Jeremy Draghi

The "Soap-Film" Method of Comparing Protein Structures

Methods for quickly and automatically comparing two protein structures have been an area of active research for several decades, and remain a major problem in bioinformatics. Although the structure of a protein is based primarily upon its sequence, structures can reveal homology, or similarity though common descent, in cases where the divergence of the sequences masks any similarity[1,2,3]. Dissimilar sequences can also be folded into similar structure in cases of analogy, or similarity by independent evolution under similar physical constraints[2]; such similarity is only revealed by structural alignments. Structural alignments are also used to validate other comparisons, such as sequence alignments, necessitating the highest possible standards of accuracy for structural comparisons[4]. Finally, many researchers judge the ability to compare structures essential to solving the protein folding problem[1,2,3], often considered the "holy grail" of bioinformatics.

As alternatives to the slow and subjective process of manual alignment, numerous automated methods have been developed to compute the similarity of two protein structures. One popular method, root mean square deviation (RMSD), uses an iterative version of the dynamic programming technique used to align protein sequences. Based on an initial sequence alignment, RMSD minimizes the root mean squared distance in three-dimensional space between pairs of aligned vertices on the carbon backbone by rotating and translating one protein as a rigid body. Although this method is fast and conceptually simple, it is dependent on a sequence alignment, which itself depends on somewhat arbitrary gap penalties[1]. This dependence on alignment is clearly a weakness in cases of analogous proteins or those with distant common ancestry. Other methods align structures based on secondary structures or other motifs, and produce local alignments optimized to align regions of high similarity[1,2,3]. Although these methods are less biased by insertions and deletions than RMSD, they still require a potentially problematic sequence alignment[1].

Falicov and Cohen described a new method of structural comparison based on an earlier concept of minimizing the surface stretched between the carbon backbones of two proteins[1]. This method retains the desirable hyposensitivity to insertions and deletions found local alignment methods, but does not require an initial sequence alignment. This method, called the Area Functional with Fit Comparison (AFFC), also retains the conceptual and visual appeal of the RMSD method, while producing meaningful results in cases of low or nonexistent sequence identity[1]. A brief description of the details behind AFFC will reveals its computational and conceptual advantages, and its similarities and differences with RMSD.

First, each protein is reduced to a three-dimensional representation of its carbon backbone, which forms a continuous chain of C_α atoms. For any two curves in space, there are an infinite number of surfaces that can be drawn between them; there is always, however, exactly one surface with the minimum area. This surface is nearly identical to the shape of a soap bubble stretched between two wire curves[1], making this abstract geometric idea intuitively understandable. Falicov and Cohen approximate this surface with a series of triangle of two types: Type 1 triangles consist of two adjacent C_α atoms on protein 1 and one C_α atom from protein two, while the opposite triplets compose Type 2 triangles. Generally, Type 1 and 2 triangles alternate in areas of sequence similarity, while stretches of only one type are indicative of insertions or deletions[1]. A measure of dissimilarity, called AF, is then calculated as the mean area of the triangles divided by the mean length of the two structures. Dynamic programming is then used to calculate the minimum area, and one C_α backbone is then translated and rotated as a rigid body to achieve this minimum[1]. In addition to AF, a measure of alignment significance, called the fit comparison (FC), was also calculated. FC is defined as the ratio of the AF to the mean distance between any pair of resides on different backbones, and provides a measure of significance independent of the length of either sequence[1].

Like RMSD, AF is measured in Angstroms, and is calculated for each region of the protein. This allows both measures to be used to evaluate the similarity of subregions of two structures. Also, like RMSD, AFFC violates the principles of optimality on which dynamic programming is based; the most common solution to this issue, and the one which Falicov and Cohen use, is to iterate the procedure of optimizing and moving the proteins until some specified threshold is reached[1]. Unlike RMSD, however, AFFC bases its optimum on a simple average of areas, not on the mean squared deviations, and AFFC makes no assumptions as to which C_α atoms in one protein correspond to C_αs in the other[1]. For these reasons, AFFC is more tolerant of insertions and deletions.

The authors of AFFC hypothesized that their method should correlate well with RMSD for proteins with nearly identical sequences, but that the two methods should diverge with lower sequence identity[1]. For a database sets of multiple NMR structures for each of a set of proteins with the same sequence, the two methods did indeed correlate perfectly[1]. For the FSSP database, a subset of the PDB containing only proteins with less than 30% sequence identity to any other protein in the set, the scores generated by the two methods for pair-wise comparisons, as predicted, displayed little correlation[1].

To establish confidence in the alignments generated by AFFC, the authors compared the clustering into families of proteins with known structural similarity with the family clusters established with dynamic programming methods of local alignment[1,2,3]. Detecting families for all sizes of proteins required the use of AF for large proteins and CF for shorter sequences, but established families did correctly cluster together for sets of α, β, trypsin-like serine proteases, and α/β proteins. Without proving its practical superiority, as opposed to conceptual appeal, over older methods like RMSD, AFFC does pass an important test by clustering these proteins in accordance with other methods and general opinion. This, however, does not prove AFFC is, as its authors expect it to be, superior in the domain of proteins with low sequence identity, and it is in this domain that other methods are least trusted.

Despite this lack of overwhelming empirical evidence, there are certainly theoretical reasons for taking this "soap-film" method seriously. Particularly appealing is the lack of a dependence on an underlying sequence alignment, an approach warranted by more than just the imperfection of current sequence alignment tools. Aligning sequences is motivated by our understanding of the biochemistry by which insertions, deletions and substitutions alter DNA sequences; these mechanisms make the approach of sliding sequences past each other and inserting gaps logical[4]. We do not, however, have a similar understanding of the process of protein folding, nor can we guarantee that the optimal sequence alignment should correspond to the optimal structural alignment[4]. None of these reasons, of course, prove that AFFC is a superior method, but I think it is a step in the right direction, and that, based on these considerations, the ultimate method of structure comparison will be similarly global and free of dependence on sequence alignment.

References

1. Falicov, Alexis and Cohen, Fred E. A Surface of Minimum Area Metric for the Structural Comparison of Proteins. *J. Mol. Biol.* (1996) 258: 871-892.
2. Orengo, Christine A. and Taylor, William R. A Local Alignment Method for Protein Structure Motifs. *J. Mol. Biol.* (1993) 233: 488-497.
3. Holm, Liisa and Sander, Chris. Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol*. (1993) 233: 123-138.
4. Godzik, Adam. The structural alignment between two proteins: Is there a unique answer? *Protein Science* (1996), 5: 1325-1338.