Genetic Fusion

Takashi Ikegami[a]
The Graduate School of Science and Technology, Kobe University, Rokkodai, Nada-ku, Kobe 657, Japan

Kunihiko Kaneko[b]
Institute of Physics, College of Arts and Sciences, University of Tokyo, Komaba, Meguro-ku, Tokyo 153, Japan
(Received 25 June 1990)

Genetic fusion is introduced as a model for evolution. In fusion two genomes are combined to generate a longer genome. Representing each species by a binary genetic sequence, we introduce a fitness function on the bit sequence. As the evolutionary dynamics, we incorporate mutation, genetic fusion, and reproduction in proportion to fitness. It is found that genetic fusion leads to the appearance of module-type sequences and duplicated genes. The time necessary to find a sequence with larger fitness is largely reduced by the inclusion of genetic fusion, which suggests the application of our algorithm to optimization problems.

PACS numbers: 87.10.+e

Theoretical studies on mechanisms of evolutionary processes are of importance not only for theoretical biology but also from the viewpoint of information processing. Fitness in evolution is believed to depend on the genetic sequence. The dependence of this fitness on sequences can be complicated, and has been referred to as a rugged landscape.[1] A spin-glass model[2] provides a simple example which generates this class of landscape. A typical model is the Sherrington-Kirkpatrick (SK) model.[2] In the model, a spin S_i (=1 or −1) is assigned to a site i. Energy is defined on a binary sequence $[S_i]$, given by $E=\sum_{i,j} J_{i,j} S_i S_j$, where a spin S_i interacts with all the other spins S_j. The coupling $J_{i,j}$ is set at a fixed random value distributed over positive and negative values. It is known that the model provides many metastable states and a rugged landscape.[2] The application of this class of model to evolution has already been discussed by Anderson,[3] where the SK model is applied to prebiotic evolution.

Another important mechanism in evolution is sexual recombination of sequences by mating. This mechanism has frequently been used in the so-called genetic algorithm.[4] Here two genetic sequences are spliced and joined; for example, parents $[\sigma_1,\sigma_2,\sigma_3,\sigma_4,\dots]$ and $[\tau_1,\tau_2,\tau_3,\tau_4,\dots]$ lead to the offsprings $[\sigma_1,\sigma_2,\tau_3,\tau_4,\dots]$ and $[\tau_1,\tau_2,\sigma_3,\sigma_4,\dots]$. This recombination is especially useful to escape from a local minimum in configuration space and has been applied to optimization problems.

In the present Letter, we focus on another mechanism of evolution; genetic incorporation involving different species, which also changes the length of the genetic sequence itself. This mechanism may have been especially important in the early stages of evolution and it is probably still being used in present-day organisms.[5]

There is much evidence to confirm this genetic incorporation. Ohno has put forward a theory of gene duplication.[6] In his theory, some parts of a genetic sequence may incur duplication error during the process of reproduction, as

$$[\sigma_1,\sigma_2,\sigma_3,\sigma_4,\sigma_5,\sigma_6,\dots] \rightarrow [\sigma_1,\sigma_2,\sigma_3,\sigma_1,\sigma_2,\sigma_3,\sigma_4,\sigma_5,\sigma_6,\dots],$$

where σ_i represents a gene at a site i. The duplication theory is confirmed in real DNA. Multiple repetition of short DNA sequences is commonly observed.[5]

Besides the above gene duplication, a biological system has transposable genetic elements. They are self-reproductive elements such as phages, plasmids, and transposons. For example, bacteria can evolve to be antibiotic resistant with the use of R plasmids. Plasmids can insert a particular sequence into other plasmids. Even in eukaryotes, the existence of a cell symbiotic partner is known to have escalated the tempo of evolution, as has been discussed in the cell symbiosis of mitochondria.[7]

Then general questions arise: Why are the above genetic incorporation algorithms adopted in real evolution? Can we construct a simple model for these processes? If so, does the model give rise to practical merits over simple genetic algorithms? In the present Letter, we give a partial answer to these questions, by introducing an abstract model of the above cell symbiosis and gene duplication. The process in our model is called "genetic fusion," here borrowing from molecular biology.

First, we represent each species by a bit string of genes σ_i, which take the value 0 or 1. The length of this sequence is not fixed. Species are assumed to evolve in a fixed environment. For the sake of simplicity, we adopt a spin-glass-type energy for our fitness function. Instead of the SK model, we use the following translational-invariant version of it:

$$E=\sum_{i,j} J_{|i-j|} S_i S_j , \tag{1}$$

where $S_i=2\sigma_i-1$, and J_m is set at a fixed random value distributed over $[-1,1]$.[8]

Introducing an ensemble of species, we take the fol-

© 1990 The American Physical Society

lowing processes for the evolutionary dynamics.

(i) Mutation: $\sigma_1,\sigma_2,\sigma_3,\sigma_4,\sigma_5,\ldots \rightarrow \sigma_1,\sigma_2,1-\sigma_3,\sigma_4,\sigma_5,\ldots$. Here we assume that there occurs at most only a single mutation for each sequence. The rate of this single-point mutation error is chosen to be μ.

(ii) Genetic fusion: When two species $[\sigma_1,\sigma_2,\sigma_3,\ldots]$ and $[\tau_1,\tau_2,\tau_3,\ldots]$ meet, a new species $[\sigma_1,\sigma_2,\sigma_3,\ldots,\tau_1,\tau_2,\tau_3,\ldots]$ can be created. We assume that the fusion occurs only if the energy of the fused species is lower than that of the original two species. The partner for each fusion is arbitrarily chosen with equal probability from the population ensemble. If the original two lengths are l_1 and l_2, the fused sequence has a length of l_1+l_2.[9]

(iii) Reproduction: Each species generates offspring in proportion to its fitness. We assume that the fitness depends on energy as $\exp(-\beta E)$. With this reproduction process only, the number of some species would increase indefinitely. In order to avoid this divergence and to take account of the competition among individuals, we include a suppression mechanism. For the suppression, there can be a variety of algorithms. We choose one of the following two algorithms.

One is competition within the same species. In the present Letter the following dynamics is adopted: $N'(i)=N(i)\exp[-\beta E(i)]/[1+gN(i)^2]$, where $N'(i)$ is the number of the species i at the next step. The denominator comes from the competition within the same species.

The other method is to normalize the total population size to N_0 if the size exceeds some threshold value. We simply take the form

$$N'(i)=N_0N(i)\exp[-\beta E(i)]\Big/\sum_j N(j)\exp[-\beta E(j)]$$

here. In this case, all the species compete for the same niche. The latter algorithm is especially useful in the application to optimization problems.[10]

Our dynamics consists of successive operations of the above three algorithms (i)–(iii). Starting from an ensemble of species with small bit lengths (up to 2), we first apply mutation and genetic fusion. Then reproduction is applied to the ensemble. This completes the first time step. We continue these procedures for many time steps.

In Fig. 1, the energies of existing species are plotted as a function of time. A downward solid line indicates a fusion process between two species, while a dashed line denotes the genesis of a new species by mutation.

As is noted in Figs. 1 and 2, a species (often of small energy) is frequently utilized as a fusion "partner" by many other species. We call such species a "module" species and its sequence a module sequence.

Heuristically, a module sequence satisfies the following conditions: First, the sequence itself is metastable among the sequences with identical length; mere mutation cannot produce a species of smaller energy. Second, the fusion between the module sequence and many other

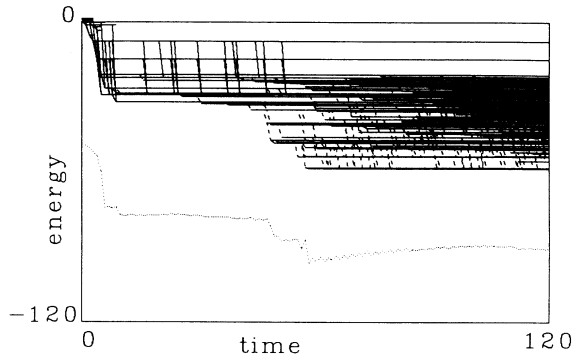

FIG. 1. Each species is depicted as a function of time and energy. A solid line shows a fusion process and a dashed line shows a mutation process. The dotted line indicates the average value of energy, plotted by shifting the value vertically by −50. The mutation rate is set at 0.03 with $\beta=0.1$ and $g=0.05$.

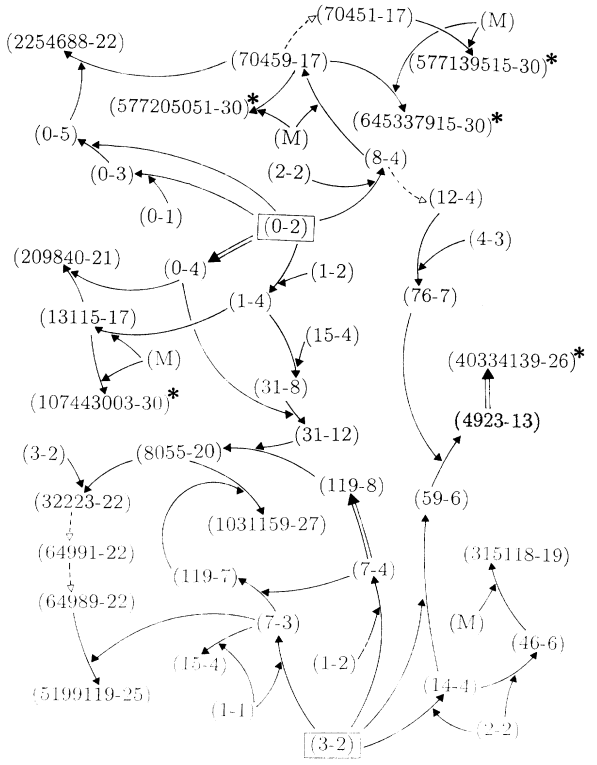

FIG. 2. An example of phylogeny of the species generated by our dynamics: The mutation rate is set at 0.02 with $\beta=0.1$ and $g=0.025$. The species at an arrowhead is generated from the species at the tail by fusing with the other species. A double solid line shows a duplication of the species of the arrow tail. A dotted line denotes the mutation process. Here the species (4923-13) is a module. Fusions with this module [denoted by the symbol (M)] occur successively. Note also the fusion of the module with itself. A species generated by the duplication of (M) is depicted with the asterisk on its right shoulder.

TABLE I. Bit sequence of module and remaining species. Examples of module species are depicted. Each of these species is expected to be in a local minimum. Each species (l-m) is specified by a decimal code l and its length m. The decimal code is obtained by the conversion of its binary sequence. Examples of remaining species after 30 time steps are also exhibited. Note that they are composed of the modules in tandem. The symbol $[A']$ denotes a sequence which is different from $[A]$ by one bit.

	Species	Sequence	Energy
		Modules	
A	4923-13	1001100111011	−13.796
B	4095-12	111111111111	−13.066
C	4064-13	0111111100000	−7.475
		Remaining species	
	645100347-30	$[A'][1011][A]$	−55.714
	577139515-30	$[1000][A'][A]$	−46.592
	16777215-29	$[00000][B][B]$	−43.532
	40334134-26	$[A][A]$	−36.634
	16777215-24	$[B][B]$	−31.838
	33550304-25	$[B][C]$	−26.564

species easily leads to species of better fitness. For this condition, the energy of a module species should not be too small. In our simulation, module species have small fitness and thus have small population size.[11]

There can simultaneously be several different module species. If this is the case, species are categorized into phylogenic groups according to whether they share a common module or not. Examples of such module sequences in our simulation are listed in Table I. All of the persistent species originate in one of these module species.

It often happens that a module sequence $M_1 = [\sigma_1, \ldots, \sigma_k]$ combines with itself. This leads to gene duplication, as $M_1 \rightarrow [M_1, M_1]$. The chain-duplication process also often occurs as $[M_1] \rightarrow [M_1, M_1] \rightarrow [M_1, M_1, M_1]$ or as $[M_1] \rightarrow [M_1, M_1] \rightarrow [M_1, M_1, M_1, M_1]$. This class of gene duplication is frequently seen in biological evolution, as is stressed by Ohno.[6]

In our simulation, the evolutionary process includes duplication, fusion, and mutations. Thus a genetic sequence does not have a complete repetitive structure, but consists of fragments of repetitive parts. This is true in our simulation (see Table I) and in real data.[6]

If the mutation rate is small, few (often a single) module species are allowed in our simulation. Possible fusions consist of those between the same module species. The identical module sequence is repetitively made use of in the fusion process. In other words, the genetic duplication is frequently seen in a low-mutation regime. If the mutation rate is higher, fusion with many other species can happen. The obtained species consist of a combination of a variety of sequences.

Our genetic fusion can provide an algorithm for the search of a lower energy (or larger fitness) in a rugged landscape, faster than the conventional genetic algorithm or simulated annealing.

Let us compare our genetic fusion and a mere mutation process. The simple mutation process corresponds to a traditional Monte Carlo method.

In Fig. 3, we have plotted the minimum energy of the species versus CPU time for simulations with and without genetic fusion (both have the same mutation rates). In the simulation of mutation only, we have started from ten randomly chosen sample species with bit length of 30. The advantage of our genetic fusion is clearly seen.[12]

The advantage comes from the following two features.

(i) Our fusion process tries to find a minimum by combining small fragments. Thus it is effective when an optimal solution is well approximated by a combination of smaller parts and modules (e.g., a partial solution).

(ii) With mere mutation, our system is easily trapped by a local minimum. It may require a long time to hit a global minimum. The present genetic fusion includes a global jump in the configuration space. Our system can skip out of the local minimum.

The second merit is also present in the conventional genetic algorithm, where sexual recombination leads to a global jump. Owing to the first merit, our fusion clearly shows predominance over the conventional algorithm.[4]

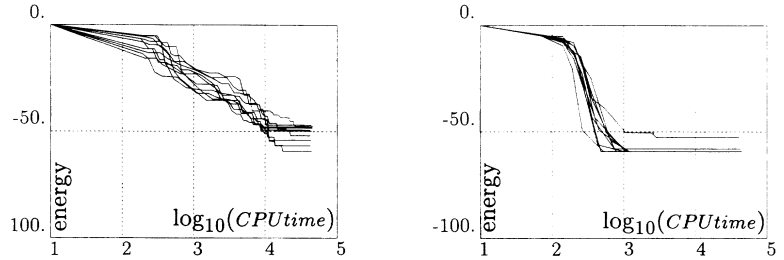

FIG. 3. The lowest-energy values at each CPU time are overlaid for ten different samples. With fusion (right-hand panel) lower energy is attained faster than without fusion (left-hand panel). Both processes are simulated with a mutation rate of 0.1 with $\beta = 0.1$ and $g = 0.1$.

The fusion process is powerful, if the optimal solution can be constructed as a combination of partial solutions. It is promising to apply our approach to other optimization problems. We believe that it is better than the mere mutational process in general.

The evolutionary process is often believed to be organized in a genealogical tree. This is true, if the process consists only of mutations and sexual recombination within species. If genetic fusion is a principal mechanism in evolution, the genealogy is organized in a rhizomelike structure.

The fusion process introduces deviation from a treelike structure, as is shown in Fig. 2. A descendant can have more than one ancestor. The mixture of ancestors gives rhizome complexity. For example, the species 645337915-30 consists of the ancestors 10 and 00 and two occurrences of 1001100111011 (the module species). The rhizome complexity can be defined through the frequency of ancestor sequences, which will be useful to discuss the deviation from ultrametricity.[2]

In the present Letter, we have introduced a novel dynamical model for evolution. We have discussed the appearance of the module sequence, gene duplication, and a possible practical merit of our dynamics in optimization problems.

For future studies, it may be important to include splitting of a sequence into two subsequences.[13] If we include this process, some "good" parts in a long sequence can be added to other creatures, as is often used in genetic technology.

In our present model fusion always occurs if the energy condition is satisfied. Since fusion itself is an error in the maintenance of genes, it must be more realistic to introduce a small rate of fusion error. With this extension, the evolutionary process exhibits much clearer stepwise changes than in Fig. 1: rapid-change eras and quasistationary phases, as has been discussed in Ref. 14 as punctuated equilibrium theory.[15]

Evolution provides a novel viewpoint for biological information processing such as autocatalytic,[16] immune, and neural networks.[17] The idea of genetic fusion will be useful to these problems.[18]

The authors would like to thank Y. Oono for critical comments, P. Davis for critical reading of the manuscript, and D. Farmer, C. Langton, and S. Rasmussen for useful discussions. This work was partially supported by Grant-in-Aids for Scientific Research from the Ministry of Education, Science and Culture of Japan. One of the authors (T.I.) is indebted to the Japan Society for the Promotion of Science for financial support.

[a]Electronic address: ikegami@gradis.scitec.kobe-u.ac.jp.
[b]Electronic address: d34205@tansei.cc.u-tokyo.ac.jp.

[1]See, for the evolution in a rugged landscape, S. A. Kaufmann and S. Levin, J. Theor. Biol. **128**, 11 (1987).

[2]*Spin Glass Theory and Beyond,* edited by M. Mezard, G. Parisi, and M. A. Virasoro (World Scientific, Singapore, 1988).

[3]P. W. Anderson, Proc. Natl. Acad. Sci. U.S.A. **80**, 3386 (1983).

[4]J. Holland, in *Escaping Brittleness in Machine Learning II,* edited by R. S. Mishalski, J. Carbonell, and T. M. Miechell (Kaufman, Ohio, 1986).

[5]See, e.g., J. M. Smith, *Evolutionary Genetics* (Oxford, Univ. Press, Oxford, 1989), Chap. 11.

[6]S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, Berlin, 1970).

[7]L. Margulis, *Symbiosis in Cell Evolution* (Freeman, San Francisco, 1981).

[8]We choose this model, not because of direct biological significance, but because it seems to have many local minima in a rugged landscape with translational invariance of gene sites. Detailed studies of the landscape of our model are left for future work.

[9]In the model, nothing stops the increase of bit length. We put a limit here externally at 30.

[10]Here we use the first algorithm for the reproduction, except for Fig. 3. Results, however, are essentially the same for either algorithm.

[11]Although the second condition itself can depend on the distribution of other species, it is rather robust with respect to the change of initial distribution in our simulation.

[12]After the fusion process hits the maximal length (30), mere mutation can occur to decrease the energy. The large plateau in CPU time comes from this saturation.

[13]This introduction of rearrangement is important to discuss the evolution of chromosomes (partitioned units of genetic codes in eukaryotes), since they split or join through evolution.

[14]N. Eldredge and S. J. Gould, in *Models in Paleobiology,* edited by T. J. M. Schopf (Freeman, San Francisco, 1972).

[15]See also K. Lindgren (to be published), for the application of gene duplication to evolution in the iterated prisoner's dilemma.

[16]J. D. Farmer, S. A. Kauffman, and N. H. Packard, Physica (Amsterdam) **22D**, 187 (1986).

[17]G. M. Edelman, *Neural Darwinism* (Basic Books, New York, 1987).

[18]T. Ikegami, Ph.D. thesis, University of Tokyo, 1989 (unpublished).