Biopolymer folding: when the details don't matter

The "folding problem" consists in predicting a structure at a given level of resolution from a particular sequence of amino acids or nucleotides. A solution to this problem provides information about the distribution of structures over the space of possible sequences, and hence about the accessibility of structures through mutation and selection. Without this information biological evolution at the molecular scale cannot be understood.

Recently, Tang *et al.* [1] showed that the distribution of structures over sequences is far from uniform in a toy model of protein folding based on "lattice proteins". Few structures are realized by a large number of sequences, while others are rare (or not realized at all). The purpose of this comment is to point out that the same result was shown to hold for RNA secondary structures [2]. Tang *et al.* have provided us with the data of their Figure 2A [1] which we display as a rank-ordered frequency plot (Fig.1) together with our RNA data based on the folding of all sequences of length 27 and 30 over the binary GC alphabet [3]. Qualitatively the same picture emerges from large samples of sequences of length 100 over the biophysical AUGC alphabet [2].

The similarity of these two results is particularly remarkable, because (i) two rather different classes of biopolymers are being considered (polypeptides versus polynucleotides), (ii) two different levels of "structure" are involved (three dimensional cube-filling self-avoiding walks versus planar graphs), and (iii) two different abstractions of folding are used (a toy model stripped to the essentials versus folding based on empirical energies and widely used to interpret laboratory data). These statistical RNA results are robust with respect to small variations in the energy set, changes in the alphabet and the folding strategy [4]. They even hold when taking into account the Boltzmann-weigthed ensemble of structures for each sequence [5].

In view of the results of Tang's group, one is tempted to suggest that some universal properties of folding are emerging. In [2] we reported on two additional statistical features which could be candidates for universals: (i) shape space covering, that is, the property that all frequent structures are found in a small (compared to sequence length) neighborhood of any random sequence, and (ii) the existence of neutral networks, that is, the property that sequences differing in at most two point mutations, while folding into the same structure, form a graph that percolates through sequence space. It should be feasible to check whether these findings extend to lattice proteins.

The folding of biopolymer sequences into structures is an important constraint that has shaped evolutionary history at the molecular scale, and is equally effective in test-tube evolution and for viruses in the wild. The existence of folding universals raises the intriguing possibility that abstract toy models, which are far too inaccurate for predicting the correct structure of a particular sequence, may nevertheless yield substantial and empirically verifiable insights into molecular evolution.

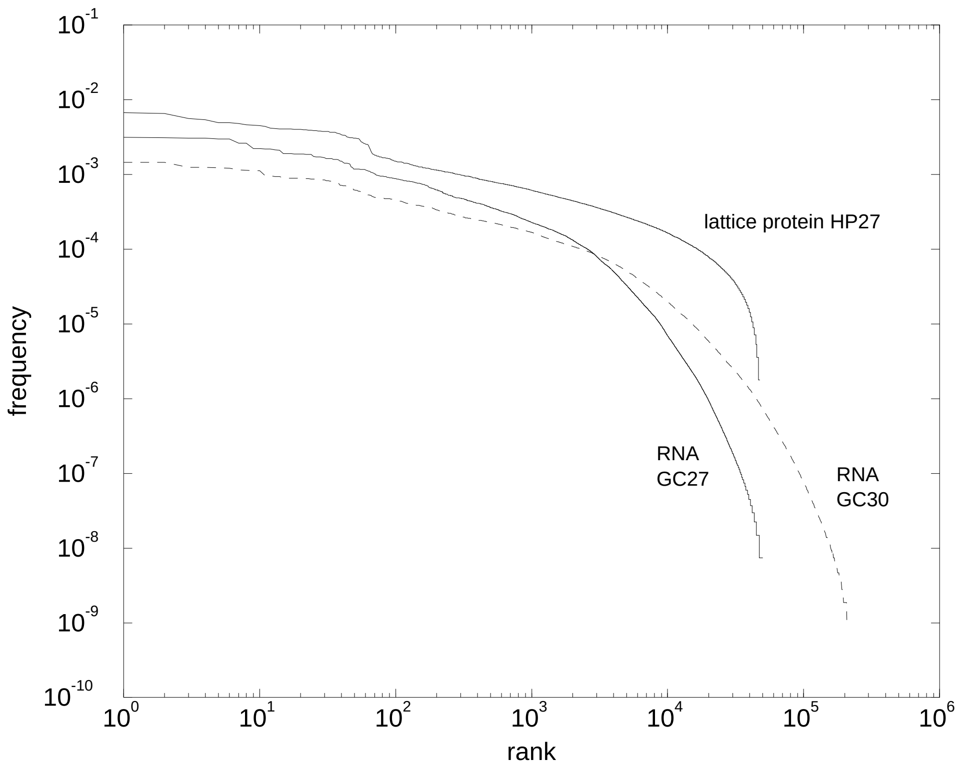

Figure 1: Curves labelled "RNA GC27 (GC30)": For each of the 2^{27} (2^{30}) possible sequences of length 27 (30) made of G and C the minimum free energy secondary structure was determined using an algorithm based on empirical energy data [6, 7]. "Frequency" denotes the number of sequences N_s realizing a particular structure s divided by the total number N of sequences folded, $f = N_s/N$. Structure frequencies f are rank ordered, that is, the highest frequency has rank $r = 1$, the next highest has rank $r = 2$, and so on. The RNA data can be fitted well by a power law of the form $f(r) = a(b + r)^{-c}$ (Zipf's law). A qualitatively similar picture is obtained for long sequences over the full AUGC alphabet.
Curve labelled "lattice protein HP27": Compact ground state structures (filling a $3 \times 3 \times 3$ cube) of binary sequences of length 27 made from hydrophilic (H) and hydrophobic (P) units. Only sequences with a unique ground state are considered. Data from Figure 2A of [1], courtesy C. Tang.

References

[1] Li, H., Helling, R., Tang, C., and Wingreen, N. *Science* **273**, 666–669 (1996).

[2] Schuster, P., Fontana, W., Stadler, P. F., and Hofacker, I. *Proc. Roy. Soc. (London) B* **255**, 279–284 (1994).

[3] Grüner, W., Gigerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Stadler, P. F., and Schuster, P. *Monatshefte f. Chemie* **127**, 355–374 and 375–389 (1996).

[4] Tacker, M., Stadler, P. F., Bornberg-Bauer, E. G., Hofacker, I. L., and Schuster, P. *Eur. Biophys. J.* (1996). In press.

[5] Bonhoeffer, S., McCaskill, J. S., Stadler, P. F., and Schuster, P. *Eur. Biophys. J.* **22**, 13–24 (1993).

[6] Zuker, M. and Stiegler, P. *Nucleic Acids Research* **9**, 133–148 (1981).

[7] Jaeger, J. A., Turner, D. H., and Zuker, M. *Proc. Natl. Acad. Sci. USA* **86**, 7706–7710 (1989).

Peter Schuster, Walter Fontana, Peter F. Stadler, and Ivo L. Hofacker
Institute for Theoretical Chemistry
University of Vienna
Währingerstrasse 17
A-1090 Vienna, Austria