J Mol Evol (1996) 43:165–169

JOURNAL OF MOLECULAR EVOLUTION

Exploring Phenotype Space Through Neutral Evolution

Martijn A. Huynen

Theoretical Biology and Biophysics, Los Alamos National Laboratory, MS-K710, Los Alamos, NM 87545, USA
Center for Nonlinear Studies, Los Alamos National Laboratory, MS-B258, Los Alamos, NM 87545, USA
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Received: 23 December 1995 / Accepted: 17 March 1996

Abstract. RNA secondary-structure folding algorithms predict the existence of connected networks of RNA sequences with identical secondary structures. Fitness landscapes that are based on the mapping between RNA sequence and RNA secondary structure hence have many neutral paths. A neutral walk on these fitness landscapes gives access to a virtually unlimited number of secondary structures that are a single point mutation from the neutral path. This shows that neutral evolution explores phenotype space and can play a role in adaptation.

Key words: RNA secondary structure — Neutral Evolution — Adaptive evolution — Genotype-phenotype relation — Sequence space — Fitness landscape

Introduction

Ever since Sewall Wright introduced the metaphor of an ''adaptive landscape'' (Wright 1932) the view on adaptive evolution has been dominated by the image of an uphill walk of a population on a mountainous fitness landscape in which it can get stuck on suboptimal peaks. The neutralist perspective that evolution at the molecular level is dominated by nonadaptive, neutral changes (Kimura 1983) has hardly changed this picture. A notable exception is Maynard-Smith's argument that the distribution of functional proteins in (neutral) networks in sequence space could facilitate adaptive evolution (Maynard-Smith 1970). In a letter to Kimura, Sewall-Wright addressed the question, to what extent can adaptive evolution benefit from neutral evolution? ''Changes in wholly nonfunctional parts of the molecule would be the most frequent ones but would be unimportant, unless they occasionally give a basis for later changes which improve function in the species in question which would then become established by selection'' (Provine 1986). Here we study the potential for Wright's scenario: Can neutral evolution facilitate adaptive evolution by increasing the number of phenotypes that can be reached with a point mutation from an original phenotype?

An answer to this question requires detailed knowledge of genotype–phenotype relations. We use the mapping from RNA sequence to RNA secondary structure as a paradigm for such a relation. The secondary structure of a single-stranded RNA sequence is its pattern of complementary base pairings (Watson-Crick and G-U pairs). As opposed to the protein case, the secondary structure of RNA sequences is relatively well defined; it provides the major set of distance constraints that guide the formation of tertiary structure and covers the dominant energy contribution to the 3D structure. RNA secondary structure represents a qualitatively important description of RNA, as is reflected in its conservation in evolution and in the independent generation of ribozymes that share both secondary structure and function (Ekland et al. 1995). This is not to say that secondary structure is the only determinant of function; it is, however, often essential.

For relatively short RNA sequences the secondary structure can be predicted using algorithms that minimize free energy of the base pairing and of a number of other structural elements like hairpin loops and internal loops (Zuker and Stiegler 1981). Using these algorithms one can analyze how the secondary structures are dis-

tributed over sequence space. The distribution of secondary structures over sequence space shows some remarkable features: There is a large redundancy in the mapping from RNA sequences to RNA secondary structures: For a sequence of length N, there are 4^N sequences, whereas there are about 1.8^N structures (Waterman 1978; Schuster et al. 1994). Furthermore, within the total set of secondary structures, only a small fraction dominates sequence space. These relatively frequent secondary structures form networks in sequence space, where a network is defined as a set of sequences that fold into the same structure and that are connected in sequence space through point mutations (Schuster et al. 1994). On the other hand, RNA landscapes are extremely rugged; randomly changing only 15% of the nucleotides gives rise to a change in secondary structure that is no less than the difference between the secondary structures of random sequences (Fontana et al. 1993; Huynen et al. 1993). The whole picture can be captured in the term ''smoothness within ruggedness''; on average the landscape is very rugged, but there exist paths that percolate through sequence space on which the secondary structure remains unaltered. These properties of RNA landscapes are insensitive to whether the folding algorithm is thermodynamic, kinetic, or maximum matching (Tacker et al. 1996).

Perpetual Innovation Along the Neutral Net

We use the mapping between RNA sequence and RNA secondary structure as a toy model for the mapping between genotype and phenotype. Within this model, point mutations that leave the secondary structure unaltered are called neutral mutations. A network of identical secondary structures in sequence space is called a neutral net. Neutral evolution within this model is the process where a population moves, due to mutations, genetic drift, and selection, through sequence space over the neutral net of a specific secondary structure. To answer the question of whether neutral evolution can change the potential for adaptive evolution, it is essential to know whether the secondary structures that neighbor the neutral net are different for various positions on the net. That is to say: Does neutral evolution give access to new, previously unencountered structures? We performed a walk on the neutral net for a $tRNA^{phe}$ clover-leaf secondary structure and recorded all the different secondary structures that were encountered as one-point mutants of the net (Fig. 1). The total number of structures encountered increased linearly in time, with every neutral mutation giving access to, on average, 18.1 new structures. The number of new accessible structures per mutation we call the ''rate of innovation.'' One can compare the rate of innovation along the neutral net with the number of new structures observed along a purely random walk through sequence space, where the secondary structure is not conserved. For sequences of the length of $tRNA^{phe}$ (76 nucleotides), we observed an average of 39 new accessible structures per step in the random walk. Hence, restricting the walk to the neutral net reduces the rate of innovation by a factor of two to three. The reduction in the rate of innovation, however, is insignificant in comparison to the fraction of sequence space that can be accessed from the neutral net. An upper bound to the number of sequences that fold into a $tRNA^{phe}$ secondary structure is the number of sequences that are *compatible* with a $tRNA^{phe}$ secondary structure; i.e., the number of sequences that have the potential for base pairing at the positions at which the $tRNA^{phe}$ secondary structure has base pairs. A $tRNA^{phe}$ secondary structure has 20 base pairs and 36 unpaired positions. Given that there are 6 different base pairs (G–C, C–G, A–U, U–A, G–U, U–G), there are $6^{20} \times 4^{36} = 1.7 \times 10^{37}$ sequences compatible with a $tRNA^{phe}$ secondary structure. In a set of randomly generated $tRNA^{phe}$ compatible sequences we observed that less than one in 10,000 had a $tRNA^{phe}$ minimum free energy structure ($P < 0.001$). The fraction of sequence space that folds into a $tRNA^{phe}$ secondary structure can then be estimated to be less than $0.0001 \times 1.7 \times 10^{37}/4^{76} = 3 \times 10^{-13}$. Hence the reduction in the rate of innovation that is caused by restricting the random walk to sequences with a $tRNA^{phe}$ secondary structures is negligible compared to the reduction in accessible sequences. These results show that Wright's scenario is plausible; neutral changes can ''set the stage'' for a mutation that leads to a better-adapted structure. Moreover, for any reasonable time scale we do not observe a saturation in the number of secondary structures that can be observed along the neutral walk. This is the main point of this article.

Besides depending on secondary structure, RNA function also relies on primary (e.g., the anticodon in tRNAs) and tertiary structure. Analyzing a rate of innovation with respect to other properties goes beyond the scope of this paper, and is, as far as tertiary structure is concerned, computationally not feasible. We can analyze to some extent, however, how extra constraints on the primary and tertiary structure will affect the innovation rate at the level of the secondary structure. Extra constraints on the structure will in general lower the rate of innovation: Less potential for neutral mutations reduces the rate at which the walk progresses through sequence space and encounters new structures. We observed that conserving nucleotides at the anticodon positions led to rates of innovation of 17.7 (GGG) and 16.6 (AAA). The tertiary base-pair interactions that can be simulated to some extent are pseudoknots. Conserving two or three pair of complementary bases between the first and third hairpin-loop of the clover-leaf led to innovation rates of 19.2 and 19.1, respectively. Finally, the simultaneous conservation of both a GGG in the anticodon loop and three complementary bases between the first and third hairpin-loop led to an innovation rate of 19.7. These results show that extra constraints do not always reduce

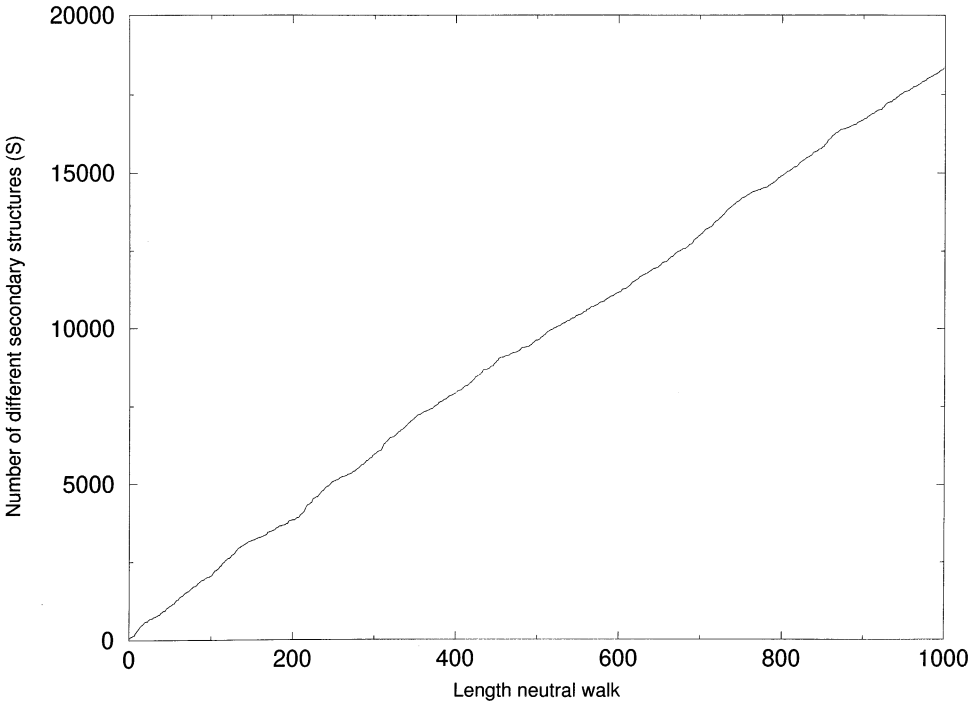

Fig. 1. Perpetual innovation along the neutral net. A random walk is performed on a neutral net for sequences that fold into a tRNA^{phe} clover-leaf secondary structure, and the number of different secondary structures that neighbor the neutral walk is counted. The algorithm works as follows: An initial sequence (length 76) is chosen that folds into a tRNA^{phe} secondary structure. All the one-point mutants of the sequence are generated and their secondary structures are calculated. The number of different secondary structures is counted (*S*), and they are stored in a set (*Q*). The one-point mutants that retain the tRNA^{phe} secondary structure are called neutral mutants. One of the neutral mutants is chosen randomly. For this neutral mutant once again the secondary structures of all its one-point mutant neighbors are calculated. Of these secondary structures, the ones that are not present in *Q* are added to *Q*, and their number is added to *S*. Once again one of the neutral mutants is chosen randomly, and the procedure is repeated. Thus *S* gives the cumulative number of different secondary structures that are encountered along a random walk over a neutral net. The results show that there is a linear increase in the number of different secondary structures that are encountered along the neutral walk; in this case there are, on average, 18.1 new structures to be discovered for every step. Thus, there is perpetual innovation in secondary structures observed along the net.

the rate of innovation. Constraints on the single-stranded regions force the neutral walk to accept more mutations in double-stranded regions, which increases the rate of innovation. What is more important than the actual rates of innovation, however, is that they remained linear under all the constraints that were tested.

Our model of neutrality only concerns the minimum free energy structure. Biologically functional secondary structures are often ‘‘well-defined’’: base pairs in their minimum free energy structure have, compared to the base pairs in the secondary structures of random sequences, a relatively high probability of occurring within the entire Boltzmann ensemble (Huynen et al. 1996b). Since only a subset of the sequences with a specific minimum free energy structure fold into that structure with a relatively high probability, constraints on this probability would reduce the neutrality and are expected to reduce the rate of innovation.

Correlations in Structure Space

One might argue that the existence of a rate of innovation is no more than a logical result of the fact that neutral networks percolate sequence space. Although the presence of percolating networks is certainly a necessary condition for the above, it is not sufficient. One could envision a scenario in which the secondary structures that neighbor the net are part of a relatively small subset of all the secondary structures. Evolution then could get stuck on a neutral ridge instead of a neutral peak. To assess whether there is something like a small subset of secondary structures that neighbors tRNA^{phe} secondary structures, we performed the same neutral walk experiment as above, but now, instead of counting new, unobserved structures, we counted structures that had been observed before. We determined how the number of equal structures that neighbor the net depends on the number of neutral mutations between the sequences on the net. In Fig. 2 we show that the number of neighboring secondary structures that two sequences that fold into a tRNA^{phe} structure have in common depends strongly on the distance between those two sequences in sequence space and saturates to a low level of about six common neighboring secondary structures as the sequences approach the distance between random sequences. Hence there appear indeed to be some structures that are often

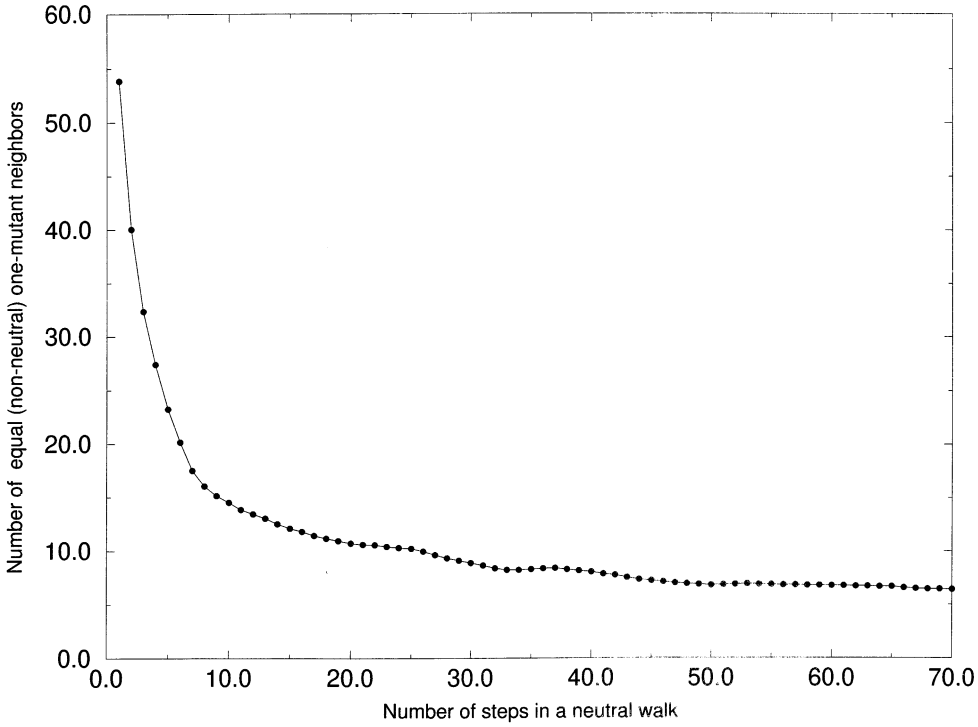

Fig. 2. Conservation along the neutral net. A random walk is performed in the same manner as described in the legend of Fig. 1. Instead of counting new, unobserved structures, the number of structures that have been observed before is counted. Plotted is how the number of neighboring secondary structures tRNA^{phe} structures have in common depends on the number of mutations that separates them in a neutral walk. The number of steps in a neutral walk is an upper bound on their distance in sequence space, given the possibility of back mutations or mutations that do not increase the distance between the sequences. For short walks on the neutral net the overlap in secondary structures is quite high, given that the number of different secondary structure that neighbor any tRNA^{phe} secondary structure is about 80. The overlap in secondary structures saturates to a level of about six common structures for any two tRNA^{phe} structures that are far apart on the net. The figure shows conservation along the neutral net: tRNA^{phe} secondary structures in general have a few neighboring secondary structures in common.

close to tRNA^{phe} structures. We generated 1,000 independent sequences that fold into a tRNA^{phe} secondary structure using a reverse folding algorithm (Hofacker et al. a). In Fig. 3 we plot the frequency distribution for secondary structures that neighbor these sequences. Some structures are nearly always neighboring the tRNA^{phe} secondary structures (in more than 80% of the sequences). Not surprisingly, these structures turn out to be relatively similar to the tRNA^{phe} secondary structure; the most frequent ones are at a distance of one base pair to the tRNA^{phe} structure. However, most of the structures that neighbor tRNA^{phe} structures are unique; out of the total of 6×10^4 different secondary structures that neighbor the 1,000 independent tRNA^{phe} structures, only 3,099 occurred more than once. The recurrence of structures that are similar to the structure on the neutral net is in accordance with earlier results where it was shown that most of the secondary structures that neighbor a single sequence/secondary structure are relatively similar to that structure (Huynen et al. 1993; Fontana et al. 1993).

The structures that occur relatively often near a tRNA^{phe} secondary structure complement the picture that was sketched in the first part of this article, where we observed perpetual innovation along the neutral net. Here we observe that besides the innovation, there is also conservation. The tRNA^{phe} secondary-structure neutral net casts a shadow in structure space of a set of relatively similar secondary structures. This type of correlation in structure space shows that the phenomenon of perpetual innovation is not trivial. If the correlation were too high, neutral evolution would only be able to reach a very limited subset of structures from any single neutral net. As it turns out this is not the case; neutral evolution at the level of secondary structure gives access to a virtually unlimited number of structures and can thus play an important role in adaptive evolution. A scenario of adaptive evolution where a population evolves over a (suboptimal) neutral net until it encounters another net with a better secondary structure, after which it relocalizes to this new net, was observed in simulations of adaptive evolution on a fitness landscape that was based on RNA secondary structure (Huynen et al. 1996a).

Acknowledgments. The concept of a shadow in structure space that is cast by a neutral net was developed in collaboration with Walter Fontana. I thank Paulien Hogeweg for useful discussions. I thank Peter Stadler, Alan Perelson, and the referees for comments on the manuscript. Part of this work was done under the auspices of the U.S. Department of Energy and supported by the Center for Nonlinear Studies at Los Alamos National Laboratory and by the Santa Fe Institute.

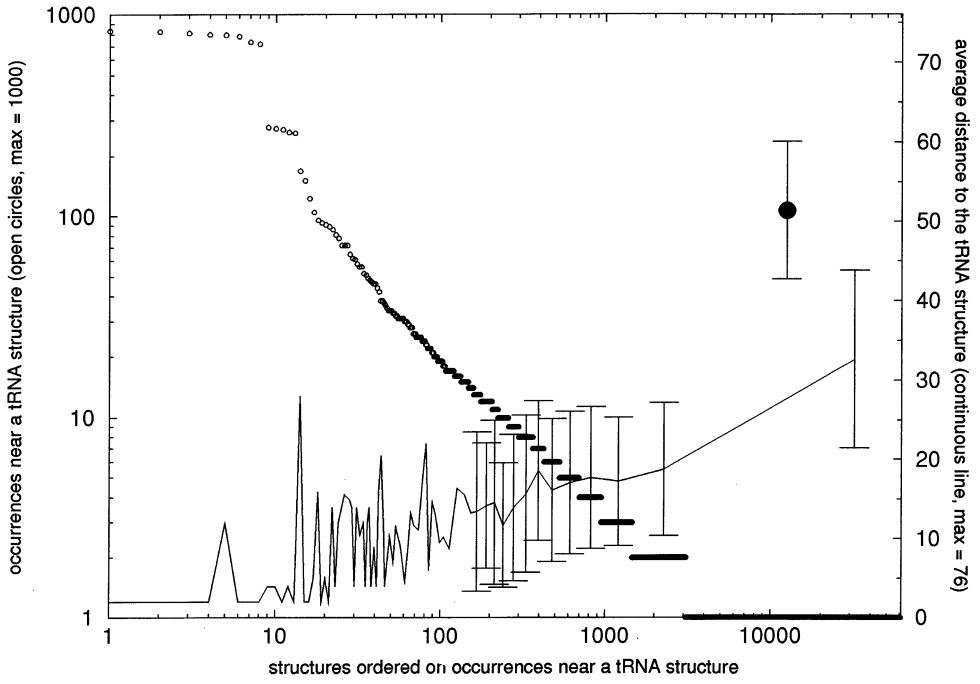

Fig. 3. Correlations in structure space. One thousand sequences that fold into a tRNA^{phe} secondary structure were generated using the reverse folding algorithm (Hofacker et al. a.). For these sequences the secondary structures of all the one-point mutants were calculated. The one-point mutants were compared for secondary structures that occurred as the neighbor of different sequences that fold into a tRNA^{phe} secondary structure. The plot shows the frequency distribution of the structures, ordered on their frequency of occurrence near a tRNA^{phe} secondary structure (*open circles*, scaling on the left ordinate). Some structures neighbor a large fraction of the tRNA^{phe} structures, giving the distribution a ''shoulder.'' The overall distribution follows a Zipf's law: such a distribution, where the logarithm of the frequency of a structure decreases linearly with the logarithm of the rank of that frequency, has also been observed for the secondary structures of all the sequences of a given length (Schuster et al. 1994). For comparison we calculated the secondary structures of 80.000 random sequences. In this set only 20 structures occurred twice, and none occurred more than twice. We also plot the average distance between the secondary structures and the tRNA^{phe} secondary structure per ''frequency class'' (*continuous line*, scaling on the right ordinate). In the lower-frequency classes multiple secondary structures are present. The *error bars* represent 1 SD for frequency classes that have 15 or more different secondary structures. The *extra error bar with the circle in the center* gives the average distance between the secondary structure of random sequences. Distance between secondary structures is measured by taking the Hamming distance between string representations of secondary structures (Konings and Hogeweg 1989). The results show that secondary structures that frequently neighbor the tRNA^{phe} secondary structure have a relatively similar structure. The most frequent structures are at a distance 2, which corresponds to one base pair. Structures that are less frequent have on average a larger distance to the tRNA^{phe} secondary structure. For the least-frequent structures the average distance to the tRNA^{phe} is about three-fifths that between the secondary structures of random sequences.

References

Ekland EH, Szostak JW, Bartel DP (1995) Structurally complex and highly active RNA ligases derived from random RNA sequences. Science 269:364–370

Fontana W, Konings DAM, Stadler PF, Schuster P (1993) Statistics of RNA secondary structures. Biopolymers 33:1389–1404

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster PA (a) Vienna RNA Package. pub/RNA/ViennaRNA-1.03 ftp.itc.univie.ac.at (Public Domain Software)

Huynen MA, Konings DAM, Hogeweg P (1993) Multiple coding and the evolutionary properties of RNA secondary structures. J Theor Biol 165:251–267

Huynen MA, Stadler PF, Fontana W (1996a) Smoothness within ruggedness: the role of neutrality in adaptation. Proc Natl Acad Sci USA 93:397–401

Huynen MA, Perelson A, Vieira W, Stadler PF (1996b) Base pairing probabilities in a complete HIV-1 genome. J Comp Biol 3:253–274

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Konings DAM, Hogeweg P (1989) Pattern analysis of RNA secondary structures, similarity and consensus of minimal-energy folding. J Mol Biol 207:597–614

Maynard-Smith J (1970) Natural selection and the concept of a protein space. Nature 255:563–564

Provine WB (1986) Sewall Wright and evolutionary biology. University of Chicago Press, Chicago

Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc R Soc Lond (Biol) 255:279–284

Tacker M, Stadler PF, Bornberg-Bauer E, Hofacker IL, Schuster P (1996) Robust properties of RNA secondary structure folding algorithms. Working Paper 96-04-016, Santa Fe Institute

Waterman MS (1978) Secondary structure of single-stranded nucleic acids. Studies on foundations and combinatorics. Adv Math Suppl Studies 1:167–212

Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Jones DF (ed) Proceedings of the sixth international congress on genetics, vol 1. pp 356–366, Brooklyn Botanic Garden, New York

Zuker M, Stiegler P (1981) Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9:133–148