New and Notable

The Inverse Protein Folding Question and Simulated Molecular Evolution

Lynda B. M. Ellis
Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, Minnesota, USA

The inverse protein folding question put simply is: how does one design sequence(s) that will fold to a desired structure? Empirical methods use secondary structure prediction in reverse to design desired structures, and short α-helical and other polypeptides have been synthesized using this approach (for example, Hill et al., 1990). More theoretical methods include the contact-lattice of Yue and Dill (1992) or the 3-D profiles of Bowie et al. (1991), which attempt to distill the essence of a given three-dimensional structure into a format that can be used to design or predict that structure.

A quite different method is that of molecular evolution, the change in molecules as a result of mutation and selection. To simulate molecular evolution, it is relatively easy to introduce random mutations in a DNA or protein sequence, but it is harder to do so in a realistic manner. It is even more difficult to develop a reasonable selection algorithm. For example, if we wish to "evolve" any given protein domain, what are the important attributes of primary sequence that produce such a domain? And how do we judge whether one sequence is closer to the target domain than another? Reliable simulated evolution of protein domains could be one approach to inverse protein folding.

Schneider and Wrede (in their article in this issue) have applied neural network techniques to this question, with interesting results. They start with the goal of evolving a 12-residue *Escherichia coli* peptidase cleavage site, as defined by known leader sequences from 24 *E. coli* protein precursors. The amino acid properties selected to model this site were hydrophobicity, hydrophilicity, polarity, and side chain volume. The neural network (or rather networks, since the final result was a modular multi-network system) was first trained to predict leader sequences using these four properties. Then its output was used as a measure of leader-site quality in the selection phase of simulated evolution.

In the mutation phase, each new residue is a function of the old residue, a position-specific mutability, a guassian-distributed random number, and a selected amino acid distance metric. Five different distance metrics were tested; a context-specific matrix designed specifically for peptidase cleavage sites, based on the same four properties used to train the neural networks, led to the best (highest quality) sequence within a fixed number of optimization cycles.

The highest-quality leader sequence thus produced, in the words of the authors, "can be regarded as an idealized sequence representing the 'optimal' amino acid motif." Much work remains before such a claim can be substantiated for this or any other predicted sequence. And a 12-residue sequence is only the first step to a larger motif or a complete domain. However, if this or alternative techniques can be developed that produce optimal motifs and domains, the inverse protein folding question will be answered.

Received for publication 30 November 1993 and in final form 30 November 1993.

0006-3495/0/02/275/ $2.00

REFERENCES

Bowie, J. U., R. Luthy, and D. Eisenberg. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science (Wash. DC).* 253:164–170.

Hill, C. P., L. W. Anderson, W. F. DeGrado, and D. Eisenberg, 1990. Crystal structure of α1: implications for protein design. *Science (Wash. DC).* 249:543–546.

Schneider, G., and P. Wrede. 1993. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: *de novo* design of an idealized leader peptidase cleavage-site. *Biophys. J.* 66: 335–344.

Yue, K., and K. A. Dill. 1992. Inverse protein folding problem: designing polymer sequences. *Proc. Natl. Acad. Sci. USA.* 89:4163–4167.