ELSEVIER

Il Farmaco 57 (2002) 947–951

IL FARMACO

www.elsevier.com/locate/farmac

Review

Structural proteomics: lessons learnt from the early case studies

Martin Norin [a], Michael Sundström [b,*]

[a] *Biovitrum, Department of Structural Chemistry, Nordenflychtsvägen 62:6, SE-112 76 Stockholm, Sweden*
[b] *Biovitrum AB, Lindhagensgatan 133, B20:9, SE-112 76 Stockholm, Sweden*

Received in revised form 8 January 2002; accepted 19 January 2002

Abstract

The genomics efforts have identified a large number of novel genes and thus provided a pool of interesting but not functionally characterized target proteins. It has been suggested that structural proteomics will significantly impact the success rate of functional characterization of such identified genes and proteins by providing structure–function hypotheses by fold and feature recognition and analysis. Structural proteomics initiatives, both in academic and industrial settings, are today generating protein structures at an unprecedented rate although relatively few large-scale efforts have been displayed in the public domain. However, a number of individual studies have provided a 'road-map' for selected approaches that hold the promise to significantly impact the process of deriving function from structure.

Keywords: Structural proteomics; Structure–function prediction; Protein production; NMR; Protein crystallography; Molecular modeling

1. Introduction

The structural proteomics research discipline has significantly impacted the strategies for which three-dimensional (3-D) structure determination of protein targets are carried out. The recent development of high throughput (HT) methodologies and technologies has enabled novel structural data to be generated with efficiency and speed by significant process improvements in protein production, X-ray crystallography and biomolecular nuclear magnetic resonance (NMR). Major improvements for protein production include HT 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 leading to that freezing of crystals at liquid-Nitrogen temperature (cryofreezing), multiple-wavelength anomalous dispersion (MAD) phasing, robotization, 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 as well as automated spectra assignment and structure determination. Improvements in the structure determination methodology using NMR and crystallography have made structure determination straightforward if the protein sample has appropriate 'structural properties', which includes good solubility, favorable aggregation state and homogeneity leading to that crystals of diffraction quality are easily formed or that the protein sample gives good NMR spectra.

Despite major improvements in experimental techniques relating to the ability to carry out true HT structure determination and prediction, a number of limiting factors are easily identifiable. Some of the most significant are, (i) the efficient and rational production of proteins with 'structural properties'; (ii) the manual intervention and time required for X-ray crystallographic data collection and evaluation; and (iii) the time required for data collection and spectral interpretation using NMR approaches.

* Corresponding author
E-mail addresses: martin.norin@biovitrum.com (M. Norin), michael.sundstrom@actar.se (M. Sundström).

0014-827X/02/$ - see front matter
PII: S0014-827X(02)01212-0

2. Populating structural space

The currently available structures cover only a fraction of the 3-D-structural space and thus restrict attempts to functionally annotate full proteomes. Depending on the organism, it has been estimated that 20–70% of the proteome can be modeled at the 30% threshold level (reviewed by Sánches [1]).

Recently, the New York Structural Genomics Research Consortium (http://www.nysgrc.org/) attempted to assess the impact of novel 3-D-structures by documenting the number and quality of the comparative models for related proteins as measured by sequence homology. For each novel structure, ~100 models (at least at the fold level) of related proteins devoid of a previous structural template could be generated.

In a more comprehensive study from the laboratory of Chris Sander, [2] an attempt was made to estimate the number of structures needed to model 90% of currently known protein families at >30% sequence identity. Using unique structures from the PDB compared with available complete genome sequences, the authors concluded that structural templates were available for segments of ~30% of sequences from genomes and that only 5–10% of all amino acid residues could be modeled. The study suggested that ~16000 additional structures were needed to model 90% of the 'global' proteome, given that the selection of structural targets was optimized. The remaining 10% would require substantial additional investments as it contains a high degree of singletons, i.e. sequences without relatives, or families with few members.

3. Protein expression and purification

From an historical perspective, protein crystallography and biomolecular NMR target proteins were often chosen for their 'structural properties' and for the ease with which they could be purified in milligram amounts. Later on, the protein biochemistry process was significantly improved for studies of high priority proteins. Here, a heavy protein engineering effort is usually made to optimize the construct prior to attempting crystallization or NMR data collection by improving the behavior of the protein sample with regards to solubility, aggregation, subtype enrichment and stability in solution.

In structural proteomics programs, it is quite common that the physical properties of expressed proteins are sub-optimal regarding their ability to form well-diffracting crystals or behavior in solution for NMR studies. The single most critical bottleneck reported is solubility (for example, see Christendat et al. [3]). Other crucial factors involve the availability of methods for rapid and accurate quality control, such as analysis of purity, homogeneity and structural integrity.

In structural proteomics research programs, one often selects target proteins that are easily expressed and purified and have favorable 'structural properties'. An often-used approach is to work with a limited number of expression constructs of proteins that lack current structural template. Since most of the selected and structure determined proteins are likely to contain previously non-described structural features, a high basic discovery value can be generated from a limited work-load (for a recent review see Edwards et al. [4]). The obvious risk with such approaches is that an enrichment of certain folds with 'favorable properties' could occur, and that a biased set of 3-D-structures will be generated.

Structural proteomics efforts, by necessity, use affinity and detection tags to allow rapid and streamlined protein purification. Smaller tags, such as histidine-clusters, can often be kept throughout the structure determination whereas larger tags need to be removed prior to NMR and X-ray studies. One drawback with this approach is that larger fusion tags are quite likely to give misleading data by solubilizing misfolded protein constructs or proteins from multi-component complexes, which lack their natural interaction partner(s) needed for functional integrity. Thus, it is our belief that optimization for the best fusion tag using solubility screens, needs to be accompanied by early quality analysis to assure that the constructs chosen for further studies are biologically relevant. The minimal criteria includes standard quality assurance methods such as N-terminal sequence, two-dimensional (2-D)-structural analysis using Circular Dichroism and mass spectroscopic analysis, but optimally a predicted activity should be tested using a standardized set of in vitro assays.

4. Biomolecular NMR approaches

Biomolecular NMR has emerged as a key methodology in structural proteomics projects (reviewed by Montelione et al. [5]). Despite the inherent limitation to not be able to determine the structure of larger proteins, it has been estimated that ~20% of the proteome of yeasts and other organisms, will fall within the size limit for NMR studies (~300 residues for a monomeric protein). Thus, a potentially more serious limitation for NMR as a key structural proteomics methodology arises from the sample property requirements needed, which include very high solubility in aqueous solutions, monomeric state and stability in solution over extended data collection time (Fig. 1).

5. High throughput protein crystallography

HT crystallography has been facilitated by improved phasing and model building methods, decreased sample requirements through miniaturization as well as robotization and automation from the crystallization stage to structure determination. Recent improvements in the integration of methods as well as user interfaces, have converted protein crystallography structure determination into an easily accessed methodology for a larger number of research groups. In addition to the numerous academic initiatives, especially with regards to the consortia efforts, companies such as Integrative Proteomics (http://www.integrativeproteomics.com/), Structural GenomiX (http://www.stromix.com/) and SYRRX (http://www.syrrx.com/) are now developing and attempting to commercialize HT capabilities for protein expression, crystallization, image analysis for automated crystal detection and structure determination.

HT approaches for structure determination using crystallography often use seleno-methionine (Se-Met) labeled protein samples and data collection using single or multi-wavelength anomalous diffraction (SAD, MAD) to retrieve phase information [6]. Traditional crystallographic methods, however, remain important as many protein structures are solved by multiple isomorphous replacement (MIR) with heavy metal derivatives. In addition, an increasing fraction of molecular replacement derived structures are expected, as the number of suitable templates continuously increase. A complementary approach uses halides bound to protein crystals. Dauter et al. [7] reported the crystal structure of a carboxyl proteinase devoid of structural template that was solved using phase information derived from bromine peak absorption. The phase information was compared with that obtained from MAD and displayed similar structural quality, which shows that halide phasing has the potential to become a complementary approach for HT structure determination.

6. Functional prediction from experimental data

With increased population of 3-D-structural space, it has been speculated that evolutionary relationships and functional assignment of single proteins and families will be greatly facilitated. In addition to fold and feature recognition and analysis, a limited but significant number of cases have shown that direct crystallography-derived electron density for 'native' ligands or co-factors bound to the protein can be observed. When such data are available, the generation of a functional hypothesis is greatly facilitated.

One example of obtaining evolutionary relationships from structures is the study by Shapiro and Scherer [8] in which a clear structural similarity between a complement 1q (C1q) family protein family member and tumor necrosis factor (TNF) could be observed. The structural analysis/comparison could clearly identify highly similar folding topologies, conserved key residues and similarity

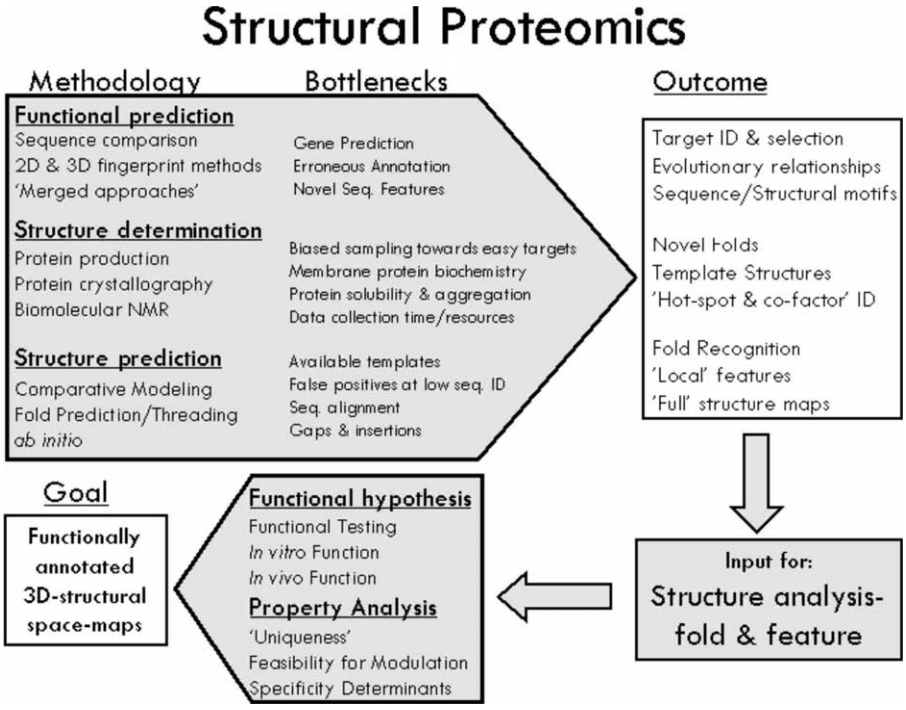

Fig. 1. The illustration gives an overview of key factors impacting research programs in structural proteomics and the outcome expected from successfully performed efforts.

of trimer interface, thus establishing an evolutionary link between the TNF and C1q protein families.

The study by Zarembinski and coworkers [9] is probably the best example of a direct structure-to-function prediction in which the crystal structure of an unannotated protein (MJ0577) from *Methanococcus jannaschii* showed a bound ATP in the high resolution electron density maps, suggesting that the protein was an ATP-ase or an ATP-mediated molecular switch. The hypothesis could be confirmed experimentally by biochemical test methods. Further, the structural analysis of the co-factor binding motif could be used to suggest other ATP binding sequences among many homologous, but unannotated, proteins in this family.

Other studies have addressed the structure-based assignment of function with various degrees of success [10–14]. One representative example was shown in the study of the YjgF protein from *Escherichia coli* [15]. Here, the high-resolution crystallographic structure revealed its closest structural homologues to be Chorismate Mutase and members of the Protein Tyrosine Phosphatase family. Standard biochemical assays subsequently used to test the hypothesis, failed to show any enzymatic activity suggested by structural similarity. Hence, the assignment of function was unsuccessful but a number of biochemical tests for further studies could be envisioned.

In contrast to 'structural proteomics case studies of single proteins', structure determination efforts of a large number of proteins from the archaeon *Methanobacterium thermoautotrophicum* whose genome comprise 1871 open reading frames, represent the best-published case study [3]. Here, $\sim$400 out of $\sim$900 (representing $\sim$25% of the organism's proteome) novel and predicted soluble target proteins were chosen for HT structure determination and structure-to-function prediction efforts. The protein components were expressed and purified in a streamlined approach and subsequent structure determination efforts included the use of both NMR ($<$20 kDa) and crystallographic methods at various laboratories. Approximately 20% of the target proteins were found to be suitable candidates for structure determination. Another interesting observation was that poor expression and solubility of the proteins accounted for close to 60% of the failures. The first ten structures published revealed five of them to contain a bound ligand or a putative ligand-binding site inferred by structural homology. Thus, a high fraction of the generated structures suggested a function to be experimentally verified.

7. Functional prediction from theoretical approaches

Until today most proteins that have been experimentally determined are quite well characterized in terms of biological function. However, we expect that the structural genomics projects will deliver an increased number of structures of proteins with unknown functions. In addition, the biological functions of many multifunctional proteins are only partly understood. Theoretical approaches to computationally mine for the structural and functional relationships should in the near future significantly contribute to highlight potential novel functions of proteins.

The functional machinery of a protein is usually built from a limited number of specific amino acid residues in precise 3-D relative positions ('hot-spots'). The PROSITE [16] (http://www.expasy.ch/prosite) database contains sequence patterns of functional sites, but has no 3-D information. To improve the usefulness of such data in structural terms, Thornton and colleagues analyzed the 3-D features of sequence patterns in the PROSITE database to define templates for structure-based mining for functional sites [17]. In another example, Fetrow and Skolnick used 3-D descriptors ('Fuzzy Functional Forms') based on intramolecular distances between amino acid residues to assign a potential oxidoreductase active site for the serine/threonine phospatase-1 subfamily [18,19].

In a recent and especially interesting example, Sowa et al. [20] presented a method to propose novel sites that is independent of any prior knowledge of specific amino acid features. By applying an 'Evolutionary Trace' (ET) method by combining protein family clusters in structural and sequence data they proposed an important site for the regulation of G protein signaling that was confirmed though site-specific mutagenesis.

8. Conclusions

Methodology and approaches developed for structural proteomics efforts have fundamentally changed the structural research discipline and are already contributing to the population of 3-D-structural space. However, a likely key limitation of the current approaches is the low efficiency of structure determinations of more 'difficult' targets. However, commercial and academic structural proteomics efforts should within the near future generate novel structures that will by far exceed our collected results to date.

Over the last few years, structure determination methods have so greatly improved, that the key factor for sustained progress has shifted towards the development and use of efficient production systems of biologically relevant proteins. A necessary development of the protein production systems includes the ability to handle traditionally difficult targets such as protein complexes, membrane proteins and enzymes requiring post-translational modifications for functional integrity.

The currently published case studies have shown that structure determination and analysis will not necessarily mean that function can be directly concluded. However, structural data suggesting an unproven function, allow a rationale for directed efforts to validate a postulated function using experimental assays with greater efficiency and less resource usage than what has been previously possible.

The population of 3-D-structural space is a crucial determinant to successfully generate full structural proteome maps. As the number structure determined proteins increases, a vast experimental pool of data to will be available to test and improve modeling approaches and hopefully lead to a major improvement in the accuracy and speed of such efforts.

Structural proteomics efforts have impacted both applied and basic research. Ongoing directed efforts within larger soluble protein families already characterized for function as well as 'unbiased' efforts on prokaryotic proteomes are already providing significant insights into their mechanism of action as well as their suitability as potential drug targets.

References

[1] R. Sánches, et al., Protein structure modeling for structural genomics, Nat. Struct. Biol. 7 (2000) 986–990.
[2] D. Vitkup, et al., Completeness in structural genomics, Nat. Struct. Biol. 8 (2001) 559–566.
[3] D. Christendat, et al., Structural proteomics of an archaeon, Nat. Struct. Biol. 7 (2000) 903–909.
[4] A.M. Edwards, et al., Protein production: feeding the crystallographers and NMR spectroscopists, Nat. Struct. Biol. 7 (2000) 970–972.
[5] G.T. Montelione, et al., Protein NMR spectroscopy in structural genomics, Nat. Struct. Biol. 7 (2000) 982–985.
[6] L.M. Rice, et al., Single-wavelength anomalous diffraction phasing revisited, Acta Crystallogr. D 56 (2000) 1413–1420.
[7] Z. Dauter, et al., Practical experience with the use of halides for phasing macromolecular structures: a powerful tool for structural genomics, Acta Crystallogr. D 57 (2001) 239–249.
[8] L. Shapiro, P.E. Scherer, The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor, Curr. Biol. 8 (1998) 335–338.
[9] T.I. Zarembinski, et al., Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics, Proc. Natl. Acad. Sci. USA 95 (1998) 15189–15193.
[10] J.R. Cort, et al., Structure-based functional classification of hypothetical protein MTH538 from *Methanobacterium thermoautotrophicum*, J. Mol. Biol. 302 (2000) 189–203.
[11] C. Colovos, et al., The 1.8 Å crystal structure of the YcaC gene product from *Escherichia coli* reveals an octameric hydrolase of unknown specificity, Structure 6 (1998) 1329–1337.
[12] C.D. Lima, et al., Structure-based analysis of catalysis and substrate definition in the HIT protein family, Science 278 (1997) 286–290.
[13] B. Stec, et al., MJ0109 is an enzyme that is both an inositol monophosphatase and the 'missing' archaeal fructose-1,6-bisphosphatase, Nat. Struct. Biol. 7 (2000) 1046–1050.
[14] F. Yang, et al., Crystal structure of *Escherichia coli* HdeA, Nat. Struct. Biol. 5 (1998) 763–764.
[15] K. Volz, A test case for structure-based functional assignment: the 1.2 Å crystal structure of the yjgF gene product from *Escherichia coli*, Prot. Sci. 8 (1999) 2428–2437.
[16] K. Hofmann, et al., The PROSITE database, its status in 1999, Nucleic Acids Res. 27 (1999) 215–219.
[17] A. Kasuya, J. Thornton, Three-dimensional structure analysis of PROSITE patterns, J. Mol. Biol. 286 (1999) 1673–1691.
[18] J. Fetrow, et al., Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity, J. Mol. Biol. 281 (1998) 949–968.
[19] J. Fetrow, et al., Structure-based functional motif identifies a potential disulphide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily, FASEB J. 13 (1999) 1866–1874.
[20] M. Sowa, et al., Prediction and confirmation of a site critical for effector regulation of RGS domain activity, Nat. Struct. Biol. 8 (2001) 234–237.