

From fold to function

John Moult* and Eugene Melamud

A number of recent advances have been made in deriving function information from protein structure. A fold relationship to an already characterized protein will often allow general information about function to be deduced. More detailed information can be obtained using sequence relationships to already studied proteins. Methods of deducing function directly from structure, without the use of evolutionary relationships, are developing rapidly. All such methods may be used with models of protein structure, rather than with experimentally determined ones, but model accuracy imposes limitations. The rapid expansion of the structural genomics field has created a new urgency for improved methods of structure-based annotation of function.

Addresses

Center for Advanced Research in Biotechnology, University of Maryland, Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD 20850, USA

*e-mail: moult@umbi.umd.edu

Current Opinion in Structural Biology 2000, **10**:384–389

0959-440X/00/\$ – see front matter

© 2000 Elsevier Science Ltd. All rights reserved.

Abbreviation

PDB Protein Data Bank

Introduction

Structure has an important role to play in providing insight into protein function. With the advent of structural genomics [1], this role has greatly increased in significance, as more and more structures are now being obtained in advance of any other sort of experimental information. Also, the completion of large-scale genome sequencing projects is resulting in the discovery of hundreds of thousands of new protein sequences [2]. The vast majority of these proteins have not been studied experimentally, nor will they be for a long time. Thus, annotation of their function relies on bioinformatics techniques. Structure provides a more powerful evolutionary link than sequence information alone, allowing the detection of more remote relationships and also providing a means of identifying detailed functional differences within protein families. New techniques, based on *ab initio* function analysis of structure without the aid of evolutionary information, are beginning to have a significant impact. We discuss the role of structure in understanding function, examine existing computational techniques for extracting function information and suggest how these may develop in future. Related reviews have recently appeared [3**,4**].

Function from protein structure

What sort of information about function can structure provide? The function of a protein can be usefully divided

into three types: phenotypic, cellular and molecular [5]. Phenotypic function corresponds to the influence that a protein has on the properties of an organism as a whole, cellular to the role of the protein in the cell system and molecular to those properties related directly to interactions between the protein molecule and other molecular species. Information about protein structure provides insight into molecular function. A classical example is provided by members of the chymotrypsin class serine protease family [6]. In humans, these proteins have a very wide variety of phenotypic and cellular functions, ranging over food digestion, sperm processing and regulation of the immune response, but have closely related molecular functions. All are proteases, with a common catalytic mechanism, and the different phenotypic and cellular roles are the result of differences in specificity of peptides cleaved. They also share a significant degree of sequence similarity. Structure provides essential information about these differences in the specificity and, therefore, about the detailed differences in molecular function. Other high-throughput experimental techniques [7], such as transcription profiling [8], detection of protein–protein interactions [9], cellular location tracking [10] and gene disruption [10], provide complementary information about phenotypic and cellular roles. Throughout this review, ‘function’ refers to molecular function.

Structure provides information about molecular function at different levels of detail. We consider three different levels of function resolution. At a low resolution, the class is known — the protein is an enzyme, a DNA-binding protein or a cell surface receptor, for example. At medium resolution, the type of function is known — the enzyme is a protease or a β -lactamase, say; the DNA-binding protein is a repressor. At high resolution, the specificity is known — for example, it’s a protease with trypsin specificity, it’s a β -lactamase with specificity for penicillins rather than for cephalosporins, it’s the *lac* repressor. Depending on how much is already known from other techniques, low- or medium-resolution function information may be very valuable, or add nothing. Different types of analysis of structure provide functional information at the different resolutions.

Categories of function from structure

The amount and resolution of the functional information that may be deduced from a structure depends on the relationship of the structure and its features to those already known.

We consider three situations: function information based on a fold relationship to an already known structure; function from analysis of the features of a protein structure; and function from models of protein structure.

Function from fold relationships

Low- and medium-resolution function information is provided in a straightforward manner by homologous fold relationships. A fold relationship between proteins arises in two different ways: the proteins have a common evolutionary ancestor (sometimes these are referred to as homologous folds [11] or as members of the same structural superfamily [12]) or the proteins have converged to the same fold and have no common evolutionary history (an ‘analogous’ fold relationship or just a fold relationship, in the language of SCOP [12]). As low- and medium-resolution function is often conserved within a superfamily, recognizing a homologous fold relationship does provide some functional information and functional hypotheses so generated can often then be verified by looking for conserved key sequence features, not in themselves statistically significant. For example, the structure corresponding to the *ycaC* gene was found to have a fold similar to that of the amidohydrolase family, consistent with a remote sequence relationship [13•]. Inspection of the structure determined that the same catalytic apparatus was present and indicated a small substrate — medium-resolution function information. When nothing else is known, this resolution of functional information can be very valuable. Analysis of recent PDB entries using the SCOP classification [12] shows that about half of the new sequence family entries were members of known superfamilies, so this method is widely applicable.

Function information from analogous fold relationships is less specific. For example, the ‘ferredoxin’ fold, one of the 10 superfolds listed by Thornton and co-workers [14], has so far been found as a ferredoxin, as an RNA-binding protein, as a protease propiece and as an amino acid metabolism regulatory domain. Although there are no concrete rules for choosing between these or a fifth function, inspection of the structure and comparison with the known examples will generally allow a distinction to be made.

Methods of comparing the folds of two structures are well developed and used routinely. Some of the more popular are Dali [15], VAST [16] and CE [17]. An extensive set of references to current methods is provided in a recent paper [18]. The different methods have different strengths and weaknesses for suggesting remote fold relationships and it is advisable to use more than one.

Function from analysis of structural features

Identification of ligand-binding sites

A simple and powerful method of identifying ligand-binding sites is to scan the surface of a protein molecule for clefts. In a systematic study, Laskowski *et al.* [19] found that, in 83% of cases examined, the largest cleft is the known primary binding site for small ligands and, in a further 9% of cases, the second largest cleft is the primary binding site. One of a class of methods that find clefts and cavities by filling vacant space with appropriately sized spheres was used. In practice, a GRASP [20] surface

representation of a protein is often sufficient to identify these probable small-ligand-binding sites and the signal may be enhanced by displaying local curvature. For example, the putative binding site in a structural genomics target protein, HI1434, was easily found in this way [21].

Identification of ligand-binding specificity

Whereas computational methods for identifying sites of ligand binding are well developed, those for finding exactly which ligands bind are still emerging. In principle, the methods that have been developed for drug design are applicable — testing a library of small molecules found *in vivo* for compatibility with the binding site. Because this list of compounds is relatively short, compared with the number of possible drug molecules, and because some binding properties may already have been inferred from sequence family relationships, this may be less challenging than drug design. On the other hand, conformational changes occurring upon binding natural ligands (induced-fit effects) may complicate the process.

Existing drug design software, such as DOCK [22] and HOOK [23], have potential. These two methods represent broadly complementary approaches. DOCK, one of the most popular drug design algorithms, uses a sphere-filling approach similar to the method used by Laskowski *et al.* [19] for finding clefts. The shape so defined is then matched against the possible conformations of compounds in a library. Possible ligands are docked into the site and the electrostatic compatibility of ligand and site is used as a final filter. In contrast, HOOK uses a library of small functional groups, such as carbonyl groups and nonpolar rings. Possible subsites for binding these groups are identified by constructing a grid in the site of interest and evaluating the energetic and steric fit of each group type at each grid point. The library of possible ligands is then scanned for combinations of groups consistent with the site mapping.

Established docking methods are not able to take into account conformational changes to the protein that may accompany binding. These induced-fit effects are common upon the binding of natural ligands and may differ between members of a family. Emerging new methods, such as SLIDE (SK Volker, LA Kuhn, personal communication), are able to address this issue.

It may also be possible to develop methods that make use of the relationship between ligands within a family — limited variation between different steroids or different peptide sequences, for example. Specific binding roles of particular protein groups can often be assigned for members of the family for which the structure is known and changes in these groups provide immediate clues as to the probable changes in the ligand. No effective automatic methods have yet appeared.

Metal-binding sites are a special case of ligand specificity. Visual inspection of the structure, together with knowledge

of the binding specificity, given a probable functional region, is quite effective. In addition, potential sites may be scanned for known three-dimensional motifs. These data are available for the most common metals, for example, calcium [24] and zinc [25]. Second shell methods based on local hydrophobic moment analysis [26] and methods based on three-dimensional cluster analysis [27] may also be useful.

Identification of macromolecular-binding sites

The identification of macromolecular-binding sites is a challenging aspect of function analysis that has made significant progress recently. There are four methods available at present.

First, protein DNA- or RNA-binding sites and interaction with membrane surfaces can often (but not always) be identified from a region of positively charged surface. A GRASP [20] surface potential map is ideal for this purpose. Second, protein interfaces have particular residue compositions. Analysis of known protein-protein interfaces shows that these patterns are distinctive, with a mixture of salt bridge and nonpolar interactions predominating [28], and do have some predictive power [29]. Third, a new searchable library of protein surfaces [30] provides a catalog of interaction surfaces that have previously been observed. Fourth, the evolutionary conservation pattern method described below can be particularly effective.

Use of evolutionary conservation patterns

Analysis of the conservation of residue types within a protein family, as a function of position on the protein surface, has emerged as a powerful method for identifying functionally important regions of a protein. If a sufficiently large and diverse set of sequences is available, an evolutionary tree may be built and the degree of variation in residue type plotted on the surface of the structure, at different evolutionary separations [31*]. This method has been shown to be effective at identifying ligand-binding sites and identifying a protein-protein interface in a *bona fide* prediction example [32].

Identification of enzyme catalytic mechanisms

When a new structure is an enzyme, the wealth of accumulated knowledge of catalytic mechanisms may be used. Many components of enzyme mechanisms have evolved multiple times. For example, the catalytic triad of the chymotrypsin class serine proteases is seen in the subtilisin family and the serine/oxyanion hole component is also found in the β -lactamases and related enzymes [33], among others. Thus, when a new structure is that of an enzyme, there is a reasonable chance that the mechanism, or at least some components of it, has been seen before.

A thorough knowledge of enzyme mechanisms and inspection of a new structure will often be sufficient to recognize a relationship to known catalytic machinery. PROSITE patterns [34] also provide useful clues. Work is in progress to produce a catalog of known three-dimensional motifs

involved in enzyme catalysis [35]. A newly determined structure may then be scanned against the catalog. Efficient scanning methods have already been developed [27,35]

Function from models of protein structure

The role of structure modeling in the functional annotation of sequences

Most functional annotation of newly discovered proteins is based on sequence relationships: a detectable sequence relationship to a previously annotated protein implies an evolutionary relationship, which, in turn, is assumed to imply a functional relationship. It is widely recognized that these methods have limitations and that, within any sequence family, there will be variation of function, at least at high resolution. The term 'ortholog' is used to imply that two proteins have the same function and 'paralog' to imply that, although sequence related, the two proteins have, in some sense, different functions, at least at high resolution. The copying of function annotation from one member of a sequence family to another tacitly assumes the pair of proteins are orthologs. Although sequence-related annotation methods are some of the most powerful and successful tools in computational biology, the resulting errors of function assignment are wide spread [36–38]. More sophisticated sequence analysis methods, particularly principal component analysis [39], can sometimes subdivide a family of sequences into probable orthologous sets. In general, however, the problem of reliably identifying paralogs and assigning function to them is unsolved.

Structural information derived from modeling provides a powerful way of both detecting more remote evolutionary relations than is possible from sequence and of obtaining higher resolution information than is possible from sequence alone. The usefulness of the approach depends critically on the reliability of the models. One method of assessing the state of the art in protein structure modeling is provided by the CASP experiments [40,41]. These community wide experiments collect *bona fide* predictions of soon to be solved protein structures from a large number of modeling groups worldwide (approximately 4000 predictions from 98 groups in the most recent experiment). Subsequent evaluation of these models by comparison with the corresponding experimental structures, using numerical methods and the judgement of independent assessors, provides a comprehensive view of the strengths and limitations of the modeling methods.

Function from fold recognition

It is often the case that although no sequence relationship can be detected between a protein of interest and any known structure, the fold will nevertheless turn out to have been seen before. Fold recognition methods use sequence information, together with a library of known folds, to try to detect such relationships. Currently, approximately 30% of new families entering the SCOP database have new folds, implying that perhaps some 30%

of sequences with no known fold could potentially be assigned using these methods. A wide variety of fold recognition methods have been developed. The results from the most recent CASP experiment [42] show that fold recognition is clearly more sensitive than sequence methods at detecting remote evolutionary relationships. Broadly speaking, current methods are effective at detecting homologous fold relationships some of the time. In CASP3, no single method came close to identifying all the folds, but results from a set of a few top-scoring methods would include the correct fold nearly always. The methods are currently much less effective at detecting analogous fold relationships; however, as these are less powerful at providing functional insight, this is not so important for the present discussion.

As with the direct comparison of folds discussed above, recognizing a homologous fold relationship does usually allow low- or medium-resolution function to be deduced. So, fold recognition methods are useful. They have been extensively applied to genome sequence datasets, providing tentative lists of functions for otherwise unannotatable ‘hypothetical’ proteins [43–45]. The deduction of higher resolution function information is severely limited by the accuracy of the models that can be built on the basis of fold recognition. The most serious source of error is alignment of the sequence of interest onto the experimentally known structure serving as the template for the model: an alignment mistake of one residue position introduces an error of about 3.8 Å in the backbone. Typically, these models have less than 50% of the residues correctly aligned [46]. This limitation should not be allowed to detract from appreciation of the power of fold recognition for obtaining low- and medium-resolution information.

Function from models based on close sequence family relationships

For all pairs of natural proteins so far encountered, a clear sequence relationship implies similar structures [47]. Thus, an approximate structural model may be built for any sequence clearly related to that of an already known structure. The model may then be analyzed using the function identification methods described above. Typically, proteins related by significant pairwise sequence similarity have similar function at the medium-resolution level, so that models need to deliver high-resolution function information, such as differences in ligand-binding specificity, in order to be useful. The appropriate function analysis tools described above may be used. Deduction of function is, of course, complicated by the relatively large size of the errors in current models, compared with those in high-quality crystal structures. The CASP experiments [41] provide detailed information on the accuracy of the structures produced with comparative modeling methods [48]. These show that the most serious source of error is alignment of the sequence of interest onto the structure of the experimentally known

structure serving as the template for the model [48]. Over the course of the CASP experiments, alignment quality for comparative models has improved [46]. As a rule of thumb, the best modeling groups will now usually produce correct alignments at higher than about 30% sequence identity but, at lower values, alignment still presents a serious challenge. Because of this, high-resolution function information can usually only be obtained from models based on this or on higher sequence identity.

Other sources of model error, particularly sidechain conformation and the backbone of ‘loops’, also will affect interpretation of function. However, analysis of the accuracy with which the best CASP models reproduce the features needed for interpreting ligand-binding specificity (C DeWeese-Scott, J Moulton, unpublished data) gives encouraging results and suggests that, providing uncertainty in a model is taken into account, these models will be very useful for obtaining high-resolution function information. For example, the approximate specificity of the fatty-acid-binding protein CRABP [49] could have been deduced from CASP models based on 42% sequence identity. There is also an example of a *bona fide* prediction of fatty-acid-binding specificity based on comparative modeling in the literature [50].

Full exploitation of comparative models to deduce high-resolution function requires the development of new software tools that systematically consider the uncertainty in atomic position. Particularly for ligand-binding specificity, such tools are clearly possible. The development of these tools, as well as improved comparative modeling methods, is particularly important in view of plans for large-scale structural genomics projects. These call for solving a set of representative experimental structures, such that ‘useful’ models can be built for the vast majority of all soluble proteins [51*]. Eventually, all of that modeling will be in the comparative modeling regime.

Function from models based on *ab initio* structure prediction

In principle, models of protein structure can be produced directly from sequence, without the need for fold or sequence relationships to already known structures. Methods for producing such *ab initio* models have shown significant progress over the CASP experiments [52]. Still, the quality of these structures is currently far below that needed to recognize functional features. Nevertheless, some optimism has been expressed that it is possible to devise algorithms able to identify functionally significant three-dimensional motifs, using ‘fuzzy functional forms’ [53].

Conclusions

Methods of deducing function from protein structure are already widely and effectively used. New methods will extend the scope significantly and the very large number of new unannotated sequences and structures will promote increased activity in the field.

Acknowledgement

This work was supported in part by the informatics component of National Institutes of Health grant P01 GM579890.

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. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW: **Structural genomics: beyond the human genome project.** *Nat Genet* 1999, **23**:151-157.
2. Genomes Online Database on World Wide Web URL: <http://igweb.integratedgenomics.com/GOLD/>
3. Shapiro L, Harris T: **Finding function through structural genomics.** •• *Curr Opin Biotechnol* 2000, **11**:31-35.
A recent review that also discusses methods of deriving function from structure.
4. Skolnick JF, Kolinski JSA: **Structural genomics and its importance for gene function analysis.** *Nat Biotech* 2000, **18**:283-287.
Another recent review that discusses methods of deriving function from structure.
5. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y: **Predicting function: from genes to genomes and back.** *J Mol Biol* 1998, **283**:707-725.
6. Schultz GE, Schimer RH: *Principles of Protein Function.* New York: Springer-Verlag; 1979.
7. Brent R: **Genomic biology.** *Cell* 2000, **100**:169-183.
8. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.
9. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al.*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403**:623-627.
10. Ross-Macdonald P, Coelho PS, Roemer T, Agarwal S, Kumar A, Jansen R, Cheung RH, Sheehan A, Symoniatis D, Umansky L *et al.*: **Large-scale analysis of the yeast genome by transposon tagging and gene disruption.** *Nature* 1999, **402**:413-418.
11. Russell RB, Saqi MA, Bates PA, Sayle RA, Sternberg MJ: **Recognition of analogous and homologous protein folds – assessment of prediction success and associated alignment accuracy using empirical substitution matrices.** *Protein Eng* 1998, **11**:1-9.
12. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247**:536-540.
13. Colovos C, Cascio D, Yeates TO: **The 1.8 Å crystal structure of the ycaC gene product from *Escherichia coli* reveals an octameric hydrolase of unknown specificity.** *Structure* 1998, **6**:1329-1337.
An example of fold similarity revealing information about the function of an enzyme.
14. Orengo CA, Jones DT, Thornton JM: **Protein superfamilies and domain superfolds.** *Nature* 1994, **372**:631-634.
15. Holm L, Sander C: **Dali: a network tool for protein structure comparison.** *Trends Biochem Sci* 1995, **20**:478-480.
16. Gibrat JF, Madej T, Bryant SH: **Surprising similarities in structure comparison.** *Curr Opin Struct Biol* 1996, **6**:377-385.
17. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.** *Protein Eng* 1998, **11**:739-747.
18. Szustakowski JD, Weng Z: **Protein structure alignment using a genetic algorithm.** *Proteins* 2000, **38**:428-440.
19. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM: **Protein clefts in molecular recognition and function.** *Protein Sci* 1996, **5**:2438-2452.
20. Honig B, Nicholls A: **Classical electrostatics in biology and chemistry.** *Science* 1995, **268**:1144-1149.
21. Zhang HH, Li K, Banajei Z, Fisher L, Grishin K, Eisenstein NVE, Herzberg O: **Crystal structure of YbaK protein from *Haemophilus influenzae* at 1.8 Angstrom resolution: functional implications.** *Proteins* 2000, **11**:25-30.
22. Briem H, Kuntz ID: **Molecular similarity based on DOCK-generated fingerprints.** *J Med Chem* 1996, **39**:3401-3408.
23. Eisen MB, Wiley DC, Karplus M, Hubbard RE: **HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site.** *Proteins* 1994, **19**:199-221.
24. Einsphar H, Bugg CE: **Calcium and its role in biology.** In *Metal Ions in Biological Systems*, vol 17. Edited by Sigel H. New York: Marcel Dekker Inc.; 1984:51-97.
25. Christiansen DW: **Structural biology of zinc.** *Adv Protein Chem* 1991, **42**:281-355.
26. Yamashita MM, Wesson L, Eisenman G, Eisenberg D: **Where metal ions bind in proteins.** *Proc Natl Acad Sci USA* 1990, **87**:5648-5652.
27. Zhu ZY, Karlin S: **Clusters of charged residues in protein three-dimensional structures.** *Proc Natl Acad Sci USA* 1996, **93**:8350-8355.
28. Jones S, Thornton JM: **Protein-protein interactions: a review of protein dimer structures.** *Prog Biophys Mol Biol* 1995, **63**:31-65.
29. Jones S, Thornton JM: **Prediction of protein-protein interaction sites using patch analysis.** *J Mol Biol* 1997, **272**:133-143.
30. SPIN-PP: Surface Properties of Interfaces on World Wide Web URL: <http://trantor.bioc.columbia.edu/cgi-bin/SPIN/>
31. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257**:342-358.
32. Lichtarge O, Bourne HR, Cohen FE: **Evolutionarily conserved G-alpha-beta-gamma binding surfaces support a model of the G protein-receptor complex.** *Proc Natl Acad Sci USA* 1996, **93**:7507-7511.
33. Herzberg O, Moul J: **Penicillin binding and degrading enzymes.** *Curr Opin Struct Biol* 1991, **1**:946-953.
34. Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999.** *Nucleic Acids Res* 1999, **27**:215-219.
35. Wallace AC, Laskowski RA, Thornton JM: **Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases.** *Protein Sci* 1996, **5**:1001-1013.
36. Brenner SE: **Errors in genome annotation.** *Trends Genet* 1999, **15**:132-133.
37. Zhang X, Smith TF: **The challenges of genome sequence annotation or 'the Devil is in the Details'.** *Nat Biotech* 1997, **15**:1222-1223.
38. Bork P, Bairoch A: **Go hunting in sequence databases but watch out for the traps.** *Trends Genet* 1996, **12**:425-427.
39. Casari G, Sander C, Valencia A: **A method to predict functional residues in proteins.** *Nat Struct Biol* 1995, **2**:171-178.
40. Supplement: **Third meeting on the Critical Assessment of Techniques for Protein Structure Prediction.** *Proteins* 1999, **37**(suppl 3):1-239.
41. Moul J: **Predicting protein three-dimensional structure.** *Curr Opin Biotechnol* 1999, **10**:583-588.
42. Murzin AG: **Structure classification-based assessment of CASP3 predictions for the fold recognition targets.** *Proteins* 1999, **37**:88-103.
43. Rychlewski L, Zhang B, Godzik A: **Functional insights from structural predictions: analysis of the *Escherichia coli* genome.** *Protein Sci* 1999, **8**:614-624.
44. Jones DT: **GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences.** *J Mol Biol* 1999, **287**:797-815.
45. Fischer D, Eisenberg D: **Predicting structures for genome proteins.** *Curr Opin Struct Biol* 1999, **9**:208-211.
46. Venclovas C, Zemla A, Fidelis K, Moul J: **Some measures of comparative performance in the three CASPs.** *Proteins* 1999, **37**(suppl 3):231-237.

47. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5**:823-826.
48. Jones TA, Kleywegt GJ: **CASP3 comparative modeling evaluation.** *Proteins* 1999, **37**:30-46.
49. Kleywegt GJ, Bergfors T, Senn H, Le Motte P, Gsell B, Shudo K, Jones TA: **Crystal structures of cellular retinoic acid binding proteins I and II in complex with all-trans-retinoic acid and a synthetic retinoid.** *Structure* 1994, **2**:1241-1258.
50. Xu LZ, Sanchez R, Sali A, Heintz N: **Ligand specificity of brain lipid-binding protein.** *J Biol Chem* 1996, **271**:24711-24719.
51. NIGMS Structural Genomics Target Workshop on World Wide Web URL:
 - http://www.nih.gov/nigms/news/meetings/structural_genomics_targets.htmlA report describing plans for structural genomics, which will have a large impact on the field of function analysis of structure.
52. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I: **Analysis and assessment of *ab initio* three-dimensional prediction, secondary structure, and contacts prediction.** *Proteins* 1999, **37**:149-170.
53. Fetrow JS, Skolnick J: **Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases.** *J Mol Biol* 1998, **281**:949-968.