Available online at www.sciencedirect.com
ScienceDirect
BioSystems 90 (2007) 475–485
BioSystems
www.elsevier.com/locate/biosystems

New structural variation in evolutionary searches of RNA neutral networks

Sumedha [a,*], Olivier C. Martin [a,b], Andreas Wagner [c]

[a] *Univ. Paris-Sud, LPTMS, UMR8626, CNRS, Orsay F-91405, France*
[b] *Univ. Paris-Sud, UMR8120, Laboratoire de Génétiqu Végétaldu Moulon, CNRS, Gif-sur-Yvette F-91190, France*
[c] *Department of Biochemistry, Blgd. Y27, University of Zurich Winterthurerstrasse 190, CH-8057 Zurich, Switzerland*

Received 27 September 2006; received in revised form 28 November 2006; accepted 28 November 2006

Abstract

RNA secondary structure is an important computational model to understand how genetic variation maps into phenotypic (structural) variation. Evolutionary innovation in RNA structures is facilitated by neutral networks, large connected sets of RNA sequences that fold into the same structure. Our work extends and deepens previous studies on neutral networks. First, we show that even the 1-mutant neighborhood of a given sequence (genotype) G_0 with structure (phenotype) P contains many structural variants that are not close to P. This holds for biological and generic RNA sequences alike. Second, we analyze the relation between new structures in the 1-neighborhoods of genotypes G_k that are only a moderate Hamming distance k away from G_0, and the structure of G_0 itself, both for biological and for generic RNA structures. Third, we analyze the relation between mutational robustness of a sequence and the distances of structural variants near this sequence. Our findings underscore the role of neutral networks in evolutionary innovation, and the role that high robustness can play in diminishing the potential for such innovation.

Keywords: RNA neutral networks; Robustness; Evolutionary innovation

1. Introduction

A properly formed RNA secondary structure is necessary for the biological functions of many RNA molecules, and a variety of algorithms exist to determine RNA secondary structure from an RNA sequence (Hofacker et al., 1994; Tacker et al., 1996; Zuker, 2000). For these reasons, RNA secondary structure is an important computational model to understand how genetic variation maps into phenotypic (structural) variation (Fontana, 2002; Fontana and Schuster, 1998a; Schuster et al., 1994), and thus to understand the evolutionary dynamics of molecular innovations.

Our point of departure are RNA sequences that adopt a specific minimum free energy (mfe) secondary structure, which we can think of as necessary for some hypothetical biochemical process. This process might involve catalysis or just specific binding to some molecule. Some variant of this structure – perhaps very rare in the space of all possible structures – may greatly improve this biological function, or it may even lead to a new function. The question is how to find such a variant, *if we are not allowed to destroy the original structure* during an evolutionary search for this innovation. Part of the answer lies in the fact that the sequences folding into a given structure form one or a few "neutral networks" that can be traversed by single point mutations, and that span most of sequence space. This holds at least for generic structures, structures into which a sufficient number of sequences fold (Schuster et al., 1994).

* Corresponding author. Tel.: +33 1 69153604; fax: +33 1 69156525.
E-mail address: sumedha.sumedha@gmail.com (Sumedha).

0303-2647/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2006.11.007

An evolutionary search that must not destroy the original structure is effectively restricted to the 1-mutant neighborhood of a neutral network (Fontana and Schuster, 1998b). (We define a k-mutant neighborhood of a sequence G_0 as containing all sequences that differ from G_0 in at most k residues.) Any such evolutionary search would start at one sequence. The 1-mutant neighborhood of this sequence contains only a limited number of structural variants, and thus only limited potential for evolutionary innovation. However, since the sequences folding into a structure are *connected* in a neutral network, the search can explore a great many sequences and their 1-mutant neighbors, without ever leaving the original structure (Schuster et al., 1994).

Among the substantial body of work that has explored the relation between sequence and structure space (Fontana and Schuster, 1998b; Huynen, 1996; Huynen et al., 1996; Reidys et al., 1997; Schuster et al., 1994; van Nimwegen et al., 1999), one paper (Huynen, 1996) is of particular relevance. That paper focused on a specific, biologically important RNA structure, that of phenylalanine tRNA ($tRNA^{Phe}$). It showed that an exploration of this structure's neutral network through a random walk encounters an ever-growing repertoire of new structures in its neighborhood, a repertoire that does not become exhausted even for very long random walks. In addition, the 1-neighborhoods of distant sequences on the neutral network share very few structural variants. Another important result from previous work is that a thermodynamically stable and mutationally robust sequence encounters few structural innovations in its neighborhood (Ancel and Fontana, 2000). We here extend and deepen these previous analyses. First, we show that continual structural innovation is a property not only of biologically important, but also of generic structures. Second, we statistically explore the relationship between mutational robustness and structural innovation, not only by counting the *number* of structural variants, but also by analyzing their *distances* to a reference structure. Importantly, we do so for generic structures, and not just for biologically important structures.

2. Results

2.1. Around any one sequence, substantial structural variation is abundant

Consider a reference sequence (genotype) G_0 and its structure (phenotype) P. Among the 1-mutant neighbors of G_0 (there are $3n$ of them) some fraction will adopt a structure different from P. One would think that most structural variants will only differ slightly from P, because base-pair stacks resist structural changes in response to single base changes, be it for both biological and random RNA structures (Higgs, 1993).

To find out whether this is the case, or whether a local exploration around a given sequence generates a great diversity of new structures, we took the following approach. We sampled randomly chosen genotypes G_0 with a given structure P, and determined the distribution of the *structure distance* D between P and the structures of the 1-mutant neighbors of G_0 (see Section 4 for details). Specifically, we used two biological structures in this approach. The first structure is a 54-mer hammerhead motif of an RNA in peach latent mosaic viroid (Ambros et al., 1998). To ascertain that our results were not artifacts of the specific structure we chose, we also determined the same distance distribution averaged over many randomly chosen structures of length 54 (see Section 4). The second biological structure was a phenylalanine tRNA of length 76, where we similarly determined for comparison the distance distribution for many randomly chosen 76-mer.

Fig. 1a and b show these distance distributions. A key qualitative feature is similar for the two biological structures ($n = 54$ and 76) and the random structures, in that the most probable distance is the smallest possible distance ($D = 2$). Thus, variants of any one structure tend to be similar to it. However, we also note that *most* of the distributions' mass is located at moderate to large distances. For example, the median structure distances for the 1-mutant variants of the biological structures are $D = 14$ (hammerhead) and $D = 24$ (tRNA). This means that a typical single point mutation affects 7–12 base pairs, even in these short structures. All this is not just an artifact of a particular measure of structure distance: we also see it for the "bond" distance (see Section 4; Fig. 2), where median structure distances are $D = 12$ (hammerhead) and $D = 22$ (tRNA).

In sum, even at the smallest possible sequence distance, one observes a broad spectrum of structures with varying distance from a reference structure. Needless to say, however, the number of structures accessible from *anywhere* on a neutral network is much larger than that found in a single sequence's 1-neighborhood. For instance, for our 54-mer hammerhead motif we found 259,689 distinct structures among 10^6 randomly generated innovative 1-mutants of sequences on the neutral network. Furthermore, this number rises linearly with our sampling size, suggesting that the *total* number of distinct structures is many times larger still.