

Showcase on Research

Structural Genomics: Protein Structures for the Masses?

Carmel Walsh, David Hume, Bostjan Kobe and Jenny Martin

School of Molecular and Microbial Sciences and Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD 4072

Now that the human genome sequence has been released to reveal the first draft of the DNA code that makes us what we are (1,2), we are left with the mind-boggling task of sorting out exactly what it all means. We know that human chromosomes incorporate about 3 billion base pairs coding for upwards of 30,000 proteins, but we only know the function of a few hundred of these gene products.

The next logical step in the process of understanding our genetic make-up is to determine the function of each of these encoded proteins. This is where structural genomics comes into play. The objective is deceptively simple: to structurally characterise at least one member of every protein family and use the structure to infer function (3). The resulting library of protein folds would represent every possible structural permutation that can be adopted by a string of amino acids.

Eventually, each unique fold would be associated with discrete chemical and biological functions, so that as new protein sequences are discovered and categorised by fold type, their functional relevance will also be defined.

Although the concept sounds simple, putting it into practice is far more challenging. For starters, although there are currently about 15,000 structures from hundreds of different organisms in the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Databank (PDB) database (<http://www.rcsb.org/pdb>) (4,5) it took many decades to get to this point (Fig. 1).

There is good reason for this extended time frame. Structural biology is a technically demanding, expensive and time consuming business. How then can individual laboratories be purportedly contemplating solving the structures of 5,000 novel proteins in the next five years? Clearly major technological developments have to be made to achieve the goal. The second problem is one of target selection.

Structural genomics aims to define all

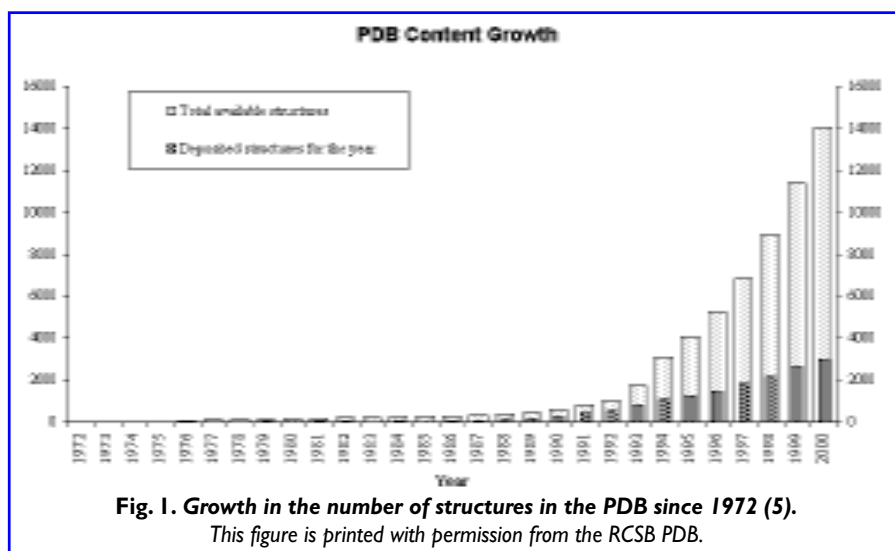


Fig. 1. Growth in the number of structures in the PDB since 1972 (5).

This figure is printed with permission from the RCSB PDB.

unique protein folds, but how do we ensure that proteins selected will have unique folds? The third problem is the over-riding concept of structural genomics itself, inferring function from structure. Ironically, if we succeed in selecting protein targets that have novel folds, then it becomes much tougher to sort out function.

The reason is that we cannot refer back to the function of proteins that share the same fold. The three big hurdles to overcome in structural genomics can thus be summarised as follows: (i) acceleration of structure determination by crystallography and NMR, (ii) selection of targets with novel folds, and (iii) inference of function from structure.

In this article we will look at some of the structural genomics efforts around the world and investigate how these problems are being addressed.

How many structures is enough?

The PDB currently holds about 15,000 protein structures that can be represented by a mere about 700 unique folds. Furthermore, despite the fact that the number of structures deposited each year is increasing exponentially, the proportion of new folds is steadily decreasing (Fig. 2). Put simply, this means that many pro-

teins with unique functions share similar structures.

The question is, how many protein structures are needed to complete the structural fold library? Ideally, one member of every sequence-defined protein family (usually characterised as having 30-35% identity) would be required to enable homology modelling of all other protein structures (4). To achieve this, 10,000–20,000 additional non-redundant protein structures would be required (6).

This is a daunting concept given how long it has taken to accumulate 15,000 redundant structures and given that only about 300 new structural folds were deposited to the PDB in 2000 (5). Clearly the rate of structure determination and the selection of targets needs to be improved if structural genomics is to succeed.

International rescue

Structural genomics initiatives encompass dozens of academic and commercial laboratories around the globe and represent a huge cross-section of the scientific community. In the United States, the National Institutes of Health recently announced funding of US\$150 million over the next five years for seven structural

Showcase on Research

Structural Genomics: Protein Structures for the Masses? (contin.)

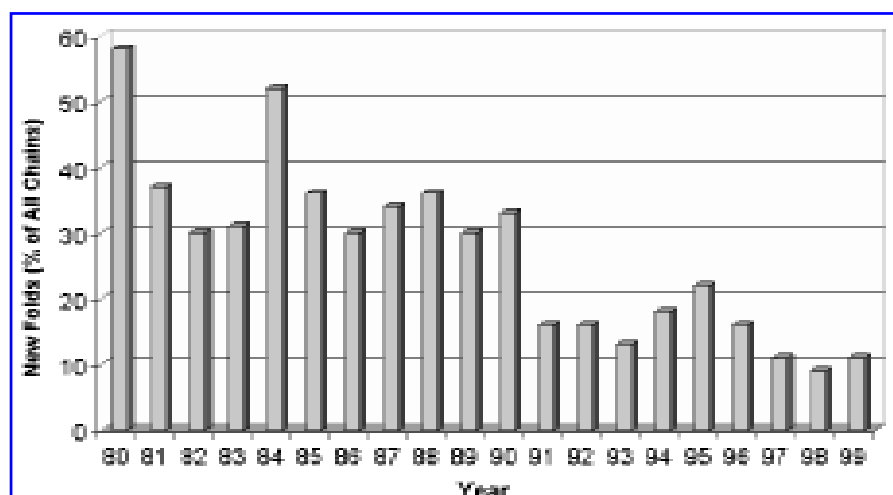


Fig. 2. Proportion of "new folds" for a given year (5).

This figure is printed with permission from the RCSB PDB.

genomics consortia. The funds will be used for solving new structures and to promote development of enabling technologies such as high-throughput protein production and high-throughput X-ray crystallography (7). The seven NIH-funded consortia each consist of five institutions, including laboratories across the USA and as far away as New Zealand.

This reflects not only a degree of cooperation between groups, but also the difficulty currently facing researchers in gathering the necessary resources (both physical and personnel) (7). The USA is also home to Structural Genomix and Syrrx, two San Diego-based companies that aim to solve the structures of thousands of protein drug targets.

In Europe, there are two European Commission-funded structural genomics groups focussing on the development of computational techniques to accelerate NMR and X-ray structure determination (8). In addition, there are nationally funded French and German projects that have identified target protein families for investigation and which will also develop high throughput technologies (8).

Meanwhile in Japan there are two major projects (9,10). The RIKEN Structural Genomics Initiative will use both crystallography and NMR techniques and will prioritise potential targets for structural genomics based on their potential biological and/or medical relevance (10).

The second Japanese project is funded by the Ministry of International Trade and Industry (MITI) (10). This group is targeting human membrane proteins (10). Mem-

brane proteins are notoriously difficult to work with, yet constitute a high proportion of the genome and represent the targets for many drugs. Clearly, success in this area will have knock-on effects for many other branches of science.

A major push in all these efforts is the development of new technologies to improve the speed of structure determination. Over the past ten years significant inroads have been made. Recombinant techniques, MAD phasing and automated chain tracing have dramatically reduced the time it takes to solve novel protein crystal structures.

Theoretically at least, if you can produce just one crystal of a protein derivatised with selenomethionine (SeMet) and you have access to a synchrotron, your new protein structure can be ready in a matter of days to weeks. This SeMet/synchrotron combination seems to be the approach most structural genomics initiatives are using to accelerate the process of structure determination. But it is not enough. Enabling technology is now being developed to automate all stages of the structure determination process, from high-throughput protein expression and cloning through to purification and crystallization (11). Efforts are also being directed to automating data measurement, model building and structure refinement.

Choices, choices, choices...

How do we decide which proteins to study? Most groups have chosen targets they consider will not only yield a high

number of unique structural folds, but which will also provide a useful legacy to the scientific research community. For example, there are initiatives aimed at characterising the entire proteome of pathogens such as *Mycobacterium tuberculosis*, *Haemophilus influenzae* and *Mycoplasma*. The outcomes could provide hundreds of potentially useful drug design targets – an important point considering the problems of antibiotic resistance. Other projects focus specifically on human proteins, opening the door for potential treatments of a plethora of diseases that are currently not well understood.

The choice of targets involves careful consideration of the benefits that the structures will yield (12). There is no doubt that broad-spectrum approaches will yield useful structures leading to interesting and beneficial outcomes. However, an alternative strategy of target selection, such as that shown in Fig. 3, may yield a higher proportion of structures that will be useful downstream.

The key difference between the two alternate selection methods lies in the genes that are chosen early on. For the "traditional" selection strategy, an entire proteome of an organism or cell type is chosen. Filters are applied to remove structural homologs so that only one representative of each protein family is chosen for structural analysis. This gives the required breadth of structural diversity but may only produce a small percentage of total structures that will be useful for other applications such as drug design and understanding of disease.

The alternative strategy outlines a method where 'interesting' genes are selected from the proteome at the initial steps. These may be genes implicated in pathogenesis, immunity, metabolic pathways or disease states. This results in a smaller number of targets, but with a higher proportion being relevant to drug design and understanding of disease.

Yes it's pretty, but what does it do?

A key aspect of structural genomics is that function can be elucidated from structure. Indeed structural genomics completely reverses the standard procedure in protein characterisation – structure is now a starting point, not an end point to understanding what a protein

Showcase on Research

Structural Genomics: Protein Structures for the Masses? (contin.)

does. The question is, how can function be derived from structure?

When a new protein structure is unexpectedly shown to have a known protein fold, the function of the unknown protein may be guessed by referring to the function of the structurally homologous partner(s). This 'knowledge-based' approach has proven to be remarkably powerful (13). The conundrum is that the driving force behind structural genomics is to select proteins on the basis of having unique folds. But if the new protein structure has a new protein fold, how then do we infer function?

Even in these cases, it is useful to compare the novel fold to known protein structures to see what it resembles most. Hwang *et al.* (14) did just that when they solved the structure of Mj0226, a novel *Methanococcus jannaschii* protein of unknown function. They found that one part of the protein was structurally similar to

proteins that had a variety of functions but which all bound nucleotides. Sure enough, Mj0226 binds nucleotides too, though the nucleotide binding mode and the biochemical function of the protein are both novel.

Other information can also be used. For instance, proteins can be categorised functionally into several broad groups: structural, enzymatic/metabolic, receptor, channel and nucleotide binding. Each of these groupings have their own structural idiosyncrasies and these can be used as a guide in categorising function. Thus, a hydrophilic protein shaped like a long filament is unlikely to be a membrane-bound receptor protein, but it could well be a structural protein. Analysis of electron density of novel proteins sometimes reveals the presence of bound ligands, such as ATP (15), and this can obviously help in function prediction. However, these are only loose guidelines and clearly a better

system of structure-function relations is necessary (13).

One approach is to use computer modelling to predict function, followed by laboratory based experiments to confirm (or otherwise) the prediction. Until now, there has been little need for *ab initio* methods of function prediction, so relatively little has been done (13). However, it is known that 70% of enzyme active sites can be identified by locating the largest cleft on the protein surface (16) and methods are being developed to do that. Similarly, neural networks can be used to identify protein-protein interaction domains with a 75% success rate (16,17).

Whilst these examples show potential, it is clear that as structural genomics moves forward, new methods for prediction of function will need to be developed. Even though computer simulation will be increasingly important, experimental techniques like the yeast two-hybrid system

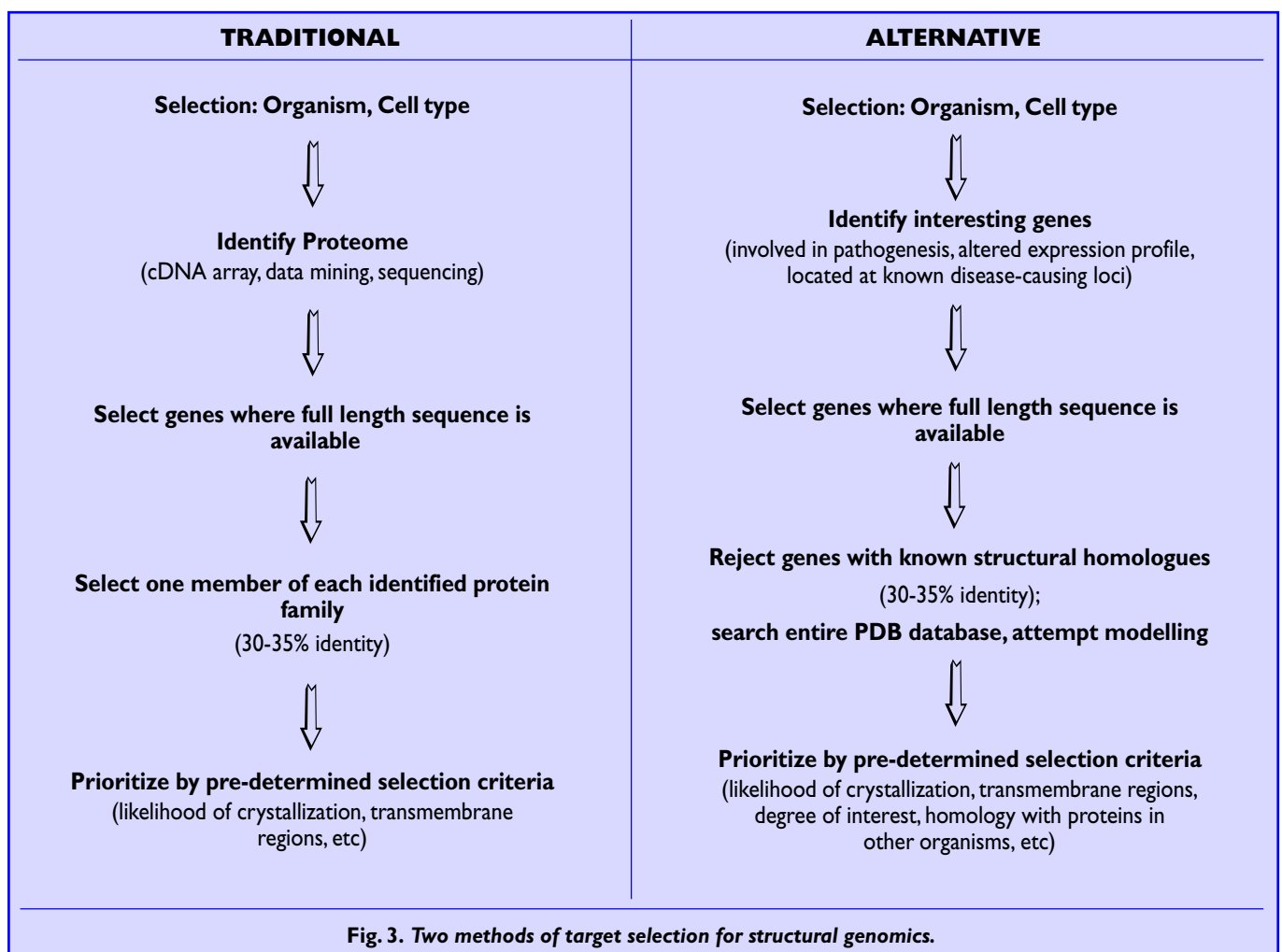


Fig. 3. Two methods of target selection for structural genomics.

Showcase on Research

Structural Genomics: Protein Structures for the Masses? (contin.)

and mutagenesis to prevent complex formation will continue to play a major role.

Structural genomics in Australia?!

There are many reasons why structural genomics has not taken off in Australia. Lack of strategic research funds, lack of a local synchrotron, the wide geographic spread of structural biology groups to name but a few. Despite these minor inconveniences, the authors have initiated an embryonic structural genomics program at the University of Queensland. We plan to use a focussed approach to investigate the structures of proteins expressed in activated macrophages.

By using directed target selection, we may even begin to reverse the ever-widening gap between the international effort in structural genomics and Australia's contribution. We anticipate that, in time, the program will deliver useful outcomes for biotechnological development. The program represents a fresh approach to drug design for infectious disease, it will provide a better understanding of the immune system, and could lead to the development of potential anti-cancer therapies.

This is an exciting new direction for structural biology in Australia. Other exciting developments, such as the recently announced construction of an Australian

synchrotron, will further boost Australia's profile in a field that appears set to revolutionise the future of biochemistry and molecular biology.

References

- Venter, J. C. et al. (2001) *Science* **291**, 1304-1351
- Lander, E. S. et al. (2001) *Nature* **409**, 860-921
- Shapiro, L., and Lima, C. D. (1998) *Structure* **6**, 265-267
- Burley, S. K. (2000) *Nat. Struct. Biol.* **7 Suppl.**, 932-934
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Research* **28**, 235-242
- Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M. A., Madhusudhan, M. S., Mirkovic, N., and Sali, A. (2000) *Nat. Struct. Biol.* **7 Suppl.**, 986-990
- Terwilliger, T. C. (2000) *Nat. Struct. Biol.* **7 Suppl.**, 935-939
- Heinemann, U. (2000) *Nat. Struct. Biol.* **7 Suppl.**, 940-942
- Yokoyama, S., Matsuo, Y., Hirota, H., Kigawa, T., Shirouzu, M., Kuroda, Y., Kurumizaka, H., Kawaguchi, S., Ito, Y., Shibata, T., Kainosho, M., Nishimura, Y., Inoue, Y., and Kuramitsu, S. (2000) *Prog. Biophys. Mol. Biol.* **73**, 363-376
- Yokoyama, S., Hirota, H., Kigawa, T., Yabuki, T., Shirouzu, M., Terada, T., Ito, Y., Matsuo, Y., Kuroda, Y., Nishimura, Y., Kyogoku, Y., Milki, K., Masui, R., and Kuramitsu, S. (2000) *Nat. Struct. Biol.* **7 Suppl.**, 943-945
- Edwards, A. M., Arrowsmith, C. H., Christendat, D., Dharamsi, A., Friesen, J. D., Greenblatt, J. F., and Vedadi, M. (2000) *Nat. Struct. Biol.* **7 Suppl.**, 970-972
- Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W., and Swaminathan, S. (1999) *Nat. Genet.* **23**, 151-157
- Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., and Orengo, C. A. (2000) *Nat. Struct. Biol.* **7 Suppl.**, 991-994
- Hwang, K. Y., Chung, J. H., Kim, S.-H., Han, Y. S. and Cho, Y. (1999) *Nat. Struct. Biol.* **6**, 691-696
- Zarembinski, T. I., Hung, L.-W., Mueller-Dieckmann, H.-J., Kim, K.-K., Yokota, H., Kim, R., and Kim, S.-H. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 15189-15193
- Laskowski, R. A., Luscombe, N. M., Swindells, M. B., and Thornton, J. M. (1996) *Protein Sci.* **5**, 2438-2452
- Jones, S., and Thornton, J. M. (1997) *J. Mol. Biol.* **272**, 121-32, 133-43.



