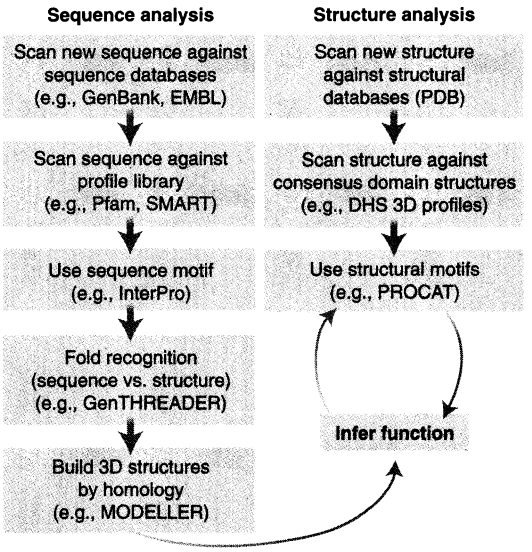
From Genome to Function

Janet M. Thornton

After the completion of the genome sequences, the challenge ahead for all biologists is to use the data to interpret the function of the protein, the cell, and the organism. There is no doubt that gathering, archiving, ordering, and classifying the data will be at the heart of this process, but bioinformatics must also set a framework from which we can extend our understanding of life and evolution. For the first time, biology can be regarded as finite, with a given set of molecular players whose interactions will determine the fate of the individual organism. New approaches to measure and identify all RNA (transcriptome) and protein molecules (proteome) in a cell will allow us to identify the critical players and the sequence of interactions of a given event. In doing so, scientists hope to gain a molecular understanding of the overarching biological processes, such as reproduction, aging, evolution, and, of course, the causes (and thereby cures) of diseases.

Traditionally, most biologists have not used, or indeed valued, computational approaches and modeling in their work, because most biological systems are very complex and the interactions among their components were still being discovered. However, the recent flood of sequence data and the vision of a complete model of life have created new challenges: First, to find the genes and determine or predict their structures and functions. Then, to understand their biological roles and model the complete cell or organism in silico. But what approaches are available and how good are they? Given a sequence or structure and using a computer, what can we deduce about a protein's function?



From sequence and structure to function.

The methods to analyze a novel sequence to predict its structure and function (see figure, this page) have become increasingly sophisticated (and often user friendly over the Web). The accuracy of each step absolutely depends on robust statistics, which provide a measure of significance.

Finding a homolog

The first step in assessing a new protein sequence is always to see whether the string of amino acids is similar to a known sequence: that is, to scan sequence databases [GenBank, European Molecular Biology Laboratory (EMBL)] for relatives. PSI-BLAST (*1*), which iteratively scans a sequence database to automatically build protein-specific profiles, is able to detect distant relations and reliably provide statistical significance. However, data derived from protein structures show that even PSI-BLAST can only identify about 2/3 of all evolutionary relationships (*2*).

An alternative approach to finding a sequence relative is to scan the sequence against a library of protein domain families. As more sequence and structure data have been gathered, it seems increasingly likely that there are a limited number of ancestral protein domains (*3*), which have duplicated and evolved into large families with great structural and functional diversity (*4*). Further diversity is introduced by mixing and matching these domains during evolution. These protein families can be thought of as the "elements of the periodic table of biology," from which biological complexity is created. In these libraries of protein domains [such as Pfam (*5*), SMART (*6*), and COGS (*7*)], a sequence alignment for each domain is constructed, which allows a novel sequence to be matched rapidly to domains already in the library. In the computer, a family is encoded as a profile or a Hidden Markov model (*8*) (HMM) that can sometimes detect more distant relatives than PSI-BLAST.

Many protein functional sites are well conserved and exhibit specific sequence motifs, which can provide surprisingly sensitive and specific search tools. Motif libraries, combined in the InterPro (*9*) resource, are sometimes useful in recognizing distant relatives or annotating a sequence. A new sequence can be rapidly scanned against these libraries to help identify functional sites within it.

If no sequence relative is found, or, as is more likely, no structural data are found for the family of interest, then a library of protein domain structures can be scanned to attempt fold recognition (i.e., compare sequence to structure rather than sequence to sequence) (*10*). The sequence is optimally aligned with each fold in turn. The match is scored using information derived from sequence similarity measures, secondary structure prediction, and empirical energy functions (from observed residue separations in proteins of known structure). The prediction is the matched protein structure that gives the best score. In blind tests (*11*), these methods are often able to identify more distant relatives than sequence comparisons alone. Theoretically, these methods aim to recognize all proteins with similar folds, but in practice only those with a common ancestor are found. The order of genes on a genome or pathway analysis can also be helpful for some proteins (*12*). These searching methods will generate a sequence alignment of the query and the matched sequence from which a structure may be built.

Building a model

If the sequence searches reveal similarity to a protein with a known structure, then a model of the new sequence can be built on the basis of the related protein's structure. In most proteins, the linear chain of amino acids folds up into a specific, compact, three-dimensional (3D) structure, which is essential for biological function. Local regions of the chain fold into secondary structures

The author is in the University College Department of Biochemistry and Molecular Biology, and the Crystallography Department, Birkbeck College, London WC1E 6BT, UK. E-mail: thornton@biochem.ucl.ac.uk

(e.g., α helix and β strand), which combine in motifs of increasing complexity (e.g., two, three, or four helices) to form the native state. Homology, or comparative, modeling yields a model based on a sequence alignment to a structure (*13*).

Modeling methods have improved as we have learned to "copy" the structure of the related protein more accurately. The quality of the model is directly related to the sequence similarity between the target sequence and the parent sequence on which the model is based. In blind tests, accurate models can be routinely built for sequences that are at least 35% identical (*14*). At this level of similarity, the overall tracing of the chain is accurate and the residues are located correctly, but the side chain conformations and the structure of long loops may be inaccurate. At lower levels of sequence identity, the alignment is often inaccurate, but reasonable models may still be built. Because most proteins belong to large families, this approach is proving to be of immense value. Large-scale modeling pipelines (*15*, *16*) can now rapidly generate models for all sequences that can be reliably aligned to any protein of known structure.

Recently, there has been progress in predicting the structures of proteins directly from sequence. Knowledge-based fragment assembly uses small bits of protein matched from the structure database and assembles them into folds (*17*). This approach combines improvements in secondary structure prediction (*18*), which now averages almost 80% accuracy, with classical and empirical energy potentials, which pack the motifs together.

But how accurately can we infer function of a given sequence or structure? Many protein domain families have duplicated and evolved to perform a wide variety of functions, which may or may not be related. As with homology modeling, the quality of transfer of functional information depends on the sequence similarity. Below 35% sequence identity between two proteins, the function often changes and care must be taken in transferring functional information (*4*). Detailed knowledge of functional residues and their roles can help in predicting change of function. Some families appear very specialized, whereas others exhibit functional promiscuity.

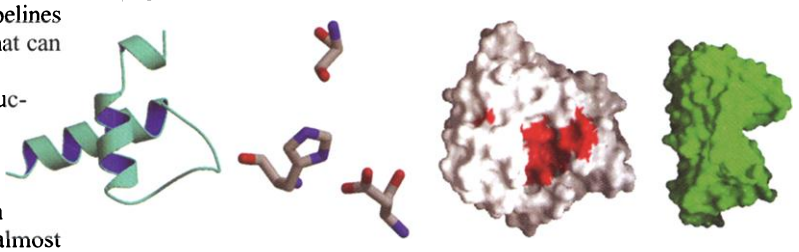
It will be increasingly common to elucidate a protein's function from its structure (*19*). The approaches used for analyzing protein structures mirror those used in sequence analysis and exploit many of the same algorithms. An outline of the steps involved in structure analysis is also shown (see figure, previous page) and is arranged to emphasize the similarity to sequence analysis.

Every new structure determined is now compared to each structure already in the Protein Data Base (PDB), generating a similarity score (*20*, *21*). Sequence comparison is performed in one dimension, but structure comparisons of shapes and topologies are much more complex. Even when no evidence for similarity has been detected from the sequence, it is common to find that a new structure is very similar to one already deposited. For example, last year less than 10% of the "newly determined" structures were novel. Furthermore, structural similarities are often associated with functional similarity and evolution from a common ancestor.

In an effort to understand more about the relation between protein sequence, structure, and function, classifications of structural domains, which cluster proteins according to their evolutionary and structural relationships (*22*, *23*), have been developed. These classifications provide the most accurate view of protein families because the tertiary structures reveal the most distant relations, but they have limited coverage due to the relatively sparse structural data. From these classification schemes, consensus sequence and structure models (cast as profiles or HMMs) are being developed, which can be used for scanning the databases (*23*). Such models, based on structure alignments, should be the most accurate.

As with sequence analysis, it is apparent that there are 3D structural motifs, which characterize the function of a protein (*24*) (see figure, this page). For example, almost all enzyme active sites are located in a deep cleft in the protein, a property which can be used to locate the active site. Similarly, the catalytic triad, first observed in chymotrypsin, is found in many different proteins with different functions. Several DNA binding motifs have been characterized, which can also suggest function. These coordinate motifs are surprisingly sensitive and specific and are therefore useful as template tools for scanning the structural databases. 3D-template libraries of these functional motifs are under construction.

In practice, sequence and structural analyses proceed hand-in-hand, because although the structural data can reveal relations hidden at the sequence level, the sheer number of sequences (>500,000) compared with the number of structures (only 12,000) provides distinguishing characteristics, stepping stones between members of a family. Ultimately, the sequence and structural libraries will coalesce as structural information for all protein domains is obtained. We need large, fully integrated databases that enumerate protein sequences and structures, their relations, functions and, eventually, their interaction partners. These data will constitute the basic dictionary and thesaurus of molecular biology. But even though a dictionary is essential to understanding a language, comprehending the words in context will require much more. Similarly, even when the biochemical function of a protein can be deduced, its biological role in the cell or organism often remains obscure. Elucidating protein function is the central focus of biology today, and computational approaches can only become more important in this challenge.



Function-related structural motifs. (Left to right) A helix-loop-helix DNA binding motif, a catalytic triad, a protein-protein interface defined by a conserved surface patch, and an enzyme cleft.

References and Notes

1. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
2. A. A. Salamov, M. Suwa, C. A. Orengo, *Protein Sci.* **8**, 771 (1999).
3. C. Chothia, *Nature* **357**, 543 (1992).
4. A. E. Todd, C. A. Orengo, J. M. Thornton, *J. Mol. Biol.* **307**, 1113 (2001).
5. A. Bateman *et al.*, *Nucleic Acids Res.* **28**, 263 (2000).
6. J. Schultz, F. Milpetz, P. Bork, C. P. Ponting, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 5857 (1998).
7. R. L. Tatusov *et al.*, *Nucleic Acids Res.* **28**, 33 (2000).
8. A. Krogh, I. Mian, D. Haussler, *Nucleic Acids Res.* **22**, 4768 (1994).
9. R. Apweiler *et al.*, *Nucleic Acids Res.* **29**, 37 (2001).
10. D. T. Jones, W. R. Taylor, J. M. Thornton, *Nature* **358**, 86 (1992).
11. J. Moult *et al.*, *Proteins (Suppl.)* **3**, 2 (1999); see http://predictioncenter.llnl.gov/casp4/.
12. M. Huynen, B. Snel, W. Lathe III, P. Bork, *Genome Res.* **10**, 1204 (2000).
13. P. A. Bates, M. J. E. Sternberg, *Proteins (Suppl.)* **3**, 47 (1999).
14. A. C. R. Martin, M. W. MacArthur, J. M. Thornton, *Proteins (Suppl.)* **1**, 14 (1997).
15. M. C. Peitsch, *Biochem. Soc. Trans.* **24**, 274 1996.
16. R. Sanchez, A. Sali, *J. Comp. Phys.* **151**, 388 (1999).
17. K. T. Simons, C. Strauss, D. Baker, *J. Mol. Biol.* **306**, 1191 (2001).
18. D. T. Jones, *Curr. Opin. Struct. Biol.* **10**, 371 (2000).
19. S. A. Teichmann, A. Murzin, C. Chothia, *Curr. Opin. Struct. Biol.*, in press.
20. S. Dietmann *et al.*, *Nucleic Acids Res.* **29**, 55 (2001).
21. F. M. G. Pearl *et al.*, *Nucleic Acids Res.* **28**, 277 (2000).
22. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J. Mol Biol.* **247**, 536 (1995).
23. J. E. Bray *et al.*, *Protein Eng.* **13**, 153 (2000).
24. A. C. Wallace, N. Borkakoti, J. M. Thornton, *Protein Sci.* **6**, 2308 (1997).

25. I would like to thank M. MacArthur and A. Todd for help with the illustrations.