Sequences and topology
Challenges for algorithms and experts
Editorial overview

Douglas L Brutlag* and Michael JE Sternberg†

Addresses
*Beckman Center, B400, MC 5307, Department of Biochemistry, Stanford University, Stanford, CA 94305-5307, USA;
e-mail: brutlag@stanford.edu
†Laboratory of Biomolecular Modelling, Imperial Cancer Research Fund, Lincoln's Inn Fields, London WC2A 3PX, UK;
e-mail: m.sternberg@icrf.icnet.uk

Current Opinion in Structural Biology 1996, 6:343–345

For more than a decade, it has been widely recognized that genome studies will yield a rapidly increasing number of nucleic acid and protein sequences requiring computer methods for storage and interpretation. Although many of the tools of sequence analysis (e.g. alignment algorithms by dynamic programming) have been available for many years, the challenges of performing these tasks with sensitivity and accuracy continues to stimulate innovation. This collection of reviews highlights several recent developments in sequence analysis. By contrast, only in the last few years has the number of known protein topologies solved by crystallography and NMR been increasing rapidly and this has now focussed interest on similar problems to those previously identified for sequences. Here we report on recent methods scanning structural databases for three-dimensional similarities. These computer tools for sequence and structural comparisons are now being used, together with expertise, to construct libraries of sequence motifs and fold families. Included in this issue are reviews on newly identified sequence and structural motifs. The growth of sequence data emphasized to a wider community the importance of predicting protein structure from sequence. Similarly, as more protein structures are solved, there will be an increasing need to model their associations, and, accordingly, we end our collection of reviews with one on recent developments in protein docking.

Our ability to identify relationships between sequence and topology has recently advanced for three major reasons. Firstly, many novel representations of sequences and sequence families have been developed. During the 1970s and 1980s, sequences, superfamilies and phylogenies were represented primarily by evolutionary models such as sequence alignments, amino acid substitution matrices and phylogenies. These representations attempted to recapitulate evolutionary steps but gave little insight into structural or functional features of the gene product.

In fact, it has been known for 10 years that classical sequence alignments do not correlate well with alignments of protein structure [1]. Several groups have recently approached this problem by representing near-optimal rather than optimal alignments (see review by Vingron, pp 346–352; [2–4]). This work has shown that the correct structural alignments are indeed among the near-optimal sequence alignments. These methods have also revealed so-called 'reliably aligned' regions correlated with structural elements. The problem remains of discovering which of the near-optimal alignments are the structurally or biologically relevant ones.

With the increase in the number of known protein structures, topologies and sequence families homologous to the known structures, sufficient information now exists to build secondary and tertiary structure-specific substitution matrices. Sometimes there is sufficient information from a single protein family to derive specific scoring systems for each individual position in a protein or a motif (blocks, profiles [5], templates [6], hidden Markov models), as discussed in the reviews by Henikoff (pp 353–360) and Eddy (pp 361–365). One can also detect structural features by examining correlations of the residues observed at two positions within a protein family or even within a short motif of common structure or function [7,8].

The second major advance that has accelerated our understanding of the relationship between sequence and structure is the development of automatic methods for discovering patterns and protein motifs from sequence families. The methods developed in the field of machine learning can extract conserved residues, discover pairs of correlated residues, and find higher order relationships between residues as well. Most of the methods for learning protein motifs from families of sequences derive directly from the fields of machine learning and signal processing. The include perceptions, discriminant analysis, neural networks, Bayesian networks, hidden Markov models, minimal length encoding, and context-free grammars to cite a few. In fact, if one wished to accelerate progress in understanding sequence–structure relationships, one need only investigate the methods that are currently being used and developed in the machine learning and signal processing area. The power of these methods is in both their ability to represent structural features rather than strictly evolutionary steps, and their ability to discover motifs from sequences automatically. Important methods

for evaluating and validating novel protein motifs have also derived from the machine learning area.

The third major advance in extracting structural information from sequences involves the conversion of probabilistic or quantitative representations to more discrete ones. Motifs are usually discovered by statistical methods (likelihood, mutual information, minimal length encoding, correlation, et cetera). However, it is only when the specific residues being conserved or correlated are examined that the biological meaning of the statistical observation becomes apparent (see review by Bork and Koonin, pp 366–376). While probabilistic methods are excellent for discovering weak interactions or conservation of structure, they are not very good for searching databases or even individual new protein sequences for such structures. It is their very robustness at discovering weak relationships in protein sequence families that cause them to lack precision (and therefore generate false positive results) when used for database searching. It is only when these quantitative observations are reduced to specific interactions that can be understood at the chemical or energetic levels that we obtain the biological relevance of the motif.

The identification of common protein structures provides valuable information about architectural principles and evolutionary relationships, and can assist in studies on sequence relationships and motifs. Recently there has been considerable interest in developing rapid methods to compare protein structures. In their review, Gibrat, Madej and Bryant (pp 377–385) describe how several groups are adopting a similar approach to achieve this goal. Rather than a complete description of the position of each Cα atom, proteins are represented by their secondary structure elements which are superposed to find the best substructure between two proteins. The next step is to evaluate whether the similarity is significant. Most approaches evaluate the number of possible alignments for these substructures between the two proteins to weight the score or a significance value attached to the score. These tools now enable the entire Brookhaven data bank to be compared and clustered revealing a taxonomic hierarchy of protein structures. The results are available at several World Wide Web sites given in the review.

In contrast to the use of a single algorithm to identify structural similarities, Murzin (pp 386–394) combines his own knowledge, literature reports and the results of algorithms to update the SCOP (structural classification of proteins) database [9]. In his review, Murzin reports on several new structural superfamilies together with the possible evolutionary implications. Three families are described in which there is a relationship between a factor (a protein that exercises control via specific interaction with another molecule) and an enzyme. In addition, this review highlights how structural comparisons can reveal a hidden motif from the identification of sequence alone. Indeed, the recognition of a hidden motif can assist in the prediction of structure/function relationships.

The importance of the identification of a sequence and structural motif is illustrated by the RING finger in the review by Borden and Freemont (pp 395–401). The RING finger was originally identified by sequence analysis in a few, previously unrelated proteins as a novel cysteine-rich motif that liganded two zinc atoms. The NMR structures of two RING fingers have now been determined. The two proteins have low sequence identity (15%) outside the zinc-binding regions. Although there are certain highly conserved regions between the two proteins involved in zinc binding, the two proteins have only a roughly similar fold. A structural superposition yields a 4.5 Å root mean square difference for 25 Cα atoms. Thus, attempts to predict the structure of one protein from another by comparative modeling are likely to yield a poor result. The review also highlights a major problem facing sequence and topology work — the function of the RING finger remains undetermined, although it might be involved in multiprotein complexes.

As more protein structures are solved, it is becoming increasingly important to model their interactions with other biomolecules. Lengauer and Rarey (pp 402–406) report the recent advances in computational approaches in both protein–protein and protein–ligand docking. Recently, a blind test of several protein–protein docking algorithms has been reported [10]. The coordinates of β-lactamase and its inhibitor in the unbound state were supplied and workers were challenged to predict the structure of the complex that had been determined crystallographically. All six groups that applied algorithms obtained one solution that was between 1.1 Å and 2.5 Å from the crystal structure. Most groups supplied a few solutions (3–15), some of which were incorrect. Most of the algorithms used rigid-body docking, suggesting that for some protein–protein complexes, modeling will be able to yield valuable biological information.

The docking challenge was an example of the importance of organizing blind tests of algorithms. The reader is referred to the issue of *Proteins: Structure, Function and Genetics* that reports the results of the Asilomar Protein Structure Prediction Challenge [11]. Comparative modeling, fold recognition and *ab initio* predictions were tested. In general, comparative modeling faired poorly, except for closely homologous structures. The recognition of a common fold in the absence of a relationship detectable from sequence alone yielded encouraging results. The commonality of folds could often be recognized but the actual sequence alignments were generally wrong. Actually evaluating whether two different folds can be considered sufficiently similar to deem a threading result a success remains problematic. Interesting *ab initio* tertiary structure predictions were successfully performed using expert knowledge to interpret the results of algorithms.

The different contributions to this issue illustrate the ranges of approaches to advancing our understanding of the interrelationships between sequence, structure and function. Certain problems today (e.g. protein docking) are best solved by algorithms. However, in many areas (e.g. assembly of fold families) the complexity of the information has not yet been adequately incorporated into algorithms and workers should not shy away from incorporating expertise to modify the results of algorithms. A major challenge in the area is to use more advanced computational methods to incorporate this expertise into the present day algorithms.

References

1. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, 5:823–826.
2. Zuker M: **Suboptimal sequence alignment in molecular biology: alignment with error analysis.** *J Mol Biol* 1991, 221:403–420.
3. Saqi MAS, Sternberg MJE: **A simple method to generate non-trivial alternate alignments of protein sequences.** *J Mol Biol* 1991, 219:727–732.
4. Naor D, Brutlag DL: **On near-optimal alignments of biological sequences.** *J Comput Biol* 1994, 1:349–366.
5. Fischer D, Rice D, Bowie JU, Eisenberg D: **Assigning amino acid sequences to 3-dimensional protein folds.** *FASEB J* 1996, 10:126–136.
6. Overington J, Donnelly D, Johnson MS, Šali A, Blundell TL: **Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds.** *Protein Sci* 1992, 1:216–226.
7. Korber BT, Farber RM, Wolpert DH, Lapedes AS: **Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis.** *Proc Natl Acad Sci USA* 1993, 90:7176–7180.
8. Klingler TM, Brutlag DL: **Discovering structural correlations in alpha-helices.** *Protein Sci* 1994, 3:1847–1857.
9. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of protein databases for the investigation of sequences and structures.** *J Mol Biol* 1995, 247:536–540.
10. Strynadka NCJ, Eisenstein M, Katchalski-Katzir E, Shoichet B, Kunta I, Abagyan R, Totrov M, Janin J, Cherfils J, Zimmerman F *et al.*: **Molecular docking programs successfully determine the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase.** *Nat Struct Biol* 1996, 3:233–238.
11. Lattman EE (Ed): **Protein structure prediction: a special issue.** *Proteins* 1995, 23:295–460.