

# STRUCTURAL GENOMICS: PITFALLS AND PROSPECTS

A. BRIDGES AND O. JENKINS

GENE EXPRESSION SCIENCES, GLAXOSMITHKLINE, HARLOW (UK)

## Presentation given at the ESRF Workshop "High Throughput Structural Biology", 20-21 February 2001.

*A discussion of the issues and feasibility of truly high-throughput expression, purification and structural determination using current technologies, highlighting potential bottlenecks in the existing process of gene to structure.*

The solving of the human genome sequence, and the genomes of many other species, has resulted in a huge increase in the number of clinically-relevant targets (proteins) of interest to pharmaceutical companies. The production of targets for structural analysis has typically been a very slow process, mainly due to the very high purity and large quantities of protein required. To give the highest impact in a drug design process, structural information should ideally be available when lead compounds are identified from high-throughput screens. However, the solving of structures of proteins of interest can frequently occur at a very late stage within drug development projects permitting only limited input to structure-based drug synthesis and optimisation.

Large pharmaceutical companies are often interested in targets from diverse protein classes, ranging from microbial enzymes to kinases, membrane bound proteins such as 7-transmembrane receptors and ion channels. Expression of these proteins increases in difficulty - microbial proteins are relatively easy to express and crystallise, soluble mammalian proteins can be much harder to obtain, and integral membrane proteins are extremely difficult to generate in an active, folded form in sufficient quantities for structural determination.

There are many different systems available for expression of genes of interest. *E. coli* is the quickest and most economical, being easy to scale to large volumes and suitable for generation of labelled protein. However, *E. coli* does not perform post-translational modifications

that may be required for functional activity. Over-expression in yeast is also fast and straightforward to scale, though the first choice for mammalian proteins is often baculovirus expression. This is much more expensive than yeast and *E. coli*, especially at large scale, but this system does perform post-translational modifications such as phosphorylation. Expression of eukaryotic proteins in mammalian cells will usually produce the native protein, but is a very slow, expensive procedure and only infrequently produces proteins in high enough yield for structural determination.

Generation and expression of proteins for structural analysis has usually been performed as an iterative process. One to three constructs - alternative forms of the gene, either truncated or mutated - are generated from the full-length gene and expressed in the 'best bet' system, usually *E. coli*. The first construct will often be that generated for high-throughput screening, which is not necessarily ideal for crystallisation. If the constructs express well, and can be supplied to the crystallographers in the required purity and concentration, a second round of construct generation is not necessary unless crystallisation fails or the crystals are of poor quality. However, if the proteins fail at any step during the supply (purification or expression), further rounds of construct generation or use of an alternative expression system is required. This iterative process clearly lengthens the time taken to produce structures, decreasing its impact on the drug design process.

It is clear that the time from gene to structure must be improved to take advantage of the many new targets in the completed human genome. Homology modelling can produce putative structures for genes within families, but the structures currently deposited in the PDB (Brookhaven Protein Data Bank) represent only a fraction of the populated folds, characteristic tertiary structure elements, such as helix-loop-helix,  $\beta$ -barrel, thought to exist in natural proteins [1]. This limits the usefulness of homology modelling as a predictive tool. There are many initiatives worldwide to determine structures of unknown proteins from the various genomes, and it is estimated that solving a minimum of 10,000 structures within 10 years should produce at least one example of the folds from each family of proteins [2].

To determine such a large number of structures within a relatively short time frame requires a re-think of the current methods of protein generation. Instead of working with one to three constructs, and performing several rounds of generation and purification to generate all of the constructs of interest to the crystallographers, multiple constructs could be produced in one go (including constructs designed for high-throughput screening). These could be cloned into powerful *E. coli* expression systems, and expressed in multiple *E. coli* strains. There are many new strains designed to improve expression and folding of proteins. One or two affinity tags could be included (both N- and C-terminal) to assist in purification, and protease cleavage sites

(such as TEV or PreScission) could also be encoded to remove large flexible regions from the N-termini of the proteins which would potentially interfere with crystallisation. In addition, site-directed mutagenesis could be employed to remove undesirable surface residues or loops. Using this approach, it could be anticipated that many constructs of a particular protein in a form suitable for crystallisation may be generated.

What would be the implications of this increase in the number of proteins for crystallisation? It is estimated that the production of 1000 crystal structures per year would require approximately 10,000 proteins to be screened for crystallisation (about 800 crystallisation screens per month). If each crystallisation screen comprised 200-500 conditions, then each screen would require 2-5mg of protein at a concentration of 10mg/ml and 95% purity. In order to collect sufficient X-ray data to solve this quantity of crystal structures a dedicated MAD (Multiple-wavelength Anomalous Diffraction) beamline would be essential. Using current structure solution methods, automated electron density map interpretation and refinement packages each structure would take approximately 2-4 weeks to complete. A

team of 50-100 protein crystallographers would be required to undertake a task of this magnitude.

The key to completion of this number of structures is miniaturisation and automation at every step of the process. Cloning, expression and purification is relatively straightforward to automate; robots exist to perform liquid handling, but expression analysis would be time consuming by electrophoresis (SDS-PAGE), so an alternative must be developed. Crystallisation can also be automated (possibly at the nanolitre scale), but crystal mounting and freezing is currently a time-consuming and manual procedure, and therefore a serious bottleneck in this process. Finally, new programs need to be developed to automate structure solving and perform docking in a high-throughput mode.

It is not impossible to do high-throughput structural determination. Indeed, there are several biotech companies (*e.g.* Syrrx, Structural Genomics) with business plans based on structural genomics. Large pharmaceutical companies have a slightly different focus, with needs closer to 'functional' genomics. Although crystal structures of pure

proteins can be used for compound docking and SAR (Structure-Activity Relationship studies), more information for drug design and optimisation can be obtained from structures of protein/ligand or protein/ inhibitor complexes. This is particularly true if inhibitors of many structural classes are available for co-crystallisation. However, this is just a difference in emphasis in the needs of structural genomics projects versus drug discovery, and the methods discussed in this article can be applied and adapted to achieve both sets of aims. ■

#### REFERENCES

- [1] S. K. Burley, *Nature Struc. Biol.* 7, 932-934 (2000).
- [2] J. C. Norvell and A. Z. Machalek, *Nature Struc. Biol.* 7, 931 (2000).

#### ACKNOWLEDGEMENTS

We would like to thank all of the members of the Structural Biology, Gene Expression Sciences, Protein Biochemistry and Computational Chemistry departments at GSK in Harlow that took part in the discussions described in this article.

# AUTOMATED DATA COLLECTION AND PROCESSING FOR MACROMOLECULAR CRYSTALLOGRAPHY

A.G.W. LESLIE

MRC LABORATORY OF MOLECULAR BIOLOGY, CAMBRIDGE (UK)

**Presentation given at the ESRF Workshop  
"High Throughput Structural Biology", 20-21 February 2001.**

*Fully automated data collection and processing is an achievable and highly desirable objective for modern synchrotron protein crystallography beamlines. A possible scheme for reaching a high level of automation with modest programming resources is outlined.*

**S**tructural genomics initiatives such as that recently funded by the National Institute of General Medical Sciences in

the USA will lead to a dramatic increase in the rate at which macromolecular structures are determined. This increase

in throughput will be achieved by applying automation to almost all steps in the structure determination pathway,