

## Structural Genomics, Round 2

As NIH plans to extend its high-speed structural biology program for another 5 years, researchers remain divided on how to best allocate its shrinking budget

Five years ago, facing some opposition, the U.S. National Institutes of Health (NIH) in Bethesda, Maryland, launched an ambitious effort that some have compared in scale and audacity to the Human Genome Project. Its ultimate goal: to obtain the three-dimensional structures of 10,000 proteins in a decade. Like the genome project, this effort, called the Protein Structure Initiative (PSI), could transform our understanding of a vast range of basic biological processes. And just as the genome project attracted debate and dissent in its early days, the initiative split the structural biology community. The effort is now approaching a critical juncture, and the debate is heating up again.

The project is nearing the end of its pilot phase, a 5-year effort to develop technologies that has begun to transform labor-intensive, step-by-step procedures into a production-line process. Now, the initiative is poised to move into the production phase, dubbed PSI 2. In the next few months, NIH is expected to designate three to five centers, each of which could receive grants of about \$12 million a year to crank out protein structures at an unprecedented clip. It will also pick a handful of smaller labs to work on problems that have so far proven difficult to solve, such as how to obtain the structures of proteins embedded in cell membranes. Officials at the National Institute of General Medical Sciences (NIGMS), which is bankrolling the initiative, are reviewing proposals for the two types of grants, and the winners are expected to be announced this summer.

But, in a debate eerily similar to the one that roiled the genome community a decade ago, structural biologists are divided on how fast to proceed—especially in the light of constraints on NIH's budget. The central issue is whether the technology is far enough along to justify the move to mass production, or whether the emphasis should continue to be on technological development.

Brian Matthews, a physicist at the University of Oregon, Eugene, and chair of PSI's external advisory board, argues that the time is ripe to move ahead in cataloging thousands of new structures. "This information will be broadly applicable to biology and medicine," he says. Raymond Stevens, a structural biologist at the Scripps Research Institute in La Jolla, California, agrees that "the technology that has come out so far has been truly impressive." But he has strong

reservations about PSI 2's planned emphasis on mass-production of structures. "It's premature to start production centers until better technologies are in place," Stevens says.

This is not just an academic debate. The PSI could determine whether a key goal of structural genomics is achievable: the development of computer models to predict the structure of a new protein from its amino acid sequence. The initiative could also provide insights into how proteins interact to choreograph life's most fundamental processes and help researchers identify important new drug targets.

### Picking up the pace

In one respect, the scientists who planned the human genome project had it easy. Gene sequencing relies chiefly on one technology: reading out the string of letters in DNA. By contrast, producing protein structures requires mastering nine separate technological steps: cloning the correct gene,

overexpressing the gene's protein in bacteria, purifying it, coaxing it to form a crystal, screening out the best crystals, bombarding them with x-rays at a synchrotron, collecting the diffraction data as the rays bounce off the protein's atoms, and using those data to work out the protein's precise structure. (Researchers turn out a smaller number of structures using another technique known as nuclear magnetic resonance spectroscopy.)

Initially, the nine centers participating in the pilot phase of PSI had trouble dealing with that complexity (*Science*, 1 November 2002, p. 948). But structural genomics teams have now automated every step. "It took these groups a couple of years to get all the hardware in place," says Matthews. "But I think [the PSI's first phase] has been very successful."

Among the advances is a robot being built at the Joint Center for Structural Genomics (JCSG) in San Diego, California, that can run 400,000 experiments per month to find just the right conditions to coax given proteins to coalesce into high-quality crystals. Synchrotron facilities too have seen vast improvements in robotics. Setting up a crystal for measurement has historically been a cumbersome process, typically



**Pure speed.** Researchers at the Midwest Center for Structural Genomics use robotic gear to speed protein purification.

CREDIT: MIDWEST CENTER FOR STRUCTURAL GENOMICS

taking hours of fine-tuning. JCSG researchers and others have now created robotic systems to carry out this work, enabling data collection on up to 96 crystals without interruption. “That has been a tremendous benefit,” says JCSG chief Ian Wilson, a structural biologist at the Scripps Research Institute.

As the technologies advanced the centers accelerated their output. They produced 350 structures in PSI’s fourth full year, up from just 77 in the first year, and are on track to complete 500 this year. That pace is still well short of the initial goal of 10,000 structures in 10 years—that goal was little more than an optimistic guess, PSI leaders now say—but it’s a big step forward and should be fast enough to accomplish most of the effort’s scientific goals. Equally important, says John Norvell, who directs NIGMS’s PSI program, the average cost of each structure has dropped dramatically, from \$670,000 in the first year—a number inflated by the cost of purchasing and installing robotic gear—to \$180,000 in year 4. This year, Norvell expects that the cost will drop to about \$100,000 per structure. By comparison, he adds, traditional structure biology groups typically spend \$250,000 to \$300,000 for a structure, although some of the proteins they tackle are far more complicated than those PSI has taken on.

The types of proteins targeted by PSI are, however, one bone of contention. Traditional structural biology groups tend to go after similar proteins in important families, such as kinases, that participate in many biological pathways. And they often determine the structure of complexes of one protein bound to different molecular targets, in order to tease out the details of how the protein functions. As a result, 87% of the structures deposited in the major global protein database are closely related to those of other proteins.

The PSI, however, was set up to acquire structures from as many of the estimated 40,000 different protein families as possible. Indeed, 73% of the structures the PSI centers have solved so far have been “unique,” which by the PSI definition means that at least 30% of the gene sequence encoding a protein does not match that encoding any other protein. The idea behind casting such a broad net is to acquire structures from representatives of each family in the hope that this will enable computer modelers to predict the structure of other family members. Already, the data suggest that there is not as much

## A Dearth of New Folds

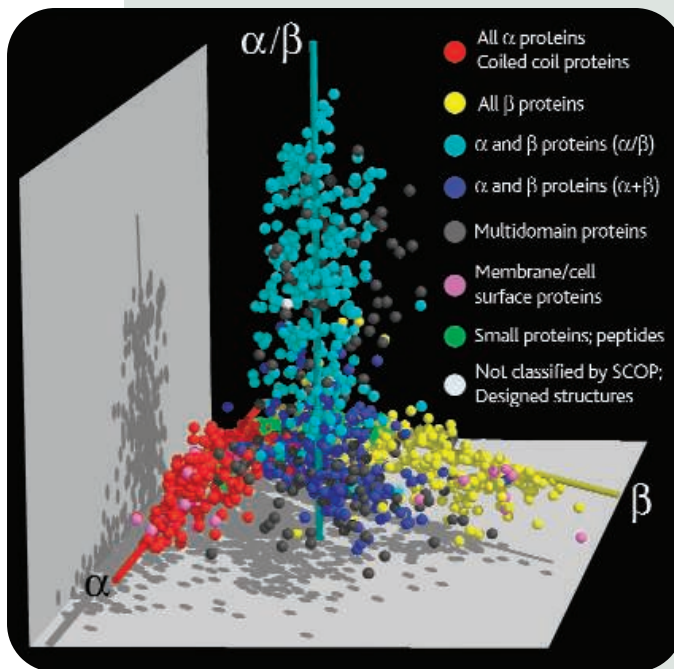
The Protein Structure Initiative (PSI) has already come up with one surprise: Proteins apparently come in a relatively limited variety of shapes. The initiative is targeting “unique” proteins, ones in which the DNA that encodes them differs markedly from that for proteins

with a known structure. Researchers expected that many if not most of those proteins would have structural patterns never seen before, but the vast majority look quite familiar.

The general shape of a protein once it assumes its three-dimensional (3D) form is known as a fold. So far PSI groups have found that only 12% of their completed structures sport new folds. “The number of folds will be considerably less than previously thought,” says Ian Wilson, a structural biologist at the Scripps Research Institute in La Jolla, California, and head of its Joint Center for Structural Genomics. This means that proteins with vastly different patterns of amino acids adopt similar 3D shapes. That, Wilson says, is critical information for computer modelers working to predict the structures of proteins based only on their DNA sequence.

Researchers are also mapping

**Protein landscape.** This graph reveals how proteins cluster into four structural classes.



out how all these unique proteins relate to one another. In a report published online on 10 February by the *Proceedings of the National Academy of Sciences*, researchers at the University of California, Berkeley, and the Lawrence Berkeley National Laboratory (LBNL) in California compared nearly 2000 different protein structures, calculating the difference in shape between each protein and all of the others in the collection. They then graphed the results, showing similar structures as close to one another. They found that the global protein structure landscape is a bit like the cosmos, where galaxies cluster together amid vast regions of emptiness.

That map does have sharp features, however, says study author Sung-Hou Kim, an LBNL structural biologist and the head of the Berkeley Structural Genomics Center. It shows the four main classes of protein structures—known as  $\alpha$  helices,  $\beta$  strands, and proteins with mixtures of  $\alpha$  and  $\beta$  domains called  $\alpha+\beta$  and  $\alpha/\beta$ —as four elongated arms emerging from a common center. The map, Kim says, suggests that much of the protein structure space is empty because proteins with certain shapes are architecturally unstable. That in turn suggests that structural genomics groups are unlikely to find any new structural classes of proteins. Says Kim: “I would be very surprised if they did.”

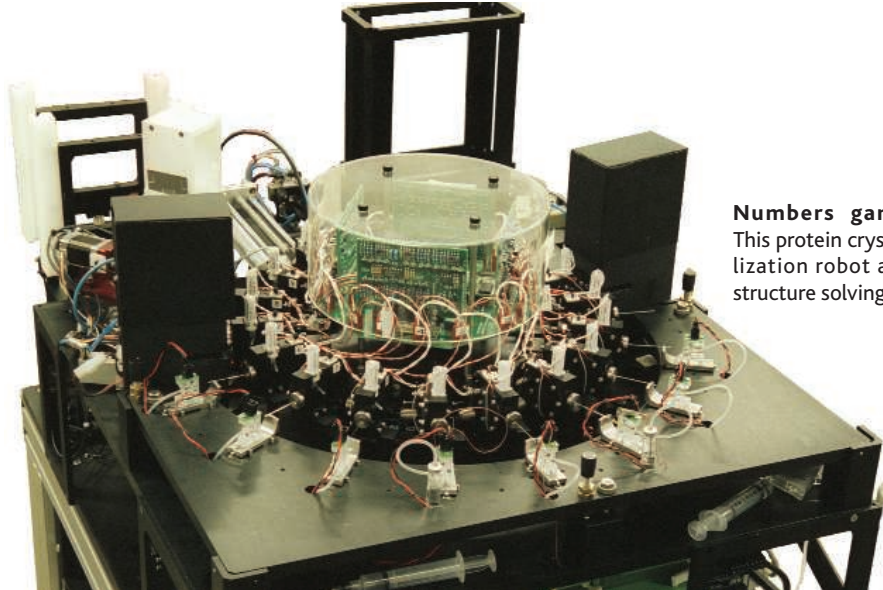
—R.F.S.

structural variation between families as many biologists expected (see sidebar).

Some structural biologists argue, however, that this approach has limited value, and that the tens of millions of dollars currently going to structural genomics centers would be better spent on traditional structural biology groups. Yale biochemist Thomas Steitz, for example, says that most of the structures PSI groups have produced so far are “irrelevant” to understanding how the proteins work because they are not

bound to their targets. The PSI focuses on bacterial rather than eukaryotic proteins, he also complains.

Berg acknowledges that “tension certainly exists,” between traditional structural biologists and structural genomics groups. Although some PSI 2 centers will likely focus on producing structures of protein complexes and eukaryotic proteins, he notes that NIH’s structural genomics effort was never set up to go after the same type of information as conventional structural biol-



**Numbers game.**  
This protein crystalization robot aids structure solving.

ogy. Rather, the goal was to explore the far reaches of the protein landscape. “To my mind the most important message is structural genomics and structural biology are largely complementary and synergistic,” Berg says.

Berg and others add that the PSI has already provided numerous important biological insights. For example, the Northeast Structural Genomics Consortium (NESGC) recently solved the structure of a protein that adds a methyl group to ribosomal RNA and in the process confers antibiotic resistance to bacteria. That structure, says NESGC director Guy Montelione, has suggested inhibitory compounds that could revive current antibiotics and spawned a separate research program on the topic. Another structure revealed details of the way plants bind a signaling molecule called salicylic acid, challenging conventional wisdom on the functioning of plants’ immune systems. “Not only are we spinning out new science, but new science initiatives,” Montelione says.

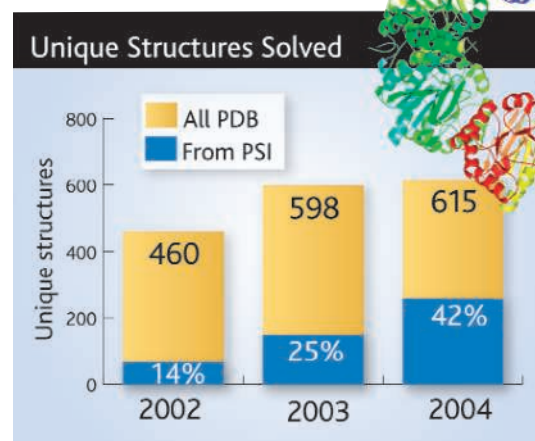
## Chapter 2

What comes next is, however, a matter of debate. NIGMS officials had expected to scale up a handful of the current PSI centers to full-scale production facilities and fund as many as six additional technology centers, each tackling a separate bottleneck.

A tight NIH budget has already forced NIGMS officials to rethink those plans, however. They had hoped to boost the current PSI budget of \$68 million to \$75 million next year, the first year of PSI 2. But they are now anticipating a decline in funding, to \$64.5 million. The cut is likely to force structural genomics leaders to rein in their goals, and ultimately it could extend the date by which they complete the program. “That’s clearly going to be a problem,” Matthews says.

Those budget cuts will also make it tough for PSI leaders to strike the proper balance between production and technology development in the next phase. Each of the existing pilot centers currently receives some \$8 million a year. The plan for PSI 2, Norvell says, had been to spend \$12 million a year on each production center. That means five centers would eat up nearly the entire budget for PSI 2’s first year, leaving little for technology development. If that happens, “I think we’ll regret it in 5 years,” Stevens says.

Stevens points out that many technical problems remain. For



example, even though PSI centers have increased their output of protein structures, their success rate in turning targeted genes into solved structures has remained essentially unchanged. At each stage in determining a structure—cloning the gene, expressing the protein, and so on—researchers take a hit. For example, only 57% of cloned genes are successfully expressed as proteins, and of those, only 28% can be purified. “It’s like doing chemical synthesis” that involves numerous steps, says Wilson. “If you have a 90% success rate at each step, that’s not going to give you much material out at the end.”

In the end, Stevens notes, only 2% to 10% of the proteins targeted by PSI centers wind up as solved structures. In view of this “pretty poor success rate,” Stevens argues that the phase 2 efforts should focus more on technology development. “I think structural genomics can do even better if the technologies are allowed to mature further,” he says.

Not many of Stevens’s colleagues agree. “Clearly we have to capitalize on the production centers we’ve already invested in,” says Wilson. Thomas Terwilliger, a structural biologist at the Los Alamos National Laboratory in New Mexico and head of the TB Structural Genomics Consortium, adds that the limited success rate isn’t a major issue because if one protein in a family doesn’t yield a structure, researchers can typically find another one that does. Furthermore, Montelione points out that the new production centers will spend about one-third of their funds on improving the technology. Stevens counters that “there will be so much pressure to produce structures that any technology developments will take a significant back seat to the structure focus.”

Berg says “it’s hard to imagine funding fewer than three of the large-scale [production] centers.” At \$12 million apiece, that would still leave \$28.5 million—more than \$4 million for each of the six proposed technology centers. The balance between production and technology development is “still very much up in the air,” says Berg, and will depend on the outcome of the reviews of the grant proposals. The NIGMS advisory committee will then decide which centers to fund in May and announce their decision in early July.

Whatever the outcome, it’s now unlikely that the PSI effort will achieve the initial goal of solving 10,000 protein structures by 2010.

With budget cutbacks and continued technical challenges, the final tally will probably be somewhere between 4000 and 6000, about the number that PSI leaders now believe computer modelers will need to accurately predict structures of related family members. Still, that means the program will solve structures for only a small fraction of the estimated 40,000 protein families. “This mixed bag of production and technology development will require another cycle, another 5 years to finish the job,” says Montelione. So will there be a PSI 3? That debate is just starting.

—ROBERT SERVICE