both cases the procedure continues until the population size increases to $2K$, then half of the cells are discarded at random and the start of a new generation is assumed. In all runs the process was continued for up to 300 generations. However, for relatively large mutation rates the problem with the simulations is an intolerable waiting time before a chosen protocell has a relative fitness large enough for template replication and division. Therefore, for any particular. simulation we reasonably assumed that the population collapses if the average fitness of protocells was lower than 0.05. At the protocell level, the fitness function we used was

$$w = \frac{\sum_{i=1}^{d}(1/g_i)}{\sum_{i=1}^{d}(1/A_i)}, \tag{1}$$

where d is the number of different metabolic gene types, g_i is the number of copies of the i-th metabolic gene type, and $A_i = \sum_{j=1}^{g_i} A_{ij}$ is the total contribution of all its j variants to protocell's fitness, where

$$A_{ij} = \frac{(1/\mathrm{e}^{(\upsilon - c_{ij})^2} - 1/\mathrm{e}^{(\upsilon^2)})}{(1 - 1/\mathrm{e}^{(\upsilon^2)}).}, \tag{2}$$

Here υ is the number of nucleotides of gene i, and c_{ij} is the number of correct (wild type) nucleotides. The fitness of the protocell exponentially decreases from $w_{max} = 1$ to 0 depending on the number of mutant nucleotides per gene (Fig. 2).

To simulate the HPC, we assumed that each member of the cycle is a template with two different joined parts (genes), one functions as a specific replicase (R) and the other as a metabolic gene (M) that is a target for replication (see Fig. 1). This follows the logic of Eigen & Schuster (1979) who suggested that the member replicators in the hypercycle should carry out complementary (e.g. metabolic) functions in a

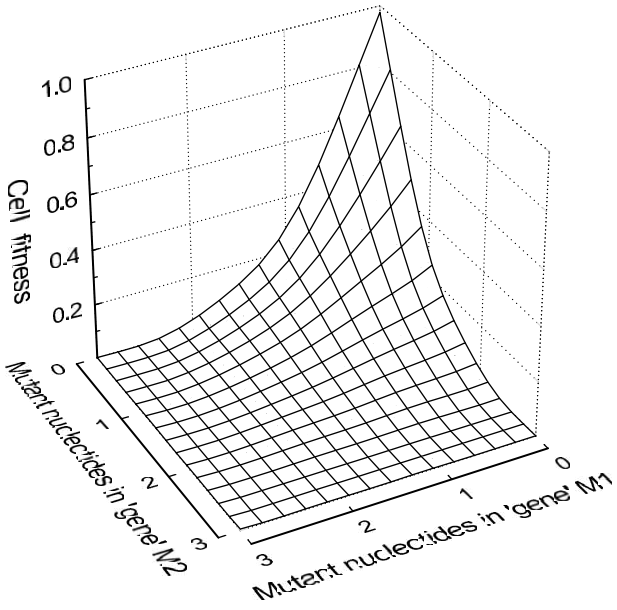


FIG. 2. Fitness function of a protocell according to the number of deleterious mutant nucleotides in metabolic genes essential for survivorship.

fully developed (and compartmentalized) system. Two types of templates are present, M_1R_1 and M_2R_2. At generation t_0 all protocells start with 15 copies of each template, i.e. a total of 60 genes. After a particular template replicase ($R.$) is randomly chosen (the dot stands for 1, 2), it will bind a template ($M.R.$) also chosen at random and eventually replicates it in accordance with its target affinity toward the replicase (because the replicase and the target are two physically independent molecules, the $M.R.$ templates were chosen without replacement). Replication occurs serially along the template and can take place in either two ways: self-replication of templates by their specific replicase, or replication by a different replicase. Their respective dynamics are

$$\frac{\mathrm{d}(M_{i(k)}|R_{i(l)})}{\mathrm{d}t} = \mu_{i(k,l)} M_{i(k)} R_{i(l)},$$

$$i = 1, 2, \quad k = 0; \ldots, 3, \quad l = 0; \ldots, 3,$$

$$\frac{\mathrm{d}(M_{i(k)}|R_{j(l)})}{\mathrm{d}t} = \mu_{i,j(k,l)} M_{i(k)} R_{j(l)}, \quad i \neq j, \quad j = 1, 2, \tag{3}$$

where the μ's are the replication rates determined by the entries in (k,l) matrices (k,l are the four mutational states; see below). The probabilities of replicating a template depend on the previous growth rates and are calculated as follows:

$$P(M_{i(k)}|R_{i(l)}) = \frac{\mathrm{d}(M_{i(k)}|R_{i(l)})/\mathrm{d}t}{\sum_{i,j,k,l}\mathrm{d}(M_{i(k)}|R_{j(l)})/\mathrm{d}t},$$

$$P(M_{i(k)}|R_{j(l)}) = \frac{\mathrm{d}(M_{i(k)}|R_{j(l)})/\mathrm{d}t}{\sum_{i,j,k,l}\mathrm{d}(M_{i(k)}|R_{j(l)})/\mathrm{d}t}. \quad (4)$$

Therefore, replication of each member depends on its own concentration and that of the preceding one.

To simulate a "conceptually analogous" version for the SCM, we assumed that each protocell has three kinds of templates: a non-specific replicase (R), and two target metabolic genes (M_1, M_2; see Fig. 1). At generation t_0 all protocells start with 20 copies of each template, i.e. a total of 60 genes as before. The replication process can also take place in either of the two ways: replication of target metabolic templates or self-replication of R (sampling of templates is also done without replacement), and their dynamics are now, respectively,

$$\frac{\mathrm{d}(M_{i(k)}|R_{(l)})}{\mathrm{d}t} = \mu_{ij(k,l)}M_{i(k)}R_{(l)},$$

$$i = 1,2,\ \ j = 1,2, i \neq j, k = 0,\ldots,3, l = 0,\ldots 3,$$

$$\frac{\mathrm{d}(R_{(l)}|R_{(m)})}{\mathrm{d}t} = \mu_{s(l,m)}R_{(l)}R_{(m)},$$

$$s = 1,2,\ \ m = 0,\ldots,3. \quad (5)$$

The probabilities of replicating a template are calculated as

$$P(M_{i(k)}|R_{(l)}) = \frac{\mathrm{d}(M_{i(k)}|R_{(l)})/\mathrm{d}t}{\sum_i\sum_k\sum_l \mathrm{d}(M_{i(k)}|R_{(l)})/\mathrm{d}t + \sum_m\sum_l \mathrm{d}(R_{(l)}|R_{(m)})/\mathrm{d}t},$$

$$P(R_{(l)}|R_{(m)}) = \frac{\mathrm{d}(R_{(l)}|R_{(m)})/\mathrm{d}t}{\sum_i\sum_k\sum_l \mathrm{d}(M_{i(k)}|R_{(l)})/\mathrm{d}t + \sum_m\sum_l \mathrm{d}(R_{(l)}|R_{(m)})/\mathrm{d}t}. \quad (6)$$

Replication rate constants (μ's) obviously play a very important role in the dynamics of the model and the way we incorporated them should be clearly understood. The replication rates are simply the products of the target affinities multiplied by the replicase activities, which are specific values with a sigmoid pattern depending on the number of mutated sites. Thus, the replicase activity is a column vector with the values 1, 0.8, 0.2 and 0 for a replicase with $l = 0, 1, 2,$ and 3 deleterious mutant nucleotides, respectively. The putative target affinities of a set of wild type and/or randomly mutated templates are defined by the entries in 4×4 matrices:

$$\begin{bmatrix} M_{i(0)}|R_{.(0)} & M_{i(0)}|R_{.(1)} & M_{i(0)}|R_{.(2)} & M_{i(0)}|R_{.(3)} \\ M_{i(1)}|R_{.(0)} & M_{i(1)}|R_{.(1)} & M_{i(1)}|R_{.(2)} & M_{i(1)}|R_{.(3)} \\ M_{i(2)}|R_{.(0)} & M_{i(2)}|R_{.(1)} & M_{i(2)}|R_{.(2)} & M_{i(2)}|R_{.(3)} \\ M_{i(3)}|R_{.(0)} & M_{i(3)}|R_{.(1)} & M_{i(3)}|R_{.(2)} & M_{i(3)}|R_{.(3)} \end{bmatrix}, \quad (7)$$

where the figures in parentheses indicate "number of mutant nucleotides" and the dot in $R_{.(l)}$ stands for i,j. In the case of no mutation [upper left corner in eqn (7)] we have up to four target affinities in the HPC and up to three in the SCM. Depending on their relative values, we can model for selfish and/or cooperative templates. Allowing for mutation considerably increases the parameter space, and there is no *a priori* rationale as how to fix target affinities.

At this point we must digress by pointing out inaccuracies in the literature concerning the effect of mutations at the compartment level in a non-hypercyclic ensemble of genes (Niesert *et al.*, 1981; Niesert, 1987): it is not necessarily true for a mutant gene with an increased target affinity toward the replicase to be lethal and, conversely, for a reduced target affinity not to have important consequences on the average fitness of the population. In any case, it is not our aim here to explore the parameter space in eqn (7), although it is important to stress that the dynamics of genes within compartments can be very complex when mutations are taken into account.

Because we intended to compare "conceptually analogous" versions of HPC and SCM, we proceeded as follows with eqn (7). Target affinities for self-replication of M_1R_1 and M_2R_2 in HPC were considered to be equal and the same as the target affinity for self-replication of R in the SCM, and target affinities for replication by a different replicase (i.e. $M_{i(k)}|R_{j(l)}; i \neq j$) were assumed to be different. Therefore, we have to define three 4×4 matrices of target affinities. As a particularly interesting case we modelled for higher autocatalysis in the HPC when all templates are wild types. Thus, target affinities were obtained from uniformly distributed random numbers over the interval (0, 1) with the constraint $\mu_{12} < \mu_1 = \mu_2 > \mu_{21}$ (see Fig. 1), which corresponds to a selfish replicase in the SCM. (It is important to realize that usually for the hypercycle one would assume $\mu_2 < \mu_{12}$ and $\mu_1 < \mu_{21}$, to ensure hypercyclic dominance. The particular choice in our case guarantees comparability with the stochastic corrector model. Note that the condition $\mu_1 = \mu_2$ ensures that there will be no internal competition in the hypercycle.) All other entries in the matrices of target affinities when deleterious mutations are incorporated into the model were also obtained from uniformly distributed random numbers but with no constraints (see Appendix). From these matrices we numerically estimated the fitness functions (i.e. the probabilities of replication) of the different templates within HPC- or SCM protocells. A total of 12 000 protocells of each type, with an initial random ensemble ($\leqslant 60$) of wild-type and mutant genes were generated. For each protocell, if a randomly chosen template could be replicated according to its probability of replication [i.e. in agreement with eqn (4) for HPC protocells and eqn (6) for SCM protocells] we assigned a value of one and actually replicated it; otherwise we assigned a value of zero. This procedure was repeated for 25 times. From the resulting two matrices of $12\,000 \times 25$ rows with ones and zeros, we estimated the different probabilities of replication (i.e. the proportion of ones for each combination of template target and replicase), and the results are plotted in Fig. 3(A,B).

We expect from the prior arguments that the fitness functions for metabolic genes when replicated by a non-specific replicase in the HPC be roughly similar to those in the SCM [cf. Fig. 3(A1, B1) and (A2, B2)], and the fitness functions when each member in the HPC serves as a template that can catalyse its own replication to be similar to that obtained for self-replication of R in the SCM [cf. Fig. 3(A3–A4, B3). A close inspection of Fig. 3(A,B) confirms that our model certainly allows for a fair assessment of the behaviour of HPC and SCM when faced with deleterious mutation pressure.

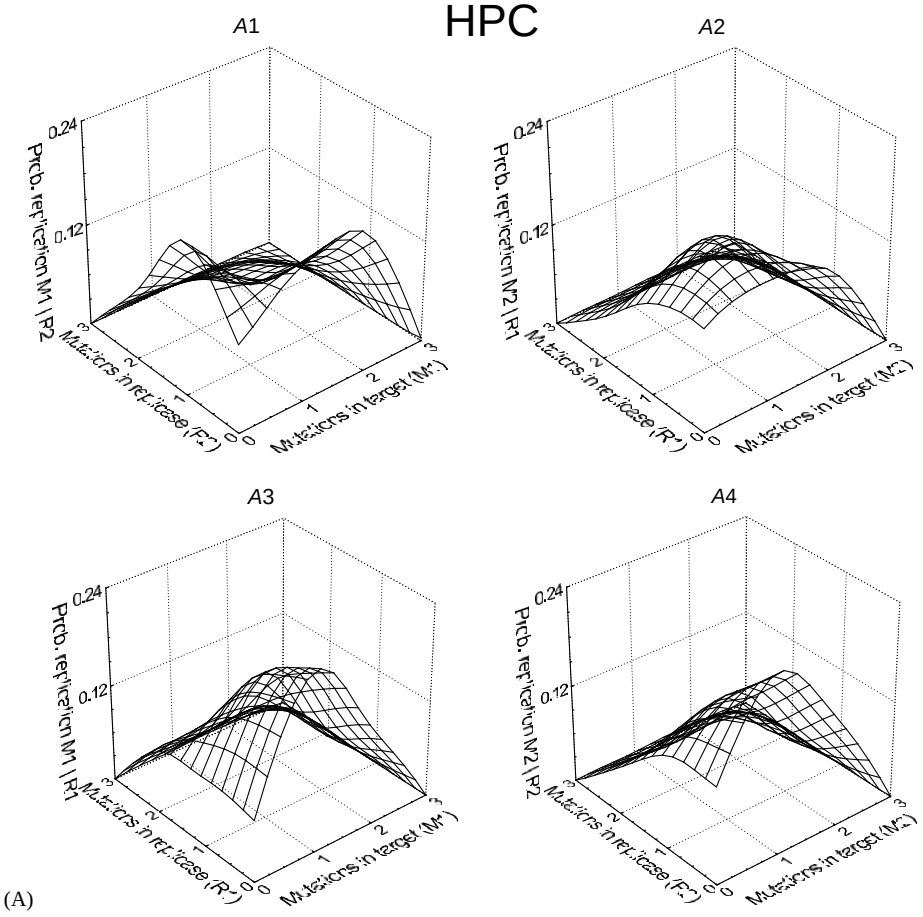
Results from Simulations

It was the objective of the simulations to evaluate the realm of viability of HPC and SCM under realistic mutation rates before the evolution of efficient enzymic machinery for replication. Twelve independent runs were obtained for each set of conditions, and numerical results are plotted in Figs 4 and 5 for mutation rates up to 0.025 per nucleotide per replication.

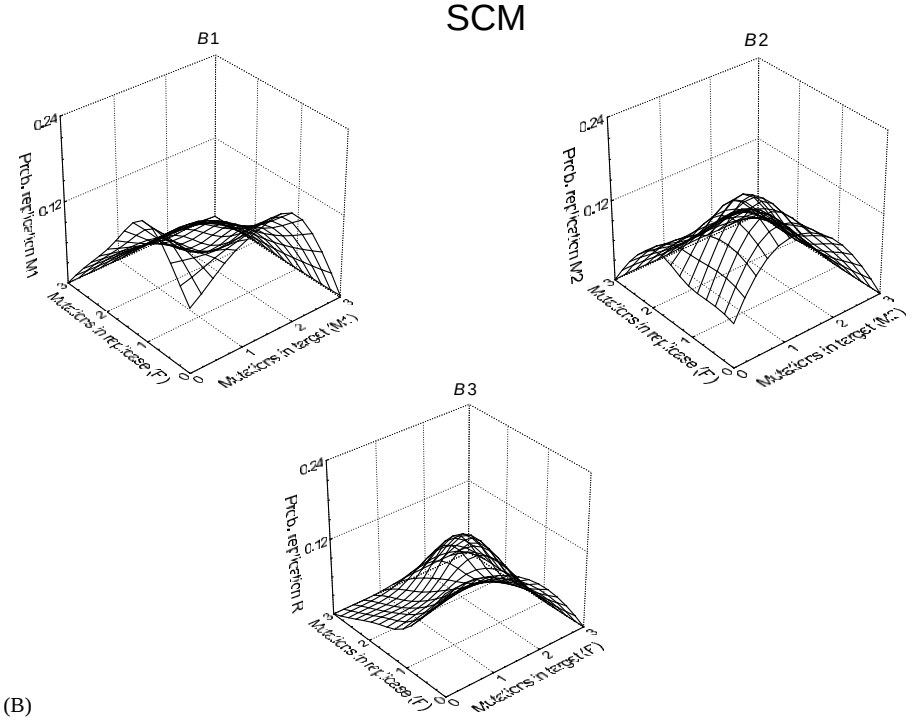
The most striking result is that both kinds of populations can survive when $u = 0.025$, which roughly stands for deleterious mutation rates in the range of 4–5 per protocell assuming a compartment of ~ 60 genes. Thus, an important conclusion is that survivorship of HPC-containing or SCM-containing protocells seems to be quantitatively consistent with a reasonable assumption on the mutation parameters of early replication regardless of the way we have modelled protocell growth and division (i.e. discrete or continuous). There are, however, important qualitative differences when comparing both systems.

Below $u \approx 0.01$, HPC performs better than SCM. The reason is that with low mutation rates

FIG. 3. Within-protocell fitness functions (i.e., probabilities of replication) for each particular template type within (A) HPC protocells or (B) SCM protocells. The plots show the fitness surfaces after least-squares fittings of the points. Note that the probabilities of replicating $M_{1(k)}|R_{2(l)}$(A1) or $M_{2(k)}|R_{1(l)}$(A2) in HPC protocells are similar to the probabilities of replicating the corresponding metabolic gene by the non-specific replicase in SCM protocells (B1 and B2, respectively).


HPC
A1
A2
A3
A4
Prob. replication M1 | R2
Prob. replication M2 | R1
Prob. replication M1 | R1
Prob. replication M2 | R2
0.24
0.12
0
Mutations in repl. case (R*)
Mutations in target (M*)

(A)


SCM
B1
B2
B3
Prob. replication M1
Prob. replication M2
Prob. replication R
0.24
0.12
0
Mutations in repl. case (R)
Mutations in target (M)

(B)

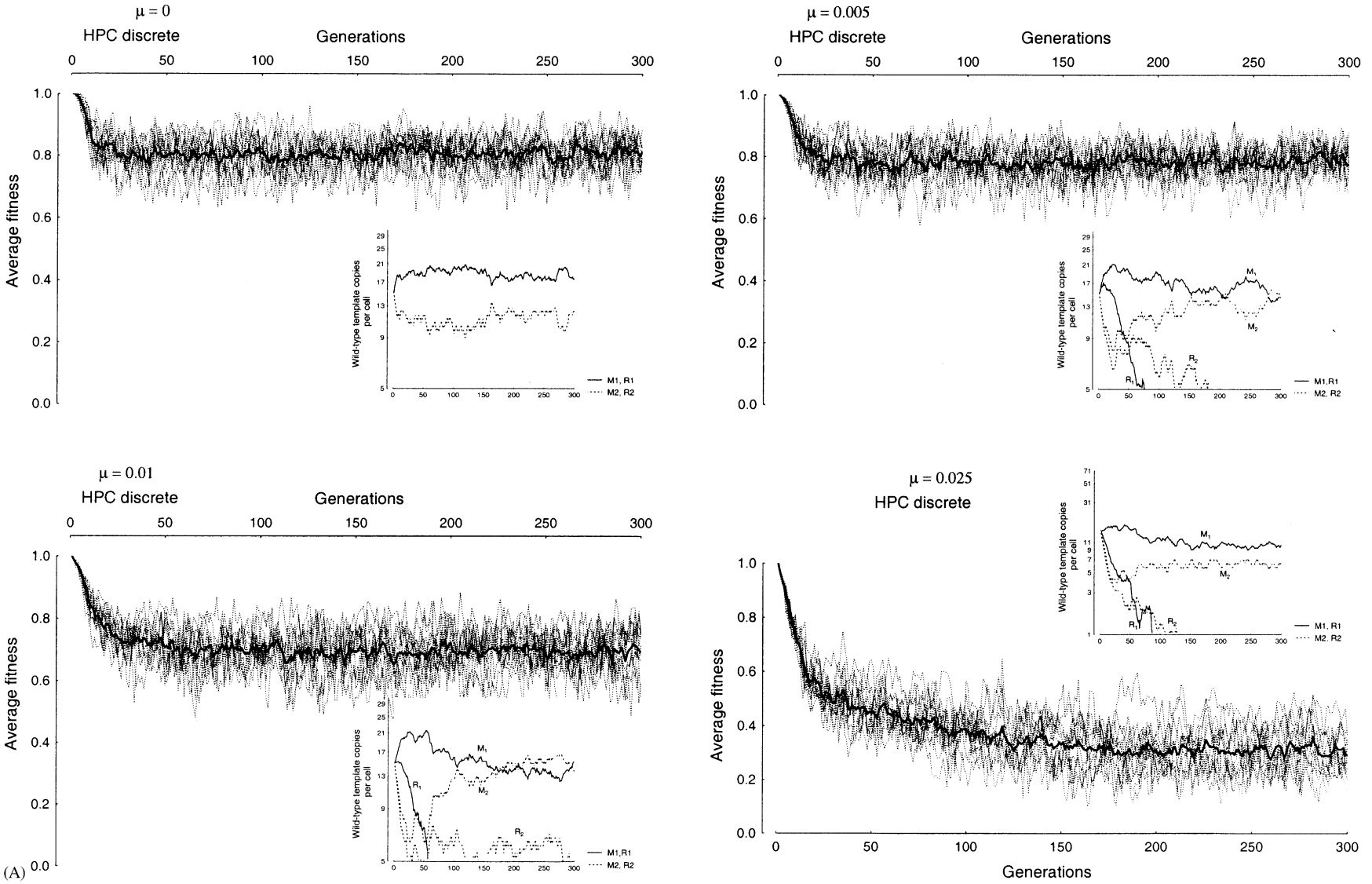
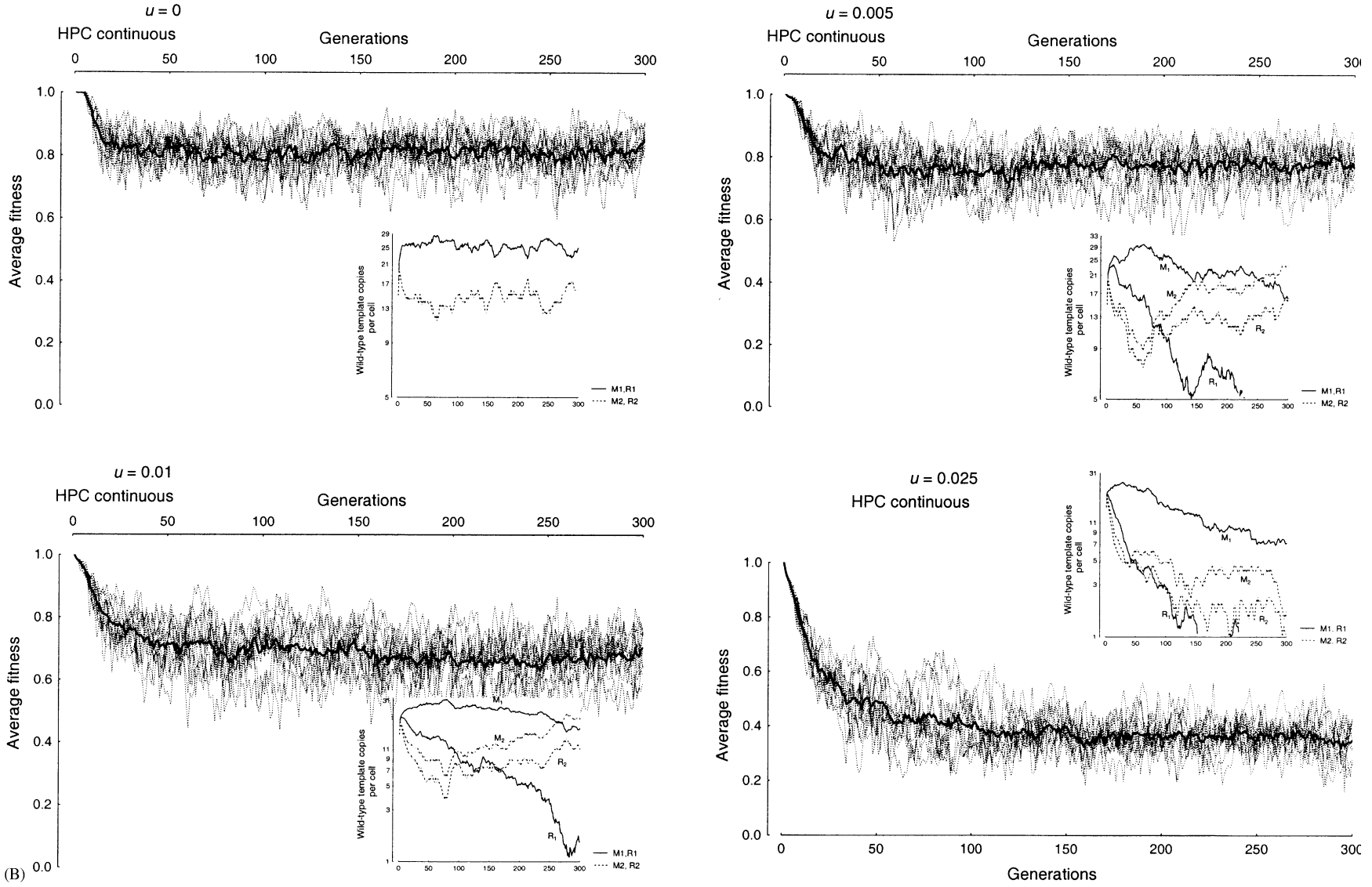
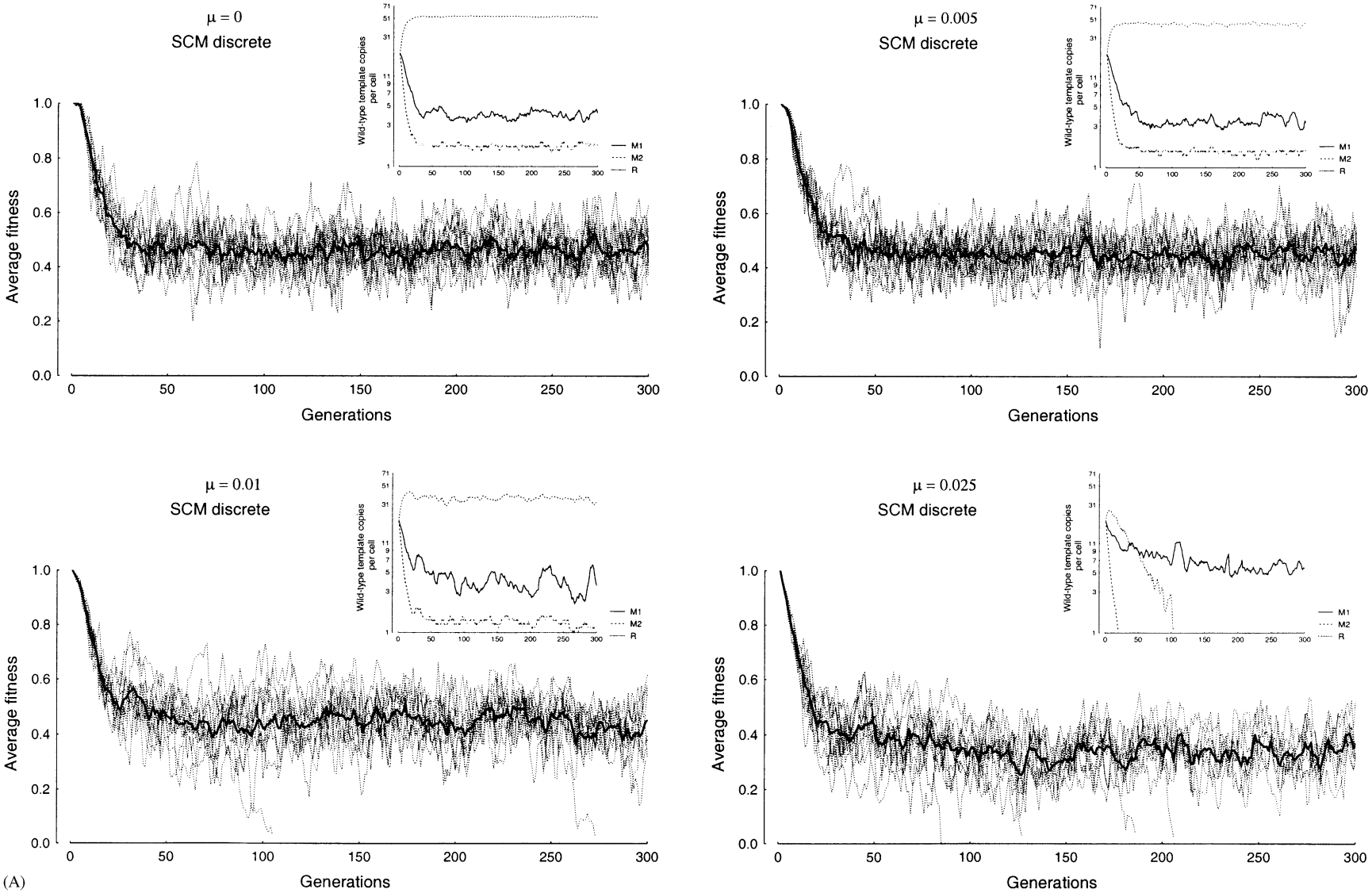


FIG. 4. Results of the Monte Carlo simulations for deleterious mutation rates per nucleotide per replication up to 0.025 in HPC protocells. For each condition, 12 independent runs were achieved and all trajectories (dotted lines) for the average fitness of the population are indicated together with the total average (solid line). Each plot embodies the average number of wild-type copies of each gene in the protocells thorough generations (ordinate is in logarithmic scale to enhance visibility). (A) Discrete version. (B) Continuous version.



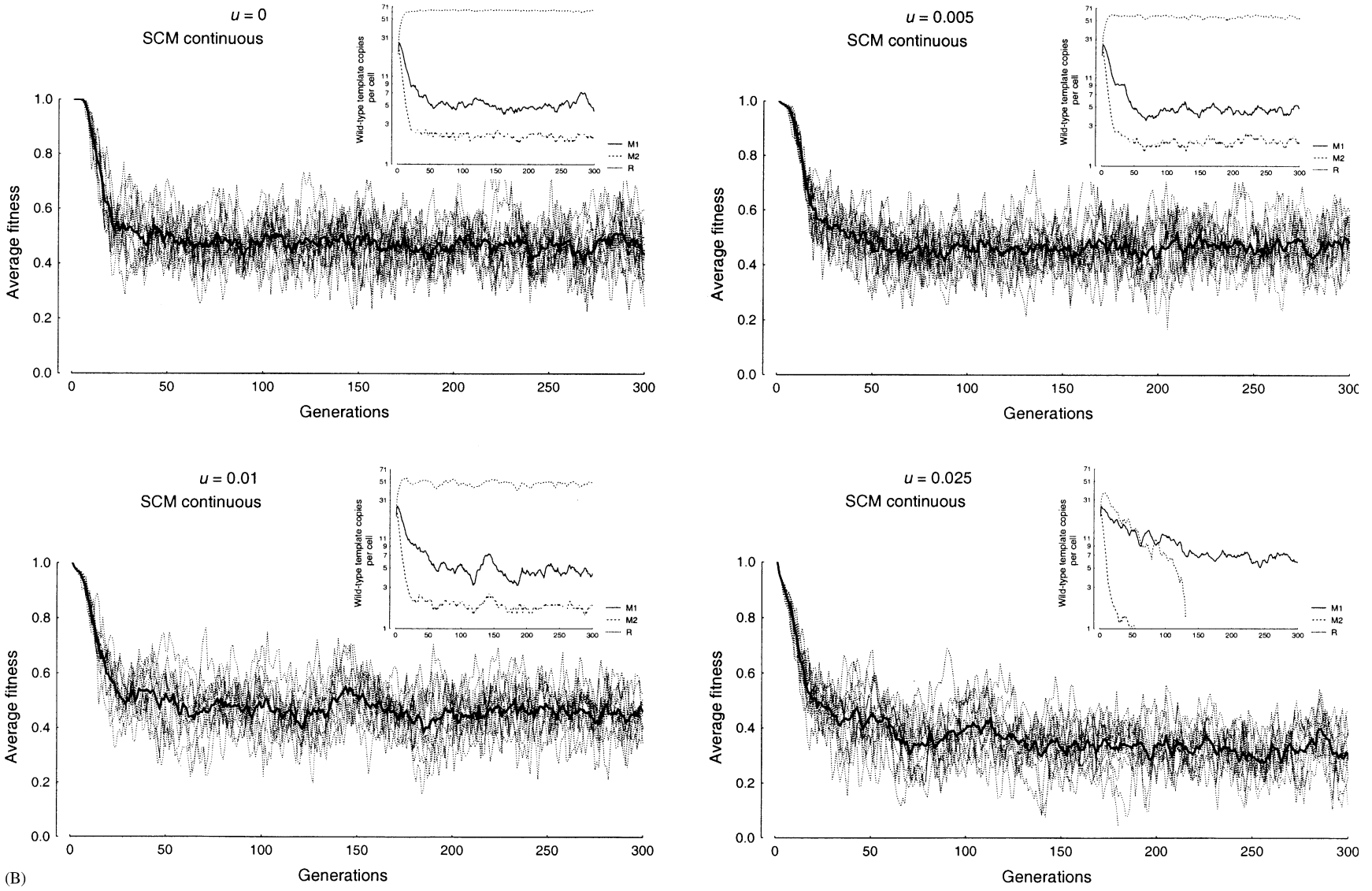
(B)

FIG. 4. (*Continued*).



(A)

FIG. 5. Results of the Monte Carlo simulations for deleterious mutation rates per nucleotide per replication up to 0.025 in SCM protocells. Average fitnesses in all 12 independent runs (dotted lines) are indicated together with the total average from the surviving runs (solid line). Plots for the average number of wild-type copies of each gene in the protocells are also shown (ordinate in logarithmic scale). (A) Discrete version. For $u = 0.01$ two runs were lost before generation 300, and the same happened in four runs with $u = 0.025$. (B) Continuous version. Only one run was lost before generation 300 when $u = 0.025$.



(B)

FIG. 5. (*Continued*).

the assortment load (i.e. the drop in average fitness due to the random loss of any essential gene after stochastic assortment of templates in the two daughter protocells) is higher than the mutation load and stronger in the SCM than in the HPC. Thus, in the SCM the selfish replicase outgrows the metabolic genes and the chance of losing an essential gene after protocell division is correspondingly higher because the segregation of the metabolic genes is independent of that of the replicase (R) template, whereas in the HPC both M and R genes are linked. This result was already anticipated because we have modelled for a selfish replicase (see above), and its immediate selective advantage is readily apparent in the graphs embodied in Fig. 5. There we have plotted the average number of wild-type genes in the protocells. With no mutation the average number of R templates in SCM protocells swiftly increases from 20 to ~ 55, whereas the average number of essential metabolic genes stabilizes at ~ 5 for M_1 and at ~ 2 for M_2. Thus, the chance of a daughter protocell to lack a metabolic gene after stochastic assortment of templates is about 1/2 and, therefore, the costs of the irregular reduction mechanism drop the average fitness to ~ 0.5 at equilibrium. On the other hand, HPC protocells retain an average of ~ 25 M_1R_1 templates and ~ 15 M_2R_2 templates at equilibrium so their probabilities of transmission of at least some copies to each daughter are large enough (Fig. 4).

As u increases from 0.01, the average fitness of HPC protocells progressively decreases, while the fall in average fitness of SCM protocells is far less clear-cut (cf. Figs 4 and 5). In fact, with $u = 0.05$, which still represents a realistic error rate before replication could be reasonably accurate, HPC protocells quickly collapse within the first 20 (for the continuous case) or 40 (for the discrete case) generations. Conversely, about 30% of the runs for SCM protocells stabilize at an average fitness of ~ 0.2 (data not shown). Put in other words, the mutational load (Haldane, 1937; Crow, 1970), defined as $L = (w_0 - w_1)/w_0$ where w_0 is the fitness with no deleterious mutations, is always lower in the SCM than in the HPC. With $u = 0.05$ the mutational load can become stable at ~ 0.6 in the SCM (but see below), whereas in the HPC it is almost 1.

The reason for a higher mutational load in the HPC could be easily understood by realizing that any of its members is a self-replicating RNA-like molecule that also catalyses specifically the replication of the next member. Therefore, for every template M_iR_i a poor replicase (i.e., a replicase with deleterious mutant nucleotides) can easily "hitchhike" a fittest wild-type metabolic gene. Both replicases (R_1 and R_2) need to be simultaneously functional in the HPC due to its double autocatalytic nature, and this results in a major difference with the SCM (Fig. 5). Consider the graphs embodied in Fig. 4(A,B) when $u = 0.025$. The average number of wild-type metabolic genes at equilibrium is still significant (~ 9 for M_1 and ~ 4 for M_2), but the wild-type replicases (mainly R_1) are eventually lost. With increasing mutation rates ($u = 0.05$), the lost is very fast and the HPC population finally collapses. The main cause for the selective inferiority of the HPC at high mutation rates is due to the following reasons: (i) one of the replicase genes becomes non-functional and just takes "space" away from the functioning one (total cell size is based on the total number of genes), and (ii) this gene is in strong linkage disequilibrium with one of the metabolic genes. Consequently, the combinations: bad replicase–good metabolic gene and good replicase–bad metabolic gene, are unbreakable. This by definition cannot happen in the SCM.

Since the SCM necessarily assumes that the number of different genes per protocell must be small in order to avoid an unsupportive assortment load, a potential caveat with our simulations is to have modelled the minimum HPC of two members to make both systems comparable. It has been previously pointed out (Szathmáry, 1989b), however, that the HPC is a wasteful means of information integration, and shortcut mutations can be produced that make the cycle arbitrarily short. In addition, the former argument concerning the mutational load clearly suggests that populations of protocells with large HPCs are more liable to collapse as u increases. Preliminary results with an HPC of three members suggest that this is indeed the case (unpublished). Thus, our conclusion that a population of SCM protocells can tolerate

higher deleterious mutation rates than a population of HPC protocells seems to be robust.

In the simulations we assumed a constant K of 150 protocells and, therefore, did not allow the population size to decline as a consequence of the resulting reduction in average fitness (Lynch *et al.*, 1995, 1999). Concerning the population of SCM protocells, there are two important sources of genetic load: assortment load and mutational load (Fig. 5). It is essential to realize that the assortment load could be easily reduced in a population of SCM protocells initially heterogeneous for target affinities [i.e. there is between-protocell variability for the entries in eqn (7)] and, hence, replication rates. Thus, selection at the compartment level would quickly increase the proportion of cells where all metabolic and replicase templates happen to grow at nearly similar rates within each protocell. In other words, selection would reduce the variability in replication rates (target affinities), with the consequent decline of assortment load (unpublished results).

When $\mu_S = \mu_{12} = \mu_{21}$ (see Fig. 1) the average fitness for SCM protocells with no mutation stabilizes at ~ 0.86. Similarly, the average fitness stabilizes at ~ 0.77 with u as high as 0.025 (i.e. $L \sim 0.11$). Now the relatively small decrease in average fitness is only due to the mutational load (cf, $L \sim 0.28$ in Fig. 5) because all entries in eqn (7) were assumed to be equal. In summary, a population of SCM protocells will likely be able to persist and flourish in spite of a high input of deleterious alleles and, hence, solve the error threshold problem.

Discussion

Starting from a world of naked replicating molecules, it took a series of not yet well-understood transitions or steps (Maynard Smith & Szathmáry, 1995; Eigen & Schuster, 1982; Szathmáry & Maynard Smith, 1997) to arrive at the first organisms that formed the earliest identified bacterial fossils (Schopf, 1993). The minimal living system can be considered to be a chemical supersystem based on three subsystems: the metabolic network, the replicating template(s), and the enclosing membrane (Gánti, 1987). A cell-like structure was a wonderful invention to align the immediate benefits of each gene with those of the whole genome. However, an unsolved question is whether or not in early protocells there was an assembly of cooperating (mutualistic) replicators (i.e., a hypercycle) or just a group of competing genes that happened to be together.

It has been said that once cellularization appeared, it solved the very problem that initiated the development of hypercycle theory and, therefore, we no longer need hypercycles at all (Bresch *et al.*, 1980; Szathmáry, 1989b). But as noted above, there have been neither rigorous comparisons of hypercyclic vs. non-hypercyclic protocell populations, nor detailed analyses of their tolerable deleterious mutation rates. Two main results have emerged from our Monte Carlo simulations: (i) both HPC- and SCM protocells are efficient information integrators whose performance obviously degenerates as the mutation rate increases but anyway could tolerate a very high input of deleterious mutations, and (ii) the HPC performs better in terms of average fitness for low mutation rates, but the mutational load by itself is always lower in a population of SCM protocells. But for HPC protocells we have assumed the minimum hypercycle of only two members and, therefore, our conclusion that they can cope with high mutational load should be taken with caution (see above).

Two important connections to earlier work must be spelled out. First, as was recognized by Szathmáry & Demeter (1987), the stochastic corrector model at the reproducing compartment level is formally analogous to Eigen's (1971) quasispecies model for replicating macromolecules; with analogous terms for reproduction, copying fidelity, and mutation. The compartmentalized hypercycle model presented here also forms a quasispecies distribution at the compartment level, due to the lack of fusion between compartments (asexual or clonal reproduction). There have been attempts to tackle with the error threshold problem, originally defined for an infinite, deterministic model (Eigen, 1971), in the case of a finite population prone to stochastic demographic effects. A good recent example is the treatment by Campos & Fontanari (1998). Note that the models

presented above tackle an analogous problem by *in extenso* simulation of finite (small) populations.

Second, the most respectable previous attempt to analyse error propagation in the hypercycle (Campos *et al.*, 2000) suffers from the shortcoming that the mutant tail forms a subpopulation of autocatalytic replicators not interacting catalytically either with the original hypercycle or among themselves, thus reducing competition that the original hypercycle is faced with (mutants grow slowly and parasites or alternative cycles are excluded). The present study does not rest on such artificial restrictions.

Campos *et al.* (2000) also complained about the use of kinetics with non-integer exponents in the original formulation (Szathmáry & Demeter, 1987) of the stochastic corrector model. This complaint is unjustified, at least for two reasons. First, the terms applied there did not cheat by prohibiting internal competition, which was the central problem that led to the hypercycle model. Second, here we show that a more detailed model, lacking such lump terms, behaves essentially in the same way. The stochastic corrector model is shown to have smaller mutational load but a higher segregation load for two metabolic genes, giving a higher average fitness to the compartmentalized hypercycle. We suspect that increasing the number of genes will increase the mutational load of the latter to such an extent that even the average fitness of the HPC will fall below that of the stochastic corrector. Work to test this prediction is in progress.

Sloppy as it is, the stochastic corrector model is a more parsimonious model than the hypercycle. SCM integrates genetic information and overcomes the danger of information decay before replication was reasonably accurate. We have now strong reasons to use it as a relatively robust system to model the first major transitions in evolution (Maynard Smith & Szathmáry, 1995). Thus, the Monte Carlo model in this paper will also be used to analyse important problems such as the efficiency of information integration of larger systems harbouring more genes, or the role sex played in the dynamics of fragmented genomes, and the spread of linkage (Maynard Smith & Szathmáry, 1993; Santos, 1998).

M.S. acknowledges the European Science Foundation for providing financial support during his stay at Collegium Budapest. E. Sz. was supported by the National Research Fund (OTKA no. T 035223) of Hungary.

REFERENCES

Bartel, D. P. & Unrau, P. J. (1999). Constructing an RNA world. *Trends Genet*. **15,** M9–M13.

Bresch, C., Niesert, U. & Harnasch, D. (1980). Hypercycles, parasites and packages. *J. theor. Biol*. **85,** 399–405.

Campos, P. R. A. & Fontanari, J. F. (1998). Finite-size scaling in a quasispecies model. *Phys. Rev. E* **58,** 2664–2667.

Campos, P. R. A., Fontanari, J. F. & Stadler, P. F. (2000). Error propagation in the hypercycle. *Phys. Rev. E* **61,** 2996–3002.

Crow, J. F. (1970). Genetic loads and the cost of natural selection. In *Mathematical Topics in Population Genetics* (Kojima, K. I. ed.), pp. 128–177. Berlin: Springer-Verlag.

Eigen, M. (1971). Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58,** 465–523.

Eigen, M. (1992). *Steps towards Life: a Prospective on Evolution*. Oxford: Oxford University Press.

Eigen, M. & Schuster, P. (1979). *The Hypercycle: a Principle of Natural Self-organization*. Berlin: Springer-Verlag.

Eigen, M. & Schuster, P. (1982). Stages of emerging life—five principles of early organization. *J. Mol. Evol*. **19,** 47–61.

Eigen, M., Schuster, P., Gardiner, W. & Winkler-Oswatitsch, R. (1981). The origin of genetic information. *Sci. Am*. **244,** 78–94.

Friedberg, E. C., Walker, G. C. & Siede, W. (1995). *DNA Repair and Mutagenesis*. Washington: ASM Press.

FTN77 (1997). *Salford software Ltd*. Tucson, AZ.

Gánti, T. (1975). Organisation of chemical reactions into dividing and metabolizing units: the chemotons. *Biosystems* **7,** 15–21.

Gánti, T. (1987). *The Principle of Life*. Budapest: OMIKK.

Grey, D., Hutson, V. & Szathmáry, E. (1995). A re-examination of the stochastic corrector model. *Proc. R. Soc. London* B **262,** 29–35.

Haldane, J. B. S. (1937). The effect of variation on fitness. *Am. Nat*. **71,** 337–349.

Jeffares, D. C., Poole, A. M. & Penny, D. (1998). Relics from the RNA world. *J. Mol. Evol*. **46,** 18–36.

Johnston, W. K., Unrau, P. J., Lawrence, M. S., Glasner, M. E. & Bartel, D. P. (2001). RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* **292,** 1319–1325.

Kauffman, S. A. (1993). *The Origins of Order*. Oxford: Oxford University Press.

Lynch, M., Conery, J. & Bürger, R. (1995). Mutation accumulation and the extinction of small populations. *Am. Nat*. **146,** 489–518.

LYNCH, M., BLANCHARD, J., HOULE, D., KIBOTA, T., SCHULTZ, S., VASILIEVA, L. & WILLIS, J. (1999). Perspective: spontaneous deleterious mutation. *Evolution* **53,** 645–663.

MATLAB V.5.3. (1999). *The Language of Technical Computing*. The Math Works, Inc. URL: http://www.mathworks.com. Natick, MA.

MAYNARD SMITH, J. (1979). Hypercycles and the origin of life. *Nature* (*London*) **280,** 445–446.

MAYNARD SMITH, J. & SZATHMÁRY, E. (1993). The origins of chromosomes I. Selection for linkage. *J. theor. Biol.* **164,** 437–446.

MAYNARD SMITH, J. & SZATHMÁRY, E. (1995). *The Major Transitions in Evolution*. Oxford: Freeman & Co.

MICHOD, R. (1983). Population biology of the first replicators: the origin of genotype, phenotype and organism. *Am. Zool.* **23,** 5–14.

NIESERT, U. (1987). How many genes to start with? A computer simulation about the origin of life. *Origins Life* **17,** 155–169.

NIESERT, U., HARNASCH, D. & BRESCH, C. (1981). Origin of life: between Scylla and Charybdis. *J. Mol. Evol.* **17,** 348–353.

SANTOS, M. (1998). Origins of chromosomes in response to mutation pressure. *Am. Nat.* **152,** 751–756.

SCHOPF, J. W. (1993). Microfossils of the early archean apex chert: new evidence of the antiquity of life. *Science* **260,** 640–646.

SZATHMÁRY, E. (1989a). The integration of the earliest genetic information. *Trends Ecol. Evol.* 4, 200–204.

SZATHMÁRY, E. (1989b). The emergence, maintenance, and transitions of the earliest evolutionary units. *Oxford Surv. Evol. Biol.* **6,** 169–205.

SZATHMÁRY, E. & DEMETER, L. (1987). Group selection of early replicators and the origin of life. *J. theor. Biol.* **128,** 463–486.

SZATHMÁRY, E. & MAYNARD SMITH, J. (1997). From replicators to reproducers: the first major transitions leading to life. *J. theor Biol.* **187,** 555–571.

APPENDIX

Target affinities for self-replication of $M_{1(k)}R_{1(l)}$ and $M_{2(k)}R_{2(l)}(k = 0, \ldots, 3; l = 0, \ldots, 3)$ in the HPC were considered to be the same and equal to target affinities for self-replication of $R(R_{(k)}|R_{(l)})$ in the SCM, and they are shown in matrix **A**. Target affinities for replication of $M_{1(k)}|R_{2(l)}$ are given in matrix **B**, and those for replication of $M_{2(k)}|R_{1(l)}$ in matrix **C**. Note that $\mathbf{B}(1,1) < \mathbf{A}(1,1) > \mathbf{C}(1,1)$, so replication of R in the stochastic corrector model assuming no mutation will be higher than the replication of the metabolic genes (i.e., we are modelling for a selfish replicase).

$$\mathbf{A} = \begin{bmatrix} M_{1(0)}|R_{1(0)} = 0.3757 & M_{1(0)}|R_{1(1)} = 0.4669 & M_{1(0)}|R_{1(2)} = 0.7021 & M_{1(0)}|R_{1(3)} = 0.7129 \\ M_{1(1)}|R_{1(0)} = 0.8413 & M_{1(1)}|R_{1(1)} = 0.6937 & M_{1(1)}|R_{1(2)} = 0.5345 & M_{1(1)}|R_{1(3)} = 0.7034 \\ M_{1(2)}|R_{1(0)} = 0.4432 & M_{1(2)}|R_{1(1)} = 0.2307 & M_{1(2)}|R_{1(2)} = 0.9334 & M_{1(2)}|R_{1(3)} = 0.7042 \\ M_{1(3)}|R_{1(0)} = 0.3117 & M_{1(3)}|R_{1(1)} = 0.5484 & M_{1(3)}|R_{1(2)} = 0.7991 & M_{1(3)}|R_{1(3)} = 0.5322 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} M_{1(0)}|R_{2(0)} = 0.3006 & M_{1(0)}|R_{2(1)} = 0.7972 & M_{1(0)}|R_{2(2)} = 0.7933 & M_{1(0)}|R_{2(3)} = 0.0711 \\ M_{1(1)}|R_{2(0)} = 0.5953 & M_{1(1)}|R_{2(1)} = 0.3178 & M_{1(1)}|R_{2(2)} = 0.7454 & M_{1(1)}|R_{2(3)} = 0.8377 \\ M_{1(2)}|R_{2(0)} = 0.6571 & M_{1(2)}|R_{2(1)} = 0.4725 & M_{1(2)}|R_{2(2)} = 0.8757 & M_{1(2)}|R_{2(3)} = 0.4751 \\ M_{1(3)}|R_{2(0)} = 0.1669 & M_{1(3)}|R_{2(1)} = 0.1369 & M_{1(3)}|R_{2(2)} = 0.0758 & M_{1(3)}|R_{2(3)} = 0.0179 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} M_{2(0)}|R_{1(0)} = 0.2532 & M_{2(0)}|R_{1(1)} = 0.2551 & M_{2(0)}|R_{1(2)} = 0.812 & M_{2(0)}|R_{1(3)} = 0.2655 \\ M_{2(1)}|R_{1(0)} = 0.8574 & M_{2(1)}|R_{1(1)} = 0.8598 & M_{2(1)}|R_{1(2)} = 0.5430 & M_{2(1)}|R_{1(3)} = 0.5062 \\ M_{2(2)}|R_{1(0)} = 0.4051 & M_{2(2)}|R_{1(1)} = 0.2171 & M_{2(2)}|R_{1(2)} = 0.8459 & M_{2(2)}|R_{1(3)} = 0.9403 \\ M_{2(3)}|R_{1(0)} = 0.4426 & M_{2(3)}|R_{1(1)} = 0.5348 & M_{2(3)}|R_{1(2)} = 0.4156 & M_{2(3)}|R_{1(3)} = 0.2900 \end{bmatrix}.$$