J. theor. Biol. (1993) **165**, 251–267

Multiple Coding and the Evolutionary Properties of RNA Secondary Structure

MARTIJN A. HUYNEN†§||, DANIELLE A. M. KONINGS‡ AND PAULIEN HOGEWEG†

† *Bioinformatics Group, University of Utrecht,* 3584 *CH Utrecht, The Netherlands and* ‡ *MCD Biology, University of Colorado, Boulder, CO* 80309, *U.S.A.*

(*Received and Accepted on* 12 *March* 1993)

This article evaluates evolutionary properties of the transition from RNA primary sequence to RNA secondary structure. It focuses on the restrictions that the conservation of a protein code in an RNA sequence puts on its potential to evolve towards a specific secondary structure. Restricting the mutations to those that do not affect the coding for a protein restricts both the accessibility and the connectivity of the sequence space. The accessibility is restricted because only certain point mutations are allowed. The connectivity is restricted because no insertions and deletions are allowed. Simulating an evolutionary search process for a specific secondary structure shows that (i) the reduction of allowable point mutations allows for adaptation to some large-scale topology, but strongly reduces the possibility of small-scale adaptations, (ii) the abolition of insertions and deletions has very little effect on the results of the search process.

During the evolutionary search process for a secondary structure with a specific topology and a high frequency of base-pairing the quasispecies moves into a subspace in which the similarity between secondary structures of neighboring sequences is relatively high. Increased similarity between second structures of neighboring sequences is also found in the Rev responsive element (RRE) in the lentiviruses Caprine arthritis-encephalitis virus and Visna virus. In these viruses a biased nucleotide frequency in the RRE region suggests that selection for the RRE RNA secondary structure affects the amino acid sequence of the *env* gene. Our results show a variation in the ruggedness of fitness landscapes which are based on a high degree of epistatic interactions. Fitness landscapes play an essential role, not only in biotic evolution, but also in all kinds of optimization processes (Hill Climbing, Simulated Annealing, Genetic Algorithms, etc). Variation in their ruggedness should therefore be taken into account in the analysis of these processes.

Introduction

Species with a low quality of replication can only maintain a short genome (Eigen & Schuster, 1979), that may be too small to store all the necessary information in a sequential manner. To increase the amount of information that can be stored, the quantity of information per length unit has to be increased; i.e. parts of the genome have to code for multiple functions. Such multiple coding has indeed been observed in viruses, which are notorious for their high mutation rate. A well-known example is

§ Present address: Theoretical Biology, T-10 Mall Stop K710, Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.

|| Author to whom correspondence should be addressed.

0022–5193/93/022251+17 $08.00/0

the overlapping of protein-coding regions such as is found in the bacteriophage φX174 (Barell *et al.*, 1976). Here the multiple coding occurs at one functional level, i.e. at the protein-coding level. It might, however, also occur at different functional levels. In some lentiviruses like HIV, Caprine arthritis-encephalitis virus (CAEV) and Visna virus, the Rev response element (RRE), which is an RNA secondary structure involved in regulating the transport of unspliced mRNA to the cytoplasm, is located within the coding region of the *env* gene (Saltarelli *et al.*, *1990*; for a review of HIV: Cullen, 1991). A shortage of "genome space" is not the only reason why multiple coding occurs; there may be a need for the spatial proximity of different "codes". In a number of retroviruses, pseudoknots which are located in the overlap of reading-frames facilitate frameshifts during translation (reviewed in Hatfield & Oroszlan, 1990). Another example of multiple coding can be found in the coding region of histone genes where the balancing of G and C contents points to a constraint of the secondary structure of the mRNA (Huynen *et al.*, 1992). Moreover, multiple coding may be favored by evolutionary dynamics. As is shown by Hogeweg & Hesper (1992), parts of the genome that are relatively well conserved can serve as recognition sites for the development of new functions.

In this paper we address the question, to what extent does multiple coding affect, or even frustrate, the functionality of its constituents? We tackle this question by analyzing in what way multiple coding affects the evolutionary search process. Evolution can be viewed as the migration of a quasispecies through a sequence-space. The subspace that can be reached directly from the sequences within the quasispecies is determined by the Genetic Operators, i.e. the types of mutations (point mutations, insertions, deletions, recombination, etc) that can be used. Multiple coding can influence the extent to which these Genetic Operators can be used. The presence of coding for a protein within a piece of RNA/DNA will, if the protein sequence has to be preserved, greatly reduce the number of allowable point mutations. Moreover, it will reduce possibilities for insertions or deletions since they are likely to cause frameshifts.

The mapping from RNA primary sequence to RNA secondary structure is a nice example of a non-linear genotype–phenotype relation in which the constituents of the genotype (RNA nucleotides) have to be well co-adapted in order to create some desired phenotype. The mapping is particularly intriguing since it is highly redundant; in other words, multiple primary sequences give rise to the same secondary structure (suboptimal foldings not included). Therefore restrictions on the primary sequence do not necessarily limit the secondary structures that can be attained. In the past decade algorithms have been developed which give reasonable estimates of the secondary structure of relatively short RNA molecules (<300). This makes it possible to study the evolution of RNA secondary structure by simulation (Fontana & Schuster, 1987). In the present paper we will focus on the restrictions that coding for a protein imposes on the possibility of development of RNA secondary structure. This does not mean that we believe that the development of multiple codes might not be in a different order, or simultaneous. This approach will enable us to get some idea about the evolutionary consequences of multiple coding. We first analyze the amount of change that the various Genetic Operators cause in secondary structure.

Second, we simulate an evolutionary process in which different "sets" of Genetic Operators are used, and analyze their outcomes. Finally, we discuss to what extent constraints at the protein level affect the selection of RNA secondary structure in biotic (multicoding) RNA sequences.

Methods

RNA SECONDARY STRUCTURE

The Enfold algorithm (Hogeweg & Hesper, 1984) is used to analyze the effect of different Genetic Operators and to simulate the evolution of RNA secondary structure. The RNA secondary structure in CAEV and Visna virus is analyzed using the Zuker-algorithm (Zuker, 1989) based on the Fontana-algorithm (Fontana *et al.*, 1993). Although the latter gives more accurate predictions of secondary structure, it takes more CPU time and was therefore not used for the large-scale simulations. The parameter set from Jaeger *et al.* (1989) was used in both algorithms. We cross-checked our expectation that our results were not affected by the difference between the algorithms by using both algorithms for the analysis of the strings produced by the simulated evolution using the Enfold-algorithm.

To compare RNA secondary structures we represented the structures as strings, in which every position has a symbol depending on its direction of base-pairing (upstream or downstream from the hairpin loop), if a base is not base-paired the symbol depends on whether it is in a hairpin loop or not (Konings & Hogeweg, 1989). Dissimilarity between the strings is given by their nominal distance from each other, that is by the number of different symbols at corresponding positions. Where alignment is included in the comparison of secondary structure, an algorithm is used based on Needleman & Wunsch (1970), with a penalty of 1 per position per gap.

GENETIC OPERATORS

Point mutations are substitutions of nucleotides in the RNA string. Because there is no checking for back mutations, in the long run the number of mutations will yield a slight overestimate of the nominal distance between sequences. In our analysis of the effect of mutations that do not change the amino acid code we took only the redundancy of third positions into account.

Deletions and insertions take place by removing one nucleotide from the RNA string, and putting one back. The sites are chosen randomly and independently. One insertion and deletion thus creates a shift of one nucleotide over the distance between the insertion and the deletion. When two sites are chosen randomly within a string, the mean distance between the sites is about one-third of the length of the string, in a string of length 150 nucleotides one insertion/deletion thus causes a mean shift of 50 nucleotides.

For cross-over one piece of the RNA string is substituted by another piece. The amount of change in the primary sequence will then depend on the length of the piece of RNA that is substituted and on its similarity to the new piece of RNA. To determine the maximum amount of change in secondary structure that is caused by cross-over in random RNA strings, we let the new piece of RNA differ at every position from the piece it has replaced. This way the amount of change in the primary sequence after a cross-over of length L equals the amount of change in the primary sequence after L point mutations, the only difference being that the change caused by cross-over is localized, whereas the change caused by point mutations is spread out over the entire string. By comparing the amount of change in secondary structure after cross-over with that after point mutations we obtain an idea of the "locality" of the transition from an RNA sequence to RNA secondary structure. During the simulation of evolutionary processes the part that is crossed-over comes from another individual in the population. In this case the nominal distance between the primary sequence before and after the cross-over depends on the heterogeneity in the population.

EVOLUTION

A Genetic Algorithm (Holland, 1975) was used to simulate the evolutionary selection process. A population consists of 100 RNA strings (GNOMES) of 150 nucleotides. At each time-step 10 GNOMES are removed from the population. The chance of being removed is proportional to the relative fitness of the GNOMES, i.e. "non-survival of the non-fittest". Thereafter reproduction takes place. From the remaining population 10 GNOMES are randomly chosen and copied to create 10 new GNOMES. The Genetic Operators change the primary sequences of these new GNOMES. After point mutations and/or insertions/deletions, (equal) cross-over between the new GNOMES takes place and they are put into the population. The secondary structure and fitness of the newly formed GNOMES are then determined. In the initial population all GNOMES are identical, a setting which can prevent premature convergence on local optima and which is biologically more relevant than the traditional setting for Genetic Operators in which all initial strings are chosen independently (Huynen & Hogeweg, 1989). Each simulation is run for 2500 time-steps, 25 simulations are performed for every set of the Genetic Operators.

SELECTION CRITERIA

The topology that is selected is a secondary structure with four stacks and three hairpin loops (Fig. 1). The three hairpin stacks do not necessarily have to stem from one "master" stack. The actual fitness is determined by multiplying the lengths of the four different stacking regions. The lengths of the stacks are multiplied in order to create selection towards stacks of equal length. Bulges and internal loops are not included in the length of the stacks; therefore the selection criterion favors a high level of base-pairing.

FREQUENCIES OF OCCURRENCE OF GENETIC OPERATORS

The frequency of occurrence of a Genetic Operator per newly formed string has a *Poisson distribution. The mean number of point mutations per new* GNOME *is* 1,

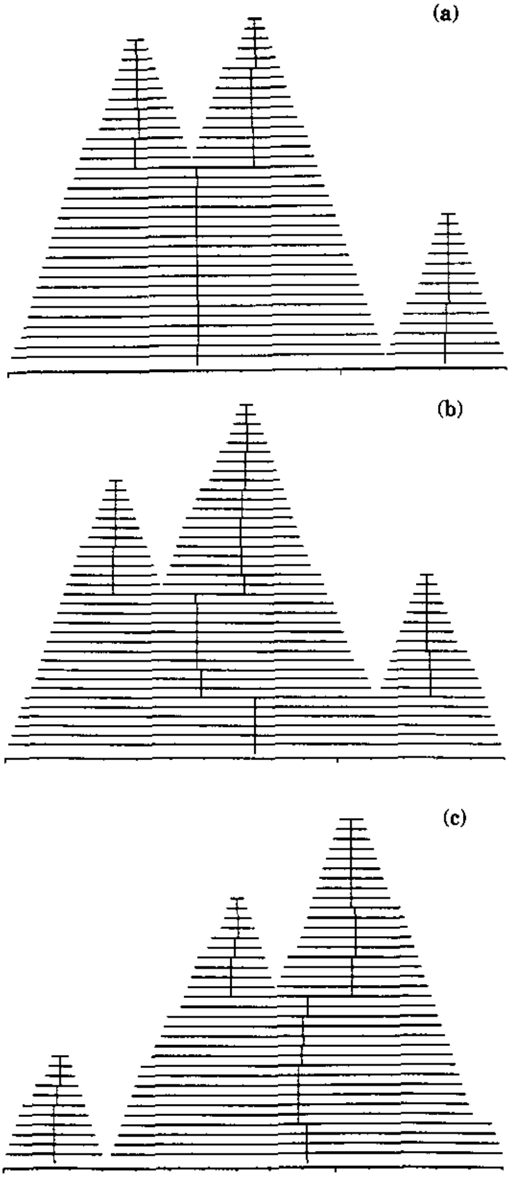


FIG. 1. Secondary structures after evolution for a "four-stack/three-hairpin loop" structure. The secondary structure shown is that of the run (one out of 25) which gave the median result with respect to fitness. The representation of secondary structure is according to the "mountain range" representation (Hogeweg & Hesper, 1984). Each base pair is shown by a horizontal line whose length spans the distance of that representation. Hairpin loops appear as flat tops, interior loops and bulges as intermediate plateaux, helices as sloping hillsides, and branching regions as valleys. Vertical lines within the pattern show the point midway between the nucleotides paired by the horizontal line. As is shown, evolution with different sets of Genetic Operators produced reasonable adaptation to the required RNA secondary structure. (a) Evolved with point mutations, insertions/deletions and recombination. (b) Evolved with point mutations and recombination. (c) Evolved with point mutations that do not change amino acid coding and recombination.

unless insertions and deletions are present; then it is 0·5. Also if the amino acid code is not allowed to change, the mean number of mutations per new GNOME is 1. The mean number of insertions/deletions per new GNOME is 0·25. The mean number of recombination events per new GNOME is 1.

RNA SEQUENCES

Both for the analysis of the effect of Genetic Operators and for the simulation of an evolutionary selection process, the nucleotides in all positions were chosen randomly and independently in the initial sequences, giving equal probability to all nucleotides. The *env* sequence of CAEV is from Saltarelli *et al.* (1990), the *env* sequence of VisNa is from Sonigo *et al.* (1985).

Results

THE EFFECTS OF GENETIC OPERATORS ON RNA SECONDARY STRUCTURE

One way to analyze the effect of Genetic Operators on secondary structure is to calculate the relation between the (dis)similarity of primary sequences and the dissimilarity of secondary structures (Fontana *et al.*, 1993). This relation is calculated with regard to the changes caused by the various Genetic Operators in random sequences. The Genetic Operators are: point mutations, point mutations that do not affect the coding capacity of the string, insertions and deletions, and recombination (Fig. 2). As is shown for all Genetic Operators, small changes in an RNA primary sequence can give rise to large changes in RNA secondary structure.

Point mutations

Point mutations that do not affect the coding capacity change secondary structure less than do the same number of point mutations in arbitrary positions (Fig. 2). This is due to a restriction on the positions where mutations can occur (only one in every three positions) and to the fact that if no amino acid changes are allowed, transitions (A to G.or C to U) occur more often than transversions (A, G to C, U). The latter have a more drastic effect on secondary structure. The restriction in the types of mutations that can occur affects the mean dissimilarity after only one mutation. The restriction on the positions where mutations can occur starts to lower the mean dissimilarity after about 12 mutations (8% of the sequence length). The limitation in the mean dissimilarity between secondary structures is small compared to the limitation in dissimilarity between primary sequences. That is, the mean dissimilarity between two sequences that code for the same protein, is less than one-third of the dissimilarity between two random sequences, whereas the dissimilarity between the secondary structures of these protein-coding sequences is nearly nine-tenths of that of random sequences. The latter ratio is the one between the levels of saturation of the curves of the different types of mutations.

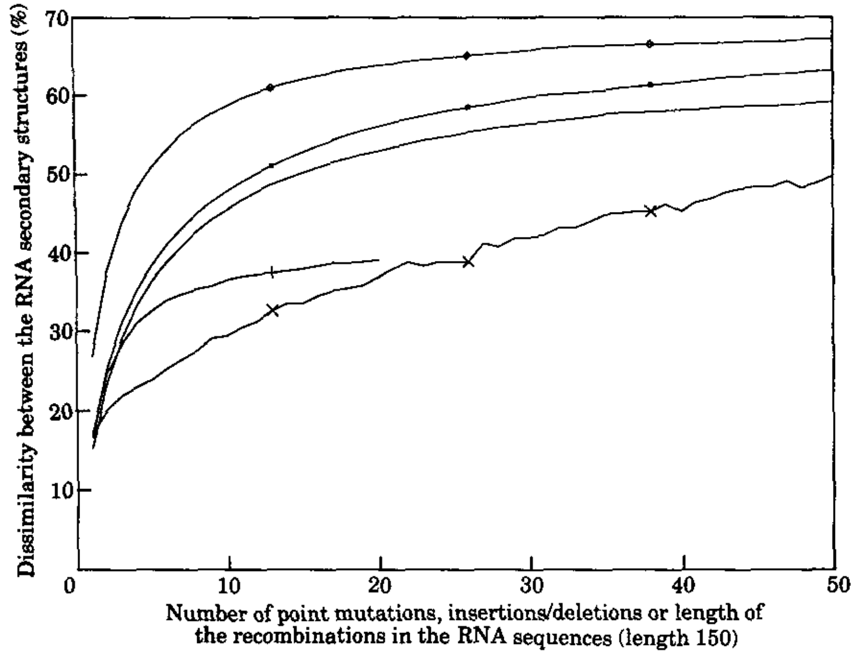


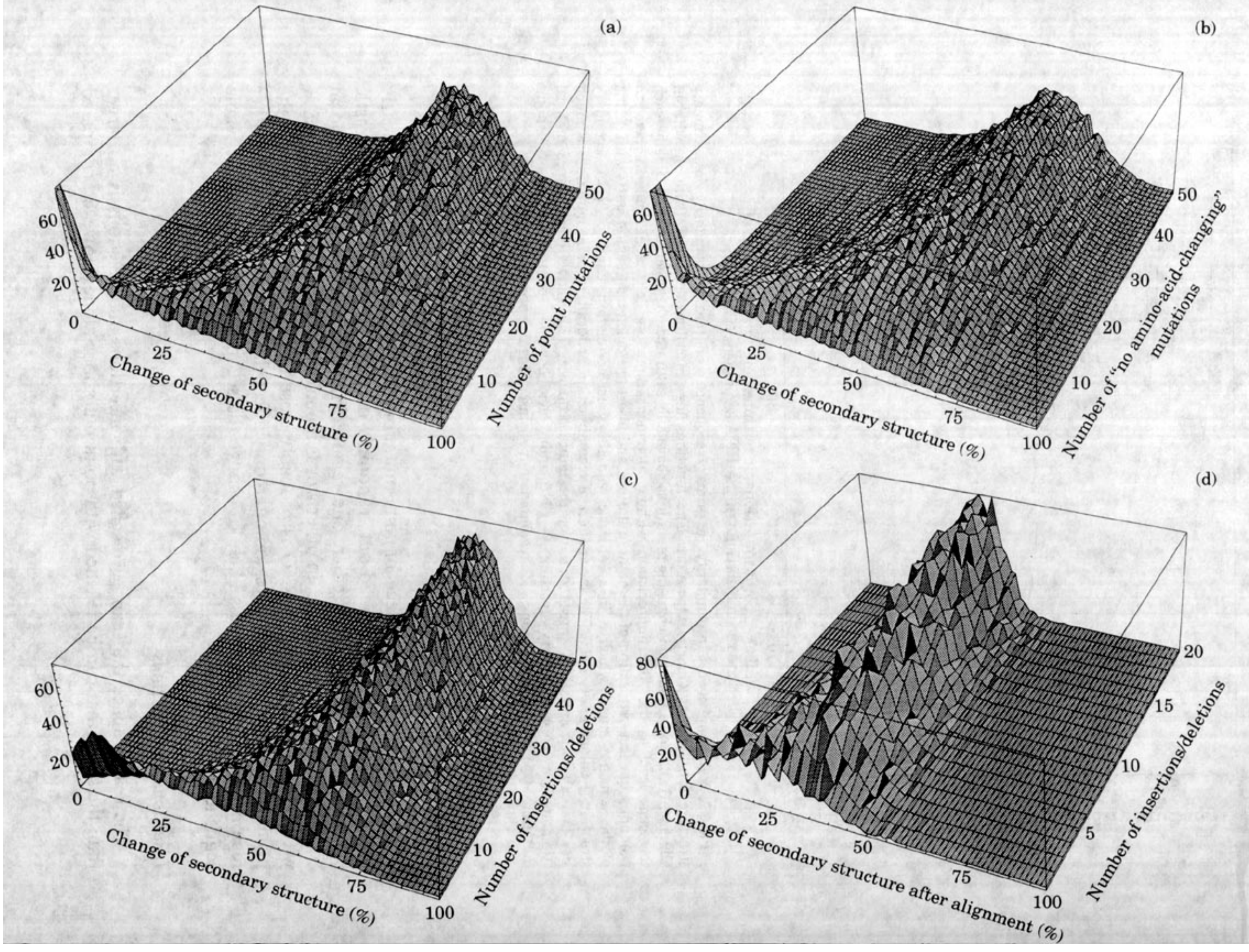
FIG. 2. Relation between dissimilarity in primary sequence and mean dissimilarity in secondary structure for different Genetic Operators. The functions for "insertions/deletions" (with or without alignment) and "no amino-acid-changing mutations" have fully saturated, the other functions finally reach the same level as the "insertion/deletion without alignment" function, i.e. the dissimilarity between random sequences. Saturation of the point mutations and insertions/deletions function starts within 10% of the sequence-length, which is in accordance with the short correlation length of RNA secondary structure (Fontana *et al.*, 1993). The mean effect of recombination or localized point mutations is much less than that of mutations spread out over the entire sequence. This shows the presence of local "domains" in RNA secondary structure. (—), No amino acid changing point mutations; (□), point mutations; (◇), insertions/deletions; (×), recombination; (|), insertions/deletions with alignment.

Recombination

The comparison of the effects of point mutations and recombination shows clearly that mutations that are "spread out" randomly over the string have a more drastic effect than mutations localized in one region. This reflects the presence of local "domains" in second structure, in other words if nucleotide P interacts with Q, and R with S, then if P is close to R, Q is probably close to S. Therefore changes in the primary sequence within one region often only affect the interactions with one other region, and are therefore less likely to change the secondary structure of the whole sequence.

Insertions and deletions

The search space that can be reached with insertions and deletions is no larger than the space that can be reached with point mutations alone. However, compared


(a)
60
40
20
0
25
50
75
100
Change of secondary structure (%)
10
20
30
40
50
Number of point mutations
(b)
60
40
20
0
25
50
75
100
Change of secondary structure (%)
10
20
30
40
50
Number of "no amino-acid-changing" mutations
(c)
60
40
20
0
25
50
75
100
Change of secondary structure (%)
10
20
30
40
50
Number of insertions/deletions
(d)
80
60
40
20
0
25
50
75
100
Change of secondary structure after alignment (%)
5
10
15
20
Number of insertions/deletions

to point mutations, insertions and deletions clearly increase the connectivity of the search space, because they create larger "jumps" through secondary structure space. This is reflected in the high dissimilarity after one insertion/deletion, and in the saturation of the "dissimilarity relation" after a relatively small number of steps (Fig. 2). Alignment of the secondary structures drastically increases the similarity, but does not change the shape of the relation between changes in sequence and secondary structure (Fig. 2).

RNA landscapes resulting from different Genetic Operators

The frequency distribution of changes in secondary structure, a so-called RNA landscape (Peter Schuster), has been calculated for insertions and deletions, point mutations, and for point mutations that do not change the amino acid coding (Fig. 3). The distribution of changes in secondary structure caused by mutations that do not change the amino acid coding hardly differs from that caused by point mutations in all positions. Although restricting the point mutations to those that do not change the amino acid code shifts the mean dissimilarity of secondary structures to a lower value (Fig. 2), this does not affect the maximum dissimilarity, i.e. it does not affect the possibility to obtain a maximum difference between secondary structures after a certain number of mutations.

Although the mean nominal distance between secondary structures after a small number of insertions/deletions looks more or less like the mean nominal distance after twice as many point mutations (Fig. 2), the distribution of the changes is different. The "peak" of the distribution is shifted towards a higher dissimilarity. The main effect of the alignment of the secondary structures is, besides a general increase in similarity, a small peak at dissimilarity 2 after one insertion/deletion. Here alignment corrects the shift in secondary structure that is caused by a shift in the primary sequences.

Robustness of RNA landscape properties

Fontana *et al.* (1993) have obtained results for point mutations in arbitrary positions similar to those depicted in Figs 1 and 2. The algorithm they used for secondary structure comparison algorithm differs from the one used here: RNA secondary structures are represented as trees, and dissimilarity is based on the

FIG. 3. Frequency distribution of the number of changes in secondary structure after point mutations or insertions/deletions in the primary sequence. The frequency of a number of changes after a number of mutations is given by the z direction, which represents the promillage of the total. A low, odd number of changes in secondary structure is very unlikely. To smooth this effect, the odd and even numbers of changes have been added pairwise, starting at 0. As is shown, secondary structure space is highly connected, that is, maximum dissimilarity between secondary structures can be achieved after one point mutation or one insertion/deletion. (a) Frequency distribution for point mutations, dissimilarity of secondary structures is determined by direct comparison (no alignment). (b) Frequency distribution for point mutations that do not change amino acid code, dissimilarity of secondary structures as in (a). (c) Frequency distribution for insertions/deletions, dissimilarity of secondary structures as in (a). (d) Frequency distribution for insertions/deletions, dissimilarity of secondary structures determined using alignment. Note that the "number of insertions/deletions" axis has a length of only 20.

number and kind of "tree editing" steps required to change one secondary structure into another. Thus properties, like the shape of the curve that describes the mean dissimilarity in secondary structure after a number of changes in primary sequence and the shape of the frequency distribution of changes in secondary structure, do not depend on the specific algorithms used to determine or compare secondary structure. It would be interesting to find out whether the frequency distributions of changes in phenotype in response to changes in genotypes, as shown above, are general properties of "highly epistatic" genotype–phenotype mappings like the mapping from RNA primary sequence to RNA secondary structure.

EVOLUTION OF SECONDARY STRUCTURE WITH DIFFERENT GENETIC OPERATORS

Simulations of evolutionary processes with different Genetic Operators were performed for the development of a specific topology of the secondary structure. The three "sets" of Genetic Operators that were used during the different runs are: (i) insertions/deletions and point mutations, (ii) point mutations, and (iii) point mutations that do not change the protein that is coded for. Cross-over was used in all runs. The topology that was selected for was a four-stack, three-hairpin structure, like the RRE in the lentiviruses CAEV and Visna virus. The final shapes of the best individual in the population show that even in the most restricted evolution (only allowing for point mutations that do not change the amino acid code and crossing over) there is a reasonable capability of adapting to the required secondary structure (Fig. 1). The fitness of the RNA sequences depends not only on the presence of a large-scale secondary structure but also on the frequencies of base-pairing. The final fitnesses and frequencies of base-pairing show that allowing for point mutations with insertions/deletions or point mutations without insertions/deletions during evolution makes little difference. There is, however, a significant difference between these two

TABLE 1

Fitnesses and base-pairing frequencies after evolution for a "four-stack/three-hairpin" RNA secondary structure, using different Genetic Operators, and in random sequences. The means and the standard deviations of 25 runs are shown

	Fitness		Base-pairing	
	$\bar{x}$	S.D.	$\bar{x}$	S.D.
Genetic Operators used during selection				
Point mutations, insertions/deletions and cross-over	56 592	8305	85·0%	1·7%
Point mutations and cross-over	53 598	9076	83·1%	3·1%
No amino acid-changing point mutations, and cross-over	34 381	7216	73·9%	4·2%
Random sequences	2245	2996	48·8%	8·9%

The fitness is the product of the four stack lengths. The maximum fitness (the maximum total stacklength is 70, thus the maximum fitness is 17 . 17 . 18 . 18 = 93 636) is never reached: this is probably due to the low correlation of the fitness landscape. Evolution in which only the no amino acid-changing point mutations are allowed gives rise to a lower percentage of base-pairing than evolution in which all positions are free to change.

on the one hand and the result of the most restricted evolution on the other hand (Table 1). Although the latter is capable of adapting to a large-scale topology, adaptation on a smaller scale, i.e. by removing bulges or internal loops from stacking regions is less easy, as is reflected in the frequencies of base-pairing.

THE SHAPE OF THE SECONDARY STRUCTURE LANDSCAPE AROUND SELECTED SEQUENCES

Starting from the sequences that were selected using different sets of Genetic Operators for the "four-stack, three-hairpin loop" topology, we calculated the relation between dissimilarity of primary sequence and secondary structure, as is caused by point mutations (Fig. 4). The selected sequences are clearly more similar to their "neighbors" than random sequences are to their neighbors. This effect is the least pronounced in products of the evolution in which the protein coding had to be preserved. There is little difference between the results of the other, less restricted evolutions with respect to their dissimilarity relation.

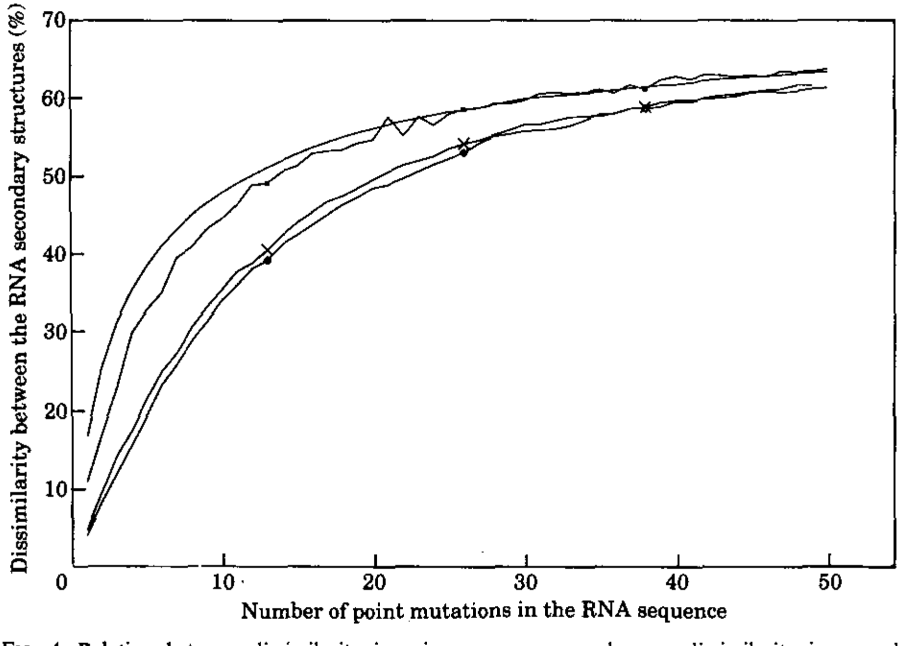


FIG. 4. Relation between dissimilarity in primary sequence and mean dissimilarity in secondary structure for point mutations. The initial sequences in the analysis have evolved for a specific topology using different "sets" of Genetic Operators: point mutations, insertions/deletions and recombination, point mutations and recombination, no amino acid-changing point mutations and recombination. As is shown, the evolved sequences show less change in secondary structure after point mutations than do random sequences. (—), Random initial sequence; (□), after evolution with no amino acid changing point mutations; (◇), after evolution with insertions/deletions and point mutations; (×), after evolution with point mutations.

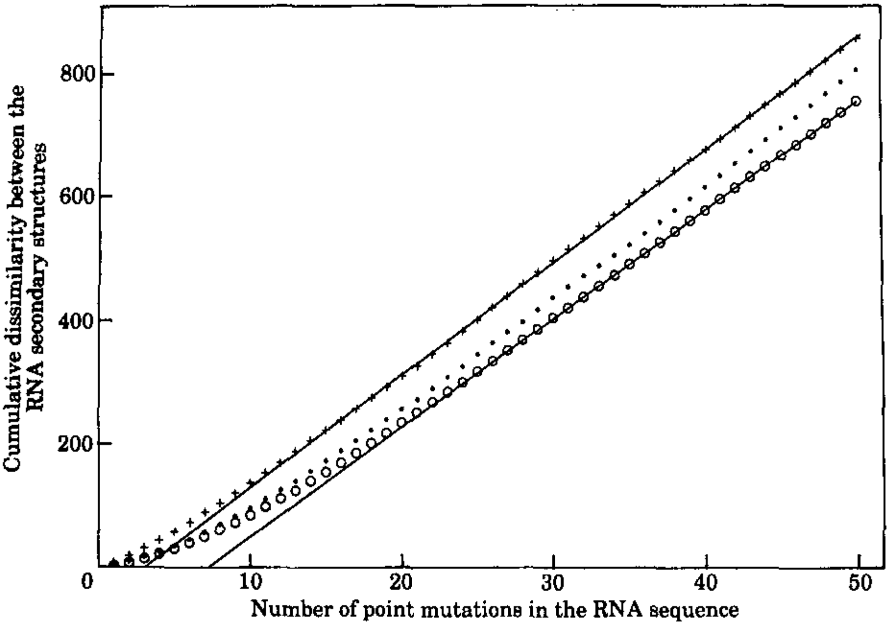


FIG. 5. Cumulative distances in secondary structure for point mutations. Starting sequences are selected for a "four-stack/three-hairpin loop" secondary structure, using different sets of Genetic Operators. The lines show where the curves become "straight", i.e. where the similarities between secondary structure of "neighboring" primary sequences become constant. The sequences that have evolved with no amino acid-changing point mutations require fewer point mutations before the dissimilarity in secondary structure of "neighboring" sequences becomes constant than the sequences that have evolved without restrictions. (○), After evolution with point mutations and insertions/deletions; (♦), after evolution with point mutations; (+), after evolution with no amino acid-changing point mutations.

THE DISTANCE FROM SELECTED SEQUENCES TO RANDOM SEQUENCES

As is shown in Fig. 4, selection for a topology with a high degree of base-pairing increases the similarity between secondary structure of the fittest individual in the population and that of its nearest neighbors. If this is a general feature of the (sequence) subspace around the fittest individual, this effect should also be visible in the correlation between the neighbors of the fittest. By examining how many mutations are needed before the similarity between neighbors in a random walk, starting from the optimum, approaches the similarity of neighbors in a random walk, starting from a random sequence, one can estimate how many mutations the optimum is away from a "random" sequence (with respect to secondary structure). The cumulative nearest-neighbor distances in a random walk, starting from sequences evolved with the different Genetic Operators, are given in Fig. 5. The difference between the most restricted and the least restricted (with point mutations and insertion/deletion) evolution is significant. An RNA sequence that has evolved without restrictions needs about twice as many mutations to come into a region where the distances between secondary structure are constant as an RNA sequence that has evolved with restrictions. In the least restricted case about 15% of the

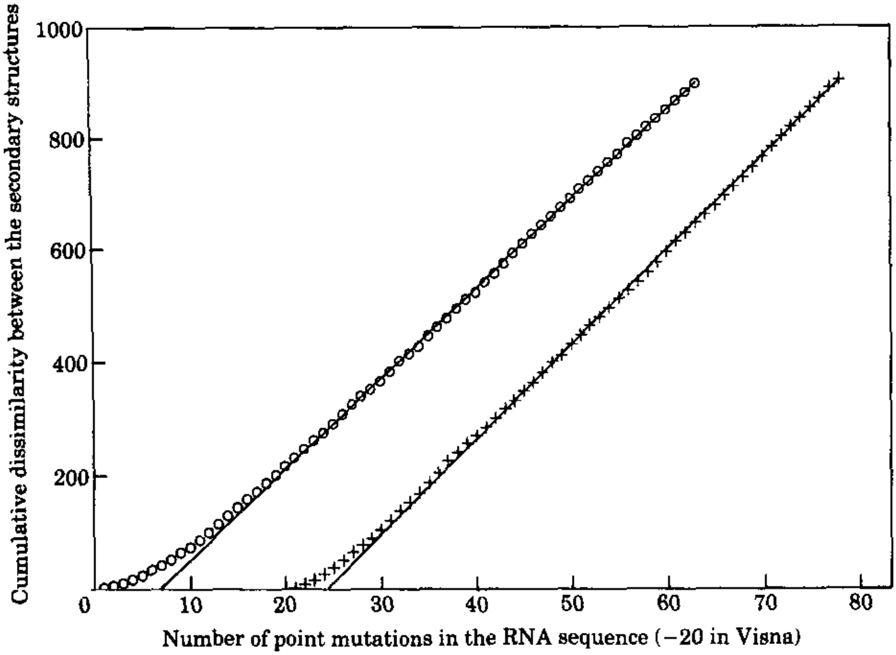


FIG. 6. Cumulative distances between secondary structure for point mutations. Starting sequences are the Rev response elements in CAEV and Visna virus. (○), RRE in CAEV; (+), RRE in Visna.

sequence has to be changed (about 22 mutations). The fact that it takes about 15% of change in the primary sequence to go from a selected secondary structure to a more or less random sequence does not mean that it would take only 15% of change to go from any random sequence to a secondary structure like the one that has been selected. This 15% is more like a lower bound. The number of changes between the original sequence (from which the evolution started) and the final sequence after evolution with only point mutations varies between 0·5 and 0·66 times the length of the sequence (data not shown). This gives an upper bound, since one expects some neutral drift during evolutionary search.

THE DISTANCE FROM BIOTIC RNA SECONDARY STRUCTURES TO "RANDOM" SEQUENCES

The "cumulative distance plot" has also been calculated for some biotic functional secondary structures (Fig. 6). The RRE secondary structure is formed by a primary sequence that also codes for part of the *env* protein in lentiviruses. The length in CAEV is 193 and in Visna virus it is 176. The nucleotide sequence that codes for the RRE in these lentiviruses is relatively well conserved within the *env* gene, with respect to both silent and non-silent substitutions (Saltarelli *et al.*, 1990). The cumulative distance plots of the RRE or CAEV and Visna virus both show that the distance between secondary structures of neighboring sequences increases as one moves away from the initial sequence, as has been shown for the molecules that have been

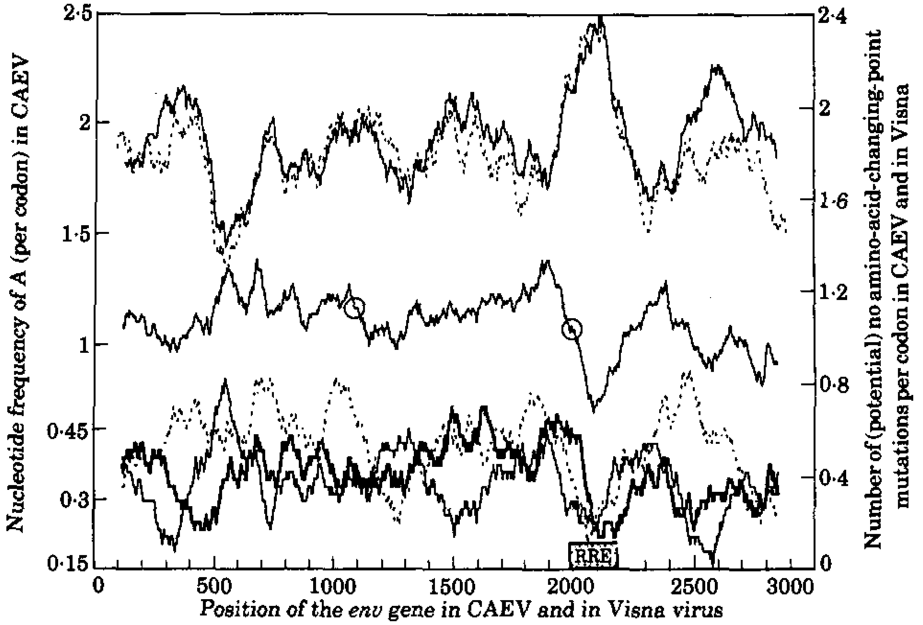


FIG. 7. Redundancies of amino acid-coding of the *env* gene in CAEV and Visna virus (upper part of the figure), and nucleotide frequencies in CAEV. The redundancies are expressed by the number of mutations per codon (out of nine possible mutations) that would not change the amino acid coding. The graph shows the average redundancies and frequencies of a window of 64 codons in CAEV and 59 codons in Visna virus, i.e. the size of the Rev response elements in CAEV and Visna. The *env* genes were aligned using Needleman & Wunsch (1970). Note that the frequencies of adenine per position per codon and the frequency of adenine per codon are not on the same scale. In the RRE region the adenine frequency is relatively low, *this is found for the first and third coding positions and to a lesser extent also for the second* coding positions. The low adenine frequency is mainly compensated by a relatively high C frequency. High frequencies of A in first and second coding positions bring about a low redundancy for amino acid coding, and low G + C frequencies reduce secondary structure base-pairing potential (Huyen *et al.*, 1992; Fontana *et al.*, 1993). Selection for the development of conservation of the secondary structure of the RRE is therefore likely to have caused nucleotide frequencies which are biased relative to the rest of the sequence. (○), A in all coding positions; (—), A in first coding positions; (–), A in second coding positions; (---), A in third coding positions; (–), redundancy CAEV; (····), redundancy Visna.

selected for a specific secondary structure. The number of mutations required to reach a constant similarity between neighboring secondary structures relative to the sequence length is, for RRE in CAEV, larger than the number for the artificial multiple coding sequences reported here. This suggests that the nucleotide sequence has been flexible in more than only the redundant third codon positions in order to form the particular secondary structure. Comparing the curves quantitatively is however not appropriate, since we know little about the selection parameters for the biotic sequences.

It is interesting to note that the RRE in CAEV and in Visna lies in the most redundant area with respect to amino acid coding. That is, out of nine potential mutations per codon in this area a mean of 2·4 will not change the protein coding; for the rest of the *env* gene this mean is 1·8 (Fig. 7). The relative high redundancy is related to a bias of nucleotide frequencies in the RRE relative to the rest of the *env* gene (A 24%, C 24%, G 28%, U 24% in RRE; A 38%, C 16%, G 25%, U 21% in the rest of *env*). The high adenine frequency in the "not-RRE part" of the *env* gene might be due to a mutation bias of the reverse transcriptase, as was found in HIV-I (Preston *et al.*, 1988; Vartanian *et al.*, 1991). The biased nucleotide frequencies in the RRE region (relative to the skewed nucleotide frequencies in the rest of the *env* gene) might be caused by the necessity to have a high G+C content and a balanced G/C ratio in order to form stable secondary structures with a high degree of base-pairing. A balancing of the G/C ratio that points to a constraint on RNA secondary structure has also been observed in histone genes of warm-blooded vertebrates (Huynen *et al.*, 1992). The bias of nucleotide frequencies, which is predominantly present in the first and third coding positions (Fig. 7), suggests that the development or conservation of the RRE secondary structure has affected the amino acid sequence of the *env* gene.

Another argument supporting the idea that the protein-coding sequence has been affected by the need for the secondary structure of RRE is that the variation that exists between sequences of CAEV and Visna virus is not neutral, and appears to be "coadapted". That is, primary sequences that are intermediate between the sequences of RRE in and CAEV and in Visna, both with respect to silent or non-silent substitutions and with respect to conservative or non-conservative amino acid changes, do not give rise to a (minimal energy) secondary structure like RRE in CAEV or Visna virus (data not shown).

Discussion

Restricting the mutations that are allowed during the evolution of RNA to those that do not change the protein code still lets an arbitrary RNA molecule adapt its secondary structure to some large-scale topology. This is reflected in the amount of change that these "restricted" mutations cause in secondary structures of random RNAs and in the results of evolutionary adaptation that uses these restricted mutations. Adaptation on a smaller scale appears to be less easy, as is shown in the frequency of base-pairing that can be achieved. If all positions in an RNA molecule are allowed to change during evolution, the addition of the possibility for insertions and deletions in the string does not significantly increase the amount of base-pairing that can be achieved. In a way this is surprising, since allowing for insertions and deletions appears to be an excellent way of removing bulges in stacking regions. For the specific selection criteria used here, this removal connects local optima with the global optimum. One way of removing such bulges using point mutations alone, is by local shifts in the interactions. Such shifts have been deduced from biotic data in sequences in which no insertions/deletions had occurred (Saltarelli *et al.*, 1990; Konings, 1992).

Our analysis shows that during evolution for a specific topology the characteristics of the RNA landscape change. In selection for a structure with four stems and three hairpins, the quasispecies moves to a subspace of the landscape in which the correlation between the secondary structures of the neighbors is significantly higher than in the space where evolution started. This increase can depend on two factors:

(i) the specific topology: selection for a high level of base-pairing (as in this case) creates relatively stable secondary structures in which mutations will only disrupt local structures, creating small bulges in otherwise undisrupted stacking regions. (ii) Evolution for the "flattest" peaks: evolution has a general tendency to move quasispecies to peaks with a relatively high correlation, because they have relatively large "basins" of attraction and/or because a high correlation increases the fitness of the quasispecies. The latter effect has been observed in a "double-peaked" landscape where the quasispecies ends up on the lower, relatively flat peak, provided that the mutation frequency is sufficiently high (Schuster, 1989). If this effect were to be present in our simulations, it would of course imply a correlation between similarity of secondary structure and similarity of fitness. Since the fitness is a function of the secondary structure this correlation is highly likely. We cannot determine whether evolution for the flattest peaks plays a role in the simulations reported here because of the dominant effect of the specific topology that is selected for. Recent findings show, however, that evolution for the flattest peaks does play a role in selection for RNA secondary structure (Huynen & Hogeweg, in preparation).

The correlation between neighbors in a landscape can be regarded as a measure of its "ruggedness". As shown by Kauffman (1989), the ruggedness of a landscape together with the quality of replication play a crucial role in whether a population can reach an optimum. Thus, if the ruggedness of the landscape covered by the quasispecies changes during evolution, then for optimum performance, parameters concerning the mutation rate should also be allowed to change, not only for a specific problem but also during the process of solving it. Suggestions have been made about how to adapt the Genetic Operators during evolutionary search (for a review see Davis, 1989) and the relation between the performance of a Genetic Algorithm and the statistical features of the fitness landscape has been stressed (Manderinck *et al*., 1991). The changes of the statistical features of the landscape during the evolutionary search process and the implications of these changes have, however, never been taken into account, and deserve our attention.

We thank S. M. McNab for linguistic advice. Danielle Konings thanks the W. M. Keck foundation for its generous support of RNA science on the Boulder campus. The investigations were supported by the Foundation for Biological Research (BION), which is subsidised by The Netherlands Organization for Scientific Research.

REFERENCES

BARELL, B. G., AIR, G. M. & HUTCHISON, C. A. (1976). Overlapping genes in bacteriophage φX174. *Nature, Lond.* **264,** 34–41.

CULLEN, B. R. (1991). Human immunodeficiency virus as a prototypic complex retrovirus. *J. Virol.* **65,** 1053–1056.

DAVIS, L. (1989). Adapting operator probabilities in genetic algorithms. In: *Proceedings of the Third International Conference on Genetic Algorithms* (Schaffer, J. D., ed.) pp. 61–69. San Mateo: Morgan Kaufmann.

EIGEN, M. & SCHUSTER, P. (1979). *The Hypercycle: A Principle of Natural Self-Organization.* Berlin: Springer.

FONTANA, W., KONINGS, D. A. M., STADLER, P. F. & SCHUSTER, P. (1993). Statistics of RNA secondary structure. *Biopolymers,* in press.

FONTANA, W. & SCHUSTER, P. (1987). A computer model for evolutionary optimization. *Biophys. Chem.* **26,** 123–147.

HATFIELD, D. & OROSZLAN, S. (1990). The where, what and how of ribosomal frameshifting in retroviral protein synthesis. *TIBS* **15,** 186–190.

HOGEWEG, P. & HESPER, B. (1984). Energy directed folding of RNA sequences. *Nucl. Acids Res.* **12,** 67–74.

HOGEWEG, P. & HESPER, B. (1992). Evolutionary dynamics and the coding structure of sequences: multiple coding as a consequence of crossover and high mutation rates. *Comp. Chem.* **16,** 171–182.

HOLLAND, J. H. (1975). Adaptation in natural and artificial systems. Ann Arbor, MI: University of Michigan Press.

HUYNEN, M. A. & HOGEWEG, P. (1989). Genetic Algorithms and information accumulation during the evolution of gene regulation. In: *Proceedings of the Third International Symposium on Genetic Algorithms* (Schaffer, J. D., ed.). San Mateo: Morgan Kaufmann.

HUYNEN, M. A., KONINGS, D. A. M. & HOGEWEG, P. (1992). Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure. *J. molec. Evol.* **34,** 280–291.

JAEGER, J. A., TURNER, D. H. & ZUKER, M. (1989). Improved predictions of secondary structures for RNA. *Proc. natn. Acad. Sci. U.S.A.* **86,** 7706–7710.

KAUFFMAN, S. A. (1989). Adaptation on rugged fitness landscapes. In: *Lectures in the Sciences of Complexity* (Stein, D. L., ed.) pp. 619–712. New York: Addison Wesley.

KONINGS, D. A. M. & HOGEWEG, P. (1989). Pattern analysis of RNA secondary structure. *J. molec. Biol.* **207,** 597–614.

KONINGS, D. A. M. (1992). On the coexistence of multiple codes in messenger RNA molecules. *Comp. Chem.* **16,** 153–163.

MANDERINCK, B., DE WEGER, M. & SPIESSENS, P. (1991). The genetic algorithm and the structure of the fitness landscape. In: *Proceedings of the Fourth International Conference on Genetic Algorithms* (Belew, R. K. & Booker, L. B., eds) pp. 143–150. San Mateo: Morgan Kaufmann.

NEEDLEMAN, S. B. & WUNSCH, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. molec. Biol.* **48,** 443–453.

PRESTON, B. D., POIESZ, B. J. & LOEB, L. A. (1988). Fidelity of HIV-1 reverse transcriptase. *Science* **242,** 1163–1171.

SALTARELLI, M., QUERAT, G., KONINGS, D. A. M., VIGNE, R. & CLEMENTS, J. E. (1990). Nucleotide sequence and transcriptional analysis of molecular clones of CAEV which generate infectious virus. *Virology* **179,** 347–364.

SCHUSTER, P. (1989). Optimization of RNA structure and properties. In: *Molecular Evolution on Rugged Landscapes: Proteins, RNA and the Immune System* (Perelson, A. S. & Kauffman, S. A., eds) pp. 47–71. Redwood City, CA: Addison Wesley.

SONIGO, P., ALIZON, M., STASKUS, K., KLATZMANN, D., COLE, S., DANOS, O., RETZEL, E., TIOLLAIS, P., HAASE, A. & WAIN-HOBSON, S. (1985). Nucleotide sequence of the Visna lentivirus: relationship to the AIDS virus. *Cell* **42,** 369–382.

VARTANIAN, J.-P., MEYERHANS, A., ÅSJÖ, B. & WAIN-HOBSON, S. (1991). Selection recombination, and G → A hypermutation of human immunodeficiency virus type 1 genomes. *J. Virol.* **65,** 1779–1788.

ZUKER, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science* **244,** 48–52.