

STRUCTURAL BIOLOGY

Protein Structure Initiative: Phase 3 or Phase Out

The production-line approach to finding protein structures is rapidly filling up databases. But is it the data researchers want, and is it worth the cost?

In the early 1990s when structural biologist Andrzej Joachimiak was working in the labs of Paul Sigler and Arthur Horwich at Yale University, he and six colleagues worked together for more than 2 years to solve the x-ray crystal structure of a protein known as GroEL. To obtain such structures, researchers must arrange copies of a protein into the regular pattern of a crystal and then ricochet beams of x-rays off it to map out the position of each of the protein's atoms. At the time, the structure was hailed for offering a host of insights into how GroEL carries out its role as a "chaperone" helping other proteins fold into their proper three-dimensional shapes. But GroEL's large size made it a bear to solve.

Today, as head of the Midwest Center for Structural Genomics, a consortium of investigators at eight institutions in the United States and Canada, Joachimiak and his colleagues churn out some 180 such structures a year, an average of one every 2 days. Not all are as difficult as the GroEL structure, but Joachimiak estimates that recent technological

advances would allow them to solve something as complex as GroEL within about 2 months. That, Joachimiak says, "is a true revolution."

But some in the field say the revolution has gone far enough. Joachimiak's center is one of four high-throughput structural biology centers participating in the Protein Structure Initiative (PSI), a big-science project funded by the U.S. National Institute of General Medical Sciences (NIGMS), part of the National Institutes of Health (NIH) in Bethesda, Maryland. PSI is doing for protein structures what the Human

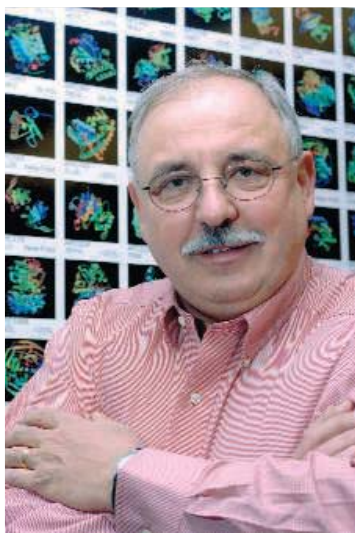
Genome Project did for sequencing: turning it into a mass-production exercise. Already, PSI's four main centers and six smaller ones have turned out nearly 3000 protein structures and over the past 7 years have contributed about 40% of all the novel structures deposited in the Protein Data Bank (PDB), a global repository for protein structures.

But with PSI now halfway through its second 5-year phase, critics say the cost of the program is too high. This year, NIGMS will spend approximately \$80 million on

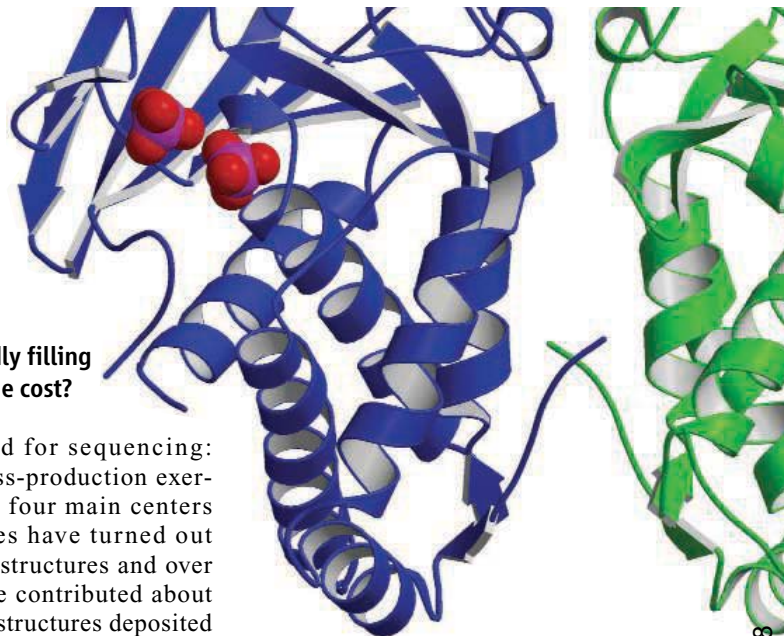
PSI. By the end of phase 2 in July 2010, the total tab will be more than three-quarters of a billion dollars. At a time when NIH funding is flat, many critics argue that the money is better spent on traditional small-scale structural biology projects, ones geared toward solving particular questions about the detailed working of proteins highly relevant to biology and medicine. In December, that message was underscored by an external review committee of prominent biologists charged with assessing PSI. Among the report's conclusions: "The large

PSI structure-determination centers are not cost-effective in terms of benefit to biomedical research." Structural biologist Gregory Petsko of Brandeis University in Waltham, Massachusetts, echoed the sentiment in an editorial last year in *Genome Biology*, in which he labeled PSI "an idea whose time has gone."

PSI proponents have plenty of counterarguments, and the debate shows no signs of waning. "It's a real hot point in the community," says Janet Smith, a structural biologist at the University of Michigan, Ann Arbor, who led the recent PSI review



Speedster. Andrzej Joachimiak and his colleagues at the Midwest Center for Structural Genomics are pushing the pace for solving protein structures.



panel. "It's a fairly contentious topic, and opinion tends to run high," she adds. In the midst of this debate, NIGMS officials will have to decide soon on PSI's fate. The current round of PSI funding is scheduled to run through July 2010. If agency officials want to continue uninterrupted funding for the project, they must send out a request for proposals sometime next year, according to NIGMS Director Jeremy Berg. That means they will likely need to decide by the end of this year whether PSI has a future. At this point, Berg says, funding for PSI 3 "is not a given."

A family affair

Whereas genomics can reveal the sequence of amino acids in a protein, structural biology tells us how that sequence folds up into a particular shape, which is key to a protein's function. These structures have long been seen as a treasure trove of information about life's molecular machines. By revealing structures through x-ray crystallography and nuclear magnetic resonance spectroscopy, structural biologists glean insights into how they operate. In some cases, those insights can discover the likely function of an unknown protein, lead to a deep understanding of how misshaped proteins cause disease, and potentially reveal a path to new drug treatments. For example, resolving the structure of the HIV-1 protease led to the creation of the first protease inhibitors used to fight AIDS.

Structural biologists have traditionally taken a hypothesis-driven approach to their science, asking questions about proteins known to be of interest. PSI, by contrast, chose a novel and somewhat controversial strategy: a "discovery-based" approach primarily targeted at proteins from different structural classes, or "families," throughout

CREDITS (TOP TO BOTTOM): NEW YORK STRUCTURAL GENOMIX RESEARCH CONSORTIUM, PSI-1; ARGONNE NATIONAL LABORATORY

Downloaded from www.sciencemag.org on March 25, 2008

Twister. The catalytic part of a human phosphatase enzyme.

the protein landscape. Members of each family fold up into similar shapes, often adopting similar functions, such as proteases, kinases, and phosphatases. One

major goal of PSI has been to obtain structures of representatives of as many of these families as possible, in particular the large families that have the most members. Proponents argue that each structure could be the key to many more: information on how the sequences fold into proteins should enable computational biologists to create “homology models,” detailed simulations of closely related family members for which no physical structure exists, and thereby glean insights into their function (see sidebar, p. 1612).

Success was far from certain. When the project started in July 2000, perhaps the biggest question was whether PSI centers would be able to automate all of the many steps involved in mapping proteins. Unlike genomics, which relies on speeding up one technology—reading the sequence of DNA’s nucleotide bases—PSI leaders had to speed up numerous technologies including cloning

genes into microbes, expressing and purifying proteins, coaxing them to form crystals, testing their quality, collecting x-ray data, and solving the structure. “Early on, we didn’t know whether we were going to be able to build these pipelines,” says Ian Wilson, a structural biologist at the Scripps Research Institute in San Diego, California, who heads one of PSI’s four large centers, the Joint Center for Structural Genomics in San Diego.

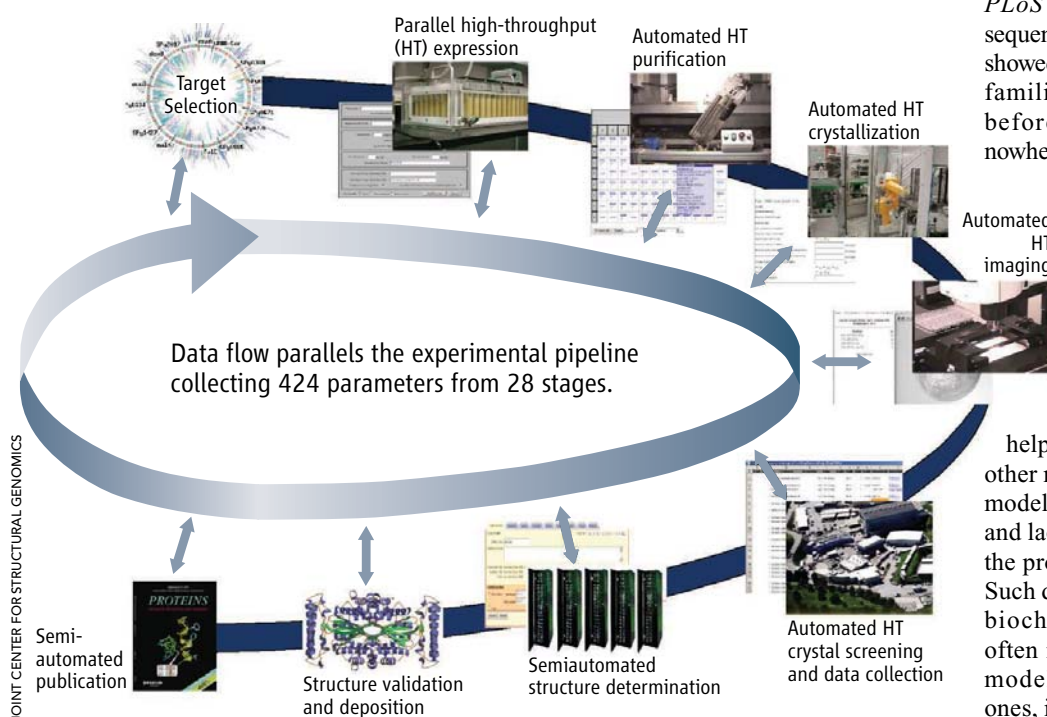
But the recent review panel concluded that PSI’s technology development had been “highly successful,” with advances dramatically speeding all phases of structure determination. In many cases, the panel’s report concludes, PSI has fostered technology that can be adopted by more traditional structural biology efforts. The centers not only developed new technology, they’ve applied it effectively, too: The large centers now crank out an average of 135 protein structures each per year.

PSI proponents argue that this production-line approach has dropped the cost of solving structures from about \$250,000 apiece in 2000 to about \$66,000 today. But PSI’s success is not just about the bottom line, they argue. It’s also revealing a diversity in protein structures never seen before. A 2006 analysis in *Science* by Steven Brenner and John-Marc Chandonia of Lawrence Berkeley National Laboratory (20 January 2006, p. 347) in California found that PSI centers account for about half of the novel

structures submitted to PDB. These are structures for which their immunocid sequence overlaps with that of any other proteins by less than 30%. Another study in the *Proceedings of the National Academy of Sciences* last year by Michael Levitt at Stanford University in Palo Alto, California, showed this trend continuing, and that PSI centers reversed an earlier steady decline among structural biologists in the number of novel structures being added to PDB. Wilson and several colleagues argued in an editorial in the January issue of *Structure* that the novelty of the PSI structures is a great benefit to the community because it provides data complementary to traditional structural biology rather than simply answering the same sets of questions.

But researchers are still divided over just how useful all this new information is. “I had reservations from the outset,” says Petsko, who says he objected because protein structures are only useful when they can answer specific biochemical questions about the detailed workings of a protein. The recent PSI assessment report echoes this criticism, calling PSI’s strategy of focusing primarily on novelty “seriously flawed.” One problem, Smith and her co-authors argue, is that the number of new protein families identified by gene-sequencing efforts worldwide continues to grow more rapidly than the number of protein structures being produced. A team of researchers reported last March in *PLoS Biology*, for example, that a random sequencing of DNA from the world’s oceans showed that more than half of all the protein families they found had never been seen before, suggesting that researchers are nowhere near completing their survey of the diversity of protein families. That makes the challenge of obtaining representative structures from each family “an open-ended problem,” say the authors of the assessment report.

What is more, the assessment panel concluded that although having a protein structure can help computer modelers make models of other members of that protein family, those models almost always have a low resolution and lack detail of the precise location of all the protein’s different amino acid residues. Such detail is key to nailing down the exact biochemical workings of a protein and often its specific function. “The ability to model structures, particularly complex ones, is very far from being able to connect most PSI structures to function,” the report states. Even if an accurate model can be



Building a bigger pipeline. PSI groups created a series of new technologies to speed up the many steps involved in determining a protein’s structure, such as robots to purify and crystallize proteins.

CREDIT: JOINT CENTER FOR STRUCTURAL GENOMICS

made, using that to discern a protein's function is not a straightforward task. A structure, Smith says, "is a little bit of data" that can be used to discern a protein's function. "But it's not as much as folks had hoped it would be."

On top of these problems, critics say PSI's data are not getting picked up by the

broader community of biologists. In part, they argue that's because only a relatively small fraction of this broader community knows how to use this type of structural information. The bottom line, Smith says, is that "the number of structures provided [by PSI] is not providing a boon to biology." By contrast, she adds, when the Human

Genome Project began to release its data, it was instantly seized upon: "There was no need to ask, 'Was this worthwhile?'"

Function follows form

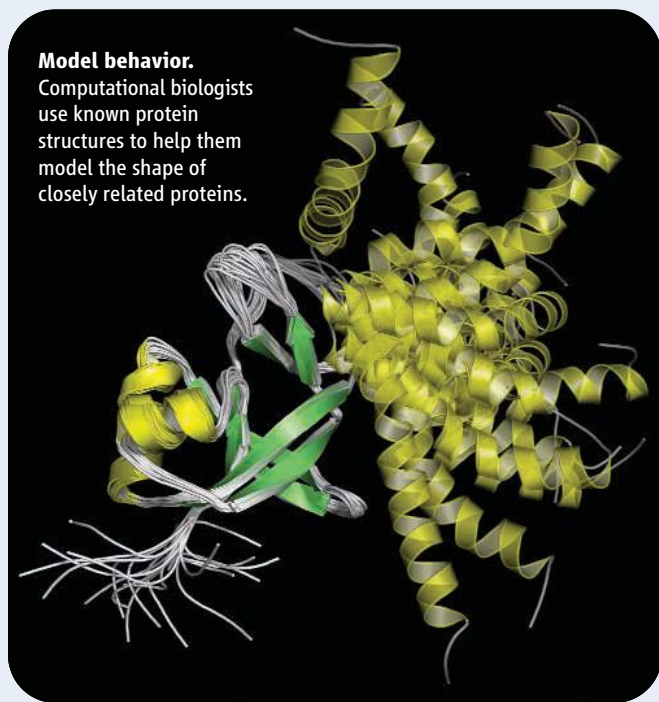
PSI leaders counter that although it's true that the number of protein families is growing rapidly, most of the newly discovered

Researchers Hone Their Homology Tools

The Protein Structure Initiative (PSI) is churning out new protein structures at a pace never seen before. But even the hundreds of structures the initiative unveils each year don't make much of a dent in the millions of proteins and multiprotein complexes thought to be out there. One hope for PSI, however, is that the proteins it has solved will give researchers insights into the structures and functions of some of those whose shapes are unknown. For such work, computational biologists employ "homology models," which use solved structures as templates for computer models of the three-dimensional (3D) shapes of closely related proteins.

Model behavior.

Computational biologists use known protein structures to help them model the shape of closely related proteins.



How well do homology models work? Not well enough, according to members of a review panel that issued a mixed report card on PSI in December. Although such models often get the general shape of related proteins correct, they typically lack the atomic-scale resolution needed to gain specific insights into how a protein does its job—or even what job it does. "The large numbers of new structures determined by the PSI effort have not led to significant improvements in the accuracy of homology modeling that would allow modeling of more biologically relevant proteins, complexes or conformational states," the report concluded.

But computer modelers say that conclusion misses the mark on several counts. First off, they point out, it was never a stated goal of PSI to improve the accuracy of homology models. "This was a complete red her-

ring," says John Moulton, a computational biologist at the University of Maryland Biotechnology Institute in Rockville. Moulton says the initial intention was simply to allow computational biologists to apply existing models to a larger number of target proteins. And that, he says, has undoubtedly occurred. Of all the structures submitted to the global Protein Data Bank, PSI now contributes about 40% of all the "novel" protein structures—those significantly different from any solved previously. And according to one recent estimate, those allow for the creation of more than 40,000 homology models that could otherwise not be made.

That said, Moulton and others argue that PSI is actually now beginning to contribute to the improvement of homology models themselves. In its second phase, PSI has supported two small centers geared toward improving computer models and has also supported individual computer-modeling groups. That bioinformatics support was perhaps "a little slow" in coming, says Andrej Sali, a computational biologist at the University of California, San Francisco. But he and others argue that this support, together with the increased number of structures, has helped spur advances in the basic algorithms to improve the accuracy of models.

Whether due to PSI or not, Moulton and others say there's plenty of evidence homology models are improving. For starters, they point to a biennial competition among computational biologists to predict the structure for a series of proteins. The Critical Assessment of Structure Prediction (CASP), which began in 1994, will hold its eighth competition later this year. The first "was embarrassing," says Moulton, who heads the CASP competitions. Few of the early models even came close to figuring out the actual structure of their target proteins, which were also simultaneously solved by x-ray crystallography for comparison. But by 2002, 60% of the models got close enough to the final structures to add useful information. By 2006, that number had climbed to 80%. "I don't want to say modeling was improving only because of the PSI," Moulton says. But the added structures in the database, he argues, are making a "very significant contribution." Adds David Baker, a computational biologist at the University of Washington, Seattle: "Homology modeling is definitely getting better."

In another key advance, improved computer models are making it easier for x-ray crystallographers to solve their structures. Experimentalists solve these structures by firing powerful beams of x-rays at protein crystals and tracking how those x-rays ricochet off their targets. These data give them much of what they need to nail down the position of all the atoms in the protein. But for a complete 3D picture, researchers typically compare the original data with another set taken from a closely related protein. Combining the two data sets is usually enough to finish the job. Not all proteins have close relatives that have been solved. But in a *Nature* paper last November, Baker's team showed that it was possible to use newer high-resolution homology models as the close relative to help researchers solve the x-ray structures. "It's not an established method yet," Baker says. However, he argues, it shows the synergy that can occur between high-quality experimental data and computational models.

—R.F.S.

families have only a few members. The majority of proteins are found in a small number of large families that are the focus of PSI's targeting. Gaetano Montelione, a structural biologist at Rutgers University in Piscataway, New Jersey, argues that as a result of focusing on large families the impact of PSI structures is increasing, because each solved structure carries more leverage, or ability to model a greater number of related structures, than those solved along traditional lines. Even with the limitations of current computer models, "the large information leverage provided by determining the first structural representatives from very large sequence families is tremendously enabling to biomedical research," Montelione writes in a February 2008 response to the PSI assessment report. Although homology models may not always reveal a protein's function, Montelione and others argue, in many cases it can offer important clues to guide future biochemical experiments designed to nail down that function.

Many researchers also dispute the claim that many PSI structures lack biological relevance. A fraction of PSI targets are chosen for biological interest. And any of its structures' relevance, as with the value of any basic research, takes time to grow, they argue. "The benefits we will see 2 to 3 years from now will be very great," Wilson says, and will include a growing understanding of how protein families evolved and the evolutionary connections between different families. And as for the dissemination of PSI data, PSI leaders say that a new knowledge base (kb.psi-structuralgenomics.org/KB) that came online earlier this month should improve matters dramatically.

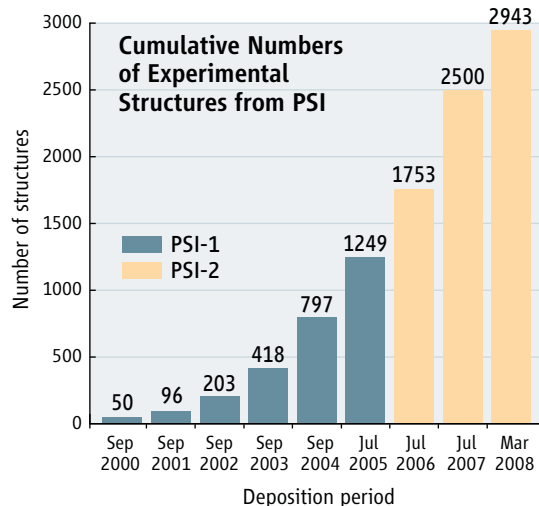
David Baker, a computational biologist at the University of Washington, Seattle, adds that the large number of PSI structures is also making possible an emerging approach to designing new therapeutics. Baker's group uses the full gamut of PDB structures to help them design better protein-based inhibitors to toxins, as well as vaccines for diseases such as HIV. When the group designs their proteins, they start with the shape of the target they are trying to block. Then they conduct a computer scan through all the known protein structures in PDB—including PSI structures—looking

for as many proteins shaped to fit into those targets as possible. Then they set their computational program loose to refine those matches and design a novel protein for an optimal fit. And the more close matches they have, the more accurate and effective the designed protein tends to be. As such, Baker says, the value of the database will only grow. "These structures are really going to help protein design," Baker says. "I don't think that was anticipated originally."

A question of value

Smith and others say they readily agree that PSI is producing good science, but they question whether it's worth the cost. "It's how do you get the most bang for your buck," says Philip Cole, a pharmacologist who specializes in signal transduction at the Johns Hopkins School of Medicine in Baltimore, Maryland.

Still, Cole and others worry that even if PSI isn't funded for a third phase, there's no guarantee that money saved



Off the charts. PSI centers are determining structures at a pace never seen before. But critics doubt that the impact has kept pace.

will flow to traditional structural biology groups. That's not how science funding works. "If PSI were to be discontinued, the money would go back to the general pool within NIGMS," Berg says. Structural biology funding, he adds, accounts for about 10% of the NIGMS budget, with



The messenger. Janet Smith led a recent panel of biologists that criticized PSI for having a limited impact on the broader community of biologists.

traditional single-investigator grants taking up about 6.3%. So doing away with PSI would likely increase the share of funding for individual structural biology grants from about 6.3% to perhaps 6.7%, Berg says. What is more, structural biologists currently working on PSI would then be competing for those funds. So the net result could wind up being "a pretty big negative" for the community, Berg says.

So what's next? Berg says NIGMS is currently evaluating all of its large-scale projects to decide which ones to continue. The PSI assessment panel argued against continuing the project in its present form. "Future effort might be focused on smaller projects with much higher experimental coupling to biological function and improving computational methods of analyzing and predicting protein structure," the report concluded. In his response to the report, Montelione agreed that connecting more directly with the priorities of biologists "needs to be a priority" in designing PSI 3.

Others agree that perhaps the best solution is to focus more tightly on protein targets with known biological relevance, such as multiprotein complexes, proteins that are embedded in cell membranes, and proteins from disease-causing microbes. "This can evolve," says Joel Sussman, a structural biologist at the Weizmann Institute of Science in Rehovot, Israel. "Now you can use these enormous platforms [built in the PSI centers] to tackle biological problems." Whether the broader biological community can agree on such a compromise will largely depend on whether NIGMS sees budget increases anytime soon. Says Wilson: "When the money gets tight, the knives come out."

—ROBERT F. SERVICE