Trying to Crack the Second Half of the Genetic Code

Inspired by practical problems in biotechnology and medicine, researchers are attempting to figure out the rules that govern protein folding

RONALD Schoner and his associates at Lilly Research Laboratories in Indianapolis tried to produce bovine growth hormone by inserting the gene into bacteria and getting the bacteria to synthesize huge quantities of the protein. But instead of getting the nice soluble protein they wanted, they got a clumpy mess.

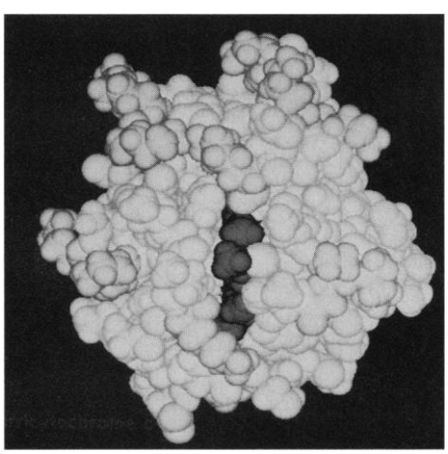
It is a problem that is plaguing biotechnology firms. The companies want to make known proteins and they want to make new hybrid proteins, such as monoclonal antibodies hooked to toxins to destroy cancer cells. Yet, says Irwin Kuntz of the University of California in San Francisco, "the outcomes are not always as they expected."

The reason these firms are having such difficulty is that molecular biologists have not yet cracked the second half of the genetic code: what are the rules that determine how a linear amino acid sequence will fold into a protein? The first half of the code—deciding how a sequence of DNA bases is translated into a sequence of peptides—was fairly straightforward and was reported more than two decades ago. The second half has been unsolved for so long, says Jonathan King of the Massachusetts Institute of Technology, "that for a long time people forgot that it was a problem." Now it is becoming hard to ignore.

Not only biotechnology firms but also molecular biologists are pushing to learn protein-folding rules. The difficulty is that DNA sequencing methods are far ahead of the study of protein structure. Molecular biologists are quickly getting the sequences of thousands of genes. Now they would like to know what the genes code for and what the gene products look like. For a typical protein, this means deciphering why some sequences of a polypeptide chain fold into an α-helical conformation, other regions form a β-sheet, silk-like conformation, and still other sequences form turns and loops. And, in addition, it means understanding how all these structures that make up a protein pack together.

Finally, there is an increasing realization by some physicians who study genetic diseases that some inherited disorders may be caused by defective dynamics of protein folding. The more that is known about protein-folding rules, the better these diseases will be understood.

The disease connection arises in studies of collagen disorders. Collagen, King notes, is the one protein for which the folding rules are relatively clear. The protein is a rigid rod made of three extended strands that are twisted together. The amino acid sequence consists of repeating units of three. At every first position there is a glycine, at the third position there is a proline or hydroxyproline 25% of the time, and in the second position, says King, "a great deal of variation can be tolerated." Glycine, the smallest amino acid, is irreplaceable because larger amino acids would prevent the protein chains from packing tightly together. The proline ring makes an extra covalent bond to the backbone of the chain, which keeps it straight and extended. But in the second position, says King, "the chains face outward, so there is lots of room" for a variety of different amino acids.



Science Source/Photo Researchers, Inc.

Protein folding patterns are hard to predict. *The sequence of amino acids determines the protein's three-dimensional shape, but rules for going from an amino acid sequence to a folded protein are unknown.*

When collagen is made, the chains line up and zip together. Anything that interrupts this dynamic zipping of the chains destroys the collagen structure. The results can be deadly because collagen is the single most plentiful protein in the body. It gives mechanical strength to skin and it underlies bone and teeth. It is, says Peter Byers of the University of Washington in Seattle, "what holds the body together."

Byers is now finding that patients with certain connective-tissue disorders, including osteogenesis imperfecta, Marfan's syndrome, and Ehlers-Danlos syndrome, have mutations in collagen genes that prevent this zipping up of the molecule. "I think the collagen story is very important," says King. But the rules for collagen structure cannot be extended to other proteins. In most proteins, "the chains change direction many times, giving the proteins their globular character," King points out. Collagen, with its triple-stranded rod, is truly in a class by itself.

Yet, stimulated by their pressing need to know protein-folding rules and by new techniques that may make the search for rules easier, researchers are starting to work on the problem of protein folding again, and some are optimistic that they will eventually solve it. "In principle," says Robert Baldwin of Stanford University, "we have the tools to solve the problem." Others, including Kuntz believe that there really is no exact solution. "Proteins generally have built-in lifetimes," he remarks. "They probably are meant to have alternative structures. If they are not engineered for maximum stability, then probably any one sequence may code for several structures."

At first, the protein-folding problem sounded easy. Biochemists knew that a protein's structure is determined by its amino acid sequence. So, it seemed, all that was needed was to analyze the relation between a protein's amino acid sequence and its final structure to deduce the protein folding rules. But it turns out not to be that simple. "The structures of hundreds of proteins are known to atomic dimensions, and the amino acid sequences of these proteins are known. How come we don't know the rules?" King asks.

Different researchers answer King's question in different ways. King's own response

is to suggest that proteins go through fleeting intermediate stages as they fold, and these intermediate stages may be quite different from the end product. Yet without knowing these intermediates, researchers can find it difficult, if not impossible, to predict the end product.

King gives the analogy of virus protein coats. "Many viruses have shells as their final structure. To make a shell, they first build a double shell—they make a scaffolding and then they remove it. If you only see the final structure, you would never guess there was a scaffolding." The common assumption that intermediates are less complex than the final structure is, in this case, fundamentally wrong.

If intermediates are so transitory, how do you trap one and determine its structure? Thomas Creighton and his colleagues at the Medical Research Council's Laboratory of Molecular Biology in Cambridge, England, found intermediates in the folding of pancreatic trypsin inhibitor because the intermediates have disulfide bonds that make these arrangements easier to isolate. Creighton learned, he says, that the three disulfide bonds in the final structure are not the same as the disulfide bonds formed as intermediates. The protein, he says, makes "wrong disulfides" and then removes them to form the correct ones.

It is like folding the flaps of a cardboard box, Creighton explains. "To fold them together, you have to put the flaps in a specific order and then distort them."

Now Creighton is collaborating with others in a number of different laboratories to try to learn the structure of the pancreatic trypsin inhibitor intermediates. He is using nuclear magnetic resonance because the structures are too flexible to form crystals.

Baldwin and his colleagues also are trying to trap protein folding intermediates, this time for the protein ribonuclease A. "We have reasonably good evidence that intermediates are observable and that they are hydrogen bonded," he says. "But we don't have much information on their structure." Baldwin is now using spectral methods, including nuclear magnetic resonance, to try to see these structures.

Another approach is to use x-ray crystallography to try to determine common structures in proteins. Proteins do not come in infinite varieties. So it is possible to do what Michael Rossman of Purdue University calls "protein taxonomy" to classify them by shapes. Rossman, for example, finds similar folds in a number of viral proteins—the proteins form a structure that Jane Richardson of Duke University describes as a jelly-roll. So Rossman and his associates are asking what is it about the protein sequences that leads to these similar folds. They are looking, says Rossman, "for the fingerprints of particular types of proteins."

King is one of a number of investigators using genetic methods to vary amino acids of a protein in order to determine which are important in controlling the folding process. "Our experimental data indicate that some amino acids are more important than others," he says. "If you try to determine folding rules without knowing which amino acids are most important, you are running a little blind," he says. King's group has characterized mutants that do not fold properly at high temperatures, but do when the temperature is lowered.

Protein folding is now a challenge that cannot be ignored.

Working with Lila Gierash and her colleagues at the University of Delaware, King's group is also looking at the effects of sequence alterations on the structures of small model peptides. "We're encouraged," says Gierash. "The [model] sequences seem to be associated with folding steps, and the genetic data are providing us with clues to those steps." Moreover, she adds, "King's data strongly suggest that local sequence alterations can influence the overall structure of proteins. We may be able to capture the rules."

C. Robert Matthews of Pennsylvania State University also is using specific mutants, this time to look at the final step in the formation of tryptophan synthetase. The protein has four α helices on its periphery. "It's a very common structural motif. Several dozen proteins are like it," says Matthews. Just before the protein takes its final form, "something happens to get one of the helices out of the way. The core of the protein rearranges and the helices snap back down." The question, then, is what does this final intermediate look like and how is it predictable from the protein's sequence? Matthews says he now has a mutant that seems to form a more stable version of this intermediate, which should enable him to examine the intermediate in detail. Still, Matthews notes, "we're looking at a step where things are pretty well organized. It's a final polishing step. We still don't know what happens in the early stages." Yet the methods he and others are using "in principle, could work" to get the full sequence of folding events, he remarks.

A final approach, which uses a combination of everything that is known about protein chemistry, is to try to get computers to predict protein structure. So far, the efforts are a very qualified success. "We don't believe models in the sense that you believe x-ray structures," Kuntz notes.

The effort began in the 1960's, when Harold Scheraga of Cornell University began using computers to try to decipher protein-folding information from amino acid sequences alone. But the project did not work, and Scheraga quickly learned why. The difficulty was that he was trying to decide which protein conformations were most likely by looking for conformations that represented minimal energy states. But there were a very large number of low-energy states possible and the computer simply could not pick through them. "There is no computer even today that could get you through it," says Kuntz.

Next, researchers realized that the problem could be greatly simplified. Rossman, Richardson, Michael Levitt of the Weizmann Institute in Jerusalem, and others discovered that they could classify proteins by shapes and that the same few shapes keep occurring over and over again. By then it was the early to mid-1970's. "What these ideas really led to was heuristic approaches," says Kuntz. "People began to attempt to build models based on what proteins look like."

A number of groups, including Richardson's, are testing their predictions by synthesizing defined amino acid sequences that ought to form, for example, β-helical barrels or α-helices, and determining whether their predictions are correct.

Another approach taken by several computer scientists, including Richard Feldman of the National Institutes of Health and Jonathan Greer of Abbott Laboratories, was to work on "protein extensions." The idea is that proteins with very similar sequences also have very similar three-dimensional structures. So they tried to determine what proteins should look like by comparing their sequences to sequences of proteins whose structures are known. This technique is now being used by investigators at industrial firms, including Genentech and Merck Sharp & Dohme.

At Genentech, for example, a group led by Ronald Wetzel starts with a protein whose crystal structure is known. In one case, they started with lysozyme, according to Dennis Kleid of Genentech. Then, says Kleid, "we change some of the amino acids and try to guess [by using computer programs] what the protein will look like. Our guesses are not always correct, but we're learning from that."

A second idea, developed by Frederic

Richards at Yale, Frederick Cohen at the University of California in San Francisco, Michael Sternberg at the University of London, and David Phillips at Oxford University, is to use what is called a "combinatorial tertiary structure." They try all possible ways to put together a protein from its amino acid sequence, based on rules saying when particular sequences are likely to form helices or pleated sheets, for example. Then they examine the resulting million or so structures with the computer. Most of these potential structures are completely unreasonable. They sift through the remaining reasonable ones by using all they know about the biochemical function of the protein to decide which of the structures is most likely correct.

For example, Cohen and Sternberg tried this method for the protein myoglobin. Out of a million possible structures, only 100 were reasonable. But only two of those reasonable structures could possibly be correct because only two could bind heme groups, as myoglobin does.

The two approaches, says Kuntz, "are not a wonderful success, but they are certainly the best thing going." Kuntz, Cohen, and their colleagues Robert Langridge and Thomas Ferrin are now starting a research program that will combine protein extensions and combinatorial tertiary structures for computer predictions of protein structures.

So the work continues. Everyone thinks that the protein-folding problem is worth pursuing, and even the most optimistic see no solution immediately in sight. "Nobody seems to be on the road to where they can say they will do it in 3 to 5 years," Creighton observes. "At present, there is not enough progress to say it is solved even to a first approximation," says Baldwin. "What is needed now is luck and very clear thinking." ■ GINA KOLATA

Do California Quakes Portend a Large One?

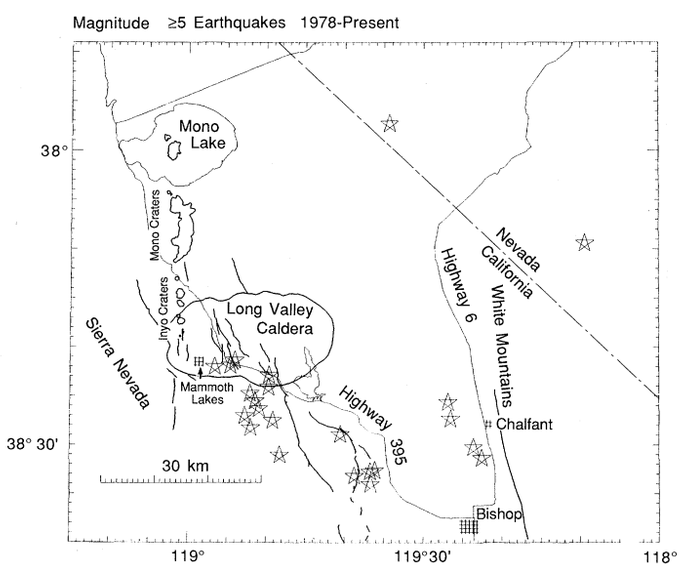
Far from the San Andreas fault, three lines of evidence hint at a large earthquake striking within the next decade

AMONG the flurry of earthquakes last month in California, seismologists took particular interest in the sequence of quakes in Chalfant Valley near the Nevada border east of Yosemite National Park. At first glance, the other shocks implied little about future events, but these eastern California earthquakes strengthened an already suggestive argument that a large earthquake of magnitude 7 or almost 8 could hit California soon. Rather than striking the closely watched San Andreas fault, the expected shock would occur on the less closely monitored but far less densely populated California-Nevada border. One of California's three great earthquakes in historic times struck just to the south in Owens Valley in 1872, followed since by three large events to the north. Geophysicists think they see signs that the next in the sequence could strike at anytime.

The case for expecting a large earthquake in the near future depends on the application of three relatively standard forecasting techniques to a new sort of locale. The most extensively applied forecasting technique is the recognition of a seismic gap—a fault section waiting to break in an earthquake. In this case the suspect fault section is near the White Mountains, just north of Bishop, California. This is about midway in a zone of seismic activity extending from southern California, where the North American and Pacific plates are sliding past each other, to north-central Nevada, where the Great Basin is being stretched apart.

Although not a single, well-defined fault like the San Andreas, this seismic belt was broken by earthquakes of magnitude 6.8 and greater in 1872, 1915, 1932, and 1954, activity rivaling that on the San Andreas. Only two unbroken sections remain among these failed sections, the largest being the 130 kilometers of the belt along the White Mountains. By analogy with the way earthquakes completely rupture the faults along coastal Mexico and Japan section by section, Robert Wallace of the U.S. Geological Survey (USGS) in Menlo Park has suggested that there is a high potential for a major earthquake in the White Mountains seismic gap. Because the interval between breaks has ranged from 22 to 43 years and the last break was 32 years ago, the next break could come at anytime, Wallace reasons.

The Chalfant Valley earthquakes of last month focused attention on the White Mountains seismic gap because they enhance a pattern of moderate seismic activity encircling that gap that is familiar elsewhere as a harbinger of a large earthquake. In 1983 Alan Ryall of the Center for Seismic Studies in Arlington, Virginia, who was then at the University of Nevada, and his colleagues pointed out that since 1978 the level of moderate seismic activity in the general area of the gap had been 20 times that during the previous decade. And that heightened activity seemed to be forming a partial circle about the gap. Such circles or doughnut patterns of moderate earthquakes had formed about the sites of future large events in Japan and elsewhere, as noted by Kiyoo



A Mogi doughnut?

The pattern of recent seismicity in the vicinity of the White Mountains resembles the roughly circular pattern, called a Mogi doughnut, known to precede some large earthquakes. The earthquakes of magnitude 5 and greater since 1978 include the four near Chalfant this July. Heavy lines denote faults active during the past 10,000 years or volcanic features.