Structural Proteomics: Large-Scale Studies

Rob M Ewing, Centre for Biological Sciences, University of Southampton, Southampton, UK
Declan A Doyle, Centre for Biological Sciences, University of Southampton, Southampton, UK

Based in part on the previous version of this eLS article ‘Structural Proteomics: Large-scale Studies’ (2006) by Martin Norin and Michael Sundström.

Introduction

The genetic code is the fundamental blueprint of life on earth. The simple combination of only four nucleotides as deoxyribonucleic acid (DNA) codes for all of the living organisms on our planet. While DNA is the code, sequences of amino acids and the proteins that they form are the nanomachines that build organisms and maintain them. The combination of the roughly 20 amino acids can produce millions of novel three-dimensional (3D) structures although biology is more selective in how it combines groups of amino acids. The chemical properties of the amino acids are used for specific purposes, and the most important concerning protein structure is the hydrophobic effect which is used to generate a globular-shaped protein. In this case, side chains that are hydrophobic in nature are gathered together in the centre of the protein thus generating a core. Additional side chains with different chemical and physical properties provide attractive and/or repulsive forces, flexibility, charge, mass and volume as well as the ability to cross-link thus providing multiple tools for specific functions. It is the exact 3D arrangement of these amino acids that ultimately determines the function of a protein hence determining a protein’s 3D structure is important in understanding its biological role. Although techniques for protein structure determination have been around for over 50 years (Kendrew and Perutz, 1957), large-scale or high-throughput determination of protein structures is more recent and is the topic of this review. Whole genome sequencing was the major scientific advance that set in motion the development of structural proteomics. Questioning what all of the proteins do in any one of the sequenced genomes naturally comes from having the exact DNA sequences. The ultimate goal of structural proteomics (or genomics) is to provide the structural basis for functional annotations of all proteins within an organism.

The initial drive of the structural genomic organisations (SGO) was to develop high-throughput techniques for as many proteins as possible, with a particular emphasis on novel protein folds. These new methodologies have been extensively used and improved so that the next phase has focused on applying these techniques and developing new approaches to the more difficult targets such as integral membrane proteins. This can be seen in the increase in groups working specifically on membrane proteins (Table 1). It should be noted that at present, there are no SGOS that as their main focus concentrate on protein complexes, protein/DNA complexes or protein/RNA complexes.

As of the beginning of 2015, the number of publicly available protein structures in the Research Collaboratory for Structural Bioinformatics (RCSB) protein data bank (www.rcsb.org) originating from structural genomic projects is ~13,200 PDB entries of which ~2700 are of human origin and ~1300 contain a ligand. The number of entries that are membrane protein structures is 111, which represents ~0.8% of the total output for structural genomics organisations. In comparison, all combined membrane protein structure depositions is, at present, 1.5% of the total. This...
Table 1 Structural proteomics initiatives

<table>
<thead>
<tr>
<th>Name</th>
<th>Strategic goals</th>
<th>Web site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center for Eukaryotic Structural Genomics</td>
<td>Method and technology development. HT structure determination with focus on Arabidopsis thaliana</td>
<td><a href="http://www.uwstructuralgenomics.org">www.uwstructuralgenomics.org</a></td>
</tr>
<tr>
<td>Joint Center for Structural Genomics</td>
<td>Novel structures from Caenorhabditis elegans and human proteins involved in cell signalling</td>
<td><a href="http://www.jcsg.org">www.jcsg.org</a></td>
</tr>
<tr>
<td>Midwest Center for Structural Genomics</td>
<td>Streamlined and cost-effective processes. Structures of targets of unknown fold and proteins from disease-causing organisms</td>
<td><a href="http://www.mcsg.anl.gov">www.mcsg.anl.gov</a></td>
</tr>
<tr>
<td>TB Structural Genomics Consortium</td>
<td>Structure determination and analysis of proteins from Mycobacterium tuberculosis</td>
<td><a href="http://www.webtb.org">http://www.webtb.org</a></td>
</tr>
<tr>
<td>Center for High-Throughput Structural Biology</td>
<td>Technology development in structural genomics</td>
<td><a href="http://www.chtsb.org">http://www.chtsb.org</a></td>
</tr>
<tr>
<td>Ontario Center for Structural Proteomics in Toronto</td>
<td>Genome-scale structural biology. Function from structure. Provides protein samples for various structural research groups worldwide</td>
<td><a href="http://www.uhnres.utoronto.ca/proteomics">www.uhnres.utoronto.ca/proteomics</a></td>
</tr>
<tr>
<td>Membrane Protein Structural Biology Initiative GPCR Network</td>
<td>Structure determination of integral membrane proteins</td>
<td><a href="http://mpsbc.org/">http://mpsbc.org/</a></td>
</tr>
<tr>
<td>RIKEN Structural Genomics/roteomics Initiative</td>
<td>Structural determination of the medically important GPCR family</td>
<td><a href="http://cmpd.scripps.edu/index.html">http://cmpd.scripps.edu/index.html</a></td>
</tr>
<tr>
<td>Transmembrane Protein Center</td>
<td>Large-scale structural biology of prokaryotes (replication, repair, transcription and translation) and eukaryotes (cell growth and differentiation genetic systems)</td>
<td><a href="http://www.rsgi.riken.go.jp/rsgi_e/">http://www.rsgi.riken.go.jp/rsgi_e/</a></td>
</tr>
<tr>
<td>Oxford Protein Production Facility</td>
<td>Method and technology development for integral membrane proteins</td>
<td><a href="http://www.uwmembraneproteins.org/index.html">http://www.uwmembraneproteins.org/index.html</a></td>
</tr>
<tr>
<td>Center for Membrane Proteins in Infectious Diseases</td>
<td>High-throughput production of proteins and protein crystals by automating and miniaturising</td>
<td><a href="http://www.oppf.ox.ac.uk">www.oppf.ox.ac.uk</a></td>
</tr>
<tr>
<td>TransportPDB</td>
<td>Structure determination of viral, bacterial and human proteins involved in pathogenesis</td>
<td><a href="http://mpid.asu.edu/">http://mpid.asu.edu/</a></td>
</tr>
<tr>
<td></td>
<td>High-throughput functional assay development and structural characterisation of integral membrane proteins</td>
<td><a href="http://192.231.106.23/">http://192.231.106.23/</a></td>
</tr>
</tbody>
</table>

points to the successful push by the structural genomics groups with this difficult group of protein structures. (See also: Protein Structure)

Another important group of proteins are those of our own, the human proteome. Recent estimates suggest that the number of protein-coding human genes may be even fewer (<20,000) than previously estimated (Ezkurdia et al., 2014). However, the number of distinct human proteins is much higher than this as multiple protein isoforms may be encoded by a single gene, and most proteins are subject to some form of post-translational modification. Structural knowledge of each one of these variants is important in understanding their biological roles hence more than one structure may be required for the complete understanding of each individual target. If we consider drug development, the structures of many protein/ligand complexes are required in order to fully explore and exploit the protein landscape. Hence, human proteins are relatively heavily populated in SGO PDB depositions with ~20% of their total output being human in origin.
The most precise and accurate information on the structure of a particular protein or a protein complex can be obtained from experimental methods, such as X-ray crystallography and nuclear magnetic resonance (NMR). Advances in the technology and methodology are beginning to produce atomic resolution structures using single-particle electron microscopy (Liao et al., 2013); hence, this technique will be expected to add to the numbers in the future.

All structural genomics projects aim at systematically mapping the protein structural space, either targeting specific organisms (e.g. Homo sapiens, thermophilic bacteria, Caenorhabditis elegans and Mycobacterium tuberculosis), different protein classes (e.g. membrane proteins, metabolic enzymes, kinases and proteases), targets of specific diseases or biological function relevance or targeting proteins that have the potential of providing examples of novel structure folds (note that novel experimental protein structures provide templates for structure predictions of homologous proteins). However, the technological challenges are common to any of these strategies. The key limiting factors are difficulties obtaining pure soluble protein material, growing protein crystals, the manual intervention and time required for X-ray crystallographic data collection and evaluation and the time required for data collection and spectral interpretation using NMR approaches.

Technological developments driven by the structural genomics approach include high-throughput (HT) parallel cloning and multivariate approaches for expression and purification, core domain identification using proteolysis methods and the use of expression and detection tags. Protein crystallography has undergone a dramatic series of improvements: freezing of crystals at liquid-nitrogen temperature (cryofreezing), single-wavelength anomalous dispersion (SAD) and multiple-wavelength anomalous dispersion (MAD) phasing, crystallisation in nanolitre volumes, novel crystallisation techniques, robotisation, automated data collection and the use of synchrotron beamlines have been adopted as standard methodologies. The improvements in structure determination by biomolecular NMR using isotope-enriched protein samples include the use of high-field spectroscopy instrumentation, cryogenic probes and automated spectra assignment and structure determination. Figure 1 summarises the experimental flow in structural proteomics, the current bottlenecks and technology developments.

**Figure 1** Experimental flow in structural proteomics, the current bottlenecks and important technology developments.

---

**Collection of Protein Folds in the Proteome**

The initial idea behind structural proteomics was to generate useful 3D structures of entire proteomes by a combination of experimental structure determination and modelling. Once a structure has been determined, it was assumed that related proteins (>30% amino acid sequence identity) would adopt the same conformation and therefore modelling of additional family members based on the original structure would be sufficient to structurally describe the remaining members. Considering that many proteins are made up of multiple structured domains, the key question then would be: how many structural templates are needed to model most proteins or their domains?

It turns out that only ~2000-folds would cover 70% of all structural domains from 203 genomes. However, to be able to generate models for the remaining family members based on the 30% sequence identity cut-off would require ~90,000 structures (Marsden et al., 2006). Realistically, this underestimates the total number of structures required to completely describe, at a molecular level, all that is going on in a cell. The additional factors that are not considered are how multiple domains within the same protein interact, alterations due to post-translational modifications, multiple isoforms and conformational changes associated with ligands and protein–protein complexes. An example of the requirement for multiple structures is seen in the structural analysis of human 14-3-3 family. These proteins which are important in cell signalling have a sequence identity of >60% between family members. The combination of the structures demonstrated...
that only with the generation of multiple structures was there sufficient information to describe the flexibility between the monomers and within the phospho-peptide-binding pocket (Yang et al., 2006). Another important line of knowledge that requires an increase in the number of structures of the same protein fold is when considering the changes that occur as a result of molecular evolution (Inoue et al., 2014; Spudich et al., 2014).

In the pharmaceutical industry, protein modelling is applied throughout the value chain from the discovery of target proteins to the generation of lead molecules to the prediction of pharmacological effects in clinical trials. In many instances, multiple protein/ligand structures are required to fully describe and understand all of the interactions and binding properties (Cousido-Siah et al., 2014; Gazzard et al., 2014). So, even though the initial number of required folds is relatively small, the complexity of living organisms ensures that structural genomics will be in demand in the future.

Finally, databases are of course critical repositories for the large volumes of information associated with structural proteomics. For protein folds, database mining tools to store, organise and identify protein folds are becoming more and more important as the number of protein structures grows (Sippl et al., 2008). A novel structure that has been determined may be scanned against databases of known structures such as the DALI and CATH resources (see Web Links) (See also: Protein Structure Prediction and Databases).

Protein Production

The success of HT structure determination and subsequent structural analysis is totally dependent on high-throughput protein production. Other critical factors involve the availability of methods for rapid and accurate analysis of purity, homogeneity and structural integrity.

For a research effort in structural proteomics, one can pick the ‘winners’, that is, target proteins that with minimum amount of effort are easy to express with the appropriate characteristics and give good quality NMR spectra or form diffracting crystals. Thus, in the initial phase of structural proteomics, expression and purification steps are streamlined so that multiple constructs of the same target can be used. The risk with this approach is that certain folds could become overrepresented in time and that other target types will not appear until a directed effort is attempted (e.g. causing a biased sampling of the structural space). Even the general properties of proteins from different kingdoms can affect the ability of a protein to crystallise and therefore allow its structure to be determined. Eukaryotic proteins are significantly less likely to crystallise than bacterial proteins owing to their larger inherent flexibility (Mizianty et al., 2014). This is likely to improve over time as newly developed approaches and techniques successfully circumvent these problems.

HT approaches, by necessity, utilise affinity and detection tags to allow rapid protein screening and purification. These tags can range from the small hexahistidine cluster to the large maltose-binding protein, all of which generally need to be removed before NMR and X-ray studies (Bird et al., 2014; Elsliger et al., 2010; Makowska-Grzyska et al., 2014). Structural studies by NMR require the tag to be small and to not interfere with the target protein. An example of such an approach was the identification of a solubility enhancement tag (SET) from the protein GB1 domain (Zhou and Wagner, 2010). In the test cases reported, the SET tag improved the characteristics of the expressed proteins in terms of solubility and stability and did not interact with the target proteins.

Regardless of the choice of fusion partners, either smaller tags or larger proteins such as green fluorescent protein (GFP) may give misleading data by solubilising poorly behaving expression constructs or protein components lacking their natural interaction partner. Thus, ‘blind’ optimisation for the best fusion tag using solubility screens needs to be accompanied by functional assays to assure that the constructs chosen for further studies are biologically relevant.

High-throughput Protein Crystallography

The five basic steps in structure determination by X-ray crystallography are cloning, expression, purification, crystallisation and structure determination. Application of novel molecular biology techniques such as ligation independent cloning and miniaturisation of the expression to crystallisation steps greatly speeded up the generation and number of crystals that can be produced. The intense X-ray flux at synchrotrons is the fastest and best place to collect data, especially from the relatively small-sized crystals that are produced as a result of miniaturisation of the crystallisation process. Technical advances in automated crystal mounting has removed this slow and potentially error prone manual step (Smith and Cohen, 2008). In addition, the advance in computing power, speed and connectivity is making remote data collection much more prevalent (McPhillips et al., 2002; Smith et al., 2010; Stepanov et al., 2011). Smaller crystals require new methods of handling and collection of suitable quality diffraction data that can be used to solve its structure. Automated methods of crystal mounting are one such advance (Cipriani et al., 2012; Heidari Khajepour et al., 2013; Wagner et al., 2013). The combination of fast, continuous read out detectors and free electron laser X-rays are allowing the collection of data from micrometre- and even nanometre-sized crystals (Chapman et al., 2011; Yoshikawa et al., 2014). Not surprisingly, these small crystals are sensitive to radiation damage and therefore only a small proportion of data is able to be collected from any one crystal. This requires software development in order to scale and merge these potential millions of data set fragments (Foadi et al., 2013; Hunter and Fromme, 2011). All of these advancements are increasing not only the throughput of structure determination but also the type of protein targets that can now be included such as the membrane proteins and large protein complexes.

From Structure to Function

Proteins sharing the same folding may have quite different functions, and prediction of protein function from structure is
challenging (Redfern et al., 2008). Other studies have concluded that precise function seems to be conserved down to 40% sequence identity, whereas a broader definition of a functional class is conserved down to 25–30% identity (Todd et al., 2001; Wilson et al., 2000). In a limited but significant number of cases, direct electron density for ‘native’ ligands or co-factors bound to the protein could be observed in structures derived from X-ray crystallography. When such data are available at high resolution, hypothesis generation on the function of the protein often can be more straightforward.

A good example of direct functional annotation from structure was previously reported (Zarembinski et al., 1998). In this study, the crystal structure of an unannotated protein, MJ0577, from Methanococcus jannaschii clearly revealed a bound ATP in the 1.7 Å electron density maps, suggesting that MJ0577 was an ATPase or an ATP-mediated molecular switch. The structure-based hypothesis could subsequently be confirmed by biochemical experiments. In addition, the structural analysis of the ATP-binding motif could be used to suggest other putative ATP-binding sequences among the many homologous, but previously unannotated, proteins in this family.

Although a few studies on structure-based assignment of single proteins from experimental structures have emerged, the structural proteomics effort on the archaeon Methanobacterium thermoaerotrophicum is a good case study (Christendat et al., 2000). Here, 424 out of 900 target proteins, predicted to be soluble and without a template in the Protein Data Bank, were chosen for structure determination and subsequent functional assignment. The selected proteins represented around 25% of the organism’s proteome (1871 open reading frames). The targets were cloned, expressed and purified in a streamlined approach and attempts were made to solve the structures by both NMR (<20 kDa) and crystallographic methods at various laboratories. Approximately 20% of the target proteins were found to be suitable candidates for structure determination.

Furthermore, the study revealed that poor expression and solubility of the proteins accounted for close to 60% of the failures. It was also observed that NMR data collection and crystallisation were the two major time and resource consumers in the process. Ten structures (including MTH538 discussed above) by NMR and X-ray were simultaneously published. Five of the ten structures contained a bound ligand or a ligand-binding site that could be inferred from structural homology. Thus, many of the structures suggested a number of functional assays that could be used to provide insights of function.

Computational prediction of protein function from structure has been and continues to be an important area of investigation. Although protein sequence alone can provide many clues as to the function of a protein, 3D structural information is particularly useful for identifying distant relationships between proteins that suggest functional roles (Watson et al., 2007). The ProFunc server (see Web Links) predicts protein function by combining sequence level features of proteins with structural features such as protein folds, surface topology and motifs (Laskowski et al., 2005). While no one feature or method is able to always perform the best protein function prediction, it was found that for prediction of function for a large set of new protein structures from a structural proteomics study, secondary structure matching (SSM) (Krissinel and Henrick, 2004) in which unknown protein structures are aligned with known protein structures, provided the best overall prediction of function (Watson et al., 2007) (See also: Protein Structure Prediction and Databases).

### Structural Proteomics and Systems Biology

Understanding the function of individual proteins is a key goal of structural proteomics. Most proteins, however, function as components of macromolecular complexes or networks. Understanding protein function therefore requires an understanding of the interactions and interrelationships between proteins and the global organisation of proteins into networks, which is a principal theme of systems biology. By resolving large numbers of protein structures, structural proteomics has an important part to play in systems biology approaches, and recent efforts have begun to integrate available protein structures with other types of ‘omics’ data to better resolve cellular networks at a structural level. Indeed, an editorial in one of the principal proteomics journals stated that structural proteomics should be defined as the systematic study of relationships between biological macromolecules (Stevens and Yates, 2007). As an example of this approach, the central metabolic network of a bacterium that lives in hot springs, Thermotoga maritima, was reconstructed by integrating biochemical information about metabolic reactions with known and predicted protein structures (Zhang et al., 2009). An interesting finding from this study was that the central metabolic network of Thermotoga is dominated by a surprisingly small number of protein folds. Integration of structural proteomics data with protein–protein interaction networks and genetic information on human diseases has also proved to be a powerful approach, in this case allowing predictions to be made about the effects of disease-causing mutations on protein–protein interactions and networks. For example, integration of the thousands of known disease-linked, Mendelian mutations in human with protein–protein interaction networks and protein structures was used to show how mutations at different protein interaction interfaces of the same protein may cause different diseases (Das et al., 2014). These types of study, using structural proteomics data, are somewhat limited by the numbers of available protein structures. However, the advances in experimental and computational determination of protein structures outlined in this article will continue to contribute large numbers of high-resolution protein structures that can be used in these integrative studies (Lu et al., 2013). See also: Interaction Networks of Proteins

### Related Articles

- **Industrialization of Proteomics: Scaling Up Proteomics Processes**
- **Macromolecular Structure Determination: Comparison of Crystallography and NMR**
- **Mass Spectrometry in Protein Characterization**
Molecular Entry Point: Strategies in Proteomics
Protein Characterisation in Proteomics

References


Ezkurdia I, Juan D, Rodriguez JM, et al. (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Human Molecular Genetics 23 (22): 5866–5878. DOI: 10.1038/hmg/ddu309.


Hunter MS and Fromme P (2011) Toward structure determination using membrane-protein nanocrystals and microcrys-


Smith CA, Card GL, Cohen AE, et al. (2010) Remote access to cry-


Further Reading


Web Links

http://www.rcsb.org/ - The Protein Databank (PDB) – A worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

http://www.cathdb.info - CATH – Protein Structure Classification. CATH is a database for classification of protein domain structures into families.


http://www.els.net - ProFunc – server for predicting protein function from structure.