CURRENT OPINION
www.current-opinion.com

Overview of structural genomics: from structure to function

Chao Zhang[†] and Sung-Hou Kim[*]

The unprecedented increase in the number of new protein sequences arising from genomics and proteomics highlights directly the need for methods to rapidly and reliably determine the molecular and cellular functions of these proteins. One such approach, structural genomics, aims to delineate the total repertoire of protein folds, thereby providing three-dimensional portraits for all proteins in a living organism and to infer molecular functions of the proteins. The goal of obtaining protein structures on a genomic scale has motivated the development of high-throughput technologies for macromolecular structure determination, which have begun to produce structures at a greater rate than previously possible. These new structures have revealed many unexpected functional and evolution relationships that were hidden at the sequence level.

Addresses
Department of Chemistry and Calvin Laboratory, Lawrence Berkeley National Laboratory, University of California at Berkeley, Berkeley, CA 94720, USA
[*]e-mail: shkim@cchem.berkeley.edu
[†]Current address: Plexxikon, Inc., 91 Bolivar Drive, Berkeley, CA 94710, USA

Current Opinion in Chemical Biology 2003, **7**:28–32

This review comes from a themed issue on
Proteomics and genomics
Edited by Matthew Bogyo and James Hurley

1367-5931/03/$ – see front matter

DOI 10.1016/S1367-5931(02)00015-7

Abbreviations

AdoMet	*S*-adenosyl-methionine
FAD	flavin adenine nucleotide
HAP	6-*N*-hydroxyaminopurine
ITP	inosine triphosphate
NAD^+	nicotinamide adenine dinucleotide
PDB	Protein Data Bank

Introduction

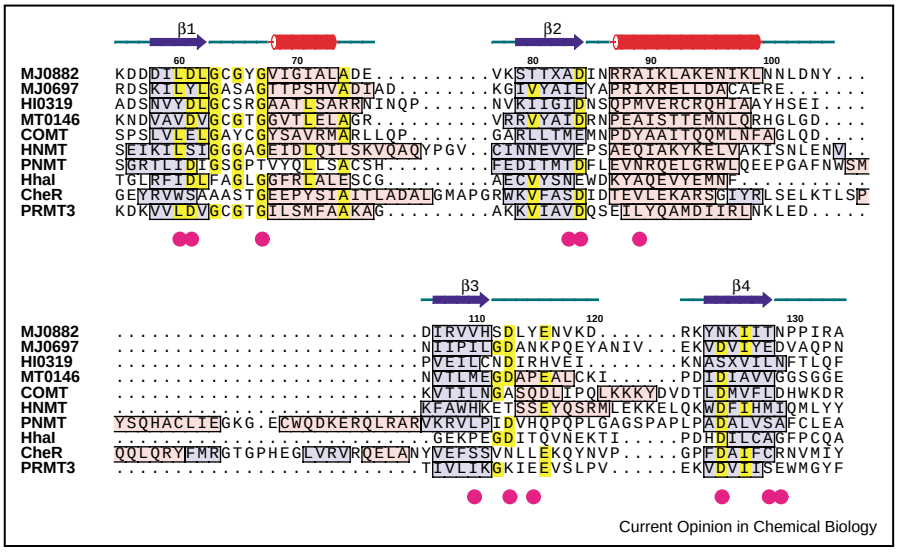
The genomes of more than 100 prokaryotic and eukaryotic organisms have been sequenced (http://www.ncbi.nih.gov/Genomes/ and http://www.tigr.org/). In all genomes sequenced to date, a large portion of these organisms' predicted protein-coding regions encode polypeptides of unknown biological functions (also called hypothetical proteins). A major challenge is to find ways to reliably and rapidly determine the molecular (biochemical and biophysical) and cellular functions of these proteins. One approach for assigning the molecular function of a protein is first to determine its three-dimensional structure by either X-ray crystallography or NMR, and then to compare the solved structure against known structures in the protein structure databases. If there are one or more significant structural homologues, the hypothetical protein is predicted to have molecular functions similar to those of the homologues, despite the absence of sequence similarities. The predictions can then be tested experimentally. The molecular function provides a basis for searching for the cellular function of the protein in combination with other genomics and proteomics techniques (e.g. expression profiling, protein interaction mapping, gene knockout). This method, structural genomics [1], is far more sensitive than primary sequence comparisons because proteins performing similar or related functions, albeit having insignificant sequence similarity, may have similar structure or fold [2–4].

Since the publication of the first structural genomics test case in 1998 [1], in which the crystal structure of a hypothetical protein revealed its molecular function, many similar studies have been carried out [5–7]. The compelling results from these pilot studies helped to initiate a major international effort to obtain protein structures on a genomic scale [7,8], with multi-institutional collaborations formed all over the world [9–11]. Although infrastructure building and technology development are still the main focus of structural genomics programs [12–18], a considerable number of protein structures have already been produced, some of them coming directly out of semi-automated structure-determination pipelines [17,19,20•,21,22•]. A search of the October 2002 release of the Protein Data Bank (PDB) [23] returned 117 PDB entries containing the key words 'structural genomics'. Given the delay between structure deposition and release, such a list is likely to represent only a fraction of the structures solved by the structural genomics programs over the past three years. Here we select a few examples from the list to illustrate the type of structural insights that would be expected from a structural genomics project and, in particular, how information flows from the atomic coordinates to a functional characterization of a protein.

From structure to function

The assignment of biochemical activity to a protein of unknown function is most straightforward when the new structure resembles that of proteins whose functions are known. Structural similarities yield powerful clues to biochemical function that are not evident from sequence alone. For example, although there was no detectable sequence similarity between MJ0882, a hypothetical protein from *Methanococccus jannaschii* (MJ), and any of the

Figure 1



The structure-based sequence alignment of the AdoMet-binding region of methyltransferases. Included in the alignment are four methyltransferases identified by the structural genomics approach, MJ0882, MJ0697, HI0319/YecO and MT0146/CbiT, and six previously known methyltransferases, catechol O-methyltransferase (COMT), histamine *N*-methyltransferase (HNMT), phenylethanolamine *N*-methyltransferase (PNMT), cytosine-5-methyltransferase Hhal, chemotaxis receptor methyltransferase CheR, and protein arginine methyltransferase 3 (PRMT3). Secondary structure elements of MJ0882 are marked above the alignment. β-Strands and α-helices of individual proteins are highlighted in blue and red shades, respectively. Conserved residues are shown in yellow shade. The residues important for AdoMet binding are indicated by purple circles.

known methyltransferases, the crystal structure of MJ0882 (PDB code: 1dus; deposited in January 2000 [24]) revealed an *S*-adenosyl-methionine (AdoMet)-dependent methyltransferase fold. The methyltransferase activity of MJ0882 inferred from the structure was subsequently confirmed by biochemical experiments. Structural genomics has also led to the discovery of two other unsuspected methyltransferases. The first is a previously unannotated protein from *Haemophilus influenzae*, HI0319/YecO, whose structure had a methyltransferase fold and a bound *S*-adenosyl-homocysteine (the methylation by-product) [25]. The second protein, MT0146/CbiT from *Methanobacterium thermoautotrophicum*, was originally annotated as a precorrin decarboxylase. The structure of MT0146, however, showed the canonical AdoMet-depdendent methyltransferase fold, suggesting a reclassification of the enzymatic function of the protein [26•]. In addition, there is at least one earlier report of an AdoMet-dependent methyltransferase suggested by structure [27]. MJ0697 is likely to be a rRNA methyltransferase on the basis of its homology to the yeast protein fibrillarin, which is essential for pre-rRNA maturation [28]. The fact that methyltransferases have been identified repeatedly in structural genomics projects suggests that many methyltransferases may have been overlooked by the current genome annotations. A structure-based sequence alignment of MJ0697, MJ0882, HI0319 and MT0146 with other methyltransferases of known structures revealed moderate sequence conservation in the core AdoMet-binding region of the enzymes (Figure 1). A hidden-Markov model [29] built using such an alignment has demonstrated much higher sensitivity in detecting unannotated methyltransferases (Zhang C, Kim S-H, unpublished data).

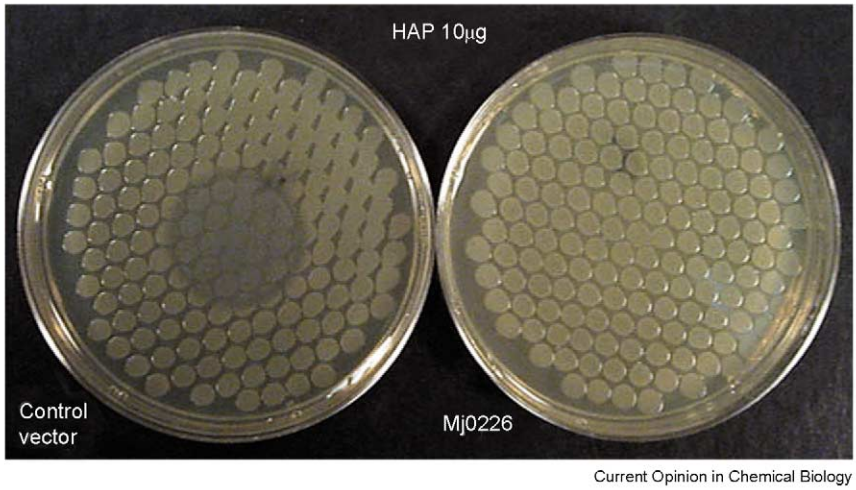
The unexpected presence of a ligand in the structure of a hypothetical protein can also help to infer its biochemical function and can be readily tested experimentally. This was the case with protein MJ0577 where ATP was fortuitously co-crystallized with the protein, which immediately suggested a possible role of the protein in ATP hydrolysis [1]. The ATP-binding pocket of MJ0577 contains some of the motifs commonly found in nucleotide-binding proteins, but has a different sequential arrangement of the motifs compared with others, and thus has evaded the detection of existing motif-based search methods. In another example, the MT0150 protein was found to be co-purified and co-crystallized with NAD^+, and the structure later solved has a nucleotide-binding fold similar to several nucleotidyltransferases [19]. Additional biochemical studies have confirmed that MT0150 has nicotinamide mononucleotide adenylyltransferase activity, and the solved structure corresponds to the product-bound form of the enzyme. It is worth noting that the bound ligand does not have to be the natural ligand or cofactor to be useful for understanding the function. When other biological information is available, a fortuitously trapped buffer molecule can sometimes shed

light on the possible biochemical function of the protein. For example, the structure of the *Thermotoga maritime* protein TM0423 included a Tris buffer molecule bound in the enzyme's active site, which seemed to mimic a glycerol substrate and suggested that TM0423 is a glycerol dehydrogenase [20•].

Structure-based function assignments can suggest possible biochemical function and, sometimes, the cofactors or substrates, of a hypothetical protein. In cases when the protein has known cellular function, the knowledge of the biochemical function helps to reveal the molecular mechanism for a cellular process. The study of the thymidylate synthase-complementing proteins offers an interesting example where independent projects converge to give functional and structural characterization of an important protein family. Found almost exclusively in organisms that lack thymidylate synthase, thymidylate synthase-complementing proteins synthesize the essential DNA precursor thymidylate by an alternative pathway. The exact mechanism of the pathway had been unknown until very recently; a combination of biochemical and structural studies revealed a novel flavin-dependent mechanism for thymidylate synthesis [20•,30•]. The structure of a member of the thymidylate synthase-complementing proteins from *T. maritime*, TM0449, has been solved as part of a structural genomics programme [20•]. The structure revealed a large pocket in the centre of a TM0449 tetramer and a bound flavin adenine nucleotide (FAD) molecule in each of the four equivalent putative active sites in the pocket. Independent biochemical assays of another member (ThyX) of the same protein family indicated that the activity of the enzyme was dependent on reduced flavin nucleotides [30•]. These results, together with the fact that FAD-binding residues are highly conserved in the ThyX protein family, suggest that FAD is the genuine cofactor for the alternative thymidylate synthesis pathway. The delineation of this mechanism has important implications for the evolution of DNA synthesis machineries and the design of new antimicrobial strategies.

Even when there are no bound ligands or close structural homologues of a hypothetical protein, the three-dimensional structure of the protein can, sometimes, suggest one or more testable molecular and cellular functions. The case of MJ0226 particularly illustrates how knowledge of a structure can lead to simple experiments that provide immediate insights into both biochemical and cellular functions [5]. Although the structure of MJ0226 has a new fold, it has limited structural similarity with a group of nucleotide-binding proteins (using the program DALI [31] with a Z score below 4). Nucleotide-binding assays on MJ0226 showed that the protein interacts with both ATP and GTP and that MJ0226 has weak nucleotide triphosphatase activity. Subsequent analysis found that xanthine triphosphate (XTP) and inosine triphosphate (ITP) are better substrates for MJ0226. On the basis of the structural and biochemical information, and an observation that MJ0226 is homologous (30% sequence identity) to yeast HAM1 protein, which is required for the survival of yeast in the presence of modified bases [32], it has been proposed that the cellular function of MJ0226 may be to prevent mutations by protecting DNA from incorporation of modified purine bases such as dXTP or dITP. This prediction has recently been confirmed by a complementation experiment (Y Pavlov, personal communication; Figure 2). When MJ0226 is overexpressed in *Escherichia*

Figure 2



MJ0226 overexpression protects *E. coli* host from toxic effect of HAP. When *E. coli* was transformed with plasmids without (left) and with MJ0226 (right), *E. coli* without MJ0226 dies where HAP is spotted (central circular regions in both photos) but *E. coli* with MJ0226 survives.

coli, the host strain is protected both from toxic and mutagenic effects of base analogue 6-*N*-hydroxyaminopurine (HAP).

Another principal goal of structural genomics is to populate the protein structure or fold space [33–35], thereby providing representative structures for all existing protein families. This implies that a large fraction of the new structures produced by structural genomics projects would not have close structural homologues in the current protein database, and thus their functional assignments remain a challenge. An analysis of the 117 PDB entries mentioned above indicates that the structures contributed by structural genomics show a much higher probability of revealing new folds or new variations of existing folds than other PDB structures deposited during the same period of time. This vindicated the target selection strategies that have been employed by various structural genomics consortia to maximize the information return of structure determination [36–40]. When a new fold is revealed, the universe of known protein folds is enriched, and once the function is determined from its structure and other means, novel structure–function relationships are established. With improved understanding of the structure–function relationships of proteins, structural bioinformatics tools can play an important role in expediting this process. Meanwhile, homology modeling using the solved structure as a template enables a structural description and function prediction of a large number of protein sequences that fall within the 'modeling distance' [41,42•,43,44].

Conclusions

In summary, structural genomics is likely to reveal to us a global view of the protein structure universe and to complement other genomic and proteomic technologies in providing molecular functions of many proteins of unknown function. This opens the perspective that, in a foreseeable future, many cellular processes can be correlated to the physics and chemistry of the individual proteins or fold domains involved in the processes.

Acknowledgements

We thank all authors who have deposited their structures under the structural genomics category. We acknowledge the support of the NIH grant GM62412 for most of the structures cited in this article. We thank Dr Y Pavlov of the National Institute of Environmental Health Sciences and Dr Ye Sun Han of Korea Institute of Science and Technology for the complementation experiment shown in Figure 2.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, Kim SH: **Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics**. *Proc Natl Acad Sci USA* 1998, **95**:15189-15193.

2. Kim SH: **Shining a light on structural genomics**. *Nature Struct Biol* 1998, **5**:643-645.

3. Murzin AG: **How far divergent evolution goes in proteins**. *Curr Opin Struct Biol* 1998, **8**:380-387.

4. Aloy P, Oliva B, Querol E, Aviles FX, Russell RB: **Structural similarity to link sequence space: new potential superfamilies and implications for structural genomics**. *Protein Sci* 2002, **11**:1101-1116.

5. Hwang KY, Chung JH, Kim SH, Han YS, Cho Y: **Structure-based identification of a novel NTPase from *Methanococcus jannaschii***. *Nat Struct Biol* 1999, **6**:691-696.

6. Boggon TJ, Shan WS, Santagata S, Myers SC, Shapiro L: **Implication of tubby proteins as transcription factors by structure-based functional analysis**. *Science* 1999, **286**:2119-2125.

7. Campbell ID: **Timeline: the march of structural biology**. *Nat Rev Mol Cell Biol* 2002, **3**:377-381.

8. Stevens RC, Yokoyama S, Wilson IA: **Global efforts in structural genomics**. *Science* 2001, **294**:89-92.

9. Terwilliger TC: **Structural genomics in North America**. *Nat Struct Biol* 2000, **7**:935-939.

10. Yokoyama S, Matsuo Y, Hirota H, Kigawa T, Shirouzu M, Kuroda Y, Kurumizaka H, Kawaguchi S, Ito Y, Shibata T *et al.*: **Structural genomics projects in Japan**. *Prog Biophys Mol Biol* 2000, **73**:363-376.

11. Heinemann U: **Structural genomics in Europe: slow start, strong finish?** *Nat Struct Biol* 2000, **7**:940-942.

12. Dieckman L, Gu M, Stols L, Donnelly MI, Collart FR: **High throughput methods for gene cloning and expression**. *Protein Expr Purif* 2002, **25**:1-7.

13. Pedelacq JD, Piltch E, Liong EC, Berendzen J, Kim CY, Rho BS, Park MS, Terwilliger TC, Waldo GS: **Engineering soluble proteins for structural genomics**. *Nat Biotechnol* 2002, **20**:927-932.

14. Chayen NE, Saridakis E: **Protein crystallization for genomics: towards high-throughput optimization techniques**. *Acta Crystallogr D* 2002, **58**:921-927.

15. Villaseñor A, Sha M, Thana P, Browner M: **Fast drops: a high-throughput approach for setting up protein crystal screens**. *BioTechniques* 2002, **32**:184-189.

16. Karain WI, Bourenkov GP, Blume H, Bartunik HD: **Automated mounting, centering and screening of crystals for high-throughput protein crystallography**. *Acta Crystallogr D* 2002, **58**:1519-1522.

17. Rupp B, Segelke BW, Krupka HI, Lekin T, Schafer J, Zemla A, Toppani D, Snell G, Earnest T: **The TB structural genomics consortium crystallization facility: towards automation from protein to electron density**. *Acta Crystallogr D* 2002, **58**:1514-1518.

18. Bhavesh NS, Panchal SC, Hosur RV: **An efficient high-throughput resonance assignment procedure for structural genomics and protein folding research by NMR**. *Biochemistry* 2001, **40**:14727-14735.

19. Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I *et al.*: **Structural proteomics of an archaeon**. *Nat Struct Biol* 2000, **7**:903-909.

20. • Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreusch A, Spraggon G, Klock HE, McMullan D, Shin T *et al.*: **Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline**. *Proc Natl Acad Sci USA* 2002, **99**:11664-11669.
This paper describes the high-throughput structural determination pipeline designed and implemented at Syrrx Corp. and the Joint Center for Structural Genomics. The results of the application of the pipeline to the proteome of thermophillic bacterium *Thermotaga maritima* are presented. Of the two structures used as examples of the final outputs of the pipeline, the structure of thymidylate synthase-complementing protein TM0423 is of particular interest. The presence of four FAD molecules in the central active site of the TM0423 tetramer suggests a flavin-dependent mechanism in the alternative thymidylate synthesis pathway.

21. Burley SK, Bonanno JB: **Structural genomics of proteins from conserved biochemical pathways and processes**. *Curr Opin Struct Biol* 2002, **12**:383-391.

22. • Yee A, Chang X, Pineda-Lucena A, Wu B, Semesi A, Le B, Ramelot T, Lee GM, Bhattacharyya S, Gutierrez P *et al.*: **An NMR approach to structural proteomics**. *Proc Natl Acad Sci USA* 2002, **99**:1825-1830.

This paper presents an NMR approach to structural genomics. The authors found 20% of the 500 small proteins that they have tested are amenable to NMR structure determination. They found that proteins from the thermophilic bacterium *T. maritima* had more well-behaved proteins on the basis of the expression, solubility and heteronuclear single quantum coherence (HSQC) results. By contrast, thermophilic properties of *M. thermoautotrophicum* proteins do not provide a significant advantage for structural genomics of small proteins using NMR. For a portion of 12 structures reported in the paper, putative functions have been inferred.

23. Bernstein FS, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based archival file for macromolecular structure**. *J Mol Biol* 1977, **112**:535-542.

24. Huang L, Hung LW, Kim R, Kim, SH: **Crystal structure and functional analysis of a hypothetical protein, MJ0882, from *Methanococcus jannaschii***. *J Struct Funct Genomics* 2003, in press.

25. Lim K, Zhang H, Tempczyk A, Bonander N, Toedt J, Howard A, Eisenstein E, Herzberg O: **Crystal structure of YecO from Haemophilus influenzae (HI0319) reveals a methyltransferase fold and a bound *S*-adenosylhomocysteine**. *Proteins* 2001, **45**:397-407.

26. • Keller JP, Smith PM, Benach J, Christendat D, deTitta GT, Hunt JF: **The crystal structure of MT0146/CbiT suggests that the putative precorrin-8w decarboxylase Is a methyltransferase**. *Structure* 2002, **10**:1475-1487.

This paper describes an interesting example of structure-based reclassification of the enzymatic function of a protein. *M. thermoautotrophicum* MT0146/CbiT protein was originally annotated as precorrin-8w decarboxylase. The structure of MT0146 revealed an AdoMet-dependent methyltransferase fold and a bound AdoHcy. The authors suggest that all functional annotations in sequence and structure databases should cite the papers containing the primary experimental data establishing the functional assignment.

27. Wang H, Boisvert D, Kim KK, Kim R, Kim SH: **Crystal structure of a fibrillarin homologue from *Methanococcus jannaschii*, a hyperthermophile, at 1.6 Å resolution**. *EMBO J* 2000, **19**:317-323.

28. Tollervey D, Lehtonen H, Jansen R, Kern H, Hurt EC: **Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly**. *Cell* 1993, **72**:443-457.

29. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**:755-763.

30. • Myllykallio H, Lipowski G, Leduc D, Filee J, Forterre P, Liebl U: **An alternative flavin-dependent mechanism for thymidylate synthesis**. *Science* 2002, **297**:105-107.

This paper describes the biochemical assay that establishes the flavin-dependent mechanism for the alternative thymidylate synthesis pathway.

31. Holm L, Sander C: **DALI: a network tool for protein structure comparison**. *Trends Biochem Sci* 1995, **20**:478-480.

32. Noskov VN *et al.*: **HAM1, the gene controlling 6-*N*-hydroxylaminopurine sensitivity and mutagenesis in the yeast *Saccharomyces cerevisiae***. *Yeast* 1996, **12**:17-29.

33. Holm L, Sander C: **Mapping the protein universe**. *Science* 1996, **273**:595-603.

34. Burley SK: **Bonanno JB: Structuring the universe of proteins**. *Annu Rev Genomics Hum Genet* 2002, **3**:243-262.

35. McGuffin LJ, Jones DT: **Targeting novel folds for structural genomics**. *Proteins* 2002, **48**:44-52.

36. Brenner SE: **Target selection for structural genomics**. *Nat Struct Biol* 2000, **7**:967-969.

37. Frishman D: **Knowledge-based selection of targets for structural genomics**. *Protein Eng* 2002, **15**:169-183.

38. Liu J, Rost B: **Target space for structural genomics revisited**. *Bioinformatics* 2002, **18**:922-933.

39. Vitkup D, Melamud E, Moult J, Sander C: **Completeness in structural genomics**. *Nat Struct Biol* 2001, **8**:559-566.

40. Portugaly E, Kifer I, Linial M: **Selecting targets for structural determination by navigating in a graph of protein families**. *Bioinformatics* 2002, **18**:899-907.

41. Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N, Sali A: **Protein structure modeling for structural genomics**. *Nat Struct Biol* 2000, **7**:986-990.

42. • Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A, Mewes HW: **Functional and structural genomics using PEDANT**. *Bioinformatics* 2001, **17**:44-57.

This paper describes the software design of PEDANT, one of the most comprehensive structure annotation servers available (http://pedant.gsf.de/).

43. Baker D, Sali A: **Protein structure prediction and structural genomics**. *Science* 2001, **294**:93-96.

44. Norin M, Sundstrom M: **Structural proteomics: developments in structure-to-function predictions**. *Trends Biotechnol* 2002, **20**:79-84.