PERGAMON

Progress in Biophysics & Molecular Biology 73 (2000) 339–345

Progress in Biophysics & Molecular Biology

www.elsevier.com/locate/pbiomolbio

Review

Structural proteomics: prospects for high throughput sample preparation

Dinesh Christendat[a], Adelinda Yee[a], Akil Dharamsi[a,†], Yuval Kluger[c], Mark Gerstein[c], Cheryl H. Arrowsmith[a,*], Aled M. Edwards[a,b]

[a] *Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, 610 University Avenue, Toronto, Ontario, Canada M5G2M9*

[b] *C.H. Best Institute, Banting and Best Department of Medical Research, University of Toronto, 610 University Avenue, Toronto, Ontario, Canada M5G2M9*

[c] *Department of Molecular Biophysics and Biochemistry, Yale University, Connecticut, USA*

Accepted 20 June 2000

Contents

1. Background

With the near completion of many genome sequencing projects has come the sobering realisation that our understanding of biology is nowhere near complete. For example, in the worm, *C. elegans*, less than half of the predicted proteins have a known function (Consortium, 1998). The major challenge facing biologists in the next decade will be to "finish the job", that is, to ascribe a function to each of the proteins that have been discovered

*Corresponding author. Tel.: +1-416-946-2017; fax: +1-416-946-6529.
E-mail address: carrow@oci.utoronto.ca (C.H. Arrowsmith).
† Current address: Integrative Proteomics Inc., 100 College St, Toronto, ON, Canada.

PII: S0079-6107(00)00010-9

in the sequencing phase of the genome projects. As was necessary in the sequencing phase of the genome project, success in the "functional genomics" phase will require a considerable shift of experimental strategy, leaning away from hypothesis-driven approaches focussing on individual molecules to ones that are more oriented towards comprehensive, large-scale analysis. Moreover, because the analysis of protein function is considerably more complicated than gene sequencing, functional genomics will also require the integration of many different, yet synergistic, technologies.

Structural biology is emerging as one of the core areas of functional genomics because the three-dimensional structure of a protein can often provide functional clues, primarily by detecting *structural* homology with a protein of known function (Zarembinski et al., 1999; Cort et al., 1999). This method of ascribing function will prove particularly valuable since, at present, many structural and functional homologies escape detection by sequence-based approaches (Teichmann et al., 1999). The concept of determining protein structures on a genome-wide scale has been coined "structural genomics" (Kim, 1998; Shapiro and Lima, 1998), though "structural proteomics" may be more accurate — as this reflects the aim to determine protein structures and also avoids the term "structural genomics", which had been used previously in the literature to describe another research area (Heiter and Boguski, 1997).

Structural proteomics has two parallel goals, whose relative merit is a matter of debate. One approach is to determine the three-dimensional structures of a small but diverse subset of proteins that would be selected carefully to represent "a basis set" of protein folds. The idea is that by determining these structures, many other structures could be modelled using computational techniques. This approach is useful for addressing broad questions relating to the phylogenetic distribution of folds and structural features (Cort et al., 1999; Burley et al., 1999). However, it is often the subtle *differences* in protein structure that contribute to their precise function. Current modelling techniques are not accurate enough to reveal these subtleties when starting from a template that is not identical (see for example the results of the CASP competitions, Moult et al., 1997). Thus, the second goal, which we favour, is to determine a considerably larger number of structures from a number of model organisms so as to completely map the structures of their proteomes. The differences between these goals is to some degree determined by the accuracy of modelling algorithms and our relative satisfaction with knowing "folds" in constrast to "structures." In more specific quantitative terms, the former approach might consider a protein structure "known" if a homolog could be found by advanced fold recognition techniques such as PSI-blast and threading, whereas the latter approach would require a close homolog, say at 70% identity, which certainly shares the same function and active site geometry. Eventually, we expect that modelling algorithms will improve such that computational techniques will be able to "fill in the gaps" between a sufficiently populated database of high-resolution 3D structures, giving us a reasonable structure for every sequence. At this point, the aims of structural proteomics efforts will evolve towards solving the structures of protein complexes.

The major challenge of structural proteomics, however, is not to define the concept, but rather to reduce the idea to practice in a realistic and cost-effective way. Until a few years ago, the process of structure determination was not developed to the point that a high-throughput approach was feasible. However, with the recent advent of synchrotron radiation and MAD

phasing, the rapidly improving methods of nuclear magnetic resonance and the advances in computational speed, structural proteomics is becoming more realistic (Terwilliger et al., 1998; Kim, 1998; Burley et al., 1999). The major remaining barrier to success is the high-thoughput preparation of suitable samples for structure determination: well-diffracting crystals for X-ray diffraction and soluble, well-behaved samples for NMR studies. To assess the magnitude of the technical hurdles that would be encountered in generating excellent structural samples on a genome-wide scale, we initiated a pilot study of 500 proteins in the proteome of *Methanobacterium thermoautotrophicum* (MT). Based on these studies, which involved almost a third of the proteins in the genome, we were able to estimate the percentage of proteins that are immediately amenable to structure analysis versus those that will need further optimisation of expression and/or sample conditions. The former number is fairly low, especially for larger proteins and this suggests that the best strategy for target selection is to screen as many genes as possible for protein expression and solubility, rather than targeting a representative subset based on a priori "desirable" properties such as the possibility of having a new fold.

1.1. Goals of the MT pilot project

The rate-determining step in any structural proteomics project is the generation of excellent samples. The MT pilot project was initiated in order to critically assess the difficulties in this process. The goal of the project was not to determine three-dimensional structures per se, but rather to identify the bottlenecks in the process. An associated goal was to gain insight into the structural characteristics of proteins on a genome-wide scale.

MT was selected for four reasons. First, the complete genome sequence was known. Second, the genome lacks introns; therefore, the genomic DNA provided an inexpensive source of DNA. Third, MT is a thermophile, and thermophilic proteins are often easier to purify than those from mesophiles. Fourth, of the three different thermophilic organisms tested (MT, *Methanococcus jannaschii* and *Archeoglobus fulgidus*), the genomic DNA of MT proved to be the best substrate for PCR cloning.

The target selection criteria for the project involved only two exclusionary criteria. First, proteins which had known 3D structures or for which a structural homologue was known, were not included in the 500-protein list. Second, proteins with membrane-spanning domains were treated separately from non-membrane proteins, since the purification and crystallization of membrane proteins is not yet amenable to high-throughput approaches. The criteria for both finding a known structure match and for identifying a membrane-spanning region were deliberately chosen to be simple, fast, and straightforward; a pairwise sequence comparison with BLAST (Altschul et al., 1990) was used to identify structural homologues, and the TMHMM program (Sonnhammer et al., 1998) was used to identify proteins with transmembrane helices.

A total of 463 non-membrane proteins selected in this manner were divided roughly into large (>200 amino acids) and small proteins. The smaller proteins were selected so that we could explore the feasibility of using NMR in structural proteomics. A total of 37 proteins predicted to be membrane transporters were also chosen in order to explore the feasibility of cloning, expressing and purifying membrane proteins for structural studies.

2. Results

2.1. Cloning and expression

Within a nine-month period, 500 proteins from MT were cloned into T7 RNA polymerase-producing expression vectors encoding hexa-histidine fusion proteins, and the level of expression in bacteria (*E. coli*) was tested. In general, this study revealed that most ($\sim$80%) of the MT proteins could be expressed in bacteria, including most of the membrane proteins. However, efficient expression of many MT proteins required the co-expression of extra copies of three tRNAs, which are frequently used by archeons and eukaryotes but are poorly expressed in *E. coli*.

2.2. Large-scale expression for structural studies

A total of 393 of the non-membrane proteins were selected for further analysis of their solubility and suitability for structural studies using either NMR or X-ray crystallography. The results are summarised in Fig. 1. Solubility of large proteins was tested in 50 ml cultures and proteins that appeared in the soluble fraction of the cell lysate were expressed in large-scale cultures (1–2 L) in rich media. The smaller proteins destined for NMR analysis, were expressed directly in 1 L of ^{15}N-enriched minimal media. All proteins were purified using metal affinity chromatography. For "small" proteins, the ^{15}N-labelled hexa-histidine fusion proteins were concentrated by ultrafiltration and the ^{15}N-HSQC NMR spectrum taken. The large proteins required additional purification steps. In all cases, the hexa-histidine tag was removed by proteolytic cleavage and the protein subjected to size-exclusion chromatography followed by

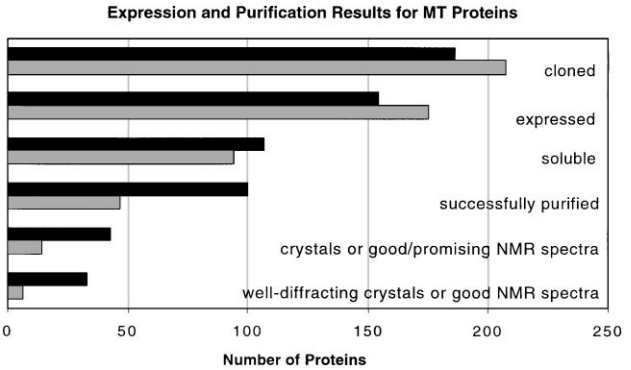

Fig. 1. Summary of expression and purification results for 393 small (<200 residues; dark bars) and large (light bars) proteins from *M. thermoautotrophicum*. A judgement as to whether a protein was expressed or not was made based on the presence of an induced band of the appropriate molecular weight in the crude *E. coli* lysate. Solubility was judged from the presence or absence of the induced band in the soluble fraction of the lysate. The lowest pair of bars represent protein samples that immediately gave well diffracting crystals or an HSQC NMR spectrum that contained the expected number of well-resolved peaks. The second lowest pair of bars also includes proteins that produced micro crystals or poorly diffracting crystals, and small proteins that had HSQC spectra with many of the features of soluble globular proteins, but for which the peaks were broader than expected (probably oligomeric proteins) or a significant number of peaks were missing.

concentration by ultrafiltration. Large proteins that "survived" this process were screened for crystal formation using factorial screens. More than 20 structure determinations of MT proteins are currently underway or completed using both X-ray crystallography and NMR. The structural results will be reported elsewhere. Membrane proteins are not amenable to the above procedures and results for these proteins will be reported elsewhere.

3. Discussion

In general, over half of the expressed proteins were soluble, with a larger proportion of the smaller proteins being soluble (69%) than the larger ones (54%). Proteins whose sequence gave a good database match to one in another organism (but not necessarily with a precise functional annotation) or to a paralog within MT were considerably more likely to be expressible and soluble. This latter finding perhaps reflects the fact that some proteins without a database match may not even be real — just being erroneous results of ab initio gene-prediction methods. Surprisingly, a retrospective analysis revealed that a number of factors that one may have thought as related to insolubility were, in fact, not at all connected with it — in particular, the occurrence of low-complexity regions (as defined by SEG, Wootton and Federhen, 1996), the predicted secondary structure composition, and simple statistics based on the overall amino acid composition. However, retrospective analysis also shows that a number of the insoluble proteins could be reliably flagged by more elaborate computational methods for membrane protein identification (Gerstein, 1998; Boyd et al., 1998).

Even if proteins were found to be in the soluble fraction of the cell lysate, there were often further losses due to precipitation during the purification process. However, in both cases roughly half of the successfully purified proteins gave promising results in the form of either protein crystals or a reasonable quality HSQC spectrum. Overall, only $\sim 10\%$ of large proteins and $\sim 15\%$ of small proteins yielded immediately to structure analysis. These figures provide what we feel is an accurate estimate of the "low hanging fruit" of the proteome.

The dominant factor that appears to be confounding the structural analysis of both large and small expressed proteins is the relatively small number of proteins which are soluble. For example, only about two-thirds of the proteins that are expressed are in the soluble fraction of the cell lysate, and of these, only a fraction remained soluble after concentration. The efficiency of harvesting the low hanging fruit can be improved by approaches that screen for the expression of soluble protein in *E. coli* (Waldo et al., 1999). However, a major goal of our subsequent studies will be to uncover the fundamental mechanisms that control the solubility properties of the remaining, less soluble proteins. We are considering three explanations. First, the solubility properties of recombinant proteins expressed in bacteria could be determined predominantly by the surface properties of globular, folded proteins. Patches of hydrophobic residues, normally used to mediate protein–protein or protein–ligand interactions, could mediate non-specific aggregation. Alternatively, the recombinant protein might be partially unfolded (even transiently), and transiently exposed hydrophobic surfaces might aggregate. Second, multi-domain proteins, when expressed in bacteria, have a propensity to aggregate, most likely due to the aberrant interaction of the hydrophobic cores of each of the domains with each other during the protein folding process. Finally, the overall charge state or charged surfaces of the protein

may affect its solubility under a given set of buffer conditions. In actual fact, all of these mechanisms may be operating. The aim of our subsequent work will be to quantify the relative contribution of these factors. The most simplistic approach is to screen for alternate buffer conditions that would improve solubility. However, this will not help with proteins that precipitate in *E. coli* or during the purification process for which the buffer conditions are limited.

Thus, we are exploring several other strategies to increase the fraction of proteins that can be expressed and purified in a soluble form. If the aggregation of the recombinant protein in bacteria results from the interaction of surface hydrophobic patches that normally mediate the interaction with another protein or ligand, then supplying the partner protein, or the specific ligand, should increase the solubility. In practice, our group has studied many proteins that form obligate heterodimers or heterotrimers, and inevitably the solubility of each of the proteins is dramatically increased by co-expression of the protein partners. The need to co-express what might be a large fraction of proteins in any genome (particularly eukaryote genomes) underscores the importance of a parallel protein–protein interaction analysis as part of a comprehensive structural proteomics effort. Finally, the percentage of soluble proteins should be increased dramatically by focussing on single domains as opposed to multidomain proteins. One of our major efforts will be to develop methods for the identification and structural analysis of single domains within complex proteins.

4. Conclusions

In summary, our biophysical and structural analysis of 500 proteins from the genome of *Methanobacterium thermoautotrophicum* has yielded the following insights.

1. The dominant experimental barrier to a proteome-wide structural analysis will be the large number of proteins that are insoluble either when expressed in bacteria or when concentrated for structural studies. The relatively low proportion of soluble proteins suggests that "targeted" selection criteria that seek to identify new folds will have a similar success rate, and that an unbiased target selection strategy might ultimately reap a higher percentage of new folds. However, further studies will be required to reveal whether the proportion of soluble proteins found here will be general or whether they are specific to the archaeon we have selected.
2. The relatively high number of small proteins in any genome, along with the rapidity of preliminary NMR analysis, suggests that NMR will have a central role in any structural proteomics effort, not only to characterize newly produced proteins, but also to determine their structures.
3. Comprehensive structural proteomics efforts must be closely linked to efforts that seek to identify protein–protein and protein–ligand interactions. This information will be critical in the analysis of many eukaryotic proteins.
4. Finally, due to the relatively high proportion of proteins that are insoluble, significant investment will be needed for the development of high-throughput technologies that will generate soluble protein or protein domains from an initial insoluble protein construct. Further development of computational methods for membrane protein identification may also prove useful in the initial screening step.

The "low hanging fruit" of the proteomes will feed hungry structural biologists for no more than 3–5 years. A parallel effort must be vested in the creation of an experimental ladder to reach the fruit in the upper regions of proteomes.

References

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.

Boyd, D., Schierle, C., Beckwith, J., 1998. How many membrane proteins are there? Protein Sci. 7, 201–205.

Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland, T., Lin, D., Sali, A., Studier, F.W., Swaminathan, S., 1999. Structural genomics: beyond the human genome project. Nat. Genet. 2, 151–157.

Consortium TceS. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282, 2012–2018.

Cort, J.R., Koonin, E.V., Bash, P.A., Kennedy, M.A., 1999. Phylogenetic approach to target selection for structural genomics: solution structure of YciH. Nuc. Acids Res. 27, 4018–4027.

Sonnhammer, E.L.L., von Heijne, G., Krogh, A., 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. In: Glasgow, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff, D., Sensen, C. (Eds.), Proceedings of Sixth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, pp. 175–182.

Gerstein, M., 1998. Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. Proteins 33, 518–534.

Heiter, P., Boguski, M., 1997. Functional genomics: It's all how you read it. Science 278, 601–602.

Kim, S.H., 1998. Shining a light on structural genomics. Nat. Struct. Biol. 5 (Suppl), 643–645.

Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K., Pedersen, J.T., 1997. Critical assessment of methods of protein structure prediction (CASP): round II. Proteins; Suppl 1:2–6

Shapiro, L., Lima, C.D., 1998. The Argonne structural genomics workshop: Lamaze class for the birth of a new science. Structure 6, 265–267.

Teichmann, S., Chothia, C., Gerstein, M., 1999. Advances in Structural Genomics. Curr. Opin. Structr. Biol. 9, 390.

Terwilliger, T.C., Waldo, G., Peat, T.S., Newman, J.M., Chu, K., Berendzen, J., 1998. Class-directed structure determination: foundation for a protein structure initiative. Protein Sci. 7, 1851–1856.

Waldo, G.S., Standish, B.M., Berendzen, J., Terwilliger, T.C., 1999. Rapid protein-folding assay using green fluorescent protein. Nat. Biotechnol. 7, 691–695.

Wootton, J.C., Federhen, S., 1996. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 266, 554–571.

Zarembinski, T.I., Hung, L.-W., Muller-Dieckmann, H.-J., Kim, K.-K., Yokota, H., Kim, R., Kim, S.-H., 1999. Structure-based assignment of the biochemical function of a hypothetical protein: A test case for structural genomics. Proc. Natl. Acad. Sci. USA 95, 15 189–15 193.