

**DISTANCE GEOMETRY,
HELIX PACKING,
&
CONTACT MAP CONGRUENCY ADVISORS**

Queen's University, School of Computing
Supervisor: Dr. J. Glasgow
By: Mireille Gomes

Acknowledgements

I would like to thank *The Natural Sciences and Engineering Research Council of Canada* (NSERC), and *The Computer Research Association's Committee on the Status of Women in Computing Research* (CRA-W), for presenting me with this research opportunity. I would also like to thank my supervisor, Dr. J. Glasgow, for her mentorship and assistance during this project.

TABLE OF CONTENTS

Acknowledgements	1
Table of Contents	2
1. Introduction	3
2. Background	5
2.1. Protein Structure from Contact Maps: A Hierarchical Approach.....	5
2.2. Contact Maps.....	9
2.3. Distance Geometry.....	10
2.4. Helix Packing & Clustering Contact Maps.....	11
3. Methodology	13
3.1. Data Used.....	13
3.2. Distance Geometry Advisor - Lower and Upper Bounds.....	14
3.3. Helix Packing Advisor.....	18
3.4. Contact Map Congruency Advisor.....	18
4. Results	20
4.1. Distance Geometry Advisor - Lower and Upper Bounds.....	20
4.2. Helix Packing Advisor.....	20
4.3. Contact Map Congruency Advisor.....	24
5. Discussion & Future Work	25
6. Bibliography	27

1. INTRODUCTION

More than half the dry weight in a cell is made up of proteins. These abundant and important molecules carry out numerous functions; for example, they are used to support the skeleton, control senses, move muscles, digest food, defend against infections and process emotions. There are more than 100,000 proteins that come in all shapes and sizes; however, they are all made up of the same set of 20 amino acids, its primary sequence. The shape of a protein is determined by the folding of this primary sequence. The shape of a protein is vital; for, entire biological systems can be ultimately explained at the basic level of protein shape fitting and interaction.

Since the functional capabilities of a protein are largely determined by its structure, the prediction of protein structure is an area of significant research. Despite recent efforts to develop automated protein structure determination procedures, structural genomics projects are slow in generating spatial distribution for complete proteomes, and folding remains unknown for many protein familiesⁱ. Alternatively, prediction algorithms provide cheap and fast methods to assign spatial distributions for many proteins. The first class of these types of methods, include threading and comparative modeling and rely on detectable homology traversing most of the modeled sequence and at least one known structureⁱⁱ. The second class of methods, *de novo* or *ab initio* methods, predict the structure from sequence alone, without considering the similarity in spatial distribution between known and unknown structuresⁱⁱⁱ.

Arriving at the native confirmation of a polypeptide chain is a substantial problem in protein structure prediction. Ideally, the predicted model will have a 3D configuration precisely congruent to the actual protein conformation. Developers of prediction methods tend to over-estimate the performance of their tool. This can be attributed to the following reason: Single forms of assessment are not sufficient to describe the functioning of a method^{iv}.

This paper addresses three heuristic methods to evaluate the predicted structures of alpha helix pairs in proteins while accounting for some of the physical, chemical and

homologous properties of helices, namely distance geometry, helix packing and contact map similarity. These three Advisors are part of a Case-based Reasoning (CBR) approach of protein structure prediction. They are amongst 20 to 30 Advisors that will ultimately be employed to determine a native conformation of a protein of unknown structure.

These particular Advisors were designed because it is believed that distance geometry, helix packing and contact maps are key contributors and identifiers of protein structure. For example, it has long been recognized that helical-helical interactions play a vital role in stabilizing membrane proteins^{v,vi}. Several research groups have been successful in determining novel protein structure using distance geometry methods (Aszodi et al., 1995^{vii}; Mumenthaler and Braun, 1995^{viii}; Nilges, 1995^{ix}). On the other hand, the contact map provides useful information about the protein's secondary structure, and it also captures non-local interactions giving clues to its tertiary structure^x.

The hypothesis behind this research is that by assessing and scoring predicted helix pair models of known helix structures, taking into consideration only a singular property of a protein sub-domain at a time, we will eventually be able to isolate the native confirmation of the entire protein structure that adheres to all known biological, physical and chemical attributes of a protein.

This report will provide background information on the CBR approach of protein structure prediction, the theory behind Contact Maps, the formalism of Distance Geometry, and the types of helix packing. It will also delineate the methodology used to create each of the three Advisors. Results of implementing each of the Advisors as part of the CBR system of protein structure prediction will be stated and discussed.

2. BACKGROUND

2.1 Protein Structure from Contact Maps: A Hierarchical Approach

In order to better understand protein-protein interactions and in turn biological pathways and functions, it is imperative that we determine the structure of proteins. Knowledge of protein structure is critical in determining the evolutionary origins of proteins, drug design, re-construction of defective proteins and synthetic protein creation. The accurate prediction of protein structure from sequence data is a fundamental problem in modern molecular biology. One approach to this problem is to first predict a contact map and structural features (such as secondary structures) from a given protein sequence, and then to reconstruct the 3D structure of the protein from its predicted contact map. A proposed method, to address the second step in this process, uses the experience embedded in the Protein Structure Databank (PDB)^{xi}. This method is hierarchical, in the sense that it applies contact maps at varying levels of the complexity hierarchy for a protein. Maps that describe the contact between secondary structures (and potentially super-secondary structures) are used to initially locate known proteins that share high-level structural properties with the input protein^{xi}. Then contacts among amino acid residues are examined to determine more detailed similarities among the input protein and substructures for proteins within the database^{xi}.

A Case-Based Reasoning (CBR) framework is applied. CBR is a paradigm that involves solving new problems by recalling old problems and their solutions, and by adapting these previous experiences, which are represented as cases. CBR is founded on the premise that similar problems have similar solutions. The underlying hypothesis of this research, and the motivation for using CBR as a problem solving tool, is that proteins with similar contact maps also have similar structures^{xi}. 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 or anti-parallel to the main diagonal. A contact map can be viewed as a translational and rotational invariant representation of the protein's topology, thus capturing much of its relevant structural information. It also

provides a “fingerprint” that can be used to efficiently compare proteins to find ones with similar structures. This approach incorporates a hierarchical search strategy that initially locates proteins that have similar secondary structures to our input protein^{xi}. Given a protein p with m secondary structures (α -helices, β -sheets and coils), its *secondary structure contact map* is defined as the $m \times m$ binary array S_p such that $S_p(m,n) = 1$ if there exists a contact in map C_p between any residue in secondary structure m and any residue in secondary structure n . Otherwise $S_p(m,n) = 0$.

Moreover, the *profile*, $P_{m,n}$, is defined for each contact (m,n) in the secondary structure contact map as the subarray of C_p such that the rows of $P_{m,n}$ correspond to the amino acid residues in secondary structure m and the columns correspond to the residues in secondary structure n ^{xi}. Profiles need only be defined for contacts along and below the diagonal of the secondary structure contact map, as the profile for contact (i,j) is equivalent to that for contact (j,i) . Note, that unlike the contact map and the secondary structure contact map, a profile contact map is not symmetric^{xi}. The motivation for defining secondary structure contact maps and profiles is to provide an efficient and detailed representation of high-level contacts that allows us to initially compare the structural similarity between proteins without considering the more complex contact maps^{xi}.

The CBR system compares the novel contact map with contact maps derived from structures in the PDB based on image retrieval techniques. Four similarity metrics are applied^{xi}: Quad-tree vector; Minkowski distance of colour and edge distributions; Gray-level co-occurrence matrix interception; and Jaccard distance^{xii}. Refer to Figure 2.1.

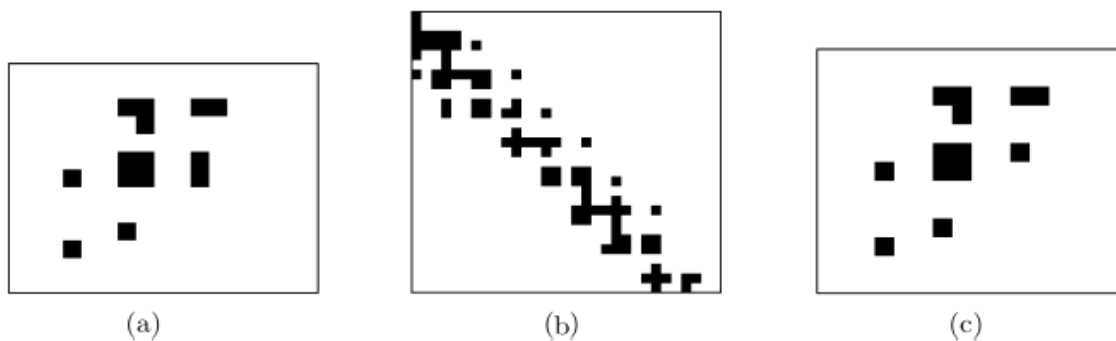


Figure 2.1.1^{xi}

Given the following contact maps, maps (a) and (c) were determined to be similar, while map (b) was considered to be dissimilar.

Once similar profile contact maps and their corresponding structures have been retrieved from the database, they are adapted to determine the relative locations of secondary structures in the protein^{xiii}. The adaptation program transfers the coordinate information of the best matched helix pairs for each novel helix pair^{xiii}. Any missing 3D coordinates are determined using biochemical knowledge^{xiii}.

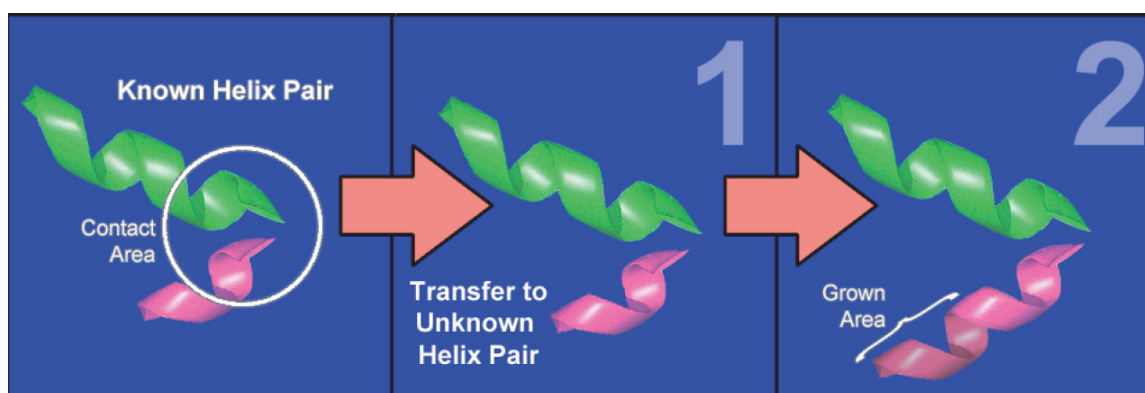


Figure 2.1.2^{xii}

Transfer of the coordinate information of the best matched helix pairs for each novel helix pair. Any areas of missing 3D coordinates are “grown” using biochemical knowledge.

The predicted structures are then evaluated by a series of knowledge-based “expert advisors^{xi}.” Each of the advisors computationally ranks the quality of each prediction based on a different criterion, such as polarity, distance geometry, etc. A neural network assigns weights to the advisors with the assumption that certain combinations of the advisors give optimal predictions^{xi}. These determined weights are applied to compute a weighted sum of individual advisor rankings in order to determine an overall ranking for each predicted structure^{xi}.

This CBR implementation is currently being tested on ideal (physical maps computed using coordinates in the PDB) and predicted contact maps. On average, the distance between the predicted and correct coordinates for the best adaptation is 1.234 Å (Ångströms)^{xii}.

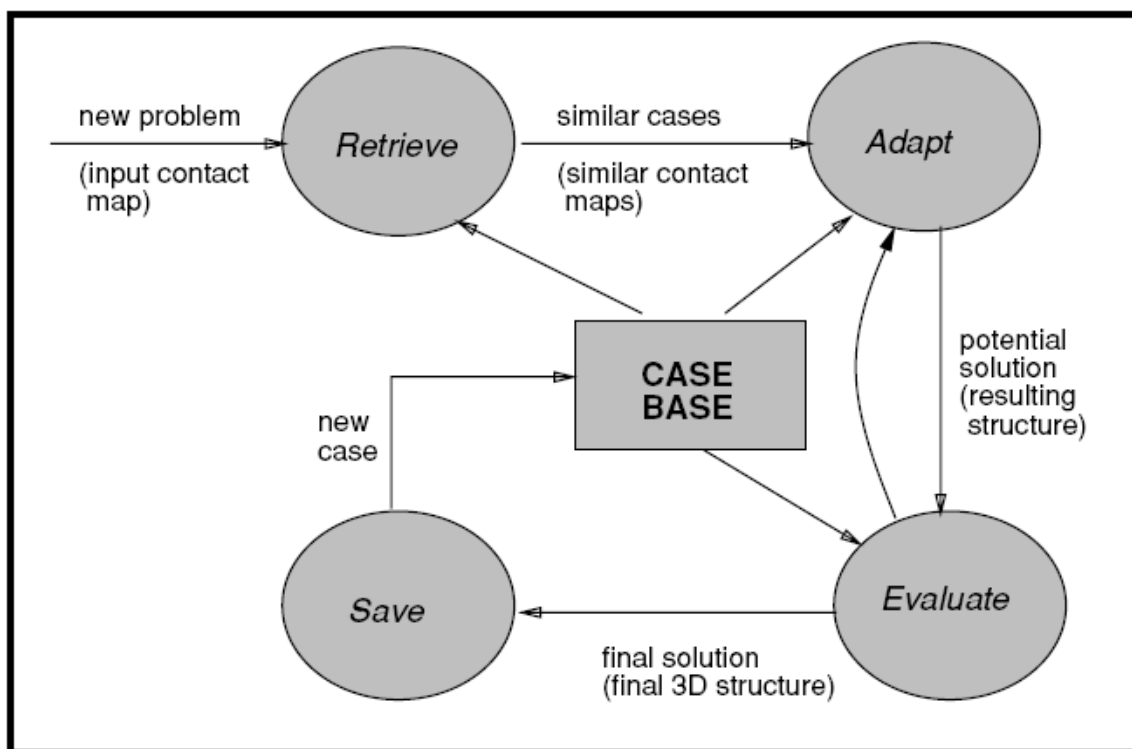


Figure 2.1.3^{xi}: Design of a Case-Based Reasoning system for determining protein structure from contact maps.

2.2 Contact Maps

Contact maps are the fundamental tool utilized in this hierarchical approach of protein structure prediction. Traditionally contact maps are created from distance maps, where a threshold value is applied to given a distance matrix to produce the Boolean values, as illustrated in Figure 2.2.1. A distance map, D , is a $M \times M$ matrix where M is the number of amino acids in a protein and D_{ij} is the distance between amino acid i and amino acid j in the protein in 3D space, usually measured in Ångströms (Å). However, the problem of predicting a distance map from the primary sequence of a protein is almost as hard as predicting the structure of the protein from the primary sequence.

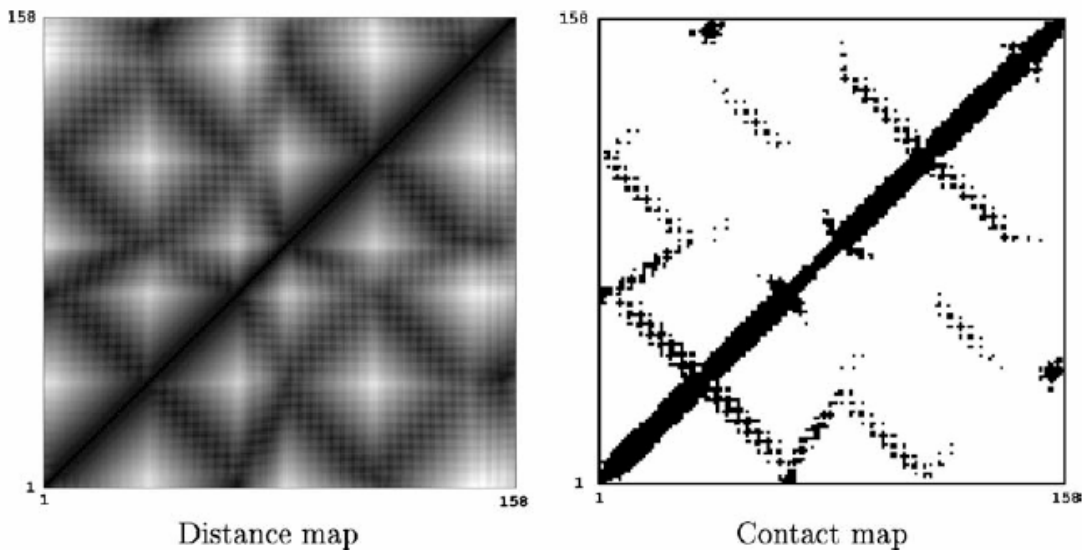


Figure 2.2.1^{xi}

Left to right, distance map and contact map for the protein Bacterioferritin (Cytochrome B1). The axes in both maps represent the residues for the protein starting at the N terminus in the left corners. In the distance map the darker areas represent closer distance, while in the contact map the dark areas represent a value of 1 where residues are in contact (within 10 Å of each other).

In consequence, this research group uses the contact maps predicted by Fariselli, P. et al. (2001)^{xiv} in Bologna, Italy. The team in Bologna uses neural networks to formulate these maps with evolutionary pathways, conserved regions of sequence, and predicted secondary structures among the information that is used as inputs^{xiv}. The method of contact map prediction is superfluous to the study being presented in this report and it is

assumed that the input contact maps given to us by the group in Bologna will be very accurate some time in the future.

Determining an ideal value to use as a threshold for creating contact maps, also poses a challenge. Various research groups are using different thresholds; such as, 8 Å^{xiv}, 9 Å^x and even 5 Å^{xv}. Our lab has chosen to adopt 10 Å between alpha carbons as the threshold value, for it proved to be the most useful for our purposes. It was observed that with values less than 10 Å we would not have enough points of contact and consequently not enough information to describe the interface between a pair of helices.

In the Advisors under consideration in this paper all contact maps are created using a threshold value of 10 Å and by utilizing the distance map methodology, after having predicted coordinates for pairs of alpha helices.

2.3 Distance Geometry

Several types of structural information, for example, distances, chemical cross-linking, neutron scattering, etc. can be expressed as intra- or intermolecular distances. The distance geometry formalism permits these distances to be assembled and three-dimensional structures consistent with them to be calculated.

Some authors, such as Mor'e, J. and Wu, Z. (1999)^{xvi}, have suggested the use of distance geometry in conjunction with optimization techniques and experimental data to predict the structure of a protein. For instance, in NMR experiments the structure of a protein is construed using distance geometry methods which are applied to a set of internuclear distances. However, owing to experimental error, only upper and lower bounds for the distances can be obtained.

In relation to the CBR approach employed in this research, some authors (Pollastri G. and Baldi P. (2002)^{xvii}, Mor'e, J. and Wu, Z. (1999)^{xvii}) have even suggested the use of distance geometry, optimization techniques and contact maps to predict protein structure.

Sometimes contact map information is supplemented with statistical data about the distances between residues and the constructed structure is refined using knowledge-based and inverse kinematics approaches^{xvii}.

2.4 Helix Packing & Clustering Contact Maps

In the bottom up approach of protein structure prediction employed in this research we start by predicting the configuration of the secondary structures. Over one-third of a globular protein is made up of alpha helices. Consequently, it is imperative that we formulate an accurate method of predicting the coordinates of alpha helices.

The face to face interaction (packing) between a pair of alpha helices is insufficiently understood in the *ab initio* method of prediction^{xviii}. Several packing models have been developed, for example, “ridges into grooves”^{xix} and “knobs into holes”^{xx}. However, these models insufficiently describe helix pair interaction and structure. The Helix Packing Advisor discussed later in this report is based on the hypothesis that if two pairs of alpha helices have similar contact map interfaces then they will pack similarly (and consequently be assigned similar packing values)^{xxi}.

In following with this hypothesis, contact maps for pairs of alpha helices and have been divided into three classes^{xxii}: The first class includes any maps with contacts in the corner. Maps that have contacts within the outer two rows or columns, but not in the corners, belong to the second class, “edge contact maps.” Finally, maps that do not have any contacts in the perimeter are grouped in the third class known as “central contact maps.”

The classification of contact maps into these three categories is done using the greedy algorithm in the order of corner, edge and central contact maps^{xxii}. A database for each of the three types of contact maps was created. In order to create this database 1078 alpha helix pairs belonging to 171 proteins from the PDB were used^{xxii}. The subsequent table (Table 2.4.1) summarizes the contents of the database:

Contact Map Class	Total Instances	Face-to-Face Packing	Not Face-to-Face Packing
Central	112	112 (100%)	0 (0%)
Edges	535	531 (99.3%)	4 (0.7%)
Corner	431	397 (92.1%)	34 (7.9%)
Doubled Corner	862	794 (92.1%)	68 (7.9%)
All Maps	1078	1040 (96.5%)	38 (3.5%)

Table 2.4.1^{xxii}

Database of clustered contact maps based on 1078 alpha helix pairs belonging to 171 proteins from the PDB.

The corner and edge contact maps are clustered further because clusters of maps share the same packing value. In order to cluster maps a similarity matrix is constructed by treating the contact map as a point (first translated into a vector) and finding the cosine distance between two contact maps^{xxii}. For example, if n contact maps are to be clustered, the similarity matrix S is size $n \times n$, where entry $S(i,j)$ is the cosine distance from contact map i to j .

The similarity matrix is further reduced using k nearest neighbours (k-nn) sparsification, where k is the number of points that are closest to a specific point (or contact map, in this case) using a particular distance measure^{xxii}. The k points with the smallest distances are found in each row of the similarity matrix^{xxii}. If point j is a member of the k-nn list of point i , then the k-nn list of point j is checked to see if point i is a member^{xxii}. In this manner there will be total mutuality in the k-nn list for each contact map.

The coordinates of a pair of alpha helices can then be determined by comparing its contact map to that of its k nearest neighbours of known structure, and transferring the corresponding spatial distribution.

3. METHODOLOGY

3.1 Data Used

The data used to implement the Distance Geometry, Contact Map and Helix Packing Advisors is training data which was taken from 329 alpha helix pairs found in the PDB. For each of the alpha helix pairs a text file containing the optimal retrieved and adapted cases from the CBR system is available. Each text file, example shown below, contains the primary sequence of the protein the helix pair is from, the residue numbers of the specific helix pairs in contact, the correct native coordinates, the source coordinates of each optimal retrieved case, and the predicted coordinates based on each set of source coordinates.

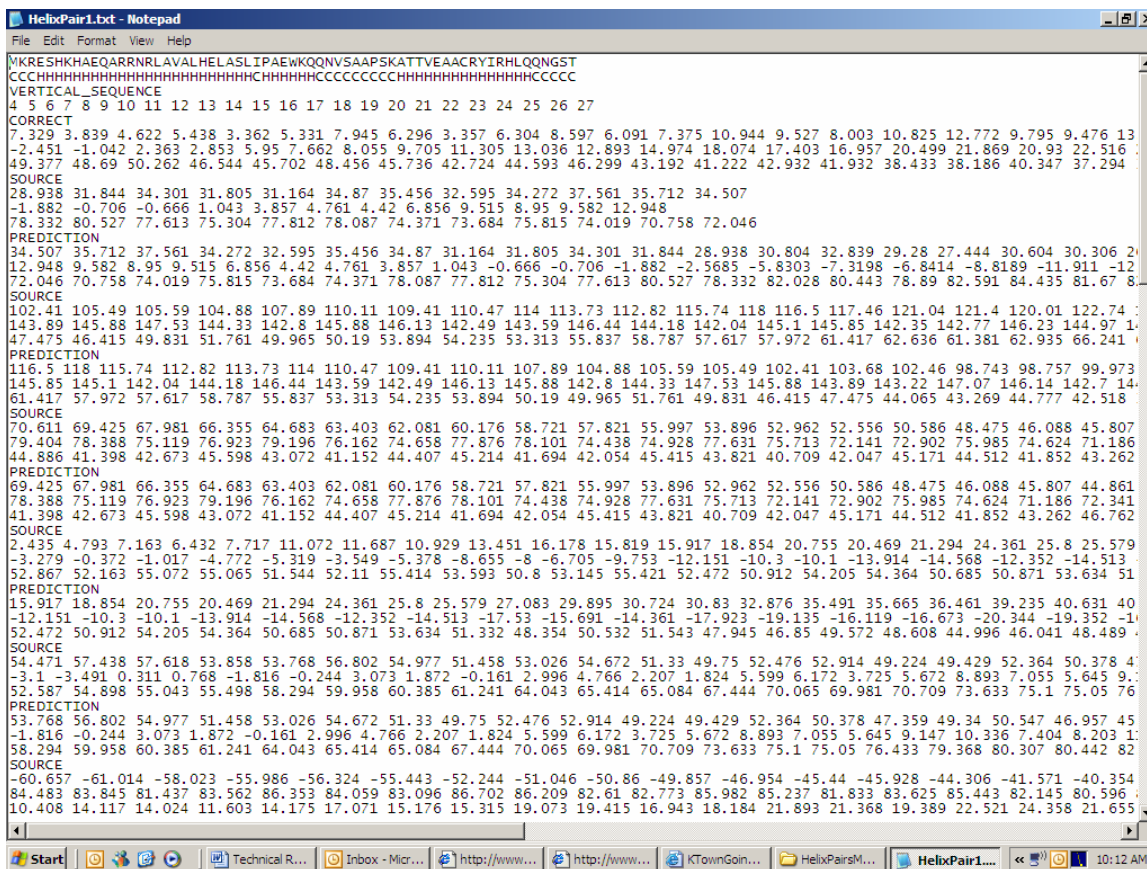


Figure 3.1.1
Screen shot of a text file used in the execution of the three Advisors under consideration in this report.

3.2 Distance Geometry Advisor - Lower and Upper Bounds

Given the predicted structure of a helix pair, the Distance Geometry Advisor verifies whether the distance between any pair of amino acids falls within a determined upper and lower bound. This is of importance because distance is a key feature in preserving biochemical properties within a protein structure. In order to compute these bounds, statistical data about minimum and average distances between alpha carbons is taken into consideration.

From literature we know that the average distance between adjacent alpha carbons along a helix backbone is 3.84 Å. The lower and upper bounds for distances between adjacent alpha carbons are thus calculated by adding or subtracting 0.5 Å respectively from this distance.

Under the assumption that residues of a helix are distributed according to a model average helix it is assumed that the alpha carbons are distributed along a helix with radius 2.5 Å, 100 degrees separation, and a distance along the axis of 1.5 Å. The lower and upper bounds for distances between non-adjacent pairs of amino acids within the same helix are obtained by adding and subtracting 0.5 Å respectively to the estimated distance between the two amino acids. The distance between the residue pair is computed as follows; where x_1 , y_1 and z_1 correspond to the coordinates of the first residue, and x_2 , y_2 and z_2 to that of the second residue. “positionResidue#” refers to the position of that residue in the alpha helix:

$$\begin{aligned}y_1 &= 2.5 * \cos(\text{positionResidue1} * (100/360) * 2 * \pi) \\z_1 &= 2.5 * \sin(\text{positionResidue1} * (100/360) * 2 * \pi) \\x_1 &= 1.5 * \text{positionResidue1} - 1.5\end{aligned}$$

$$\begin{aligned}y_2 &= 2.5 * \cos(\text{positionResidue2} * (100/360) * 2 * \pi) \\z_2 &= 2.5 * \sin(\text{positionResidue2} * (100/360) * 2 * \pi) \\x_2 &= 1.5 * \text{positionResidue2} - 1.5\end{aligned}$$

$$\text{distance} = ((x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2)^{1/2}$$

If two amino acids, one from each of the helices, in the helix pair are in contact, the lower bound is set to the allowable minimum distance according to the type of amino acids of their corresponding residues. This data is obtained by calculating the minimum distance between a pair of amino acids from a group of approximately 2400 proteins in the CATH database which belong to class 1: proteins which are mainly composed of alpha helices. While the upper bound is set to the threshold value of the contact map (10 Å in this lab's case).

If the amino acids, one from each helix, are not in contact the lower bound is set to the threshold value of the contact map and the upper bound is set to the average distance between alpha carbons (3.84 Å) times the number of carbons separating the pair of residues.

In a three dimensional Cartesian space, constraints are imposed by the Euclidean metrics in order to make sure that the distances are geometrically consistent. The triangle inequality is one such constraint:

$$D_{AC} \leq D_{AB} + D_{BC}$$

Other constraints, though important, are computationally costly to employ. Consequently, it is common to use an approximation based solely on the triangle inequality. As such, two rules to correct inconsistencies in lower (L) and upper (U) bounds are deduced from the triangle inequality:

$$U_{AC} \leq U_{AB} + U_{BC}$$

$$L_{AC} \leq L_{AB} - U_{BC}$$

This process is known as boundary smoothing and can be reduced to the equivalent problem of finding the shortest path in a digraph^{xxii}.

ALGORITHM 3.2.1: Distance Advisor Boundary Smoothing. Taken from Havel, T.^{xiii}.

```
for k = 1:Natoms do
  for i = 1:Natoms-1 do
    for j = i+1:Natoms do
      if Upper(i,j) > Upper(i,k) + Upper(k,j) then
        Upper(i,j) = Upper(i,k) + Upper(k,j)
      if Lower(i,j) < Lower(i,k) - Upper(k,j) then
        Lower(i,j) = Lower(i,k) + Lower(k,j)
      if Lower(i,j) < Lower(j,k) - Upper(k,i) then
        Lower(i,j) = Lower(j,k) - Upper(k,i)
    end
  end
end
```

All of the above information which depicts how the Distance Geometry Advisor establishes the upper and lower bounds for a pair of amino acid residues from alpha helices has been adapted from Zuviria, E. (2006)^{xxiii}.

In order to determine good scoring criteria all the “correct” helix coordinates for each helix pair text file were run through the distance advisor and each “correct” helix pair was assigned a score from 1 to -1 based on the number of residues pairs that fell within the computed upper and lower and bounds. The greater the number of residue pairs that fell within the bounds, the closer the score of that prediction to +1. The average and range of their ranking was determined, and in consequence the predicted structures were ranked relatively. See Algorithm 3.2.2 for scoring methodology.

ALGORITHM 3.2.2: Distance Advisor Scoring Methodology

PART 1 - Verify whether alpha carbons in the helix pair are within the upper & lower bounds. Assign appropriate score.

```
for c1 = 1:lengthHH
  for c2 = 1:lengthVH
    idealDistance = (lHP(c1, c2) + uHP(c1, c2))/2
    numCPairsHP = numCPairsHP + 1
    if (HP(c1, c2) < lHP(c1, c2) or HP(c1, c2) > uHP(c1, c2))
      score = score + -1
    elseif (HP(c1, c2) >= (idealDistance - 1) & HP(c1, c2) <= (idealDistance + 1))
      score = score + 1
    else
      score = score
    end %end if
  end %end for c2
end %end for c1

% Calculate average score
score = score / (numCPairsHH + numCPairsVH + numCPairsHP);
```

PART 2 - Based on the calculated score assign predicted structure a relative score as determined by the “correct” structures’ score.

```
if score >= 0.25 && score <=0.4
  score = 1
elseif score > 0 && score <0.2
  score = 0.2
elseif (score >= 0.2 && score <=0.25) || (score > 0.4 && score <=0.5)
  score = 0.4
elseif score < 0 && score > -0.0777
  score = -0.8
elseif score <-0.0777 || score > 0.6168
  score = -1
else
  score = 0
end
```

Legend:

“HP” → Helix Pair
“VH” → Vertical Helix
“HH” → Horizontal Helix
“%” → Comment

3.3 Helix Packing Advisor

The primary step in this Advisor is to generate an input contact map and a contact map for each of the predicted structures of a helix pair, using the distance matrix method discussed in section 2.2 of this report. If the input contact map (currently determined by the coordinates of the “correct” helices as stated in the text files (Fig. 3.1.1)) of the considered helix pair is a “central contact map” the Advisor affirms that the two helices should pack with 100% confidence. However, if the input contact map belongs to the corner or edge classes, then we search the database for the closest matching contact map. The confidence is then the cosine distance between the input map and its closest neighbour.

In the hierarchical approach used to predict helix pair structures from the source contact map, each source contact map will have on average 9 sets of predicted helix pair coordinates. If the packing in a set of predicted coordinates matches the packing of the input map’s closest neighbour, which by design will be the same as the input contact map, the confidence score is positive, else it is negative. Similarly, if the predicted structure of a central input contact map does not pack then it is given a negative score, otherwise the score is positive.

3.4 Contact Map Congruency Advisor

The Contact Map Advisor evaluates a predicted helix pair’s spatial distribution based on how similar the input contact map is to the prediction’s contact map. This is because in theory the contact map of the final predicted protein structure should be identical to the input map produced by Fariselli, P. et al. (2001)^{xiv}.

The contact map of the predicted structure is produced by first determining the Euclidean distance between each amino acid pair in the two helices. The distance is determined by employing the following formula: $((x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2)^{1/2}$. Subsequently if

the distance between any pair of amino acids is less than 10 Å the corresponding position in 2D array is assigned a 1, else it holds a 0.

Each position in the computed predicted structure contact map is then compared with the corresponding position in the input contact map (presently constructed using coordinates of the “correct” helix pair structure, as it appears in the text file (Fig. 3.1.1)). The total number of positions with equal values in the corresponding maps is tallied and the percentage determined. This percentage is then linearly scaled to fit a scoring range of -1 to 1, using the formula $2x - 1$, where x is the calculated percentage.

Note: Eventually this CBR approach of predicting protein structures will substitute the input contact maps yielded from the “correct” helix pair’s structure for input maps produced by Fariselli, P. et al. (2001)^{xiv} in the Helix Packing and Contact Map Congruency Advisors.

4. RESULTS

4.1. Distance Geometry Advisor

When the 329 “correct” helix pairs were run through PART 1 of Algorithm 3.2.2 the following were the scoring results:

average score = 0.3249
minimum score = -0.0777
maximum score = 0.6168
number of scores in range 0.0 to 0.2 = 24
number of scores in range 0.2 to 0.25 = 32
number of scores in range 0.25 to 0.3 = 62
number of scores in range 0.3 to 0.35 = 63
number of scores in range 0.35 to 0.4 = 73
number of scores in range 0.4 to 0.5 = 59
number of scores in range 0.25 to 0.4 = 205
number of scores in range -0.0777 to 0.0 = 2
number of scores in range -1 to -0.0777 (min.) and 0.6168 (max.) to 1 = 2

These results verify that average alpha helix specifications are for an optimized helix and rarely occur in nature; for most distances between residues fall somewhere within the range of the upper and lower bound, and not exactly in the middle of that range.

The average score after running the predicted helix structures through this advisor was 0.2911. This in turn indicates that the predicted structures obey as much of the distance geometry rules of alpha helices as correct helices themselves. Further suggesting that the CBR prediction method utilized encompasses precisely this vital aspect of alpha helix structure.

4.2 Helix Packing Advisor

The following (Fig. 4.2.1) is sample output retrieved after running the HelixPair1.txt file through this advisor. As illustrated in the data, the score is either positive or negative a certain number. This is because for edge or corner maps if the packing of the predicted structure matches the packing of the closest neighbour then the score is determined by the

positive cosine distance between the input map and its closest neighbour. Conversely, if it does not match the score is determined by the negative cosine distance.

```
>> start('e:\Lab\HelixPackingAdvisor\HelixPairsMarch22\HelixPair1.txt')
```

```
Prediction =
```

```
1
```

```
edge
```

```
score =
```

```
-0.5727
```

```
Prediction =
```

```
2
```

```
edge
```

```
score =
```

```
-0.5727
```

```
Prediction =
```

```
3
```

```
edge
```

```
score =
```

```
0.5727
```

```
Prediction =
```

```
4
```

```
edge
```

```
score =
```

```
0.5727
```

```
Prediction =
```

```
5
```

```
edge
```

score =

0.5727

Prediction =

6

edge

score =

0.5727

Prediction =

7

edge

score =

-0.5727

Prediction =

8

edge

score =

-0.5727

Prediction =

9

edge

score =

-0.5727

avgScore =

-0.0636

Figure 4.2.1

Sample output after running the HelixPair1.txt file through the Helix packing Advisor

Table 4.2.1 presents examples of results from the Helix Packing Advisor. The helix pair files in the table were randomly selected from the 329 helix pair text files available.

Helix Pair #	Type of Contact Map	Average Score
110	Corner	0.6447
98	Corner	0.6976
329	Corner	0.7276
222	Corner	0.4753
1	Edge	-0.0636
22	Edge	0.7285
50	Edge	0.7071
212	Edge	0.6197
28	Central	1
124	Central	1
248	Central	1
321	Central	1

Table 4.2.1
Results from Helix Packing Advisor

These results assert that most predicted structures have a score above 0.6 and thus pack similarly to the type of packing indicated by the input contact map. Therefore, the CBR method employed to determine the predictions, and the hypothesis that “if two pairs of alpha helices have similar contact map interfaces then they will pack similarly”^{xx} help support each other.

Another noteworthy observation was that attempting to find examples of helix pairs that pack centrally, was much more difficult than locating corner or edge helices. This in turn correlates to the packing distribution shown in Table 2.4.1. The fact that all predicted structures with central input maps pack face-to-face with 100% confidence, reaffirms that the CBR approach of assigning spatial distribution to a protein of unknown structure by first finding similar contact maps of proteins of known structure is step in the right direction of protein structure prediction.

4.3 Contact Map Congruency Advisor

When the 329 helix pair files, with an average of 9 predictions each, were run through the Contact Map Congruency Advisor the overall average score was: 0.5622786093419401 and 65% of predictions had a score between and including 0.5 and 1. This indicates that most predictions of helix pairs have structures similar to that of the native confirmation. This in turn denotes that the employed method of helix structure prediction, that is adapting coordinate information of structures with similar contact maps, is accurate and noteworthy.

5. DISCUSSION AND FUTURE WORK

In the near future the Case-based Reasoning approach will employ a neural net algorithm to evaluate the performance of each Advisor. This evaluation will depend on gathering information about what constitutes a “good” protein structure. This is being done by Glasgow et al., 2006^{xi}:

1. conducting statistical and machine learning analysis of existing databases of protein structures
2. acquisition of expert knowledge (accessed through interviews and research papers)
3. using and adapting text-book knowledge
4. calculating the root mean square distance (RMSD) between the spatial distribution of the predicted structure and the native confirmation

The lower the RMSD, the greater the similarity in protein structure of the prediction to the native confirmation. The effectiveness of the Distance Geometry, Helix Packing and Contact Map Congruency Advisors, along with all the other 20 to 30 Advisors will be assessed by comparing their scoring techniques to the RMSD. If a prediction has a low RMSD, then the Advisor should assign it a relatively high score. For a high RMSD, we should expect a score close to -1.

Through utilization of this comparative technique of an Advisor’s score to the RMSD a weighted tally of each Advisor is determined and an overall score for each prediction is calculated by employing the following formula:

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

In this formula A_i is an Advisor, m is the predicted structure, $s(A_i, m)$ is the strength of the A_i ’s assessment for m , and $w(A_i)$ is the weight assigned to Advisor A_i .

Consequently, the novel protein's final predicted structure will obey the principles of chemistry, biology and physics. It will also adhere to experimental data and the

previously determined structures, and it will be consistent with the information available from the primary sequence.

BIBLIOGRAPHY

- ⁱ Ginalski, K., Grishin, N. V., Godzik, A., and Rychlewski, L. Practical lessons from protein structure prediction. *Nucleic Acids Res* 33: 1874-1891, 2005.
- ⁱⁱ Baker, D. and Sali, A. Protein Structure Prediction and Structural Genomics. *Science*, 294 (5540): 93-96, 2001.
- ⁱⁱⁱ Friesner R. A. and J. R. Gunn. Computational studies of protein folding. *Annu. Rev. Biophys. Biomol. Struct.*, 25: 315-342, 1996.
- ^{iv} Eyrich et al. EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, 17 (12): 1242, 2001.
- ^v Popot, J. L. and Engelman, D. M. Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, 29: 4031-4037, 1990.
- ^{vi} Kahn, T. W. and Engelman, D. M. Bacteriorhodopsin can be refolded from two independently stable transmembrane helices and the complementary @ve-helix fragment. *Biochemistry*, 31: 6144-6151, 1992.
- ^{vii} Aszodi, A., Gradwell, M.J. and Taylor, W.R. In Protein folds: a distance based approach (Bohr, H. and Brunak, S., eds.), *Protein folds: a distance based approach*. CRC Press, Boca Raton, Florida, pp. 85-97, 1992.
- ^{viii} Mumenthaler, Ch. and Braun, W. Predicting the helix packing of globular proteins by self-correcting distance geometry. *Prot. Sci.*, 4: 863-871, 1995.
- ^{ix} Nilges, M. Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J. Mol. Biol.*, 245: 645-660, 1995.
- ^x Vendruscolo, M., Kussell, E. and Domany, E.. Recovery of Protein Structure from Contact Maps. *Folding and Design*, 2: 295-306, 1997.
- ^{xi} Glasgow, J., Kuo, T. and Davies, J., ``Protein Structure from Contact Maps: A Case-Based Reasoning Approach'', *Information Systems Frontiers*, 8(1): 29-36, Special Issue on Knowledge Discovery in High-Throughput Biological Domains, Springer, January 2006.
- ^{xii} Abelson, A., Davies, J., Fraser, R., Kuo, T., Zuviria, E., and Glasgow, J. Protein structure from contact maps: An hierarchical approach. *Intelligent Systems for MolecularBiology Conference (ISMB05)*, 2005.
- ^{xiii} Davies, J., Glasgow, J. and Kuo, T. Visio-spatial Case-Based Prediction of Protein Structure. *Journal of Computational Intelligence*, in press.
- ^{xiv} Fariselli P, Olmea O, Valencia A, and Casadio R. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering*, 14(11):835-843, 2001.
- ^{xv} Kettler B, and Darden L. Protein sequencing experiment planning using analogy. *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 216-224, 1994.
- ^{xvi} Mor'e, J. and Wu, Z. Distance Geometry Optimization for Protein Structures. *Journal of Global Optimization*, 15: 219-234, 1999.

-
- ^{xvii} Pollastri, G. and Baldi, P. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, 18:S62-S70, 2002
- ^{xviii} Chivian, D., Robertson, T., Bonneau, R., and Baker, D. Ab initio methods. *Methods Biochem Anal.*, 44: 547-57, 2003.
- ^{xix} Chothia, C., Levitt, M., and Richardson, D.. Helix to helix packing in proteins. *Journal of Molecular Biology*, 145: 215-250, 1981.
- ^{xx} F. Crick. The packing of α -helices: simple coiled coils. *Acta Crystallographica*, 6: 689-697, 1953.
- ^{xxi} R. Fraser. A Tale of Two Helices: A study of alpha helix pair conformations in three-dimensional space. Master's thesis, Queen's University, 2006.
- ^{xxii} Havel T. F. Distance geometry: Theory, algorithms and chemical applications. *Harvard Medical School, Boston, MA, USA*.
- ^{xxiii} Zuviria, E. Reconstruction of Protein Structures from Contact Maps. Master's thesis, Queen's University, 2006.