Prediction of Protein Function from Structure: Insights from Methods for the Detection of Local Structural Similarities

Rafael J. Najmanovich, James W. Torrance, and Janet M. Thornton
European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

Introduction

Predicting the function of a protein from its three-dimensional (3-D) structure is a major intellectual and practical challenge. Despite the details inherent in the structure, extracting knowledge about what a protein does biologically and how it does it, is not straightforward. There are currently many proteins whose structure is known, but no information as to their function is available. One approach to try to elucidate the function of such proteins relies on the detection of local structural similarities to proteins with known function. This essay summarizes recent methods for the detection of local structural similarities.

Until recently, a protein would be selected to undergo the costly and time-consuming process of having its structure determined experimentally only after other experimental techniques had been used to characterize its function at different levels (biochemical, cellular, etc.), and it would seem that knowledge of the structure could further elucidate its function. With the advent of structural genomics projects, there has been a shift in this traditional research paradigm. This shift has been driven mainly by the desire to know all protein folds, as an increase in the number of folds would increase the accuracy of homology modeling. Thus, the main factor in the selection of targets for structural genomic projects has been that of low sequence similarity to proteins of known structure. As a result of this new paradigm, there are numerous proteins whose structure is known but no information of its function is available.

In recent years, the task of predicting function from structure has benefited from the impressive successes of similarity-based methods, especially sequence-based similarity methods (1). Such methods allow one to predict to some extent the function of a protein based on its similarity to another protein, whose function has previously been determined experimentally. These methods provide an attractive shortcut that allows one to avoid the theoretical difficulties involved in understanding the processes responsible for the function of proteins at a molecular level (2). In this context, the use of similarity-based methods, in particular structural similarity-based methods, can help advance our understanding of the physical processes involved in folding and molecular recognition by allowing one to focus attention on relevant regions of the protein.

The classification of proteins based on overall structural (fold) similarities is a well-established field. Public resources such as CATH (3) and SCOP (4) can aid in our understanding of the general principles behind protein architecture. These resources are based on a combination of manual curation and automatic methods. Several methods exist for the detection of overall structural similarity; methods such as DALI (5), GRATH (6), and SSM (7) are discussed elsewhere (8,9). These approaches share some obvious similarities (particularly in the search procedure) to methods for the detection of local structural similarities and can be used to predict function from structure.

Local structural similarities, particularly in active/binding site regions, can provide a different perspective into the evolution of a protein from that obtained using the overall sequence or structure. Further insight into this field and related areas can be found elsewhere (10,11).

Detection of Local Structural Similarities

Methods for the detection of local structural similarities can be conceptually deconstructed into three components: (*i*) representation, (*ii*) search, and (*iii*) scoring. The choices made in a given method on each of the three conceptual components might be interrelated. For example, certain choices of structure representation might preclude certain search methods and impose constraints on the way scoring is performed.

Different schemes are used to describe protein structure, sometimes simplifying, but at times adding more detail. Using different types of representation, one can implicitly account for effects such as side chain flexibility and differences in the types of amino acids found to be functionally similar in the structures under comparison. This can be particularly important when looking for structural similarities in proteins that reflect related but different enzymatic mechanisms (12,13).

The starting information for predicting function from structure is the atomic coordinates of heavy atoms (nonhydrogen) of the protein. This information is deposited in the Protein Databank (PDB) (14). The simplest level of representation of the protein structure ignores all but the C_α atoms (15). Using this representation, side chain flexibility is not an issue, but important side chain-based interactions are ignored. Several combinations of atoms and pseudo-atoms (points defining chemical moieties) can be used to increase the level of detail with respect to the C_α representation (16,17). Some methods simplify the description of the protein structure by mapping the structure onto linear strings of residues in a manner that reflects the spatial relationship between those residues (18–20) rather than the residue connectivity in the amino acid sequence. Using this type of representation, one can employ sequence similarity methods to detect local structural similarities.

In principle, any computational search algorithm can be adapted to be used to search for regions of local structural similarities in proteins. Two methods however, prevail as the most common choices: (*i*) detection of subgraph isomorphism (21), a technique to detect local structurally similar co-occurring clusters of atoms in two structures and (*ii*) geometric hashing (22), where an indexed 3-D grid is used for rapid similarity searching.

The most widely used parameter to score the goodness of a match is the root mean square deviation (RMSD) of the spatial coordinates, which provides a score in Ångstroms (10^{-10} m), measuring the geometric difference between two structures. RMSD can only be meaningfully calculated when there is a 1:1 correspondence between the units being compared in each of the

substructures (atoms, pseudo-atoms, etc.), and in many cases, the goal of the search method is to find the correspondence with the smallest geometrical difference (i.e., minimal RMSD). In general, this correspondence already takes into consideration chemical properties, such that a purely geometrical measure of likeness is sufficient. Other methods used to score matches involve the calculation of probability values (*P* value) or expectation values (*E* value) that measure, respectively, the probability or number of times that a match of such a type should be expected by chance. Sequence-based methods take advantage of the well-developed procedures used to calculate the statistical relevance of pair-wise matches to gauge the statistical relevance of the structural matches they find.

The different methods can (somewhat artificially) be classified into two distinct categories: (*i*) template-based methods and (*ii*) pair-wise comparison methods. This classification is independent of the three components discussed above. Figure 1 presents a schematic guide of the applications of template and pair-wise comparison methods.

A template is a predefined portion of a structure (substructure), either extracted from an existing structure as a representative example (e.g., a set of catalytically important residues or a structural motif) or defined by other means (Figure 1a). Template-based methods search the target structure for the presence of template(s) (Figure 1b) or a query template (or a database of templates) against a database of target structures (Figure 1c). Templates are more useful when they reflect prior knowledge of their biological relevance, be it structural or functional. When a substructure is found to be similar to a template, it might be due to convergent or divergent evolution. Yet in other cases the similarity might not be of any biological functional relevance, but merely the result of convergence to a stable structure. Until we learn how to apply known chemical and physical concepts to predict function from structure, experimental verification will continue to be necessary.

Pair-wise comparison methods search two or more protein structures for the presence of any common substructures. They do not require any prior specification of a substructure. As such, these methods have the advantage of detecting previously unknown statistically relevant similar substructures (Figure 1, a–g), thus providing a method to generate templates automatically (Figure 1h). One disadvantage of pair-wise comparison methods is that the uncertainties about the biological relevance of a match are even greater than those encountered with template methods that use annotated templates, even in cases where the match is statistically relevant.

Useful Structural Similarity Web-Based Resources

Recently Torrance et al. (23) created what is possibly the only curated nonredundant library of structural templates representing catalytic sites. Their templates are a nonredundant subset of those present in the catalytic site atlas (CSA) (24), which contains a collection of manually annotated catalytic sites from the scientific literature and their close homologues present in PDB structures. Such a library can be used to predict function from structure with greater certainty, as the templates are well curated and more likely to be functionally relevant than other template libraries created with less stringent criteria. The authors provide a web-based interface called Catalytic Site Search (CSS) with which users can search a target structure against their database of annotated templates (Figure 1b).

Binkowski et al. (25) developed a pair-wise method for the detection of local motifs in protein structures focused on the residues lining the surface of clefts and internal voids. Residues are arranged in a string in their original order of appearance in the primary structure. Such arrangement precludes the detection of conserved structural motifs in which the residues appear in a different order in the primary sequence (such as in the case

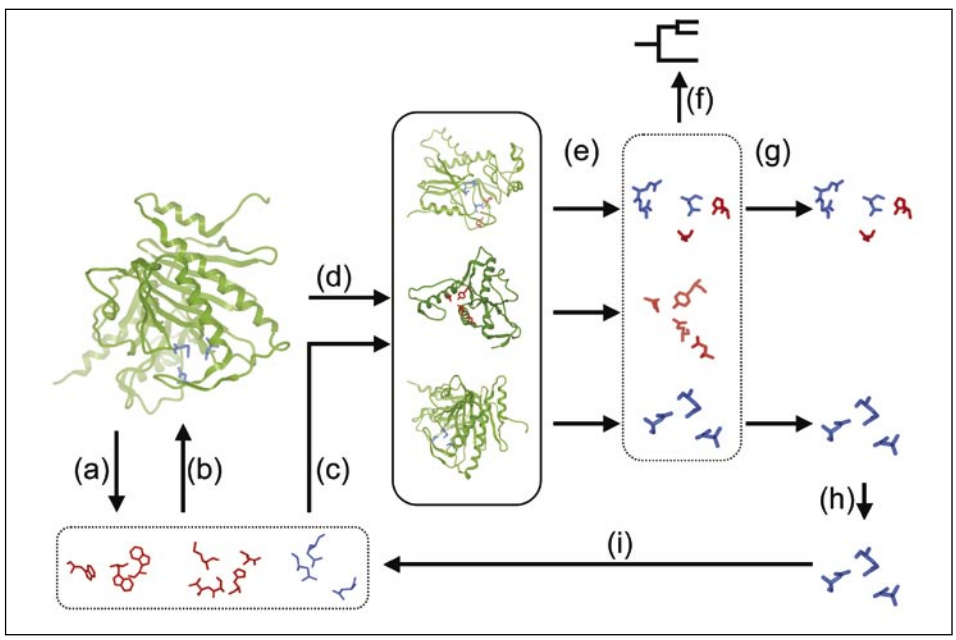

Figure 1. View of applications of methods for the detection of regions of local structural similarities. Parts a–c correspond to template-based methods, while parts d–i correspond to pair-wise comparison methods. (a) Experimental evidence can be used to generate a template from a protein structure. (b) Several methods use a database of templates, curated or automatically generated, to search a query structure for their presence. (c) Another typical application is the utilization of a specific template, extracted from an existing structure or generated by the user, to search a database of target structures. (d) Pair-wise comparison methods do not use templates: query and target proteins (perhaps a database of such structures) are compared, and (e) regions of local structural similarities are detected and scored usually on the basis of root mean square deviation (RMSD). (f) The regions of local structural similarities can be used to cluster the different structures in the database. This is especially useful if the regions being compared are restricted to binding site areas. (g) Considering the statistical relevance of matches can help focus in more biologically relevant hits. (h) Further comparison of such statistically relevant matches can be useful to define biologically relevant motifs. (i) Such automatically generated, perhaps biologically relevant, motifs can be added to a template database. Molecular graphics images were generated using ICM-Pro (Molsoft LLC, La Jolla, CA, USA).

of the catalytic triad of serine proteases). The authors created two databases, CASTp (Computed Atlas of Surface Topography of Proteins) and pvSOAR (pocket and void surface patterns of amino acid residues) that can be searched against a query structure uploaded by the user.

PINTS, a database of patterns in nonhomologous tertiary structures developed by Stark and Russell (26), uses a method (27) that is string-based but independent of the ordering of residues in the primary sequence, allowing users to search with predefined patterns of 3-D clusters of residues against databases of complete structures. The method also permits searching of entire structures against databases of 3-D patterns of residues that have been automatically generated representing particular residues likely to be functionally important.

Prediction of Function from Structure

There are a number of methods available to predict protein function by similarity. One can look for sequence similarities with proteins of known function. One can also look for global structural similarities with proteins of known function. In order to be useful for function prediction, methods for the detection of local structural similarities must not merely be capable of detecting functional similarity, they must (in at least a significant minority of cases) be more useful for detecting functional similarity than these existing methods. The most obvious case in which methods for the detection of local structural similarities should be more useful is where similarities in active sites are due to convergent evolution or where the novel protein has diverged so far from its relatives that similarities at the level of sequence and global structure are very difficult to detect. More subtly, methods for the detection of local structural similarities may be useful not only in cases where it is possible to make a prediction of function on the basis of sequence or global structure, but where their application can make the prediction with greater certainty.

Conclusions

Following the elucidation of genome sequences, the challenge is now to assign molecular function to all gene products and to create in silico models that represent the interplay of these molecules to create a living system. Using structure to help elucidate function is a relatively recent development, which has led to the creation of the novel tools described in this essay. A major challenge in the future will be to elucidate the higher order complexes and interactions present in living systems. Here too, structural studies involving computational analyses, such as those described herein, will have an important role to play.

Acknowledgments

The authors would like to thank Roman Laskowski and Richard Morris for their comments during the preparation of this manuscript. R.J.N. is funded by the Structural Genomics Consortium. J.W.T. is funded by a European Molecular Biology Laboratory studentship and is affiliated with Cambridge University Department of Chemistry.

References

1. **Laskowski, R.A., J.D. Watson, and J.M. Thornton.** 2003. From protein structure to biochemical function? J. Struct. Funct. Genomics *4*:167-177.
2. **Leckband, D. and J. Israelachvili.** 2001. Intermolecular forces in biology. Q. Rev. Biophys. *34*:105-267.
3. **Pearl, F., A. Todd, I. Sillitoe, M. Dibley, O. Redfern, T. Lewis, C. Bennett, R. Marsden, et al.** 2005. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res. *33*:D247-D251.
4. **Andreeva, A., D. Howorth, S.E. Brenner, T.J.P. Hubbard, C. Chothia, and A.G. Murzin.** 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. *32*:D226-D229.
5. **Holm, L. and C. Sander.** 1995. Dali: a network tool for protein structure comparison. Trends Biochem. Sci. *20*:478-480.
6. **Harrison, A., F. Pearl, I. Sillitoe, T. Slidel, R. Mott, J. Thornton, and C. Orengo.** 2003. Recognizing the fold of a protein structure. Bioinformatics *19*:1748-1759.
7. **Krissinel, E. and K. Henrick.** 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D Biol. Crystallogr. *60*:2256-2268.
8. **Sierk, M.L. and G.J. Kleywegt.** 2004. Deja vu all over again: finding and analyzing protein structure similarities. Structure (Camb) *12*:2103-2111.
9. **Novotny, M., D. Madsen, and G.J. Kleywegt.** 2004. Evaluation of protein fold comparison servers. Proteins *54*:260-270.
10. **Campbell, S.J., N.D. Gold, R.M. Jackson, and D.R. Westhead.** 2003. Ligand binding: functional site location, similarity and docking. Curr. Opin. Struct. Biol. *13*:389-395.
11. **Via, A., F. Ferre, B. Brannetti, and M. Helmer-Citterich.** 2000. Protein surface similarities: a survey of methods to describe and compare protein surfaces. Cell. Mol. Life Sci. *57*:1970-1977.
12. **Todd, A.E., C.A. Orengo, and J.M. Thornton.** 2002. Plasticity of enzyme active sites. Trends Biochem. Sci. *27*:419-426.
13. **Pearl, L.** 1993. Similarity of Active-Site Structures. Nature *362*:24.
14. **Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne.** 2000. The Protein Data Bank. Nucleic Acids Res. *28*:235-242.
15. **Pennec, X. and N. Ayache.** 1998. A geometric algorithm to find small but highly similar 3D substructures in proteins. Bioinformatics *14*:516-522.
16. **Shulman-Peleg, A., R. Nussinov, and H.J. Wolfson.** 2004. Recognition of functional sites in protein structures. J. Mol. Biol. *339*:607-633.
17. **Barker, J.A. and J.M. Thornton.** 2003. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics *19*:1644-1649.
18. **Karlin, S. and Z.Y. Zhu.** 1996. Characterizations of diverse residue clusters in protein three-dimensional structures. Proc. Natl. Acad. Sci. USA *93*:8344-8349.
19. **Jonassen, I., I. Eidhammer, and W.R. Taylor.** 1999. Discovery of local packing motifs in protein structures. Proteins *34*:206-219.
20. **Jonassen, I., I. Eidhammer, D. Conklin, and W.R. Taylor.** 2002. Structure motif discovery and mining the PDB. Bioinformatics *18*:362-367.
21. **Kresher, D. and D. Stinson.** 1998. Combinatorial Algorithms: Generation, Enumeration and Search. CRC Press, Boca Raton.
22. **Rigoutsos, I. and H.J. Wolfson.** 1997. Geometric hashing. Ieee Comp. Sci. Eng. *4*:9.
23. **Torrance, J.W., G.J. Bartlett, C.T. Porter, and J.M. Thornton.** 2005. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J. Mol. Biol. *347*:565-581.
24. **Porter, C.T., G.J. Bartlett, and J.M. Thornton.** 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. *32*:D129-D133.
25. **Binkowski, T.A., L. Adamian, and J. Liang.** 2003. Inferring functional relationships of proteins from local sequence and spatial surface patterns. J. Mol. Biol. *332*:505-526.
26. **Stark, A. and R.B. Russell.** 2003. Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. Nucleic Acids Res. *31*:3341-3344.
27. **Russell, R.B.** 1998. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J. Mol. Biol. *279*:1211-1227.