Meeting report

Structural genomics and structural biology: compare and contrast

John-Marc Chandonia*, Thomas N Earnest† and Steven E Brenner*‡

Addresses: *Berkeley Structural Genomics Center and †Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. ‡Department of Plant and Microbial Biology, 461A Koshland Hall, University of California, Berkeley, CA 94720-3102, USA.

Correspondence: Steven E Brenner. E-mail: brenner@compbio.berkeley.edu

Published: 25 August 2004

Genome ***Biology*** 2004, **5:**343

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/9/343

A report on the Keystone Symposium 'Structural Genomics', held concurrently with the 'Frontiers in Structural Biology' symposium, Snowbird, USA, 13-19 April 2004.

Is structural genomics a visionary undertaking heralding the future of structural biology, or merely a billion-dollar-plus folly? Two concurrent Keystone Symposia, 'Structural Genomics' and 'Frontiers in Structural Biology', brought together leading structural biologists and pioneers of the structural genomics community, providing an exciting opportunity to contrast cutting-edge advances in the two fields. The advances in structural genomics have focused on providing value for money: improved automation has resulted in expanded productivity. We expect the resulting flood of structures to provide an essential platform for future biological and medical research - just as genome-sequencing projects enabled new avenues of research, for example on the non-coding regions of DNA. In contrast, recent structural biology work has provided a tremendous amount of detail on specific biological mechanisms, including many areas that were previously beyond our technological capabilities to study.

Results from structural genomics consortia

In the US, the National Institutes of Health (NIH) are currently sponsoring nine pilot structural genomics centers through the Protein Structure Initiative (PSI). These centers are developing and deploying high-throughput techniques in preparation for the production phase of the PSI, set to begin in 2005. Gaetano Montelione (Northeast Structural Genomics Consortium (NESGC) and Rutgers University, Piscataway, USA) described the accelerating pace of structure determination in structural genomics. His center produced 9 structures in 2001 but it is currently on track to produce 75 structures in 2004; similar scaling up is reported by most other centers. The NESGC has successfully developed software to recognize innaccurate structures automatically, and to speed up the solution of structures by nuclear magnetic resonance (NMR) spectroscopy through automatic assignment of peaks on the NMR spectrum to atoms in the structure. The NESGC has also set up an automated pipeline to annotate proteins of known structure but unknown function.

The push towards high-throughput structure determination has shown the most impact initially on the automation of cloning, expression, and purification of proteins. Representatives from various PSI centers described efforts to establish an automated pipeline all the way from selection of the 'target' protein whose structure is to be solved, through to deposition of the solved structure in the Protein Data Bank (PDB [http://www.rcsb.org/pdb/]). Scott Lesley (Joint Center for Structural Genomics (JCSG), San Diego, USA) and Spencer Emtage (Structural Genomix, San Diego, USA) described the automated technology that their teams have developed for protein production: their multi-tiered approach handles the more tractable targets quickly and then applies increasingly specific approaches to the less tractable targets. The speakers who addressed protein production described primarily the adaptation of commercial robots to automate tasks; the main lesson is that one needs to multiplex as much and as early as possible in creating constructs, vectors, hosts, tags, purification schemes, and so on. With the added complexity, information management becomes particularly important. In contrast, Cheryl Arrowsmith (Structural Genomics Consortium (SGC) and Ontario Cancer Institute, Toronto, Canada) described how the SGC has deployed an effective pipeline with only limited use of robotics; too much automation has the danger of overly limiting experimental flexibility and especially the ability to adapt new protocols.

Automation of data collection for structures determined by X-ray crystallography has proven to be a success. Peter Kuhn (JCSG and Scripps, La Jolla, USA), Andrzej Joachimiak (Midwest Center for Structural Genomics (MSGC) and Argonne National Laboratory, Argonne, USA), and one of us (T.E.) described systems to automate and integrate the steps of crystal screening, data collection and processing, resulting in overall increases in both speed and accuracy of structure determination. Kuhn also described the early development of a nanocalorimeter that offers the possibility of experimental exploration of protein-protein and protein-ligand interactions with samples as small as a nanoliter. As the crystallization of protein targets remains a bottleneck, Rebecca Page (JCSG and Scripps Research Institute, San Diego, USA) described how high-throughput systematic crystallization trials at the JCSG have resulted in a database of over 325,000 crystallization experiments, which are being mined to identify proteins that crystallize more readily. Page showed that some biophysical properties of proteins - such as an unusually large or small Grand average of hydropathicity (Gravy) index, which correlates with solubility, or low complexity regions predicted by the SEG program - correlate well with crystallization difficulties. While the results are not unexpected, this is one of the first experimental studies to prove this correlation, and it enabled the JCSG to eliminate 35% of their potential targets without reducing their production of structures. Wim Hol (Structural Genomics of Pathogenic Protozoa (SGPP) and University of Washington, Seattle, USA) promoted the advantages of microcrystallization trials in plastic capillaries. These capillaries provide conditions similar to hanging-drop crystallization trials; but, as the plastic is nearly invisible to X-rays, freezing and data collection proceeds without removing the crystals from the capillary, avoiding the need for crystal handling.

Analysis of structures from structural genomics

An important goal of structural genomics is to annotate proteins of unknown function, primarily through analysis of their structures. Janet Thornton (MCSG and University College London, UK) described numerous methods for assigning function from structure, including inference of remote homology relationships, identification of ligands, and locations of electrostatic patches, pockets, and evolutionarily conserved residues; these methods are implemented in the ProFunc pipeline. Function may also be determined using high-throughput biochemical screens even in the absence of structure: Cheryl Arrowsmith described an array of binding assays that are used to determine whether each of a standard set of ligands binds to every protein purified by the SGC. Function may also be predicted by inferring homology from structure. Juswinder Singh (Biogen-Idec, Cambridge, USA) described methods for mining databases of the topologies of proteins with disulfide bonds to identify remote evolutionary relationships between proteins. Singh also presented the SIFTS database of structure-interaction fingerprints derived from known structures of ligand-receptor interactions, which his colleagues have successfully used to perform 'virtual screening' of potential ligands.

To ensure the quality of structures produced by their centers, Gaetano Montelione and Jane Richardson (Southeast Collaboratory for Structural Genomics (SECSG) and Duke University, Durham, USA) described automated tools (Procheck, MAGE, and MolProbity) they have deployed to reduce the number of errors. Procheck [http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html] checks the overall stereochemical quality of a protein structure; Mage [http://kinemage.biochem.duke.edu/kinemage/kinemage.php] is an interactive program for displaying proteins; and MolProbity [http://kinemage.biochem.duke.edu/molprobity/help/index.html] is a graphical user interface to several other structure-validation tools developed by the Richardson lab.

Janet Thornton also pointed out the relatively large diversity of structures solved by structural-genomics centers, compared with all the structures deposited in the PDB over the same time period: although 62% of all structures recently deposited in the PDB have a near-identical match already in the PDB, 63% of structures determined by structural genomics have no matches detectable from sequence. Thornton found that 14% of structures solved by structural genomics had new folds in the CATH protein structure classification database [http://www.biochem.ucl.ac.uk/bsm/cath/], 9% belonged to new superfamilies within existing fold classes, and 77% belonged to existing superfamilies. The lengths and domain organizations of structural-genomics targets were also distributed similarly to all other PDB entries.

The future of structural genomics

John Norvell (National Institute for General Medical Sciences, NIH, Bethesda, USA) presented details of the request for applications for the next phase of the PSI, which will start in July 2005. The PSI-II will consist of three or four major components: large-scale centers, which will focus on high-throughput production of structures aimed at increasing the structural coverage of proteins from sequenced genomes; specialized centers that will focus on eliminating the remaining barriers to high-throughput structure solution of challenging proteins (such as integral membrane proteins and multiprotein complexes); a centralized 'knowledge base' to disseminate results to the public, as well as to coordinate the target lists for each center; and (pending availability of funds from other NIH centers) disease-related centers that will focus on pathogenic genomes and on proteins from tissues and organs related to disease. Additional funds for biochemical analysis of structures will be available through supporting grants.

Several speakers suggested strategies for directing resources in the next phase of structural genomics. Before presenting

recent advances in NMR technology that enable larger structures to be solved, Kurt Wuthrich (JCSG and Swiss Federal Institute of Technology, Zurich, Switzerland) recommended selecting targets that provide complete coverage of several small proteomes and supplementing these targets with human and mouse proteins as well as membrane structures. Andrej Sali (NESGC and University of California, San Francisco, USA) recommends that structural genomics focuses on maximizing the number of structures that can be modeled with useful accuracy by computational methods.

Christine Orengo (MCSG and University College London, UK) recommended focusing broad coverage on the largest 1,345 protein families in the Pfam database [http://www.sanger.ac.uk/Software/Pfam/] with no structural representative, with finer coverage used sparingly to probe medically relevant families and unexplored regions of function space in sequenced proteomes. Orengo also presented unpublished data showing that although functional annotations based on single domains are generally unreliable below 40% sequence identity, annotations of a single protein based on combinations of multiple domains are accurate when based on as little as 20% identity. We (J.-M.C. and S.E.B.) quantified the 'Pfam 5,000' [http://www.strgen.org/pubs/chandonia-2004-proteins-pfam5000.pdf], a strategy to solve the structures of proteins from the largest Pfam families. This strategy would have a broad impact on our structural interpretation of sequenced proteins: obtaining the structure of one target from each of the 5,000 largest Pfam families would enable accurate fold assignment for approximately 68% of all prokaryotic proteins (covering 59% of residues) and 61% of eukaryotic proteins (40% of residues). We expressed the view that although the strategy of focusing structural genomics on a single tractable genome would have intrinsic benefits for our understanding of that organism, it would have little impact on our ability to interpret protein structures from other organisms.

Structural biology highlights

While structural genomics has become large-scale by increasing the throughput of structure production, structural biology has also become large-scale through advances that extend the range of structural biology to exploration of large macromolecular assemblies, real-time visualization of protein motions, single-molecule studies, and studies of ribozymes. Steve Harrison (Harvard University, Cambridge, USA) set the tone for the structural biology talks by describing the elegant interplay of several techniques - such as combining data from low-resolution electron microscopy with high-resolution X-ray crystallography, and real-time movies of live cells stained with immunofluorescent markers - to elucidate details of clathrin coat assembly and disassembly. The clathrin structure illustrates the importance of protein folding and unfolding in macromolecular assembly, as key components of the disassembly process are chaperone-related proteins. Harrison predicted that structural biology will tend towards more interactive experimentation in the future.

Three other speakers described technological advances in structural biology. David Agard (University of California, San Francisco) discussed the use of spatially structured illumination to extend the resolution limits of optical wide-field microscopy to better than 100 nm, with the promise of even higher resolution. Bridget Carragher (Scripps Research Institute) described how automated electron microscopy can be used to assemble medium-resolution structures of biomolecular assemblies from thousands of low-dose images. Homme Hellinga (Duke University Medical Center, Durham, USA) reported his group's remarkably successful efforts to computationally engineer the functions of a biologically active protein. One application of this work has been reagentless sensors, which combine ligand-binding and reporter functions in a single molecule. Hellinga also used a combination of computational design and directed evolution to swap the functions of two active enzymes using scaffolding from the other's (very different) folds.

Susan Marqusee (University of California, Berkeley, USA) has surveyed the *Escherichia coli* proteome to determine which proteins resist proteolysis, finding that resistance is not a result of the overall shape, rigidity, or thermodynamic stability of the native fold, but instead is a property of the energy landscape and whether the protein folds into near-native states with locally unfolded regions. Finally, two studies of individual macromolecular structures provided particularly interesting mechanistic insights. Jennifer Doudna (University of California, Berkeley) described studies of a self-cleaving ribozyme from hepatitis delta virus, whose mechanism appears to be similar to that used by protein ribonucleases. Electrostatic analysis of the active site revealed a shift in pKa (the negative log of the acidity constant) of a critical nucleotide base that enables enzymatic activity; accurate computational electrostatic analysis of this type has previously been limited mainly to proteins. James Spudich (Stanford University School of Medicine, USA) described single-molecule studies of myosin V and myosin VI, two molecular motors that move along actin filaments. Myosin VI may behave as a dynamic tension sensor, moving along actin until tension is sensed and then anchoring to maintain tension.

In conclusion, we find that structural biology and structural genomics complement each other well, if the focus of structural genomics is directed properly. If structural genomics funds are applied to decrease the cost and increase the throughput of protein production and purification, as well as to provide structural coverage of the broadest possible range of proteins, these efforts will set the stage for the next generation of structural biology research. There are no uninteresting proteins in the human genome; we may just not know what their importance is - just as we could not explore the

function of non-coding regions of genomes until complete genomes were sequenced. If the structures of most protein families can be determined, the ingenuity of structural biologists will be better focused on exploring the cellular and biochemical mechanisms of macromolecular assemblies, ultimately leading to better understanding of diseases and treatments and even to the engineering of proteins as nanomachines. Together, both fields are rapidly leading us to exciting new avenues of biomedical research.

Acknowledgements

This work is supported by grants from the NIH (1-P50-GM62412, 1-K22-HG00056) and the Searle Scholars Program (01-L-116), and by the US Department of Energy under contract DE-AC03-76SF00098.