

SEPTEMBER 22, 2005

Protein Construction May Be Governed by Simple Rules

Howard Hughes Medical Institute researchers have discovered that all the necessary information to sculpt a protein into its proper shape and function is contained in a relatively simple set of instructions encoded in its amino acid sequence.

The proof comes in a pair of papers published in the September 22, 2005, issue of the journal *Nature* from researchers in the laboratory of Rama Ranganathan, a Howard Hughes Medical Institute investigator at University of Texas Southwestern Medical Center. Ranganathan and his colleagues first used an algorithm to identify the key pattern of amino acid interactions in a particular protein family. Then, they used this pattern to build a dozen synthetic proteins from scratch that reproduced the form and function of their natural counterparts, at least in test tubes.

Scientists have known that the sequence of amino acids in a protein tells it how to fold and what to do. But the new findings may help explain where the necessary information is in the sequence and may help explain how a protein can simultaneously tolerate many mutations without apparent harm, but be crippled by a single mutation at certain sites. And the method suggests that, if biologists follow the same few simple rules as evolution, creating complex proteins may not be quite as complicated as it might seem.

"The main point of these papers is that a rather simple set of sequence rules suffices to specify a protein family."

- Rama Ranganathan

"The studies indicate that the number of crucial interactions in a protein may be smaller than previously thought—a boon for those who want to design novel proteins from scratch to fulfill a specific function," writes Jeffery Kelley of the Scripps Research Institute in an accompanying commentary in *Nature*.

"The design of artificial sequences having the capacity to fold into stable proteins with desired functions has been the holy grail of protein engineering for many years," write Robert Smock and Lila Gierasch of the University of Massachusetts, Amherst, in a perspective in the September 23, 2005, issue of the journal *Cell*. "From a protein engineering standpoint, the approach has great promise."

Ranganathan did not set out to build artificial proteins. He is interested in learning how nature designs proteins naturally through the evolutionary process of random variation and selection. But rebuilding a protein was the best way to assess a remarkably simple evolutionary hypothesis he discovered in genome databases.

"It will be really important to test whether these artificial proteins can work in living cells, embedded within the signaling networks that they were selected for through evolution," Ranganathan said. "But the main point of these papers is that a rather simple set of sequence rules suffices to specify a protein family."

Six years ago, Ranganathan and his colleagues reported a way to extract the key amino acids by looking at the record of successful products of evolution. Modern proteins are patchworks of modules, swapped among proteins over time, that do similar jobs of binding specific targets, catalyzing certain reactions, and sending signals.

Ranganathan reasoned he could compare a family of modules from several species to find which combinations of amino acids were evolutionarily selected over many millennia after they branched on the tree of life. To do so, he wrote a computer program that computes correlations between amino acids among the similar modules that have stood the test of time.

Surprisingly, only a few amino acids seemed to be tightly constrained and linked over evolutionary time. The highly constrained positions were biologically important; the protein was significantly affected when those few amino acids were mutated. Other parts of the protein were different in that they seemed to have evolved independently from each other, and were relatively unaffected by experimental mutations, he said.

Ranganathan and his colleagues found that sometimes, functionally important parts of a folded protein that are far away from the apparent center of action are detected by the evolution-based algorithm.

"A large body of work shows that proteins have a dense and local pattern of structural interactions between amino acids," Ranganathan said. "But in the evolutionary data, we see a sparse but distributed architecture of amino acids." This new way of looking at the problem could contribute to protein design efforts by providing the natural architecture of amino acid interactions. Currently, most protein design algorithms attempt to optimize all

the atomic relationships in a crystal structure of a protein, he said.

To test the global implications - that not all amino acids in a protein have equal significance for its structure and function - the Ranganathan group found the critical amino acid networks in a family of proteins and used them to build working synthetic proteins from an otherwise random sequence of amino acids. It promised to be a labor-intensive, time-consuming project. At first, no one in his lab wanted to tackle it.

To start, Ranganathan selected the WW domain family, a small module used by proteins from many diverse species to bind to target sites on other proteins and mediate protein-protein interactions. The WW domains all fold into a simple shape that resembles a hand cupped to hold water. In people, a mutation in the WW domain hinders the ability of a protein called *nedd4* to bind to an ion channel and put it where it is needed in a kidney cell, which can lead to a serious disease called Liddle's syndrome, he said.

In experiments described in the first article in *Nature*, Michael Socolich, a senior technician, and Steve Lockless, a graduate student, directed the analysis of 120 WW domains in organisms ranging from yeast to humans, using all the sequences available in the genome databases at the time. From this evolutionary analysis, they found the global pattern of conserved and tightly coupled amino acids.

From there, they created four libraries containing the blueprints for the WW domain, synthesized all the proteins in the libraries, and tested which ones folded correctly.

The first library contained a random assortment of natural WW domains to test for the innate folding rate of their experimental method. About two-thirds of the proteins folded properly. The second library contained totally random sequences from the WW alignments. None of the resulting proteins folded, as expected.

In a third library, they instructed the computer to preserve all the amino acids that had been conserved over time, but withheld the information about which amino acids needed to be coupled. None of these proteins folded properly. But in the fourth library, containing computer-generated sequences that preserved the few key longstanding amino acid relationships, about one third of the proteins folded.

The first paper proved that a few key amino acids can cooperate to control the shape of a protein. "It begins to explain how a sequence can diverge and retain the essence of a structure," Ranganathan said. "But nature builds function, not structure."

In the second *Nature* paper, postdoctoral fellow William Russ and colleagues expanded their analysis to 292 WW domains now in the growing genome

databases and tested the ability of the folded artificial WW domains to bind to the same spectrum of peptides as naturally occurring WW sequences.

When the team examined the co-evolving amino acid network in the WW domain, it contained residues on the backside of a molecule linked to the active binding site. Distant as it is from the active binding site, the residue appears to participate in the peptide binding, mutation experiments showed.

Together, the papers show that conservation plus coupling information is necessary and sufficient to recreate members of a specific protein family. At this stage, the researchers can not distinguish the rules for structure from the ones for function, or if they differ. Nor do they know the physical mechanisms underlying the statistical sequence rules.

So what about all the amino acids excluded from the co-evolving network? "I see them as the engine for forward evolution," he said. "The independence of local interactions gives the protein its capacity for novel variation. The protein conserves a small core network of key interactions which by themselves guarantee its function. That liberates the other positions to vary independently. In this way, proteins created by random mutation become robust to random perturbation. It's good to have tolerance for the process that created you."