Published in final edited form as:

IUBMB Life. 2003 ; 55(4-5): 249–255. doi:10.1080/1521654031000123385.

Target Selection and Determination of Function in Structural Genomics

James D. Watson¹, Annabel E. Todd², James Bray³, Roman A. Laskowski¹, Aled Edwards^{3,5}, Andrzej Joachimiak⁴, Christine A. Orengo², and Janet M. Thornton¹

¹EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

²Biochemistry and Molecular Biology Department, University College London, Gower Street, London WC1E6BT, UK

³Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada

⁴Biosciences Division and Structural Biology Center, Argonne National Laboratory, Argonne, Illinois, USA

⁵Clinical Genomics Centre/Proteomics, University Health Network, Toronto, Ontario, Canada

Summary

The first crucial step in any structural genomics project is the selection and prioritization of target proteins for structure determination. There may be a number of selection criteria to be satisfied, including that the proteins have novel folds, that they be representatives of large families for which no structure is known, and so on. The better the selection at this stage, the greater is the value of the structures obtained at the end of the experimental process. This value can be further enhanced once the protein structures have been solved if the functions of the given proteins can also be determined. Here we describe the methods used at either end of the experimental process: firstly, sensitive sequence comparison techniques for selecting a high-quality list of target proteins, and secondly the various computational methods that can be applied to the eventual 3D structures to determine the most likely biochemical function of the proteins in question.

Keywords

Structural genomics; target selection; function from structure; functional annotation

INTRODUCTION

One of the goals of the structural genomics projects (1, 2) is to obtain experimentally determined models of representative proteins from all protein families. This will vastly increase the number of proteins for which structural models are available, either directly from experiment, or via homology to an experimentally determined structure (3, 4, 5). It is also hoped that, along the way, novel protein folds will be discovered and will thus expand the known ‘fold space’; and indeed for some groups, ours included, this is one of the principal goals.

Many advances have been made on the experimental front: cloning and expression systems, purification and crystallization procedures, robotics and computer systems for data collection, and improvements to the software for structure determination and refinement. These advances have shortened the time it takes to go from target protein to crystal structure, and this is likely to improve even further in the future.

Because the emphasis is on high-throughput structure determination, and the attrition rate at each stage of the experimental process is high (proteins turn out not to be soluble, or do not crystallize, or the crystals are too small, and so on) the proportion of proteins that make it to the end as sets of 3D coordinates is very low. If the resultant structure is one that is already known or fails to meet any of the selection criteria (e.g. it has one of the common folds where novel folds are the primary goal), then the structure has considerably less inherent value. Therefore, it is crucial that the targets be carefully selected at the start of the process to minimize wasted time, expenditure and effort. Here we describe some particularly sensitive target selection techniques.

In cases where the targeting is directed at putative novel folds and uncharacterized proteins a side effect is that a significant proportion of the structures solved will be ‘hypothetical proteins’; that is, proteins of unknown function. For these, further effort needs to be made to determine their function in order to provide added value to the structures when depositing their coordinates with the Protein Data Bank, PDB (6). A protein structure, with no functional assignment, is of lesser interest and limited use. However, determining function of these hypothetical proteins, even knowing their 3D structure, is not always an easy task.

TARGET SELECTION

Assignment of Likely Structure to Gene Sequences

Critical to any target selection protocol is comparative sequence analysis in order to limit redundant structure determination efforts. Protein sequences that do not appear to have a relative of known structure in the PDB are prioritized so that, ideally, all new structures have a novel fold or define a new homologous superfamily of a previously observed fold. However, since tertiary structure is better conserved than sequence (7), structure determination often reveals very distant evolutionary relationships that may yield valuable functional insights. Moreover, since at least 30% sequence identity between structural template and protein is necessary for a reasonably accurate homology model (8), for diverse superfamilies multiple structures must be solved if the structural genomics community is to meet its target of mapping protein structure space to completion.

In our approach for the Midwest Center for Structural Genomics (MCSG) (summarised in Figure 1) the SSEARCH program (9) is used to identify any gene sequences (or regions of) that share 30% sequence identity with a known structure over a significant length of sequence. These regions are immediately excluded and not considered for structure determination.

For the prediction of distant evolutionary relationships between all remaining sequences and proteins in the PDB, BLAST, and several sensitive, profile-based techniques are applied: PSI-BLAST (10), the powerful Hidden Markov model (HMM) method SAM-T99 (11) and the novel SAMOSA protocol. SAMOSA (I. Sillitoe, manuscript in preparation) generates ‘3D-HMMs’ by combining several distant 1D-HMM sequence alignments according to a multiple structural alignment and these improve the coverage of homology detection even further. Those sequences without a PDB match are categorized as high priority targets for structure determination.

Fold recognition techniques, which attempt to find folds compatible with a particular sequence, can detect very distant relationships and putative fold similarities where purely sequence-based methods fail. We plan to apply GenTHREADER (12), a rapid fold recognition method, in the future on the subset of sequence regions for which a PDB match was not detected by the purely sequence-based techniques described above.

Problematic Structures

As well as excluding sequences having close structural matches, it is common practice to exclude domains corresponding to so-called problematic structures. These include low complexity (e.g. proline-rich) regions which are difficult to crystallize due to their high degree of conformational flexibility. These are identified using SEG (13). Also commonly excluded are membrane proteins despite their considerable biological interest, particularly as they currently represent a high proportion of human drug targets. Membrane protein structure determination methods are not sufficiently advanced for high-throughput approaches so we exclude domains containing two or more transmembrane helices, as predicted by MEMSAT (14, 15). SignalP (16) predicts N-terminal signal peptides and these are flagged so they can be excluded during cloning.

Sequence Clustering

The organization of all sequences into families is the basis for the rational selection of targets. This allows structural biologists to select the most suitable family representatives for structure analysis, and to select certain families with particular features, in order to minimize effort and focus their work on key areas of interest.

We have clustered all sequences from more than 100 genomes into protein families using the powerful Markov clustering method TRIBE-MCL (17). Families and individual sequences are annotated by integrating functional information from a variety of sources (e.g. Gene3D (18), GO (19), SWISSPROT (20), COG (21), WIT (<http://wit.mcs.anl.gov/WIT2/>), KEGG (22), PROSITE (23)) to guide family prioritization. Furthermore, protein-ligand complexes are invaluable in terms of the functional insights they provide and these data can direct experiments by suggesting substrates that may co-crystallize with the protein under investigation.

These and a variety of other databases can be utilized to provide information that may be significant for later functional analysis. They fall into 4 main classes:

1. Motif-based databases: PROSITE, PRINTS (24), Blocks (25)
2. Domain-based: Pfam (26), SMART (27), TIGRFAMs (28), ProDom (29), Gene3D, SUPERFAMILY (30)
3. Gene based: SWISS-PROT, COG, Systers (31)
4. Others: GO, KEGG, and WIT

Searching each of these separately is time-consuming so several 'meta-servers' such as Interpro (32), Motifscan (33) and NCBI's CDsearch (34) have been developed to gather data from several resources simultaneously.

Homology Modelling Coverage

Another consideration for target selection is the size of the family to which the protein belongs. Large families are preferred over those with few members since these ensure that many more proteins can be modelled once a template structure is obtained. Specifically, a family member with a large fraction of 'modelable' relatives is a preferred candidate for

structure determination as this will optimize the homology modelling coverage of the family. In practice, biophysical properties, such as methionine counts, also influence target choice.

Some structural genomics groups are also interested in solving the structures of ORFans, proteins found in a single organism that lack detectable homology to any other protein (35). Structural data could help unveil the mystery of these genes with regard to their origin and function. Do ORFans correspond to rapidly evolving proteins, newly evolved proteins or very distant members of known superfamilies if, indeed, they are expressed at all? Are they non-essential proteins or species determinants? Or, will many current ORFans turn out not to be ORFans at all but have homologues in genomes yet to be sequenced?

Species Distribution

The species distribution of a protein family is also considered in target prioritization. Families whose members span phylogenetically distant lineages in all three super kingdoms of life are of interest since they may have an unknown but fundamental ‘housekeeping’ function that the solved structure might help to define. Additionally, protein clusters unique to pathogenic organisms, or those that contain one or more human proteins, are prioritized owing to their medical relevance. Taxonomic data can be extracted from the NCBI Taxonomy Database (36).

Whatever the criteria used, at the end of the process a prioritized target list is produced and forms the basis for the experimental work. Research is currently under way to further prioritize certain targets if their sequence suggests they might be more amenable to structure determination – for example, they are more likely to be soluble, to crystallize, and so on.

DETERMINING FUNCTION FROM STRUCTURE

Now we turn to the other end of the structural genomics pipeline, from which emerges the experimentally determined models of protein 3D structures. In the ideal case, the function of each protein would have already been assigned during the target selection stage described above. However, if the function of the protein is not known, an effort needs to be made to determine it.

The most powerful methods are the sequence and motif methods already mentioned. However, if these fail to provide any functional information, all one is left with are the 3D coordinates of the structure itself. Can these reveal the protein’s function, or at least provide some clues to it?

In fact, there are a number of methods that predict function from structure, but they only work some of the time, and occasionally give conflicting results. They are described briefly below.

Fold Relatives

One of the most powerful comparative methods for functional inference is identification of homologous proteins through structural comparisons (37, 38). It has been known for some time that two sequences can diverge greatly while still retaining similar folds and even similar functions.

There are a number of search tools that can take a protein structure and scan it against the fold databases and retrieve the closest matches, usually with some measure of the significance of the match. The best known is DALI (39), which searches the FSSP (40) database. Others include GRATH (41) which searches the CATH (42) fold database, VAST

(43) which scans MMDB/PDB entries, and the SSM (Secondary Structure Matching) program at the European Bioinformatics Institute, EBI. Where no significant structural homologues are detected, it is possible to use fold recognition methods, such as GenTHREADER, to identify sequences in the sequence databases that are likely to adopt the same 3D fold as the target protein. The hope here is to find distant homologues of known function which might provide a hint of the target protein's function.

Care must be taken when attempting to assign function by structural homology (44) as there are many cases where relatives with very similar folds have been shown to perform very different functions. The best example is the TIM barrel fold, which has been shown to catalyse 60 different enzyme functions (45). The opposite situation in which different folds have evolved to perform the same function is also observed.

3D Motif Matching

Where the fold provides little or no functional information, the structure can still be useful as it may be possible to identify local clusters of residues in 3D that are associated with a particular function. Enzyme active sites are well studied (46, 47, 48) and the catalytic reactions they perform depend on the chemical characteristics of the residues and atoms responsible and their relative configuration in 3D. For this reason, catalytic residues are often very strongly conserved in terms of their 3D location.

The Protein Site Atlas, PSA, is a database of catalytic residues and their 3D coordinates (C. Porter, manuscript in preparation). Currently it holds 189 such motifs spanning a wide range of enzymatic functions. Using a fast 3D search program called JESS (49) it is possible to rapidly scan each one against the target protein and report all matches, together with their root-mean-square deviation (RMSD) and statistical significance. Similarly, a library of around 600 metal binding sites has been developed and, together, the lists of potential functional matches can be investigated further. Other programs that perform similar motif searching are ASSAM (50), TESS (51), RIGOR (52) and PINTS (53).

Comparing Binding Sites

A slightly different approach is to look at the target protein's surface, particularly its binding site, and compare this against a database of binding site surfaces. One approach (54) automatically extracts possible binding sites from a structure and then characterizes them using descriptors based on physiochemical properties (pseudocentres). A database of known functional cavities (Cavbase/Relibase) can then be searched for equivalents in not only shape but also physiochemical properties. A similar method compares the electrostatic surfaces of functional sites against a database called eF-site (55) in the hope of identifying distantly related binding sites. Recently, a method to predict nucleic acid binding sites based on surface patches has been published (56).

Evolutionary Data

A powerful method for identifying which residues are functionally important is to see which have been conserved over evolutionary time. One can calculate a conservation score for each residue in the sequence by comparing the residue variability at each position in a multiple sequence alignment of related proteins, as created by a program such as ClustalW (57). The residue conservation scores can then be mapped onto the protein's 3D structure. Two servers that use this method are the ConSurf (58) (<http://consurf.tau.ac.il/>) and Evolutionary Trace (59, 60, 61) servers (<http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/EtserverHome.html> or <http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html>). A java based implementation of the evolutionary trace method, JevTrace, has been developed by

Joachimiak & Cohen (62) in order to make the method more accessible and is available from <http://www.cmpharm.ucsf.edu/~marcinj/JevTrace/index.html>. These methods can identify likely functional or protein-protein interaction sites in the target protein but depend on there being a suitable number of sufficiently diverse sequence relatives in the sequence database in order to give meaningful residue conservation scores. Many hypothetical proteins do not have such a diverse set of sequence homologues.

Combining Methods

Each of the methods described above has its own success rates and limitations. Approaches that take many or all of these methods into account are more likely to give reliable functional predictions. In this respect we are developing a semi-automatic functional analysis pipeline as part of the MCSG. An outline of the pipeline is shown in Figure 2.

AN ANALYSIS OF MCSG STRUCTURES

To date, the MCSG has submitted 60 structures to the PDB, all of which had no structural homologues in the database at the time of submission. To date, we have analysed 24 of these. This has been an excellent dataset with which to test function prediction methods, as a third of the proteins are hypothetical proteins and several of the structures are novel folds (63, 64). We have met with varying levels of success in predicting function, as outlined by the pie charts in Figure 3.

The most difficult aspect of our functional analysis is to predict the cognate ligand and specific substrates. This is because even if reaction chemistry is conserved the binding site can readily mutate to allow different substrates to bind. It is also very difficult to assign the specific Enzyme Commission (E.C.) number to a new protein due to the fact that enzymes often change their function even at 40% sequence identity. Our ability to assign possible cofactors is better because particular cofactors are often associated with particular sequence patterns or folds. Identifying the most likely active site and the residues most likely to be conserved for function is the most confidently assignable feature.

The enzyme active site templates have provided strong functional matches for three of the MCSG hypothetical protein structures: a serine protease, a beta lactamase and a phosphoserine phosphatase. The first of these predictions has been experimentally confirmed (R. Sanishvili, manuscript submitted) while the latter two are currently undergoing the relevant assays.

REFERENCES

1. Terwilliger TC. Structural genomics in North America. Nat. Struct. Biol. Structural Genomics Supplement. 2000 Nov.2000:935–939.
2. Yokoyama S, Hirota H, Kigawa T, Yabuki T, Shirouzu M, Terada T, Ito Y, Matsuo Y, Kuroda Y, Nishimura Y, Kyogoku Y, Miki K, Masui R, Kuramitsu S. Structural genomics projects in Japan. Nat. Struct. Biol. Structural Genomics Supplement. 2000 Nov.2000:943–945.
3. Zhang C, Kim S-H. Overview of structural genomics: from structure to function. Curr. Opin. Chem. Biol. 2003; 7:28–32. [PubMed: 12547423]
4. Burley K. An overview of structural genomics. Nat. Struct. Biol. Structural Genomics Supplement. 2000 Nov.2000:932–934.
5. Blundell T, Mizuguchi K. Structural genomics: an overview. Prog. Biophys. Mol. Biol. 2000; 73:289–295. [PubMed: 11063776]
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]

7. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 1986; 5:823–826. [PubMed: 3709526]
8. Vitkup D, Melamud E, Moulton J, Sander C. Completeness in structural genomics. *Nat. Struct. Biol.* 2001; 8:559–566. [PubMed: 11373627]
9. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics.* 1991; 11:635–650. [PubMed: 1774068]
10. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic. Acids. Res.* 1997; 25:3389–3402. [PubMed: 9254694]
11. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics.* 1998; 14:846–856. [PubMed: 9927713]
12. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 1999; 287:797–815. [PubMed: 10191147]
13. Wootten JC, Federhen S. Analysis of compositionally biased regions in sequence databases. *Meth. Enzymol.* 1996; 266:554–571. [PubMed: 8743706]
14. Jones DT, Taylor WR, Thornton JM. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry.* 1994; 33:3038–3049. [PubMed: 8130217]
15. Jones DT. Do transmembrane protein superfolds exist? *FEBS Lett.* 1998; 423:281–285. [PubMed: 9515724]
16. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 1997; 10:1–6. [PubMed: 9051728]
17. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 2002; 30:1575–1584. [PubMed: 11917018]
18. Buchan DW, Shepherd AJ, Lee D, Pearl F, Rison S, Thornton JM, Orengo CA. Gene3D: Structural assignment for whole genes and genomes using the CATH domain structure database. *Genome Res.* 2002; 12(3):503–514. [PubMed: 11875040]
19. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* 2000; 25:25–29. [PubMed: 10802651]
20. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 2000; 28(1):45–48. [PubMed: 10592178]
21. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 2001; 29(1):22–28. [PubMed: 11125040]
22. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. *Nucleic. Acids. Res.* 2002; 30(1):42–46. [PubMed: 11752249]
23. Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* 2002; 3(3):265–274. [PubMed: 12230035]
24. Attwood TK. The PRINTS database: A resource for identification of protein families. *Brief. Bioinform.* 2002; 3(3):252–263. [PubMed: 12230034]
25. Henikoff S, Henikoff JG, Pietrokovski S. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics.* 1999; 15(6):471–479. [PubMed: 10383472]
26. Bateman A, Birney E, Cerruti L, Durbin R, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. *Nucleic Acids Res.* 2002; 30(1):276–280. [PubMed: 11752314]
27. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signalling domains. *Proc. Natl. Acad. Sci. USA.* 1998; 95(11):5857–5864. [PubMed: 9600884]

28. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O. TIGRFAMs: A protein family resource for the functional identification of proteins. *Nucleic Acids Res.* 2001; 29:41–43. [PubMed: 11125044]
29. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D. ProDom: automated clustering of homologous domains. *Brief Bioinform.* 2002; 3(3):246–251. [PubMed: 12230033]
30. Gough J. The SUPERFAMILY database in structural genomics. *Acta. Cryst. D.* 2002; 58:1897–1900. [PubMed: 12393919]
31. Krause A, Stoye J, Vingron M. The SYSTERS protein sequence cluster set. *Nucleic Acids Res.* 2000; 28(1):270–272. [PubMed: 10592244]
32. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, et al. The InterPro Consortium. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* 2001; 29(1):37–40. [PubMed: 11125043]
33. Pagni M, Iseli C, Junier T, Falquet L, Jongeneel V, Bucher P. TrEST, trGEN and Hits: Access to databases of predicted protein sequences. *Nucleic Acids Res.* 2001; 29:148–151. [PubMed: 11125074]
34. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* 2002; 30(1):281–283. [PubMed: 11752315]
35. Siew N, Fischer D. Twenty thousand ORFan microbial protein families for the biologist? *Structure.* 2003; 11:7–9. [PubMed: 12517334]
36. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2000; 28:10–14. [PubMed: 10592169]
37. Orengo CA, Todd AE, Thornton JM. From protein structure to function. *Curr. Opin. Struct. Biol.* 1999; 9:374–382. [PubMed: 10361094]
38. Moulton J, Melamud E. From fold to function. *Curr. Opin. Struct. Biol.* 2000; 10:384–389. [PubMed: 10851191]
39. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 1993; 233:123–138. [PubMed: 8377180]
40. Holm L, Sander C. Mapping the protein universe. *Science.* 1996; 273:595–602. [PubMed: 8662544]
41. Harrison A, Pearl F, Mott R, Thornton JM, Orengo CA. Quantifying the similarities within fold space. *J. Mol. Biol.* 2002; 323:909–926. [PubMed: 12417203]
42. Orengo CA, Mitchie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—A hierarchic classification of protein domain structures. *Structure.* 1995; 5(8):1093–1108. [PubMed: 9309224]
43. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. *Proteins.* 1995; 23(3):356–369. [PubMed: 8710828]
44. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. From structure to function: approaches and limitations. *Nat. Struct. Biol. Structural Genomics Supplement.* 2000 Nov. 2000:991–994.
45. Todd AE, Orengo CA, Thornton JM. Evolution of protein function, from a structural perspective. *Curr. Opin. Chem. Biol.* 1999; 3:548–556. [PubMed: 10508675]
46. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* 2002; 324:105–121. [PubMed: 12421562]
47. Todd AE, Orengo CA, Thornton JM. Plasticity of enzyme active sites. *Trends Biochem. Sci.* 2002; 27(8):419–426. [PubMed: 12151227]
48. O'Brian PJ, Herschlag D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & Biology.* 1999; 6:R91–R105. [PubMed: 10099128]
49. Barker J, Thornton JM. An algorithm for constraint-based structural template matching: application to 3D templates. *Bioinformatics.* (in press).
50. Spriggs RV, Artymiuk PJ, Willet P. Searching for patterns of amino acids in 3D protein structures. *J. Chem. Inf. Comp. Sci.* 2003; 43:412–421.

51. Wallace AC, Borkakoti N, Thornton JM. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* 1997; 6:2308–2323.
52. Madsen D, Kleywegt GJ. Interactive motif and fold recognition in protein structures. *J. Appl. Cryst.* 2002; 35:137–139.
53. Stark A, Sunyaev S, Russell RB. A model for statistical significance of local similarities in structure. *J. Mol. Biol.* 2003; 326:1307–1316. [PubMed: 12595245]
54. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* 2002; 323:387–406. [PubMed: 12381328]
55. Kinoshita K, Furui J, Nakamura H. Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics.* 2001; 2:9–22. [PubMed: 12836670]
56. Stawiski EW, Gregoret LM, Mandel-Gutfreund Y. Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.* 2003; 326:1065–1079. [PubMed: 12589754]
57. Mullan LJ. Tutorial section. Multiple sequence alignment – the gateway to further analysis. *Brief. Bioinf.* 2002; 3(3):303–305.
58. Glaser F, Pupko T, Paz I, Bell RE, Bechor D, Martz E, Ben-Tal N. ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics.* 2003; 19(1):163–164. [PubMed: 12499312]
59. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* 2003; 326:255–261. [PubMed: 12547207]
60. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* 1996; 257:342–358. [PubMed: 8609628]
61. Innis CA, Shi J, Blundell TL. Evolutionary trace analysis of TGF- β and related growth factors: implications for site-directed mutagenesis. *Protein Eng.* 2000; 13(12):839–847. [PubMed: 11239083]
62. Joachimiak MP, Cohen FE. JevTrace: refinement and variations of the evolutionary trace in JAVA. *Genome Biology.* 2002; 3(12):0077.1–0077.12.
63. Teplova M, Tereshko V, Sanishvili R, Joachimiak A, Bushueva T, Anderson WF, Egli M. The structure of the yrdC gene product from *Escherichia coli* reveals a new fold and suggests a role in RNA binding. *Protein Sci.* 2000; 9(12):2557–2566. [PubMed: 11206077]
64. Osipiuk J, Górnicki P, Maj L, Dementieva I, Laskowski R, Joachimiak A. *Streptococcus pneumoniae* ylxR at 1.35Å shows a putative new fold. *Acta Cryst., D.* 2001; 57:1747–1751. [PubMed: 11679764]

Figure 1.

Outline of the target selection protocol for MCSG. Algorithms and functional databases are italicised (see text for references). For simplicity, methods for handling domains within sequences are not addressed in the diagram.

Figure 2. Outline of the semi-automated functional analysis pipeline we have developed, illustrating the types of information that can be inferred from the combination of procedures.

Figure 3. The various 'success' levels of each of the major stages in our functional analysis pipeline are shown. The results are based on the analysis of 24 out of 60 deposited MCSG structures.