Structural Constraints and Neutrality in RNA

ULRIKE GÖBEL[a], CHRISTIAN V. FORST[b,*], AND PETER SCHUSTER[c,d]

[a,b,c]Institut für Molekulare Biotechnologie e.V., D-07708 Jena, Germany
[d]Institut für Theoretische Chemie, Universität Wien, A-1090 Wien, Austria

[*]Mailing Address: Dr. Christian Forst
Institut für Molekulare Biotechnologie
Beutenbergstraße 11, PF 100 813, D-07708 Jena, Germany
Phone: **49 (3641) 65 6459 Fax: **49 (3641) 65 6450
E-Mail: *a*: goebel@imb-jena.de
b: chris@imb-jena.de
c: pks@imb-jena.de
d: pks@tbi.univie.ac.at

Abstract

Generic properties of neutral networks of RNA secondary structures can be described by means of random graph theory. The success of this approach is dependent on details of the underlying secondary structure. Some of these dependencies are analyzed in this paper. In addition we present an algorithm, which, given a network does conform to the random graph model, allows to determine whether it is a connected graph. The algorithm is linear in time in the sequence length, this being possible because of local connectivity, a special property of graphs under the random graph model.

1 Introduction

The *neutral theory of evolution* [6, 7] emphasizes the role of drift (in addition to selection) in evolutionary scenarios where there exists more than a one fittest genotype. By drifting on a "plateau" of constant fitness [1], a population searches a broader range of the space of all genotypes, and thus may ultimately find a better optimum.

The sequence-structure mapping of RNA is an excellent example of a partly neutral landscape[12]. All sequences which do fold into the structure are equally fit on this landscape. It is possible to define a graph structure on this set of sequences by considering those sequences as adjacent which either differ by a single base exchange at an unpaired position or else by a base pair exchange. Together with that relation, the set is called the *neutral network* of the respective secondary structure.

Reidys *et al.* [11] have constructed generic fitness-mappings based on random graph (r.g.) theory. In their model, the set of sequences which do form a certain structure are considered a subset of the set all sequences which in principle can form the structure (the *compatible* sequences). The graph is constructed by selecting compatible sequences with some probability λ. Certain properties of the graphs are proven to exist above a threshold probability λ^* : Connectivity, density, and locality [10]. All three properties have the effect of facilitating drift on the network.

The topologies of the networks in the case of minimum free energy secondary structures have been determined by means of exhaustive folding [2, 3]. These data, which cover sequences up to length 30 over binary alphabets (AU, GC) are used as a reference to test predictions by the r.g. model. We show that the well-definedness of the structure bears on whether or not a network will conform to the model. Locality has also implications on time bounds for certain algorithmic problems on the networks. We present an algorithm which, based on the locality property, allows to classify a network as connected or disconnected using only a small subset of its sequences.

2 Test Set

We use a test set of 500 minimum free energy networks which each consist of two components of about equal size. The sizes of the components have to fulfill the following relations:

$(\frac{2}{3}\text{size}(1) \leq \text{size}(2) \leq \frac{3}{2}\text{size}(1)\,) \wedge (\frac{2}{3}\text{size}(2) \leq \text{size}(1) \leq \frac{3}{2}\text{size}(2)\,)$

Out of the list of all structures over the GC alphabet of length 30, rank ordered according to size, the first (largest) 500 networks fulfilling these relations were selected.

These two component networks were chosen because they have advantageous properties for testing the connectivity algorithm. This kind of topology combines a maximal fraction of sequence pairs which are not in the same component with a maximal size of the single components. Thus, the single components can be used for cross-testing: Using only pairs from the same component, a prediction algorithm should indicate connectedness. For the same network, if all pairs are considered, it should label the network disconnected.

3 Density, Connectivity, and Locality

The formal definitions from graph theory for connectivity and density are as follows:

Connected vertices: Two vertices $v, v' \in \text{v}[G]$ are called *connected* if there exists a path in G in which both vertices occur.

Connected graph: A graph G is *connected* if any two vertices $v, v' \in \text{v}[G]$ are connected.

Locally connected graph: A graph G is call *locally connected* if any two vertices $v, v' \in \text{v}[G]$ are connected by a path that is completely contained in a cylinder with axis endpoints at direct neighbors of v and v. (For a formal definition see [10].)

Dense graph: Let H be a finite graph. A subgraph $G < H$ is called *dense in* H if and only if $\overline{\text{v}[G]} = \text{v}[H]$. (Where $\overline{\text{v}[G]}$, the *closure* of the graph, consists of the vertices of the subgraph and those vertices of the surrounding graph which are connected to them by a path of length 1.)

For a neutral network, density is considered with respect to the graph of sequences which are able to form the underlying structure.

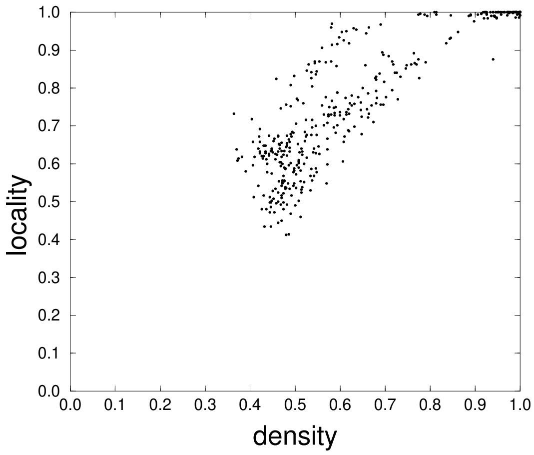

Fig. 1: Locally connected network are (close to) dense ones. Density is measured as follows: For each of 391 structures, 500 sequences were selected at random which are *compatible but not on the network*. The fraction of these sequences *which do have a neighbor on the network* is reported as the statistical density of the network.

Both density and connectivity are advantageous properties for a population of sequences which is subject to the constraint of maintaining a certain structure. On a connected network, the population can search the whole network without ever losing the structure. That does not necessarily mean that a broad range of sequences is searched: A connected network may well be very localized in sequence space, and in particular need not be dense in the set of its compatible sequences. On the other hand, a dense network may well be very fragmented.

At a pair distance of 12, our minimum free energy networks differ strongly with respect to the degree of locality (see Fig. 5).

Locality is strongly positively correlated with the density of the network. Moreover, the single components of the local networks in the test set are themselves near-dense, too. This leads to the consequence that the single components are almost everywhere close to each other. If a population which is evolving on such a network can tolerate short-term deviations from the reference structure, then it is expected to equilibrate about the two components. On the other hand, if the components are separated in sequence space, a population may become trapped in one of them.

4 Relating Network Topology to Properties of the Underlying Structure

In the r.g. model, two parameters determine the overall topology of a network: λ_u, the probability that a single base mutation is neutral with respect to the structure and λ_p, the analogous probability for a base pair mutation. The very idea of modeling generic

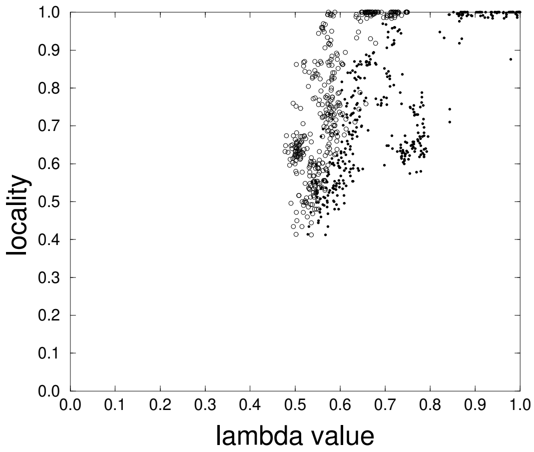

Fig. 2: Relation of degree of locality of a network and the average number of neutral neighbors, in 500 neighborhoods per network: Open circles denote the fraction of neutral single base exchanges out of all possible single base exchanges, black dots the corresponding fraction for base pair exchanges. For both types of exchange, there are threshold values above which almost all networks are local. For λ_u, this threshold is about 0.6, for λ_p it is about 0.85

features of a neutral network by a random graph approach contains the implicit notion of context dependent mutations: changing symbol a to symbol b at position i will result in a neutral neighbor in a fraction λ out of all possible contexts (symbol assignments at positions other than i). These probabilities, and hence the expected number of neutral neighbors are the same for all sequences.

The minimum free energy networks do exhibit a threshold phenomenon as predicted by theory. If the λ values are estimated by the mean fraction of neighbors of the appropriate type per sequence on the network, then the networks are dense and locally connected for $\lambda_u \geq 0.6$, and $\lambda_p \geq 0.85$. The threshold of 0.6 for λ_u is close to the one predicted by theory ($\lambda^* = 0.5$). The value for λ_p, however, is much larger than the one predicted by theory, which would be 0.5, too (Fig. 2).

If a compatible pair neighbor is not neutral, then at least one of the members of the pair engages in a different base pair in the alternative structure. That is true for all non neutral base pair mutations in 100 neighborhoods per network (391 networks) without a single exception. Non neutral single base exchanges, too, will yield a member of a base pair in 99% of all cases. Therefore, the number of possible alternative base pairs is an important number to consider. A nice representation of this is the *base pair probability matrix* which results from the partition function algorithm [9]. An entry p_{ik} of this matrix is the probability to observe position i paired with position k, over all structures of the equilibrium ensemble. A simple measure how easy alternative pairings are accessible

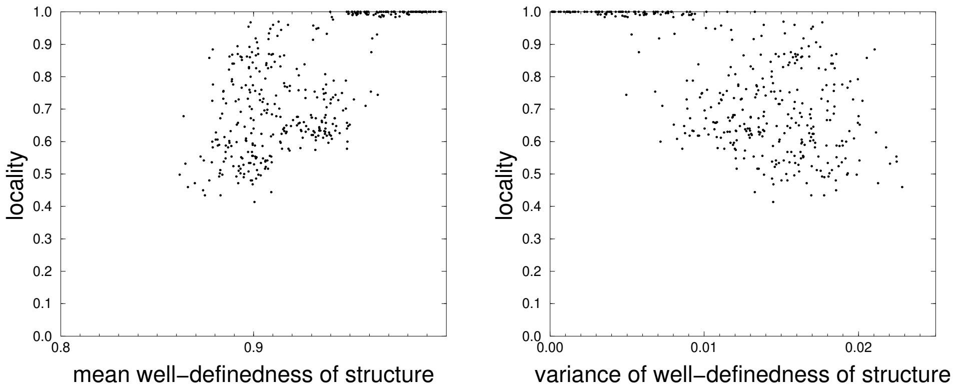

Fig. 3: Locally connected networks have structures with high well-definedness. Tn addition well-definedness in these structures does not depend much on the actual sequence. For both the mean and the variance of well-definedness, there is a threshold value beyond which the networks are no longer locally connected, but beyond which a law-like dependency on the actual degree of locality vanishes, too. The mean and variance are over 100 randomly selected sequences per network. Data for 391 networks are shown.

was derived from this matrix by [4] :

$d_k \equiv \max\{\max_i\{p_{ik}\}, 1 - \sum_i p_{ik}\}$

the *well-definedness* of the structure at position k. A value close to 1.0 indicates that this position will either pair almost surely with a specific partner, or else almost surely will be unpaired. By taking the mean of these values over all positions in the structure, one can define a well-definedness of the entire structure.

It turns out that in a range of average well-definedness of about 1.0 to 0.95 the networks are local (and also dense). Below this threshold, there is no completely local network any more. Locality is not clearly correlated with well-definedness below the threshold. The *variance* of well-definedness behaves in a similar way (Fig. 3). High variance of well-definedness means that it depends on the actual sequence which alternative structures are accessible, and which mutations will “trigger” them. This is in agreement with the fact that the number of neutral neighbors per sequence exhibits a high variance in these networks. Although there clearly is no law-like dependency of locality on variance of well-definedness beyond the threshold, the very fact that networks which do not conform to the r.g. model (which are not local) inevitably have the higher variances points to the possibility that the r.g. model could be improved by introducing a sequence dependent probability for the expected number of neutral neighbors.

The structures of locally connected networks are more stable (in terms of minimum free

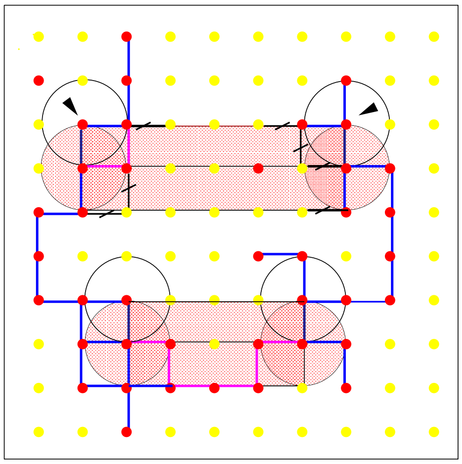

Fig. 4: Sketch of a simultaneous walk of a pair of neutral sequences in sequence space: Dark grey balls denote neutral sequences, light grey ones indicate sequences which are not neutral and hence fold in structures different to the corresponding structure of the pair. Arrows show the initial pair with open circles around as error-class. If connection is not successful by a simultaneous neutral walk a new pair is chosen. This pair is again checked for connectivity.

energy). To some extent this property can be read off directly from the structures : There is a slight but not very pronounced tendency for "good" networks to have the longer, bulge-free stacks. Structures in which the longest stack contains a bulge inevitably yield "bad" networks (data not shown).

Well-definedness measures (qualitatively) the gap between the best structure and alternative structures. It is negatively correlated with free energy of the best structure, although this is not implicit in its definition. Interestingly, the *variance* of energy is not correlated with the degree of density or locality of the network (data not shown). This can be explained as follows: the topology of the network is determined by the (binary) choice for a compatible neighbor to be neutral or not. The expected number of (non)neutral neighbors is influenced by the structure of the base pair probability matrix, with a different structure leading to a different expected number of neutral neighbors. The exact energies of the neutral neighbors matter only to the extent that there is no *consistent* tendency to more stable, and thus more well-defined, structures. This requirement can be met by a variety of energy distributions.

5 Making Use of Locality in Deciding Whether a Network is Connected

5.1 Deciding Whether a Pair of Sequences is Connected

Test-sets of networks were used as described in chapter 2. The definition of locality was implemented in a straightforward way (see Fig. 4). For a pair at distance n, there are n+2

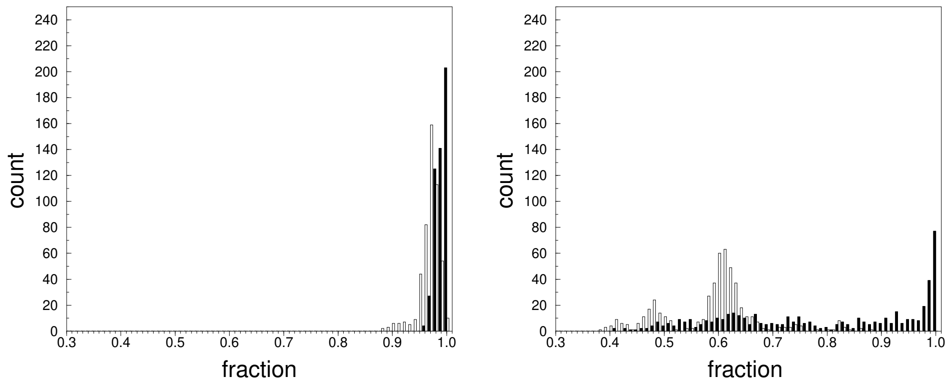

Fig. 5: At a pair distance of 3 (left), both the fraction of connected pairs that are locally connected (black bars) and the fraction of all pairs that are connected (locally or otherwise) (open bars) is close to 1.0. At a pair distance of 12 (right) the mean fraction of connected pairs approaches 0.5. The fraction of locally connected pairs among these, however, has turned into a spread out distribution, with an extensive overlap with the distribution of the connectedness fractions. Thus, one cannot tell from a given fraction of locally connected pairs whether it comes from a perfectly local network which is not connected, or else from a connected but nonlocal network, which generates a lot of false negatives.

intermediates on any shortest path that connects neighbors of the pair. They are visited from one after the other. Whenever an intermediate does not preserve the structure, a detour to the surface of the cylinder is made, in case the path is still on the axis. If it had already been mapped to the surface at a previous intermediate, the algorithm returns to that intermediate and tries to find another "bypass mutation". If none is found, the pair is considered disconnected.

A fixed number of 20 cylinders per pair is tested before the pair is finally labeled not locally connected.

For each network, 500 different pairs of sequences are tested for connectedness. As it is shown it Fig. 5, the same success rate can result from a non local connected network and from a local disconnected network. Success rate in local connection thus is not enough to estimate connectedness.

5.2 Lengths of Double Paths do Discriminate Connected Networks from Disconnected Ones

Local connectivity at low distances is high, but these distances do not yield many informative pairs (pairs in different components). On the other hand, high distances yield many informative pairs, but some networks are very badly locally connected at these

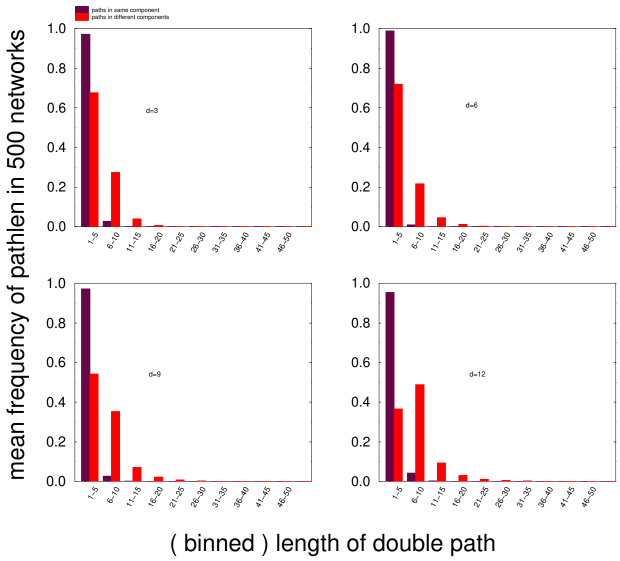

Fig. 6: Distribution of path lengths: For different hamming distances between checked pairs the length of the simultaneous neutral walk is determined. Break-off criterion for the walk is either connectivity of the checked pair, a length of walk longer than 50 steps or divergence of the pair and terminating in a dead end in sequence space.

distances. These two conflicting facts can be combined in the following way:

Start with a pair at a long distance. It will have a high probability not to be connected. If it is not, choose a pair of neutral neighbors of the initial pair , and try to connect these. The distance of the neighbors is forced to decrease, and the length of the resulting double path is restricted to 50 pairs.

For a connected initial pair, there is a probability to terminate the path by successful connection at every step, which should lead to shorter paths on average. This is indeed the case for all networks (see Fig. 6). It is possible to classify the networks as connected (disconnected) by discriminant analysis if the length distribution of the paths is used as the feature vector. The success rate of classification is about 90% for networks over structures of various lengths and (binary) alphabets (data not shown).

6 Discussion

Neutral networks of RNA secondary structures to a first approximation can be modeled as random graphs, with a fixed expected number of neutral neighbors per sequence. In this paper, we have tested some predictions from the random graph model on the neutral networks which arise if minimum free energy is the criterion for structure assignment. On minimum free energy networks, the ease of accessibility of alternative structures is critical for the expected number of neutral neighbors of a sequence. We have shown that stable, well defined structures lead to networks which conform to the r.g. model:

they are dense and locally connected (the model would in addition predict connectivity of the entire network. This by definition can not hold in our test networks, which were chosen to consist of two components of about equal size). Indeed these networks are near-regular graphs, with the expected neutrality of a given position being nearly independent of the sequence. If, however, a structure is not very well defined, then it depends on the actual sequence which alternative structures are most easily (or at all) accessible (in terms of the r.g. model, different sequences on the network belong to different *intersections* with the neutral networks of other structures). Accordingly, these structures exhibit a broad distribution of numbers of neutral neighbors per sequence. A more realistic model should account for this fact by non constant probabilities for a compatible neighbor to be neutral. Another consequence of the role of alternative pairings is the fact that, for a network to conform to the model, the expected fraction of neutral base pair exchanges has to be much higher than predicted by theory (0.85 (observed) vs. 0.5 (predicted)).

It has been reported recently [5, 8] that sequences which have been optimized for a given structure ("designed sequences") tend to have a higher energy gap between their native ground state and the next higher state, and that they tend to maximize the independence of mutational effects. It looks as if these very properties will yield neutral networks which can be modeled as random graphs according to [10]. In "badly design-able" structures, however, the network topology will exhibit micro-inhomogeneities, which will influence the fate of a population which evolves on such a network. In particular, these networks will be badly locally connected.

In locally connected networks, it is computationally cheap to give a probability value for whether or not the network is connected, without examining all sequences. In this paper, we have presented an idea how to solve this problem in the case of our minimum free energy networks, which are not guaranteed to be local. To be able to predict global properties of a neutral network, like f.e. connectivity, can be of importance in the design of SELEX type experiments [13]: Different mutational strategies might be appropriate on different topological types of networks.

References

[1] C. V. Forst, C. Reidys, and J. Weber. Evolutionary dynamics and optimization: Neutral Networks as model-landscape for RNA secondary-structure folding-landscapes. In F. Morán, A. Moreno, J. Merelo, and P. Chacón, editors, *Advances in Artificial Life*, volume 929 of *Lecture Notes in Artificial Intelligence*, Berlin, Heidelberg, New York, 1995. ECAL '95, Springer. Santa Fe Preprint 95-10-094.

[2] W. Grüner, 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. *Monatsh. Chem.*, 127:355–374, 1996.

[3] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of rna sequence struc-

ture maps by exhaustive enumeration II. structures of neutral networks and shape space covering. *Monatsh. Chem.*, 127:375–389, 1996.

[4] M. Huynen, A. Perelson, W. Vieira, and P. Stadler. Base pairing probabilities in a complete hiv-1 rna. *J.Mol.Biol.*, submitted, 1995.

[5] A. Irbaeck and H. Schwarze. Sequence dependence of self-interacting random chains. *J. Phys. A: Math. Gen.*, 28:2121 – 2132, 1995.

[6] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624 – 626, 1968.

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

[8] H. Li, R. Helling, C. Tang, and N. Wingreen. Emergence of preferred structures in simple model of protein folding. *Science.*, 273:666 – 669, 1996.

[9] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[10] C. Reidys. *Neutral Networks of RNA Secondary Structures.* PhD thesis, Friedrich Schiller Universität, Jena, May 1995.

[11] C. Reidys, P. Stadler, and P. Schuster. Generic properties of combinatory maps. neutral networks of rna secondary structures. *Bull. Math. Biol.*, in press, 1996.

[12] 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.(London)B*, 255:279–284, 1994.

[13] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249:505, 1990.