Structural genomics: inside a protein structure initiative center

Nathan Blow visits the Joint Center for Structural Genomics for a glimpse inside one of the large-scale production centers for the Protein Structure Initiative.

In 2000, the US National Institute of General Medical Sciences of the National Institutes of Health funded the Protein Structure Initiative (PSI), a ten-year project to uncover the three-dimensional shapes of a wide range of proteins. The Joint Center for Structural Genomics (JCSG), based at The Scripps Research Institute in La Jolla, California, USA, is one of four large-scale centers involved in the production phase of the PSI. Four centers focus on high-throughput protein-structure determination, six specialized centers deal with difficult-to-solve proteins, such as membrane proteins, and two others provide new approaches to molecular modeling.

Ian Wilson, director of the JCSG, thinks the timing is perfect for the PSI centers to produce large numbers of new protein structures for the research community: "With more and more DNA sequences becoming available each day, the possibilities for the future of protein structure determination are tremendous." A central goal of the PSI is to enable the prediction of three-dimensional structures for most proteins from knowledge of their corresponding DNA sequence. In principle, this can be done by inferring the structure of a protein based on the known structure of representative members of the protein's family. "Most of the big protein families have been mapped—but still for 70% of known families we have no structural data," says Adam Godzik, of the Burnham Institute for Medical Research in La Jolla, and head of bioinformatics at the JCSG. This makes for a huge number of potential target proteins if one wants to have representatives from all families and therefore raises difficult questions: 'how do you choose which families to target and then which proteins within those families to obtain structures from?'.

Ian Wilson is director of the one of four large production centers for the Protein Structure Initiative.

Targeting the protein universe

"We are dealing with a continually expanding universe of proteins, so we had to have some rules about target selection," says Wilson. For the PSI, seventy percent of the target protein families are communally selected through PSI's Target Selection Committee. "We all sit down and execute a draft to decide which families each center will get," says Wilson. "By virtue of choosing particular families we avoid overlap, but also with this selection process each center can optimize specific targets within families for themselves," says Godzik. Another 15% of target proteins are decided upon by each center, and the final 15% are community targets proposed by outside researchers.

Godzik says that it is most effective for individual centers to decide which proteins to go after within the families they have been assigned because each center relies on different 'reagent genomes'—large sets of genomic DNAs used to isolate homologous sequences. At JCSG, it is Godzik, along with his bioinformatics team, who is responsible for determining the specific proteins JCSG will work on. By aligning a protein family with all 100 genomes available at JCSG, they first identify all homologous proteins. Then, using their own software, they assign a crystallization score to each homologous gene identified within

the family—a measure of the likelihood of success of the corresponding protein in the structure determination pipeline. "We take the ones that we predict to be most likely to succeed from this tool, and then we work our way down the list," he says.

Of course, there are those cases where nothing seems good. But thanks to the rapid advances in DNA sequencing technology and the explosion of fields like metagenomics (see **Box 1**), they have a very simple remedy to this situation: a continual update of potential targets in each family. "It is for this reason that we keep adding new genomes to our collection," says Godzik.

The GNF developed high-throughput bacteria fermentation system used by JCSG for large-scale protein production. (Courtesy of JCSG.)

Proteins for all

Any protein-structure-determination center is only as good as the quantity and quality of the proteins they generate. And during any given week of the year the Crystallomics core of the JCSG produces 25–50 new purified proteins. To attain this impressive yield, JCSG researchers were forced to alter some traditional methods for protein production.

Heath Klock, a JCSG research scientist, along with Scott Lesley, the Crystallomics core leader, and several colleagues devised the first step in the protein expression pipeline. "PIPE cloning is a ligation-free PCR-based cloning strategy that is amenable to cloning thousands of inserts into expression vectors in a very short period of time," says Klock. PIPE, or polymerase incomplete primer extension, uses PCR primers with overlapping 5′ ends to amplify vectors and inserts that can then be mixed and transformed with the resulting colonies screened by colony PCR[1].

For initial protein expression evaluation, JCSG scientists often rely on microexpression screening. "The real value here is that it gives you a peek ahead," says Mark Knuth, the Crystallomics core manager. They use a Vertiga commercial shaker from Thompson Instrument Company adapted for 96-well deep blocks and capable of achieving high-density cultures rapidly. Protein expression and solubility are evaluated using small-scale purification by immobilized metal affinity chromatography. Amazingly, microexpression has a very high predictive value for successful expression in JCSG's large-scale fermentor, which is capable of generating 96 high-density cell cultures in 100-ml centrifuge tubes. "Even though we can put a lot of tubes through the large fermentor, it is still time wasted if you put things in there that you know are not going to work," notes Knuth.

The protocols for purification of proteins after large-scale expression have evolved over time. "I think that one of the things that differentiate us from other centers is that we do higher-density fermentations, but we also do somewhat less purification," says Knuth. For the majority of proteins that go to crystallization trials now, a purification on a nickel resin with subsequent cleavage by tobacco etch virus (TEV) protease and a reverse nickel purification is performed. During the early years of the PSI, JCSG worked on developing high-throughput secondary purification, but found that the material they got after these secondary purifications was equivalent, from a crystallization perspective, to the nickel resin–purified, TEV protease cleaved and then reverse nickel purified proteins.

After large-scale expression and purification, each protein is put through a series

BOX 1 STRUCTURAL GENOMICS MEETS METAGENOMICS

"When metagenomic projects, such as environmental and the human microbiome, started to emerge, millions of new sequences started to flood the databases: some fit within existing proteins families, while others appear to form entirely new families," says Wilson. It was this realization that is leading the PSI to explore metagenomic datasets at the protein level.

"At first we started to collaborate with Craig Venter on the analysis of the 17 million sequences that he collected from worldwide sampling of ocean water," says Godzik. Among those sequences, his bioinformatics team identified around 6,000 new protein families. JCSG randomly sampled 76 proteins and has obtained structures for 6 of them so far, while still working on the others. A major hurdle, however, was that all the genes had to be synthesized as no starting DNA was available.

Nevertheless, the other PSI centers are also now looking into metagenomics approaches by exploring the human gut microbiome. But Godzik and his team, who were given the task of selecting targets for the other centers, decided to change their methods. "Because synthesizing genes can be very expensive we said let's try and do this a little differently. So we followed the approach of sequencing centers," says Godzik. In the human gut, it turns out that there are several dominant species that have been identified and sequenced using standard genomic approaches. Godzik's team focused on four of these bacteria to see if there were any families of proteins that were over represented in comparison to a random genome. This would position these proteins as being gut-specific. "We identified 3,000 proteins from our analysis, which have been distributed among all the centers," says Godzik. The results are just now starting to roll in with more than 400 proteins expressed and several structures solved thus far.

For Godzik and others, these results highlight how robust the PSI pipelines are across a wide array of protein targets. "Although we are now using our high-throughput technology to characterize proteins coming from new environments, we are still obtaining the same success rate," says Godzik.

The JCSG Automated high-throughput crystallization platform, called CrystalMation, was built by Rigaku Automation. (Courtesy of JCSG.)

of assays before any attempt at obtaining crystals for structure determination. These tests include analytical sizing, mass spectrometry and SDS electrophoresis on a Bio-Rad Dodeca multigel runner to assess purity. But Knuth is quick to note that they usually rely on one test above others. "Analytical size exclusion chromatography has the greatest predictive leverage of how a protein will fare in the crystal trials."

The crystallization trials

"For crystallization we usually set up the JCSG screen, which is four blocks of 96 conditions all done at two different temperatures," says Marc Elsliger, the administrative core manager for the JCSG. Identifying these 384 conditions to test for crystallization did not happen overnight.

"These conditions were based on the testing of all available commercial screens," says Elsliger. Over the early years of PSI-1, the JCSG tested conditions to identify which gave the best hits from a wide array of proteins during crystallization process. The fruits of this effort are not only available to the JCSG for their screens, but these conditions have also been commercialized recently through Qiagen as the JCSG Core Suites I–IV and JCSG+.

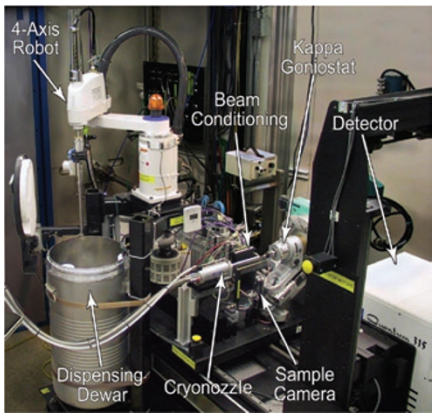
Protein crystallization trials at the JCSG rely on automated robotic systems. The original breakthrough system, developed at Syrrx/GNF, used nanoliter volumes to yield diffraction-quality crystals. They set up trials under all 384 conditions but then moved the plates into either 4 °C or 20 °C rooms for automated imaging over a 28-day period. The new JCSG CrystalMation system developed by Rigaku Automation is fully integrated with climate-controlled incubators. The introduction of the CrystalMation system boosts the capacity for crystallization trials at the JCSG to approximately 4,000 96-well plates per month.

But even with automated crystallization systems, Elsliger points out that two parts of the crystallization process proved difficult to automate. "The first is the scoring of the images to identify which of the drops contain usable crystals, and the second is the actual harvesting of the crystals." He notes that it is challenging to image a crystal in a liquid drop because of the curvature of the drop, the requirement for optimal lighting and the fact that clear crystals in a clear liquid background cause problems for scoring crystals.

At JCSG, several people have been trained to extract around 2,500 crystals per month from the drops and place these into the aluminum crystal cassettes in which they make their final journey north to the Stanford Synchrotron Radiation Laboratory (SSRL) for X-ray diffraction.

Look mom, no hands!

Once the crystals reach the Structure Determination core at SSRL, headed by Ashley Deacon, they are ready to be placed into the Stanford Auto-Mounter (SAM) system. "We were a large part of the proof of concept that users could use this on a



Automated protein crystallography beamline at Stanford Synchrotron Radiation Laboratory, showing the 4-axis robot and the dispensing Dewar of the Stanford Auto-Mounting system. (Courtesy of JCSG.)

daily basis," says Elsliger when discussing the JCSG's role in promoting the use of the automated beam line at SSRL.

The only human involvement once the crystals reach SSRL is placing the aluminum crystal cassettes into a liquid nitrogen–cooled dispensing Dewar; then the SAM systems Epson ES553S 4-axis robot takes over, opening the Dewar, obtaining a crystal and then reliably centering each crystal with the X-ray beam. Each step of this process can be observed, remotely, by JCSG scientists in La Jolla. "The value is that you can run the beam line much more efficiently," notes Elsliger. Using the SAM system, the beam line can operate 24 hours a day with only occasional human intervention when the cassettes in the dispensing dewar need to be changed. Remote operation of the beam line is not unique to the JCSG—the three other major production centers use this approach at other beam lines as well. Once the diffraction data have been obtained, automated analysis software interprets the crystallographic data for structure determination.

This also happens to be the point at which the JCSG has the most difficulty. "We take our biggest hit in going from a crystal hit to a usable dataset," says Wilson, noting that only 50% of crystals are usually big enough to be sent to SSRL, and of those around 50% will diffract to sufficient resolution to determine the structure. But the bottom line numbers show that despite the obstacles, all the centers of the PSI are making considerable headway into the protein universe. To date the JCSG has deposited more than 530 new structures in the Protein Data Bank, and structures for 335 of the original list of 1,269 largest protein families given highest priority by the PSI in 2005 have already been solved—200 from the PSI alone. "This shows that the PSI has made quite a dent in these proteins families already," says Wilson with a smile, "I find these numbers very encouraging for the future."

1. Klock, H.E., Koesema, E.J., Knuth, M.W. & Lesley, S.A. *Proteins*, published online 14 November 2007 (doi: 10.1002/prot.21786).

Nathan Blow is the Technology Editor for *Nature* and *Nature Methods* (n.blow@boston.nature.com).

SUPPLIERS GUIDE: COMPANIES OFFERING STRUCTURAL GENOMICS PRODUCTS

Company	Web address
Abbott Molecular	http://www.abbottmolecular.com
Accelrys	http://www.accelrys.com
Agilent	http://www.agilent.com
Applied Biosystems	http://www.appliedbiosystems.com
BD Biosciences	http://www.bdbiosciences.com
Beckman Coulter	http://www.beckman.com
Biacore	http://www.biacore.com
Bio-Rad	http://www.bio-rad.com
Bruker Daltonics	http://www.bdal.com
Caliper Life Sciences	http://www.caliperls.com
Cellomics	http://www.cellomics.com
Chemicon	http://www.chemicon.com
Ciphergen Biosystems Inc.	http://www.ciphergen.com
Clontech	http://www.clontech.com
Douglas Instruments Ltd.	http://www.douglas.co.uk
EMD Biosciences	http://www.emdbiosciences.com
Emerald Biosystems Inc.	http://www.emeraldbiosystems.com
Fluidigm	http://www.fluidigm.com
Formulatrix	http://www.formulatrix.com
GE Healthcare	http://www4.gelifesciences.com
Genetix	http://www.genetix.com
Genomics Solutions	http://www.genomicsolutions.com
Hamilton Robotics	http://www.hamiltoncomp.com
Hudson Control Group	http://www.hudsoncontrol.com
Imgenex	http://www.imgenex.com
Invitrogen	http://www.invitrogen.com
JEOL	http://www.jeol.com
Matrix Science	http://www.matrixscience.com
Millipore	http:///www.millipore.com
Molecular Devices	http://www.moleculardevices.com
Molecular Dimensions Ltd.	http://www.moleculardimensions.com
New England Biolabs	http://www.neb.com
Oxford Diffraction	http://www.oxford-diffraction.com
Pall Corporation	http://www.pall.com
Perkin-Elmer	http://www.perkinelmer.com
Pierce Biotechnology	http://www.piercenet.com
Promega	http://www.promega.com
Protein Sciences Corp.	http://www.proteinsciences.com
Qiagen	http://www1.qiagen.com
Rigaku Automation	http://www.rigaku.com
Roche Applied Science	http://www.roche-applied-science.com
Sigma-Aldrich	http://www.sigmaaldrich.com
Stratagene	http://www.stratagene.com
Takara Bio USA Inc.	http://www.takarabiousa.com
Tecan Group	http://www.tecan.com
Thermo Scientific	http://www.thermo.com
Varian	http://www.varianinc.com
Waters	http://www.waters.com