Evolutionary dynamics and the relation between RNA structure and RNA landscapes.

Martijn A. Huynen and Paulien Hogeweg

Bioinformatics Group, University of Utrecht
Padualaan 8, 3584 CH Utrecht, Netherlands
e-mail: mah@binf.biol.ruu.nl
Phone: (31)-30-533695

Abstract

We study evolutionary dynamics on fitness landscapes based on a genotype-phenotype relation that is given by the transition from RNA sequence to RNA secondary structure. RNA landscapes show large variations in their local "ruggedness". Evolution can on the hand (passively) be affected by these variations, in that selection for a specific secondary structure moves the quasispecies to a specific region of the landscape. On the other hand it can (actively) exploit these variations. In stable environments it moves the quasispecies towards relatively "flat" peaks, where not only the "master sequence" but also its mutants have a high fitness. In a rapidly changing environment, the situation is reversed; evolution moves the quasispecies to a region where the correlation between secondary structures of "neighbouring" RNA sequences is relatively low. Herewith it can anticipate changes in the environment. The movement towards relatively smooth or relatively rugged parts of the landscape leads to a pattern generation at the sequence and/or secondary structure level.

Patterns at the sequence or secondary structure level are tightly coupled with the local ruggedness of the fitness landscape. Since evolution can be for a specific fitness landscape as well as it can be for a specific pattern, we should be aware that patterns in products from evolutionary processes can be the result from selection for a specific landscape and that a specific fitness landscape can be the result of selection for a specific pattern.

Introduction

To understand the products of evolutionary processes, it is essential to understand the structure of fitness landscapes on which such processes operate. In biotic genomes the different parts of the genome have to be co-adapted to produce a desired behaviour, implying a high degree of epistatic interactions. Such interactions can be both between molecules, like in gene-regulation, and within molecules, like in the formation of higher order structures within RNAs or proteins. As was shown by Kauffman (Kauffman, 1989; Weinberger, 1990), a high degree of epistatic interactions leads to "rugged" fitness landscapes; i.e. the correlation of the fitnesses of genotypes that are "neighbours" in genotype space is relatively low.

As a paradigm system for evolution on a fitness landscape that involves a high degree of epistatic interactions, we study evolution on landscapes in which fitness is a function of RNA secondary structure and the genotype is the RNA sequence. The transition from RNA sequence to RNA secondary structure leads to "phenotype landscapes" that are very rugged. Changing 10% of the RNA sequence gives rise to a difference between the (minimal energy) RNA secondary structures that is hardly less than that between the secondary structures of two random RNA sequences (Fontana et al. 1993; Huynen et al. 1993). In this paper we focus on variations in the correlation of RNA landscapes. First we show that selection for a specific secondary structure with a high frequency of base-pairing leads to secondary structures that are relatively robust to mutations, i.e. a smooth landscape. Here the movement towards smooth parts of the landscape is a side effect of the specific secondary structure that is selected for. In the second part we ask the question whether and how evolution can actually exploit variations in the ruggedness of a landscape, and whether we can identify patterns in the RNA sequences that are related to the exploitation of these variations. We study evolution in two types of environments, a stable environment (a fixed fitness landscape), and a very unstable environment (a constantly changing fitness landscape). In a model with evolvable mutation rates, Kaneko and Ikegami have shown that in a fixed landscape the mutation rate decreases once a population has reached an optimum, whereas in a changing landscape it is maintained at a high level (Kaneko and Ikegami, 1992). Reducing the mutation rate after reaching an optimum in a fixed landscape increases the "inclusive" fitness of fit individuals. An alternative way of increasing this fitness is moving towards relatively flat peaks in the landscape. The latter effect has indeed been observed in a double peaked landscape where the quasispecies ends up at the lower, relatively flat peak, provided that the mutation frequency is sufficiently high (Schuster, 1989). Agur (Agur, 1987; Agur and Kersberg, 1987) shows that the smoothness of a landscape can be regulated by modulating the non-linearity of the genotype phenotype relation. In our approach the landscape is fixed, and variations in its ruggedness are solely the result of the "richness" of the RNA sequence-RNA secondary structure transition.

Methods

RNA secondary structure determination

The Enfold algorithm (Hogeweg & Hesper, 1984) with the parameter set from (Jaeger *et al.*, 1989) is used for prediction of RNA secondary structure.

Comparison of RNA secondary structures

To compare RNA secondary structures we represent 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 and Hogeweg, 1989). Dissimilarity between the strings is given by their nominal distance, that is by the number of different symbols at corresponding positions (no alignment).

Genetic Operators

Point mutations are substitutions of nucleotides in the RNA string.

For cross-over one piece of the RNA string is substituted by another piece from another string in the population.

Evolution

Selection for a fixed secondary structure (part 1 and 2 of results): A Genetic Algorithm (Holland, 1975) was used to simulate the evolutionary selection process. A population consists of 200 RNA strings (GNOMES) of 150 nucleotides (part 1) or 30 nucleotides (part 2). At each time step 20 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 20 GNOMES are randomly chosen and copied to create 20 new GNOMES. The Genetic Operators change the primary sequences of these new GNOMES. After point mutations, (equal) cross-over between the new GNOMES takes place and they enter 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 that is biologically more relevant than the traditional setting for Genetic Operators in which all initial strings are chosen independently (Huynen and Hogeweg, 1989). Each simulation is run for 2500 time steps (part 1) or 30000 time steps (part 2), each run is repeated 25 times with a different randomseed.

Evolution in a highly variable environment (part 3 of results)

In this system there are 2 populations, population A consists of 250 GNOMES, population B of 50. The sequence length is 50. Every time step all GNOMES of A randomly encounter GNOMES of B (the system is fully mixed). The absolute fitness of a GNOME in A is the similarity of its secondary structure with that of the GNOME of B it encountered. GNOMES of A reproduce every time step. GNOMES of B reproduce once in each 5 time steps, their absolute fitness is 50 (the sequence length) minus the maximum similarity of their secondary structure with that of the GNOMES of A they encountered in the past 5 time steps. Population sizes are constant, there is only competition within populations.

Selection criteria

Selection for a cloverleaf like structure (four stacks and three hairpin loops) (part 1 of the results): The absolute fitness is the product of the length of the four stacking regions. In determining the length of the stacking regions we only included the base-paired nucleotides. The fitness criterion therefore favours a high frequency of base-pairing.

Selection for a double hairpin (part 2 of the results): The absolute fitness is the product of the length of the two hairpins (with a maximum of 5 per hairpin).

The absolute fitness in the co-evolving system (part 3 of the results) is the similarity (population A) or dissimilarity (population B) of secondary structures of GNOMES that "meet".

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 in the evolution for a fixed secondary structure, it is 0.1 in the co-evolving system (part 3 of the results). It is kept relatively low in this system to give the chasing population a chance to "catch up" with the population it is chasing, and thus increase the pressure in the population that is chased to change. The mean number of cross-over events per new GNOME is 1. There is no cross-over in the co-evolving system (part 3 of the results).

Results

1) evolution for a secondary structure with a high frequency of base-pairing.

We simulated the evolution for a cloverleaf like RNA secondary structure with a high frequency of base-pairing as is found in the Rev Response Element (RRE) in the

retroviruses Caprine Arthritis Encephalitis Virus (CAEV) and Visna virus (Saltarelli *et al.*, 1990). The products of our simulations show an increase in the resistance to point mutations; i.e. the mean change in secondary structure after changes in the RNA sequence is smaller then in random sequences (figure 1). This is not only true for the selected sequences, but also for their "neighbours" in sequence space (figure 2). The same effect is observed in the RRE of CAEV and of Visna virus (figure 2). Thus evolution has moved the quasispecies to a relatively smooth part of the RNA landscape. This is mainly a result of the specific selection criterion; selection for sequences with a high frequency of base-pairing leads to long stacking regions in which mutations in the sequence will only lead to local changes in the secondary structure. One might however ask whether the movement towards relatively smooth parts of a fitness landscape is not a general phenomenon in evolution, since movement towards relatively flat "hilltops" in a fitness landscape might increase the fitness of the quasispecies. The question is particularly intriguing since the transition from RNA sequence to RNA secondary structure is a "many to one" transition, multiple sequences give rise to the same secondary structure (P. Schuster, pers. comm.). Evolution for a specific secondary structure therefore can choose between solutions (sequences) that all have the same fitness (secondary structure) but which lie in regions in sequence space with a different local ruggedness. (see also Huynen er al., 1993)

2) evolution for a simple, two hairpin secondary structure

We let sequences evolve for a 2 hairpin structure. Figure 3 shows the resistance of the secondary structures to mutations in the sequences at the time the population has adapted to the required secondary structure for the first time, and at the end of evolution (after 30000 time steps). The resistance to mutations has increased during the simulation. It appears that initially the quasi-species climbs a peak that is relatively nearby, and that selection for relatively flat peaks is more al long term process.

During evolution for a 2 hairpin structure there is a pattern formation in the sequences that is associated with the development of resistance to mutations. The sequences show both an over representation of long poly-purines and poly-pyrimidines (length > 4) , as an over-representation of mono-purines/pyrimidines (data not shown). The specific pattern that is generated is shown in figure 4. One hairpin is made of one stretch of purines on one side and one stretch of pyrimidines on the other side, whereas the other hairpin is made of alternating purines and pyrimidines. This pattern is very efficient in preventing alternative foldings. It does not only prevent that the two hairpins merge to one large hairpin, where the 5' side of the 5' hairpin interacts with the 3' side of the 3' hairpin, and the 3' side of the 5' hairpin interacts with the 5' side of the 3' hairpin, but also the merging to one small hairpin where the 3' side of the 5' hairpin interacts with the 3' side of the 3' hairpin or the 5' side of the 5' hairpin interacts with the 5' side of the 3' hairpin.

The movement towards flat peaks reduces the selection strength between mutants, and should therefore give rise to more variation in the population. Indeed if one follows the amount of variation in the populations it increases long after the populations have reached the desired secondary structure (figure 5). It is interesting to note that during the development of an increase in resistance to mutations the fitness of the whole population increases only very little (figure 5), particularly in comparison to the decrease of the effect of mutations on the fitness of sequences with the maximum fitness (figure 3). The effect of the increase in resistance to mutations seems to be partly nullified by the increase in the variation in the population.

In the evolution for a relatively complicated secondary structure we see that recombination plays an important role in the development of patterns and the increase of resistance to the effect of mutations on the fitness. Recombination leads to a selection at the subgenomic level (see also Hogeweg and Hesper, 1991). In the results presented here, cross-over leads to the formation of two hairpin structures that are relatively

independent from each other. The actual number of potential secondary structure interactions within the molecule is thereby decreased.

3) evolution in a rapidly changing environment

In order to study whether it is also possible for the quasispecies to evolve towards relative rugged parts of the landscape, we developed a system with a rapidly changing environment. The system consists of two populations of RNA strings. The fitness of strings in population A is positively dependent on the similarity between their secondary structure and the secondary structure of strings in population B, whereas the fitness of strings in population B is negatively dependent on this similarity. Population A thus "chases" population B in a Red Queen type of evolution. The situation is very like that of antibodies in an immune system (A) which try to match an epitope of a pathogen (B).

Because there is no specific secondary structure connected with the fitness criterion, we can compare the sensitivity to mutations of the selected secondary structures with that of secondary structures of random sequences. Figure 6 shows the relation of dissimilarity between sequences and between secondary structures for random sequences and for the sequences of population A and population B. The sensitivity to mutations in the chased population (B) is clearly higher than that in random sequences. Population A also shows a trend, although less pronounced, to move towards relatively rugged regions.

One can understand this higher sensitivity if one realizes that it is in the interest of sequences of B to change their secondary structure rapidly and thus stay ahead of the chasing population. The fitness landscape for B changes rapidly, but its changes are in a way predictable, since a secondary structure that gives a high fitness at time X will give a low fitness at a time X + t, so anything is better than staying at the same spot; thus there is selection for (phenotypically) rapidly evolving quasispecies. For the chasing population the situation is more complex; population A has no way of knowing where population B will move to. The sensitivity to mutations of the secondary structures in population A is affected mainly by their evolution towards the specific secondary structures of population B.

With respect to models of the interaction between an immune system and a rapidly changing virus (e.g. the AIDS model of Nowak et al, 1990) these results imply that the quasispecies of the virus will move towards regions where the effect of mutations on the higher order structure of the antigen is maximized, and thus the cross-reactivity of antibodies with antigens is minimized. The tendency to maximize the effect of mutations on the epitope of a pathogen in a pathogen-host system has also been observed in a qualitatively similar model (Agur, 1987).

The evolution towards relatively rugged parts of the landscape leads to pattern generation at the RNA secondary structure level. That is, in order to increase the effect of mutations, the amount of secondary structure (= frequency of base-pairing) is increased. In random sequences of length 50 the mean frequency of base-pairing is 34 %. In the population that is chased (B) it goes up to an average of 53 %, whereas in the chasing population (A) it becomes 42%. Although we have not been able to identify patterns at the sequence level yet, we have clear indications that the sequences of population B remain within a relatively small region of sequence space.

Discussion

The genotype-phenotype relation as is present in the transition from RNA sequence to RNA secondary structure leads to fitness landscapes with variations in their correlation. On the one hand the specific selection criterion that is used can lead to movement towards a part of the sequence space where the correlation between secondary structures is relatively high. On the other hand the evolutionary dynamics themselves can, irrespective of the specific selection criterion, move the quasispecies towards regions in

sequence-space where the correlation between the secondary structures is relatively high or low, leading to a pattern generation at the sequence level and/or the secondary structure level. Thus patterns in the sequence/secondary structure on the one hand and the local shape of the RNA landscape on the other hand are tightly coupled, and selection for one of the components will affect the other. It is therefore important when observing patterns in biotic structures to realize that they can be the result of selection for a specific fitness landscape and vice versa.

The work done on varying mutation rates (Kaneko and Ikegami, 1992; Ikegami and Kaneko, 1992) is analogous to our work on varying the correlation of the landscape if one only considers mutations that affect the phenotype. In other words by varying the correlation of the landscape one varies the number of potential phenotype-changing mutations. From a biological point of view an important difference between varying mutation rates and varying the (local) ruggedness of the landscape is that mechanisms for varying mutation rates generally affect large parts of the genome, although they do not necessarily affect the whole genome (e.g. local somatic hypermutation regions in B cells (French et al., 1989)), whereas the ruggedness of the genotype-phenotype transition in different parts of the genome can be varied locally. On the other hand an advantage of varying mutation rates is that mechanisms for this can be turned on or off on an evolutionary relatively short time scale. Both mechanisms can of course co-exist, which allows for fine tuning of the phenotype-affecting mutation rates in the whole genome.

An important question is whether these results also hold for genotype-phenotype landscapes that are based on a relation other than the RNA sequence-RNA secondary structure relation. Basically the results depend on whether there are variations in the correlation of the phenotype/fitness landscape. An obvious example of another genotype-phenotype relation is the relation between protein sequence and protein higher order structure. Because in proteins there is a greater choice of building blocks which can be used to form higher order structures (20 amino acids versus 4 nucleotides) we expect to find as least as much variation in the correlation of protein landscapes as we find in RNA landscapes.

It has often been argued that selection does not operate at a level higher than that of the individual. This precludes selection for properties, like the genotype-phenotype relation, which affect the evolutionary process itself. Here we observe selection for the genotype-phenotype relation, which is possible because the selection acts at the level of the quasispecies (Eigen and Schuster, 1979; Schuster 1989). This type of selection can be referred to as meta-selection, that is, selection for properties that affect the "evolvability". In this sense it resembles one of the explanations for the evolution of sex (for a review see Hamilton et al. 1990), namely that sex evolved because it facilitates evolutionary adaptation of hosts to resist parasites.

Acknowledgements
The investigations were supported by the Foundation for Biological Research (BION), which is subsidized by the Netherlands Organization for Scientific Research.

References

Agur Z. (1987) Resilience and Variability in Pathogens and Hosts IMA J. Math. Appl. Med. & Biol. 4: 295-307

Agur Z., & Kersberg M.(1987) The emergence of phenotypic novelties through progressive change. Am. Nat. 129: 862-875

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.

French, D., Laskov, R., and Scharff, M. (1989). The role of somatic hypermutation in generation of antibody diversity. Science 244: 1152-1157

Hamilton W.D., Axelrod R., and Tanese R. (1990) Sexual reproduction as an adaptation to resist parasites (A Review). Proc. Natl. Acad. Sci. USA 87:3566-3573.

Hogeweg, P. & Hesper, B. (1984). Energy directed folding of RNA sequences. Nucleic Acids Res. 12, 67-74.

Hogeweg, P. & Hesper, B. (1992). Evolutionary dynamics and the coding structure of sequences: multiple coding as a consequence of cross-over and high mutation rates. Computers and Chemistry 16, 171-182.

Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. Ann Arbor: 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. (1993) Multiple coding and the evolutionary properties of RNA secondary structure. J. Theor. Biol., in press.

Huynen, M. A. & Hogeweg, P. (1993) Pattern formation in molecular evolution; exploitation of variation in the correlation of RNA landscapes. (submitted)

Ikegami T. & Kaneko K.(1992) Evolution of host-parasitoid network through homeochaotic dynamics. CHAOS 2: 397-407

Jaeger, J.A., Turner, D.H. & Zuker, M. (1989). Improved predictions of secondary structures for RNA. Proc. Natl. Acad. Sc. USA 86, 7706-7710.

Kaneko K. & Ikegami T.(1992) Homeochaos: dynamic stability of a symbiotic network with population dynamics and evolving mutation rates. Physica D 56:406-429

Kaufmann, S. A. (1988). 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. Mol. Biol. 207, 597-614.

Nowak M. A., May R. M. and Anderson R.M. (1990) The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease. AIDS 4:1095-1103

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: Addison Wesley.

Weinberger E.(1990) Correlated and Uncorrelated Fitness Landscapes and How to Tell the Difference. Biol. Cybern. 63:325-336

figure 1)
Relation of the dissimilarity between RNA sequences and the mean dissimilarity between RNA secondary structures for random sequences and for sequences that evolved for a "cloverleaf" like structure (4 stacks, 3 hairpin loops) with a high frequency of base-pairing (> 80 %) . As is shown, the evolved sequences show less change in secondary structure after point mutations than do random sequences.

figure 2)
Cumulative dissimilarities in secondary structure for point mutations. Starting sequences are the Rev Response Rlements (RRE) in CAEV and Visna virus and sequences that had evolved (with a Genetic Algorithm) for a "cloverleaf" secondary structure with a high frequency of base-pairing (as in the RREs). The cumulative dissimilarity in secondary structure after a number of point mutations is calculated by adding the dissimilarity in secondary structure between the neighbouring sequences in sequence space, e.g. the cumulative dissimilarity after 2 point mutations is the dissimilarity in secondary structure between "zero" and "one" mutants plus that between "one" and "two" mutants. The curves for Visna and for the artificial sequence are shifted 20 and 40 mutations to the right respectively.
The figure shows that the dissimilarity between the secondary structures of neighbouring sequences increases as one moves away from the selected sequences; the selected sequences are thus located in a relatively "smooth" part of the RNA landscape.

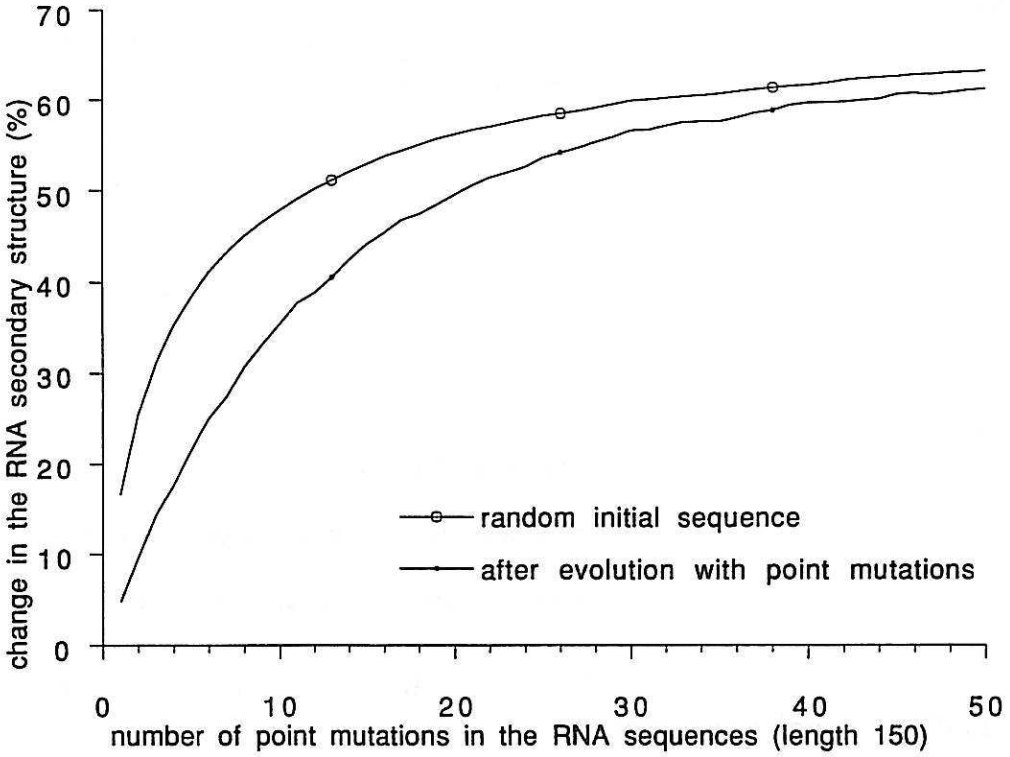
figure 3)
Decrease in fitness after point mutations in sequences that evolved for a two hairpin structure (fitness is the product of the length of the two hairpins, max=25) with point mutations and with or without crossing-over. The upper curve shows the effect of point mutations on the secondary structure of sequences that during the simulation have reached the required secondary structure for "the first time" (generally within 500 generations). The lower curve is of populations that have evolved for 30000 generations. Shown is the the mean of 25 runs. As is shown the resistance against mutations increases during evolution, especially if cross-over is used.

figure 4)
The distribution of purines and pyrimidines in the sequences of one population after evolution for a double-hairpin secondary structure. Shown is the "co-occurence" of the purines or pyrimidines between the different positions in all the sequences of one population (N=200, the original distribution wich lies between 0 and 200 is drawn between -1 and +1). The "co-occurence" of purine contents with themselves (the diagonal) is set to zero, its gray-level serves as reference. The presentation visualizes that interaction is only possible between purines and pyrimidines, and not between purines and purines or between pyrimidines and pyrimidines. White regions can only interact with black regions, and regions with alternating black and white can only interact with other regions with alternating black and white. Thus the only way the presented structure can possibly fold, even after a few mutations, is by forming two unconnected hairpins.

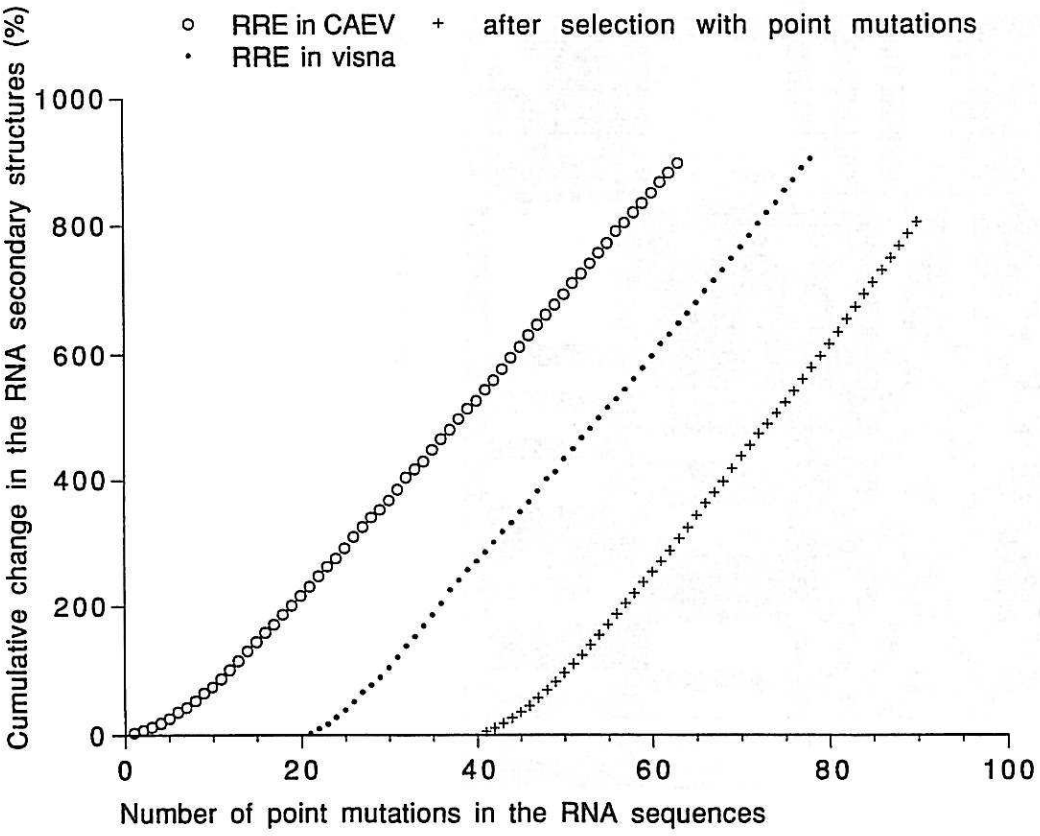
figure 5)
The amount of variation within the populations, both genotypic and phenotypic, and the mean fitness during evolution for a double-hairpin structure. The variation shown is the mean difference between two sequences per population. As the data show, the genotypic variation increases long after the desired secondary structures are reached. (all 25 populations reach the required secondary structure within 2,000 generations, more than 85% within 500 generations).

figure 6)
The effect of point mutations on the secondary structure of random sequences (equal probabilities for all nucleotides in all positions) and of sequences in a co-evolving system after 5,000 time-steps. The fitness criterion for sequences in population A is the similarity of their secondary structure to the secondary structure of sequences in population B, whereas for population B the fitness depends on the dissimilarity of its secondary structures to those of A (population A thus "chases" population B in a Red Queen like fashion). Both populations show an increased effect of point mutations on their secondary structures; the effect however is strongest in population B.

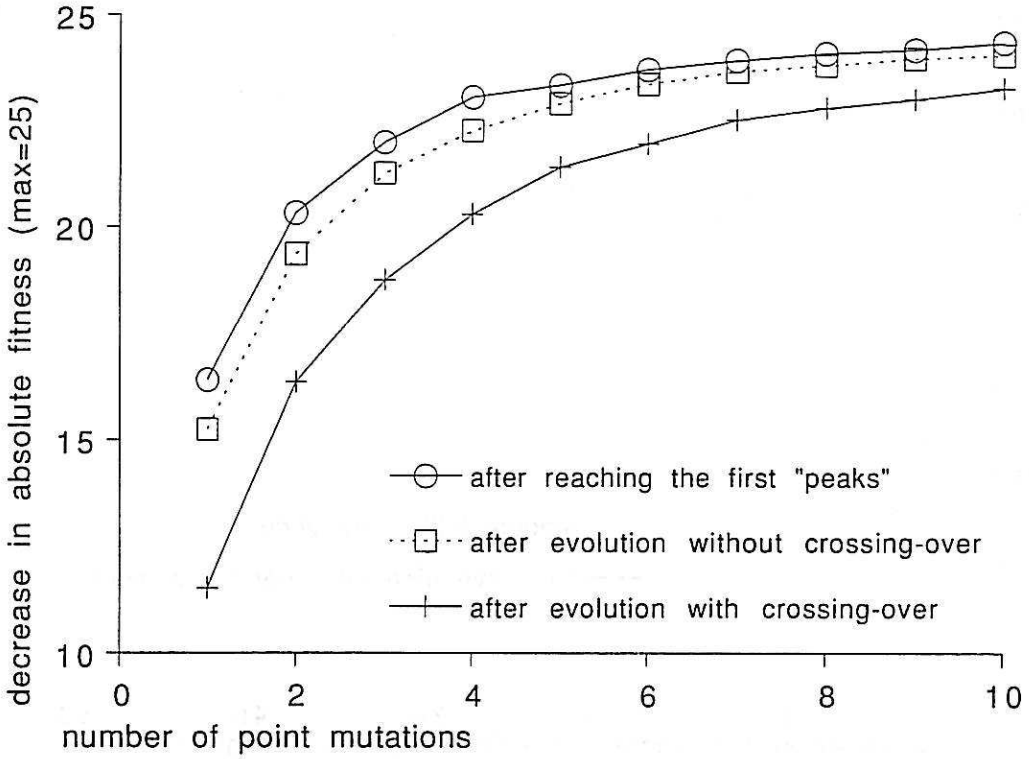
1


change in the RNA secondary structure (%)
70
60
50
40
30
20
10
0
random initial sequence
after evolution with point mutations
0
10
20
30
40
50
number of point mutations in the RNA sequences (length 150)

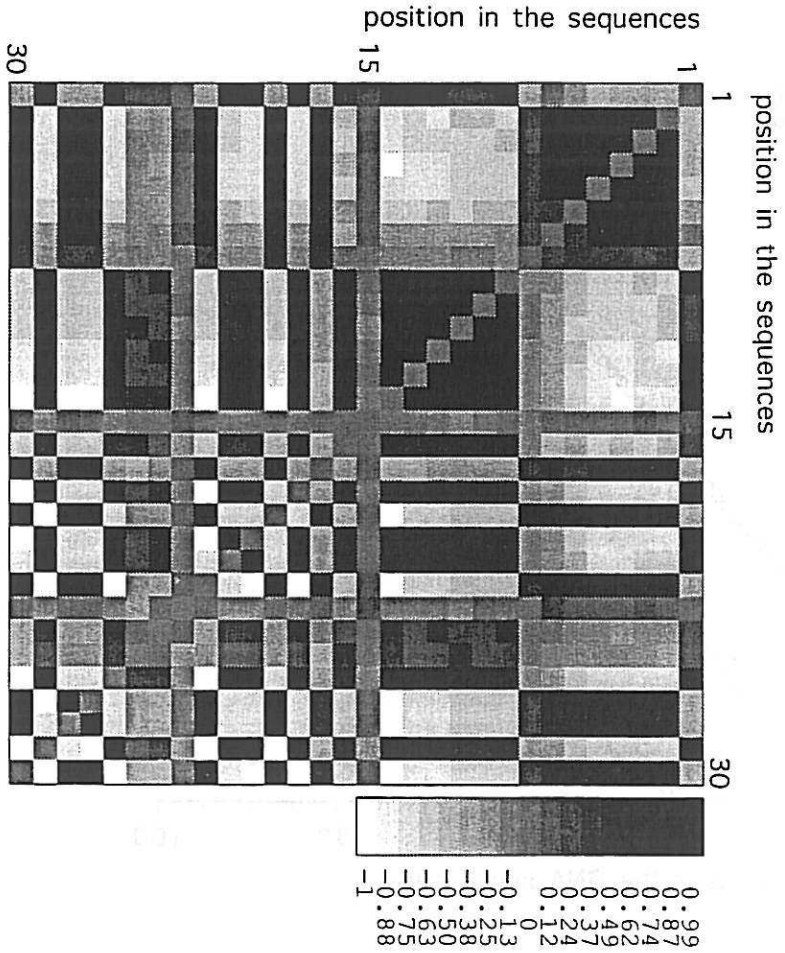
2


RRE in CAEV
RRE in visna
after selection with point mutations
Cumulative change in the RNA secondary structures (%)
1000
800
600
400
200
0
0
20
40
60
80
100
Number of point mutations in the RNA sequences

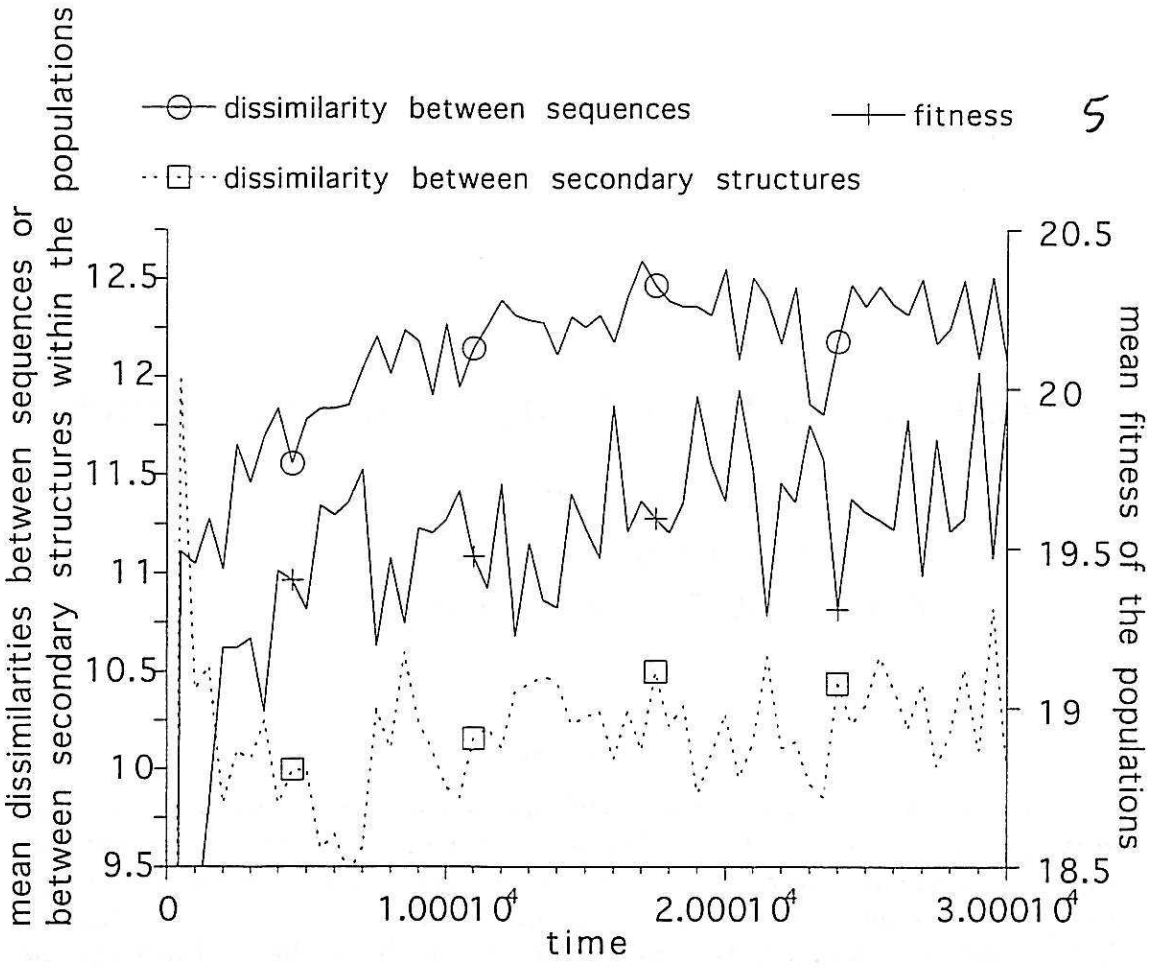
3


25
20
15
10
0
2
4
6
8
10
decrease in absolute fitness (max=25)
number of point mutations
after reaching the first "peaks"
after evolution without crossing-over
after evolution with crossing-over

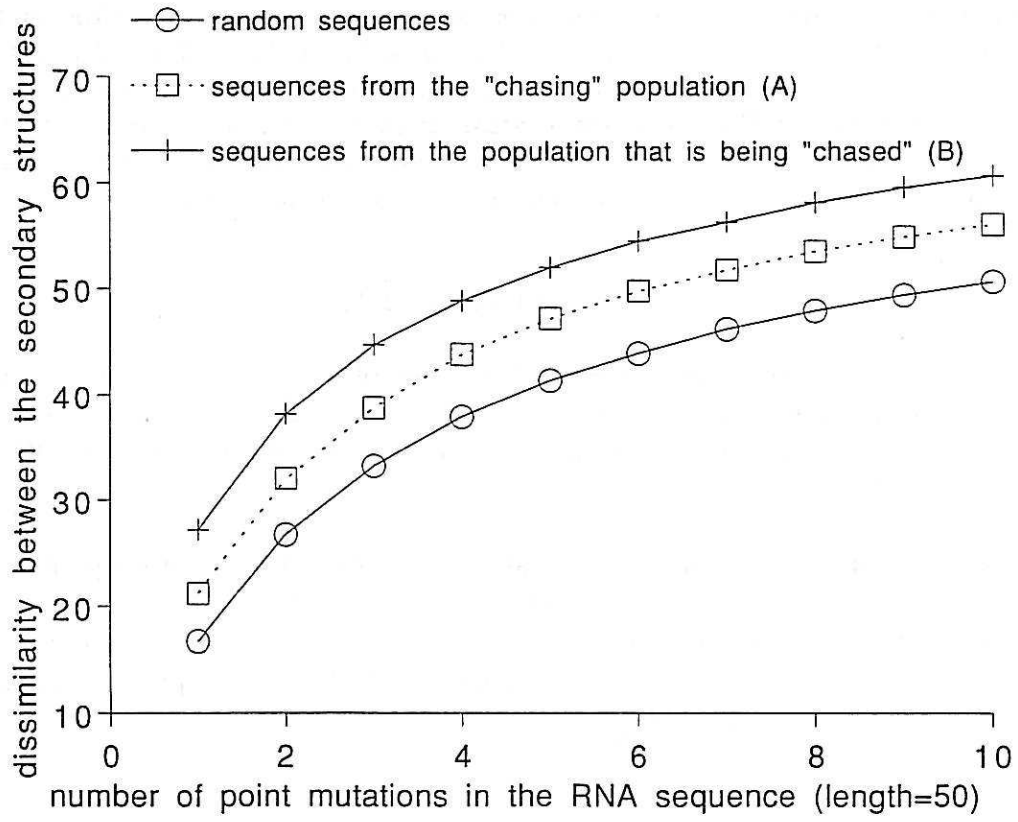
4


position in the sequences
30
15
1
1
15
30
position in the sequences
0.99
0.87
0.74
0.62
0.49
0.37
0.24
0.12
0.0
-0.13
-0.25
-0.38
-0.50
-0.63
-0.75
-0.88

5


dissimilarity between sequences
fitness
dissimilarity between secondary structures
mean dissimilarities between sequences or between secondary structures within the populations
mean fitness of the populations
12.5
12
11.5
11
10.5
10
9.5
20.5
20
19.5
19
18.5
0
1.0001 0^4
2.0001 0^4
3.0001 0^4
time

6


random sequences
sequences from the "chasing" population (A)
sequences from the population that is being "chased" (B)
dissimilarity between the secondary structures
70
60
50
40
30
20
10
0
2
4
6
8
10
number of point mutations in the RNA sequence (length=50)