NEUTRAL NETWORKS OF INTERACTING RNA SECONDARY STRUCTURES

CAMILLE STEPHAN-OTTO ATTOLINI

Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria

camille@tbi.univie.ac.at

PETER F. STADLER

Lehrstuhl für Bioinformatik, Institut für Informatik, Universität Leipzig,
Kreuzstraße 7a, D-04103 Leipzig, Germany
Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM 87501

studla@bioinf.uni-leipzig.de

Received Apr 19 2005; accepted Jul 23 2005

RNA molecules interact by forming inter-molecular base pairs that compete with the intra-molecular base pairs of their secondary structures. Here we investigate the patterns of neutral mutations in RNAs whose function is the interaction with other RNAs, i.e., the *co-folding* with one or more other RNA molecules.
We find that (1) the degree of neutrality is much smaller in interacting RNAs compared to RNAs that just have to coform to a single externally prescribed target structure, and (2) strengthening this contraint to the conservation of the co-folded structure with two or more partners essentially eliminates neutrality. It follows that RNAs whose function depends on the formation of a specific interaction complex with a target RNA molecule will evolve much more slowly than RNAs with a function depending only on their own structure.

Keywords: RNA secondary structures, co-folding, neutral mutations

1. Motivation

The properties of the folding map for a single RNA sequence have been studied in much detail in the last decade [3, 6, 4, 8, 9, 5]. These studies showed that a large fraction of point mutations are neutral in RNA molecules in the sense that the mutation does not change the base pairing pattern (secondary structure) of the ground state structure. Due to this high degree of neutrality, there are "neutral networks" of sequences folding into the same ground state structure. These neutral networks "percolate" through sequence space and contain neutral paths that connect sequences without detectable sequence similarity. This structure of the RNA folding map implies a diffusion-like behavior of evolving populations of

RNA molecules in sequence space, which conforms to Kimura's neutral theory [13]. Indeed, expressions e.g. for fixation rates, can be derived which reproduce the prediction of the neutral theory save an additional scaling factor that depends on the fraction of neutral mutations [12].

More recently, simple models of strongly interacting RNA molecules have been studied from this perspective, in which selection for a common resource is replaced by frequency-dependent fitness terms. In these models, each RNA species depends on the presence of specific catalysts. A prime example of this class of models is the hypercycle model of interacting replicators [2]. While such a system has not (yet?) been realized experimentally, there has been substantial progress in constructing RNA replicase ribozymes. We refer to [14] for a description of the state of the art. It is thus worthwhile to study the evolutionary properties of such models.

In [19], the diffusion (in sequence space) of a population of interacting replicators is studied, where the replication rates depend only on the sequence similarity of the parent molecules. A model of hypercycles with interactions depending on the secondary structures of the individual RNAs is described in [7] and later in more detail in [20]. In the latter contribution, we emphasize the importance of the neutrality of the genotype-phenotype map for both persistence of the hypercyle and for diffusion in sequence space.

In this letter we consider a more sophisticated model of RNA-RNA interactions. In previous work, the basic assumption was that the actions of each RNA molecule are determined by its own secondary structure only. For example, the replication rate of sequence x under the influence of sequence y as catalyst is $a_{xy} = a(f(x), f(y))$, i.e, a function of the (ground state) secondary structures of both molecules. Here we explore the situation when $a_{xy} = a(f(x \circ y))$, i.e., when the rate is a function of the structure of the interaction complex of the two secondary structures x and y. To this end, we study in detail the statistical properties of the *RNA co-folding map* $f : (x, y) \mapsto f(x \circ y)$ which assigns to each pair of RNA sequences the secondary structure of their thermodynamically most stable co-folding.

2. The Model

The common secondary structure $f(x \circ y)$ of two interacting RNA molecules can be computed using a simple extension of the usual dynamic programming algorithms for computing RNA secondary structures, see e.g. [11, 1]. The basic idea is to compute the secondary structure of the concatenated RNA sequences $x{+}{+}y$ (or $y{+}{+}x$), where the "loop" that contains the split between x and y does not contribute to the folding energy. We use here the program `RNAcofold` implemented in the `Vienna RNA Package` [11, 10]. If the ground state structure is unique then $f(x \circ y) = f(y \circ x)$, otherwise the structures will in general be different since the backtracking routine implemented in `RNAcofold` yields one of the set of degenerate ground state structure in a deterministic way that depends, however, on the input order.

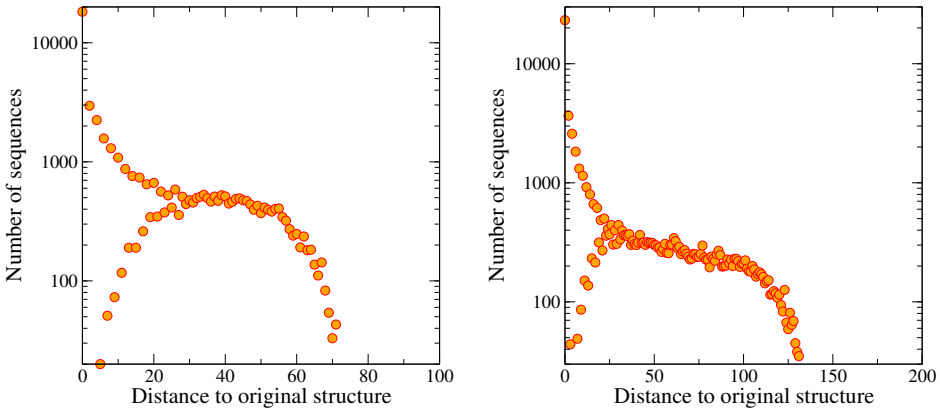

Fig. 1. Distribution of structure distances of interaction complexes before and after a single mutation mutation. (a) Point mutations. Data extracted from 600000 sequences of length 100. Fraction of neutral mutations: $\bar{\lambda} = 0.32$. (b) Compensatory mutations: 1000 sequences of length $n = 100$ cofolded with a fixed one of length also $n = 100$. On average there are 66 possible compensatory mutations, out of which $\lambda^c = 0.35$ are neutral.

In the following we will study two different versions of defining neutrality in a cofold map:

(1) We say that a mutant x' of x is neutral when $f(x' \circ y) = f(x \circ y)$ for a given partner sequence y. This scenario corresponds to RNA switches or RNAs such as microRNAs [15] that bind to target molecules in a specific way.
(2) We say that a mutant x' of x is neutral when $f(y \circ x') = f(y \circ x)$ and $f(x' \circ z) = f(x \circ z)$ for two different interaction partners y and z. This scenario corresponds e.g. to an RNA hypercycle [2]: the mutant x' simultaneously must be a template (and hence retain the structure of its complex with the catalyst z), and a catalyst (and hence be able to replicate the template y).

In the first case, i.e., cofolding of the mutating sequence with a fixed partner, we consider both point and compensatory mutations. In order to obtain accurate stastics we compute all point mutations and all compensatory mutations (where a base pair is replaced with another type of base pair) using samples of 600000 and 1000 sequences, respectively. We use the symmetric difference of the set of base pairs as a measure for the structural distance of two RNA secondary structures.

This first case is similar to folding the concatenated sequence $f(x{+}{+}y)$ instead of the co-folding complex $f(x \circ y)$, the only difference being the energy contribution from the "exterior loop" that contains the split between the two sequences. Indeed, we observe a fraction of neutral mutations $\bar{\lambda}$ similar to that reported in [8] for an individual RNA sequence.

The second case, where the RNA molecule is mutated and cofolded with two

different partners, is more imporant e.g. in the context of models of prebiotic evolution. Here a single sequence has to satisfy at least two different contraints: it has to be a recognizable template and it has to perform its catalytic function in two different contexts. In this case we sample in the following way. We randomly generate three different RNA sequences of the same length, x, y, and z, and compute $f(x \circ y)$ and $f(x \circ z)$. We then mutate x and recompute the cofolding structures $f(x' \circ y)$ and $f(x' \circ z)$ and determine the distance from the original structures. In this case, a "compensatory mutation" must be compensatory with respect to *both* $f(x \circ y)$ and $f(x \circ z)$, i.e., only base pairs shared by both cofolding structures are candidates for compensatory mutations.

We sampled approximately $300,000$ randomly generated sequences with point mutations for chain length $n = 50$, about $570,000$ of length $n = 100$, and $450,000$ of length $n = 200$. Furthermore, 3000 sequence triplets with compensatory mutations were constructed for each of the three chain lengths.

In addition to estimating the fraction of neutral mutations, we also estimated the length of neutral paths [18]. A neutral path $\mathcal{L}$ is defined a follows. Starting from a sequence x_0, a sequence of RNA sequences $\{x_i | i = 1, \ldots\}$ is constructed such that (i) $f(x_i) = f(x_0)$, i.e., the structures do not change along the path, (ii) x_i is a point mutant or compensatory mutant of x_{i-1} and (iii) the Hamming distance from the starting point x_0 strictly increases with each step. The path terminates after at most n steps when no further acceptable mutant can be found. The Hamming distance between x_0 and the last point in the path is the length L of the neutral path. Here we constructed 1200 neutral paths for sequences of length $n = 100$. In the case of one sequence cofolding with two partners, the algorithm is basically the same except that compensatory mutations must be possible in both structures and only mutations that are neutral in both interaction complexes are accepted along the path.

3. Results

The behavior of `RNAcofold` when taking into account only two sequences is very similar to that of `RNAfold` for a single RNA sequence of the same length [11]. The fraction of neutral point mutations is almost a third of the total. One difference compared to folding single sequences is that almost no point mutations change all base pairs of the structure.

In the case of compensatory mutations the situation is different since we allow mutations only in one of the two sequences. Inter-molecular base pairs can only change from **GU** to **AU** or **CG** to **UG**. Therefore, two thirds of the possible compensatory mutations are not allowed anymore and neutrality is hardly increased by them: Only 35 percent of the remaining mutations are neutral. From [17] we know that in order to change from one connected component to some other inside the neutral network, compensatory mutations may be needed. This is important from the evolutionary point of view, since a fitter structure may be accessible only from

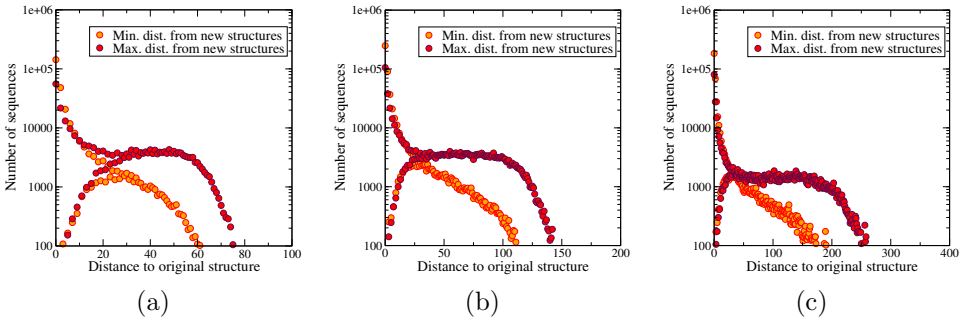

Fig. 2. Distribution of structure distances $D(f(x \circ y), f(x' \circ y))$ and $D(f(x \circ z), f(x' \circ z))$ of RNA-RNA interaction complexes. The two curves give the smaller and the larger distance of each pair of complexes. (a) 300, 000 sequences of length $n = 50$, point mutations; fraction of neutral mutations: $\bar{\lambda} = 0.185$. (b) 568, 000 sequences of length $n = 100$, point mutations, fraction of neutral mutations: $\bar{\lambda} = .186$. (c) 445, 000 sequences of length $n = 200$, point mutations, fraction of neutral mutations: $\bar{\lambda} = 0.180$.

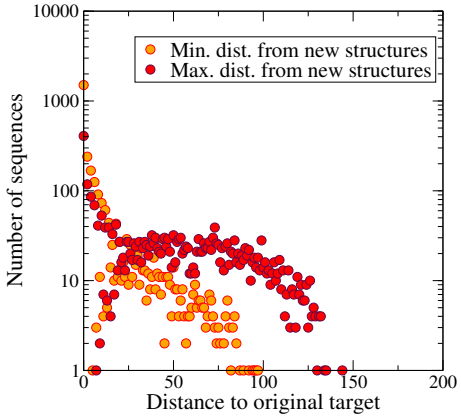

Fig. 3. Distribution of distances between interaction complexes as in Fig. 2 when compensatory mutations are used. 3000 sequences of length $n = 100$ were used. On average, we found on average only 15 possible compensatory mutations for both structures at the same time. Of these, only about 15% are neutral.

a particular connected component of the neutral network.

In the case of more than a single structural constraint, however, the situation becomes even more difficult. As shown in Fig 3b, the degree of neutrality is drastically descreased both for point mutations and for compensatory mutations. This fact is of crucial importance for models where `cofold` defines the interactions between RNA molecules.

Fig. 2 shows that neutral mutations occurring simultaneously for both cofolding structures are only about 18 percent of all possible mutations, i.e., less than two

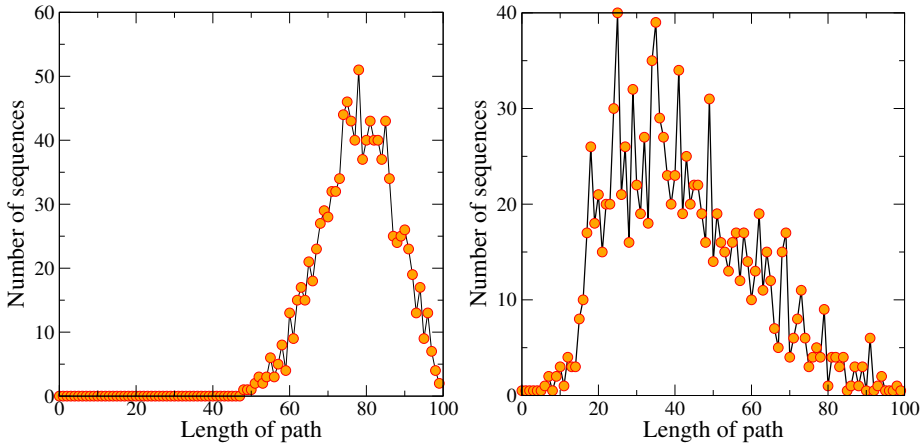

Fig. 4. (a) Length of neutral paths for 1200 sequences of length 100 cofolded with a single fixed sequence. (b) Length of neutral paths for 1200 sequences of length 100 when cofolding with two fixed sequences.

thirds of the fraction of neutral mutations in individual sequences. Note that the shape of the distributions depends very little on the sequence length. Furthermore, the distances scale linearly with chain length.

It is known that for single folding sequences, it is possible to exchange almost all nucleotides without leaving the neutral network [11, 8]. In the case we study here, the length of neutral paths for cofolding one sequence with one fixed interaction partner is shorter than in single-sequence RNA folding. The reason are the intermolecular base pairs. For some of these it would be impossible to find neutral mutations and so some bases of x cannot change without leaving the neutral network. In Fig. 4a we show the results for 1200 sequences of length $n = 100$ cofolding with fixed sequences of the same length.

The length of the path when cofolding one sequence with two different interacting RNAs is much shorter than the previous case and, of course, than in the case of folding an isolated RNA. Indeed, there are no paths along which all nucleotides of x could be replaced, Fig. 4b.

4. Concluding Remarks

An overview of the results obtained in this study is compiled in Table 1. We find that the cofolding map of two RNAs and the folding map of a single RNA secondary structure are similar, although the cofolding map is somewhat more constrained. In contrast, neutrality is drastically reduced in the case of multiple contraints, as one would expect. While the cofolding map $f(x \circ y)$ admits long neutral paths and large neutral networks, we find that the neutrality of the double-constraint co-folding map $f(x \circ y) \wedge f(x \circ z)$ is already too small ($\bar{\lambda} \approx 0.18$) to allow extensive connected neutral networks.

	Neutral mutations	Length of path
Single fold	0.33	100
Cofold with one sequence	0.32	75
Cofold with two sequences	0.18	40

Table 1. Fraction of neutral mutations and average path length for RNA sequences of length $n = 100$ and the three different definitions of neutrality.

Indeed, the connectivity thresholds for neutral networks with 4- and 6-letter alphabets (corresponding to the unpaired and pair parts of the RNA structure) are $\lambda^* \approx 0.37$ and $\lambda^* \approx 0.30$, respectively, as derived from a random graph approach [16]. While the observed neutrality $\bar{\lambda}$ is at or above the threshold for interactions with a single partner, the doubly contrained networks have an average connectivity that falls well below this threshold.

The distinction between the single-constraint and the multiple constraint situation is important from an evolutionary point of view. Neutral networks allow efficient adaptation on the corresponding fitness landscapes and imply easy evolution at the sequence level even in the limiting case of very strong stabilizing selection on the secondary structure level. This is what is observed for most functional noncoding RNA molecules, including tRNAs, rRNAs, RNAse P RNA, and microRNA precursors. In contrast, multiple constraints reduce neutrality to a point where the neutral networks decompose into disconnected components and sequence evolution becomes restricted by multiple structural constraints.

We suggest that the very slow rates of evolution of mature microRNAs could be due to multiple targets; indeed, structural constraints on the interaction complex with a single mRNA target cannot explain the almost perfect conservation of mature miRNA sequences.

Similarly, the evolution of hypercyclically coupled RNA replicators, e.g. in a prebiotic or in a laboratory setting, would be dramatically slowed down by the requirement that each RNA must properly interact as templates with its own riboreplicase and as a catalyst with other templates. This restriction implies an advantage in evolvability for systems in which template function and enzyme function are well separated, so that each RNA molecule has to satisfy only a single structural constraint.

Acknowledgments

This work was supported in part by *Consejo Nacional de Ciencia y Tecnologia*, México, the DFG Bioinformatics initative, Germany, and the European Science Foundation in the framework of COST D27 *"Prebiotic Chemistry and Early Evolution"*.

References

[1] R. A. Dimitrov and M. Zuker. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.*, 87:215–226, 2004.

[2] M. Eigen and P. Schuster. *The Hypercycle*. Springer-Verlag, New York, Berlin, 1979.

[3] W. Fontana, T. Griesmacher, W. Schnabl, P. F. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. *Monatsh. Chem.*, 122:795–819, 1991.

[4] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[5] W. Fontana and P. Schuster. Continuity in evolution: On the nature of transitions. *Science*, 280:1451–1455, 1998.

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

[7] C. V. Forst. Molecular evolution of catalysis. *J. Theor. Biol.*, 205:409–431, 2000.

[8] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. *Monath. Chem.*, 127:355–374, 1996.

[9] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II. structures of neutral networks and shape space covering. *Monath. Chem.*, 127:375–389, 1996.

[10] I. L. Hofacker. Vienna RNA secondary structure server. *Nucl. Acids Res.*, 31:3429–3431, 2003.

[11] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie*, 125:167–188, 1994.

[12] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. (USA)*, 93:397–401, 1996.

[13] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK, 1983.

[14] K. E. McGinness and G. F. Joyce. In search of an RNA replicase ribozyme. *Chem. Biol.*, 10:5–14, 2003.

[15] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10:1507–1517, 2004.

[16] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: Neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997.

[17] P. Schuster. Evolution in Silico and in Vitro: The rna model. *Biol. Chem.*, 382:1301–1314, 2001.

[18] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Roy. Soc. Lond. B*, 255:279–284, 1994.

[19] B. M. R. Stadler. Diffusion of a population of interacting replicators in sequence space. *Adv. Complex Systems*, 5(4):457–461, 2002.

[20] C. Stephan-Otto Attolini and P. Stadler. Evolving towards the hypercycle: A spatial model of molecular evolution. *Physica D*, 2004. submitted.