Visio-spatial Case-Based Reasoning: A Case Study in Prediction of Protein Structure

Jim Davies   Janice Glasgow   Tony Kuo

School of Computing, Queen's University

Kingston, Ontario K7L 3N6 Canada
jim@jimdavies.org, janice@cs.queensu.ca
613 533-6058 (phone); 613 533-6513 (fax)

Submitted to *Computational Intelligence* Oct. 2005

Running Title: Protein Structure Prediction With Visual CBR

Abstract

We show that visio-spatial representations and reasoning can be used as a similarity metric for case-based protein structure prediction. Our system retrieves pairs of α-helices based on contact map similarity, then transfers and adapts the structure information to an unknown helix pair. We show that similar protein contact maps predict similar 3D protein structure. The success of this method provides support for the notion that changing representations can enable similarity metrics in case-based reasoning.

Key words: case-based reasoning, protein structure, analogy, bioinformatics, computational biology.

1 Introduction

It is well known that the right representation greatly facilitates reasoning [Amarel, 1968] and there is a growing recognition of the need for intelligent architectures to accomodate a diversity of representations [McCarthy et al., 2002].

The guiding theory of our research is that changing representations allows reasoners to see similarities in one representation type that might be difficult to detect in another. For example teleological representations of a human face and the front of a car may have very little semantic overlap. In this research we focus on visio-spatial representations. In our example representing the headlights and eyes as circles, and the grill and mouth as a centrally-located hole allows connections to be drawn between these components.

As people often have visio-spatial experiences when solving problems [Casakin and Goldschmidt, 1999, Farah, 1988, Monaghan and Clement, 1999, Shepard and Cooper, 1988], an important step in establishing our above theory is to show that one function of visio-spatial representations is that they can be used to solve a variety of problems. In this paper we provide support for this notion in the domain of protein structure prediction. We will describe the problem, and then how we use visio-spatial reasoning on images to solve it.

1.1 Protein Structure Prediction

A primary goal of molecular biology is to understand the biological processes of macromolecules in terms of their physical properties and chemical structure. Since knowing the structure of macromolecules is crucial to understanding their functions, and all life crucially depends on protein function [Hunter, 2004], an important part of molecular biology is understanding the three-dimensional (3D) structure of proteins.

Proteins are composed of one or more chains of amino acid residues. The description of which residues appear and in what order is the protein's "primary structure". According to the laws of chemistry, the chains twist, fold, and bond at different points, forming a complex 3D shape. Subchains form regular "secondary structures", the two main types being α-helices and β-strands. The overall protein shape (which may involve several chains) is known as its "quaternary structure". A major unsolved problem for the biological sciences is to be able to reliably predict the quaternary structure from the primary. This, at the highest level, is our problem domain.

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. Many of the latter use what is known as "homology" as a similarity metric. In this context homology is the similarity of two amino acid sequences. Our work also falls in the latter category, but rather than using primary structure directly, we compare contact maps.

1.1.1 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 (carbon-alpha) 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 (in our work this theshold is $10\mathring{A}$). 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 Protein Data Bank (PDB) [Berman et al., 2000].

In our work we use idealized contact maps. That is, we generate distance maps and contact maps from the actual 3D structure from the PDB of our target proteins. We wish to show that our method can work with idealized contact maps before we start to work with predicted contact maps. Researchers have considered various approaches 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 [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. Thus we carry out our initial experiments on idealized maps generated from the PDB. Future work will include prediction of structure from predicted contact maps.

A contact map is a translational and rotational invariant, visio-spatial representation that captures some of the protein's relevant structural information. Our general hypothesis is that visual processing on contact maps enables effective retrieval of similar structures, even if homology sequence is ignored. Contact maps provide a "fingerprint" that can be used to efficiently compare proteins to find ones with similar substructures. We will refine this hypothesis when we describe our implementation.

2 Overview of the project

In this section we will describe the plan for our entire project. In the next section we will describe the implemented modules.

At the highest level, each time the system runs it takes as input 1) the contact map for the unknown (target) protein, and 2) a case library of known protein structures and contact maps. The final output consists of a location in space (*x, y, z* coordinates) of each amino acid residue in the target protein.

2.1 CBR Applied to Protein Structure Prediction

The project applies case-based reasoning (CBR) at many levels of abstraction. The secondary and super-secondary structures are identified. CBR is used to infer the coordinates for secondary structures, then for their connections to form super-secondary structures.

CBR [Kolodner, 1993, Riesbeck and Schank, 1989] is founded on the premise that similar problems have similar solutions. It is a paradigm for analogical reasoning where experiences are represented as cases in a case base, then retrieved and reused during problem solving.

Aaronson et al. [Aaronson et al., 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; an overview of these systems can be found in [Jurisica and Glasgow,].

Our system retrieves and adapts protein data from the PDB in order to construct potential 3D structural models for our target protein. These 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.

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 Figure 1.

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 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.

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.

The solution, for a novel target problem, 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.

The implementation focuses on the first step of the procedure. In particular, we retrieve similar α-helix pair contact maps and adapt the known structures to predict alignments for the unknown structures.

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 $S(s_m, s_n) > 4$.[1] 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.

The *retrieve* task returns a list of retrieved cases, ordered according to similarity. The similarity metric is a visual similarity between source and target contact maps. The *adapt* module transfers structure information from the top retrievals (called the "sources") and modifies the information according the the specifics of the target. Since these are the modules that have been implemented, we will describe them in detail in the next section.

Case representation

Our cases have 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: 1) protein name, 2) primary sequence, 3) assignment of secondary structure to residues, 4) class of structure, and 5) the protein's contact map.

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

The *solution*, for the target problem, consists of the predicted structure of the protein (the *x, y, z* coordinates for each residue). The *feedback* (if available), consists of the correct structure for the protein and the calculated Root Mean Square Distance (RMSD) measure between the predicted

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

structure and the correct structure. This distance provides a measure of "goodness" for the derived solution.

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 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.[2]

Each rationale in a *FORR*-based system is implemented as a resource-limited procedure called an *Advisor*. Some of the Advisors we are currently implementing for our system are The *side chain Advisor*, which 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, and the *contact map Advisor*, which 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 (assigns a value) to a potential problem solution. The ultimate decision of the system is based on a weighted sum of these individual comments. For more details of how this system will work see Glasgow et. al. (in press).

[2] 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 et al., 1998]. Similar to our molecular domain, these previous applications involve spatial reasoning and rely on multiple (possibly conflicting) sources of expertise.

3 Implemented Modules: Retrieval and Adaptation

Our current focus is on predicting the alignment, or relative location, in 3D space of α-helix pairs given the contacts between their residues.

Case retrieval Module

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} 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, Figure 4 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.

First we blur the images using Gaussian smoothing [Gonzalez and Woods, 1992]. This is often done to remove unwanted details and noise. Contacts are treated as black points, and points surrounding them are turned some shade of gray depending on their distance from the nearest contacts. The grayscale tone is determined by a Gaussian distribution where the contacts are the means.

The maps are then morphed a technique called *closing,* which removes low-valued points but keeps the rest of the image intact [Gonzalez and Woods, 1992].

The retrieval of similar contact maps involves a two-tiered approach. Given a query contact map, the first tier uses three general content descriptors to cull the dataset of dissimilar contact maps: quadtrees, color and edge distributions, and gray-level co-occurrence matrices.

Quadtrees have been successfully applied to image compression, comparison, and classification. The quadtree [Sullivan and Baker, 1994] is a hierarchical data structure used to represent images. For an image, a two-dimensional region is recursively decomposed into quadrants where each quadrant is a node in the quadtree.

Color distribution [Smith and Chang, 1994] is a common feature used in image retrieval. Pixel color values are put into a histogram form: colors are discretized and counted and placed in bins. Global histogram representation has the drawback of loss of location, shape, and texture information. As a result images retrieved based on similar color distributions may not be semantically related.

Edge detection [Won et al., 2002], and the features that can be extracted from it, is commonly used as a content descriptor of images. In this work we use the Canny edge detection [Canny, 1986] method. The Gaussian smoothing was necessary for this step to work, as it uses gradients and cannot be applied to binary images. Our measure of similarity based on edge detection involves comparing histograms showing the frequency of edges with angles of 0^o, 45^o, 90^o, and 135^o.

A statistical mathod that considers the spatial relationshp of pixels, the gray-level co-occurrence matrix (GLCM) [Haralick et al., 1973] is a texture analysis method from which various statistical features can be extracted. Each entry *(i, j)* in the GLCM corresponds to the number of occurrences of the pair of gray levels i and j which are a distance d apart in the original image. For example, if d is 1, then GLCM entry (1, 2) will contain the number 4 if there are four instances of gray value 1 adjacent to gray value 2 in the original image. In analysis, the GLCM are normalized so the histogram or features extracted can be compared.

A committee of these general content descriptors is used in the first tier of retrieval. Quadtrees vectors were generated from the binary, smoothed, and morphed contact maps. The color and edge distributions and gray level co-occurrence matrices were obtained from the smoothed contact maps. The committee results in a set of contact maps which are present in the retrievals of two or more general content descriptors. We 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 Module

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 input case.

Transferring locations requires a mapping function – that is, a set of alignments that determine which residues in the target structure map to which residues in the retrieved source 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 source 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 5 illustrates the portions of the helices that are determined through our mapping function and those constructed from domain knowledge (grown area).

Given this implementation and our overall hypothesis, our refined hypothesis is that CRB using contact map similarity can effectively generate accurate protein substructure predictions.

Results

We applied the retrieval and adaptation components of the CBR system to a set of 61 proteins, mostly all α chains, retrieved from the PDB.[3]

[3] the proteins were 1a0aA, 1a1z_, 1a28A, 1acp_, 1afrA, 1aj8A, 1akhA, 1akhB, 1am9A, 1aoiA,
1aoiB, 1arv_, 1auiB, 1auwA, 1bbhA, 1bcfA, 1bgp_, 1bh9A, 1bh9B, 1bu7A,

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.

N	$Mean$	Std	$MeanBest$	$Rank$
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

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

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: 1) as N, the number of retrieved cases, decreases the average RMSD of the final set of contact maps improves, 2) the *Mean Best* represents the best structure match and worsens as N decreases, and 3) 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 $\sim$ 83% of the cases.

Using the results of the retrievals 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

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_.

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

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

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 distance 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.

4 Related Work in CBR

The issue of visual knowledge in case-based reasoning has attracted the attention in of researchers in several areas. Below we relate our work to some representative case-based problem solving systems with emphasis on systems that use visio-spatial knowledge.

FABEL [Gebhardt et al., 1997] is an example of a case-based system that adapts diagrammatic cases in the domain of architectural design. In FABEL, the source diagram specifies the spatial layout of a building or similar structure. FABEL adapts source diagrams by extracting and transferring specific structural patterns to the target problem. It uses domain-specific heuristics to guide pattern extraction and transfer.

REBUILDER [Gomes et al., 2003] is a case-based reasoner that does retrieval, mapping, and transfer of software design class diagrams. The diagrams are represented structurally, not visio-spatially, however. This means that, for example, what that the connection is between two nodes is more important than the length and direction of that connection. That is, REBUILDER works with a different level of visual abstraction, a level at which only the structural relationships, such as top-ofconnectedness, between visual elements are relevant to the task. Determining the right level of visual abstraction for visual case-based problems requires additional research. The choices made by REBUILDER depend largely on the specific domains in which they operate. In REBUILDER's domain of software design class diagrams, only the structural relations appear to be important.

FAMING [Faltings and Sun, 1996] is a case-based reasoning system that uses cases describing physical mechanism parts. FAMING uses the SBF (Structure-Behavior-Function) ontology to describe the cases. The structure is described in terms of a metric diagram (a geometric model of vertices and connecting edges), a place vocabulary (a complete model of all possible qualitative behaviors of the device), and configuration spaces (a compact representation of the constraints on the part motions). Shape features can involve two objects, expressing, for example, one part's ability to touch another part. Human designers are necessary for FAMING's processing. The designer chooses which cases and functions should be used, which dimensions the system should attempt to modify, and which shape features should be unified. It uses qualitative kinematics to propose design solutions for the desired function following the designer-suggested idea. Though not described as a visio-spatial system, the important parts of physical mechanisms of the sort FAMING uses inevitably contain much knowledge that could be construed as spatial or visual.

DIVA [Croft and Thagard, 2002] is an analogical mapper that uses visio-spatial representations, using the Java Visual Object System. It does no transfer of problem solutions and uses the ACME architecture for mapping [Holyoak and Thagard, 1997].

Non-visual case-based problem-solving systems, such as CHEF [Hammond, 1990] and PRODIGY [Veloso, 1993] provide interesting points of comparison regarding the transfer process. CHEF is a case-based reasoner that transfers and adapts cooking recipes from a source to a target. The Prodigy case-based reasoning system implements the theory of Derivational Analogy [Veloso, 1993]. It models transfer using memories of the justifications of each step, allowing for adaptation of the transferred procedure. Traces, called "derivations", are scripts of the steps of problem solving, along with the justifications for why the steps were chosen over others.

The Galatea system [Davies and Goel, 2001] uses only visio-spatial representations of problem-solving procedures and transfers a source solution to a target solution. By using a sufficiently abstract visual language (Covlan) it is able to transfer problem-solving procedures between semantically distant analogs. The work on Galatea also supports the notion that visio-spatial representations are useful for problem-solving.

Previous visual CBR work in molecular biology domains include visualizing crystallographic data at different resolutions [Glasgow et al., 1993, Glasgow et al., 1995, Hennessy et al., 2000, Jurisica et al., 2001a, Jurisica et al., 2001b], in drug design [Biname et al., 2004, Glasgow et al., 2004], and in in-vitro fertilization [Jurisica and Glasgow, 2000].

Non-visual bioinformatics CBR research includes a system that finds gene sequences that produce proteins [Shavlik, 1991], a predictor for unknown regulatory regions in genes [Aaronson et al., 1993], a planner for experiments for finding protein sequences [Kettler and Darden, 1993], the prediction of angles between amino acid residues in a protein chain [Zhang and Waltz,], and the prediction of secondary structure elements in a primary sequence [Leng et al., 1993].

5 Discussion

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

In this paper we described and demonstrated the applicability of the CBR methodology to the problem of secondary structure alignment from contact maps. Our hypothesis was that CRB using contact map similarity can effectively generate accurate protein substructure predictions. Our system retrieves protein substructures based on visual similarity of 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, supporting this hypothesis. 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.

Future work will include implementation of the other modules of our system. Once the viability of the approach is shown to be effective with idealized contact maps, the predicted, error-prone contact maps can used as input.

The theory behind this work is that changing representations can provide novel similarity insights. In this work we use contact maps and treat them as binary images, applying image processing techniques to them to retrieve similar protein substructures. This is in contrast with, for example, Jurisica et al. ([Jurisica et al., 2001a]), who retrieve based on generated attributes. In the adapt module, the information transferred is purely spatial. The success of this method for α-helix pair structure prediction provides preliminary support for this theory, in that generated visio-spatial representations can provide a means to find similarity. Future work will compare the results of contact map retrieval to sequence homology retrieval to investigate in exactly which conditions contact map similarity (representing visio-spatial representations) is superior to the non-visual homology similarity metric.

6 Acknowledgements

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).

References

[Aaronson et al., 1993] Aaronson, J. S., Juergen, H., and Overton, G. C. (1993). Knowledge discovery in genbank. In Hunter, L., Searls, D., and Shavlik, J., editors, *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, pages 3–11. AAAI Press.

[Amarel, 1968] Amarel, S. (1968). On representations of problems of reasoning about actions. In Michie, D., editor, *Machine Intelligence 3*, volume 3, pages 131–171. Elsevier/North-Holland, Amsterdam, London, New York.

[Berman et al., 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). Protein data bank. *Nucleic Acids Research*, 28:235–242.

[Biname et al., 2004] Biname, J., Meurice, N., Leherte, L., Glasgow, J., Fortier, S., and Vercauteren, D. P. (2004). Use of electron density critical points as chemical function-based reduced representations of pharmacological ligands. *Journal of Chemical Information and Computer Science*, 44:1394–1401.

[Canny, 1986] Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Alanysis and Machine Intelligence*, 8(6):769–798.

[Casakin and Goldschmidt, 1999] Casakin, H. and Goldschmidt, G. (1999). Expertise and the use of visual analogy: Implications for design education. *Design Studies*, 20:153–175.

[Croft and Thagard, 2002] Croft, D. and Thagard, P. (2002). Dynamic imagery: A computational model of motion and visual analogy. In Magnani, L. and Nersessian, N. J., editors, *Model-Based Reasoning: Science, Technology, & Values*, pages 259–274. Kluwer Academic: Plenum Publishers, New York.

[Davies and Goel, 2001] Davies, J. and Goel, A. K. (2001). Visual analogy in problem solving. In Nebel, B., editor, *Proceedings of the International Joint Conference for Artificial Intelligence 2001*, pages 377–382, Seattle, WA. Morgan Kaufmann Publishers. First published Galatea paper.

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

[Epstein, 1998] Epstein, S. (1998). Pragmatic navigation: Reactivity, heuristics and search. *Artificial Intelligence*, 100:275–322.

[Epstein et al., 1998] Epstein, S., Gelfand, J., and Lock, E. (1998). Learning game-specific spatially oriented heuristics. *Constraints: An International Journal*, 2:239–251.

[Faltings and Sun, 1996] Faltings, B. and Sun, K. (1996). FAMING: supporting innovative mechanism shape design. *Computer-aided Design*, 28(3):207–216.

[Farah, 1988] Farah, M. J. (1988). The neuropsychology of mental imagery: Converging evidence from brain-damaged and normal subjects. In Stiles-Davis, J., Kritchevsky, M., and Bellugi, U., editors, *Spatial Cognition–Brain bases and development*, pages 33–59. Erlbaum, Hillsdale, New Jersey.

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

[Gebhardt et al., 1997] Gebhardt, F., Voss, A., Grather, W., and Schmidt-Belz, B. (1997). *Reasoning with Complex Cases*. Kluwer.

[Glasgow et al., 2004] Glasgow, J., Epstein, S. L., Meurice, N., and Vercauteren, D. P. (2004). Spatial motifs in design. In *Proceedings of the Third International Conference on Visual and Spatial Reasoning in Design*.

[Glasgow et al., 1993] Glasgow, J. I., Conklin, D., and Fortier, S. (1993). Case-based reasoning for molecular scene analysis. In *Working Notes of the AAAI Spring Symposium on Case-Based Reasoning and Information Retrieval*, pages 53–62, Menlo Park, California. AAAI Press.

[Glasgow et al., 1995] Glasgow, J. I., Fortier, S., Conklin, D., and Allen, F. (1995). Knowledge representation tools for molecular scene analysis. In *Proceedings of the 28th Annual Hawaii International Conference on System Biotechnology Computing Track*.

[Gomes et al., 2003] Gomes, P., Seco, N., Pereira, F. C., Paiva, P., Carreiro, P., Ferreira, J. L., and Bento, C. (2003). The importance of retrieval in creative design analogies. In Bento, C., Cardoso, A., and Gero, J., editors, *Creative Systems: Approaches to Creativity in AI and Cognitive Science. Workshop program in the Eighteenth International Joint Conference on Artificial Intelligence*, pages 37–45, Acapulco, Mexico.

[Gonzalez and Woods, 1992] Gonzalez, R. C. and Woods, R. E. (1992). *Digital Image Processing*. Addison-Wesley, New York.

[Hammond, 1990] Hammond, K. J. (1990). Case-based planning: A framework for planning from experience. *Cognitive Science*, 14(4):385–443. CHEF.

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

[Hennessy et al., 2000] Hennessy, D., Buchanan, B., Subramanian, D., Wilkosz, P. A., and Rosenberg, J. M. (2000). Statistical methods for the objective design of screening procedures for macromolecular crystallization. *Acta Crystallogr D Biol Crystallogr*, 56(Pt 7):817–827.

[Holyoak and Thagard, 1997] Holyoak, K. J. and Thagard, P. (1997). The analogical mind. *American Psychologist*, 52(1):35–44.

[Hunter, 2004] Hunter, L. (2004). Life and its molecules. *AI Magazine*, 25(1):9–22.

[Jaccard, 1908] Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 44:223–270.

[Jurisica and Glasgow,] Jurisica, I. and Glasgow, J. I. Applications of case-based reasoning in molecular biology. *AI Magazine*, 25.

[Jurisica and Glasgow, 2000] Jurisica, I. and Glasgow, J. I. (2000). Extending case-based reasoning by discovering and using image features in in-vitro fertilization. In *ACM Symposium on Application Computing (SAC*

2000) Biomedical Computing- special session on biomedical applications of knowledge discovery and data mining, Villa Olmo, Italy. CITO.

[Jurisica et al., 2001a] Jurisica, I., Rogers, P., Glasgow, J. I., Collins, R. J., Wolfley, J. R., Luft, J. R., and DeTitta, G. T. (2001a). Improving objectivity and scalability in protein crystallization: Integrating image analysis with knowledge discovery. *Intelligent Systems in Biology, Special Issue of IEEE Intelligent Systems*, pages 26–34.

[Jurisica et al., 2001b] Jurisica, I., Rogers, P., Glasgow, J. I., Fortier, S., Collins, R. J., Wolfley, J. R., Luft, J. R., and DeTitta, G. T. (2001b). Integrating case-based reasoning and image analysis: High-throughput protein crystallization domain. In *Proceedings of the Innovative Applications of Artificial Intelligence (IAAI01)*, pages 73–80, Seattle. IRIS, CITO.

[Kettler and Darden, 1993] Kettler, B. and Darden, L. (1993). Protein sequencing experiment planning using analogy. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, pages 216–224.

[Kolodner, 1993] Kolodner, J. L. (1993). *Case-based reasoning.* Morgan Kaufmann, San Mateo, California.

[Leng et al., 1993] Leng, B., Buchanan, B., and Nicholas, H. B. (1993). Protein secondary structure prediction using two-level case-based reasoning. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, pages 251–259.

[McCarthy et al., 2002] McCarthy, J., Minsky, M., Sloman, A., Gong, L., Lau, T., Morgenstern, L., Mueller, E. T., Riecken, D., Singh, M., and Singh, P. (2002). An architecture of diversity for commonsense reasoning. *IBM Systems Journal*, 41(3):530–539. John McCarthy, Marvin Minsky, Aaron Sloman, Leiguang Gong, Tessa Lau, Leora Morgenstern, Erik T. Mueller, Doug Riecken, and Moninder Singh, and Push Singh (2002). An architecture of diversity for commonsense reasoning. IBM Systems Journal, 41(3):530-539.

[Monaghan and Clement, 1999] Monaghan, J. M. and Clement, J. (1999). Use of computer simulation to develop mental simulations for understanding relative motion concepts. *International Journal of Science Education*, 21(9):921–944.

[Nilges et al., 1988] Nilges, M., Clore, G., and Gronenborn, A. (1988). Determination of the three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. *FEBS Lett.*, 229:129–136.

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

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

[Riesbeck and Schank, 1989] Riesbeck, C. and Schank, R. (1989). *Inside case-based reasoning.* Lawrence Erlbaum: Hillsdale, NJ.

[Shavlik, 1991] Shavlik, J. (1991). 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.

[Shepard and Cooper, 1988] Shepard, R. and Cooper, L. (1988). *Mental Images and their Transformations.* MIT Press, Cambridge, Massachusettes.

[Smith and Chang, 1994] Smith, J. and Chang, S. (1994). Quad-tree segmentation for texture-based image query. *Proceedings of the second ACM international conference on= Multimedia*, pages 279–286.

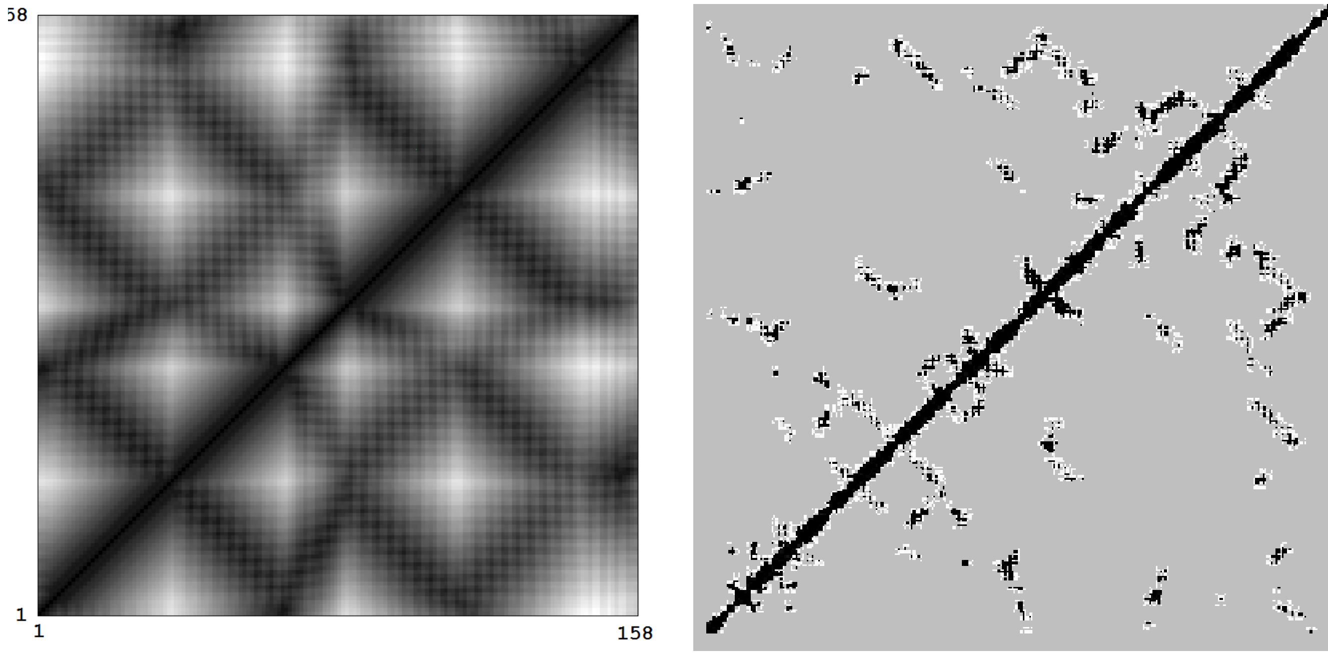
[Sullivan and Baker, 1994] Sullivan, G. and Baker, R. (1994). Efficient quadtree coding of images and video. *IEEE Transactions on Image Processing*, 3(3):327–331.

[Veloso, 1993] Veloso, M. M. (1993). Prodigy/analogy: Analogical reasoning in general problem solving. In *EWCBR*, pages 33–52.

[Vendruscolo et al., 1997] Vendruscolo, M., Kussell, E., and Domany, E. (1997). Recovery of protein structure from contact maps. *Folding and Design*, 2:295–306.

[Won et al., 2002] Won, C. S., Park, D. K., and Park, S. J. (2002). Efficient use of mpeg-7 edge histogram descriptor. *Electronics and Telecommunications Research Institute Journal*, 24:23.

[Zhang and Waltz,] Zhang, X. and Waltz, D. Protein-structure prediction using memory-based reasoning: A case study of data extrapolation. In *Proceedings of a Workshop on Case-Based Reasoning*, pages 351–355, San Francisco, California. Morgan Kaufmann.



Distance map

Contact map

Figure 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 $10\mathring{A}$ of one another). Secondary structures are easily recognizable in a contact map: α-helices appear as thick bands along the main diagonal; β-sheets appear as thin bands parallel and perpendicular to the main diagonal.

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**

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

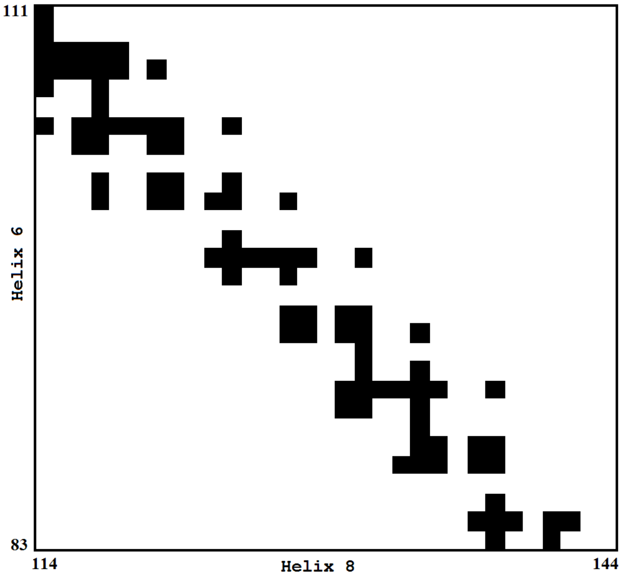


Figure 3: The sub-contact $C_{Helix-6,Helix-8}$ map for a pair of helices in protein Bacterioferritin. Since the diagonal band shows contacts that extend from the beginning of helix 6 and end of helix 8, to the end of 6 and beginning of 8, we can discern that the helices are oriented anti-parallel to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

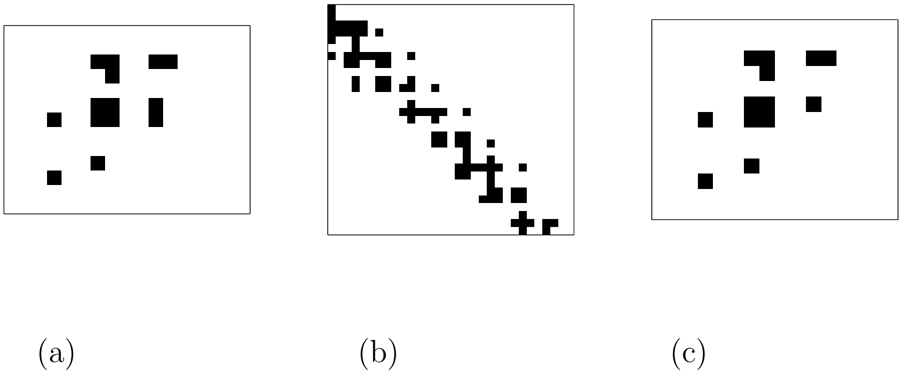


Figure 4: Illustration of similar, (a) and (c), contact maps and a map (b) that is dissimilar to the other two.

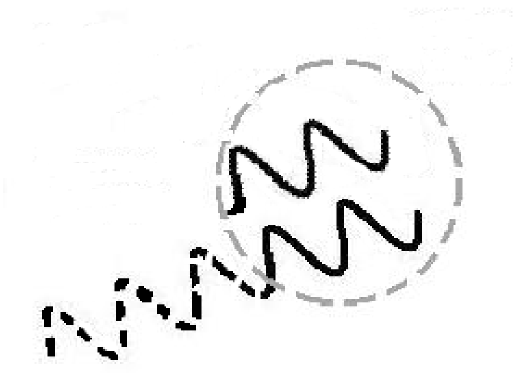


Figure 5: In this figure the lower helix is the target and the upper is the source. The dotted gray circle represents the mapping area. The locations of the target amino acid residues for which there are no cooresponding source residues are inferred based on the known geometry for helices. These "grown" areas are represented with the dotted black line.