An Intelligent System for Comparing Protein Structures

Ethan Benatan

Department of Biological Sciences
University of Pittsburgh
Pittsburgh PA 15260
ethan+@pitt.edu

Abstract

An approach to protein structure comparison is presented which uses techniques of artificial intelligence (AI) to generate a mapping between two protein structures. The approach proceeds by first identifying the seed of a possible mapping, and then searching for ways to extend the seed by incorporating corresponding elements from the two proteins. Correspondence is judged using heuristic functions which assess the similarity of the structural environments of the elements. The search can be guided by separately encoded knowledge. A prototype has been implemented which is able to rapidly create mappings with a high degree of accuracy in test cases.

Introduction

The long-term goal of this project is the construction of an intelligent system which will work as an assistant to biologists in their studies of the three-dimensional structure of proteins. Such a system will communicate with humans in terms of their domain knowledge. Typical interactions with such a system might include requests to describe how regions related to a query structure have varied through evolution, or to find examples that will help determine the stability of a proposed mutation. The system will require a great deal of flexibility in searching, comparing, and reasoning about protein structures, as well as an ability to cope with human modes of reasoning about them. Described here is one step in the construction of such a system: an approach to the comparison of protein structures that uses the principles of artificial intelligence together with domain knowledge to compare protein structures in a flexible yet efficient way.

The three-dimensional structure of proteins is more strongly conserved through evolution than is sequence. Proteins with undetectably low sequence similarity may share a similar fold (Chothia 1992; Chothia & Lesk 1987). If a protein structure is considered to consist of a set of substructures which have certain spatial and chemical relations to each other, then it can be said that what is shared by distantly related structures is some pattern of these relations. Comparing two protein structures can thus be seen as creating a mathematical mapping between their component parts with the goal of finding conserved patterns of relations, perhaps revealing a biologically meaningful similarity. If two protein structures have related folds, a useful mapping should reveal a common pattern of relations. This work serves in part to test the converse: that finding a conserved pattern of relations shared by two proteins permits the production of a useful mapping between them.

This work was supported in part by the Keck Center for Computational Biology at the University of Pittsburgh, Carnegie Mellon and the Pittsburgh Supercomputing Center.

Related Work

Structure comparison

An excellent review of methods for protein structure comparison is provided by Orengo (1992b). The method presented here uses concepts from several of the methods reviewed there. In particular, the method of Sali and Blundell (1990) and the method of Orengo and Taylor (1992a) represent multiple levels in the hierarchy of protein structure; each of these methods has been successful in meeting its design goals. The method of Sali and Blundell was designed primarily to extract data for knowledge-based modeling of proteins, and results from the method have been used in the program COMPOSER (Sali et al. 1990; Topham et al. 1990). The method of Orengo and Taylor was designed with the aim of analyzing and clustering the structures in the Protein Data Bank, and has been used for that purpose (Morris et al. 1992). Holm & Sander (1993) recently presented a novel method of structure comparison based on a distance-matrix representation. After decomposing the distance matrix from each protein into many matrices each representing the local structure of a small stretch of the backbone, their algorithm uses a Monte Carlo approach to find a near-optimal alignment of the two structures. The program was used to carry out a cross comparison of 225

representative structures from the database, leading to the discovery of several unexpected structural similarities.

An intelligent system approach. The goal of constructing an intelligent assistant system demands an approach to comparisons that is more flexible than any of the methods described above. The work presented here differs from currently available methods in two important ways, both of which stem from the AI-influenced philosophy of the approach. First, it provides a general framework for incorporating a variety of heuristic functions and search strategies; to this extent, it can be used as a tool to explore the integration and generalization of current and novel methods. Second, it guides and carries out search using knowledge which is encoded in an encapsulated, flexible form. These two features combine to provide the versatility required for an intelligent assistant system.

Artificial intelligence

The approach presented here is based on principles which are widely accepted in the field of AI. It does however owe a special debt to the structure mapping theory of analogy, as described by Gentner (1989) and implemented in the computer program SME (Falkenhainer, Forbus & Gentner 1989). Structure mapping creates an analogy between two schemas by identifying a common structure of relations within each, in a way very similar to that described here for protein structures. There are, however, significant differences between the two methods. The structure mapping theory is a very general approach to analogy, appropriate for use in knowledge-poor situations. For this reason, it ignores all unary relations (properties), and creates a mapping between two domains by considering only higher-order relations (Gentner 1989). This work, in contrast, lies within a single domain, about which much knowledge is available. It therefore incorporates unary relations and semantic interpretations of higher-order relations by using them in heuristic functions.

Flexible Representation. It is well known (Rich & Knight 1991) that the use of appropriate abstractions in representing a domain can be a major factor in the accuracy and efficiency of problem-solving. Working at the highest appropriate level of abstraction greatly reduces the amount of data one must manipulate and can greatly simplify problems of search. Abstraction is pervasive in human reasoning in general, and in reasoning about protein structure in particular: abstractions such as α-helices and β-sheets are almost invariably used to describe structures, and every molecular biologist is familiar with ribbon diagrams (Richardson 1981). One of the strengths of the approach presented here is that it permits the use of abstractions where appropriate, and enables the easy incorporation of new kinds of abstractions. For this reason, few references are made to specific types of substructures in the description of the method below. Instead, the term "element" is used to refer to a component of protein structure without specifying the particular characteristics of the component. In keeping with this philosophy, an object-oriented approach has been used in the design and implementation of the system.

The following four hierarchically related classes of elements are considered. Single *residues* represent amino acid residues in the protein. *Contiguous substructures* represent sets of residues which form contiguous structures along the protein chain. *Discontiguous substructures* are comprised of arbitrary groupings of contiguous substructures. *Prosthetic groups* represent chemical components of a protein structure other than amino acids. Some examples of possible elements are domains, turns, binding sites, heme groups and β-sheets.

Guided Heuristic Search. Heuristic search methods are used to construct a mapping between the elements of two proteins. At each step, a mapping may be produced or extended by declaring a *correspondence*—a pair of elements, one from each protein, that are mapped to each other. A search begins with the (somewhat arbitrary) identification of some correspondence; this forms the seed of a mapping. The search then continues by evaluating further pairs of elements and selecting correspondences for addition to the match. This process is guided by two kinds of knowledge—heuristic evaluation functions which are used to assess the quality of a particular partial mapping, and control knowledge which is used to determine how to search and which heuristics to employ. The program can use information about the state of the mapping in applying these kinds of knowledge. Although the approach permits the use of any kind of evaluation function, only functions which can be applied in an incremental fashion are considered in this paper. That is, rather than assess the overall value of a partial match at each step, functions are used which estimate or assess the value of extending a match by adding to it a particular pair of elements which appear to correspond to each other. This eliminates a great deal of redundant calculation and reduces the complexity of typical evaluation functions (which examine some relation between each pair of elements under consideration) from polynomial to linear in the size of the match. Experiments are presented below which use several different evaluation functions to approximate the quality of a partial match between two proteins.

Methods

A prototype of this system has been implemented as a computer program and used it to perform some preliminary tests of the approach. The goals in this testing were to explore a set of heuristics and search strategies for speed and accuracy. In addition, the construction of the prototype allowed experimentation with ways of constructing a flexible framework within which to combine these approaches. The prototype uses elements of three of the classes described above: residue, contiguous substructure, and discontiguous substructure. Contiguous substructures are defined by the termini of the elements of secondary structure–helix, strand, and turn–as reported by the authors in the Protein Data Bank entries. Contiguous regions of the chain that lie between such recognizable structures (so-called "random coils") are also considered. One type of discontiguous substructure, the β-sheet, has been implemented. The mappings reported here are resolved to the level of contiguous substructures.

Several sets of experiments have been carried out, using a small set of protein structures, in order to provide some preliminary tests of the approach. The test set included four proteins from among those used in a definitive manual comparison of the globins by Lesk and Chothia (Lesk & Chothia 1980); the structures used are presented in Table 1. Proteins were chosen so as to include both closely and distantly related structures (as judged by sequence comparisons, see Table 2). (Additional structures representing the mainly-β and α/β structural groups were also used for testing, with similar results; the data are not shown here.) Experiments were carried out using three different combinations of heuristic evaluation functions and search control strategies.

The first set of experiments considered only a set of unary relations, or attributes, of each element: the type, size, charge, hydrophobicity, amino acid sequence, and in the case of helices, amphipathy. In this set of experiments, coil regions were not mapped.

Charge, hydrophobicity and amphipathy were compared by taking the absolute difference of the attributes; size was compared by using the ratio of the number of residues in each structure, smaller over larger; sequence similarity was judged by the number of identical residues found in the two sequences, using a simple dynamic programming approach with no gap penalty. A type correspondence score of one was assigned when the type of two elements was identical, zero if they were not. The final similarity score was arrived at by taking the weighted average of these measures.

To generate the mapping, a matrix was created in which each possible pairing of elements was assigned a score estimating the overall similarity of their attributes, and a greedy approach was used to select the correspondences to use. At each step, the pair of elements with the highest the highest correspondence score was selected and removed from further consideration. The process was repeated until no more pairs were available. The procedure as a whole requires the use of a large number of weighting factors. The advantage of using weights is that it is possible to easily tune the search to look for different kinds of similarity; the drawbacks are that the weighting scheme must be carefully optimized, and that it may be difficult to understand exactly why certain correspondences were chosen (or conversely, why particular sets of weightings function poorly or well).

In the second set of experiments, a combination of both unary and binary relations was used to evaluate matches. The unary relations were those described for the first set of experiments, plus the Euclidean distance between the center of mass of the element under consideration and the center of mass of the protein. In addition, two binary

Protein	Chain	PDB ID	Reference
Normal Human deoxyhaemoglobin	α	2HHB	Fermi 1975
Normal Human deoxyhaemoglobin	β	2HHB	Fermi 1975
Annelid worm carbomonoxyhaemoglobin		1HBG	Padlan & Love 1974
Seal myoglobin		1MBS	Scouloudi 1978 Scouloudi & Baker 1978

Table 1. Proteins used in this study.

relations were used: Euclidean distance between the centers of mass of pairs of elements, and the one-dimensional distance between two elements, the latter being defined as the number of residues between them along the protein chain. The search strategy was a modified form of a steepest-ascent. In an attempt to speed up matching, a small set of elements from one protein was selected first, after which search was carried out to identify corresponding elements in the other protein. To compare distances, a similarity score S was assigned by comparing the difference between the distances to a preselected cutoff:

$$S = \max\left(0, 1 - \frac{\text{abs}(d_1 - d_2)}{cutoff}\right) \qquad (1)$$

The third set of experiments used only one kind of relation, the binary relation Euclidean distance. Each element was described by three points, being two endpoints and the center of mass. A distance matrix was constructed for each protein in which the nine possible distances between each pair of structures was computed. The search began by considering as equivalent some pair of elements (in the comparisons presented here, the center of mass of the two proteins was used); it was continued using a steepest-ascent hill-climbing strategy. At each step, the pair of elements with the highest degree of spatial correspondence was added to the match. For the results presented here, only elements of identical type were ever considered for a correspondence. Discontiguous substructures where mapped first when present, then helices, strands, and finally coil regions. The correspondence between a pair of elements, was evaluated as follows. For each element, an ordered set of distance matrices was constructed, one matrix for each element of the same protein already in the match. Each corresponding pair of distance matrices was evaluated by first applying the distance comparison function described above to corresponding elements, then calculating the weighted average of all the elements. (Weighting allows the three different types of distances, namely endpoint–endpoint, endpoint–center, and center–center, to make differently valued contributions to the measure of similarity.) This gave a single set of scores which were averaged to calculate the overall similarity of the two elements in terms of their spatial relation to rest of the match. As before, the search continued until no more pairings were possible. An advantage of this approach is that only a very few parameters are required: a cutoff distance and two weighting factors.

Results and Discussion

A summary of the results is presented in Table 2. Sequence identity data is that determined by LFASTA (Pearson & Lipman 1988). For each comparison, the table shows the number of contiguous substructures which are correctly mapped over the total number that are mapped, together with the time taken to generate the mapping. Times are shown for comparison only; this implementation was written as a prototype, without great regard for efficiency. Of the three sets of experiments shown here, only one could be said to show good performance: the set which uses as a heuristic the similarity of the Euclidean distances between corresponding elements. In fact, this method never made more than one error per mapping, and always mapped the helices perfectly.

This result is interesting for two related reasons: it is by far the fastest method, and it uses very few distances. If many distances were being used, the accuracy would not be very surprising; a complete distance matrix includes all the spatial information in the structure, so using it in a heuristic would be likely to give good (though possibly slow) results. A complete distance matrix for all Cα atoms in a protein of this size (141 residues for the α-chain of 2HHB) includes 9870 distances. Because of the simplified representation employed, far fewer distances are used by this method—in the same chain, only 15 elements are represented, yielding 105 possible pairs, for a total of 945 distances (since three points are used to represent each structure). The total number of comparisons made is also reduced by using a divide and conquer strategy which compares only like elements. The greatest efficiency gain comes from using an incremental heuristic function; instead of evaluating an entire match at each step, an $\mathbf{O}(N^2)$ operation, only distance pairs involving the particular element under consideration are evaluated, an $\mathbf{O}(N)$ operation. In the comparison of α-chain of 2HHB with 1HBG, for example, a total of only 1736 comparisons were made between 9-element distance matrices in order to generate the final mapping.

In evaluating the other two methods, it is important to consider that no effort was made to refine the weighting functions used in the heuristics. Refinement should use a much larger number of number of examples than were used here, so as to avoid overfitting the data. Future plans include refining the weighting schemes using machine learning techniques; the approach has been designed to allow this. In spite of their poor performance here, it is likely that such methods will be useful in a more mature system. Similar methods have been shown to work well after appropriate weights are selected (Orengo 1992; Sali & Blundell 1990), and there are likely to be occasions when using only distances will not be appropriate—for example, attribute information may be valuable when mapping at the residue level within a conserved site or an element of secondary structure.

Comparison	Seq. Ident.	Attributes Only		Combined Approach		Distance Only	
		Score	Time	Score	Time	Score	Time
HHBa-HHBb	43.2%	6/8	24	13/14	132	14/14	7
HHBa-MBS	27.4%	2/8	26	3/15	130	15/15	8
HHBa-HBG	n/d	1/7	25	2/14	114	13/14	7
HHBb-MBS	27.6%	1/8	27	1/14	127	14/14	7
HHBb-HBG	19.6%	1/7	28	4/14	104	12/13	6
MBS-HBG	18.9%	1/7	28	0/14	112	13/14	7

Table 2. Results of comparing proteins. The left-hand column shows the proteins being compared in each row, identified by Protein Data Bank code (Bernstein et al. 1977), with initial numerals omitted, and chain specifiers added as appropriate. The second column shows the percentage sequence identity over the entire length of the two proteins, as calculated by LFASTA (Pearson & Lipman 1988). n/d: no detectable similarity over the entire sequence. The remaining columns show first the ratio of contiguous substructures which are correctly mapped by each approach (in the form of correct mappings/total mappings), and then the time taken for the mapping, in seconds. Comparisons were carried out on a Macintosh Quadra 660av using Macintosh Common Lisp 2.0.1.

By inspecting the progress of the mappings, it is possible to discern why some mistakes are made. The distance-only method makes mistakes when trying to add the last few elements, in cases where no further correspondences are appropriate. The progress of the mapping between HHBa and HBG is shown in Figure 1. Each protein includes an element for which no corresponding element exists in the other protein. In HBG, there is a single residue between the B and C helices, while the B and C helices in HHBa are adjacent along the chain. In HHBa, there are three non-helix residues between the H helix and the C-terminal, whereas the H helix of HBG includes the C-terminal residue of the chain. After the program has correctly mapped everything else, it proceeds to map these last two remaining elements to each other incorrectly. This incorrect mapping illustrates the more general problem of deciding when no correspondence is better than a given correspondence. The decision is complicated by the use of incremental evaluation functions; the efficiency which they provide comes at the cost of making it very difficult to compare the quality of two mappings that have proceeded to different states. It can be seen from Figure 1 that the evaluation score for the incorrect correspondence is much lower than any of the scores for correct correspondences. This indicates that one simple way to prevent this kind of error would be to only accept a correspondence with a score that exceeds some cutoff value. Alternatively, the program could watch for sharp drops in score and use that information to reject correspondences. Yet another approach would be to use global heuristic functions in addition to the more efficient incremental functions, calculating the former only when there is some indication that a difficult decision must be made.

Both the attributes-only method and the combined method create erroneous mappings early, after which they do not recover (data not shown). This is a reflection of the brittle nature of greedy mapping and steepest-ascent search. It is noteworthy that the distance-based method works as well as it does while using steepest-ascent search; this indicates that the distance heuristic is very robust.

Conclusion

A new approach to structure comparison has been presented, which allows the flexible combination of a variety of heuristic methods and search control strategies. Results of testing a prototype implementation indicate that the approach is feasible and capable of providing both speed and accuracy. Though much work remains to be done, it appears that the method will be capable of providing the versatility and speed required for the long-term goal, an intelligent assistant system for working with the three-dimensional structures of proteins.

Future Work. The tests presented here are only preliminary. Work is in progress to explore a broader range of search strategies and heuristics, and to investigate how they can be combined to optimize speed

2HHB		1HBG		Order	
Name	Range	Name	Range	Matched	Score
1	1- 2	1	1- 2	10	0.8027
A	3- 18	A	3- 18	1	0.9259
2	19- 19	2	19- 22	8	0.8641
B	20- 35	B	23- 37	3	0.8684
C	36- 42	C	39- 46	6	0.7619
3	43- 49	4	47- 52	9	0.8037
E	52- 71	E	53- 72	2	0.8881
4	72- 79	5	73- 75	13	0.6596
F	80- 88	F	76- 92	7	0.6981
5	89- 93	6	93- 99	12	0.7267
G	94-112	G	100-119	4	0.8328
6	113-117	7	120-123	11	0.7659
H	118-138	H	124-147	5	0.7876
7	139-141	3	38- 38	14	0.2717

Figure 1. Output from the program, showing the progress of mapping 2HHba against 1HBG. Each row shows a corresponding element-pair, the order in which the mapping was constructed, and the evaluation score for that step. Helices are named by letter (nomenclature after Lesk and Chothia (Lesk & Chothia 1980)) and non-helix regions by number. Range indicates the positions of the termini of each element in the primary sequence.

and accuracy. Mechanisms will be provided to allow one-to-many and one-to-none correspondences, since in real proteins, a single element in one protein may fulfill roles played by two or more different elements in a related protein. Other kinds of data will be included, such as curvature, solvent accessibility and relations to prosthetic groups, and the mapping will be extended to the level of individual residues. Secondary structure assignment will be standardized using DSSP (Kabsch & Sander 1983). Longer-term goals include the use of machine learning approaches to optimizing both search strategies and heuristic functions.

Acknowledgments

The author is grateful to B. C. Wang and Bruce Buchanan for their advice. Ethan Benatan is supported by a predoctoral fellowship from the Keck Center for Advanced Training in Computational Biology at the University of Pittsburgh, Carnegie Mellon and the Pittsburgh Supercomputing Center.

References

Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T. and Tasumi, M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology* 112: 535-542.

Chothia, C. 1992. One thousand families for the molecular biologist. *Nature* 357: 543-544.

Chothia, C. and Lesk, A. M. 1987. The evolution of protein structures. *Cold Spring Harb Symposium on Quantitative Biology* 52 (399): 399-405.

Falkenhainer, B.; Forbus, K. and Gentner, D. 1989. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence* 41 (1): 1-63.

Fermi, G. 1975. Three-dimensional fourier synthesis of human deoxyhaemoglobin at 2-5tA resolution: refinement of the atomic model. *Journal of Molecular Biology* 97: 237-256.

Gentner, D. 1989. The mechanisms of analogical learning. In Vosniadou, S. and Ortony, A., eds.: *Similarity and analogical reasoning*. London: Cambridge University Press.

Holm, L. and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology* 233: 123-138.

Kabsch, W. and Sander, C. 1983. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 22: 2577-2637.

Lesk, A. M. and Chothia, C. 1980. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *Journal of Molecular Biology* 136: 225-270.

Morris, A. L.; MacArthur, M. W.; Hutchinson, E. G. and Thornton, J. M. 1992. Stereochemical quality of protein structure coordinates. *Proteins* 12 (345): 345-64.

Orengo, C. A. 1992a. Fast Structure Alignment for Protein Databank Searching. *Proteins* 14: 139-167.

Orengo, C. A. 1992b. A Review of Methods For Protein Structure Comparison. In Taylor, W. R., ed.: *Patterns in Protein Sequence and Structure*. London: Springer-Verlag.

Padlan, E. A. and Love, W. E. 1974. Three-dimensional structure of hemoglobin from the polychaete annelid, Glycera dibranchiata, at 2.5 A resolution. *Journal of Biological Chemistry* 249: 4067-4078.

Pearson, W. R. and Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Science of the USA* 85: 2444-2448.

Rich, E. and Knight, K. 1991. *Artificial Intelligence*, 2nd edition. New York: McGraw-Hill.

Richardson, J. S. 1981. The anatomy and taxonomy of protein structure. *Advances Protein in Chemistry* 34: 167-339.

Sali, A. and Blundell, T. L. 1990. Definition of general topological equivalence in protein structures. *Journal of Molecular Biology* 212: 403-428.

Sali, A.; Overington, J. P.; Johnson, M. S. and Blundell, T. L. 1990. From comparisons of protein sequences and structures to protein modelling and design. *Trends in Biochemical Sciences* 15: 235-40.

Scouloudi, H. 1978. A preliminary comparison of metmyoglobin molecules from seal and sperm whale. *Journal of Molecular Biology* 126: 661-671.

Scouloudi, H. and Baker, E. N. 1978. X-ray crystallographic studies of seal myoglobin. The molecule at 2.5 A resolution. *Journal Molecular Biology* 126: 637-660.

Topham, C. M.; Thomas, P.; Overington, J. P.; Johnson, M. S.; Eisenmenger, F.; and Blundell T. L. 1990. An assessment of COMPOSER: a rule-based approach to modelling protein structure. *Biochemical Society Symposia* 57: 1-9.