From Sequences to Shapes and back: A Case Study in RNA Secondary Structures

By

PETER SCHUSTER[a,b,c,*], WALTER FONTANA[c], PETER F. STADLER[b,c],

AND IVO L.HOFACKER[b]

[a] Institut für Molekulare Biotechnologie, Jena, Germany
[b] Institut für Theoretische Chemie, Universität Wien, Austria
[c] Santa Fe Institute, Santa Fe, USA

[*] Correspondence to:

Peter Schuster, Institut für Molekulare Biotechnologie
Beutenbergstraße 11, PF 100 813, D-07708 Jena, Germany

Phone: **49 (3641) 65 6333

Fax: **49 (3641) 65 6335

E-Mail: pks@imb-jena.de

KEYWORDS

Evolutionary optimization – inverse folding – RNA secondary structures – sequence space – shape space

SUMMARY

RNA folding is viewed here as a map assigning secondary structures to sequences. At fixed chain length the number of sequences exceeds by far the number of structures. Frequencies of structures are highly non-uniform and follow a generalized form of Zipf's law: we find relatively few common and many rare ones. Using an algorithm for inverse folding we show that sequences sharing the same structure are distributed randomly over sequence sapce. All common structures can be accessed from an arbitrary sequence by a number of mutations much smaller than the chain length. The sequence space is percolated by extensive neutral networks connecting nearest neighbours folding into identical structures. Implications for evolutionary adaptation and for applied molecular evolution are evident: finding a particular structure by mutation and selection is much simpler than expected, and even if catalytic activity should turn out to be sparse in the space of RNA structures, it can hardly be missed by evolutionary processes.

1. INTRODUCTION

Folding sequences into structures is a central problem in biopolymer research. Both robustness and accessibility of structures, as functions of mutational change in the underlying sequence, are crucial to both natural and applied molecular evolution. Test-tube evolution experiments are based on properties of RNA molecules: as sequences they are genotypes, and as spatial structures they are phenotypes (Spiegelman 1971; Biebricher 1983). Our concern is the mapping from RNA sequences into structures being the simplest, and the only tractable, example of a genotype-phenotype mapping.

An RNA sequence is a point in the space of all 4^n sequences with fixed length n. This space has a natural metric induced by point mutations interconverting sequences known as the Hamming distance (Hamming 1950 and 1986). The folding process considered here maps an RNA sequence into a secondary structure (figure 1A) minimizing free energy. A secondary structure is tantamount to a list of Watson-Crick type and **GU** base pairs and can be represented as a tree graph (figure 1B). This emphasizes the combinatorial nature of secondary structures and allows for a canonical distance measure between structures (Tai 1979). Assuming elementary edit operations with predefined costs, such as deletion, insertion and relabeling of nodes, the distance between two trees is given by the smallest sum of the edit costs along any path that converts one tree into the other (Sankoff & Kruskal 1983).

A rough upper bound on the number of minimum free energy structures (of fixed chain length n) can be obtained along the lines devised by Stein and Waterman (1978). Counting only those planar secondary structures that contain hairpin loops of size three or more (steric constraint), and that contain no isolated base pairs (stacks of two or more pairs are essentially the only stabilizing elements) one finds:

$$S_n = 1.4848 \times n^{-3/2}(1.8488)^n.$$

which is consistently smaller than the number of sequences.

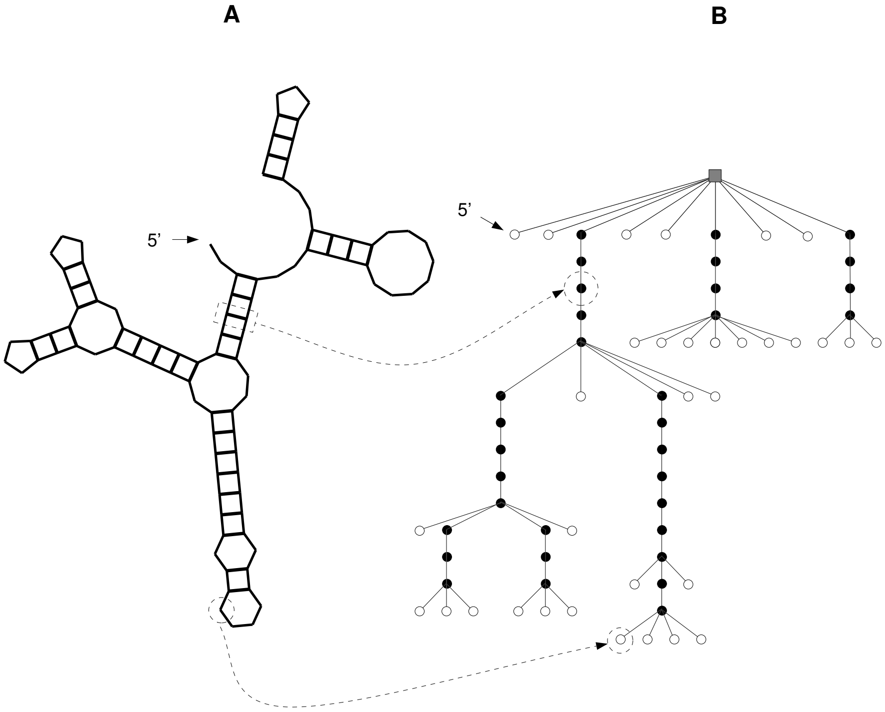

Fig. 1: **A**) A secondary structure on a sequence is any pattern of base pairs such that no bases inside a loop pair to bases outside it. Such a structure can be uniquely decomposed into structural elements that are (i) base pair stacks, (ii) loops differing in size (number of unpaired bases) and branching degree: hairpin loops (degree one), internal loops (degree two or more), and (iii) bases which are not part of a stack or a loop are termed external (freely rotating joints and unpaired ends). Each stack or loop element contributes additively to the overall free energy of the structure. These energy terms are empirically determined parameters that depend on the nucleotide sequence. A minimum free energy structure is constructed according to an algorithm proposed by Zuker and coworkers (Zuker and Stiegler 1984, Zuker and Sankoff 1981).
B) A secondary structure graph (**A**) is equivalent to an ordered rooted tree. An internal node (black) of the tree corresponds to a base pair (two nucleotides), a leaf node (white) corresponds to one unpaired nucleotide, and the root node (black square) is a virtual parent to the external elements. Contiguous base pair stacks translate into "ropes" of internal nodes and loops appear as bushes of leaves. Recursively traversing a tree by first visiting the root, then visiting its subtrees in left to right order, finally visiting the root again, assigns numbers to the nodes in correspondence to the 5'-3' positions along the sequence (Internal nodes are assigned two numbers reflecting the paired positions).

Folding can thus be viewed as a map between two metric spaces of combinatorial complexity, a sequence space and a shape space (The notion of shape space was originally used in theoretical immunology in a similar context by Perelson & Oster 1979). "Shape" refers to a discretized (and hence coarse grained) structure representation, such as the secondary structures or the tree graphs used here. The notion of secondary structure is but one among a spectrum of possible levels of resolution that can be used to define shape. It discards atomic coordinates, as well as the relative spatial orientation of the structural elements, taking into account only their number, size and relative connectedness. Nevertheless, secondary structure is a major component of whatever turns out to be an adequate shape definition for RNA: it covers the dominant part of the three-dimensional folding energies, very often it can be used successfully in the interpretation of function and reactivity, and it is frequently conserved in evolution (Sankoff *et al.* 1978, Konings & Hogeweg 1989, Le & Zuker 1990), sometimes together with a few tertiary interactions (Cech 1988). This suggests that many of the relevant intermolecular interactions that collectively set a natural scale for shape are indeed strongly influenced by the secondary structure. The observation, then, is that - at least in the present case - the shape space is considerably smaller than the sequence space (We remark that this is also true for protein models on lattices).

2. FREQUENCIES OF SHAPES AND INVERSE FOLDING

Frequencies of occurrence for individual shapes in sequence space were obtained from large samples derived by folding random sequences of fixed chain length. Ranking according to decreasing frequencies yields a distribution which obeys a generalized Zipf's law (figure 2). We are thus dealing with relatively few common shapes and many rare ones. How are the sequences which fold into the same shape distributed in sequence space? This distribution is evaluated with a heuristic inverse folding procedure, aimed at devising sequences that fold into an arbitrary predefined target shape (Hofacker *et al.* 1993). The obvious first step is to construct a compatible test sequence with nucleotide assignments such that the target shape is indeed a possible secondary structure, although typically not a minimum free energy one. We choose at random among the many compatible sequences. The next step is to decompose the minimum energy structure on the chosen test sequence into substructures, and to mutate by trial and error the corresponding subsequences. When the individual substructures are as in the target, the entire sequence is reassembled. The procedure stops if the reassembled sequence folds into the target shape. This happens in about 50% of the cases. Several sequences that fold into the same structure are sampled by starting the procedure with different compatible sequences. The average number of mutations that converted a random compatible sequence of chain length $n = 100$ into one with the desired target shape was 7.2.

The resulting ensemble of compatible sequences that fold into a predefined target has been analyzed for the target being the secondary structure of t-RNA$^{\mathrm{Phe}}$ and for three randomly constructed examples. In each case some thousand sequences were derived by the inverse folding algorithm. The distribution of pairwise distances is not distinguishable from the one expected for random compatible sequences. The properties of the sequence sample as seen by statistical geometry (Eigen *et al.* 1988a) and split-decomposition (Bandelt & Dress 1992) yield the same result: sequences folding into the same structure are randomly distributed in the space of compatible sequences.

3. STRUCTURE DENSITY SURFACES

Generalizing the previous question we ask how the possible shapes are distributed over the possible sequences. One insight is provided by considering the probability density (Fontana *et al.* 1993a,b) $P(t|h)$ of two structures being at (tree) distance t given that the underlying sequences are at (Hamming) distance h. This structure

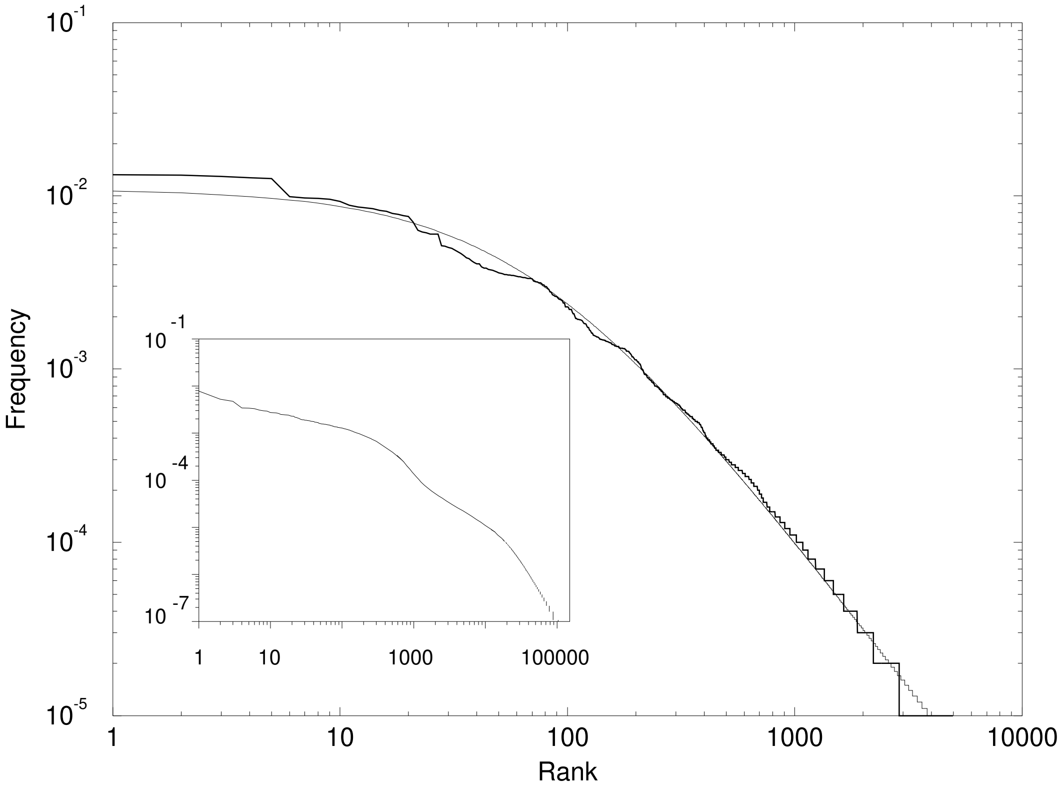

Fig. 2: The frequency distribution of RNA secondary structures. Shapes are ranked by their frequencies. The particular example shown here deals with the loop structures ([1]iagram covers 97 % of the total frequency. The frequencies follow a generalized form of Zipf's law:
$f(x) = a(b+x)^{-c}$ with x being the rank of a shape and $f(x)$ its frequency. Parameter values of the best fit (thin curve) are $a = 1.25$, $b = 71.2$ and $c = 1.73$. The frequency distribution of full secondary structures is essentially the same as shown in the insert for chain length 30. Computation of the distribution for longer chains is hardly possible since the number of structures exceeds by far the available capacities (there are about 7×10^{23} full secondary structures of chain length 100).

density surface (SDS) shows how the distribution of structure differences changes as the sequences become more and more uncorrelated with incerasing Hamming distance from the reference (figure 3 presents the SDS for sequences of chain length $n=100$). Three observations are immediate: (1) while for very small Hamming distances ($h = 1, 2, 3$) the most probable structures are identical or very similar, there is nonetheless some probability that even a single mutation substantially alters the structure; (2) beyond distance $h = 3$ identical or even closely related structures are extremely unlikely; (3) in the range $15 < h < 20$ the density becomes independent of h, thus approaching essentially what is expected for a sample of randomly drawn sequences ($h \approx 75$).

The latter indicates that the structures of a reference sequence and of its mutants at distances between 15 and 20 or larger are effectively uncorrelated. This suggests that memory of the reference structure is sufficiently lost to allow the mutants at that distance to acquire any frequent minimum energy structure - at least in its essential features. From the SDS the complete structure autocorrelation function can be recovered (Fontana *et al.* 1993a). This function is to a reasonable approximation a single decaying exponential with a characteristic length, $\ell = 7.6$ in the present case (chain length $n = 100$). From figure 3 it is seen that this corresponds essentially to the distance at which the dominant peak resulting from identical or very similar structures has disappeared.

4. SHAPE SPACE COVERING

Combination of the previous results (showing the existence of relatively few common shapes which are minimum free energy structures for sequences randomly distributed in sequence space) with the information from the SDS (showing the existence of a transition from local to global features) provides strong evidence for the existence of a neighbourhood (a high-dimensional ball) around every random sequence that contains sequences whose structures include almost all common shapes.

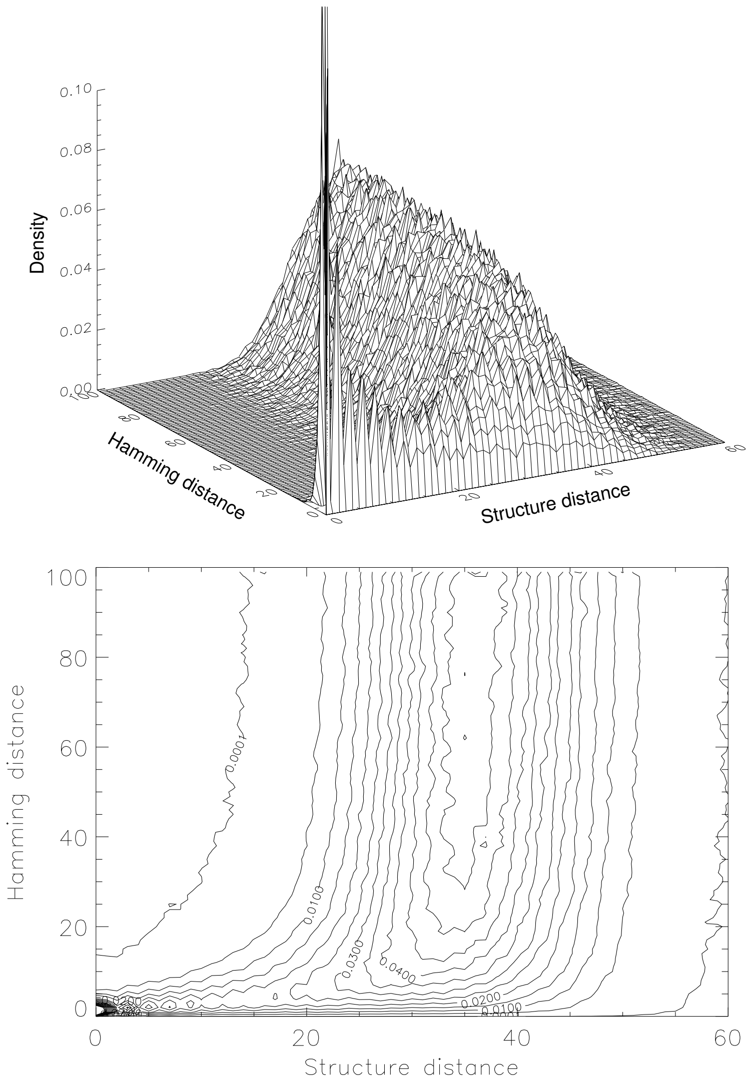

Fig. 3: The structure density surface (SDS) for RNA sequences of length $n = 100$ (upper). This surface was obtained as follows:
(i) choose a random reference sequence and compute its structure, (ii) sample randomly 10 different sequences in each distance class (Hamming distance 1 to 100) from the reference sequence, and bin the distances between their structures and the reference structure. This procedure was repeated for 1000 random reference sequences. Convergence is remarkably fast; no substantial changes were observed when doubling the number of reference sequences. This procedure conditions the density surface to sequences with base composition peaked at uniformity, and does, therefore, not yield information about strongly biased compositions.
Lower part: Contour plot of the SDS.

In order to verify the prediction of a characteristic neighbourhood covering almost all common shapes we performed a computer experiment. A target sequence is chosen at random. A second random sequence serves as initial trial sequence, and its structure as a reference structure. Next we search for a nearest neighbour of the trial sequence that folds into the reference structure but lies closer to the target. If such a sequence is found, it is accepted as the new trial sequence, and the procedure is repeated until no further approach to the target is possible. The final Hamming distance to the target is an upper bound for the minimum distance between two sequences folding into the reference structure and the structure of the target, respectively. The probability density of this upper bound to the closest approach distances determined for RNA molecules of chain length 100 is shown in figure 4 (open circles). It yields a mean value of 19.8 (It is remarkable that this value conincides with the critical Hamming distance where we observe the change from local to global features in the SDS; analogous investigations on RNA molecules with only **GC** or only **AU** base pairs have shown that the precise agreement is not generally valid).

We can also compute a lower bound for the mean value of the closest approach distance. The probability that two arbitrarily chosen bases of an RNA sequence can form a base pair is given by the number of pairings divided by the number of possible combinations of two bases: $6/16 = 3/8$ (since we have six classes of base pairs: **AU**, **GC**, **GU** and inversions). The mean number of bases that have to be changed in a random sequence, in order to form a sequence which is compatible with the target structure, (representing the lower bound) is obtained from the probability not to form a base pair by multiplication with the mean number of base pairs: $(1 - \frac{3}{8}) \times \overline{n}_{BP} = \frac{5}{8}\,\overline{n}_{BP}$. For RNA molecules of chain length 100 the mean number of base pairs is 24.34 (Fontana *et al.* 1993a), and we obtain a mean Hamming distance of 15.2 for the lower bound. From the probability density of the number of base pairs we derive a distribution of the lower bound also shown in figure 4 (open diamonds). The characteristic neighbourhood has a radius of $15 < h_c < 20$.

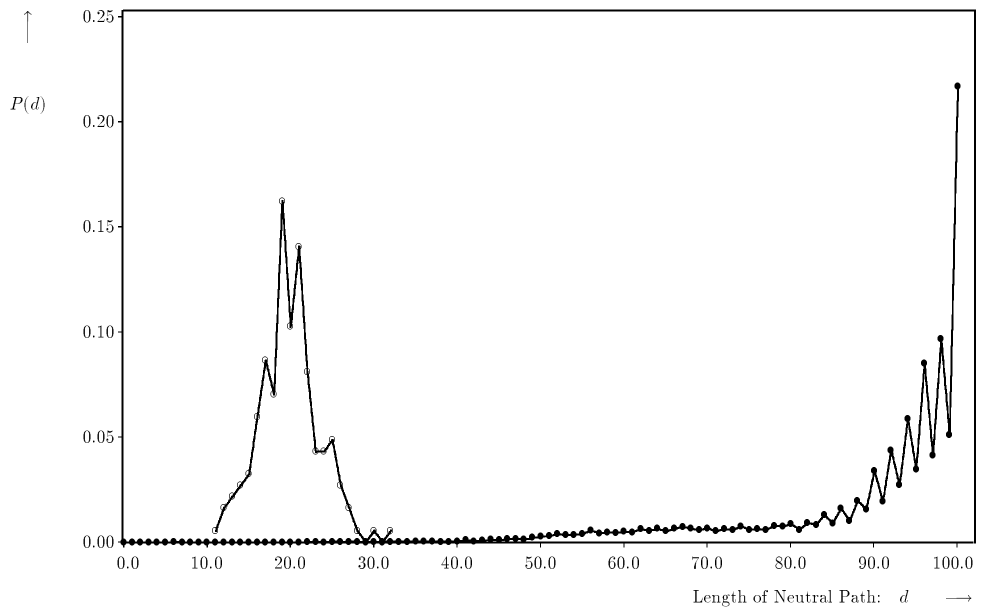

Fig. 4: Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1 which are obtained by single base exchanges in unpaired stretches of the structure and neighbours of Hamming distance 2 resulting from base pair exchanges in stacks. Two probability densities of Hamming distances are shown that were obtained by searching for neutral paths in sequence space: (i) shortest distances between trial and target sequences (○) obtained as endpoints of neutral paths approaching the target from a random trial sequence (185 targets and 100 trials for each were used), and (ii) longest distances between the reference and the end points of monotonously diverging neutral paths (•, 500 reference sequences were used).

5. NEUTRAL PATHS THROUGH SEQUENCE SPACE

The structure of the RNA shape space over the sequence space is complemented by a second computer experiment. We search for neutral paths with monotonously increasing distance from a reference sequence. A neutral path ends when no sequence that forms the same structure is found among the nearest neighbours. The probability density of the lengths of these paths is shown in figure 4 (filled circles). The vast extension of the network of neutral paths came as a surprise: 21.7 % of all paths percolate the entire sequence space and end in a sequence which has not a single base in common with the reference (The existence of extensive neutral networks meets a claim raised by Maynard-Smith 1970 for protein spaces that are suitable for efficient evolution).

6. DISCUSSION

The existence of a ball with characteristic radius around any random sequence within which almost all common shapes are found (figure 5) is a robust phenomenon of the mapping from sequences into RNA secondary structures. It depends on the ratio of sequences to structures. Changes in the base pairing alphabet, in particular the consideration of pure **GC**, or pure **AU** sequences, may cause minor alterations that can be interpreted by much smaller values of this ratio as well as by differences in the topology of sequence space. Alphabet dependencies will be published elsewhere. The major features of the shape space structure depend on the generic properties of RNA folding, in particular on the non-local nature of base pairing, but they are insensitive to the empirical energy parameters sets used in folding algorithms, as well as to the distance measure between structures used in the SDS (essentially the same SDSs were obtained with a now superseded parameter set (Fontana *et al.* 1991), and also with a different structure distance measure (Hogeweg & Hesper 1984, Huynen *et al.* 1993).

There are caveats to our approach: (1) We use a thermodynamic criterion for RNA structure formation, not one that mimics the kinetics of folding. This does not constitute a problem for short sequences up to some few hundred nucleotides. Moreover the number of possible kinetic structures is not entirely different from the number of thermodynamic structures, the principles of base pairing are the same and thus, the generic features of mappings of sequences into kinetic or thermodynamic structures will be essentially the same, too.

(2) We consider only a single minimum free energy structure for each sequence. Our approach can be carried over to ensembles comprising optimal and suboptimal

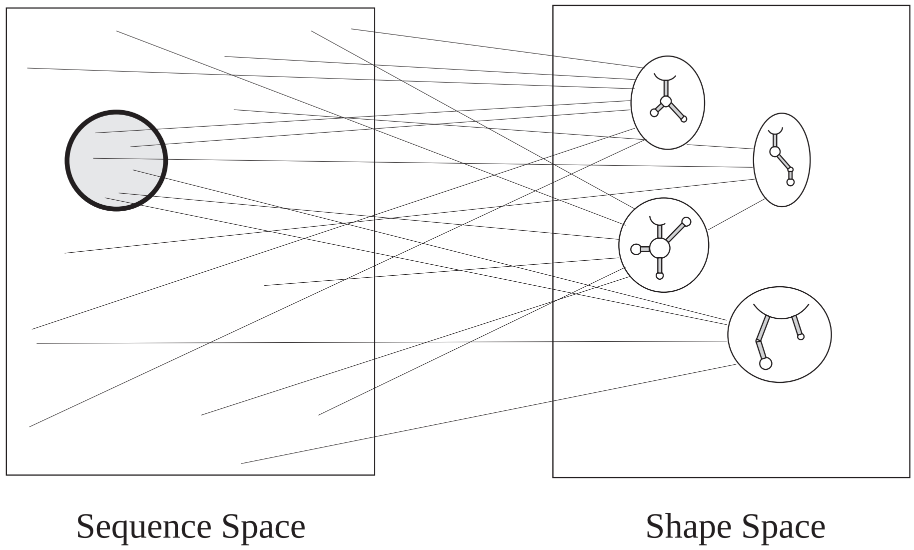

Fig. 5: A sketch of the mapping from sequences into RNA secondary structures as derived here. Any random sequence is surrounded by a ball in sequence space which contains sequences folding into (almost) all common structures. The radius of this ball is much smaller than the dimension of sequence space.

foldings, as represented by partition functions (McCaskill 1990, Bonhoeffer *et al.* 1993). "Shape", then, becomes a matrix of temperature dependent base pairing probabilities, and the concept of distance is changed accordingly. All qualitative features of the SDS remain essentially unchanged and numerical corrections are in the range of 10% (as for example in the case of correlation lengths: Bonhoeffer *et al.* 1993).

(3) We do not consider three-dimensional structure. Nevertheless, the secondary structure defines an informative scale of resolution. In addition, it constitutes an approximation to a coarse grained spatial structure (current algorithms for the modeling of RNA 3D-structures start from secondary structures, and introduce a few tertiary interactions: Major *et al.* 1991 and 1993).

The consequences of our results for natural and artificial selection are immediate. We predict that there is no need to systematically search huge portions of the sequence space. In the particular example of RNA molecules of chain length 100 the characteristic ball contains some 10^{27} sequences, which is only a fraction of 10^{-33} of the entire sequence space. Almost all structures are within reach of a few mutations from a compatible sequence (average: 7.2), and even in reasonable proximity of any non-compatible random sequence (≈ 18). The conclusion, thus, is that optimization of structures by evolutionary trial and error strategies is much simpler than often assumed. It provides further support to the idea of widespread applicability of molecular evolution (Eigen 1971, Eigen & Schuster 1979, Eigen *et al.* 1988b and 1989). The existence of networks of neutral paths percolating the entire sequence space has strong implications for (molecular) evolution in nature as well as in the laboratory. Populations replicating with sufficiently high error rates will readily spread along these networks and can reach more distant regions in sequence space.

If one were to design the ultimate evolvable molecule that carries information and is engaged in functional interactions, one would ideally require two features: (1) capability of drifting across sequence space without the necessity of changing shape, (2) proximity to any common shape everywhere. These are precisely the features that statistically characterize the mapping from RNA sequence to secondary structure.

The work reported here was supported financially by the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung* (Projects S 5305-PHY, P 9942-PHY, and P 8526-MOB), by the European Economic Community (Research Study Contract PSS*03960) and by an National Science Foundation (Washington) core grant to the Santa Fe Institute (PHY-9021437). We thank Professors Leo Buss and Andreas Dress for comments on the manuscript, and Dipl.-Chem. Erich Bornberg-Bauer for supplying the data for the insert of figure 2.

REFERNECES

Bandelt, H.-J. & Dress, A.W.M. 1992 A canonical decomposition theory for metrics on a finite set. *Adv.Math.* **92**, 47-105.

Biebricher, C.K. 1983 Darwinian selection of self-replicating RNA molecules. *Evolutionary Biology* **16**, 1-52.

Bonhoeffer, S., McCaskill, J.S., Stadler, P.F. & Schuster, P. 1993 RNA multi-structure landscapes. A study based on temparature dependent partition functions. *Eur.Biophys.J.* **22**, 13-24.

Cech, T.R. 1988 Conserved sequences and structures of group I introns: Building an active site for RNA catalysis – a review. *Gene* **73**, 259-271.

Eigen, M. 1971 Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465-523.

Eigen, M. & Schuster, P. 1979 slThe hypercycle - a principle of natural self-organization. Springer-Verlag: Berlin.

Eigen, M., Winkler-Oswatitsch, R., & Dress, A. 1988a Statistical geometry in sequence space: A method of quantitative comparative sequence analysis. *Proc.Natl.Acad.Sci.U.S.A.* **85**, 5913-5917.

Eigen, M., McCaskill, J. & Schuster, P. 1988b Molecular quasi-species. *J.Phys.Chem.* **92**, 6881-6891.

Eigen, M., McCaskill, J. & Schuster, P. 1989 The molecular quasi-species. *Adv.Chem.Phys.* **75**, 149-263.

Fontana, W., Griesmacher, T., Schnabl, W., Stadler, P.F. & Schuster, P. 1991 Statistics of landscapes based on frre energies, replication and degradation rate constants of RNA secondary structures. *Mh.Chem.* **122**, 795-819.

Fontana, W., Konings, D.A.M., Stadler, P.F. & Schuster, P. 1993a Statistics of RNA secondary structures. *Biopolymers* **33**, 1389-1404.

Fontana, W., Stadler, P.F., Bornberg-Bauer, E.G., Griesmacher, T., Hofacker, I.L., Tacker, M., Tarazona, P., Weinberger, E.D. & Schuster, P. 1993b RNA folding and combinatory landscapes. *Phys.Rev.E* **47**, 2083-2099.

Hamming, R.W. 1950 Error detecting and error correcting codes. *Bell Syst.Tech.J.* **29**, 147-160.

Hamming, R.W. 1986 slCoding and information theory. 2nd Ed., Prentice Hall: Englewood Cliffs.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M. & Schuster, P. 1993 slFast folding and comparison of RNA secondary structures. Santa Fe Institute Preprint No.93-07-044; Mh.Chem., in press.

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

Huynen, M.A., Konings, D.A.M. & Hogeweg, P. 1993 slMultiple coding and the evolutionary properties of RNA secondary structure. *J.Theor.Biol.*, in press.

Konings, D.A.M. & Hogeweg, P. 1989 Pattern analysis of RNA secondary structure. Similarity and consensus of minimal-energy folding. *J.Mol.Biol.* **207**, 597-614.

Le, S.-Y. & Zuker, M. 1990 Common structures of the 5' non-coding RNA in enteroviruses and rhinoviruses. Thermodynamical stability and statistical significance. *J.Mol.Biol.* **216**, 729-741.

Major, F., Turcotte, M., Gautheret, D., Lapalme, G., Fillion, E. & Cedergren, R. 1991 The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science* **253**, 1255-1260.

Major, F., Gautheret, D. & Cedergren, R. 1993 Reproducing the three-dimensional structure of a tRNA molecule from structural constraints. *Proc.Natl.Acad.Sci.USA* **90**, 9408-9412.

Maynard-Smith, J. 1970 Natural selection and the concept of a protein space. *Nature* **225**, 563-564.

McCaskill, J.S. 1990 The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers* **29**, 1105-1119.

Perelson, A.S. & Oster, G.F. 1979 Theoretical studies on clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J.Theor.Biol.* **81**, 645-670.).

Sankoff, D., Morin, A.-M. & Cedergren,R.J. 1978 The evolution of 5S RNA secondary structures. *Can.J.Biochem.* **56**, 440-443.

Sankoff, D. & Kruskal, J.B. Eds. 1983 slTime warps, string edits and macromolecules: the theory and practice of sequence comparisons. Addison Wesley, London.

Shapiro, B.A. & Zhang, K. 1990 Comparing multiple RNA secondary structures using tree comparisons. *CABIOS* **6**, 309-318.

Spiegelman, S. 1971 An approach to the experimental analysis of precellular evolution. *Quart.Rev.Biophys.* **4**, 213-253.

Stein, P.R. & Waterman, M.S. 1978 On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Mathematics* **26**, 261-272.

Tai, K. 1979 The tree-to-tree correction problem *J.Ass.Computing Machinery* **26**, 422-433.

Zuker, M. & Sankoff, D. 1984 RNA secondary structures and their prediction. *Bull.Math.Biol.* **46**, 591-621.

Zuker, M. & Stiegler, P. 1981 Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucleic Acids Res.* **9**, 133-148.