Inf Syst Front (2006) 8: 29–36
DOI 10.1007/s10796-005-6101-9

Protein structure from contact maps: A case-based reasoning approach

Janice Glasgow · Tony Kuo · Jim Davies

Abstract Determining the three-dimensional structure of a protein is an important step in understanding biological function. Despite advances in experimental methods (crystallography and NMR) and protein structure prediction techniques, the gap between the number of known protein sequences and determined structures continues to grow.

Approaches to protein structure prediction vary from those that apply physical principles to those that consider known amino acid sequences and previously determined protein structures. In this paper we consider a two-step approach to structure prediction: (1) predict contacts between amino acids using sequence data; (2) predict protein structure using the predicted contact maps. Our focus is on the second step of this approach. In particular, we apply a case-based reasoning framework to determine the alignment of secondary structures based on previous experiences stored in a case base, along with detailed knowledge of the chemical and physical properties of proteins. Case-based reasoning is founded on the premise that similar problems have similar solutions. Our hypothesis is that we can use previously determined structures and their contact maps to predict the structure for novel proteins from their contact maps.

The paper presents an overview of contact maps along with the general principles behind our methodology of case-based reasoning. We discuss details of the implementation of our system and present empirical results using contact maps retrieved from the Protein Data Bank.

Keywords Case-based reasoning · Protein structure · Contact maps · Secondary structure · Analogy

Funding provided by: The Natural Science and Engineering Research Council (Ottawa); Institute for Robotics and Intelligent Systems (Ottawa); Protein Engineering Network Center of Excellence (Edmonton).

J. Glasgow (✉)· T. Kuo · J. Davies
School of Computing, Queen's University, Kingston Ontario K7L 3N6
e-mail: janice@cs.queensu.ca
Tel: 613 533-6058
Fax: 613 533-6513

1. Introduction

The accurate prediction of protein structure from sequence data is a fundamental problem in modern molecular biology. Despite current endevours in experimental (crystallographic and NMR) efforts, the determination of protein structure cannot keep pace with the increase in known sequences. Attempts to predict structure vary from those that apply physical principles to those that consider known amino acid sequences and protein structures. One approach to this problem is to first predict a contact map and structural features from a given protein sequence, and then to reconstruct the three-dimensional (3D) structure of the protein from its predicted contact map. This paper addresses the second step in this process by proposing an effective method for structure determination from contact maps using knowledge derived from the Protein Data Bank (PDB) (Berman et al., 2000), along with known and derived properties of protein structure and sequence. The method is hierarchical, in the sense that it considers protein contact maps at varying levels of structural complexity. In a bottom-up fashion, we initially construct secondary structure motifs using the contact map and geometric knowledge of α-helices and β-strands. Contacts between residues in pairs of secondary structures are used to

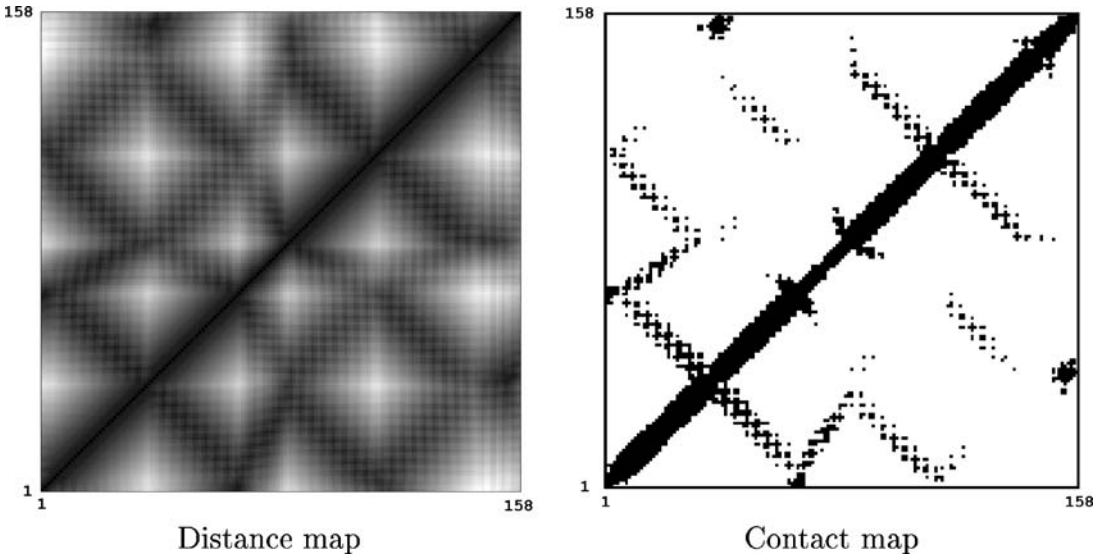

Fig. 1 Distance map and contact map for the protein Bacterioferritin (Cytochrome B1). The axes represent the residues of the protein starting from the N terminus (bottom left corner). In the distance map, darker colors correspond to closer distances. For the contact map, black areas correspond to values of 1, where residues are in contact (within 10 Å of one another)

predict the alignment for the pairs based on substructures in the PDB with similar contact maps. Similarly, we propose that super-secondary structure and tertiary structure alignments can be predicted based on structures retrieved from the PDB using contact maps at higher levels of the hierarchy.

To solve novel structures from contact maps we apply Case-Based Reasoning (CBR), which is founded on the premise that similar problems have similar solutions. The underlying hypothesis of our research, and our motivation for using CBR as a problem solving tool, is that proteins with similar contact maps tend to have similar 3D structures. Aaronson, Juergen, and Overton (1993) suggest that analogical reasoning is particularly applicable to the biological domain, partly because biological systems are often homologous (rooted in evolution). As well, biologists often use a form of reasoning similar to CBR, where experiments are designed and performed based on the similarity between features of a new system and those of known systems. CBR and/or analogical reasoning has previously been applied to a number of problems in molecular biology, including gene finding (Shavlik, 1991), prediction of unknown regulatory regions in genes (Aaronson, Juergen, and Overton, 1993), planning of sequence experiments (Kettler and Darden, 1993), secondary structure prediction (Leng, Buchanan, and Nicholas, 1993) and protein crystallization (Hennessy et al., 2000; Jurisica et al., 2001). An overview of these systems can be found in Jurisica and Glasgow (2004).

Our current focus is on predicting the alignment, or relative location, in 3D space of α-helices given the contacts between their residues. The paper begins with an overview of contact maps, followed by a description of the methodology used (the CBR paradigm), and its application to the problem of protein structure recovery from contact maps. Results from testing the technique using contact maps derived from structures in the PDB are presented.

2. Methods

2.1. Contact maps

A *distance map*, D, for a protein with n amino acid residues is an $n \times n$, symmetric array where entry $D(a_i, a_j)$ is the distance between residue a_i and residue a_j, generally calculated at the coordinates of the C_α atoms for the residues. Given a distance map D, we compute a *contact map* C for the protein as a symmetric, $n \times n$ array such that:

$$C(a_i, a_j) = \begin{cases} 1, & \text{if } D(a_i, a_j) < t; \\ 0, & \text{otherwise.} \end{cases}$$

where t is a given threshold value.[1] Thus, there exists a contact between residues a_i and a_j if and only if they are within a given distance t of one another in the protein structure. Figure 1 illustrates image representations for a distance map and a contact map reconstructed from the PDB entry for protein Bacterioferritin. The contact map is calculated using a distance threshold of 10 Å.

Secondary structures, the building blocks of protein structures, are easily recognizable in a contact map: α-helices ap-

[1] For the purpose of this paper, contact maps will be computed for known structures. Future work will involve contact maps that have been predicted from sequence information.

11	0	5	8	28	0	0	0	0	0	14	43
	0	0	3	2	0	3	0	20	5	36	14
	0	0	0	0	0	0	0	6	1	5	0
	0	6	0	33	1	88	9	33	6	20	0
	3	1	0	0	0	12	4	9	0	0	0
	3	64	0	4	33	30	12	88	0	3	0
	12	71	0	32	11	33	0	1	0	0	0
	7	13	9	28	32	4	0	33	0	2	28
	0	13	4	9	0	0	0	0	0	3	8
	11	345	13	13	71	64	1	6	0	0	5
1	14	11	0	7	12	3	3	0	0	0	0
	1										**11**

Fig. 2 Secondary structure contact map for the protein Bacterioferritin containing 11 secondary structures

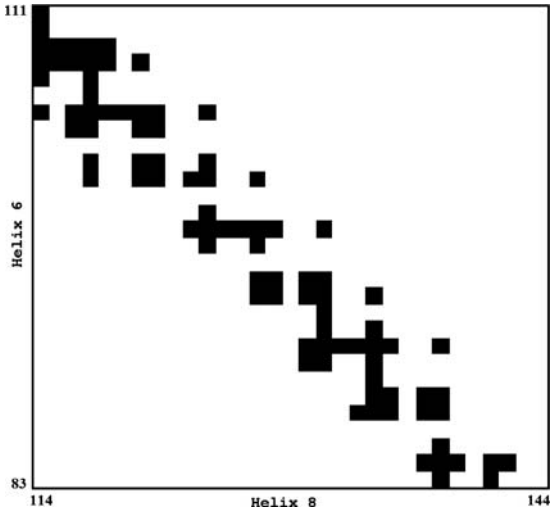

Fig. 3 The sub-contact $C_{\text{Helix}-6,\text{Helix}-8}$ map for a pair of helices in protein Bacterioferritin. Since the diagonal band shows contacts that extend from the beginning of helix 6 and end of helix 8, to the end of 6 and beginning of 8, we can discern that the helices are oriented anti-parallel to one another

pear as thick bands along the main diagonal; β-sheets appear as thin bands parallel and perpendicular to the main diagonal. A contact map can be viewed as a translational and rotational invariant representation of the protein's topology, capturing much of its relevant structural information. It provides a "fingerprint" that can be used to efficiently compare proteins to find ones with similar substructures.

Our approach incorporates an hierarchical search strategy that initially locates proteins that have similar secondary structures to our input protein. Given a protein p with j secondary structures (α-helices, β-sheets and coils), we define its *secondary structure contact map* as the $j \times j$ array S such that $S(s_m, s_n) = k$, where k is the number of contacts in map C between residues in secondary structure s_m and residues in secondary structure s_n for protein p. Figure 2 illustrates the secondary structure contact map corresponding to the contact map of Fig. 1.

To predict the alignment of sub-structures in 3D space, we consider contact maps, C_{s_m,s_n}, corresponding to pairs of secondary structures (s_m, s_n) such that $S(s_m, s_n) > 4$.[2] This map is the subarray of C such that the the rows of C_{s_m,s_n} correspond to the amino acid residues in secondary structure s_m and the columns correspond to the residues in secondary structure s_n. These maps need only be defined for contacts along and below the diagonal of the secondary structure contact map, as the map for pair (s_m, s_n) is equivalent to that for (s_n, s_m). Note, that unlike the protein contact map and the secondary structure contact map, the contact maps for pairs of helices are not generally symmetric. Figure 3 illustrates a contact map for a pair of α-helices.

2.2. Case-based reasoning

Artificial intelligence systems generally solve problems by reasoning from first principles. An alternative approach is to solve novel problems through analogy with old problems. CBR (Kolodner, 1993; Riesbeck and Schank, 1989) is a paradigm for analogical reasoning where experiences are represented as cases in a case base, then retrieved and reused during problem solving. A case represents knowledge about a particular problem solving experience and includes a problem description, a solution to the problem and (if available) feedback on the success of the solution. The case base is a repository of cases, designed to support the efficient storage and retrieval of a large number of complex cases. CBR is particularly useful in domains that are poorly understood or evolving, where knowledge is difficult to formalize.

As illustrated in Fig. 4, a CBR system consists of several components. Once a problem (a case without a solution) is submitted to the system, the first task is to *retrieve* previous experiences with similar problem descriptions. A similarity function, which is often complex and domain dependent, determines which cases are most relevant to the problem at hand. The case base may be organized, or indexed, so that only a subset of the relevant cases are considered, thus making the retrieval process more efficient. The results of retrieval are passed on to the *adapt* module. Here the solutions from the retrieved cases are modified to derive a potential solution for the new problem. Approaches to adaptation are mostly domain dependent. Machine learning or data mining techniques applied to the case base can be incorporated in the development of the adaptation module. Once a solution has been proposed, it is *evaluated* in terms of previous cases and/or domain knowledge. If the solution is not satisfactory, the system may return to the adaptation module for further modification. Once a solution is deemed satisfactory, it is applied and feedback (if available) is added to the case which is then *saved* in the case base.

[2] If there are fewer than five contacts between two secondary structures it is difficult to determine their orientation from their contacts.

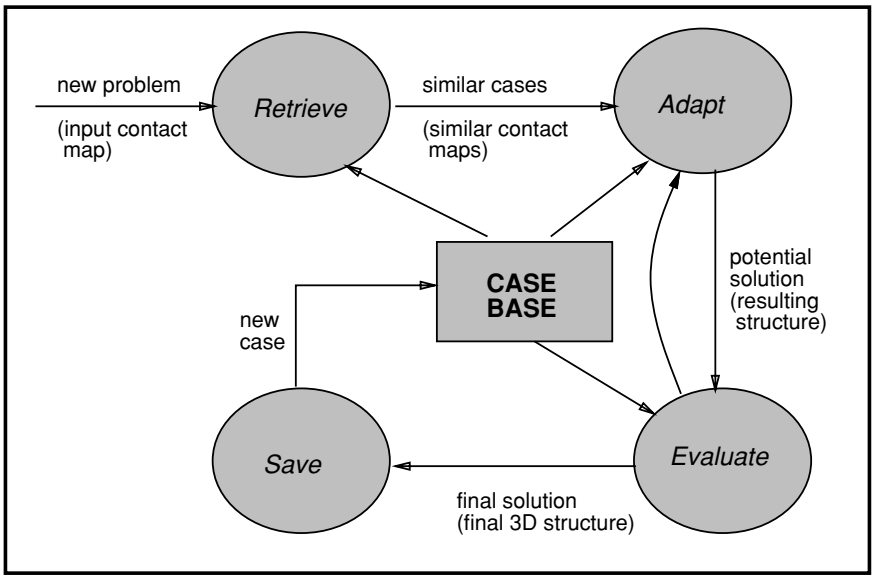

Fig. 4 Architecture of a CBR system for determining protein structure from contact maps

Our method consists of designing and implementing a CBR system that retrieves and adapts protein data from the PDB in order to construct potential 3D structural models for our input protein. All potential models are evaluated in terms of domain knowledge and the "best" structures will ultimately be used as building blocks at the next level of model building. The implementation is tested on the alignment of pairs of α helices using ideal contact maps (i.e., those computed using coordinates in the PDB).

2.3. Applying CBR to structure alignment

Below we describe the application of CBR to the problem of determining the 3D structure of a protein from a contact map. The approach incorporates a case representation that captures the contact between substructures of the protein at both the amino acid and the secondary structure levels. This allows for an efficient preliminary search of the case base to retrieve proteins that may have similar solutions, followed by a more detailed analysis of contacts between amino acids to adapt previous solutions to the new problem.

Case representation

Recall that a case has three parts: a problem description, a solution and feedback on the solution. The *problem description* - input to the system - consists of the following attributes and their corresponding values:

– protein name;
– protein sequence;
– assignment of secondary structure to residues;
– class of structure;
– contact map for protein.

Secondary structure maps and maps for pairs of secondary structures are computed using the protein contact map and secondary structure assignment.

The *solution* consists of a 3D backbone model of the protein structure computed for the input contact map (henceforth we will refer to this as the *target* map). The *feedback* (if available), consists of the correct structure for the protein and the calculated root mean square distance (RMSD) measure between the predicted structure and the correct structure. This distance provides a measure of "goodness" for the derived solution. Note that the values for attributes need not be explicitly stored in the case base. Rather, these values may be pointers to the location of the information (e.g., a pointer to the structure in the PDB) or procedures that allow for the computation of a value on an "as needed" basis (as in the contact maps for known structures).

The solution for a novel target case is a protein structure predicted from its contact map using a step-wise, hierarchical approach:

1. For each target map $C_{(s_m,s_n)}$ that contains more than four contacts, use CBR to determine an optimal alignment of the two secondary structures using experience embodied in the PDB.
2. Using the aligned pairs of secondary structures as building blocks, super-secondary and tertiary structures can be constructed by once again using a CBR approach.

This paper focuses on the first step of the procedure. In particular, we construct the individual α-helices for a protein then retrieve contact maps with known structures similar to the target maps and adapt the structures to predict alignments for the unknown structures.

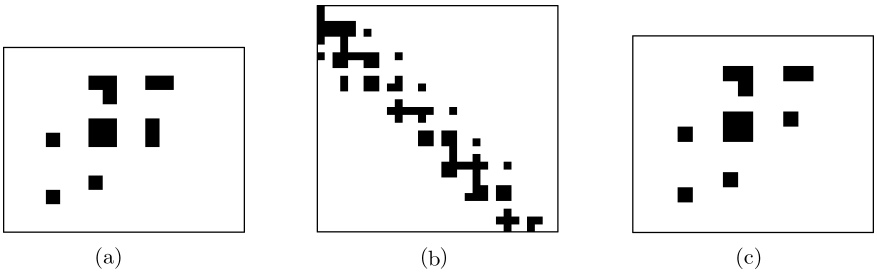

Fig. 5 Illustration of similar, (a) and (c), contact maps and a map (b) that is dissimilar to the other two

Case retrieval

Cases are organized (indexed) in the case base by class of structure: α domains, β domains and α/β domains. When initiating a retrieval, only cases that match the class of the input protein are considered. For the purpose of this paper we considered proteins in the α domain.

For each query map C_{s_m,s_n} such that $m \neq n$, we retrieve proteins that contain substructures (pairs of secondary structures) with contact maps most similar to C_{s_m,s_n}.

A similarity measure for comparing the query contact map with maps generated from structures in the PDB was derived using techniques from machine vision, where we consider the black regions to be the image within the array. We were less concerned about the dimensions of the map, than what it looked like in terms of shape and location of black regions (regions which contain contacts). For example, Fig. 5 illustrates three different maps for pairs of helices, where maps (a) and (c) are considered similar to one another, and (b) is different from the other two.

The retrieval of similar contact maps involves a two-tiered approach. Given a query contact map, the first tier uses general content descriptors to cull the dataset of dissimilar contact maps. These descriptors are: quadtrees (Sullivan and Baker, 1994), color and edge distributions (Smith and Chang, 1994; Won, Park, and park, 2002) and gray-level co-occurrence matrices (Haralick, Shanmugam, and Dinstein, 1973). A committee of these general content descriptors is used in the first tier of retrieval. The committee results in a set of contact maps which are present in the retrievals of two or more general content descriptors. It was determined empirically that 100 retrievals for each descriptor is sufficient. The results of the committee are then used in the second tier of retrieval.

For the second tier, the Jaccard's distance (Jaccard, 1908) was calculated between each contact map from the first tier and the query map. Because the maps vary in size, a sliding window approach was used to determine the best matching regions between the query and the contact maps from the first tier. The best mapping regions also provide registration of residues for evaluation using RMSD. The best 25 retrievals were then selected from the 100 as the final set of contact maps to be returned.

Adaptation

The retrieval process returns, for each query contact map, potential helix pairs from the PDB ranked in order of estimated similarity. For each query map, the adaptation phase of CBR transfers the structure information from the highest-ranking structures to the the input case.

Transferring locations requires a mapping function—that is, a set of alignments that determine which residues in the *target structure*[3] map to which residues in the retrieved structure. This is achieved by first aligning the contact maps so that the mean cell location of contacting amino acid residues in the retrieved structure aligns with the mean cell location of contacting residues in the target. Then all amino acid residues in the target structure that have corresponding residues in the known structure are given the coordinate information from these residues. Usually there remain some target residues with no coordinates (i.e., no corresponding residue in the known structure). Since α-helices tend to have a consistent structure, the missing coordinates are filled in using general domain knowledge. Specifically, each turn of an α-helix is estimated at 5.4 Å along the helix axis and each turn at 5 Å across. Using this information and the helix axis, calculated from the filled-in locations, our system is able to infer these unmatched residue locations. Figure 6 illustrates the portions of the helices that are determined through our mapping function and those constructed from domain knowledge (grown area).

Evaluation of potential structures

The adaptation component of our CBR system outputs multiple possible substructures of helix pairs. In the evaluate module, we wish to rank the potential structures using multiple sources of knowledge and expertise. One question we are faced with is how to integrate these diverse knowledge sources. This question is addressed by incorporating an architecture that will allow us to discard any of the structures that are infeasible (based on physical or chemical constraints) and determine which of the remaining structures is most likely to

[3] We use the term "target structure" to denote the predicted substructure for the current query contact map.

be closest to the correct structure. We apply *FORR* (*FO*rr the *R*ight *R*easons) (Epstein, 1994), a cognitive architecture for learning and problem solving by consensus among heuristic rationales, to integrate our multiple sources of knowledge.[4]

Each rationale in a *FORR*-based system is implemented as a resource-limited procedure called an *Advisor*. Examples of Advisors that are currently being implemented for our system are:

- The *side chain Advisor* examines pairs of residues that are in contact in the model and determines, given their possible side chain configurations, whether the predicted locations are feasible.
- The *contact map Advisor* compares the contact map of the predicted model with the contact map for the problem description.

We anticipate that the final system will have between 20 and 30 expert advisors that will participate in the evaluation process.

Each Advisor comments on (assigns a value to) a potential problem solution. The ultimate decision of the system is based on a weighted sum of these individual comments.[5] The *strength* of a comment, which is an integer value in the range [−1..1], denotes the Advisors' opinions about the particular structure. Positive strengths represent a degree of support, negative strengths a degree of opposition, and a strength of 0 denotes indifference to the structure.

[4] The FORR system has been successfully applied to the development of problem solving systems for the domain of path finding in grid-world mazes (Epstein, 1998) and for the domain of finite-board games (Epstein, Gelfand, and Lock, 1998). Similar to our molecular domain, these previous applications involve spatial reasoning and rely on multiple (possibly conflicting) sources of expertise.

[5] Advisors may vary in importance and trustworthiness; this is reflected in the weights assigned to them.

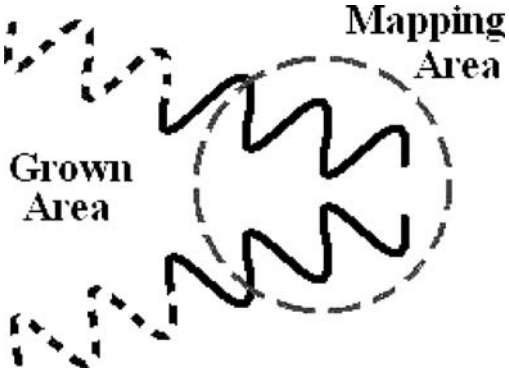

Fig. 6 The target structure for a given query contact map is predicted using the mapping area for the retrieved structure and extending the helices (grown area) based on the known geometry for helices

A voting procedure assesses a structure m using a weighted tally of the strengths of the Advisors' comments:

$$\upsilon(m) = \sum_{i=1}^{n} w(A_i) \times s(A_i, m) \quad (1)$$

where A_i is an Advisor, $s(A_i, m)$ is the strength of A_i's comment for m, and $w(A_i)$ is the weight assigned to Advisor A_i. Our approach incorporates a neural net learning algorithm to determine optimal weights for Advisors.

The evaluation process depends on gathering knowledge and data about what constitutes a "good" structure. This is being achieved through: (1) statistical and machine learning analysis of existing databases of protein structures; (2) acquisition of expert knowledge (accessed through interviews and research papers); (3) assimilation of text-book knowledge; and (4) incorporation of existing algorithms for protein structure evaluation.

The ability to evaluate the quality of a protein model is fundamental to the determination of structure from sequence. This is true whether we are deriving a structure using structure prediction or experimental structure determination methods. The quality of a protein model is based on whether it adheres to the known principles of chemistry, biology and physics, and whether it is consistent with the information available from the primary sequence, the experimental data and the previously determined structures. An evaluation process may involve considering a single model or comparing multiple competing models. Research in model evaluation for experimentally derived structures has previously focussed on verifying that the final protein model is correct (Luthy, Bowie, and Eisenber, 1992).[6] Kleywegt and Jones (1997) have proposed some quality control criteria for the assessment of intermediate protein models. The tools they and others suggest generally assume a single complete model, which is evaluated to determine what parts of the structure need to be revised or rebuilt. Our system, however, evaluates a set of partial models to identify the most promising.

Results

The retrieval and adaptation components of the CBR system were applied to to a set of 61 proteins, mostly all α chains, retrieved from the PDB:

1a0aA, 1a1z_, 1a28A, 1acp_, 1afrA, 1aj8A, 1akhA, 1akhB, 1am9A, 1aoiA,

[6] Even with techniques for evaluating the final protein model, incorrect models have been published and entered into the protein database.

Table 1 The results of the committee on 422 unique queries when the top N out of 100 are returned as the final set of contact maps

N	Mean	Std	Mean best	Rank
100	1.8604	0.8035	0.5259	7.5
50	1.6498	0.6447	0.5303	7
25	1.3944	0.5077	0.5506	5
10	1.1919	0.4166	0.6034	3

1aoiB, 1arv_, 1auiB, 1auwA, 1bbhA, 1bcfA, 1bgp_, 1bh9A, 1bh9B, 1bu7A,

1bvb_, 1c52_, 1cc5_, 1cem_, 1cktA, 1cll_, 1cpq_, 1csh_, 1cy5A, 1d9cA,

1dceB, 1dpsA, 1ea1A, 1eerA, 1eteA, 1fce_, 1fgjA, 1ft1B, 1furA, 1gakA,

1hcrA, 1hnr_, 1hryA, 1huuA, 1hyp_, 1kx2A, 1lbd_, 1lfb_, 1lis_, 1lmb3,

1mhyD, 1neq_, 1pbwA, 1pru_, 1rzl_, 1tc3C, 1tx4A, 1uxc_, 2af8_, 2hddA, 2ilk_.

For each protein, we computed the distance map, contact map and secondary structure contact map. From the contact maps, we were able to derive 422 maps that described contacts for pairs of helices.

The results of the retrieval process for 422 unique test queries are shown in Table 1. N is the number of cases retrieved; *Mean* describes the average RMSD for the queries and *Std* is the average standard deviation. *Mean Best* and *Rank* describe the average best RMSD and its median rank within the final set of contact maps. The results suggest the following:

- As N, the number of retrieved cases, decreases the average RMSD of the final set of contact maps improves.
- The *Mean Best* represents the best structure match and worsens as N decreases.
- As N increases from 25 to 50 to 100, the *Mean Best* does not change significantly.

Further examination of the 100 retrievals using the committee determined that 65.40% of the 422 queries have its best RMSD fall within the top 10 retrievals, 83.18% within the top 25 and 96.45% within the top 50. Thus, a final set of contact maps consisting of the top 25 retrievals from a set of 100 seems to be the best balance between a low average RMSD over all the retrievals and a low RMSD for the average best retrieval. This ensures all the retrievals are similar to the query and contains the best match in ∼83% of the cases.

Table 2 Experimental results when considering the top n results. RMSD denotes the mean of the best scores for each of the 422 input cases for the top n retrievals

n	RMSD
1	3.6668
5	2.2667
10	1.8814
25	1.5286
50	1.3921
100	1.3011
200	1.2507
422(all)	1.2426

Using the results of the retrieval module, we evaluated the adaptation method by comparing the *predicted* locations of the residues to the *actual* locations, as given in the Protein Data Bank (PDB) in terms of RSMD. The results when considering the top N retrievals, for N = 1, 5, 10 25, 50, 100, 200, and 422 are presented Table 2. These results suggest that we converge to a good solution when considering about the top 50 solutions.

Note that the retrieval scores for the *Mean Best* (in terms of RMSD between the correct and predicted structures) are less than the adaptation scores (which reported the distance between the retrieved structures and the correct structure). The reason for this is that the retrieval scores are based on the RMSD of only the regions of the helices in contact with each other. The adaptation method extends the helices beyond the regions of contact based on biochemical knowledge, affording more opportunity for error.

3. Discussion

We have described and demonstrated the applicability of the CBR methodology to the problem of secondary structure alignment from contact maps. Initial results suggest that the retrieve and adapt phases are successful in finding similar contact maps in the PDB and modifying these to predict the alignment of pairs of helices. The advantage and novelty of our approach lies in its use of multiple sources of knowledge, including existing structural knowledge from the PDB, expert and text book knowledge, as well as knowledge mined from the database.

Our initial results considered contact maps computed from existing structures in the PDB. Various approaches have been considered for the process of predicting contact maps for a protein from its primary sequence and structural features; these are primarily based on neural network-based methods (Fariselli et al., 2001; Pollastri and Baldi, 2002). Punta and Rost (2005) propose a contact prediction method that combines alignment information, secondary structure predictions and solvent accessibility. While results from these studies

are encouraging, they still result in maps that contain a large degree of noise. This suggests that in the second phase of prediction, we must be able to recover structure from such noisy maps. Future work will include prediction of structure from predicted contact maps.

Previous methods for the recovery of 3D structure from distance contact maps are mostly based on distance geometry and stochastic optimization techniques. Nilges, Clore, and Gronenborn (1988) applied distance maps and dynamical simulated annealing to determine the 3D structure of proteins. More recently Vendruscolo, Kussell, and Domany (1997) proposed a dynamic approach that generates a structure that has a contact map similar to the query contact map.

References

Aaronson JS, Juergen H, Overton GC. Knowledge discovery in genbank. In: Hunter L, Searls D, and Shavlik J, eds. In: *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, 1993;3–11.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, SHindyalov IN, Bourne PE. Protein data bank. *Nucleic Acids Research* 2000;28:235–242.

Epstein S. For the right reasons: The FORR architecture for learning in a skill domain. *Cognitive Science* 1994;18(3):479–511.

Epstein S. Pragmatic navigation: Reactivity, heuristics and search. *Artificial Intelligence* 1998;100:275–322.

Epstein S, Gelfand J, Lock E. Learning game-specific spatially oriented heuristics. *Constraints: An International Journal* 1998;2:239–251.

Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering* 2001;14(11):835–843.

Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics* SMC-1973;3(6):610–621.

Hennessy D, Buchanan B, Subramanian D, Wilkosz PA, Rosenberg JM. Statistical methods for the objective design of screening procedures for macromolecular crystallization. *Acta Crystallogr D Biol Crystallogr* 2000;56(Pt 7):817–827.

Jaccard P. Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles* 1908;44:223–270.

Jurisica I, Glasgow J. Applications of case-based reasoning in molecular biology. *AI Magazine* Winter, 2004.

Jurisica I, Rogers P, Glasgow J, Fortier S, Luft J, Wolfley J, Bianca M, Weeks D, DeTitta GT. Intelligent decision support for protein crystal growth. *IBM Systems Journal, Special Issue on Deep Computing for Life Sciences* 2001;40(2):394–409.

Kettler B, Darden L. Protein sequencing experiment planning using analogy. In: *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology,* 1993; 216–224.

Kleywegt GJ, Jones TA. Model-building and refinement practice. In: *Methods in Enzymology* 1997;(277):208–230.

Kolodner J. *Case-Based Reasoning.* Morgan Kaufmann, 1993.

Leng B, Buchanan BG, Nicholas HB. Protein secondary structure prediction using two-level case-based reasoning. In: *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology,* 1993;251–259.

Luthy R, Bowie JU, Eisenber D. Assessment of protein models with three-dimensional profiles. *Nature* 1992;356:83–85.

Nilges M, Clore GM, Gronenborn AM. Determination of the three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. *FEBS Lett.* 1988;229:129–136.

Pollastri G, Baldi P. Prediction of contact maps by recurrent neural network architectures and hidden context propagation from all four cardinal corners. *Bioinformatics* 2002;1(1):1–9.

Punta M, Rost B. Profcon: Novel prediction of long-range contacts. *Bioinformatics* 2005;21(13):2960–2968.

Riesbeck C, Schank R. *Inside Case-Based Reasoning*. Lawrence Erlbaum: Hillsdale, NJ, 1989.

Shavlik J. Finding genes by case-based reasoning in the presence of noisy case boundaries. In: *Proceedings of the 1991 DARPA Workshop on Case-Based Reasoning* Morgan-Kauffman, 1991.

Smith JR, Chang SF. Quad-tree segmentation for texture-based image query. *Proceedings of the second ACM international conference on = Multimedia* 1994;279–286.

Sullivan GJ, Baker RL. Efficient quadtree coding of images and video. *IEEE Transactions on Image Processing* 1994;3(3):327–331.

Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. *Folding and Design* 1997;2:295–306.

Won CS, Park DK, Park SJ. Efficient use of mpeg-7 edge histogram descriptor. *Electronics and Telecommunications Research Institute Journal* 2002;24:23.

Janice Glasgow is a Professor in the School of Computing at Queen's University, Canada, where she holds a Research Chair in Biomedical Computing. Her research in artificial intelligence involves theoretical work in computational imagery and case-based reasoning, and applications in protein structure determination and computational neuroscience.

Tony Kuo is a Masters student in the School of Computing at Queen's University. His research interests includes protein structure prediction, gene expression data mining and other areas of bioinformatics.

Jim Davies is a postdoctoral fellow at Queen's University's School of Computing. He received his Ph.D. in computer science with a focus on artificial intelligence from the Georgia Institute of Technology. He is a congnitive scientist interested in how visual knowledge can be used in analogical problem solving applied to bioinformatics.