

# NO ASSEMBLY REQUIRED

*Protein folding has perplexed generations of biochemists. Now the problem may be about to yield.*

BY GEORGE D. ROSE

**W**HILE THERE IS LIFE, THERE IS HOPE—so said Terence, and so said Cicero. But even more fundamental, wherever there is life, there are proteins. Your body contains roughly 100,000 kinds of them, half its dry weight, and they take part in every single one of your biological processes. They metabolize your food, define your form from skeleton to skin, transport oxygen and regulate respiration, arm your immune system, package and replicate your DNA, and serve as both signal and sensor for the

about at random. Return the temperature to a normal physiological level, however, and the shape will return, too, spontaneously, reliably and completely. That reversibility was revealed in 1957 in a classic experiment by the protein chemist Christian B. Anfinsen and his colleagues at the National Institutes of Health. Anfinsen and his team were studying the protein ribonuclease, an enzyme they extracted from cow pancreases. Through a series of delicate reactions, they dissolved the chemical bonds that held the molecule in its native conformation, until tests showed that

UNFOLD A PROTEIN, AND IT THRASHES AROUND AT RANDOM.

*But wad it up just right, and it becomes the stuff of life.*

network of chemical messages that interconnects your organs. And that is just a beginning; the list goes on. And on.

What role a protein takes in that grand biological opera depends on exactly one thing: its shape. For a protein molecule, function follows form. Unreeled, it is just a long string of amino acids. But wad it up just the right way, into a bundle known as its native conformation, and it becomes the stuff of life. The three-dimensional shape of a protein is a complex, exquisitely fashioned molecular ideogram involving interactions among thousands (often tens or even hundreds of thousands) of atoms.

Not surprisingly, such an intricate structure is highly susceptible to alteration. A slight tweak—a touch of heat, a few drops of bleach, a chance mutation—is all it takes to make a protein lose its potency. That is why you cannot hatch a boiled egg. It is also why people get Alzheimer's disease, Huntington's disease, cystic fibrosis, sickle-cell anemia and, probably, many forms of cancer. And yet a protein is also surprisingly robust. Within seconds after it is extruded from the cellular apparatus called a ribosome, it folds unerringly into its proper form. No blueprint guides it, and barring extreme measures, no edict can dissuade it. Just how a protein assembles itself into its native conformation is a mystery. Biochemists call it the protein-folding problem, and it is arguably the simplest yet deepest unsolved problem in biology.

Protein folding, of course, is no problem at all for a protein. It is more like a molecular reflex. Unfold the protein—say, by heating it carefully—and it will lose its conformational identity; the shapeless molecules just thrash

the ribonuclease was completely inactive. Then Anfinsen rinsed off the solvent and watched the reaction that ensued. Over the course of a day, the ribonuclease spontaneously regained both its structure and its activity.

The implications were profound. Until then biochemists had not thought of protein folding as a single chemical reaction. For all they knew, it might be as shadowy and complicated a process as is protein synthesis, which involves an entire cellular manufacturing plant culminating in a web of molecular assembly lines. Anfinsen showed that matters were much more straightforward than that. There were no cells in his beakers, only purified ribonuclease. No phantom forces could be at work; the secret of protein folding must reside within the chemical composition of the protein itself. Know what a protein is made of and, in principle, you know what it looks like.

Inspired by that prospect, investigators have spent the second half of this century searching for the rules that govern protein folding—the universal molecular grammar that will enable them to predict the structure of a protein from first principles. It has been a paradoxical quest; the same knowledge that makes the goal conceivable also drives home how hard it is to achieve. But now biochemists are starting to believe that the reach of theory will not forever exceed the grasp of method. There is a gathering mood of expectancy, a pervading sense that the problem is ready to yield. As one step toward a solution, my colleagues and I at the Johns Hopkins University School of Medicine in Baltimore, Maryland, have developed a computer program



endeavor worthy of the nineteenth-century French taxonomist Georges Cuvier—structural biologists such as Jane S. Richardson of Duke University in Durham, North Carolina, have been hard at work naming and categorizing the labyrinth of molecular structures.

**S**OME INVESTIGATORS HOPE THAT MOLECULAR taxonomy can help them predict protein folds. Because folds fall into major classes, those workers point out, perhaps they can function the way fingerprint types do, to match the amino-acid sequence of an unsolved protein against potential look-alikes in a data base. By “threading” the trial sequence along the folds of the known proteins, a computer could identify chemically plausible candidate structures. The bigger the library, the better the chance of finding a match. Such an empirical method is only as good as the algorithm that compares new sequences with the stored templates. And totally novel folds, of course, will always slip through the net.

To avoid the limitations of empirical methods, other investigators are seeking to predict protein structures from first principles. The key to the approach is energy. Every time you twist a protein chain, you set up a new pattern of attractions and repulsions among its atoms. That pattern determines how much energy is needed to untwist the protein. In theory, every protein engaged in folding should home in on one optimal energy—the lowest—just as a ball at the top of a hill will head for its lowest potential energy by rolling toward the bottom. (A real ball, of course, probably would not make it to the foot of the hill: it would roll into a ditch on the way. No one knows whether proteins can get similarly stuck in higher than optimal energy states, or whether they all reach absolute

**F**OR THE PAST DECADE MY COLLEAGUES AND I have tried to explain how the myriad forces within a protein—some local, some long-range—combine so that distant parts of a protein chain end up as next-door neighbors. Typically, we started with too many variables, too little insight and a few hunches, and many promising ideas failed to pan out. But even the false starts were instructive and helped lead to the work we are engaged in today.

First we focused on hydrogen bonds, relatively weak secondary links that form between certain atoms already bound in a molecular structure. Such bonds were Linus Pauling’s favorite candidate for the driving force behind protein folding. In 1992 the biochemist Douglas F. Stickle, now at Washington University in Saint Louis, Missouri, and I took a molecular inventory of the hydrogen bonds in forty-two proteins. More than two-thirds of hydrogen bonds, we discovered, connect an amide hydrogen in the backbone of one amino-acid residue with the carbonyl oxygen in the backbone of another residue, which may be many links down the chain of amino acids. When a protein folds, those distant residues come together in space in a unique pattern. But all amino-acid backbones are the same; with so many identical rivals to choose from, why should the distant residues always pair up in exactly the same way? Apparently hydrogen bonds between backbone atoms could not explain such juxtapositions. Hydrogen bonds involving side chains were just as disappointing. Almost all of them, our survey showed, are too short-range to account for big folds. Some other force must be at work.

Next we turned to the hydrophobic effect, the tendency of certain amino-acid side chains to shrink from water. The key to the “phobia” is carbon. Unlike the readily sol-

### TO THE TRAINED EYE, PROTEIN PATTERNS ARE AS PLEASING AS *the lines of a Mondrian painting or the Verrazano-Narrows Bridge.*

rock bottom.) That optimal-energy principle is just another way of putting the Anfinsen hypothesis, so it ought to work. But for reasons that remain unclear, the approach has met with only limited success to date, despite the best efforts of many outstanding physical chemists.

In my laboratory at Johns Hopkins my colleagues and I have performed many computer experiments aimed at coming to grips with the protein-folding problem. In our work we draw both on the empirical approach and on first principles. Our guiding principle is a characteristic of proteins that the biophysicist Gordon M. Crippen of the University of Michigan in Ann Arbor and I discovered independently in the late 1970s. The convolutions of protein molecules, we noticed with some surprise, are organized as an architectural hierarchy: domains can be divided into subdomains, which can be divided into sub-subdomains, and so forth. Consequently, not all chemical interactions in a protein have equal priority; at different stages of folding, some are much more important than others.

uble nitrogen and oxygen atoms, carbon atoms in proteins tend to move to the interior of the protein structure, where they shield one another from the surrounding water. That observation was the starting point for another molecular survey that I conducted, with Glenn J. Lesser, a hematologist-oncologist currently at the Bowman Gray School of Medicine at Wake Forest University in Winston-Salem, North Carolina. Lesser and I wanted to find out whether carbon atoms are more water shy in some amino acids than in others. If so, we reasoned, the biases could give important clues about the overall shape of the protein. Unfortunately, carbon atoms turned out to be more or less interchangeable: on average, they bury 87 percent of their available surface area, regardless of which amino acid they reside in. Another dead end.

Searching for broader structures that might explain folding, with the structural and molecular biologist Leonard G. Presta, currently a senior scientist at Genentech in San Francisco, I decided to take a closer look at the alpha-helix. The coils of a helix are held together by a regular, repetitive pattern of hydrogen bonds that connect each amino-

acid residue to a residue four places ahead of it and to another one four places behind it. Toward the ends of the coil, however, that pattern breaks down. The frayed ends, Presta and I conjectured, might contain signals that direct the backbone into nonhelical conformations. I set out to track them down, working with Rajeev Aurora, a molecular biologist at Johns Hopkins University; Edwin T. Harper, a biochemist at the Indiana University School of Medicine in Indianapolis; and Jeffrey W. Seale, a biochemist now at the University of Texas in San Antonio. We discovered that, like the finished ends of a hemp rope, many alpha-helices are whip-tied with chemical motifs

the heart of the energy function, the subroutine that decides whether a trial conformation is a “keeper.” The detailed energetics of proteins are enormously complex. To do them justice, LINUS would have to tease out every strand in an inconceivably tangled web of intermolecular forces. The key to a practical energy function, then, is knowing what to ignore. I worked on that problem with Eaton E. Lattman, professor of biophysics and biophysical chemistry at Johns Hopkins. We realized that we could save ourselves a lot of trouble by splitting the problem of protein folding into two questions. One is why a protein adopts its native structure instead of unfolding—the ques-

## A PROTEIN GETS ITS SHAPE FROM AN INCONCEIVABLE TANGLE OF FORCES.

*The key is knowing which of them to ignore.*

that supply extra hydrogen bonds. Such helix-capping motifs help explain why helices form in some segments of a protein chain but not in others.

FOR THE PAST TWO YEARS MY COLLEAGUE THE organic chemist Rajgopal Srinivasan and I have been developing a simple computer program to fold a protein from first principles. The program is called LINUS, at once an acronym for “local independently nucleated units of structure” and a tribute to Linus Pauling. Starting at one end of a protein, LINUS jiggles the protein chain a few residues at a time, twisting them at random into various poses. A subroutine calculates the energy of each pose; if one turns out to be clearly better than others, LINUS “freezes” the residues involved, locking them into position during future permutations. Then the program continues down the chain and repeats the process with another set of residues. After many thousands of repetitions, LINUS arrives at a shape with optimal energy—the conformation of the protein.

LINUS repeats its inspection tour of a protein many times, each time calculating the energy over a slightly wider range of residues. Because of the multiple passes, the program favors structures resulting from short-range interactions—LINUS’s way of simulating the hierarchical nature of protein folding.

What makes it all possible is two drastic simplifications. One has to do with the way LINUS decides which folds to evaluate. In an ideal world, one with infinitely fast computers, LINUS would look at every possible fold and only then select the best one. In practice that is impossible; the numbers are just too large. Instead the program employs a sampling technique called a Monte Carlo method. At each point in its path down the protein, instead of running through all possible twists and turns, LINUS picks a few of them at random. If a pose looks good to the energy function, LINUS keeps it—but the program also keeps a few “bad” ones in reserve, just in case the “good” poses lead to dead ends down the line. In that way LINUS can reliably ferret out key conformations without having to test them all.

The other simplification is even more radical. It lies at

tion of stability. The other is why it folds the way it does—the question of specificity.

The stability of a protein can be defined as the amount of energy needed to unfold it, just as the stability of a chair can be defined as the effort needed to tip it over. The specificity of a protein is more like the external architecture of a house. The two questions are quite different. My cedar-and-glass contemporary and your Brooklyn brownstone may have similar stabilizing features (foundations, beams, and so on); but adding another cross-member to further stabilize my house will not make it look more like yours.

IN WRITING LINUS, SRINIVASAN AND I REASONED that evolution has already taken care of the problem of stability: any proteins we encounter *must* be stable, or they would not exist. Consequently, LINUS ignores stability and focuses all its efforts on specificity. The payoff for that selective ignorance is that the energy function can be heretically simple. It need not calculate the energy of a fold to two decimal places, or even to one. All it needs is four basic rules:

1. Two atoms cannot be in the same place at the same time.
2. Each hydrogen bond counts as one energy point.
3. Amino-acid side chains are classified as hydrophobic (water-avoiding and oil-seeking), hydrophilic (water-seeking) or amphipathic (mixed). Each interaction between hydrophobic side chains gets two points; between hydrophobic and amphipathic side chains, one point; all others, zero.
4. In real proteins, amino-acid residues are not perfectly flexible. Side chains and other obstacles get in the way, with the result that a residue finds certain angles off limits. In LINUS’s energy function, amino acids that stray into such “underpopulated regions of conformation space” are penalized one energy point.

In short, almost everything LINUS does is based on rules of thumb and extremely rough approximations. How well does our simpleminded protein generator stack up against the real folds of actual proteins? Very well indeed. LINUS correctly, albeit imprecisely, predicts much of the secondary structure, supersecondary structure and large fragments of tertiary structure—not bad for a young program. It is an exciting time for Srinivasan and me, though we never lose sight of the tendency of the protein-folding

problem to humble its most ardent devotees.

**L**INUS IS NOT THE SOLUTION TO THE protein-folding problem. Indeed, it is not even clear what a solution would look like; people working on the problem disagree. Some investigators insist on nothing less than the ability to predict, for any protein, a conformation that rivals the accuracy of a high-resolution X-ray structure. Others might settle for a sketchy bird's-eye view of the protein backbone wending its way through space. The molecular biophysicist Frederic M. Richards of Yale University, codiscoverer of the structure of ribonuclease, has a keen sense of the field and of the disparate goals of the people in it. At a recent meeting on protein folding he predicted, with characteristic wit, that progress in the next several years would convince 50 percent of us that the problem had been solved.

If Richards is right, many biochemists and molecular biophysicists now working will witness a solution during their scientific careers. It would be hard to overestimate the importance of such an event. A solution to the folding problem will drive activities ranging from the Human Genome Project to rational drug design, protein engineering and nanotechnology. In the long run, it could even change the concept of *self*. Cognition and consciousness are conditioned inescapably by the material that houses them. Our proteins enable us to see but limit what can be seen. Ultimately, the divine fire that illuminates our lives is mediated by the chemistry of our proteins. They are us. ●

---

*GEORGE D. ROSE is a professor of biophysics and biophysical chemistry at the Johns Hopkins University School of Medicine in Baltimore, Maryland. He thanks Rachel Povereny for her contributions to this article. This article is dedicated to the memory of Christian B. Anfinsen, who died on May 14, 1995.*