Computers Chem. Vol. 16, No. 2, pp. 171–182, 1992
Printed in Great Britain.

0097-8485/92 $5.00 + 0.00

EVOLUTIONARY DYNAMICS AND THE CODING STRUCTURE OF SEQUENCES: MULTIPLE CODING AS A CONSEQUENCE OF CROSSOVER AND HIGH MUTATION RATES*

P. HOGEWEG and B. HESPER
Bioinformatica, University of Utrecht, Padualaan 8, 83584CH Utrecht, The Netherlands

(Received 4 November 1991; in revised form 20 December 1991)

Abstract—In this paper we explore the influence of the dynamics of evolution on coding structures of sequences. We show that, in systems with crossover, high mutation rates cause the most conserved subsequences to be preferentially used as recognition sites for newly evolving sequences. In other words: "multiple coding" evolves in these systems. Multiple coding often does not increase the fitness of the population; nevertheless it is selected. By contrast, in systems without crossover, a low mutation rate causes multiple coding to be avoided, so that only single coding evolves. Again this "choice" is not reflected in the fitness of the population, but is dictated by the evolutionary dynamics. We conclude that the genetic operator crossover turns evolutionary processes in pattern detectors rather than optimizers.

1. INTRODUCTION

1.1. Evolution and the analysis of biomolecular sequences

DNA and RNA sequences contain a wealth of traces of the evolutionary processes by which they are shaped, hence evolutionary considerations play an important role in the informatic analysis of biomolecular sequences. On the one hand much work in computational molecular biology is concerned with reconstructing evolutionary pathways (phylogenies) on the basis of patterns of variation in sequences, and on the other hand the evolutionary history of sequences is used to unravel the functional significance of subsequences. Sequences which are similar in a variety of organisms are called "conserved" and are supposed to "code" for certain functions. Moreover, assumed conservation of secondary and tertiary structures of RNA and protein sequences is used in the prediction of these higher order structures from primary sequences.

In most research concerned with molecular sequences and evolution the observed molecular sequences are taken as a starting point for the investigations, and "conservation" and (random) "diversification in time" are the properties of evolutionary processes used in the analysis of the sequences; this analysis may lead to inferences about the evolutionary process. In this paper we will report on research which proceeds in the opposite direction: we start by defining a (simple) evolutionary process and investigate the patterns in the coding structure of the (artificial) sequences it generates; these patterns can be used to "debug" our ideas about evolutionary processes, and may serve as "search images" for the analysis of molecular sequences.

1.2. Coding structure of sequences

The term "coding structure of sequences" refers to the arrangement of "codes" in the sequence, using the work "code" in the meaning: a pattern which is recognized by some part of the system under consideration, such that the subsequent behaviour of that subsystem is determined by the recognition. Possible coding structures are: "one to one" or "single" coding: one recognition site is used for one process, and the process has no alternative recognition sites; "many to one" coding: various recognition sites are used for the same (or similar) processes; "one to many" or "multiple" coding: the same (or overlapping) subsequences are used as the most crucial recognition site(s) for several processes. (Note that "DNA makes RNA makes protein" implies that the same sequence is involved in multiple processes and thus should harbour codes for these processes. Trifonov (1989, 1991), Konings *et al.* (1987), Huynen *et al.* (1992) emphasized this phenomenon under the heading "multiple coding" or "multiple constraints", we use here the stronger definition as stated above.

The relationships between evolutionary dynamics and coding structure can be studied from different perspectives:

(1) How does the coding structure of sequences influence evolutionary optimization, i.e. how does it influence the shape of "fitness landscapes"

* The preliminary version of this work was presented during the *Open Problems in Computational Molecular Biology Workshop*, Telluride Summer Research Center, Telluride, CO, 2–8 June 1991.

(see e.g. Kauffman and Smith, 1986; Kauffman, 1989; Fontana and Schuster (1987), Fontana *et al.*, 1989, 1990; Schuster, 1991).

(2) Do evolutionary processes bias the coding structure and if so, how?

The two questions can be dealt with in combination (e.g. Kauffman and Johnson, 1991), but at present we think it useful to study them (also) separately. So we explore the second type of question and focus on the circumstances under which evolutionary dynamics will tend to generate either single coding or multiple coding. (Note: our experimental setup does not allow one to many coding, or other, more complicated, coding structures.)

Single coding is often assumed to be the default situation. It seems to be the most efficient coding structure because it allows for independent optimization of each process. Multiple coding is then seen as hampering evolutionary optimization, which is to be avoided as much as possible. ("Make the best of a bad job".) Nevertheless multiple coding, in the strong sense defined above, is observed in biotic sequences and is often striking. One example is the conserved sites of tRNA: they are conserved in all tRNA (prokaryotic and eukaryotic alike) and function in the transfer process; in eukaryotes the same sites are used as recognition sites for Pol III, i.e. in transcription. Moreover, the same sites appear to bind several DNA binding factors and, in reconstructing the evolutionary history of U1 snRNAs, we noted the crucial roles of again the same sites in the folding of ancestral U1 molecules (Hogeweg and Konings, 1985). Other examples include regulation sites (coded in terms of conserved secondary structure) within the protein coding region in e.g. Lentiviruses (Saltarelli *et al.*, 1990; Konings, 1992), and overlapping reading frames in various viruses. In the experiments we report here, the tRNA examples may serve as a prototype.

2. METHODS

Our method for studying the problems sketched above is to define paradigm systems of "evolving sequences" and to observe the coding structure which evolves. The paradigm system is not supposed to match any real system closely, but it tries to study the consequences of a simple model of evolutionary dynamics in isolation. In particular, because we want to study exclusively the informatic constraints the evolutionary dynamics imposes on the coding structure of a sequence, we ignore not only the fact that sequences have physical–chemical properties, but even make an effort to exclude any implicit non-evolutionary properties in our experiments. Note that calling recognition sites "codes" invites this (odd) behaviour: the relationship between "code" and "meaning" is supposed to be an arbitrary convention. In order not to confuse this paradigm (digital) system with a "wet" system, its constituent entities will be denoted with capitals.

Our evolving GENOME consists of a 4 letter SEQUENCE which represents two GENEs. GENE-1 (here of length 50) evolves first, and is supposed to confer fitness ($F1$) to its GENOME according to its match on some preassigned (4 letter) SEQUENCE. This match is weighted, relative to a preassigned weight vector, which is constructed by assigning a large weight (in the reported experiments 30) to a stretch (here of length 10) of adjacent positions, and by randomly assigning small weights (here between 0 and 5) to all other positions. The stretch represents a "code" (as defined above) for GENE-1. GENE-2 (here of length 20) evolves later. It is supposed to need the recognition of the presence of GENE-1 for its function. Recognition is by matching a subsequence to GENE-1. The fitness conferred to the GENOME depends on the length of the match (LM) (up to a maximum), i.e. by the HILL function:

$$F2 = \frac{LM^2}{K^2 + LM^2}, \quad K = 12.$$

Total fitness of the GENOME is simply the weighted sum of the fitness of both genes: $F = F1 + R \times F2$, $R = 0.5$ or varying.

Evolution of the sequences is defined by a simple "genetic algorithm"; a population of GENOMEs (population size 100) reproduce and compete. During reproduction the sequences are subjected to two "genetic operators":

- "mutation": randomly chosen "letters" (from the 4 letter alphabet) are put in randomly chosen positions in the GENOME. Mutation rate (MU) is defined as the probability of a (next) mutation taking place in the GENOME under consideration (i.e. mutations take place until a drawing from a uniform distribution between 0 and 1 is greater than MU). Note that a mutation may not change the letter, or it may be a "back mutation". MU is varied from 0.1 to 0.85, i.e. between *ca* 0.1 and 5 mutations per reproduction.
- "Crossover"; if crossover occurs, reproduction is by 2 GENOMEs and GENE-1 is taken from one GENOME, GENE-2 from the other one, i.e. crossover is always between the 2 GENEs, they are not supposed to be adjacent. Crossover probability is set to CO = 0 (no crossover) or CO = 0.5 (equal probability that the new GENOME is assembled from one to two parent GENOMEs).

Competition (selection) takes place through "non-survival of the non-fittest" as well as "reproduction of the fittest".

- "Non-survival of the non-fittest": At every generation 10 GENOMEs (10% of the population) decay. Probability of decay is governed by

$$\frac{1/F_i}{\Sigma_j\, 1/F_j},$$

where F_i and F_j are the fitness of GENOME i or j and the sum is taken over the entire population.

- "Reproduction of the fittest": Probability of reproduction is likewise governed by $F_i/\Sigma_j F_j$; when crossover occurs two parent GENOMEs are thus selected to reproduce one offspring.

  Note that both selection processes are weak, i.e. selection is subjected to high noise levels.

In order to exclude influences of particular sequences of letters (which greatly determine the probability of matching (see e.g. Pevzner *et al.*, 1989)), i.e. to exclude all implicit "physical properties", we perform our experiments in "pairs of runs" which use the same (randomly constructed) target SEQUENCE, weightvector and initial GENOME. In one run of the pair the CODE of GENE-1 (e.g. the high weights) is located at position 10–20, and in the other run the CODE is located at position 30–40. We study where GENE-2 will match GENE-1.

This basic paradigm system is extended in the following ways:

1. longer and shorter GENOMEs are used.
2. Both GENEs evolve simultaneously.
3. 1-D spatial structure is added by allowing a decayed GENOME to be replaced only by offspring from adjacent GENOMEs.
4. The ratio of contribution in both GENEs is varied, starting with a very low contribution of GENE-2 which is (very slowly) increased.

(a)

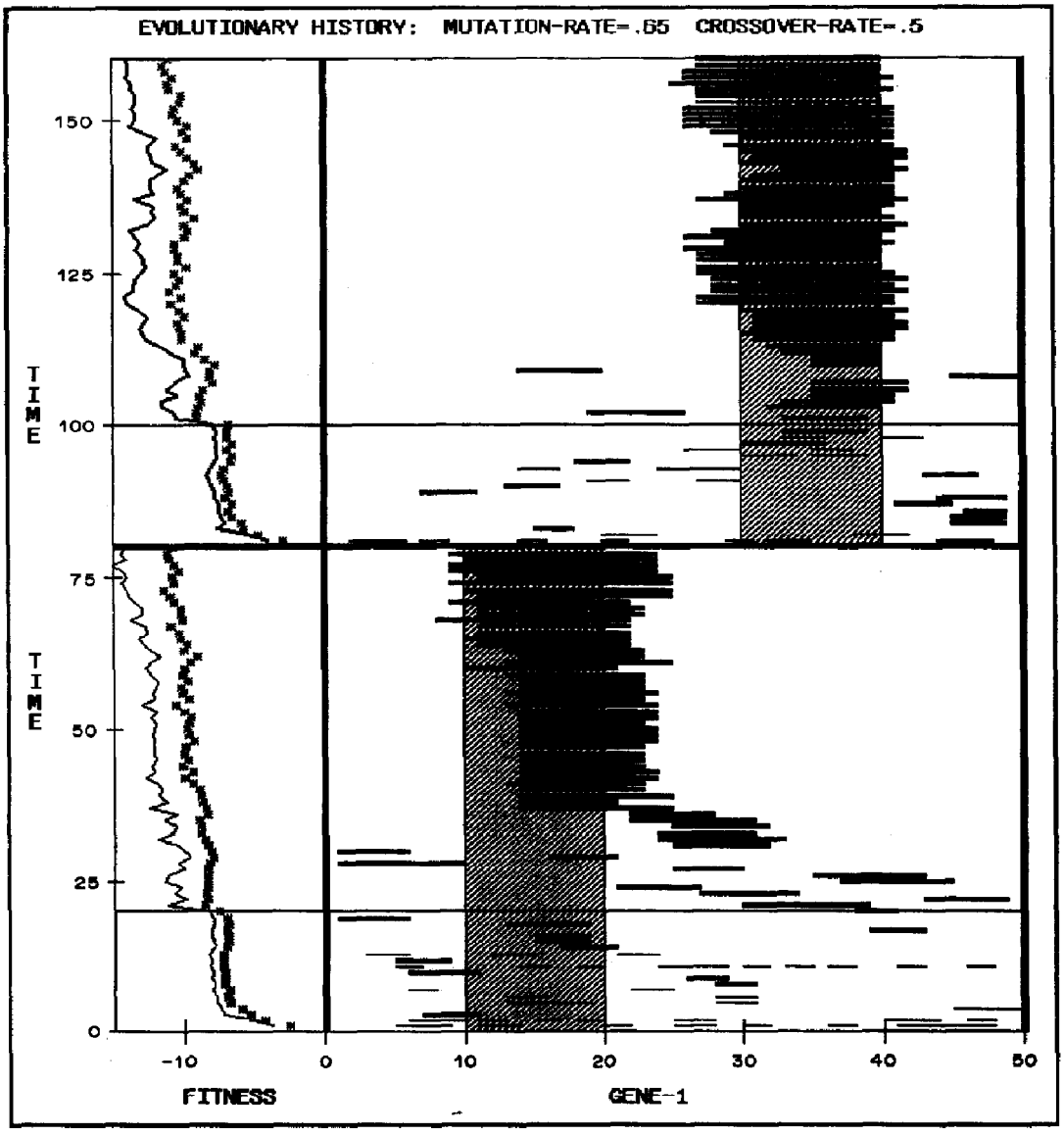

Fig. 1(a)—*legend on p. 175*

We will report mainly on the basic model, but will refer to these extensions where appropriate.

3. RESULTS

3.1. *Evolution of multiple coding: a case study*

Figure 1 documents the evolutionary history of one pair of runs with crossover and a mutation rate of MU = 0.65 (i.e. *ca* 1.4 mutations/SEQUENCE/reproduction). Figure 1(a) shows the fitness of the fittest SEQUENCE, the mean fitness and the location of the maximal match of GENE-1 and GENE-2 (by horizontal bars) of the best SEQUENCE over time (per 100 generations). Fairly soon after the switching on of GENE-2 the location of the match converges to the location of the (primary) CODE of GENE-1 (this location is indicated by the hatched areas). In other words: multiple coding evolves.

Examining the final population [Figs 1(b)–(d)] we observe:

— multiple coding is fixated in the population, but the least fit has a poor match or, sometimes, alternative maximal matches [Fig. 1(b)].
— variation in the population is fairly large, except for the (multiple) CODE regions, i.e. the CODEs show up as relatively conserved regions [Fig. 1(d)];
— the fitness landscape is "rugged" on the scale of the region occupied by the final population, i.e. similarity with the fittest sequence and fitness are unrelated except for a few of the fittest GENOMEs which sit near the "top" of the fitness landscape to which the population has evolved [Fig. 1(c)].

(b)

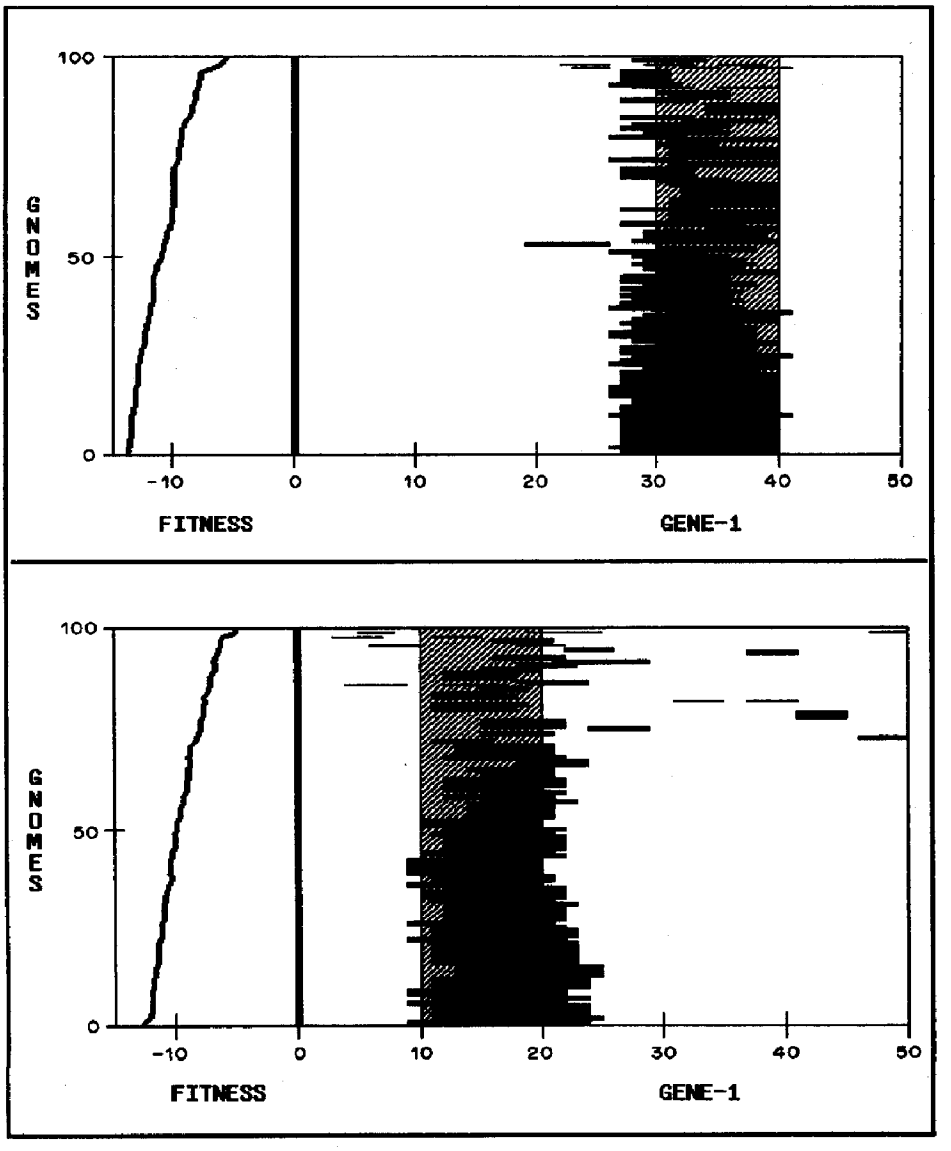

Fig. 1(b)—*legend opposite*

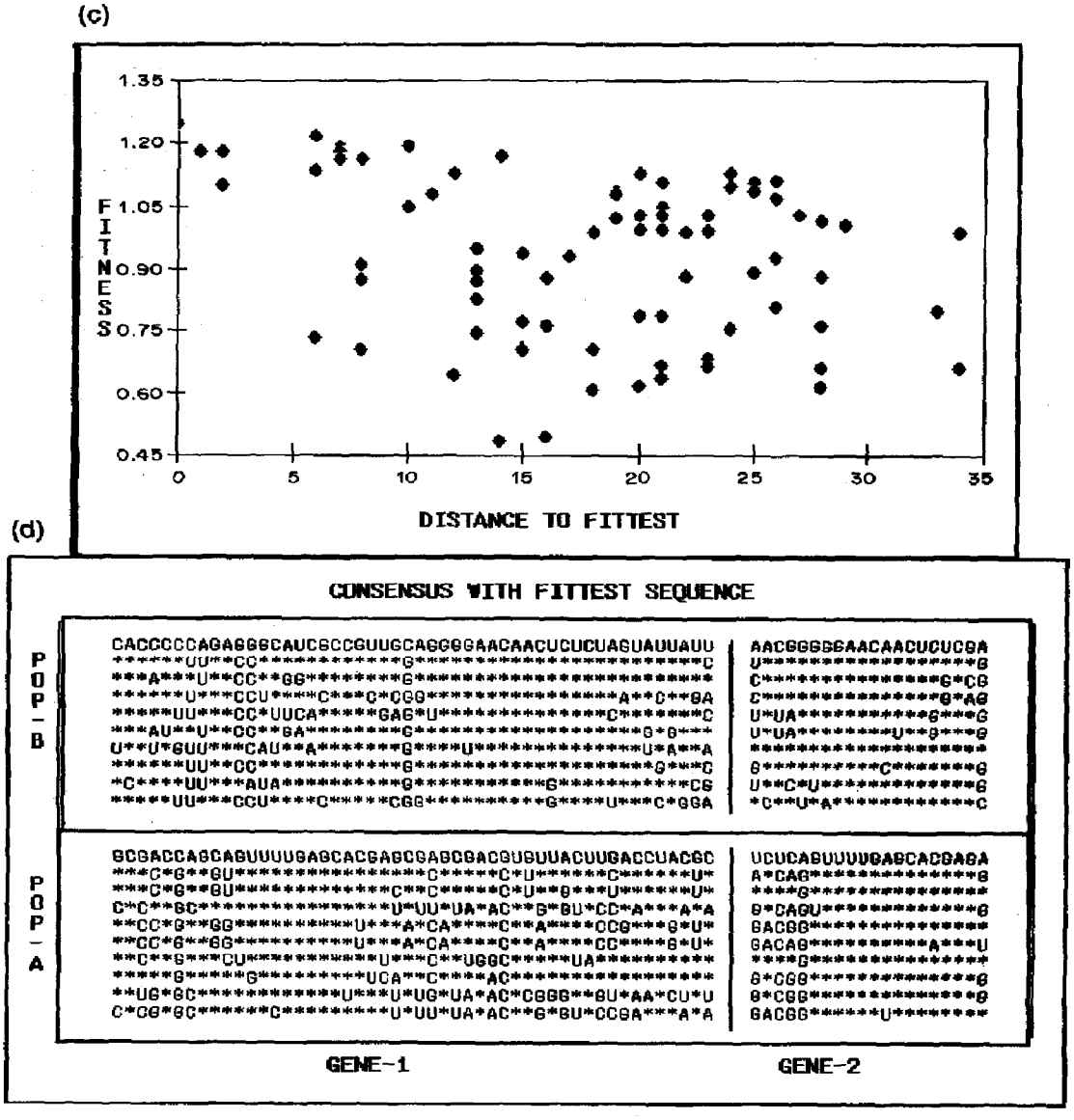

Fig. 1(c, d)

Fig. 1. Case study of the evolution of multiple coding. The evolutionary history of one pair of runs with crossover and a mutation rate of MU = 0.65 (i.e. *ca* 1.4 mutations/SEQUENCE/reproduction). For 2000 generations the fitness is only determined by GENE-1; in the next 6000 generations it is determined by both GENE-1 and the maximal match of GENE-2. The runs are initialized identically (i.e. with the same initial population, target sequence and basic weight vector) except for the position of the CODE (i.e. the high weights) for GENE-1: it is localized either at position 10–20 or at position 30–40. (a) Evolutionary development in time (*y*-axis, bottom to top). Fitness on the negative *x*-axis; dots: mean fitness; line: maximum fitness. Positive *x*-axis represents GENE-1. Match of GENE-2 to GENE-1 of fittest sequence is depicted by showing its location and length by a horizontal bar. (b) Final populations of both runs, ordered according to fitness. Representation as for (a). (c) Fitness landscape. *x*-axis: Manhattan distance to fittest sequence. *y*-axis: fitness. (d) Sequence variation: fittest sequence and every 10th sequence of the ordered final populations is given. Conserved sites are detectable.

Figure 2 generalizes the findings of the case study by summarizing the results of 25 replicate pairs of runs with the parameters of the case study above: in the vast majority of cases multiple coding evolved preferentially in the sense that there is a considerable overlap of the two codes.

3.2. Multiple coding with crossover and high mutation rates

The evolution of multiple coding depends on the mutation rate. Under crossover, a clear preference for multiple coding exists for mutation rates between MU = 0.5 and MU = 0.85, (i.e. for an average number of mutations between 1 and 5) (Fig. 3). The preference for multiple coding at high mutation rates is present, but less pronounced when the crossover operator is not used in the evolutionary process (Fig. 4). For the highest mutation rates (MU = 0.8 to 0.85) multiple coding is a necessary condition for a stable selection of GENE-2; for the lower mutation rates this is not the case. Examination of the evolutionary histories reveal that multiple coding develops whenever the selection is relatively slow, but that arbitrary coding develops when soon after GENE-2 is switched on a relatively very long match (accidentally) occurs and takes over the population before it competes with another coding structure. Indeed, at lower mutation rates, (i.e. MU = 0.1 to MU = 0.45) the absence of coding preference may be for lack of alternatives before a code is fixated in the population.

3.3. Single coding at low mutation rates without crossover

Interestingly, without crossover and at low mutation rates, multiple coding is clearly avoided, i.e. single coding is preferred. This is because under these circumstances mutation is a limiting factor for evolutionary adaptation. By avoiding the CODE of GENE-1 the search for a match between GENE-1 and GENE-2 exploits the mutations in both genes, i.e operates under a double mutation rate, and is therefore more likely to find a match. On the other hand, when crossovers do occur a low mutation rate is much less a limiting factor, because the search may operate in parallel, combining partial solutions (matches) (cf. Holland, 1976). However, recombination by crossover favours multiple coding (see below). Thus the apparent absence of coding preference at low mutation rates with crossover (Fig. 3) can be the result of opposing coding preferences. Because of this parallel operation, adaptation is much more effective with crossover than without, also for high mutation rates: without crossover only poor selection for GENE-2 occurs at MU $\geqslant$ 0.7 (Fig. 4.)

3.4. Multiple coding and fitness

According to the definition of fitness, fitness is not *a priori* affected by the coding structure: the same fitness scale exists for both multiple and single coding. Nevertheless the coding structure is conceivably determined by attainable or maintainable fitness. In other words, are fitness and multiple coding correlated? It turns out that for relatively low mutation rates, in which a preference for multiple coding occurs, (e.g. MU = 0.5) there is no correlation (correlation coeff. $C = 0.13$, $N = 30$) between the overlap of the codes and maximal fitness in the population. In the case of very high mutation rates (e.g. MU = 0.85) there is a correlation: without multiple

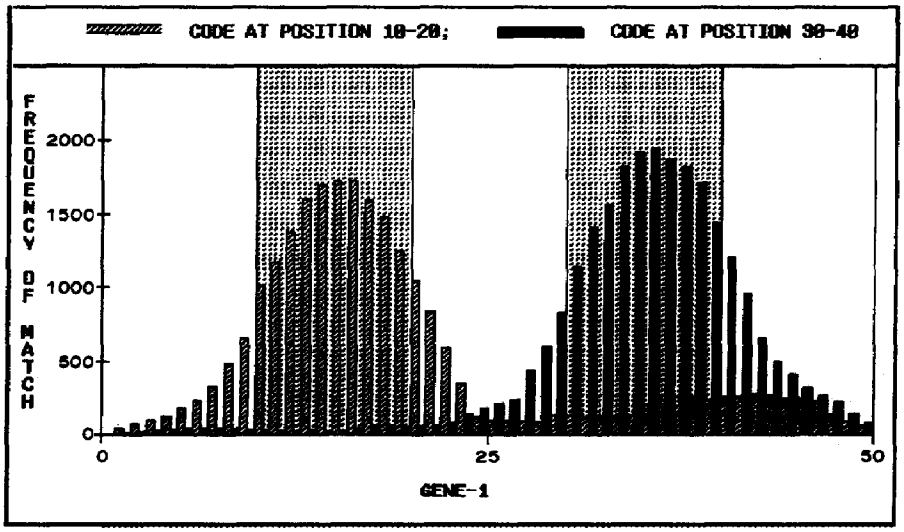

Fig. 2. Frequency of occurrence of multiple coding (MU = 0.65; CO = 0.5) x-axis represents GENE-1; shaded regions are the alternative locations of the CODE of GENE-1 (pos. 10–20 and pos. 30–40). Bars represent the frequency in which position of GENE-1 is part of the longest match with GENE-2. Shaded bars: primary CODE at position 10–20. Dark bars primary CODE at position 30–40. Results of 25 pairs of runs summed.

coding the matching of the genes cannot be maximized by selection, and accidentally arising long matches cannot be maintained. At intermediate mutation rates (e.g MU = 0.65) we see very high fitness values co-occurring with considerable overlap of the codes, but the fitness attained in most evolutionary histories occurs with single or multiple coding alike (except for the fact that single coding is much rarer). Thus we conclude that the preference of multiple coding is not primarily due to the attainability or maintainability of a higher fitness but is due to some other aspect of the evolutionary dynamics.

3.5. Multiple coding and hybrid unfitness

Competition between single and multiple coding can be observed nicely by imposing a spatial structure on the population. We used a 1-D spatial structure in which a decaying GENOME is replaced by offspring of nearby GENOMEs (chosen from 3 on either site according to our fitness criterion). In this system we observe that different single coding solutions can coexist in the population even if there is quite some fitness difference, but a multiple coding solution takes over the population quickly. This is because a hybrid offspring which inherits GENE-2 from the multiple

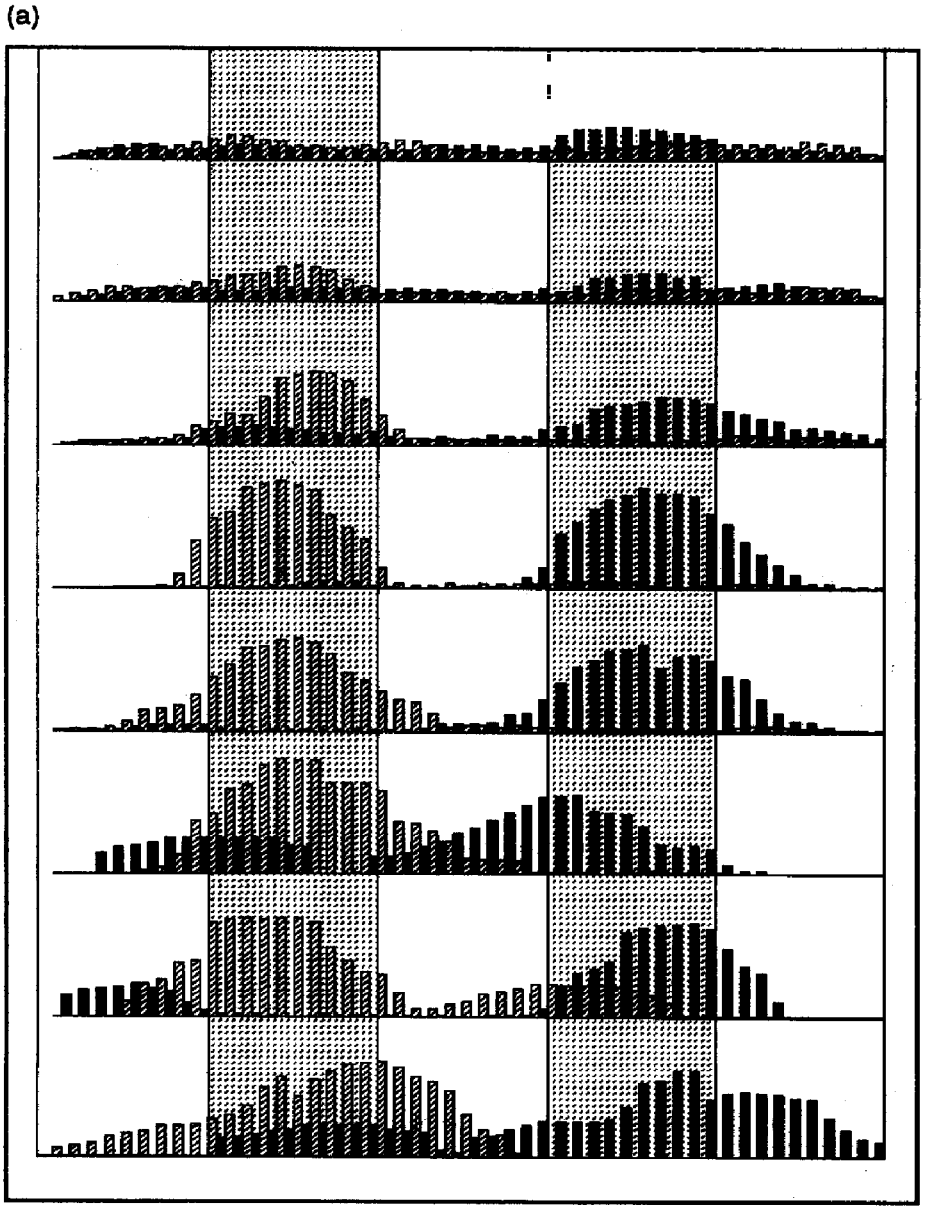

Fig. 3(a)—*legend overleaf.*

(b)

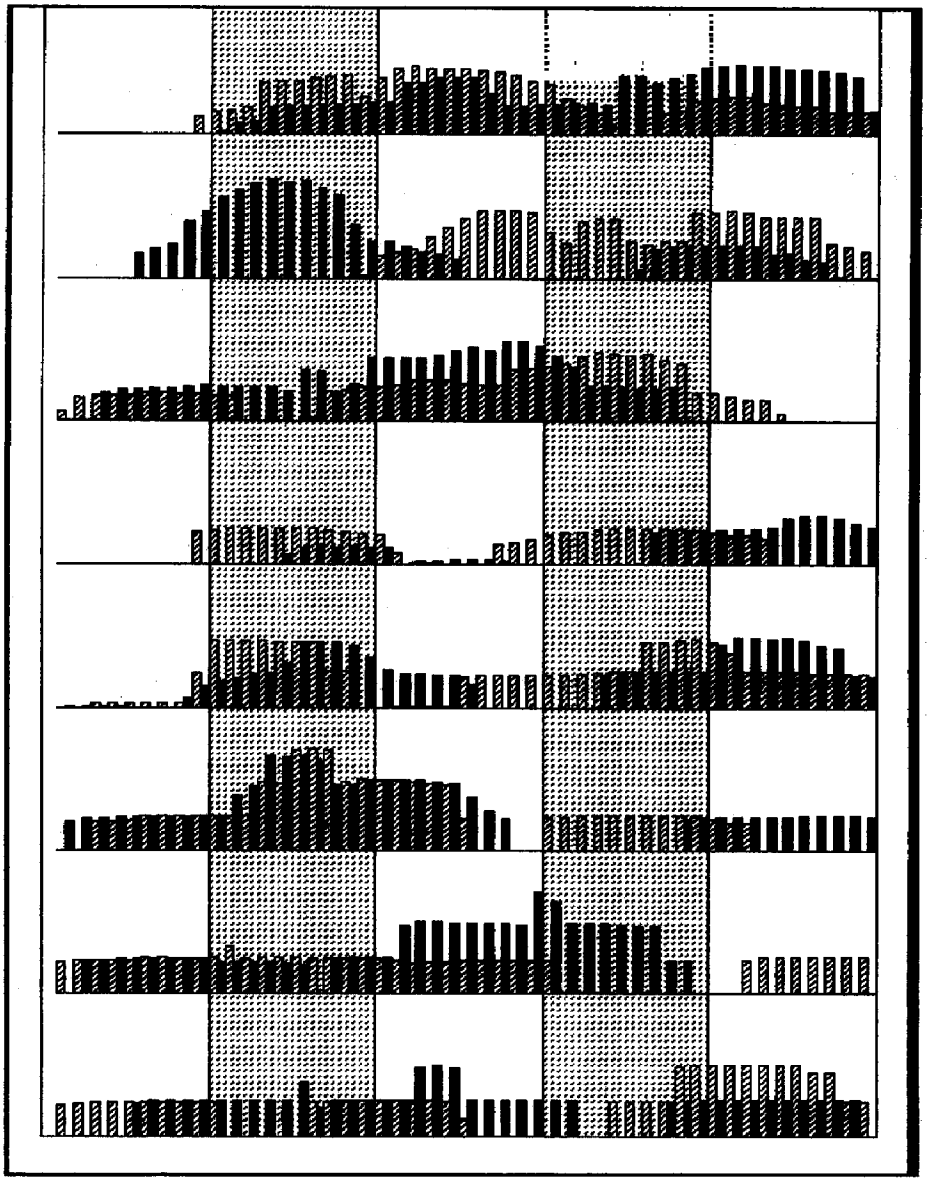

Fig. 3(b)

Fig. 3. Dependence of coding structure on mutation rate: CROSSOVER (CO = 0.5). Representation as in Fig. 2. Five pairs of runs summed for each mutation rate. (a) Top to bottom: MU = 0.85 to MU = 0.5 (i.e. average number of mutations/generation 5 to 1): multiple coding is weakly favoured. (b) Top to bottom: MU = 0.45 to MU = 0.1: arbitrary coding.

coding GENOME is likely to do well since the sequence to which it should match is present in the other GENOME because of its coding function for GENE-1. Contrariwise an offspring which inherits its GENE-2 from a single coding GENOME will generally have very low fitness, i.e. it exhibits "hybrid unfitness". Hybrid unfitness spatially isolates subpopulations which can coexist indefinitely. We conclude that in a population with recombination, relatively conserved parts of the GENOME are to be preferred as CODEs because they allow the spread of new GENEs in the population.

3.6. Multiple coding and intrapopulation variation

In the system described above, multiple coding evolves if and only if there is a large intrapopulation variation due to high mutation rates [Fig. 1(d)]. Should we conclude that low intrapopulation variation multiple coding cannot be caused by the evolutionary dynamics? The spatial experiments (section

3.5) show that this is not the case: multiple coding evolves in populations which are very homogeneous in the end. This is because in the spatial system: (1) multiple coding is selected at much lower mutation rates (i.e. MU = 0.2–MU = 0.5) due to the less effective selection caused by the small subpopulation from which replacements are selected; (2) small subpopulations with multiple coding take over the entire population.

4. DISCUSSION

We will discuss now how our results fit into and extend different conceptual frameworks which have been proposed previously in order to understand evolutionary dynamics.

4.1. Information threshold

Eigen and Schuster (Eigen and Schuster, 1979; Eigen *et al.*, 1988, 1989; Nowak and Schuster, 1989) have shown that evolutionary optimization can only take place if the mutation rate does not exceed a certain threshold; above this threshold evolution is just equal to random walk. They point out that the error rate per nucleotide is fixed at a certain level by physical–chemical properties; this means that the length of the sequences, and therewith the amount of

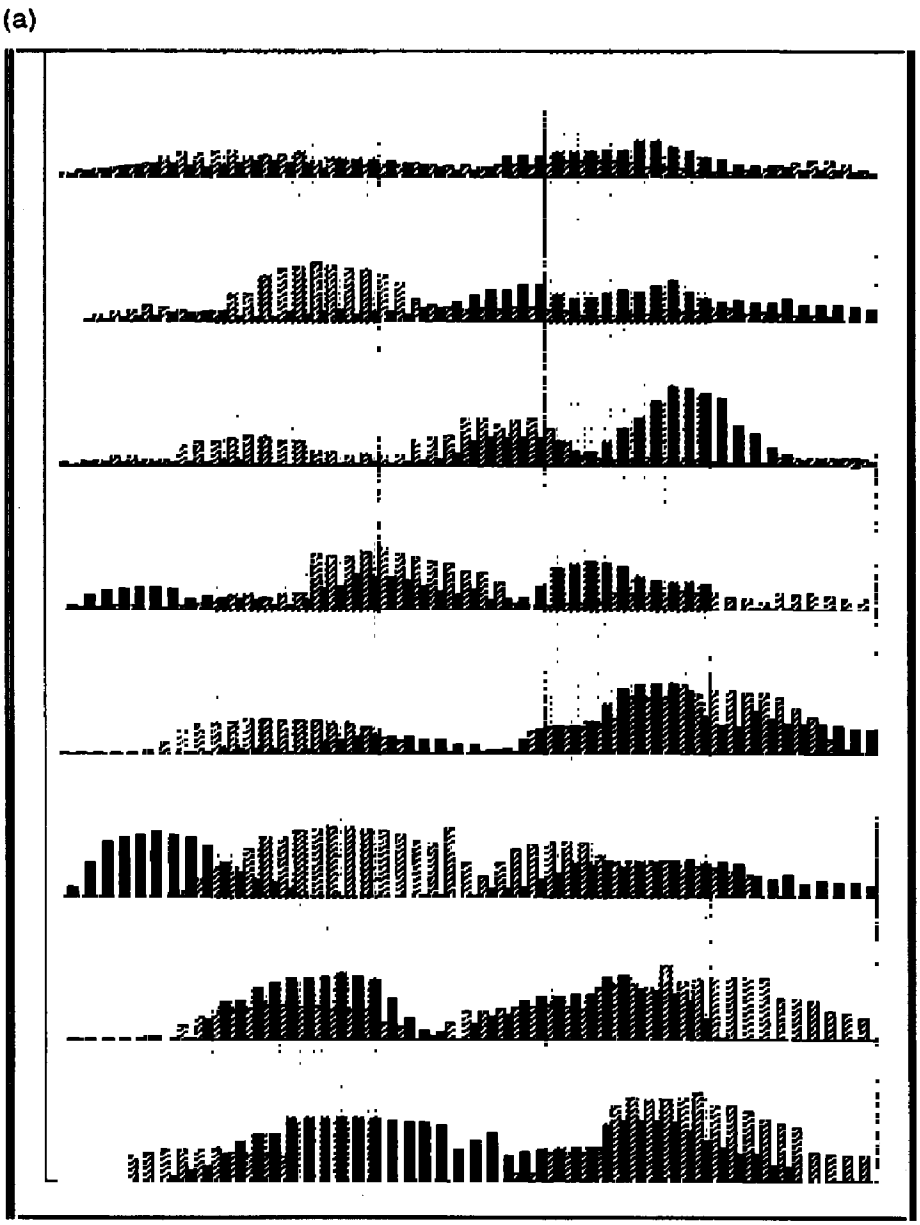

Fig. 4(a)—*legend overleaf.*

(b)

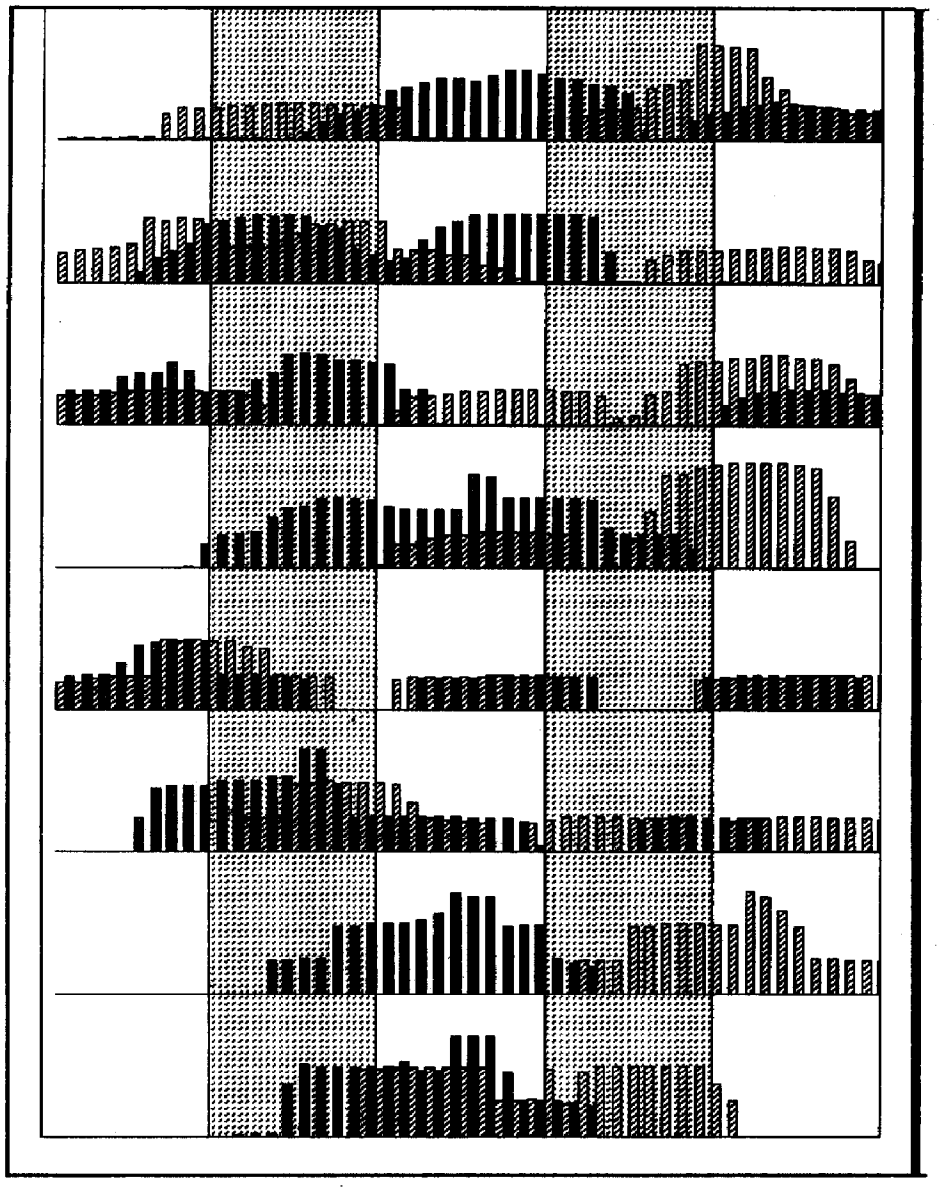

Fig. 4(b)

Fig. 4. Dependence of coding structure on mutation rate: NO CROSSOVER (CO = 0) Representation as for Figs 2 and 3. Five pairs of runs summed for each mutation rate. (a) Top to bottom: MU = 0.85 to MU = 0.5 (i.e. average number of mutations/generation 5 to 1): multiple coding is weakly favoured. (b) Top to bottom: MU = 0.45 to MU = 0.1: single coding favoured (multiple coding avoided).

information stored, is limited, hence the concept of information threshold.

Our results fit into this conceptualization nicely. Multiple coding evolves if and only if the evolutionary process operates not too far from the information threshold; at lower mutation rates coding is initially arbitrary, and at very low mutation rates, it even selects unique coding preferentially. This means that at low mutation rates the length of the crucial parts of the sequence increases, and that the evolutionary process moves toward the information threshold, where multiple coding is first preferred and later (i.e. at higher mutation rates) delays the onset of non-evolvability.

Nevertheless the information threshold is apparently not the mechanism which causes multiple coding because:

(1) multiple coding is preferred before the storage of information is limited; and

(2) multiple coding is enhanced by including a second genetic operator into the evolutionary process, i.e. crossover; in contrast the evolutionary model which gives rise to the information threshold is based on point mutations only and we have seen that multiple coding is only weakly preferred under these circumstances.

4.2. *Crossover as genetic operator*

Holland (1976) has recognized the power of crossover as a genetic operator. The general optimization technique called "genetic algorithms" is mainly based on this insight. By crossover different schemata can evolve in parallel, and be combined into a better solution. Thus, global optimization may be achieved in rugged fitness landscapes. Our experiments show that crossover indeed enhances optimization but also multiple coding, although multiple coding does not constitute a global optimum (its fitness is not higher than that of other peaks which represent unique coding solutions). So, the effect of this operator in an evolutionary process should be reexamined. Our 1-D spatial experiments clearly show that the multiple coding GENOME is less likely to suffer from "hybrid unfitness" when crossed over with some other GENOME, because the CODE of GENE-1 will be identical in nearly all cases. Thus, a gene which recognizes an invariant part of the genome will spread. We conclude that the crossover genetic operator turns evolutionary processes into pattern detectors rather than optimizers.

4.3. *Evolution as pattern processing*

Evolution is most often seen as an optimization process. It is well known that pattern detection or pattern recognition can be achieved by optimization procedures. For example pattern detection in neural networks is based on optimization (e.g. Hopfield, 1984; Ackley *et al.*, 1985). Recently we have proposed that it might be a good heuristic to view evolution as pattern processing (Hogeweg and Hesper, 1990) rather than as optimization. We argued that such a shift in viewpoint does not require new assumptions about evolutionary processes, but helps to highlight properties of evolutionary dynamics which cannot be expressed in terms of higher fitness. The preference for multiple-multiple coding reported here is an example. The evolutionary process can be said to "recognize" preferentially the relative invariances in a variable world (multiple coding at high mutation rates and crossover), and relative variability in an invariant world (single coding at low mutation rates). In this test situation invariance could only be achieved by being important for fitness: hence multiple coding. In more complex situations other sources of invariance exist, e.g. by self-organization where many different initial conditions or interaction structures may converge to similar macropatterns. Therefore we expect evolutionary processes to "recognize" such invariances. We have observed such recognition of macropatterns in the case of spatially interacting cyclically catalysing molecules (Hypercycles) (Boerlijst and Hogeweg, 1991a,b).

5. CONCLUSIONS

We have shown that in a system in which no external biases are imposed on an (artificial) Darwinian evolutionary process, the dynamics of the evolution itself imposes patterns on the coding structure of the replicating units. In particular the evolutionary process appears to evolve toward a situation in which multiple coding is favoured by the evolutionary dynamics. The preference for multiple coding is not reflected in the achieved fitness. We have proposed that viewing evolution as a pattern detection process rather than as an optimization process, might be helpful for detection of such additional patterns in evolution. Obviously evolutionary processes operate under external constraints, imposed by the physical–chemical properties of the molecules involved. Nevertheless, because the coding structure evolves and is not imposed, such constraints may play a less important role than is often assumed. In any case, knowing the biases imposed by the evolutionary process itself may provide us with useful search images for the study of its final product: present day molecular sequences.

Acknowledgement—We thank Martijn Huynen for useful discussions.

REFERENCES

Ackley D. H., Hinton G. E. and Sejnovski T. J. (1985) A learning algorithm for Boltzmann machines. *Cognitive Sci.* **9**, 147–169.

Boerlijst M. A. and Hogeweg P. (1991a) Spiral wave structure in prebiotic evolution: Hypercycles stable against parasites. *Physica D* **48**, 17–28.

Boerlijst M. A. and Hogeweg P. (1991b) Selfstructuring and selection: spiral waves as a substrate for prebiotic evolution. In *Artificial Life II* (Edited by Langton C. G., Taylor C., Farmer J. D. and Rasmussen S.), pp. 255–276. Addison Wesley, Reading, MA.

Eigen M. and Schuster P. (1979) *The Hypercycle: A principle of Selforganisation.* Springer Verlag, Berlin.

Eigen M., McCaskill J. S. and Schuster P. (1988) The molecular quasispecies. *J. Phys Chem.* **92**, 6881–6891.

Eigen M., McCaskill J. S. and Schuster P. (1989) The molecular quasispecies. *Adv. Chem. Phys.* **75**, 149.

Fontana W. and Schuster P. (1987) A computer model for evolutionary optimisation. *Biophys. Chem.* **26**, 123.

Fontana W., Schnabl W. and Schuster P. (1989) Physical aspects of evolutionary optimisation and adaptation. *Phys. Rev. A.* **40**, 3301.

Fontana W., Schuster P., Stadler P and Weinberger E. (1990) Characterisations and quantitative evaluation of RNA folding landscapes. Preprint.

Hogeweg P. and Konings D. A. M. (1985) U1 snRNA: the evolution of its primary and secondary structure. *J. Mol. Evol.* **21**, 323–333.

Hogeweg P. and Hesper B. (1990) Evolution as pattern processing: TODO as substrate for evolution. In *From Animals to Animats* (Edited by Meyer J. A. and Wilson S. W.), pp. 492–497. MIT Press, Cambridge, MA.

Holland J. H. (1976) *Adaptation in Natural and Artificial Systems*. Univ. of Michigan Press, Ann Arbor.

Hopfield J. J. (1984) Neurons with graded response have collective computational properties like those of two-state neurons. *PNAS* **81**, 3088–3092.

Huynen M. A., Konings D. A. M. and Hogeweg P. (1992) Equal G and C contents in histone genes indicate selection on mRNA secondary structure *J. Mol. Evol.* In press.

Kauffman S. A. and Smith R. G. (1986) Adaptive automata based on darwinian selection. *Physica D* **22**, 68–82.

Kauffman S. A. (1989) Adaptation in rugged fitness landscapes. In *Lectures in the Sciences of Complexity* (Edited by Stein D. L.), pp. 527–618. Addison Wesley, Reading, MA.

Kauffman S. A. and Johnson S. (1991) Coevolution to the edge of chaos: coupled fitness landscapes, poised states and coevolutionary avalanches. In *Artificial Life II* (Edited by Langton C. G., Taylor C., Farmer J. D. and Rasmussen S. pp. 325–370. Addison Wesley, Reading, MA.

Konings D. A. M. (1992) *Comput. Chem.* **16**, 153–163.

Konings D. A. M., Hesper B. and Hogeweg P. (1987) Evolution of primary and secondary structures of the 2Ea mRNA's of the Adenovirus. *Mol. Biol. Evol.* **4**, 300–314.

Nowak M. and Schuster P. (1989) Error thresholds of replication in finite populations: mutation frequences and the onset of Muller's Rachet. *J. Theor. Biol.* **137**, 375.

Pewzner P. A., Borodovsky M. Y. and Mirornow A. A. (1989) Linguistics of neeleotide sequences I: The significance of deviations from mean statistical characteristics and the prediction of the frequency of the occurrence of words. *J. Biomol. Structure Dynamics* **6**, 1013–1026.

Saltarelli M., Querat G., Konings D. A. M., Vigne R. and Clements J. E. (1990) Nucleotide sequence and transcriptional analysis of molecular clones of CAEV which generate infectious virus. *Virology* **179**, 347–364.

Schuster P. (1991) Complex optimisation in an artificial RNA world. In *Artificial Life II* (Edited by Langton C. G., Taylor C., Farmer J. D. and Rasmussen S.). pp. 277–292. Addison Wesley, Reading, MA.

Trifonov E. N. (1989) Viewpoint: the multiple codes of nucleotide sequences. *Bull. Math. Biol.* **51**, 417–432.

Trifonov E. N. (1991) Sequence ontogenesis and spacial separation of overlapping messages. In *Open Problems of Computational Molecular Biology—Book of Extended Abstracts and Topics for Discussion*, pp. 58–62.