PERGAMON

Progress in Biophysics & Molecular Biology 73 (2000) 289–295

Progress in Biophysics & Molecular Biology

www.elsevier.com/locate/pbiomolbio

Review

Structural genomics: an overview

Tom L. Blundell*, Kenji Mizuguchi

Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK

Contents

1. Introduction

The availability of complete sequences for the genomes of many bacteria (see, for example, http://www.tigr.org/tdb/mdb/mdb.html) and some model eukaryotes including yeast (http://genome-www.stanford.edu/Saccharomyces/; Mewes et al., 1997), worm (http://www.sanger.ac.uk/Projects/C_elegans/; *C. elegans* Sequencing Consortium, 1998) and fly (http://www.fruitfly.org/; Adams et al., 2000), together with the emerging draft sequences for man, has focused attention on the functions of the gene products, the proteins. In this respect, the three-dimensional structure of the protein is likely to be helpful, as its function is almost always mediated by structure, and tertiary structure is more conserved in evolution than sequence. Efforts to define the

*Corresponding author. Tel.: +44-1223-333628; fax: +44-1223-766082.
E-mail address: tom@cryst.bioc.cam.ac.uk (T.L. Blundell).

PII: S0079-6107(00)00008-0

three-dimensional structures of all gene products revealed by genome sequencing are known as structural genomics.

The idea of structural genomics can be realised only if structures can be defined quickly and cheaply (Service, 1999). This has focused attention on high-throughput methods for protein expression, purification, characterisation and structure determination. With the increasing success of these high-throughput methodologies, structures of gene products might be defined quite straightforwardly perhaps, within a week, from gene to protein structure. The percentage of structures that might be defined quickly in this way will depend on the extent of post-translational modification and whether they are multidomain proteins which may not be accessible to NMR analysis and may need dissecting into domains or complexing with nucleic acids, proteins or polysaccharides before they can be crystallised. Thus, many structures will take longer, and may be more expensive to determine.

In this volume, we have brought together a number of reviews that address the challenges of structural genomics. The first challenge is target selection. Should all gene products be studied? Or should we select just those where nothing is known about function, or those that are expressed in particular cells or in certain disease states? For all aspects of target selection, it will be necessary to derive as much knowledge as possible from a bioinformatics approach. The second challenge will be to express and purify the proteins in a state suitable for structural analysis. The third will be to accelerate methods for structure determination, principally by developing high throughput X-ray crystallography and NMR analysis. Finally, it will be necessary to develop methods for inferring function and mechanism from structure if we capitalise on the new knowledge about the three-dimensional structures.

2. Target selection

In accepting that the main objective of structural genomics is to learn more about function, it is useful to remember that function will mean very different things to the evolutionary biologist, to the biologist involved in whole organisms and their communities, to the geneticist and to the biochemist or cell biologist. In each case function will be operationally defined. For structural genomics it is probably best defined as the interactions of the protein with other molecules, whether small or large, and whether substrates, receptors, regulators or adaptors.

Analyses of genome sequences have shown that functions of very few proteins have previously been identified from genetics or biochemistry, although functions for about 50% can be inferred with reasonable confidence from knowledge of close homologues (see, for example, Adams et al., 2000). For other proteins, clues about function may be gained from distant sequence patterns or motifs that are characteristic of a superfamily (Bork and Koonin, 1998). The functions of other proteins, perhaps as many as 30%, may not be defined (Rubin et al., 2000), often because they are members of families with undefined functions. Alternatively they may be difficult to recognise on the basis of sequence alone, or they have no relatives in sequenced genomes.

In view of the very large numbers of sequences coding for proteins, particularly in eukaryotes, it is necessary to prioritise or select targets. One approach is to select representatives of each homologous family or even of each superfamily. The total number of homologous families can be operationally defined as the number of representative sequences whose neighbours (e.g., those

within 30% sequence identity, the similarity level often required for accurate comparative modelling) jointly cover a certain percentage (e.g., 90%) of the sequence space. This number is probably ten thousand or more for a reasonable coverage (Burley et al., 1999; see also http://www.nigms.nih.gov/news/meetings/structural_genomics_targets.html and Linial and Yona (2000, this volume)), but these homologous families can be clustered into a thousand or so superfamilies, the exact number depending on how many single member families there turn out to be. We can assume that the great majority, although certainly not all, of the members of a homologous family will have similar functions. For superfamilies that have divergently evolved, with retention of general tertiary structure but perhaps less than 25% sequence identity, a related function may be retained. Selection of primary targets could be further focused by choosing representative structures of families only where there is no clue to function. Alternatively we could choose only "core families", those that are common to most genomes. These and other approaches are discussed in Linial and Yona (2000, this volume) and Heger and Holm (2000, this volume).

For complex eukaryotic genomes such as those of mouse or human, there are likely to be more than 20,000 gene products that cannot be recognised by the best methods of bioinformatics. For such complex genomes it might be best to focus on groups of gene products that are characteristic of particular, perhaps diseased cells. Structural genomics should, therefore, be coordinated with good programmes of proteomics, where proteins expressed in particular cells are separated by polyacrylamide gel electrophoresis (PAGE) and identified with mass spectrometry linked to sequence databases and bioinformatics software (Yates, 2000).

Identification of homologous families and superfamilies requires the development of very sensitive approaches to sequence search and alignment. Databases such as SCOP (Murzin et al., 1995), CATH (Orengo et al., 1997), FSSP (Holm and Sander, 1996), CAMPASS (Sowdhamini et al., 1998) and HOMSTRAD (Mizuguchi et al., 1998) that classify protein structures into homologous families, superfamilies and folds will be an important reference source. Databases of aligned sequences will be invaluable; they can be based on either sequence comparisons such as Pfam (Bateman et al., 2000), or structure comparisons such as FSSP (Holm and Sander, 1996), CAMPASS (Sowdhamini et al., 1998) and HOMSTRAD (Mizuguchi et al., 1998). Such databases as well as various algorithms are discussed in Linial and Yona (2000, this volume) and Heger and Holm (2000, this volume). The BLAST family of programs (Altschul et al., 1990) can be used to scan DNA and protein sequence databases and multiple sequence alignments generated using, for example, CLUSTALW (Higgins et al., 1996). For more distant members of superfamilies, methods using profile hidden Markov models (e.g., HMMER (http://hmmer.wustl.edu/)) or fold recognition or threading procedures will be needed (e.g., GenTHREADER (Jones, 1999), FUGUE (http://www-cryst.bioc.cam.ac.uk/~fugue/)). It is probably also wise to construct models where there are homologues of known structure (Sanchez et al., 2000). The models will have value in their own right, but they are also useful for validation of the fold recognition procedure.

3. Protein expression, purification and characterisation

Perhaps the best way to optimise chances of success in structural genomics is to choose a thermophilic organism with a small genome. Such an approach has been pioneered by Sung-Hou Kim at Berkeley, who has chosen *Methanococcus jannaschii*, a microbe that lives in the high

temperatures and pressures around volcanic vents (Kim et al., 1998). The choice of organism maximises the chance of expression in *Escherichia coli* of stable, soluble proteins, suitable for X-ray or NMR analysis. However, exploratory analyses of several organisms demonstrate that only a small percentage of gene products will easily give soluble proteins that have a well-defined structure as indicated by circular dichroism, fluorescence spectroscopy or differential scanning calorimetry, and an even smaller percentage will give crystals (see Christendat et al., 2000, this volume). Most groups, such as that of John Moult at the Center for Advanced Research in Biotechnology in Maryland studying *Haemophilus influenzae*, the microbe responsible for meningitis, have established conditions for expression of about 50 proteins in their first year of activity but solved the structures of only a few proteins, although more are promised. As the projects scale up, there will be an increasing need to automate the expression of the proteins; cell free systems may be helpful in this respect (Kigawa et al., 1999). There will also be a need to have proteins expressed as selenomethionine proteins for multiple anomalous dispersion methods in X-ray analysis or labelled with ^{13}C and/or ^{15}N for NMR structure determination (see Heinemann et al., 2000, this volume).

In general, the first task is to recognise and remove those low complexity sequences that are unlikely to form globular structures (Write and Dyson, 1999) but can be found in as many as 15,000 proteins in the SWISS-PROT database (Romero et al., 1998). Second, it will be necessary to identify sequences of proteins residing in cellular membranes, which are predicted to constitute 20% of the gene products in *Drosophila* (Adams et al., 2000). These membrane proteins will need a different approach to expression, purification and crystallisation. Although the success in crystallisation of intrinsic membrane proteins is also improving, this has been mainly for those that, like channels or electron transfer molecules, do not involve conformational change as part of their mechanism of action. However, many membrane proteins have regions that are extrinsic to the membrane, and these can be expressed separately if they can be identified. This has indeed been a very successful approach for cell surface receptors and adhesion molecules. Nevertheless, this will require special attention and care.

Many proteins contain multiple domains, often with flexible linkers. This will demand special care in identifying domain boundaries and removing large flexible regions. A further challenge arises from post-translational modification, which can give rise to heterogeneity especially in cases of partial auto-phosphorylation with protein kinases and in complex glycosylation of extracellular proteins. Finally, many extracellular proteins have complex patterns of disulphide bridges, which give rise with bacterial expression to inclusion bodies that are difficult to refold. Many of these factors point towards expression in insect, yeast or mammalian cells and will undoubtedly slow progress. This situation is likely to be more often the case for higher eukaryotic gene products. Indeed, more multidomain proteins have been identified in *Drosophila* and *C. elegans* compared to yeast and in particular, *Drosophila* has bigger and more heterogeneous multidomain proteins, with a larger number of extracellular domains than *C. elegans* (Rubin et al., 2000).

4. Structure determination

The time and cost of successful structure determination urgently needs reducing. For X-ray analysis the high cost of beam time at synchrotron sources has already been a driver to improve

the speed and sensitivity of detectors. However, much needs to be done in automation of all steps in the analysis (Heinemann et al., 2000, this volume). The introduction of robots for crystallisation has received much attention, but they are little used in practice. This may be partly due to the poor psychology of entrusting valuable protein to a robot, but partly due to the good sense of carrying out small-scale screens to scope the problem. With automated protein production this should change; automated screening of videos for identifying crystals should further contribute to this process. An urgent need is to automate crystal handling and changing devices. Multiple anomalous dispersion methods combined with solvent modification have done much to speed up the phase calculation. For good quality electron density maps, we are already close to methods for automatic interpretation.

For NMR structure analysis, there has been less emphasis on large central facilities and probably less coordination and cooperation with analysis of the data. The development of 900 MHz and higher field machines is increasingly making it difficult to integrate them into laboratory buildings and the cost is driving most agencies to fund central facilities. The Japanese government through RIKEN is planning a large centre with several 800 and 900 MHz machines in one centre (Yokoyama et al., 2000, this volume).

5. Identification of function

The value of the three-dimensional structure of a protein for suggesting function will depend on knowledge inferable from homologues. Where close homologues or even distant members of a superfamily with known function have been identified, the three-dimensional structure may provide a basis for new hypotheses about specificity of interactions with substrate, ligand or other partner mediating function. Where no obvious relatives can be identified, knowledge of the three-dimensional structure may give clues about a superfamily relationship and so suggest ideas about function or even key residues that can be tested.

So far very little attention has been paid to the general problem of identifying function, given three-dimensional structure. Much can be done by identifying functionally important residues from sequence variation, perhaps by using environment-specific substitution tables as suggested by Overington et al. (1990) or by evolutionary trace analysis (Lichtarge et al., 1996). Further algorithms also need to be implemented for transferring knowledge of binding sites of known superfamily members to new structures of other members, perhaps in the manner developed by Russell et al. (1998). Even if the structure shows no similarity to any other proteins, some structural features such as electrostatic surface and the spatial clustering of disease mutants can present hypotheses about function, which subsequently can be tested using other biological techniques (Boggon et al., 1999).

6. The organisation of structural genomics initiatives

There are two current approaches to the organisation of structural genomics initiatives. The distributed approach has been supported by the NIH in the USA (http://www.structuralgenomics.org/). The emphasis has been on linking large numbers of laboratory groups with complementary skills to provide a coordinated programme. The strength of this approach is its

flexibility and its ability to respond to new ideas from many bright people; it is illustrated by the initiatives of Moult and collaborators in Maryland. Its weakness is perceived by some as its inability to gain critical mass in carrying out large-scale expression, preparative biochemistry, crystallisation, X-ray or NMR analysis. Thus, the centralised facility/factory approach has been supported in Japan with the RIKEN NMR structure determination project (Yokoyama et al., 2000, this volume) and Protein Structure Factory in Berlin (Heinemann et al., 2000, this volume).

7. Applications to drug discovery

Structural genomics is likely to play an important role in drug discovery, particularly in target identification. There will be much interest in the smaller genome sequences of pathogens such as viruses, fungi and bacteria. The identification of HIV proteinase from the viral genome and its targeting for AIDS antivirals has already demonstrated the success of this way forward. With respect to the human genome, the complexity will demand a focused approach. For many this will be an extension of the systems approach where new sequences of homologues of known and useful targets are selected from genome sequencing projects for structural analysis. Alternatively comparative genomics can be used to identify key regulatory or signalling proteins from genetics of model organisms such as *Drosophila melanogaster* or *C. elegans* that also occur in the human genome. A more profitable approach may be to identify genes that are associated with the onset of disease using a proteomics approach.

The challenge with any approach to drug discovery where information is collected without a necessary focus on a disease target is that the information must be cheap. Even if a protein structure can be defined at an average cost of $100,000, the real cost will be still higher, depending on the percentage of proteins in the genome that are useful drug targets. Currently, the pharmaceutical industry worldwide is investigating about 500 targets. Drews (2000) estimates that there may be 100 multifactorial diseases that pose major medical problems, which implies about 1000 targets if we assume 10 genes per disease. These, Drews argues, will each be linked to perhaps five other proteins in physiological or pathological circuits. The good news is that there are probably ten times as many targets as in use today. The bad news is that not all will be easily inhibited by small molecules that make most useful drugs — druggable in the current jargon. Furthermore, less than 10% of the genome will be useful, implying a real cost of ten times $100,000, or $1 million per target. Even with a worldwide drug discovery budget in big pharma of over $40 billion, this may be a little high. It is clear that structural genomics companies will need to focus and it is likely that structural genomics start-up companies will provide information to a large number of big pharma.

References

Adams, M.D., et al., 2000. The genome sequence of *Drosophila melanogaster*. Science 287, 2185–2195.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., Sonnhammer, E.L., 2000. The Pfam protein families database. Nucleic Acids Res. 28, 263–266.

Boggon, T.J., Shan, W.S., Santagata, S., Myers, S.C., Shapiro, L., 1999. Implication of tubby proteins as transcription factors by structure-based functional analysis. Science 286, 2119–2125.

Bork, P., Koonin, E.V., 1998. Predicting functions from protein sequences — where are the bottlenecks? Nat Genet. 18, 313–318.

Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland, T., Lin, D., Šali, A., Studier, F.W., Swaminathan, S., 1999. Structural genomics: beyond the Human Genome Project. Nature Genetics 23, 151–157.

The *C. elegans* Sequencing Consortium 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science 282, 2012–2018.

Christendat, D., Yee, A. Dharamsi, A., Kuger, Y., Gerstein M., Arrowsmith, C.H., Edwards, A.M., 2000. Structural proteomics: Prospects for high throughput sample preparation. Prog. Biophys. Mol. Biol. 73, 339–345.

Drews, J., 2000. Drug discovery: a historical perspective. Science 287, 1960–1964.

Heger, A., Holm, L., 2000. Towards a covering set of protein family profiles. Prog. Biophys. Mol. Biol. 73, 321–337.

Heinemann, U., Frevert, J., Hofmann, K.-P., Illing, G., Maurer, C., Oschkinat, H., Saenger, W., 2000. An integrated approach to structural genomics. Prog. Biophys. Mol. Biol. 73, 347–362.

Higgins, D.G., Thompson, J.D., Gibson, T.J., 1996. Using CLUSTAL for multiple sequence alignments. Methods Enzymol 266, 383–402.

Holm, L., Sander, C., 1996. Mapping the protein universe. Science 273, 595–602.

Jones, D.T., 1999. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287, 797–815.

Kigawa, T., Yabuki, T., Yoshida, Y., Tsutsui, M., Ito, Y., Shibata, T., Yokoyama, S., 1999. Cell-free production and stable-isotope labeling of milligram quantities of proteins. FEBS Lett. 442, 15–19.

Kim, K.K., Hung, L.-W., Yokota, H., Kim, R., Kim, S.-H., 1998. Crystal structures of eukaryotic translation initiation factor 5A from *Methanococcus jannaschii* at 1.8 Å resolution. Proc. Natl. Acad. Sci. USA 95, 10419–10424.

Lichtarge, O., Bourne, H.R., Cohen, F.E., 1996. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257, 342–358.

Linial, M., Yona, G., 2000. Methodologies for target selection in structural genomics. Prog. Biophys. Mol. Biol. 73, 297–320.

Mewes, H.W., et al., 1997. Overview of the yeast genome. Nature 387, 7–65.

Mizuguchi, K., Deane, C.M., Blundell, T.L., Overington, J.P., 1998. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci. 7, 2469–2471.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., 1997. CATH — A hierarchic classification of protein domain structures. Structure 5, 1093–1108.

Overington, J.P., Johnson, M.S., Sali, A., Blundell, T.L., 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc. R. Soc. Lond. B. Biol. Sci. 241, 132–145.

Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Garner, E., Guilliot, S., Dunker, A.K., 1998. Thousands of proteins likely to have long disordered regions. Pac. Symp. Biocomput. 3, 437–448.

Rubin, G.M., et al., 2000. Comparative genomics of the eukaryotes. Science 287, 2204–2215.

Russell, R.B., Sasieni, P.D., Sternberg, M.J.E., 1998. Supersites within superfolds. Binding site similarity in the absence of homology. J. Mol. Biol. 282, 903–918.

Sanchez, R., Pieper, U., Mirkovic, N., de Bakker, P.I., Wittenstein, E., Sali, A., 2000. MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res. 28, 250–253.

Service, R.F., 1999. Wiggling and undulating out of an X-ray shortage. Science 285, 1342–1346.

Sowdhamini, R., Burke, D.F., Huang, J.-F., Mizuguchi, K., Nagarajaram, H.A., Srinivasan, N., Steward, R.E., Blundell, T.L., 1998. CAMPASS: a database of structurally aligned protein superfamilies. Structure 6, 1087–1094.

Write, P.E., Dyson, H.J., 1999. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. J. Mol. Biol. 293, 321–331.

Yates, J.R., 2000. Mass spectrometry from genomics to proteomics. Trends Genet. 16, 5–8.

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., Kuramitsu, S., 2000. Structural genomics projects in Japan. Prog. Biophys. Mol. Biol. 73, 363–376.