Which Came First, Protein Sequence or Structure?

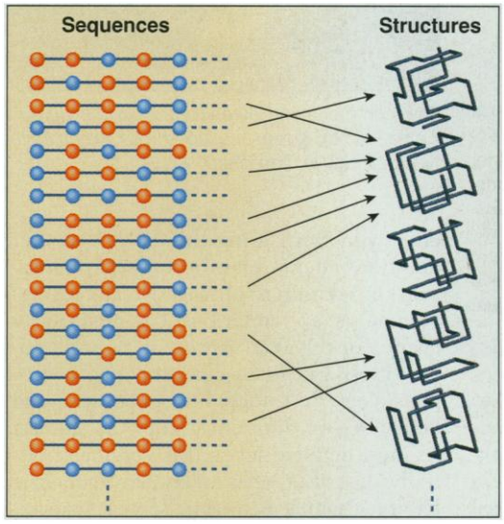
Mehran Kardar

Each protein is composed of a specific sequence of amino acids (its primary structure). However, its functionality is determined by its full three-dimensional structure (secondary and tertiary structure). The relation between the one-dimensional sequence and the final structure is part of the puzzle of protein folding. The possible number of sequences is enormous: 20^{400} potential proteins of length 400 can be constructed from 20 amino acids. Statistical mechanicians are thus intrigued by why only a very small fraction actually occurs in nature. One hypothesis is that the naturally selected sequences are special because they code for structures that have unique and stable native states, allowing for easy folding. On page 666 of this issue, Li *et al.* (*1*) provide an alternative perspective: The observed protein structures are special because they can be easily coded (designed) and are stable against mutations in the sequence.

The recent theoretical interest in understanding the folding of proteins is partly inspired by developments in the statistical mechanics of polymers and glasses. The hope is that some important emergent features of a complex system can be captured by simple models. Thus, rather than focusing on the chemistry of the amino acid side chains, simple models of proteins attempt to unravel organizing principles starting from simple interactions on a lattice. For example, considerable modeling has focused on polymers of length 27, occupying all sites of a $3 \times 3 \times 3$ cube. [This model was first suggested by Shakhnovich and Gutin (*2*) and is now considered standard.]

One principle that was embraced early on by the practitioners of this approach was that of "foldability." The great majority of sequences have multiple ground states and hence (assuming kinetic accessibility) may fold into different structures. Such sequences are unlikely candidates for coding functional (model) proteins. Potentially good sequences are those with a unique ground state, preferably separated by a large gap from the first excited state. The latter ensures the thermodynamic stability of the ground-state structure against thermal fluctuations and other perturbations.

Whereas "foldability" focuses on the sequence, selecting potentially functional ones, "designability," the principle introduced by Li *et al.* (*1*), is based on the structure of the resulting protein (see figure). This concept is quantified by measuring the number of sequences that uniquely fold into a particular structure (foldability is thus implicitly included). A great technical achievement of these authors is that they are able, for the first time, to compute the energies of all 103,346 structures, for all 2^{27} possible sequences of 27-mers (consisting of simplified sequences of polar and hydrophobic entities) on the cube. This exhaustive enumeration enables them to make a plot [figure 2 in (*1*)] of the designability of various structures. (Some structures are not designable as they do not correspond to the ground state of any sequence; the best structure is obtained from 3794 sequences.)



Designability of each structure is measured by the number of sequences that uniquely produce it as a ground state. Well-designed structures, such as the second from the top, have folds similar to real proteins. The red beads are hydrophobic and the blue beads are polar. (Courtesy H. Li, NEC Research, Princeton, NJ)

Several interesting patterns emerge from the enumeration. (i) At the tail of the distribution, there are structures that are highly designable: the number of sequences that fold into them is much greater than expected from simple probability distributions. (ii) These structures have, on average, a larger gap to their first excited state, making them thermodynamically more stable. Figure 3 of Li *et al.* (*1*) indicates that this gap suddenly jumps up beyond a designability of 1400 sequences, an unexpected and remarkable result. (iii) The well-designed structures are also more robust against simple changes in the sequence (random mutations). Thus, a major claim of the Li *et al.* paper is that the designability principle unites several properties (thermodynamic stability and mutational plasticity) occurring in real proteins.

In a partial answer to the question of "why proteins look like proteins," the authors find that well-designed structures have subunits and symmetries reminiscent of the secondary and tertiary structures of real proteins. (The secondary structure refers to the intermediate organization of proteins into subunits such as α helices or β sheets.) The intriguing similarity between some local motifs of two-dimensional compact polymers and the secondary structures in three dimensions was pointed out by Yee *et al.* (*3*). Some imagination is necessary to see these patterns in the small 27-mers, and it is more usual to compare the similarity in complexity and topological structure of proteins and 27-mers (*4*). Thus, the direct association of designability to the structural elements such as "superfolds" (*5*) is a bold new direction for protein-folding studies.

Although one needs to be cautious of results obtained from small chains, the apparent coincidences uncovered so far suggest that there may be some deep truth in the designability principle. Like most fertile concepts, it immediately suggests various tests and experiments to check its viability. Given a statistical mechanician's mistrust of tails of distributions, it is important to check that the results are independent of such details as the particular type of interactions employed by Li *et al.* (*1*). Recent work finds that the designability of a conformation does depend on the nature of interactions between monomers (*6*). While similar interactions lead to similar degrees of designability, different interactions yield different patterns. A major success for "designability" would be to generate new superfolds that guide in unraveling the structures of real proteins—an ambitious, but perhaps not unrealistic, goal.

The author is in the Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: kardar@mit.edu

References

1. H. Li, R. Helling, C. Tang, N. Wingreen, *Science* **273**, 666 (1996).
2. E. Shakhnovich and A. Gutin, *J. Chem. Phys.* **93**, 5967 (1990).
3. D. P. Yee, H. S. Chan, T. F. Havel, K. A. Dill, *J. Mol. Biol.* **241**, 557 (1994).
4. J. N. Onuchic, P. G. Wolynes, Z. Luthey-Shulten, N. D. Socci, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3626 (1995).
5. C. Orengo, D. T. Jones, J. M. Thornton, *Nature* **372**, 631 (1994).
6. V. S. Pande, A. Yu. Grosberg, T. Tanaka, in preparation; *J. Chem. Phys.* **103**, 9482 (1995).