

## PROTEIN BIOCHEMISTRY

## Tackling biology's big question

Working from opposite ends of the protein folding problem, two research teams have developed powerful mathematical strategies that offer the potential to greatly clarify the relationship between primary sequence and native structure.

For years and years, the greatest question in structural biology has remained: 'how is all the necessary information specifying native protein structure contained in its primary amino acid sequence?' Because there is no satisfying solution, structural biologists spend months or even years crystallizing a protein and determining its structure. Protein designers must refine multiple generations of structures until the desired fold and function is obtained.

The major stumbling block is that protein structures are extremely complicated to mathematically model because of the sheer number of interactions between amino acids. Recently, two groups reported strategies to address this 'numbers' problem. David Baker and colleagues at the University of Washington describe a new computational method to predict high-resolution structures (Bradley *et al.*, 2005), and Rama Ranganathan and colleagues at the University of Texas Southwestern Medical Center show that artificial proteins can be designed using principles of cooperative evolutionary conservation (Socolich *et al.*, 2005).

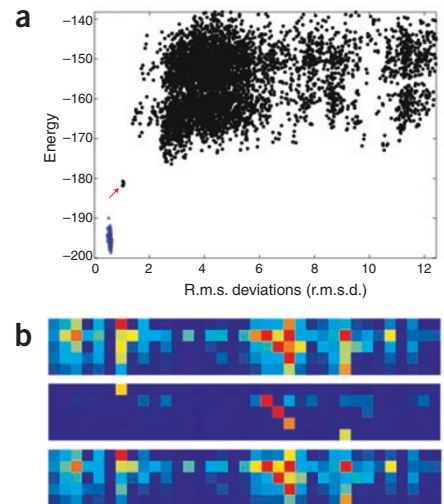
As structural biologists increase their understanding of protein folding, computational biologists improve their predictive algorithms. But because potential folding space is so enormous, it must be constrained to reduce the calculation to a reasonable timescale. Unfortunately, limiting the search often means that the true energy minimum is overlooked. Baker describes this dilemma with an analogy: "Imagine an explorer landing on a planet and having to find the lowest elevation point. If they land on the wrong continent, they'll never find it." To avoid this problem, Baker and colleagues predict low-energy conformations for several sequence homologs,

which are mapped to the target protein. "By using many explorers, we can search many different landscapes, and it is likely that at least one of them will find a minimum pretty close to the true minimum," says Baker.

Even with this advance, further refinement of the low-energy conformations is time-consuming. The Baker lab does have an interesting solution to this problem, however, in the form of Rosetta@home, a distributed computing project (<http://www.boinc.bakerlab.org/rosetta>), to which people from all over the world have donated time on their personal computers. Although this computational method can predict the structure of small, simple proteins such as ubiquitin (Fig. 1a) with high accuracy, Baker hopes that ultimately, they will be able to predict any protein structure.

Approaching the protein folding problem from the opposite end, the Ranganathan lab is interested in elucidating the key intramolecular interactions of specific folds that will allow them to design artificial functional proteins. Whereas traditionally researchers have used consensus sequences as scaffolds for new structures, Ranganathan and colleagues concur that the specific interactions between these conserved residues are more important for encoding a particular fold (Fig. 1b). "We know that both the stability of proteins and their function depend on cooperative interactions between amino acids," says Ranganathan. "There are networks of mutually evolving amino acids that are strongly associated with the core function of a protein family."

By using statistical coupling analysis on multiple sequence alignments of the three-stranded  $\beta$ -sheet WW (Trp-Trp) domain family, they showed that it was possible to design artificial proteins that fold into functional WW domain structures based only on the evolutionary conserved coupling of amino acids. This finding was unexpected, even to Ranganathan, who says, "The information content of protein sequences is sur-



**Figure 1** | Mathematical approaches to the protein folding problem. (a) Energy sampling of ubiquitin starting from an extended chain (black; red arrow, lowest energy structure) and starting from a native-like structure (blue). Reprinted with permission from *Science*. Copyright 2005, AAAS. (b) Evolutionary statistical coupling matrices for five positions (rows) in the WW domain for natural sequences (top), consensus sequences (middle) or sequences based on coupled conservation (bottom). Reprinted with permission from *Nature*.

prisingly low, which indicates that there are a vast number of degenerate solutions for building protein folds."

Whether one's interest is in predicting protein structures or designing new proteins, these two groups demonstrate that improved computational searching methods and an appreciation of evolutionary conservation should help us better understand the relationship between primary sequence and native structure.

Allison Doerr

## RESEARCH PAPERS

Bradley, P. *et al.* Toward high-resolution *de novo* structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).

Socolich, M. *et al.* Evolutionary information for specifying a protein fold. *Nature* **437**, 512–518 (2005).