RNA BASED EVOLUTIONARY OPTIMIZATION

PETER SCHUSTER

Institut für Molekulare Biotechnologie, D-07745 Jena, Germany, Santa Fe Institute, Santa Fe, NM 87501, U.S.A., and Institut für Theoretische Chemie der Universität Wien A-1090 Wien, Austria

(Received July 2, 1993)

Abstract. The notion of an RNA world has been introduced for a prebiotic scenario that is dominated by RNA molecules and their properties, in particular their capabilities to act as templates for reproduction and as catalysts for several cleavage and ligation reactions of polynucleotides and polypeptides. This notion is used here also for simple experimental assays which are well suited to study evolution in the test tube. In molecular evolution experiments fitness is determined in essence by the molecular structures of RNA molecules. Evidence is presented for adaptation to environment in cell-free media. RNA based molecular evolution experiments have led to interesting spin-offs in biotechnology, commonly called 'applied molecular evolution', which make use of Darwinian trial-and-error strategies in order to synthesize new pharmacological compounds and other advanced materials on a biological basis.

Error-propagation in RNA replication leads to formation of mutant spectra called 'quasispecies'. An increase in the error rate broadens the mutant spectrum. There exists a sharply defined threshold beyond which heredity breaks down and evolutionary adaptation becomes impossible. Almost all RNA viruses studied so far operate at conditions close to this error threshold. Quasispecies and error thresholds are important for an understanding of RNA virus evolution, and they may help to develop novel antiviral strategies.

Evolution of RNA molecules can be studied and interpreted by considering secondary structures. The notion of sequence space introduces a distance between pairs of RNA sequences which is tantamount to counting the minimal number of point mutations required to convert the sequences into each other. The mean sensitivity of RNA secondary structures to mutation depends strongly on the base pairing alphabet: structures from sequences which contain only one base pair (**GC** or **AU** are much less stable against mutation than those derived from the natural (**AUGC**) sequences. Evolutionary optimization of two-letter sequences in thus more difficult than optimization in the world of natural RNA sequences with four bases. This fact might explain the usage of four bases in the genetic language of nature.

Finally we study the mapping from RNA sequences into secondary structures and explore the topology of RNA shape space. We find that 'neutral paths' connecting neighbouring sequences with identical structures go very frequently through entire sequence space. Sequences folding into common structures are found everywhere in sequence space. Hence, evolution can migrate to almost every part of sequence space without 'hill climbing' and only small fractions of the entire number of sequences have to be searched in order to find suitable structures.

1. The concept of an RNA world

RNA molecules were found to be unique in chemistry and biology since they can not only act as templates for their reproduction but are also catalysts for several reactions involving other RNA molecules or oligopeptides. Template induced reproduction of RNA molecules occurs in cells that were infected by RNA viruses or viroids (For the latter case see e.g. Riesner and Gross, 1985). Enzyme free template induced RNA synthesis has been studied extensively by Leslie Orgel (1992). Replication of RNA molecules under non-equilibrium conditions provides the basis for evolutionary adaptation in the sense of Darwin's mechanism as shown in the

Origins of Life and Evolution of the Biosphere **23**: 373–391, 1993.

next two sections. The discoveries of Thomas Cech and Sidney Altman (Cech, 1986; Guerrier-Takada *et al.*, 1983; Guerrier-Takada and Altman, 1984; Symons, 1992) have shown that, in contrast to previous belief, RNA molecules can catalyze several classes of reactions with similar high efficiency and specificity as protein enzymes. The notion of 'ribozymes' was created for such RNA based catalysts. The catalytic activities of RNA molecules are all related to cleavage or formation of nucleotide or peptide bonds. For RNA substrates ligation and cleavage are highly sequence specific. The capability of RNA to catalyze peptide bond cleavage and formation (Noller, 1991; Noller *et al.*, 1992; Piccirilli *et al.*, 1992) is of particular interest since it can be interpreted as a hint that primordial ribosomes might have worked without proteins. Although the catalytic repertoire of ribozymes is rather poor compared to that of protein enzymes, it comprises the key reactions that are necessary to produce polynucleotide from oligomers and monomers. It is generally believed nowadays that there was a period in the origin of life on Earth when RNA served as both, genetic material and specific catalyst (See e.g. Gilbert, 1966; Joyce, 1991). In other words RNA has the capability to be genotype and phenotype. A world of RNA molecules driven by a steady supply of suitable energy rich compounds represents also a kind of biological toy universe which allows to study the evolutionary process in its simplest form. Such an **RNA world** is a biological adaptive system of minimal complexity.

Is there room for an RNA world in the current view of chemical evolution on the primordial Earth? In Figure 1 we show a sketch of a plausible sequence of scenarios leading to the onset on RNA based Darwinian evolution (for more details see e.g. Schuster, 1981; Mason, 1991). The major problem with RNA in chemical evolution is to seen in its highly elaborate molecular structure: for RNA template induced reactions molecules of high stereochemical purity are required, and chiral compounds such as ribose have to be present as pure single antipods (racemic mixtures of D- and L-ribose incorporated into polynucleotides are prohibitive for template induced replication). It was suggested therefore that less complicated molecules which could nevertheless act as templates preceded the RNA world (Orgel, 1986; Joyce, 1989). Recent progress in **template chemistry** (Orgel, 1992) makes this suggestion more and more plausible. RNA molecules, on the other hand, are the first template molecules which share most of their chemical properties with DNA the predominant genetic information carrier at present. Information transfer from DNA to RNA and vice versa, known as transcription and reverse transcription, is routine in present day biochemistry. We can speculate therefore that a phylogenetic extrapolation of present day genetic information may go back as far as to the first efficiently replicating species of a primordial RNA world. In short, RNA molecules seem to appear late in chemical evolution and very early in biological evolution.

According to the analysis of microfossils carried out by William Schopf (Schopf and Packer, 1992) the first sure remnants of microorganisms are dated 3.4×10^9 years ago. These microorganisms were most likely photosynthetic and related to present day cyanobacteria. Needless to say, there was a long way with many inter-

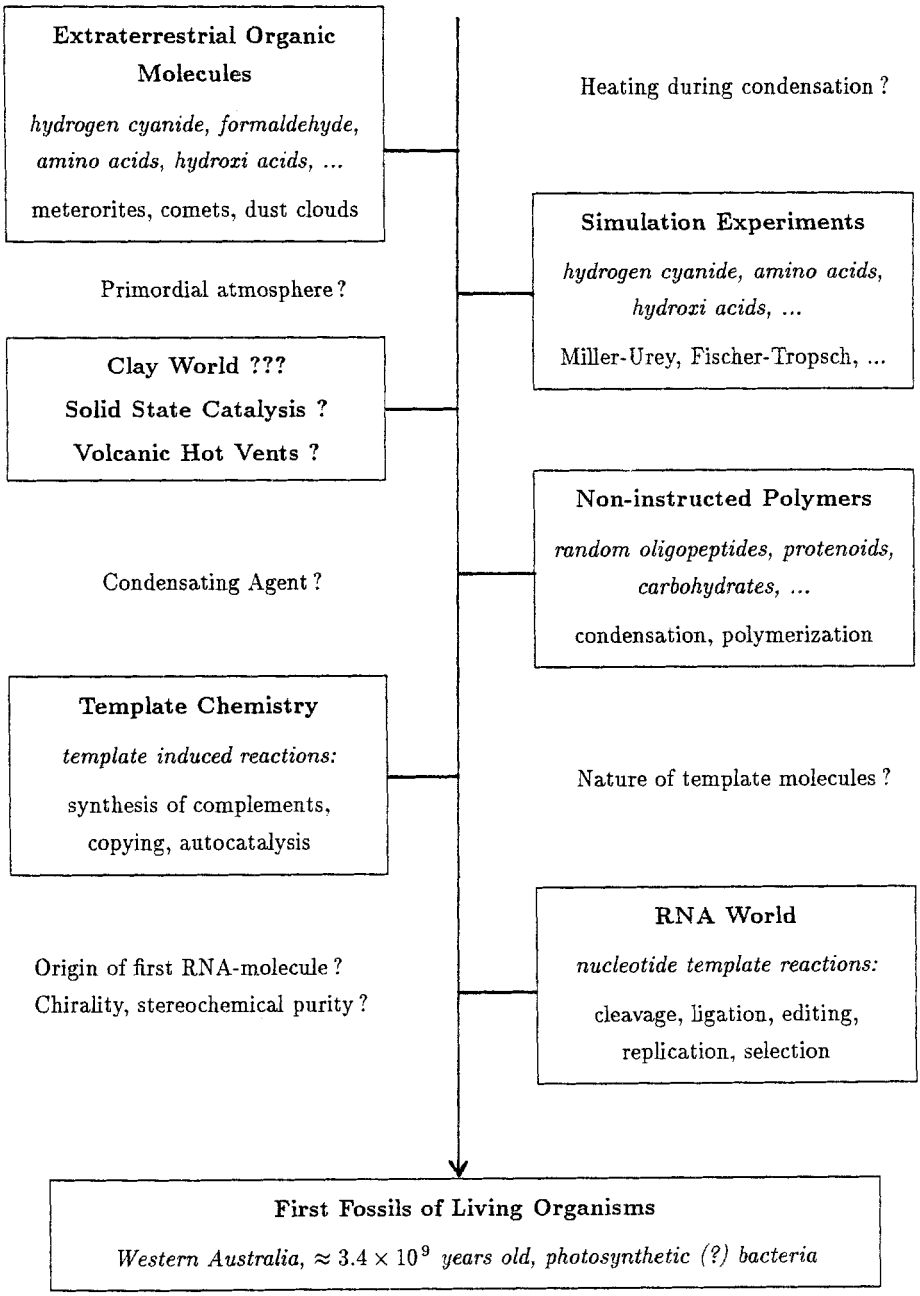

Fig. 1. Some facts and open questions about the origin of life.

mediate steps to go from a primitive Darwinian system like the RNA world to a fully developed phytosynthetic organism (Eigen and Schuster, 1985). The flow chart sketched in Figure 1 does not refer explicitly to these other 'worlds' which inevitably lay in between.

Evolution experiments in the pure RNA world have to face severe problems: template induced RNA synthesis is slow, does not work for many templates under the conditions applied so far, and the accuracy of base incorporation is low (of the order of one error in hundred nucleotides). Another problem concerns separation of double strands which does not occur readily under the conditions of template induced synthesis. In order to explore the capabilities of a minimal complexity adaptive system we are, however, not restricted to a pure RNA world. Why not use protein catalysts as highly specific and efficient environmental factors? Addition of enzymes leads to a new toy universe for studies on evolution that might be characterized as a **protein assisted RNA world**. Sol Spiegelman (1971) has indeed shown that such a universe can be created in the test tube, and indeed, it shows all features of the Darwinian scenario. RNA dependent RNA polymerases, commonly called RNA replicases, occur in nature. Spiegelman used an enzyme isolated from bacterial cells of *E. coli* which were infected by the RNA bacteriophage Qβ. This enzyme replicates RNA fast and with much higher accuracy than achieved without a protein catalyst: many thousands of generations can be studied in experiments lasting less than a day, and the accuracy of replication is about one error in 3000 nucleotides.

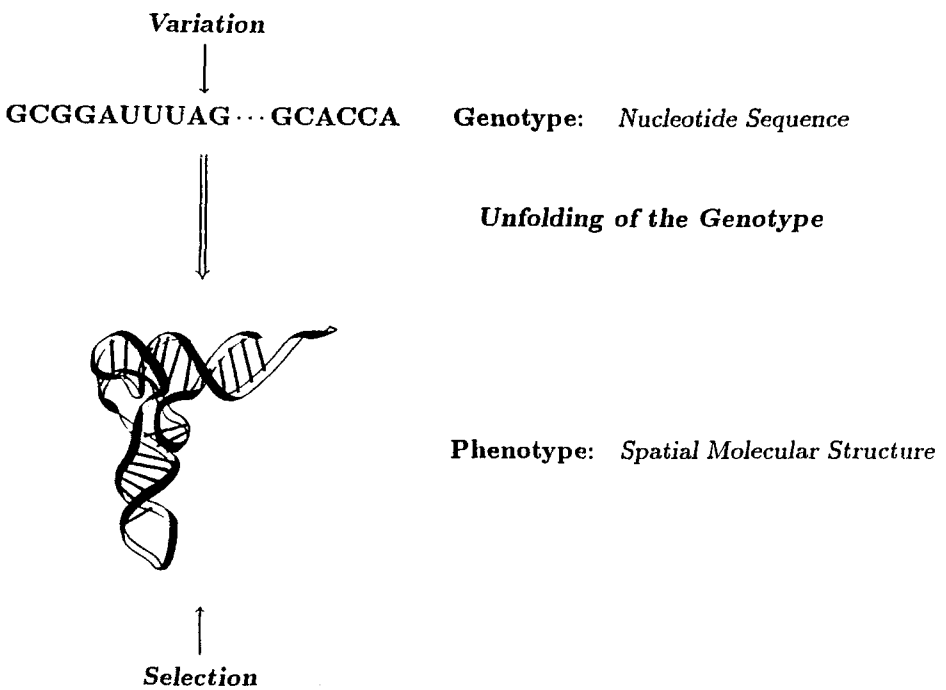

Fig. 2. Molecular genotypes and phenotypes in RNA evolution experiments.

2. Molecular genotypes and phenotypes

The dichotomy between genotypes and phenotypes is basic to biological evolution. Molecular biology has shown that the information for the unfolding of organisms (which represent the phenotypes) is contained in the DNA (which is the genotype). The Darwinian principle of evolution is based on heridity, variation and selection. According to our present day knowledge all inheritable variations are caused by changes of the genotype (i.e. by changes in the sequence of nucleotide bases of the DNA), and selection acts through differential reproductive success of phenotypes. RNA molecules in cell-free evolution experiments are both genotypes and phenotypes. As we see in Figure 2 the genotype is again a sequence of nucleotide bases, this time in the form of an RNA molecule. Following Sol Spiegelman we consider the spatial structure of the RNA molecule as its phenotype. Indeed, this structure is

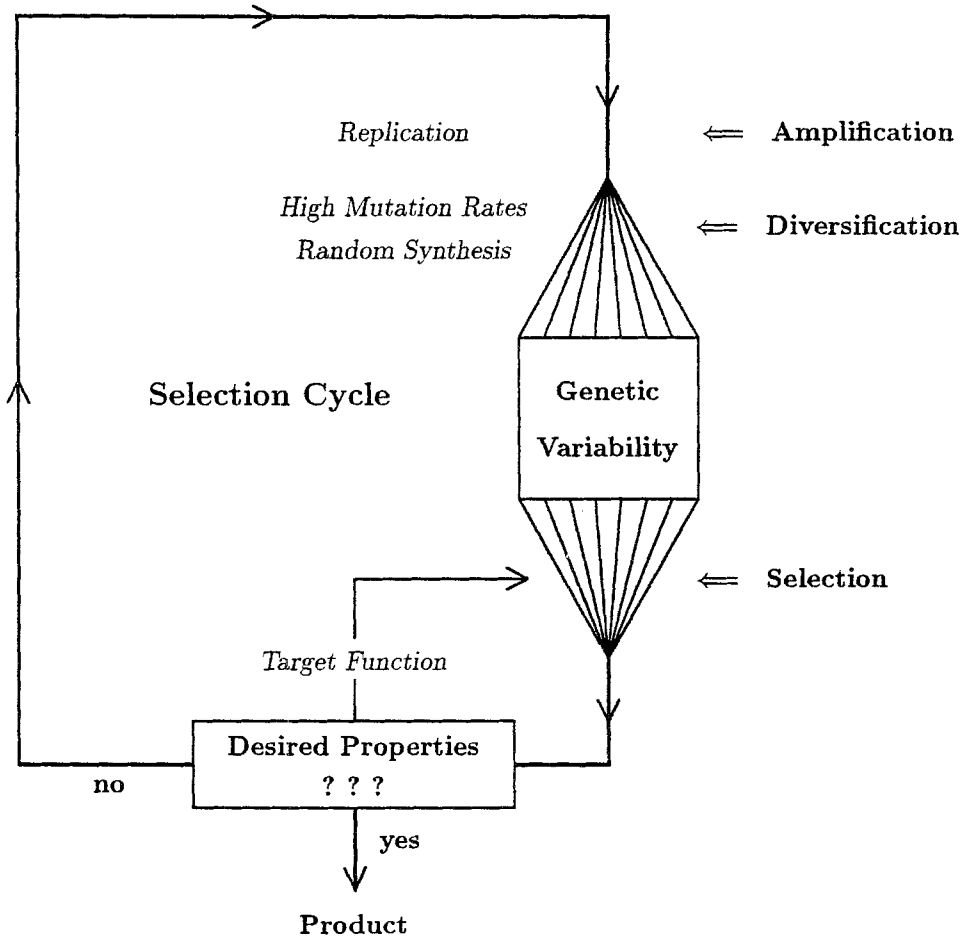

Fig. 3. A selection technique of applied molecular evolution based on encoding of the function to be developed into the selection constraint.

evaluated by the selection process. Systematic kinetic studies on RNA replication and test tube evolution carried out by Christof Biebricher in Manfred Eigen's laboratory (Biebricher *et al.* 1983, 1984, 1985) revealed the mechanism of RNA replication. The rate and equilibrium constants which determine the outcome of selection under different conditions are known now. They can be determined independently from selection experiments by direct measurements. In addition minimum requirements for efficient replication were discovered. They consist in sequence regularities and sufficient secondary structure to facilitate double strand separation by the enzyme.

Selection experiments may be carried out by discontinuous renewal of the consumed material as in the serial transfer technique (Spiegelman, 1971) or under the continuous constraint maintained in elaborate evolution reactors (Husimi *et al.*, 1982; Husimi and Keweloh, 1987). An new selection technique was used by

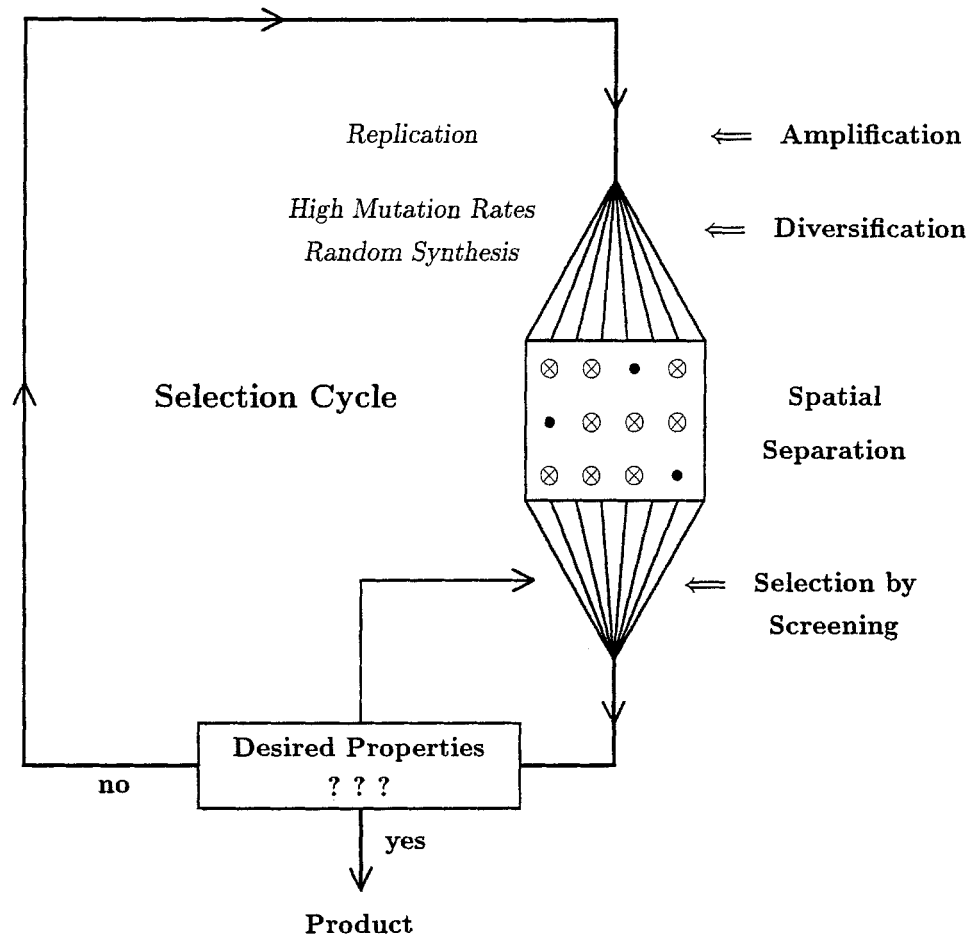

Fig. 4. A selection technique of applied molecular evolution based on massively parallel screening.

John McCaskill in his molecular evolution experiments in capillaries (Bauer *et al.*, 1989a): the capillaries contain a medium suitable for replication, RNA is injected and a wave front spreads through the medium. The front velocity of the traveling wave increases with the replication rate and hence, faster replicating mutants are selected by the wave propagation mechanism.

The capability of RNA molecules to evolve by natural selection is used now in applied molecular evolution to produce biomolecules with new properties. This novel evolutionary technology was originally suggested by Eigen and Gardiner (Eigen and Gardiner, 1983). Somewhat later Kauffman suggested the use of random sequences in selection experiments (Kauffman 1986; for a recent review see Kauffman, 1992). One approach to the problem is suitable for 'batch' experiments (Figure 3). The essential trick of this technique is to encode the desired functions into the selection constraint. As indicated in Figure 3 sufficient variation is introduced into populations either by artificially increased mutation rates or by partial randomization of sequences. The synthesis of oligonucleotides with random sequences is routine nowadays. Several examples of successful applications of molecular selection techniques to biochemical problems are found in the current literature (Horwitz *et al.*, 1989; Tuerk and Gold, 1990; Ellington and Szostak, 1990; Beaudry and Joyce, 1992). In reality it will often be impossible to encode the desired function directly into the selection constraint. Then spatial separation of individuals and massively parallel screening provides a solution (Figure 4). This technique, however, requires highly sophisticated equipment which is currently under development (Bauer *et al.*, 1989b).

3. Adaptation to environmental changes

The most frequently occurring genotypes (commonly called 'master sequences') can be isolated from populations and analyzed during the course of molecular evolution experiments. These investigations show how the structure of RNA molecules is adjusted in order to cope with changes in the environment. Such environmental changes are caused by adding suitable substances to the replication medium which deteriorate the conditions for reproduction. For example, heterocyclic dyes (ethidium bromide, acridiniuum orange, etc.) which intercalate between Watson-Crick base pairs and thus interfere with replication therefore, can be used (Kramer *et al.*, 1974), or ribonucleases (RNA cleaving enzymes) can be applied (Strunk, 1992). Mutants which compensate for the change in the environment by having fewer binding or cleavage sites than the wild type, appear and are enriched in the populations.

Replication errors lead to new molecular species whose replication efficiency is evaluated by the selection mechanism. The higher the error rate, the more mutations occur and the more viable mutants appear in the population. The stationary mutant distribution is characterized as 'quasispecies' since it represents the genetic reservoir of asexually replicating populations. An increase in the error rate thus leads to a

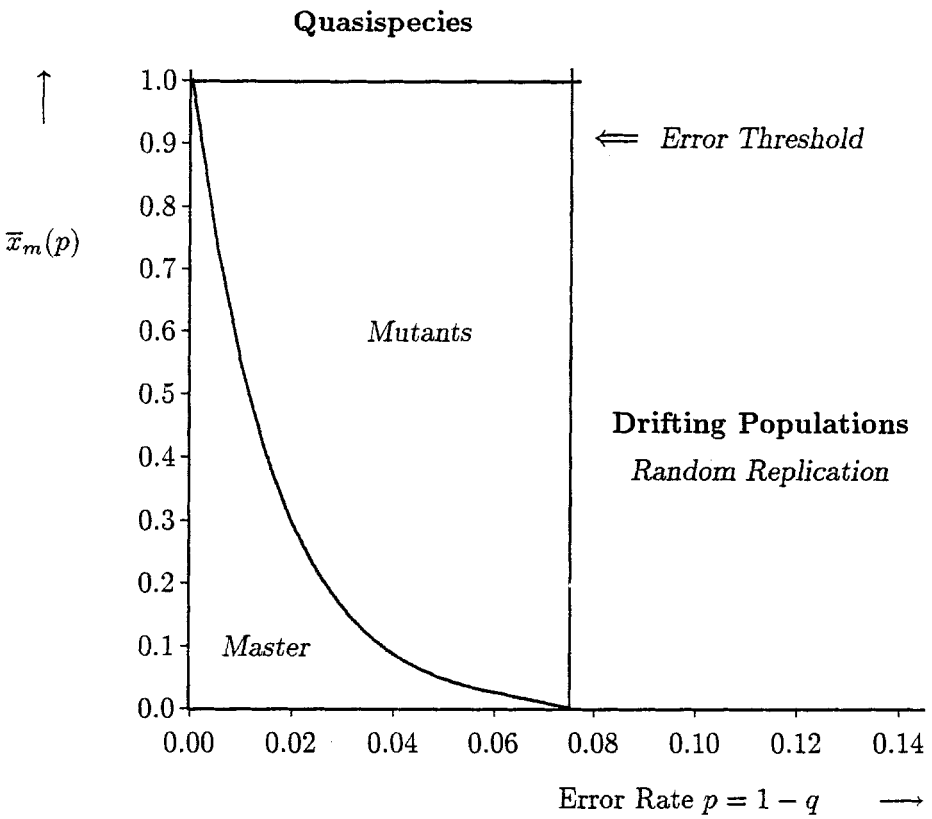

Fig. 5. Evolution at the error threshold of replication. The fraction of the most frequent species in the population, called the **master sequence**, is denoted by $x_m(p)$. It becomes very small at the error threshold. Accordingly, the total fraction of all mutants, $1 - x_m(p)$, approaches one at the critical mutation rate.

broader spectrum of mutants and makes evolutionary optimization faster and more efficient in the sense that populations are less likely caught in local fitness optima. There is, however, a critical error threshold (Eigen, 1971; Eigen and Schuster, 1979; Swetina and Schuster, 1982): if the error rate exceeds the critical limit heredity breaks down, populations are drifting in the sense that new RNA sequences are formed steadily, old ones disappear, and no evolutionary optimization according to Darwin's principle is possible (Figure 5). Variable environments require sufficiently fast adaptation and species with tunable error rates will adjust their quasispecies to meet the environmental challenge. In constant environments, on the other hand, such species will tune their error rates to the smallest possible values in order to maximize fitness.

Viruses are confronted with extremely fast changing environments since their hosts developed a variety of defense mechanisms ranging from the restriction enzymes of bacteria to the immune system of mammals and man. RNA viruses have been studied extensively. Their multiplication is determined by enzymes that do not allow

large scale variations of the replication accuracies. They vary the error rate by changing the length of their genomes, and adjust the RNA chain lengths to optimal values, which correspond to maximal chain lengths (Eigen, 1971; Eigen and Schuster, 1977):

$$\nu_{max} = \frac{\ln \sigma}{p}. \tag{1}$$

Herein ν_{max} is the maximal chain length that a master sequence can adopt and still allow for stable replication over many generations, $\sigma \geqslant 1$ is the superiority of this master sequence in the stationary population and p is the error rate per (newly incoporated) base and replication event. The superiority expresses differential fitness between the master sequence and the average of the remaining population. In the limit $\lim \sigma \rightarrow 1$ we are dealing with 'neutral evolution' (Kimura, 1983). Experimental analysis of several RNA virus populations has shown, that almost all chain lengths are adjusted to yield error rates close to the threshold value. Thus, RNA viruses appear to adapt to their environments by driving optimization efficiency towards the maximum.

4. Evolutionary stability of RNA structures

No physical process can occur with ultimate accuracy and hence, mutations are unavoidable. Given a certain mutation rate, we may ask what are the differences in stability of RNA structures against mutation. How likely does a change in the sequence result in an actual change in the structure? In other words we try to estimate the fraction of neutral mutants in the neighbourhood of a typical sequence ('neutral' is commonly used for sequences which have properties that are indistinguishable for the selection process; we shall use the term here in the narrower sense of identical structures). This question is of statistical nature and can be answered only by proper application of statistical techniques.

Firstly, the RNA sequences have to be ordered in a natural way. In case point mutations (single base exchanges) are predominant, the Hamming distance (d_h, counting the number of positions in which two aligned sequences differ) is an appropriate measure of the relationship between two sequences, since it counts the minimum number of point mutations required to convert one sequence into another. Moreover, the Hamming distance d_h induces a metric on the sequence space. The space of binary sequences of chain length ν=4 is shown in Figure 6 as an example.

RNA secondary structures are first approximations to the spatial structures of RNA molecules. They are understood as a listing of the Watson-Crick-type base pairs in the actual structure and may be represented as planar graphs (Figure 7). We consider RNA secondary structures as elements of an abstract 'shape space'. Again a measure of relationship of RNA structures which induces a metric on the shape space can be found (Fontana *et al.*, 1991, 1993a, b). We derived this

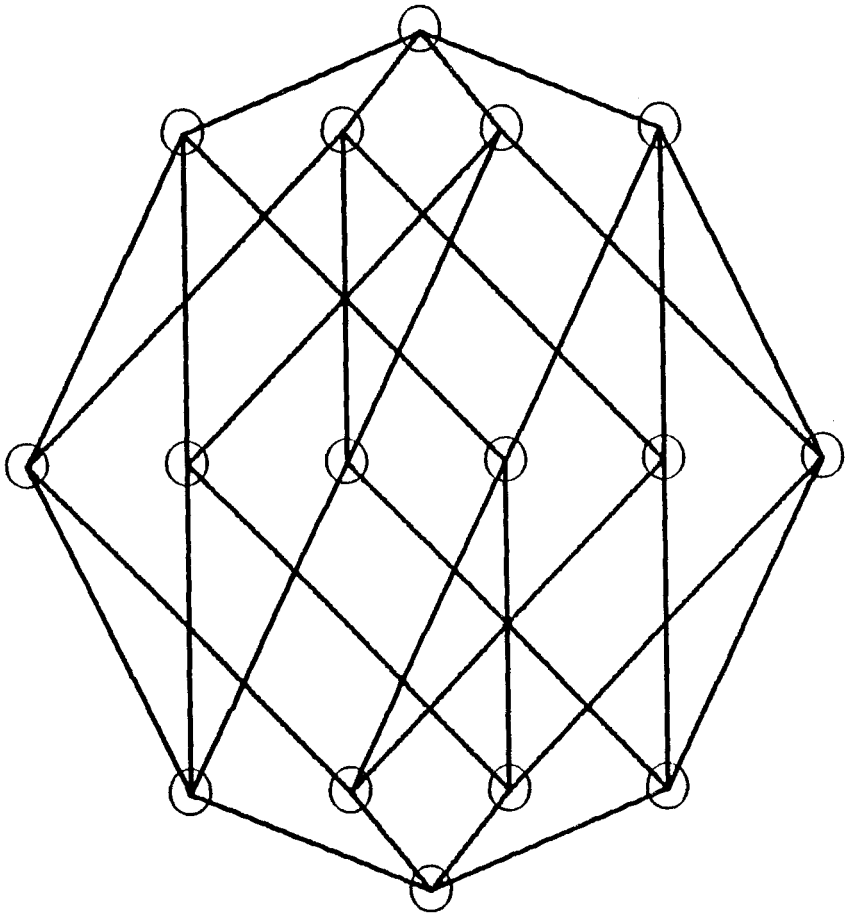

Fig. 6. The sequence space of binary (AU or GC) sequences of chain length $\nu = 4$. Every circle represents a single sequence of four letters. All pairs of sequences with Hamming distance $d_h = 1$ (these are pairs of sequences that differ in one position) are connected by a straight line. The geometric object obtained is a hypercube in four-dimensional space and hence, all positions (and all sequences) are topologically equivalent.

distance measure from trees which are equivalent to the structure graphs, and accordingly it is called the 'tree distance', d_t. RNA folding thus can be understood as a mapping from one metric space into another, in particular from sequence space into shape space. A path in sequence space corresponds uniquely to a path in shape space (The inversion of this statement is not true as we shall see in the next sextion 5).

The whole machinery of mathematical statistics and time series analysis can now be applied to RNA folding. In particular, an autocorrelation function of structures based on tree distances (d_t) is computed from the equation

$$\rho_t(h) = 1 - \frac{<d_t^2(h)>}{<d_t^2>}. \qquad (2)$$

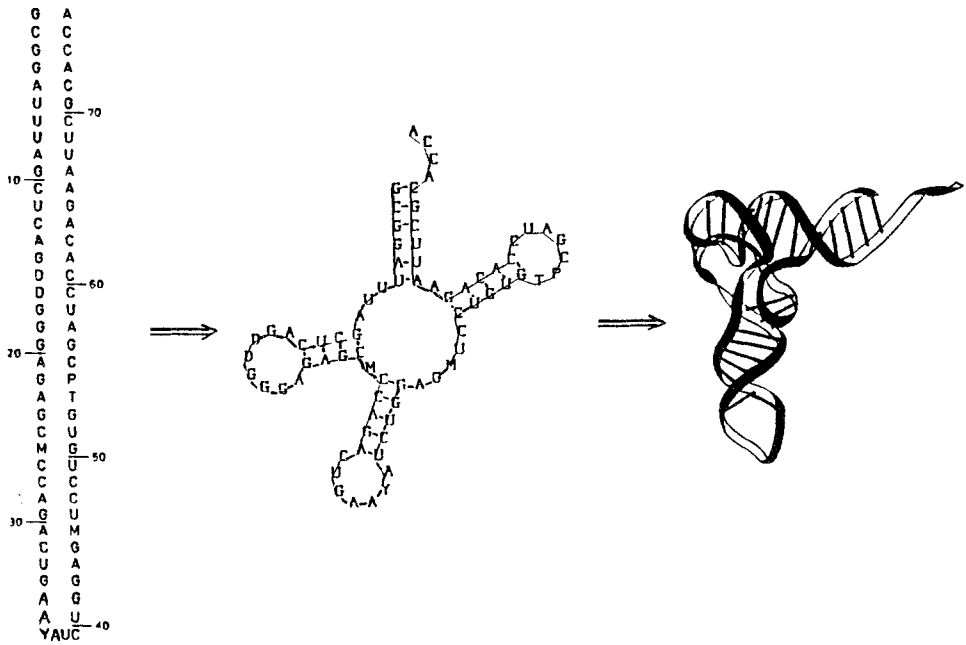

Fig. 7. Folding of an RNA sequence into its spatial structure. The process is partitioned into two phases: in the first phase only the Watson-Crick-type base pairs are formed (which constitute the major fraction of the free energy), and in the second phase the actual spatial structure is built by folding the planar graph into a three-dimensional object. The example shown here is phenylalanyl-transfer-RNA (t-RNA^{phe}) whose spatial structure is known from X-ray crystallography.

Mean square averages are taken here over all sequences in sequence space ($<d_t^2>$), or over all sequences in the mutant class h of the reference sequence ($<d_t^2(h)>$), i.e. over all sequences at Hamming distance h from the reference). The autocorrelation functions can be approximated by exponential functions, and correlation length (l_t) are estimated from the relation: $\ln(\rho_t(l_t)) = -1$.

The correlation length is a statistical measure of the hardness of optimization problems (see e.g. Eigen *et al.*, 1989). The shorter the correlation length, the more likely is a structural change occurring as a consequence of mutation. The correlation length thus measures stability against mutation. In Figure 8 correlation lengths of RNA structures are plotted against the chain length. An almost linear increase is observed. Substantial differences are found in the correlation lengths derived from different base pairing alphabets. In particular, the structures of natural (**AUGC**) sequences are much more stable against mutation than pure **GC**-sequences or pure **AU**-sequences. This observation is in agreement with structural data obtained for ribosomal RNA (Wakeman and Maden, 1989). It provides also a plausible explanation for the use of two base pairs in nature: optimization in an RNA world

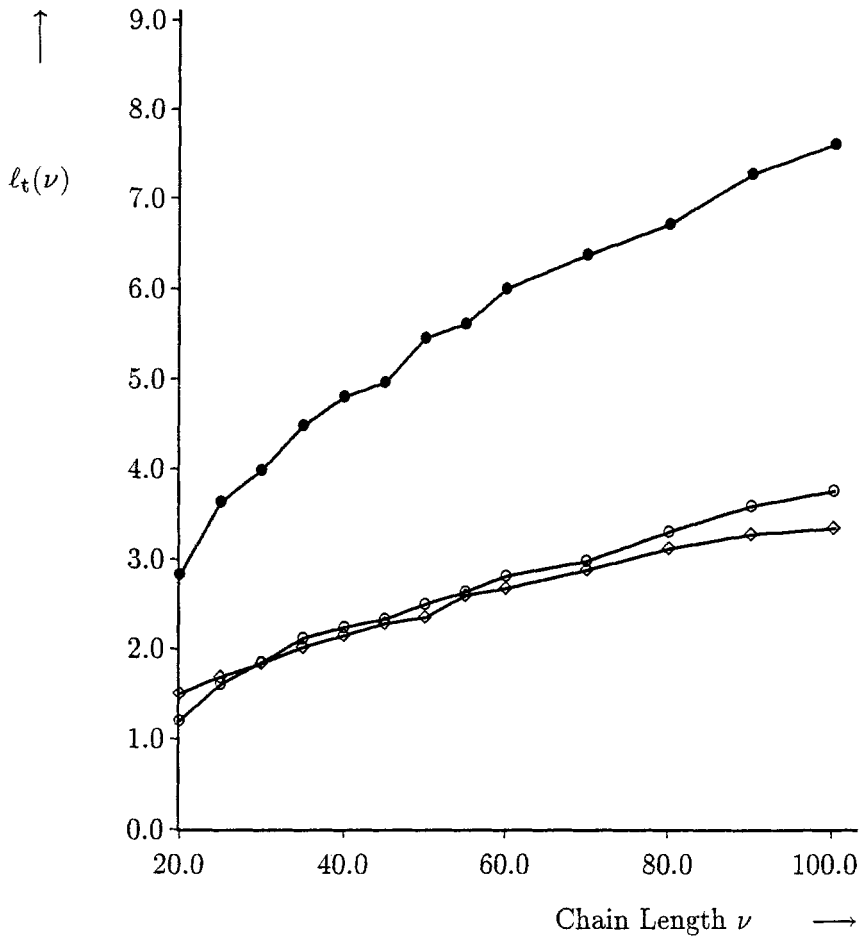

Fig. 8. Correlation lengths of structures (l_t) of RNA molecules in their most stable secondary structures as functions of the chain length ν. Values are shown for binary pure **GC**-sequences (◇), for binary pure **AU**-sequences (○)), and for natural **AUGC**sequences (●). The correlation lengths are computed from (ln $q_t(h)$, h)-plots by means of a least root mean square deviation fit.

with only one base pair would be very hard, and the base pairing probability in sequences with three base pairs is rather low and hence most random sequences of short chain lengths ($\nu<50$) do not form thermodynamically stable structures. The choice of two base pairs thus appears to be a compromise between stability against mutation and thermodynamic stability. An alternative explanation for the usage of two base pairs in nature was published recently (Szathmáry 1991, 1992): the current alphabet is understood to be optimal in an RNA world where replication fidelity decreases and catalytic efficiency increases with alphabet size. Both hypotheses are experimentally testable and hence we may expectt a decision in favor of one of the two alternatives in the future.

5. How to search for RNA structures

The sequence space is a bizarre object: it is of very high dimension (its dimension coincides with the chain length of RNA: $25 < \nu < 500$ for RNA molecules in test tube experiments, $250 < \nu < 400$ for viroids, and $3500 < \nu < 30\,000$ for (most) RNA viruses), but there are only few points on each coordinate axis (κ points; κ is the number of digits in the alphabet: $\kappa = 2$ for **AU** and **GC**, $\kappa = 4$ for **AUGC**). The number of secondary structures which are acceptable as minimum free energy structures of RNA molecules is much smaller than the number different sequences: in case of natural (**AUGC**) molecules we have about $1.485 \times \nu^{-3/2}(1.849)^{\nu}$ structures for 4^{ν} sequences (Schuster *et al.*, 1993). The mapping from sequence space into shape space thus cannot be invertible. We cannot expect therefore that our intuition which is well trained with invertible maps in three-dimensional space will guide us well through sequence and shape spaces. In order to obtain information on optimization of RNA structures and properties search algorithms were conceived (Fontana and Schuster, 1987; Wang, 1987) and computer simulations on realistic landscapes based on RNA folding were carried out (Fontana and Schuster, 1987; Fontana *et al.*, 1989). Here we shall adopt another strategy and try to develop an appropriate statistics to deal with such abstract objects as RNA shape space.

The information contained in the mapping from sequence space into shape space is condensed into a two-dimensional, conditional probability density surface

$$S(t, h) = \text{Prob}\,(d_t = t \mid d_h = h). \tag{3}$$

This structure density surface (SDS) expresses the probability that the secondary structures of two randomly chosen sequences have a structure distance t provided their Hamming distance is h. An example of a structure density surface for natural sequences of chain length $\nu = 100$ is shown in Figure 9. We recognize an overall shape that corresponds to one half of a horseshoe and superimposed on it rugged details. The contour plot illustrates an important property of the structure density surface: at short Hamming distances ($1 \leqslant \nu < 16$) the probability density changes strongly with increasing Hamming distance, but further away from the reference sequence ($16 < \nu < 100$) this probability density is essentially independent of the Hamming distance h. The first part reflects the local features of sequence-structure relations. Up to a Hamming distance of $h = 16$ there is still some memory of the reference sequence. Then, at larger Hamming distances the structure density surface contains exclusively global information which is independent of the reference.

In order to gain more information on the relation between sequences and structures an inverse folding algorithm which determines the sequences that share the same minimum free energy secondary structure was conceived and applied to a variety of different structures (Schuster *et al.*, 1993). The frequency distribution of structures has a very sharp peak: relatively few structures are very common, many structures are rare and play no statistically significant role. The results obtained show in addition that sequences folding into the same secondary structure are, in essence,

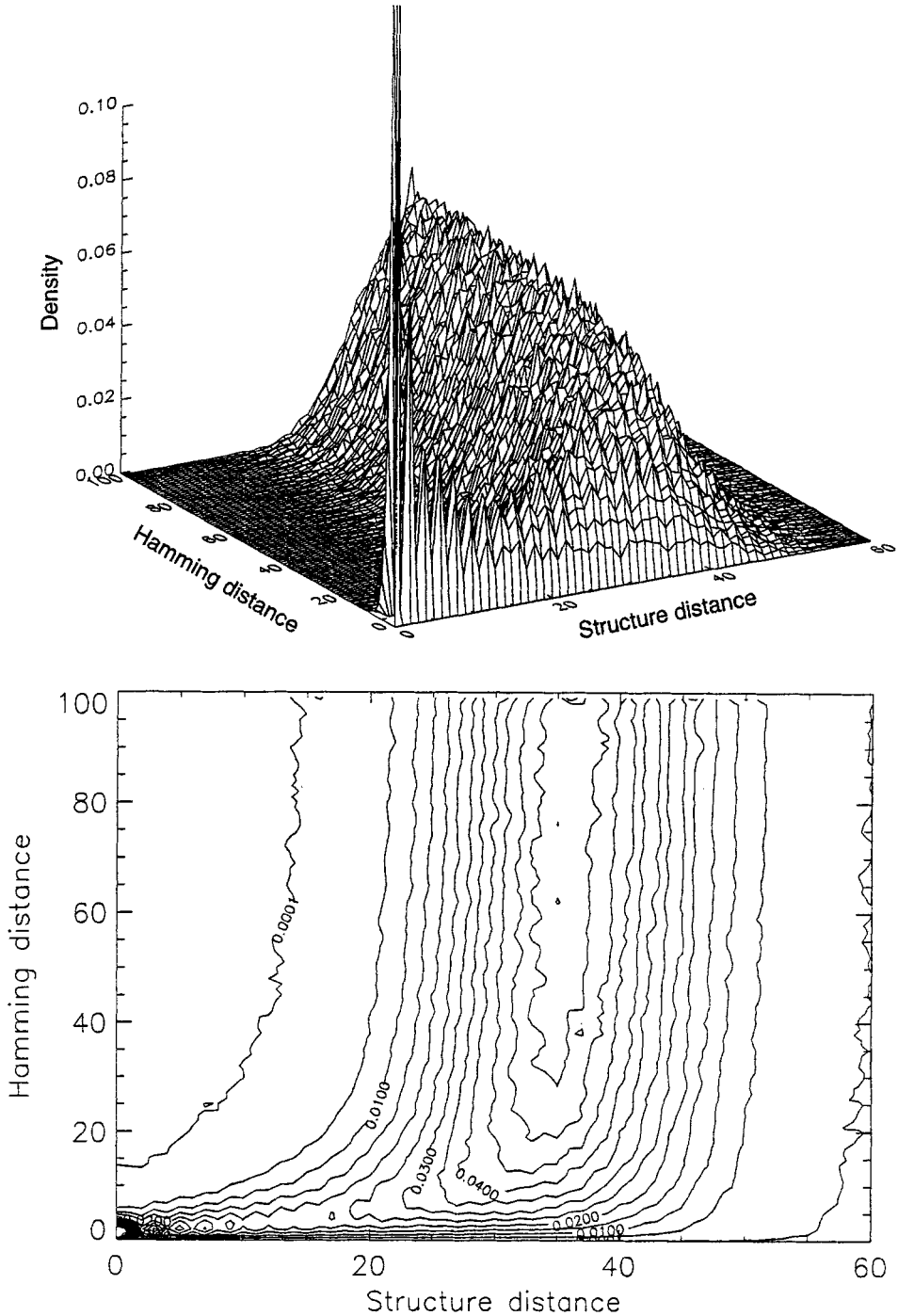

Fig. 9. The structure density surface $S(t, h)$ of natural **AUGC**-sequences of chain length $\nu = 100$. The density surface (upper part) is shown together with a contour plot (lower part). In order to dispense from confusing details the contour lines were smoothed. In this computation a sample of 1000 reference sequences was used which amounts to a total sample size of 10^6 individual RNA foldings.

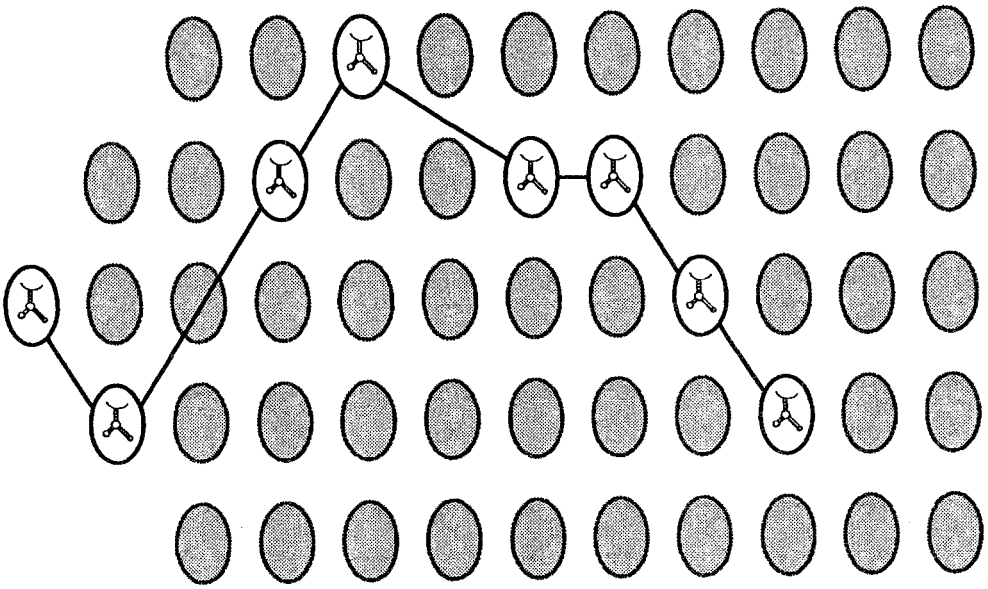

Fig. 10. Percolation of sequence space by neutral networks. A neutral path connects sequences of Hamming distance $h = 1$ (single base exchange) or $h = 2$ (base pair exchange) which fold into identical minimum free energy structures. The sketch shows a neutral path of length $h = 9$. The path ends since no to identical structure was found with $h = 10$ and $h = 11$ from the reference.

randomly distributed. Consequently all common structures are found in relatively small patches of sequence space. For natural sequences of chain length $\nu = 100$ a sphere of radius $h \approx 20$ (in Hamming distance) is sufficient to yield a global distribution of structure distances.

In order to complement this illustration of the RNA shape space, a computer experiment was carried out which allows an estimate of the degree of selective neutrality (two sequences are considered neutral here if the fold into the same secondary structure). As indicated in Figure 10 we search for 'neutral paths' through sequence space. The Hamming distance from the reference increases monotonously along such a neutral path but the structure remains unchanged. A neutral path ends when no further neutral sequence is found in the neighbourhood of the last sequence. The length l of a path is the Hamming distance between the reference sequence and the last sequence. Clearly, a neutral path cannot be longer than the length ν ($l \leqslant \nu$). The length distribution of neutral path in the sequence space of natural RNA molecules of chain length $\nu = 100$ is shown in Figure 11. It is remarkable that about 20% of the neutral paths have the maximum length, and thus lead through the whole sequence space to a sequence which differs in all position from the reference, but shares its structure.

Combination of information derived from Figures 9 and 11 provides insight into the structure of the shape space of RNA secondary structures which may be cast into four statements:

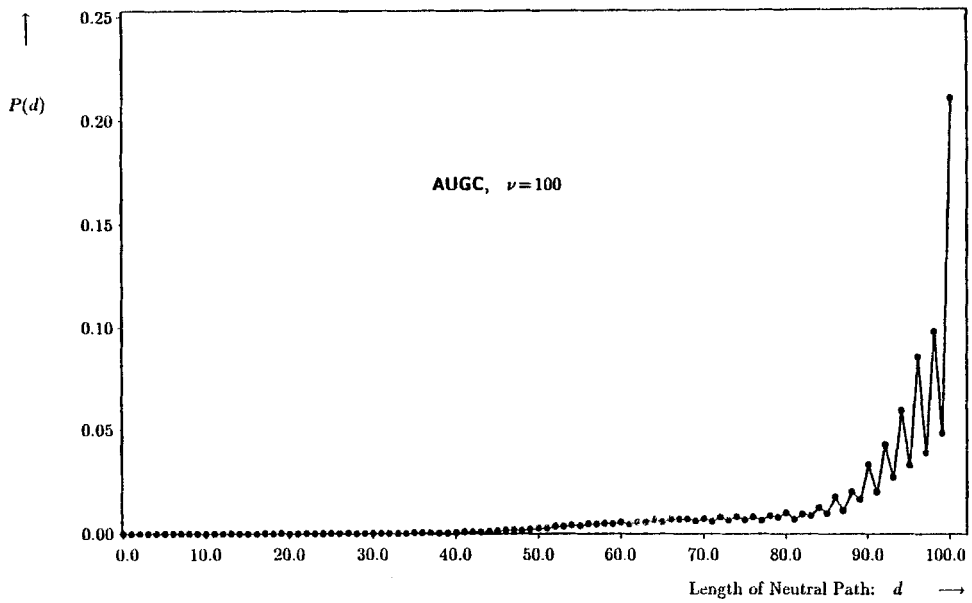

Fig. 11. Length distribution of neutral paths starting from random AUGC-sequences of chain length $\nu = 100$. A neutral path connects pairs of sequences with identical structures and Hamming distance $d_h = 1$ (single base exchange) or $d_h = 2$ (base pair exchange). The Hamming distance to the reference sequence is monotonously increasing along the path.

(1) sequences folding into one and the same structure are distributed randomly in sequence space,

(2) the frequency distribution of structures is sharply peaked (there are comparatively few common structures and many rare ones),

(3) sequences folding into all common structures are found within (relatively) small neighbourhoods of any random sequence, and

(4) the shape space contains extended 'neutral networks' joining sequences with identical structures (a large fraction of neutral path leads from the initial sequence through the entire sequence space to a final sequence on opposite side – there are $(\kappa - 1)^\nu$ sequences which differ in all positions from an initial sequence).

Combining the two statements (1) and (3) we may visualize the mapping from sequences into structures as illustrated by the sketch shown in Figure 12.

These results suggest straightforward strategies in the search for new RNA structures. It provides little advantage to start from natural or other preselected sequences since any random sequence would do equally well as the starting molecule for the selection cycles shown in Figures 3 and 4. Any common secondary structure with optimal functions is accessible in a few selection cycles. The secondary structure of RNA is understood as a crude first-order approximation to the actual spatial structure. Fine tuning of properties by choosing from a variety of molecules sharing the same secondary structure will often be necessary. In order to achive this goal it is of advantage to adopt alternations of selection cycles with low and high error

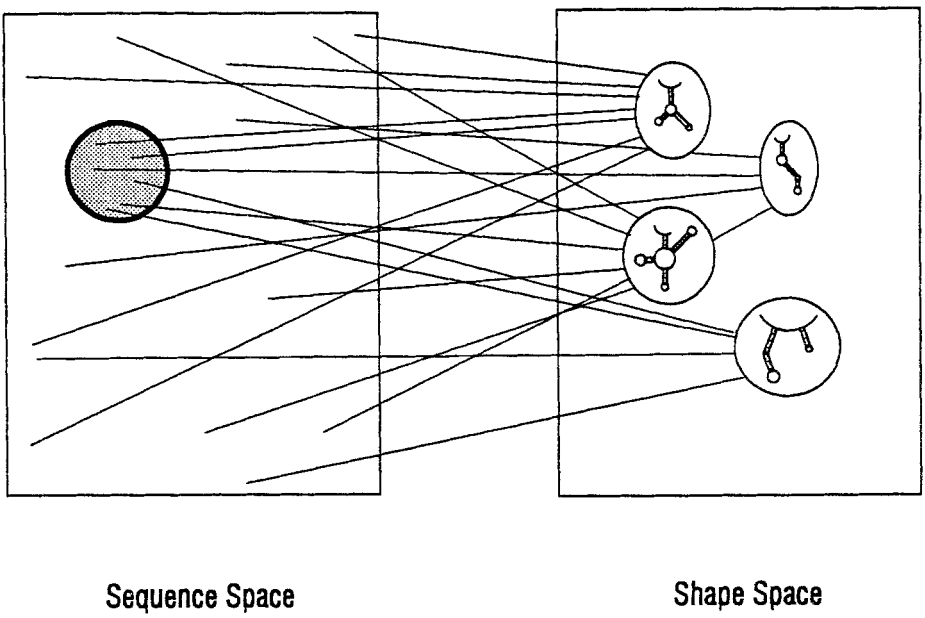

Fig. 12. A sketch of the mapping from sequences into RNA secondary structures as derived here. Any random sequence is surrounded by a ball in sequence space which contains sequences folding into (almost) all common structures. The radius of this ball is much smaller than the dimension of sequence space.

rates. At low error rates the population performs a search in the vicinity of the current master sequence (the most common sequence which is usually also the fittest sequence). If no RNA molecule with satisfactory properties is found, a change to high error rate is adequate. Then the population spreads along the neutral network to other regions in sequence space which can be explored in detail after tuning the error rate low again.

The structure of shape space is highly relevant for evolutionary optimization in nature too. Since long neutral paths are common, populations drift readily through sequence space whenever selection constraints are absent. This is precisely what is predicted by the 'neutral theory of evolution' (Kimura, 1983), and what is observed in molecular phylogeny by sequence comparisons of different organisms. The structure of shape space provides also a firm answer to the old probability argument against the possibility of succesful adaptive evolution (Wigner, 1961). How should nature find a given biopolymer by trial and error when the chance to guess it is as low as $1{:}\kappa^{\nu}$? Previously given answers (Eigen, 1971; Eigen and Schuster, 1979) can be supported and extended by precise data on the RNA shape space. The numbers of sequences that have to be searched in order to find adequate solution are many orders of magnitude smaller than those guessed on naive statistical grounds. If one of the common structures has a property which increases fitness it can hardly be missed in an evolutionary search.

Acknowledgments

The statistical analysis of the RNA shape space presented here is joint work with Drs. Walter Fontana and Peter F. Stadler and will be described elsewhere in detail. Financial support by the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung* (Projects S 5305-PHY and P 8526-MOB) and by the Commission of the European Communities (Contract PSS*0396) is gratefully acknowledged.

References

Bauer, G. J., McCaskill, J. S. and Otten, H., 1989a: *Proc. Natl. Acad. Sci. USA*, **86**, 7937.
Bauer, G. J., McCaskill, J. S. and Schwienhorst, A., 1989b: *Nachr. Chem. Tech. Lab.*, **37**, 484.
Beaudry, A. A. and Joyce, G. F., 1992: *Science*, **257**, 635.
Biebricher, C. K., Eigen, M. and Gardiner, W. C., jr., 1983: *Biochemistry*, **22**, 2544.
Biebricher, C. K., Eigen, M. and Gardiner, W. C., jr., 1984: *Biochemistry*, **23**, 3186.
Biebricher, C. K., Eigen, M. and Gardiner, W. C., jr., 1985: *Biochemistry*, **24**, 6550.
Cech, T. R., 1986: *Sci. Am.*, **255/5**, 76.
Domingo, E., 1990: *Futura*, **3/90**, 6.
Eigen, M., 1971: *Naturwissenschaften*, **58**, 465.
Eigen, M., and Gardiner, W., 1984: *Pure & Appl. Chem.*, **56**, 967.
Eigen, M., McCaskill, J. and Schuster, P., 1988: *J. Phys. Chem.*, **92**, 6881.
Eigen, M., McCaskill, J. and Schuster, P., 1989: *Adv. Chem. Phys.*, **75**, 149.
Eigen, M. and Schuster, P., 1977: *Naturwissenschaften*, **64**, 541.
Eigen, M. and Schuster, P., 1979: *The hypercycle – A principle of natural self-organization*, Springer-Verlag: Berlin.
Eigen, M. and Schuster, P., 1985: *J. Mol. Evol.*, **19**, 47.
Ellington, A. D. and Szostak, J. W., 1990: *Nature*, **356**, 818.
Fontata, W. and Schusterm, P., 1987: *Biophys. Chem.*, **26**, 123.
Fontana, W., Schnabl, W., and Schuster, P., 1989: *Phys. Rev. A*, **40**, 3301.
Fontana, W., Griesmacher, T., Schnabl, W., Stadler, P. F. and Schuster, P., 1991: *Mh. Vhem.*, **122**, 795.
Fontana, W., Stadler, P. F., Bornberg-Bauer, E. G., Griesmacher, T., Hofacker, I. L., Tacker, M., Tarazona, P., Weinberger, E. D. and Schuster, P., 1993a: *Phys. Rev. E*, **47**, 2083.
Fontana, W., Konings, D. A. M., Stadler, P. F. and Schuster, P., 1993b: *Biopolymers*, **33**, 1389.
Gilbert, W., 1986: *Nature*, **319**, 618.
Gurrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. and Altmen, S., 1983: *Cell*, **35**, 849.
Guerrier-Takada, C. and Altman, S., 1984: *Science*, **233**, 285.
Horwitz, M. S. Z., Dube, D. K. and Loeb, L. A., 1989: *Genome*, **31**, 112.
Husimi, Y., Nishgaki, K., Kinoshita, Y. and Tanaka, T., 1982: *Rev. Sci. Instrum.*, **53**, 517.
Husimi, Y. and Keweloh, H.-C., 1987: *Rev. Sci. Instrum.*, **58**, 1109.
Joyce, G. F., 1989: *Nature*, **338**, 217.
Joyce, G. F., 1991: *The New Biologist*, **3**, 399.
Kauffman, S. A., 1986: *J. theor. Biol.*, **119**, 1.
Kauffman, S. A., 1992: *J. theor. Biol.*, **157**, 1.
Kimura, M., 1983: *The neutral theory of molecular evolution*, Cambridge University Press: Cambridge (U.K.).
Mason, S. F., 1991: *Chemical evolution. Origin of the elements, molecules and living systems*, Clarendon Press: Oxford (U.K.).
Noller, H. F., 1991: *Ann. Biochem.*, **60**, 1991.
Noller, H. F., Hoffarth, V. and Zimniak, L., 1992: *Science*, **256**, 1146.
Orgel, L. E., 1986: *J. theor. Biol.*, **123**, 127-149.
Orgel, L. E., 1992: *Nature*, **358**, 203.

Piccirilli, J. A., McConnell, T. S., Zaug, A. L., Noller, H. F. and Cech, T. R., 1992: *Science,* **256**, 1420.
Riesner, D., and Gross, H. J., 1985: *Ann. Rev. Biochem.*, **54**, 531.
Schopf, J. W., and Packer, B. M., 1992: *Science*, **237**, 70.
Schuster, P., 1981: *Prebiotic Evolution*, in: Gutfreund, H., ed. Biochemical Evolution, pp. 15–87. Cambridge Univ. Press: Cambridge (U.K.).
Schuster, P., Fontana, W., Stadler, P. F., and Hofacker, I. L., 1993: *How to search for RNA structures* Preprint.
Spiegelman, S., 1971: *Quart. Rev. Biophys.*, **4**, 213.
Symons, R. H., 1992: *Ann. Rev. Biochem.*, **61**, 641.
Strunk, G., 1992: Doctoral Thesis, Universität Braunschweig.
Swetina, J., and Schuster, P., 1982: *Biophys. Chem.*, **16**, 329.
Szathmáry, E., 1991: *Proc. Roy. Soc. London B*, **245**, 91.
Szathmáry, E., 1992: *Proc. Natl. Acad. Sci. USA*, **899**, 2614.
Tuerk, C. and Gold, L., 1990: *Science*, **249**, 505.
Wakeman, J. A., and Maden, B. E. H., 1989: *Biochem. J.*, **258**, 49.
Wang, Q., 1987: *Biol. Cybern.*, **57**, 95.
Wiger, E., 1961: *The logic of personal knowledge*, in: Shills, E., ed. Essays presented to Michael Polanyi on his seventieth birthday 11th March 1961. Free Press: Glencoe (ILll.).