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

Peter Schuster
Walter Fontana
Peter F. Stadler
Ivo L. Hofacker

SFI WORKING PAPER: 1993-07-045

SFI Working Papers contain accounts of scientific work of the authors and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu
From Sequences to Shapes and back:
A Case Study in RNA Secondary Structures

By

Peter Schuster*,†,‡,§, Walter Fontana‡, Peter F. Stadler†,‡,
and Ivo L. Hofacker†

* Institut für Molekulare Biotechnologie e.V., Jena, Germany
† Institut für Theoretische Chemie, Universität Wien, Austria
‡ Santa Fe Institute, Santa Fe, USA

§ Correspondence to:
Professor Peter Schuster, Institut für Molekulare Biotechnologie e.V.
Postfach 100813, D-07708 Jena, Germany
Phone: **49 (3641) 85 2200; Fax: **49 (3641) 85 2203 and 33 13 25
E-Mail: pks@imb-jena.de

Keywords
Evolutionary optimization – inverse folding – RNA secondary structures –
– sequence space – shape space
RNA folding is viewed 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: 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 space. All common structures can be accessed from an arbitrary sequence by a number of mutations much smaller than the chain length. Evolutionary optimization has to search only a small fraction of all sequences in order to find a suitable target structure and thus could hardly miss it. The sequence space is percolated by extensive neutral networks connecting nearest neighbors folding into identical structures.

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 (1). Our concern is the mapping from RNA sequences into structures being the simplest 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 (2). The folding process considered here maps an RNA sequence into a secondary structure (Fig.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 (Fig.1B). This emphasizes the combinational nature of secondary structures and allows for a canonical distance measure between structures (3). 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 (4).

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 (6). 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.

Folding is thus 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 (7).). “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 3-dimensional folding energies, very often it can be used successfully in the interpretation of function and reactivity, and it is frequently conserved in evolution (8), sometimes together with a few tertiary interactions (9). 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 (This is also true for protein models on lattices).
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 Stiegler (5).

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).
Frequencies of occurrence for individual shapes in sequence space were obtained from large samples derived by folding random sequences. Ranking according to decreasing frequencies yields a distribution obeying a generalized Zipf’s law (Fig. 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 (10). 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. 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 100 into one with the desired target shape was 7.2. The resulting ensemble of several thousand sequences that fold into a predefined target has been analyzed for the target being the secondary structure of t-RNA^Phe and for three randomly constructed examples. The distribution of pairwise distances is not distinguishable from the one expected for random compatible sequences. Statistical geometry (11) and and split-decomposition (12) also cannot distinguish the sample of sequences obtained by inverse folding from a random sample.

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 (14) $P(t|h)$ of two structures being at (tree) distance $t$ given that the underlying sequences are at (Hamming) distance $h$. This structure density surface (SDS) shows how the distribution of structure differences changes as the sequences become uncorrelated (Fig. 3 presents the SDS for chain length 100).
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 (13) of $10^5$ RNA molecules of chain length 100 which are derived from secondary structures by further coarse graining that eliminates all details concerning stack lengths and loop sizes. The diagram 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 $7 \times 10^{23}$ full secondary structures of chain length 100).
Three observations are immediate: (i) 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; (ii) beyond distance $h = 3$ identical or even closely related structures are extremely unlikely; (iii) 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 distance 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 (14). This function is to a reasonable approximation a single decaying exponential with a characteristic length, $\ell = 7.6$ in the present case. From Fig.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.

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 that the transition from local to global features occurs around some characteristic Hamming distance $h_c \approx 2.5 \times \ell$) 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. In case of RNA molecules of chain length 100 this ball appears to have a radius of $15 < h_c < 20$.

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
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.
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 Fig.4 (open circles). It yields a mean value of 19.8.

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 types 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: $(1 - \frac{3}{8}) \times \bar{n}_{BP} = \frac{5}{8} \bar{n}_{BP}$. For RNA molecules of chain length 100 the mean number of base pairs is 24.34 (14), and we obtain a mean Hamming distance of 15.2 for the lower bound. The characteristic neighbourhood has a radius of $15.2 < h_c < 19.8$.

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 Fig.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 (15) for protein spaces that are suitable for efficient evolution).

The existence of a ball with characteristic radius around any random sequence within which almost all common shapes are found is a robust phenomenon of the
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).

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 (16), and also with a different structure distance measure (17).).

There are caveats to our approach. (i) We use a thermodynamic criterion for RNA structure formation, not a kinetic one. This does not constitute a problem for short sequences up to some few hundred nucleotides. Moreover the number of possible kinetic structures is only slightly larger than 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. (ii) We consider only a single minimum free energy structure for each sequence. Our approach can be carried over to ensembles comprising optimal and suboptimal foldings, as represented by partition functions (18). "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 length (14,18)).

(iii) We do not consider 3-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 (19).).

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 ($\approx 18$). 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 (20). 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: (i) capability of drifting across sequence space without the necessity of changing shape, (ii) proximity to any common shape everywhere. These are precisely the features that statistically characterize the mapping from RNA sequence to secondary structure.

References and Notes


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