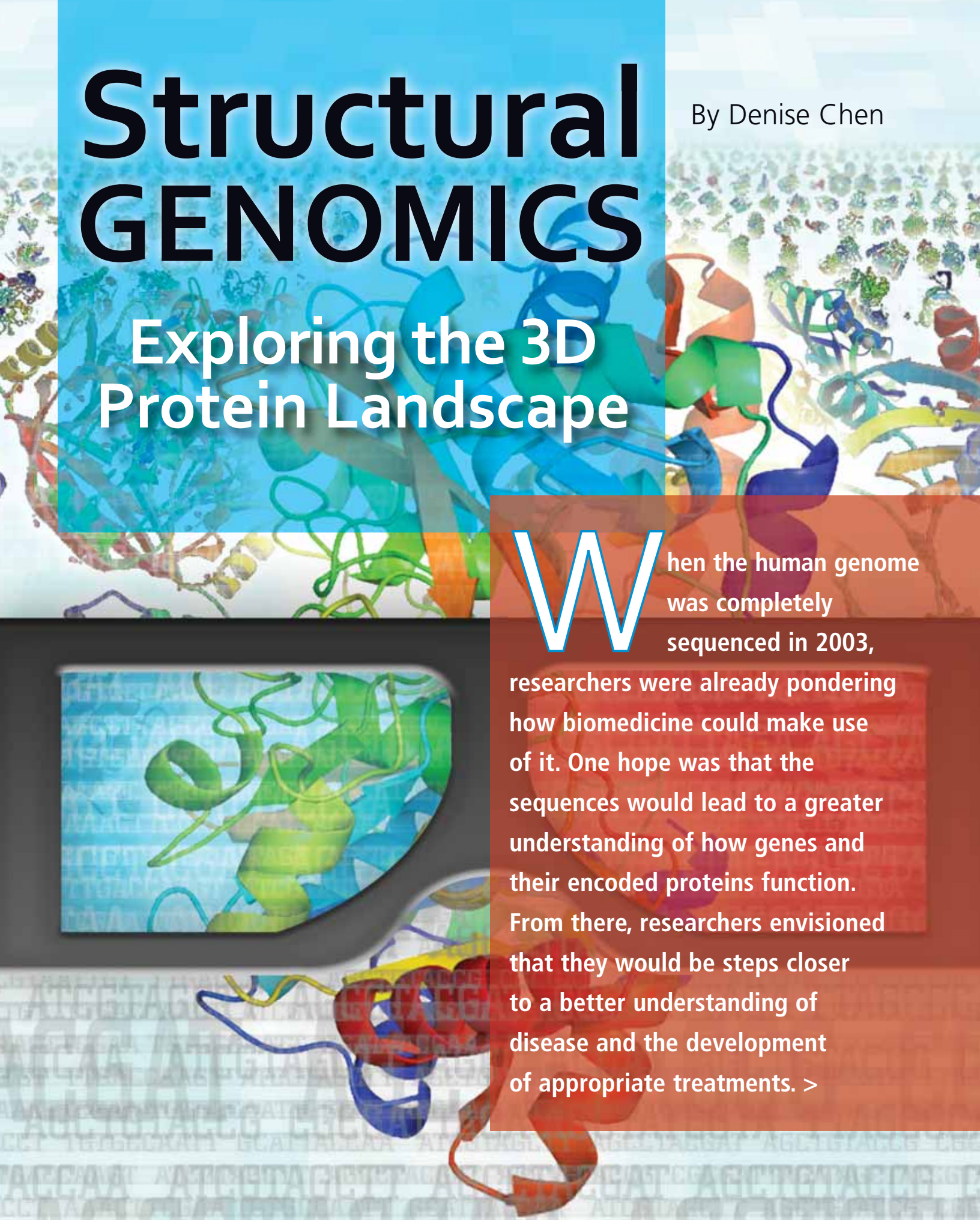
Structural GENOMICS

Exploring the 3D Protein Landscape

By Denise Chen



When the human genome was completely sequenced in 2003, researchers were already pondering how biomedicine could make use of it. One hope was that the sequences would lead to a greater understanding of how genes and their encoded proteins function. From there, researchers envisioned that they would be steps closer to a better understanding of disease and the development of appropriate treatments. >

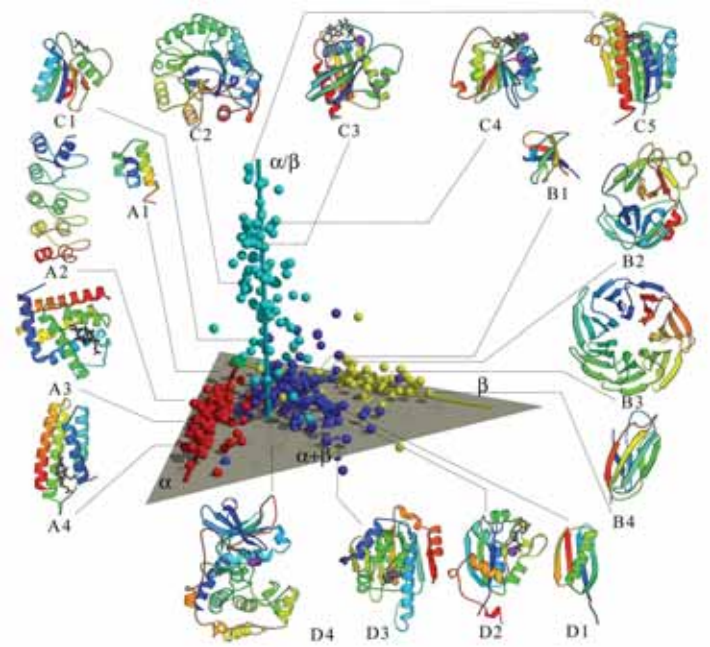
But a fuller understanding of proteins' functions within the human body depends on determining those proteins' structures. And as the number of known gene sequences grew, many scientists realized they could not catch up simply by determining protein structures one by one. So a group of scientists embarked on a strategic plan to uncover the three-dimensional structures of all the proteins that these genes encode.

This endeavor is called structural genomics. "The original question for which structural genomics came into being was: 'Can we translate the sequence of everything into the structure of everything?'" says **Peter Preusch, PhD**, acting director of the Protein Structure Initiative at the National Institute for General Medical Sciences (NIGMS).

The primary motivator of structural genomics is the sheer speed with which genomic sequence data is accumulating. Structural determination in traditional structural biology laboratories can't possibly keep up, researchers say. Fortunately, unlike sequences, which are nearly infinite in number, "there may be a finite number of different shapes that proteins actually adopt to perform their functions in the cell," says **Ian Wilson, DPhil**, professor of structural biology at The Scripps Research Institute and director of the Joint Center for Structural Genomics (JCSG).

In fact, a 1992 *Nature* paper estimated that the majority of proteins belong to no more than 1,000 families. Thus, researchers reasoned that it might be possible to unveil the universe of protein structures through a combination of experimental structure determination and computational structure prediction. And although upwards of 10,000 protein families have now been identified, uncovering the protein structure universe remains feasible.

In pursuit of this goal, ten years ago, NIGMS made a major investment to fund and spearhead a coordinated public structural genomics effort in the United States. Known as the Protein Structure Initiative (PSI), the program established four research centers and several specialized centers. The plan: to determine structures faster and cheaper; improve computational methods for predicting protein models; and ultimately develop innovative strategies for delivering useful structural information to the greater biological community.



Dimensions of the Protein Universe. Protein structures are displayed here along axes signifying secondary protein structure elements: strictly α helices or β sheets, both α and β, or combinations of α and β. The more complex and highly structured proteins reside at the extreme ends of the axes. In 1992, researchers estimated the number of protein families at around 1000, but the size of the protein universe has turned out to be much larger than predicted--as exemplified by the 23rd release of the Pfam database listing over 10,000 protein families. Source: NIGMS image gallery: http://images.nigms.nih.gov/index.cfm?event=viewDetail&imageID=2367. Courtesy of Berkeley Structural Genomics Center, PSI.

"The original question for which structural genomics came into being was: 'Can we translate the sequence of everything into the structure of everything?'" Preusch says.

In each of these areas, the PSI has made great strides. Before the PSI launched, determining the structure of a relatively complex protein was a major task, requiring the efforts of a graduate student for several months or even years, says **Keith Hodgson, PhD**, professor of chemistry at Stanford and head of the JCSG structure determination unit. Today, at each of the four main PSI centers, "a structure is turned out every few days," he says. At the same time, protein structure prediction helps fill in the gaps between known and unknown structures, bringing us closer to knowing the "structure of everything." This increased coverage of the structure space is transforming the field of biology, making it possible to assemble all of the structures in a particular pathway and visualize the interplay between them; or screen multiple structures to determine what they will bind; or carefully study the structures of proteins involved in disease.

Progressing through the Pipeline: From Sequence to Structure

Structure determination consists of multiple steps including cloning, expressing, and purifying a protein, finding appropriate conditions for crystallizing the protein, performing structural analy-

sis by techniques such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, and analyzing the resulting data. This can be a long and arduous process requiring an enormous investment of time, labor, and money with no guarantee of success. Thus, a key initial goal of the PSI was to make this process more efficient.

In the PSI's first five years (2000-2005, often referred to as PSI I: Pilot Phase) four large-scale pilot centers were created: the Joint Center for Structural Genomics (JCSG), the Midwest Center for Structural Genomics (MCSG), the New York SGX Research Center for Structural Genomics (NYSGXRC), and the Northeast Center for Structural Genomics (NESG). These centers developed high-throughput X-ray crystallography and NMR spectroscopy pipelines. Using automation and robotics, they consolidated and refined all of the individual protein production and structure determination steps. "It really is like a pipeline where you start at one end with a sequence, and out of the end of that pipeline comes a three dimensional structure," Hodgson says.

During the PSI's second five years (2005-2010)—known as PSI 2: Production Phase—the centers' pipelines churned out large quantities of previously unknown protein structures. Six specialized centers were also established to focus on the structural determination of proteins that are technically trickier to crystallize, such as membrane proteins, protein complexes, and eukaryotic proteins. The goal: to cover the "structure space."

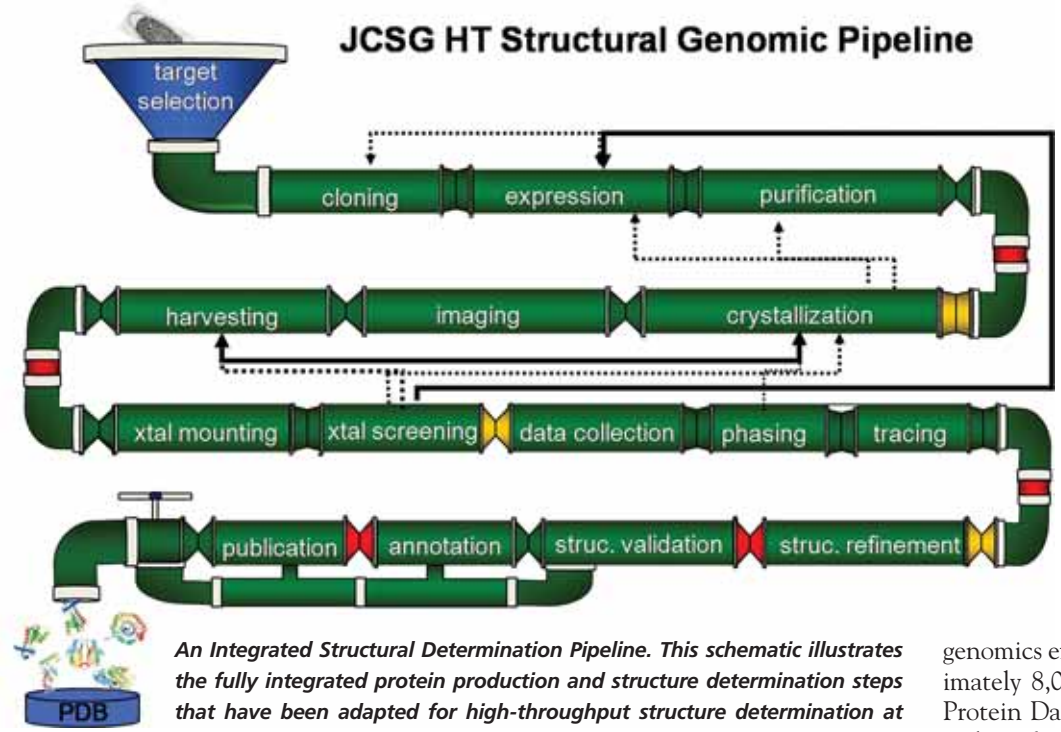
In picking protein targets, the PSI sought to complement what others were doing. "We've been trying to look into areas where nobody's really looking yet," Wilson says.

In picking protein targets, the PSI sought to complement what others were doing. "We've been trying to look into areas where nobody's really looking yet," Wilson says. Thus, novel protein targets that share less than 30 percent sequence identity with proteins of known structure comprise 70 percent of the focus at each center; the remaining targets are proteins deemed important by the biological research community.

To generate a target list of novel proteins, the PSI bioinformatics group used the publicly available Pfam (protein families) database and other automated protocols. "You've got to use informatics and computational biology to do structural genomics—first of all, to pick the right targets," says **John Norvell, PhD**, former director of the PSI at NIGMS. Pfam groups protein families by functional domains found within protein sequences. It uses protein sequence alignment entries and hidden Markov models to probabilistically determine how well a particular protein sequence matches with known families. In this way, PSI researchers identified protein sequences belonging to families with little structural information and targeted those for structural determination.

The approach has paid off. Working together to conquer the list of target proteins, the PSI centers reached their goal of solving more than 3,000 novel structures during PSI 2. In fact, over the last ten years, worldwide structural genomics efforts have deposited approximately 8,000 protein structures in the Protein Data Bank (PDB), the primary archive for structural biology.

Structure determination using PSI pipelines can now even be done by offsite researchers from the general science community. "You can run the synchrotron beamline, collect the data,



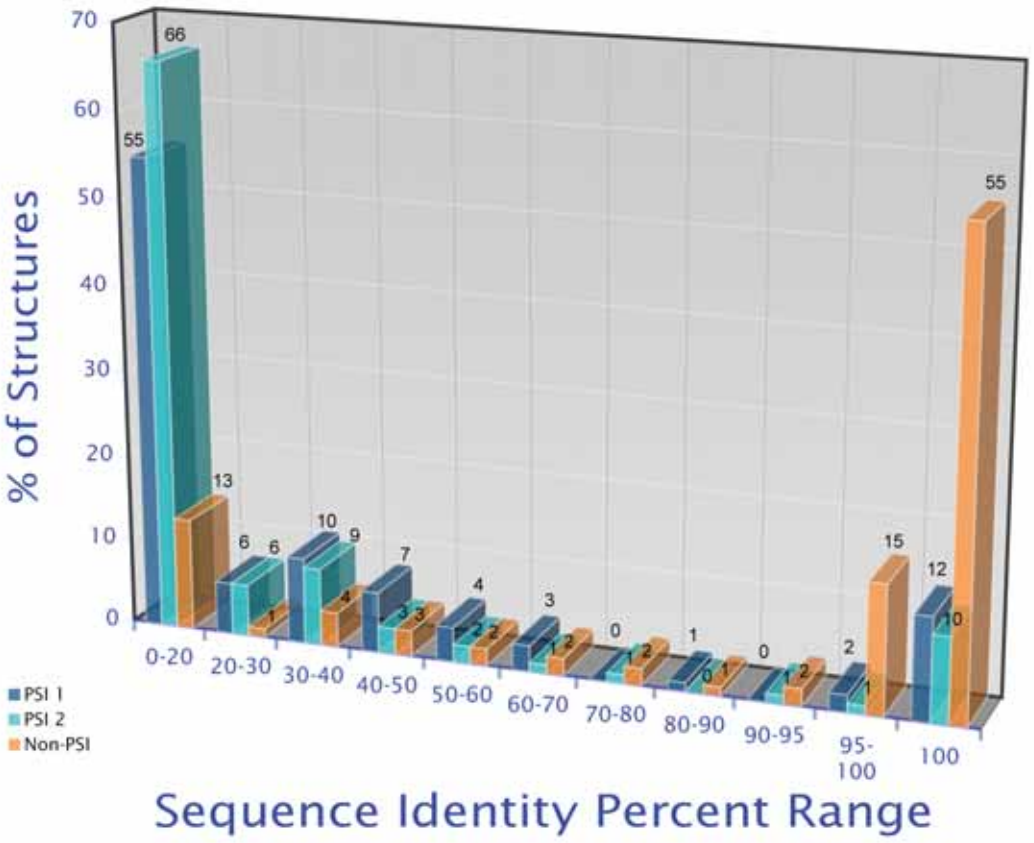
An Integrated Structural Determination Pipeline. This schematic illustrates the fully integrated protein production and structure determination steps that have been adapted for high-throughput structure determination at the PSI large-scale centers. All the steps of the pipeline are tied together by a common bioinformatics framework that enables feedback. For example, if the crystal screening step cannot identify usable crystals for structure determination, this information will be communicated to an earlier stage such as the crystallization step, where appropriate modifications will be made that help increase the procedure's likelihood of success. Courtesy of Marc Elsliger and Ian Wilson, JCSG.

"Ten years ago, there were people who would almost remember all the structures in the PDB. … But we're exactly in this stage where this type of old style analysis is no longer sustainable," Godzik says.

carry out the whole experiment from your own laboratory sitting in front of your own desktop or laptop," Hodgson says. "All you have to do is get your samples here and you can get FedEx to do that for you."

Where will scientists go from here? "Ten years ago, there were people who would almost remember all the structures in the PDB," says **Adam Godzik, PhD**, associate professor of bioinformatics at the Burnham Institute for Medical Research in La Jolla, California. "At some point this breaks down—you can memorize 300 structures, 500 structures, but you can't memorize 50,000. We're exactly in this stage where this type of old style analysis is no longer sustainable."

"What is still lacking are tools and, in some sense, even concepts of how to analyze large numbers of structures," Godzik says. Out of the small handful of structural alignment programs for doing just that, the Godzik group has written two of them, including the Flexible structure AlignmenT by Chaining Aligned Fragment Pairs with Twists (FATCAT) method. FATCAT improves upon its predecessor by accounting for structure flexibility and rearrangements. However, Godzik says, the structure alignment field is still very young and many concepts remain to be refined.



The Focus on Novel Families. This graph shows the percentage of structures contributed to the PDB by the PSI and other sources from 2000 to 2008. Each group of three bars represents how similar the sequences of the new contributions are to known structures. As shown here, the PSI determined the structures for novel protein families at a far greater rate than did other researchers during this time period. For example, of the total PSI deposits during this time (divided into the PSI1 and PSI2 phases, represented by the blue and aqua bars, respectively), most share less than 30 percent sequence identity with known structures (leftmost bars). On the other hand, over half of non-PSI deposits (represented by the orange bars) during this same timeframe had 100 percent sequence identity with known structures (rightmost bars). From "Investigators' White Paper" from the Future Structural Genomics Initiatives meeting held by NIGMS in October 2008. Courtesy of Peter Preusch.

Expanding the Protein Universe: Improved Tools for Structure Prediction

While structure determination has been evolving, so too has a complementary field: structure prediction. "To make a real impact, you've got to pick the right targets and then use modeling to expand the structural information to many more sequences," says Norvell. Thus, structural genomics can leverage structure prediction to help fill in the gaps.

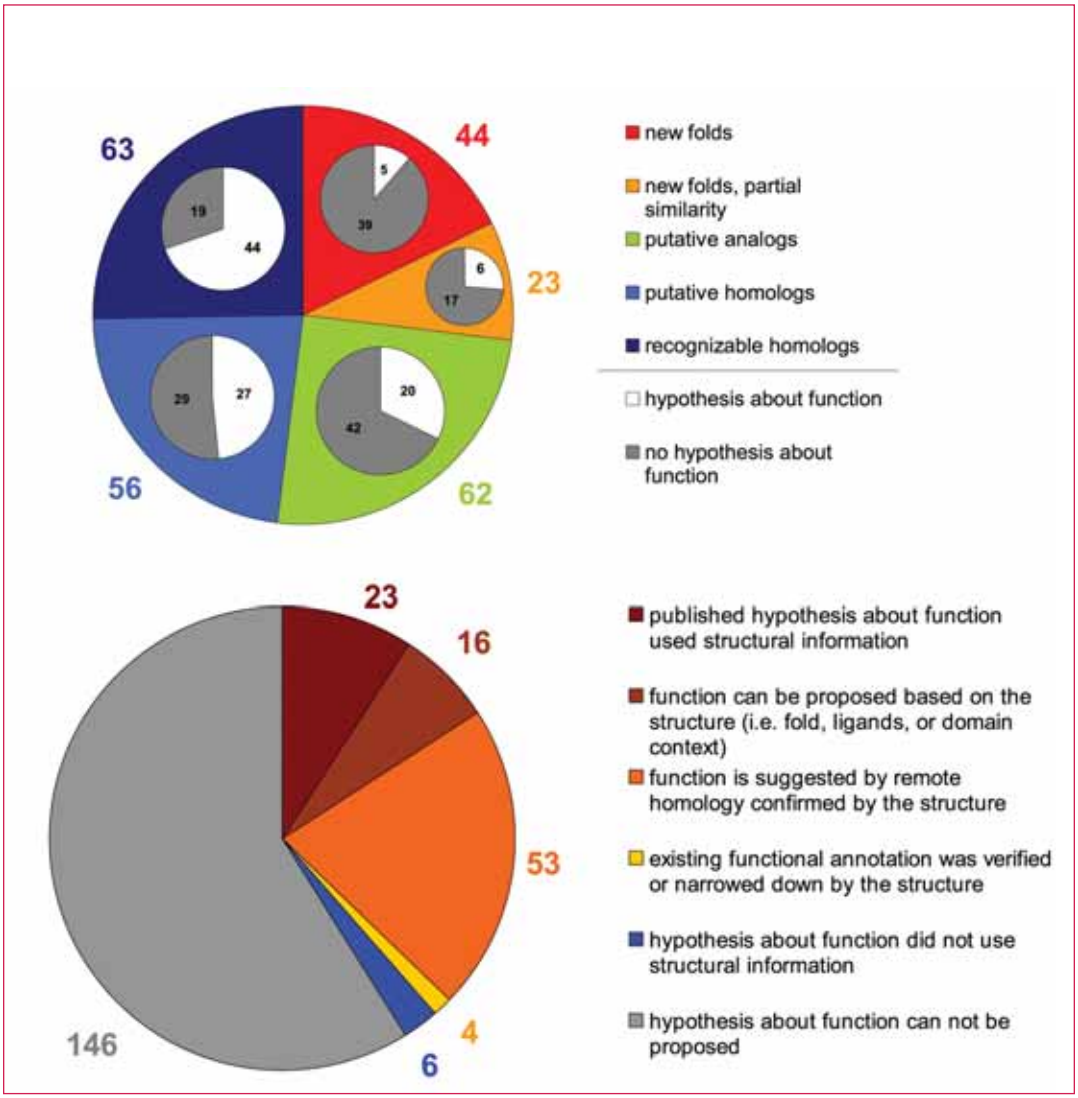
Structure prediction aims to accurately predict protein structures directly from their primary sequences, without wet lab experimentation. This prediction can be done by taking a

"knowledge-based" approach which gathers hints from known structures used as templates or a "physics-based" approach which starts from scratch using first principles to explore the possibilities of protein folds. The knowledge-based approach, also called homology modeling, is essentially "designing new buildings as better old buildings," says **Michael Levitt, PhD**, professor and chair of computational structural biology at Stanford. "The idea is that it has worked, so you can reuse it in a different combination."

High-throughput structure determination efforts have increased the number of known protein folds in sequence alignment databases, making it more likely that a protein with unknown structure will produce matches with sequences of known proteins that can serve as a template to then predict higher quality structures. Thus, structural genomics efforts contribute to homology modeling. "You're running on the same computers, same codes, but the database on which it runs is much larger now," says **Nir Kalisman, PhD**, a postdoctoral researcher of structural biology and computer science at Stanford.

In turn, structural genomics has benefited from structure prediction efforts, which leverage known structural information to fill in gaps in the structure space. "The main idea is that we really can get large scale coverage of all the structure space by sampling strategically, getting experimental structures of particular representatives, and then modeling around that using homology modeling techniques," says **John Moult, DPhil**, professor at the University of Maryland Biotechnology Institute.

Currently, scientists are able to learn from structures generated from the best of both the structure prediction and the structure determination worlds. "For any structure that's determined using X-ray crystallography or NMR, the model that you get is very highly reliable, the gold standard," says **Helen Berman, PhD**, professor of chemistry at Rutgers University and director of the PDB. Homology modeling, on the other hand, might be less certain, but still provides useful information, she says. Moult agrees: "A rough structural



Illuminating Protein Function via Structure. The Pfam database currently contains 2,247 families of "hypothetical proteins"—proteins with unknown functions or that are uncharacterized. In a 2009* PLOS Biology *paper, researchers looked at 248 of these families that were solved by the PSI to better understand regions of the yet unexplored protein universe that these families represent. The top pie chart breaks down the hypothetical proteins into subgroups based on their structural similarity and homology to known structures, ranging from proteins composed of new folds (red slice) to proteins with recognizable homology to known structures (dark blue slice). Within each of the five slices are mini pie charts showing the percentage of structures within each category for which hypotheses about their functions exist (white). What emerges is a relationship between structural similarity and homology and hypotheses about function: the greater the degree of structural similarity and homology to known structures, the more likely a functional hypothesis can be formed for that protein family. The lower pie chart further demonstrates that known structural information can facilitate inferences of function. From Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, et al. 2009 Exploration of Uncharted Regions of the Protein Universe.* PLoS Biology *7(9): e1000205. doi:10.1371/journal.pbio.1000205.

"To make a real impact, you've got to pick the right targets and then use modeling to expand the structural information to many more sequences," Norvell says.

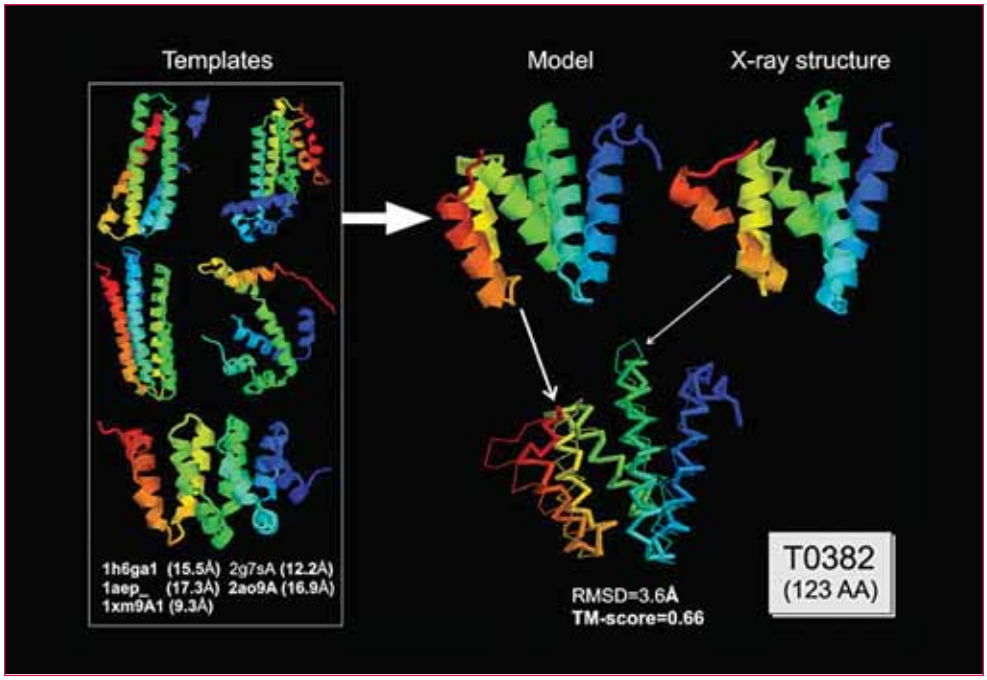
model based on the distant relationship by homology will be enough to give you some idea—albeit a crude low resolution idea—about function." And of all the publicly available gene sequences in GenBank, Moult estimates that more than 50 percent could be modeled at some rough level.

Yet Moult believes that for something like structure-based drug design, a very atomically detailed protein model is still required. Improving structure prediction to that level of precision will require further advances in computational methods and a better understanding of physical chemistry, he says.

On the other hand, a recent study by **Andrej Sali, PhD**, professor of bioengineering and therapeutic sciences at the University of California, San Francisco, and his colleagues illustrated that homology modeled proteins do nearly as well as X-ray crystal structures at deducing proteins' functions.

In order to directly evaluate the strength of current methods, the structure prediction community holds a competition every two years—the Critical Assessment of Techniques for Protein Structure Prediction (CASP). "CASP gives an objective way for many groups and methods to be compared on a level field," Kalisman says. Since 1998, the top performing modeling tool in CASP is ROSETTA, developed by **David Baker, PhD**, professor of biochemistry at the University of Washington. ROSETTA uses a fragment assembly method, taking short fragments from existing protein structures as guidance for modeling an unknown structure. The structures produced by ROSETTA get closer and closer to matching crystal structures all the time.

But another barrier to structure prediction remains: a cultural one. Biologists who work with structures want to know how reliable a predicted structure is—and that's often unknown, or at least unstated. Additionally, many structure prediction programs are large software packages that require a lot of computing power and are not accessible by the non-structural biologist. It's a problem, Kalisman says, because there's little outreach from the structure prediction folks to the biology community. "Biologists could definitely benefit much more if there was a better interface between most structure prediction algorithms and how biologists can approach them."



Fragment Assembly Algorithms for Structural Prediction. At the two most recent structure prediction competitions (CASP7 and CASP8), an algorithm called I-Tasser ranked as overall winner and outperformed human expert groups. Developed by Yang Zhang, PhD, associate professor of computational medicine and bioinformatics at the University of Michigan, I-Tasser uses fragment assembly as one step in a three-step procedure to model an unknown protein structure. In this example, I-TASSER used multiple algorithms to generate five templates (left panel) with secondary protein structure elements that best matched the query protein sequence T0382. These templates were then reassembled and refined to produce a structural model that only deviated from the experimental X-ray structure by 3.6 Angstroms. Reprinted with permission from Wiley Publishers, from Zhang, Y., Template-based modeling and free modeling by I-TASSER in CASP7 (2007). **Proteins *69(Suppl 8):108-17 (2007).***

MOVING BEYOND THE PIPELINE: THE PSI'S SPEED AND PRODUCTIVITY MAKE A DIFFERENCE FOR BIOMEDICINE

By increasing the number of available protein structures at a rate much faster than previously possible, PSI leaders believe structural genomics will hasten research advances in many areas of biomedicine. Indeed, there are signs that this is already happening.

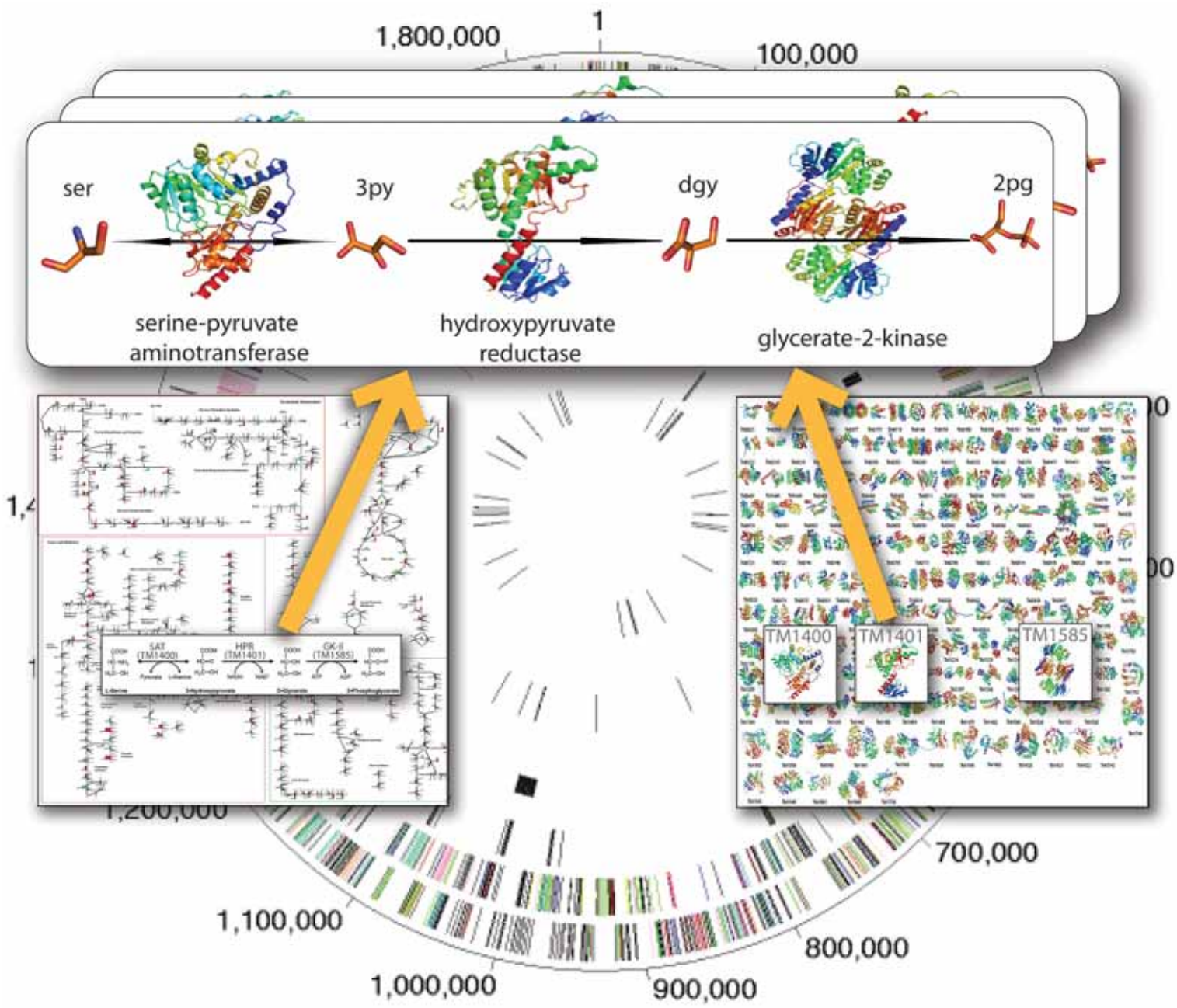
In a large collaborative project between one PSI center (NYSGXRC) and the Enzyme Specificity Consortium (ENSPEC), researchers proactively selected 535 proteins for structure determination. The proteins came from two structurally similar protein families (the amidohydrolase and enolase protein families) that catalyze a broad set of chemical reactions. To date, the NYSGXRC has completed X-ray crystallography for 75 of these proteins and modeled many more. To demonstrate the potential utility of these structures, the researchers performed *in silico* docking on one of them—the *Thermatoga maritima* amidohydrolase enzyme, Tm0936—to determine the enzyme's function, which was previously unknown. In the work, published in a 2007 *Nature* paper, thousands of configurations and conformations of molecules were docked into Tm0936 and ranked for fit. The top ranking compounds indicated that Tm0936 bound to and modified the structure of adenosine. The researchers then determined the crystal structure of Tm0936 in complex with one of the top ranking compounds and found only minor differences from the prediction, confirming its function. This is one example of how new approaches, in combination with the wealth of information from structural genomics, can lead to new insights.

Another example of the PSI's impact is the JCSG's human gut microbiome project, which focuses on impor-

tant pathways relevant to human health. Scientists hope that a better structural understanding of the proteins found in bacteria that populate the human gut will lead to the development of targeted drugs and therapies for human diseases. Even the study of proteins from deep ocean vents—far afield from the human gut—has the potential to aid in treatment of disease. For example, classical thymidylate synthase (TS) plays an important role in DNA synthesis and repair—and has been targeted in chemotherapy treatments for cancer. But the version of TS in *T. maritima* bacteria that live in thermal vents in the ocean turns out to have a completely different fold. Indeed, this enzyme has a different functional mechanism and has now been found in some pathogenic bacteria. The researchers suggest that a drug targeting the bacterial protein could prove to be a safe antibiotic because the human version is not homologous.

MAKING USE OF THE FOREST OF STRUCTURES TO ADVANCE BIOMEDICINE

Even as scientists have begun to capitalize on the large numbers of available structures, structural genomics researchers hope to take



Shaping a Metabolic Network. Godzik and his colleagues used experimentally determined and computationally modeled protein structures to reconstruct the central metabolic network of the bacterium **Thermatoga maritima.** ***By including the three-dimensional structures (lower right panel) of the proteins involved in the central biochemical reactions (lower left panel), they discovered a strong degree of conservation in protein folds that compose enzymes involved in similar reactions (top panel). From Ying Zhang and Adam Godzik, Burnham Institute. From Zhang, Y, et al., Three-Dimensional Structural View of the Central Metabolic Network of*** **Thermotoga maritima,** **Science** ***325:1544-1549 (2009), reprinted with permission from AAAS.***

things to a new level altogether.

"Classical structural biology focuses on individual proteins, so it's sort of looking at each tree separately," Godzik says. "Through changes in scale, what this becomes is looking at a forest—you suddenly see all the structures together and you start analyzing and comparing large groups of structures."

In recent work published in the September 18, 2009, issue of *Science*, Godzik and colleagues took the first leap in this direction. They constructed a comprehensive model of the metabolic network of thermophilic bacterium *T. maritima* that includes all the three-dimensional protein structures. For the first time, Godzik says, "we have a huge biological network which can be simulated and viewed as a mini cell *in silico*."

To build the model, Godzik's team had to first identify all the proteins in the metabolic network by extracting relevant information from more than 150 publications, and then subjecting that list to *in silico* analyses to identify gaps or redundancies that had to be resolved manually. Of the complete set of 478 proteins in the *T. maritima* metabolic network, 120 structures had been experimentally determined in part by the PSI JCSG. Using homology modeling, among other computational techniques, the researchers predicted the structures of the remaining 358 proteins. Standard methods produced the structures of 95 percent of these proteins, but the last few percent took a lot of effort, Godzik says. "Getting to 100 percent coverage was a huge challenge." And the quality of the predicted structures varied. For example, about 190 were comparable to low-resolution, experimental structures, while about 52 were merely approximate and others were somewhere in between.

But that effort proved worthwhile, Godzik says, because it led to a number of insights, perhaps most significantly, into the evolution of protein structures and organisms. The model they had constructed demonstrated that a small number of folds are represented in a majority of the proteins involved in the metabolic reactions of *T. maritima*. In fact, of the 478 proteins, including a total of 714 domains, there were only 182 distinct folds. And proteins involved in similar biochemical reactions have a higher probability of adopting similar folds. All of this supports the idea of structural conservation in nature, and to a much larger degree than researchers expected.

"Classical structural biology focuses on individual proteins, so it's sort of looking at each tree separately," Godzik says. "Through changes in scale, what this becomes is looking at a forest—you suddenly see all the structures together and you start analyzing and comparing large groups of structures."

With this project, researchers also challenged the conventional thinking that accompanies structure determination. "When we first submitted our paper, the first question that came from the editor was 'If this is a structural biology paper, what is the main structure you're talking about?' And we said, 'Well, there's no main structure; there are 478 main structures,'" Godzik says. "Both technological and conceptual changes are what structural genomics has brought to the table."

THE NEXT CHAPTER OF STRUCTURAL GENOMICS: STEPPING OUT INTO THE PUBLIC

In 2008, the Structural Genomics Knowledgebase (PSI SGKB) (http://kb.psi-structuralgenomics.org) was launched to integrate all the results from the PSI and make them available to the public along with an array of technology, protocols, and software. "The PDB has the structures. The SGKB has the structures and the sequences and the functional annotation and different technologies that allow you to get the structures," Berman says. "You have everything where you can find it in order to begin making new hypotheses and gaining new understanding."

The launching of SGKB signifies an important shift in the evolution of the PSI, says **Emily Carlson** of the NIGMS Office of Communications and Public Liaison. "It's gone from being a group of grants to being an actual research network where the researchers are sharing information and they're collaborating in ways that hadn't been done before. Not just within the PSI, but within the field and community in general."

By encouraging public access to solved protein structures and providing over 150 different resources at the PSI SGKB, the structural genomics community is showing its commitment to transforming structural data into meaningful information of use to the greater biological community, says **Michael Sykes, PhD**, postdoctoral researcher at The Scripps Research Institute. "It is not sufficient to determine structure for structure's sake. The scientific community needs to use these structures to make inroads into understanding the fundamental principles of biology."

The coverage of "structure space" will continue to be an aim of structural genomics, but the next phase—called PSI Biology instead of PSI 3—is shifting directions. The aim: To bring structure and function studies back together again and to connect biologists with the PSI effort, Preusch says. "The new thing is partnerships. We want to bring in people who have a biological problem of significant scope for which solving a large number of protein structures is necessary to really move the problem forward." □