From sequences to shapes and back: a case study in RNA secondary structures

PETER SCHUSTER, WALTER FONTANA, PETER F. STADLER AND IVO L. HOFACKER

1 Institut für Molekulare Biotechnologie, Beutenbergstrasse 11, PF 100813, D-07708 Jena, Germany
2 Institut für Theoretische Chemie, Universität Wien, Austria
3 Santa Fe Institute, Santa Fe, U.S.A.

SUMMARY

RNA folding is viewed here as a map assigning secondary structures to sequences. At fixed chain length the number of sequences far exceeds 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. By 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. 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, 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 1d). This emphasizes the combinatorial nature of secondary structures and allows for a canonical distance measure between structures (Tai 1979). Assuming elementary edit operations with pre-defined costs, such as deletion, insertion and relabelling 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).

An approximate upper bound on the number of minimum free-energy structures (of fixed chain length n) can be obtained along the lines devised by Stein & 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:

$$S_n = 1.4848 \times n^{(1.8488)^n},$$

which is consistently smaller than the number of sequences.

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 pre-defined 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 pre-defined target has been analysed for the target being the secondary structure of t-RNA™ and for three randomly constructed examples. In each case about 1000 sequences were derived by the inverse folding algorithm. The distribution of pairwise dis-
The frequency distribution of RNA secondary structures. Shapes are ranked by their frequencies. The particular example shown here deals with the loop structures (Shapiro & Zhang 1990) of 10³ 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: \( f(x) = ax^{-b} \), 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 inset for chain length 30. Computation of the distribution for longer chains is hardly possible as the number of structures exceeds by far the available capacities (there are about \( 7 \times 10^{25} \) full secondary structures of chain length \( n = 100 \)).

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 density surface (sds) shows how the distribution of structure differences changes as the sequences become more and more uncorrelated with increasing Hamming distance from the reference (figure 3 presents the sds for sequences of chain length \( n = 100 \)). Three observations are immediate: (i) although for very small Hamming distances \( h = 1, 2, 3 \) the most probable structures are identical or very similar, there is none the less 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; and (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 suggests that the structures of a reference sequence and 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, \( l = 7.6 \) 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 in-
Figure 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) an upper bound for the closest approach of trial and target sequences (open circles) obtained as endpoints of neutral paths approaching the target from a random trial sequence (185 targets and 100 trials for each were used); (ii) a lower bound for the closest approach of trial and target sequences (open diamonds) derived from secondary structure statistics (Fontana et al. 1993a; see this paper, §4); and (iii) longest distances between the reference and the endpoints of monotonously diverging neutral paths (filled circles) (500 reference sequences were used).

formation 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.

To verify the prediction of a characteristic neighbourhood covering almost all common shapes we did a computer experiment. A target sequence is chosen at random. A second random sequence serves as an 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 coincides with the critical Hamming distance at which 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: $6/16 = 3/8$ (as 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, 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: $(1 - \frac{3}{8})^h = \frac{5^{h/2}}{8^h}$. 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 $15 < h < 20$.

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 parameter sets used in folding algorithms, as well as to the distance measure between structures used in the sds (essentially the same sds were obtained with a new superseeded 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 a 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 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 sns are 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 modelling of RNA three-dimensional structures start from secondary structures and introduce a few tertiary interactions (Major et al. 1991, 1993)).

The consequences of our results for natural and artificial selection are immediate. We predict that there is no need to search systematically 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^{-38} 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 is thus that optimization of structures by evolutionary trial and error strategies is much simpler than is often assumed. It provides further support to the idea of widespread applicability of molecular evolution (Eigen 1971; Eigen & Schuster 1979; Eigen et al. 1988a, 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, it would ideally require two features: (i) capability of drifting across sequence space without the necessity of changing shape; and (ii) 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 5390-PHY, P9942-PHY, and P8526-MOB), the European Economic Community (Research Study Contract PSS*03960) and a National Science Foundation (Washington) core grant to the Santa Fe Institute (PHY-9021437). We thank Professor Leo Buss and Professor Andreas Dress for comments on the manuscript, and Dipl.-Chem. Erich Bornberg-Bauer for supplying the data for the insert of figure 2.

REFERENCES


Fontana, W., Stadler, P. F., Bornberg-Bauer, E. G., Gries-


Received 25 November 1993; accepted 23 December 1993.