Critical examination of residue-residue interaction used in protein structure recognition

by

Hua Wang

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Condensed Matter Physics

Program of Study Committee:
Kai-Ming Ho, Major Professor
Drena L. Dobbs
Edward Yu
E. Walter Anderson

Iowa State University

Ames, Iowa

2005

Graduate College
Iowa State University

This is to certify that the MASTER OF SCIENCE thesis of

Hua Wang

has met the thesis requirements of Iowa State University

Signatures have been redacted for privacy

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

We use a gaped structural threading method to predict the protein conformations in solution. After analysis of the results in CASP6 (Critical Assessment of Techniques for Protein Structure Prediction), we found there are some weak points in the scoring function which we should refine.

We made three attempts to improve the scoring function. First we automatically adjusted the 20 parameters of residue-residue interactions. Then we investigate 33 contact matrix potentials, in order to select the best one for energy evaluation. Last, we add the Blosum Matrix score in our energy function to take account of the sequence similarity information in consideration.

CHAPTER 1. Introduction

Proteins are the biochemical molecules that make up cells, organs and organisms. They perform a wide variety of activities in the cell. Following are a few examples of some general protein functions. Enzymes catalyze almost all biological reactions. Transport protein carries small molecules or ions. Structural protein provides mechanical support to cells and tissues. Motor protein generates movement in cells and tissues. Storage protein stores small molecules or ions. Signaling protein carries signal from cell to cell. Receptor protein used by cells to detect signal and transmit them to the cell's response machinery. Gene regulatory protein binds to DNA to switch genes on or off. Also, there are special purpose protein made by organisms with highly specialized properties(1).

Each protein has a particular shape and function. The structure of a protein often makes it possible for people to deduce its function. How proteins put themselves together forming a three dimensional structure is called "protein-folding problem". A protein is a large complex molecule made up of one or more chains of amino acids. Protein folding is the process by which a protein assumes its functional shape or conformation. All protein molecules are made up of one or more simple unbranched chains of amino acids. The chains coil into a specific three-dimensional shape that enable the proteins to perform their biological functions.

Amino acid is an organic compound containing an amino group (NH_2), a carboxylic group (COOH), and any various side groups. Proteins are polymers of amino acids joined with covalent linkage which is called peptide bond. There are precise twenty different amino acids in nature. Although there is no obvious chemical reason why other amino acids could not serve in the proteins, the nature only choose these set of amino acids and there is no change could be made on them. Amino acids comprise the protein and their side chains contribute

chemical versatility to the protein. Two amino acids have Acidic side chains. Three amino acids have basic side chains. Five have uncharged polar side chains. Ten have non-polar side chains. For small proteins, the amino acid sequence is sufficient to determine the final folded structure, or conformation of a protein, which has the minimum free energy(2).

A major technique that has been used to discover the three-dimensional (or tertiary) structure of a protein is X-Ray crystallography. A well ordered crystal of a pure protein must be grown. Another method, nuclear magnetic resonance (NMR) spectroscopy has also been used to analyze the structure of small proteins or protein domains. It is the only technique that can provide detailed information on the exact three-dimensional structure of biological molecules in solution. But the determination of the folded structure of a protein is a lengthy and complicated process. Thus prediction of native structure from amino-acid sequences alone is a major area of interest in bioinformatics. There is still no reliable method to solve protein folding problems.

When studying the protein folding problem, we need to consider both the protein and the solvent effects on the protein. Protein folding is a consequence of intermolecular forces, including hydrogen bonds, van der Waals forces, electrostatic interactions, hydrophobic interactions. Hydrophobic energy, or the solvent effects, are a major force driving the protein folding(3) (4). Protein Folding is a spontaneous process. It appears that in transitioning to the native state, a given amino acid sequence always takes roughly the same route and proceeds through roughly the same number of fundamental intermediates. First the amino acid sequence (or **primary structure**) establishes **secondary structure**, particularly alpha helices and beta sheets, which are local structures (the third type of secondary structure or local structure is loop or coil). Then, it folds into its functional shape, **tertiary structure**. Tertiary structure may involve covalent bonding in the form of disulfide bridges formed between two cysteine residues. Disulfide bond does not change the conformation of a protein, however it acts as atomic staples to reinforce its most favored conformation. Many proteins assemble together to be a multiple-subunit protein which possesses a **quaternary structure**. Van der Waals forces can play important roles in protein-protein recognition when complementary shapes are

involved.

In the treatment of protein folding, if we consider all atoms of the protein and the solvent molecules around it in detail, the conformational freedom will be too vast to calculate. Following are the reasonable simplifications we applied on geometry of molecules and interaction potentials. For a protein with N residues, its three dimensional structure is represented by a $N \times N$ matrix of pairwise, inter residue contacts or contact map. Each residue is represented by the center of its side chain atom positions. The matrix element (i, j) is 1, when the distance between residue i and residue j is less than 6.5 Å, otherwise the element is zero. Nearest-neighbor pairs along a chain are explicitly excluded in counting contacts. Miyazawa-Jernigan (MJ) matrix has been widely applied in protein design and folding simulations(5)(6) to estimate the effective inter-residue contact energies for proteins in solution. The solvent effects are included into the effective inter-residue contact energies. The effective contact energies between residues in proteins is estimated directly from the numbers of residue-residue contacts observed in protein crystal structures by regarding them as statistical averages in the quasi-chemical approximation(7). Using the method of eigenvalue decomposition, Li, Tang, and Wingreen found MJ matrix can be accurately reconstructed from its first two principal component vectors as $M_{ij} = C_0 + C_1(q_i + q_j) + C_2 q_i q_j$(8). Where C's are constants, and the 20 q values are associated with the 20 amino acids. For simplification, these twenty parameters will be noted as LTW q values later. Thus,the MJ matrix tabulating the interaction strength between any two types of amino acids with 210 independent elements can be simplified by using only twenty parameters q_i with a lot advantage in theoretical modeling of proteins.

We make predictions by using threading procedures, which employ techniques for aligning the sequence with 3D structures and evaluate how well it fits using inter-residue potentials. The one-dimensional to three-dimensional alignment is a major problem in threading. Sequence similarity is not considered in our threading code. 'Gaps' are allowed in the process of alignment step. According to Lathrop and Smith's work(9), insertions and deletions are forbidden in the secondary structure regions, and no gap penalties added in the loop regions. The gap penalties of insertion or deletion are big in secondary structure regions, and they are

small in the loop regions in the process of evaluating alignments. There is no penalties in the step of final energy calculation. In the evaluation of prediction candidates, 'relative score' is applied and defined as $E^{rel} = E^{raw} - E^{ave}(10)$. It is similar to the Z-score $((E^{raw} - E^{ave})/\sigma)$, which has been showed to be more accurate than raw score for threading method by Bryant and Altschul(11), Meller and Elber(12).

The prediction of protein structure primarily based on two procedures. One step is structure generation that is to find out some structure templates and align on these templates to get prediction structures that are closest to the native structure. The other step is to evaluate the generated models with scoring function and find out the best model to be the native structure for the query sequence. If the structure generation fails to provide high quality models, then no matter how accurate the scoring function is, we won't get good prediction. On the other hand, if the scoring function is not capable to give good evaluation, although the best model is generated among other structure models, we still can't find out the correct one. Only by maintaining the well cooperation of these two steps, can we achieve satisfied prediction of protein conformation.

Our focus is to improve the scoring function for the threading code. The first approach is to automatically adjust the twenty parameters for the twenty amino acids, since these twenty parameters are directly used in the alignment step and final energy evaluation. The criterion is to maximize the correlation between scores given by threading approach and the GDT_TS measure, which is agreed to be the most accurate score in the evaluation of similarity among 3-D structures. We found the refined twenty parameters for the twenty amino acids could only yield is too small to be desirable because it is smaller than the error introduced due to random shuffling the sequence in the calculation of energy score.

We speculated that the twenty dimensional freedom was not big enough for adjusting. Thus, we still use twenty LTW q values to we began to find alignment first, but apply a 20×20 pairwise contact potential to determine the interaction of inter-residue contacts. Due to the symmetry in the pairwise inter-residue contact potential, there are 210 independent elements in it. The time cost is very expensive for automatic adjustment in 210 dimensional

space, so we analyzed twenty nine different published matrices of protein pairwise contact potentials (CPs) and four related statistical matrices, using two data bases. The evaluations of different matrices on two data bases are consistent with a correlation of 0.88. We happily found that ten CPs can yield noticeable and desired improvement in scoring function. Third we add in blosum matrix score, and it does help the scoring function to find out the right template structure.

CHAPTER 2. Adjust q values

2.1 Background of CASP

The CASP (critical assessment of structure prediction) experiment is run every two years. From CASP1 in 1994 to CASP6 in 2004, CASP experiments have gone across for a decade and changed from solving a fascinating puzzle to a real and serious enterprise. X-ray crystallographers and NMR spectroscopists provide the prediction targets. The predictors construct models for a number of protein sequences before the experimental results are known, during a period of 3-4 months. Then the crystallographers and NMR spectroscopists have their protein structures in public, when CASP experiment is over.

The CASP team uses GDT (global distance test) to detect regions of local and global structure similarities between the native protein structure (experimental results) and submission models. This method had been thoroughly tested in previous CASP experiments. For CASP6, GDT_TS measure is again used as the principal metric of main chain accuracy. Each residue from the model submitted by the predictor is assigned to the largest set of the residues (not necessary continuous) deviating from the native structure by no more than a specified distance cutoff. GDT_TS is defined as

$GDT\_TS = (GDT\_P1 + GDT\_P2 + GDT\_P4 + GDT\_P8)/4.0,$

where GDT_Pn is an estimation of the percent of residues that can fit under distance cutoff $\leq$ n.0 Angstroms.

The targets are divided into domains, when needed, and are classified into comparative (homology) modeling, fold recognition and new fold categories. The discrimination of these categories involves the kind of fold the given target sequence adopts. Whether it adopts a **new**

fold or one of the existing folds. For the case of existing folds, which one is the most suitable fold (**fold recognition**). If the fold recognition is clear, the core problem becomes how best the predictor can find a model with the relevant information from existing homologous structures in the protein data bank(**comparative modeling**).

2.2 Automatic Optimization of q values for twenty amino acids

From the results of CASP6, we found there are some generated models that are closer to the native protein structure than the first model in our submission, but the scoring function failed to identify these models. Thus it is necessary to improve the scoring function. The alignment and final energy calculation are based on the twenty q values associated with the twenty amino acids. These twenty q values are obtained from the eigenvalue decomposition of MJ matrix done by Hao Li, Chao Tang and Ned S. Wingreen(8). In attempt to improve the accuracy of scoring function, we first try to adjust the twenty q values, in order to get refined twenty q values, bringing better performance with the threading code.

The criterion we use to adjust the twenty q values is to minimize the average error of correlation. For each target, we use the CASP6 submission models of four top groups, our group submission models and the native structure to be decoy structures which have a lot similarity among each other. The real protein structures in the protein data bank (PDB) vary a lot from each other, so they are not demanding enough to test the scoring function. Even if they belong to the same family that means their average root mean square deviation is usually under 1 Å, they are still much more different with each other than the decoy structures we choose here. Because each group can submit 5 prediction models at the most, there are around 20 structure models which form a set of selected decoy structures for each target. Then we use our scoring function to calculate decoy structures' energy sores from that target. Together with the GDT_TS measures from the CASP6 results (The native structure has GDT_TS measure 100%), we can calculate the correlation between our prediction energy scores and GDT_TS measure. Correlation between two vectors x and y is defined as

$$cor(x,y) = \frac{\langle x - \bar{x}, y - \bar{y} \rangle}{\|x - \bar{x}\| \|y - \bar{y}\|} \tag{2.1}$$

There are about twenty points, corresponding to twenty decoy structures, to determine the correlation, for each target. With $-1 \leq cor(x,y) \leq 1$, the error, defined as $1 - cor(x,y)$, is in the range of $[0,2]$. The average error is $(\sum_{i=1}^{n} error_i)/n$, where n is the number of targets.

We combine our threading code with Mina, which can find an approximate minimum of a real function of n variables. Mina uses a selective directed search of a surrounding n-dimensional grid of points to find a direction in which the function decreases. Then it proceeds in this direction as long as the function decreases. And then it determines a new direction to travel. When there is no such direction found, the search increment factor is decreased and the above process is repeated. As a disadvantage, Mina can only find a local minimum position. Whether the minimum position is local minimum or global minimum value of the function is predetermined by the initial estimation of n variables and the range limit of the adjustment for each these variables .

We let Mina to minimize the average correlation error function, and the twenty q values for the twenty amino acids are implicit variables for this function. The initial estimation of these twenty variables are the twenty q values from Li, Tang, Wingree(8) parameterization of the MJ matrix(5)(6). The range for each of these twenty variables is set from 80% to 120% of the initial value, so there won't be nonphysical sharp increase or decrease in any of these twenty directions. The initial increment is set to be 20% of the initial q value in each these twenty dimensions.

2.3 Results and Conclusion

Protein molecules, only 3 to 10 nanometers across, can self-assemble quickly. Some are as fast as a millionth of a second, but it takes long time for computers to simulate. Considering the time cost, it is not feasible to put in all the targets in the refinement of q values. We tried to use five, seven and ten targets as the input for fitting, and it turned out that ten targets in fitting gave the best result. Among the ten targets, seven are fold recognition targets and the

other three are comparative modeling targets. Eight of them have relative low correlations compared with most other targets in CASP6. The total number of decoy structures in the fitting are 220. It takes 9.74 minutes for each step of fitting, when eight nodes are used from the clusters hal2004.

When the average error vibrated up and down across a certain value at the center, we assume that the minimization is stopped. Then we took the set of refined q values with minimum average correlation error to be the initial input of the twenty variables, and start a new cycle of refinement, keeping other settings the same as before. Thus the boundary of twenty q values are reset. The average error can be further reduced for the condition that any q value hit the boundary and can't be increased or decreased as needed. In this way, we repeated the refinement five times and following six pictures show the overall process of the minimization.

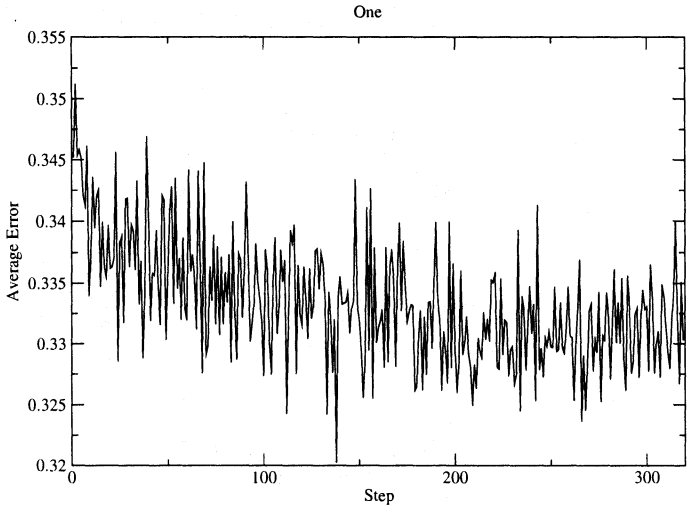

Figure 2.1 Refinement I

From thousands sets of output We selected 45 sets of refined vector q (properties of the 20 amino acids) to test. Using the first set of data base with 499 decoy structures for 23 targets, all 45 sets of vector q give average correlation between 0.721 to 0.745 . Since some targets in the first set of data base are also used in minimizing average correlation error, the

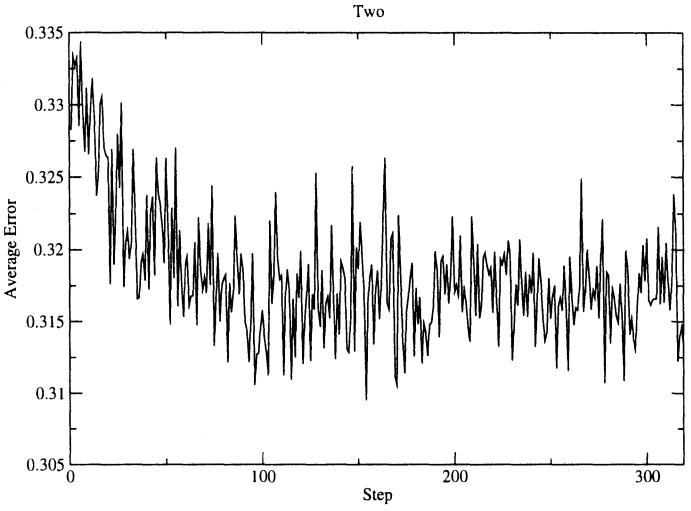

Figure 2.2 Refinement II

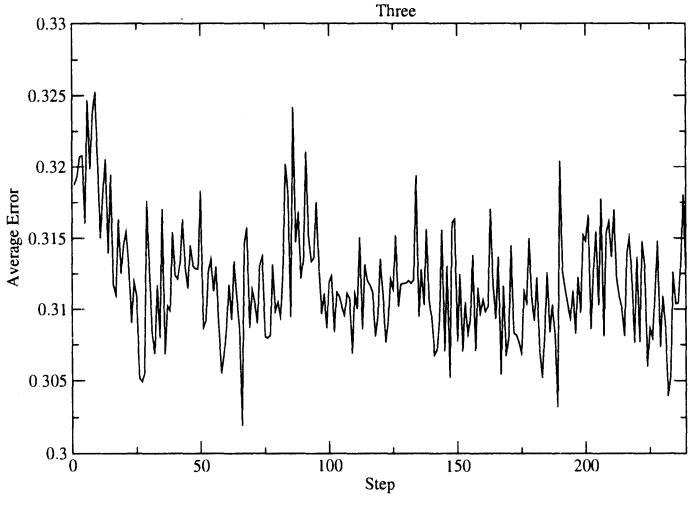

Figure 2.3 Refinement III

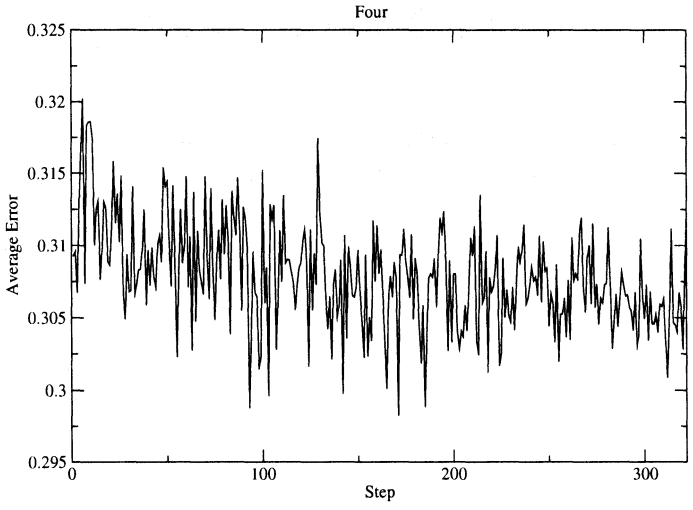

Figure 2.4 Refinement IV

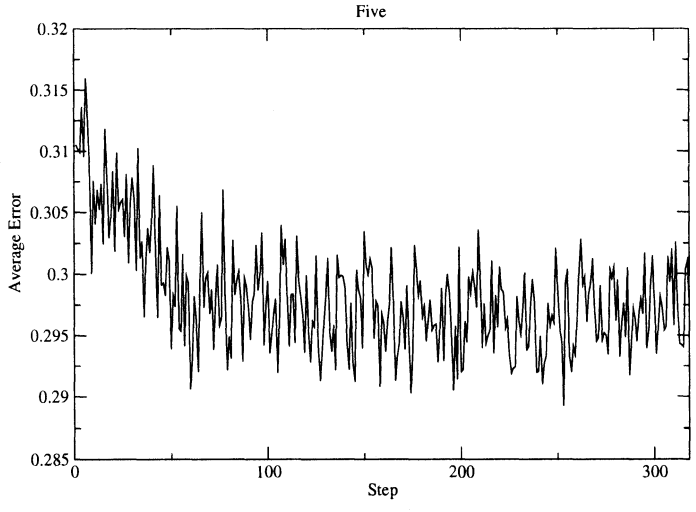

Figure 2.5 Refinement V

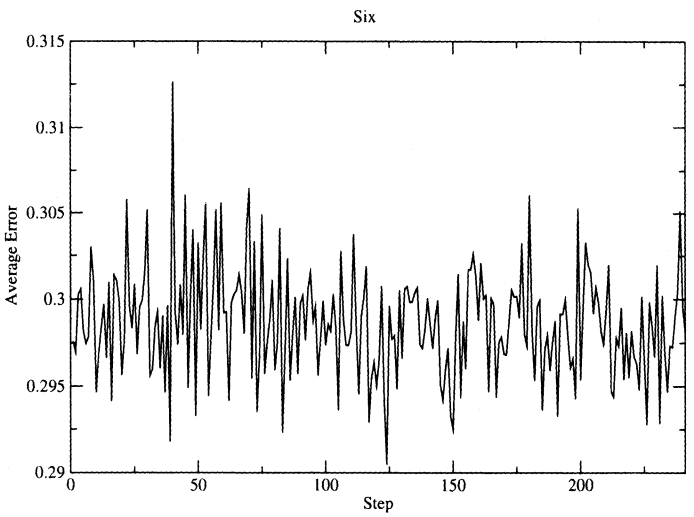

Figure 2.6 Refinement VI

refined vector q is especially favor these targets. So we set up another data base of 501 decoy structures for 25 targets, and there is no structure used either in fitting or be the same with the first set of data base. With the second data base, only eighteen sets of vector q give average correlation between 0.822 to 0.829. For the first and second data base, the average correlations of scoring function using the vector q derived by Li, Tang, and Wingreen (8) are both 0.727 and 0.829 respectively. I compared this vector q with the best refined vector q for the second set of data base in the following table.

Table 2.1 Comparing LTW vector q with refined vector q

q	LTW q	Refined q
C->L	0.9497 1.0097 1.1197 1.0697 1.1197	0.94970 0.92973 1.20928 1.15528 1.03012
V->G	0.9897 0.9797 0.9097 0.7997 0.7297	0.98970 0.90132 0.90970 0.79970 0.78808
T->D	0.7397 0.6897 0.6697 0.6997 0.6397	0.79888 0.68970 0.72328 0.69970 0.69088
E->P	0.6497 0.7897 0.6997 0.6097 0.7297	0.64970 0.72652 0.69970 0.60970 0.72970
Corr I	0.727	0.729
Corr II	0.829 (0.8287)	0.829 (0.8290)

The order of amino acids in the vector q is the following: "C" "M" "F" "I" "L" "V" "W" "Y" "A" "G" "T" "S" "N" "Q" "D" "E" "H" "R" "K" "P". Theoretically, it is always possible

to improve the correlation based on the existing vector q, but the finite times of shuffling the sequence to get E^{ave} introduces an random error in energy calculation $E^{rel} = E^{raw} - E^{ave}$. This error (around 1% of the average correlation for the data base) is small compared to the relative score, E^{rel}, but it can introduce ± 0.02 error for the average correlation, sometimes it is big enough to submerge the increase of correlation with the refined vector q. The average correlations listed in the table, using 200 times of shuffle, have the error $\pm$ 0.01.

It has been shown that only adjust the 20-dimensional vector q in the scoring function is not sufficient to improve the scoring function. Thus we begin to search if there is a contact potential matrix which can give better prediction in the energy calculation.

CHAPTER 3. Selection of the Contact Potential Matrices

3.1 Introduction of Contact Matrix Potentials

In order to find out a structure template that can closely resemble the structure of the query sequence, first we align the query sequence on a structure template, assuming the aligned residue' coordinates to be the same as the residue in the structure template. Thus after the alignment, we have a predicted structure for the query sequence. Following is to analyze the quality of the prediction by using the matrices of protein pairwise contact potentials (CPs). Due to the symmetry, there are 210 individual elements in 20×20 contact potential matrix. Each element e_{ij} represents the energy change introduced by the contact between residue i and residue j. The contact energy is defined as

$$E_c = \sum_{i,j=1}^{n} e_{ij} C_{ij} \tag{3.1}$$

where n is the length of the aligned sequence, and C_{ij} is the element of the contact matrix for the predicted model.

In our original threading code, we use the $q_i \times q_j$ to be the contact matrix element e_{ij}, where q_i, q_j are LTW q values. The 20×20 contact potential matrix is reconstructed by twenty independent variables. We want to find out if the LTW q values are the best set of twenty parameters for the alignment and if any CPs with 210 independent elements is better than the reconstructed contact matrix with twenty independent values in the scoring function. We evaluated 5 sets of q values in comparison with LTW q and 33 different published CPs(13).

3.2 Analysis

We choose 5 CPs and the corresponding twenty q values for each CP matrix. The 20-dimensional vector q is generated by minimize the sum of squares, known as the least squares problem

$$\sum_{i,j;i\leq j} [e_{ij} - \widetilde{e}(q)_{i,j}]^2 \rightarrow \min_{h,q}, \tag{3.2}$$

where the $e_{i,j}$ is the matrix element in the contact potential, and $\widetilde{e}(q)_{i,j}$ is the function of the 20-dimensional vector q defined as

$$\widetilde{e}(q)_{i,j} = C_0 + C_1 q_i q_j. \tag{3.3}$$

C_0 and C_1 are constant numbers. We draw the distribution map of the CP matrix elements, $e_{\imath,j}$, corresponding to the $\widetilde{e}(q)_{i,j}$ for each of the five CPs. The linear distribution means that the contact matrix constructed by the 20-dimensional vector q is equivalent to the CP Matrix with 210 variables. The distribution of CP Matrix B2 and MJ3 are broader than MJ3h, SJKG and SKOa Matrices.

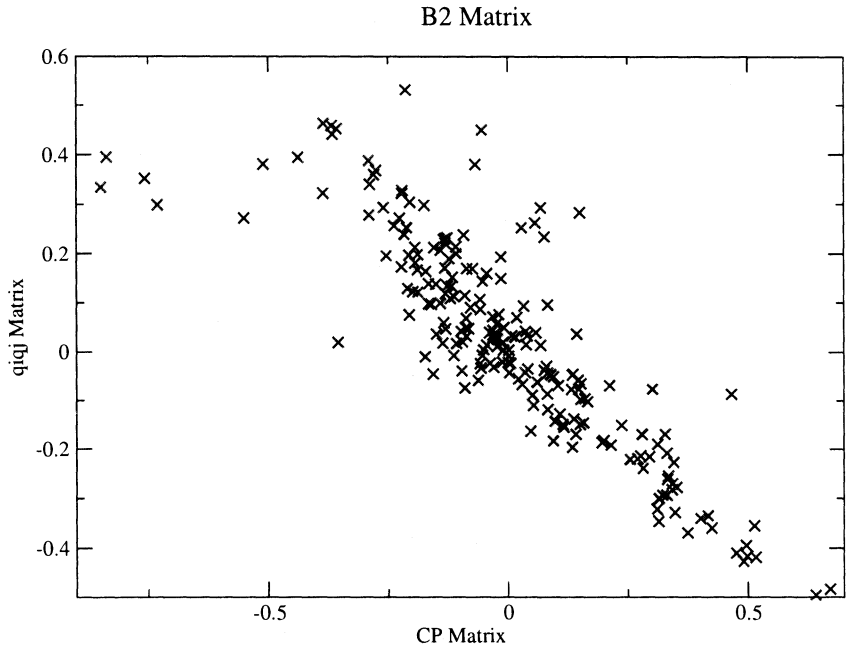

Figure 3.1 Distribution map for B2

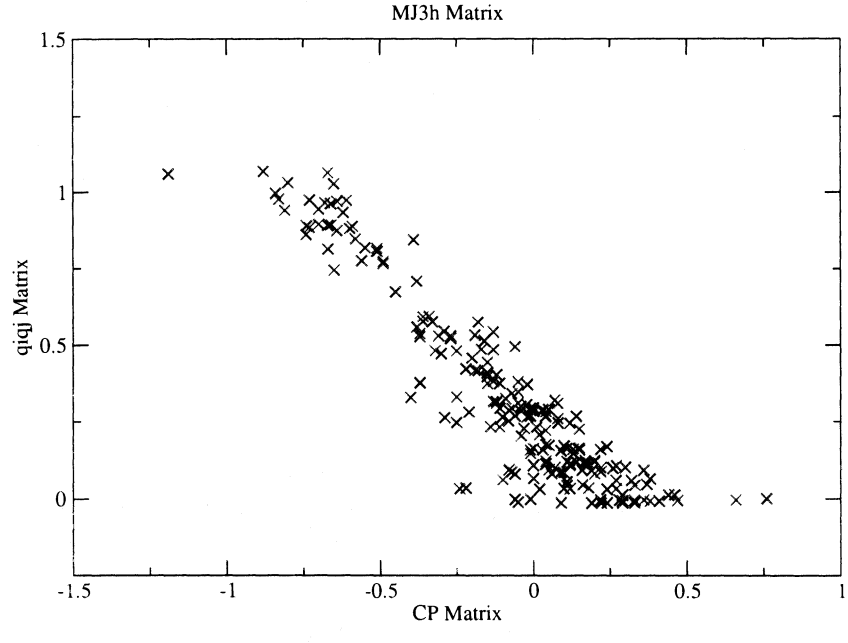

Figure 3.2 Distribution map for MJ3h

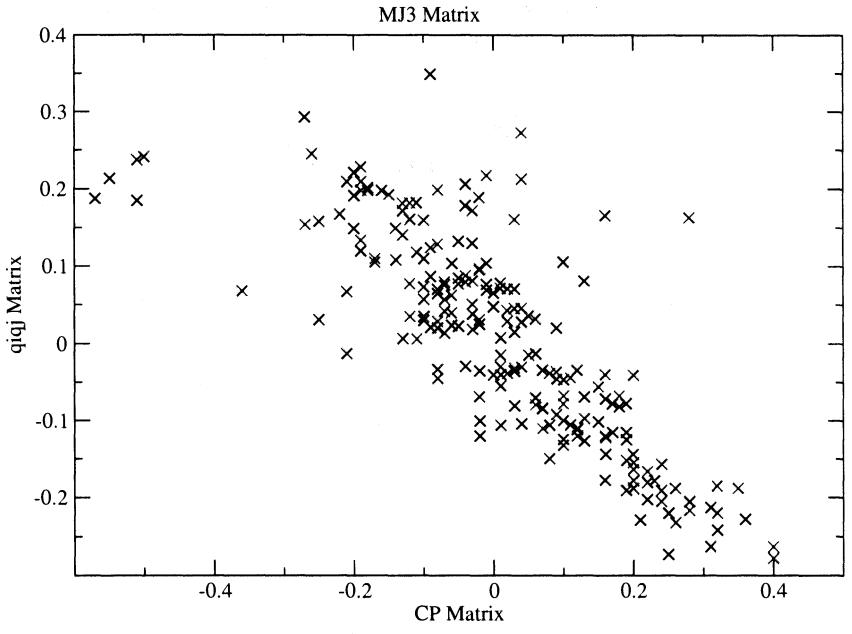

Figure 3.3 Distribution map for MJ3

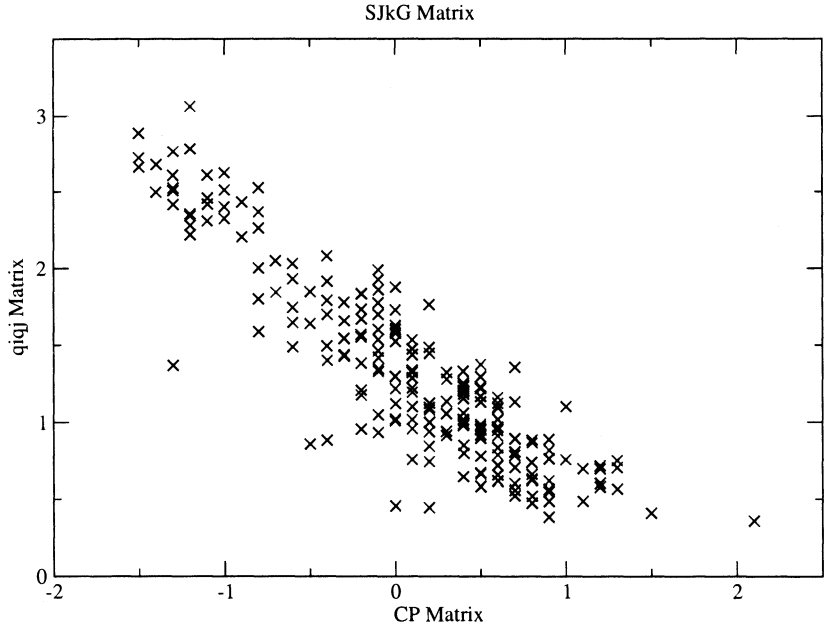

Figure 3.4 Distribution map for SJKG

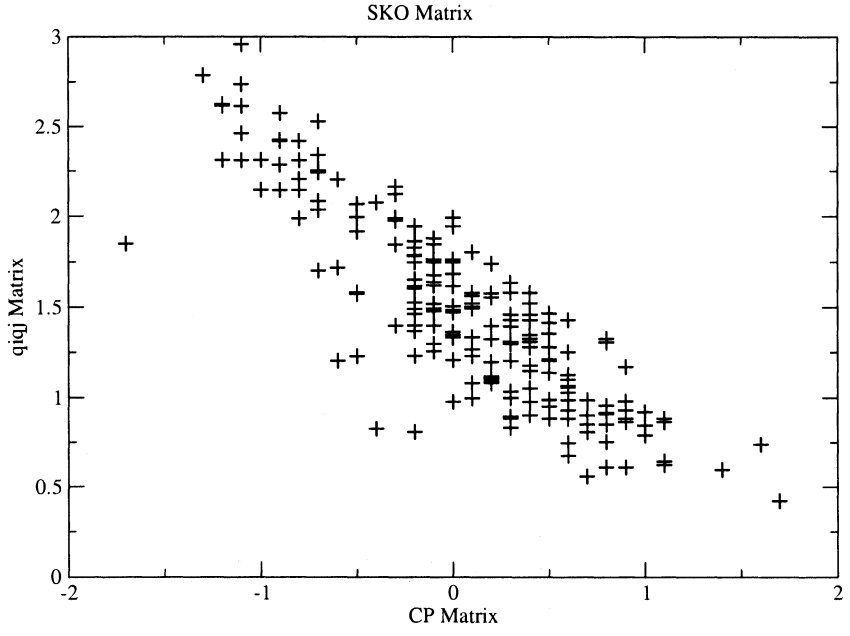

Figure 3.5 Distribution map for SKOa

The table 3.1 shows all possible combinations of the two possible choices in finding alignment, whether to apply LTW vector q or vector q got from equation 3.2 from each CP Matrix, and two possible choices in evaluating contact energy score, whether to apply CP Matrix or the matrix reconstructed by twenty q vector correspond to the CP Matrix. Comparing column two with column three, and column four with column five, under same matrix in evaluating contact energy, LTW q vector is better than other q vectors in finding the alignment. Comparing column two with column five and column three with column four, we find out under same alignment, CP Matrix is better than the reconstructed matrix of 20-dimensional q vector. Also, the average correlation of CP Matrix MJ3h, SJKG and SKOa are more desirable. This is consistent with the five distribution maps.

Table 3.1 Table 1

Matrix	Average Correlation			
	LTW q / Matrix	Matrix q / Matrix	Matrix q / $q_i \times q_j$	LTW q / $q_i \times q_j$
B2	0.694	0.660	0.649	0.681
MJ3h	0.728	0.727	0.677	0.668
MJ3	0.703	0.683	0.660	0.682
SJKG	0.737	0.730	0.716	0.721
SKOa	0.734	0.728	0.725	0.731

Based on above conclusions, we tested the performance of thirty three CPs in evaluating the contact energy while still use LTW q to find the best alignment of the query sequence. Besides data base I, we also set up data base II with 501 structures from twenty five targets to repeat the test.

3.3 Results

Twenty nine CPs together with four matrices were tested with two sets of databases. Using each matrix, the correlation between the contact potential energy scores and GDT_TS scores was obtained for each data base. For each matrix the performance in two data bases are showed in figure 3.6 with blue cross. Data of the correlations are listed in the table 3.2. The red diamond shaped data is the performance of our code. We classified the thirty three matrices into four groups: A, B, C, D. Matrix TEs is at the boundary of group A and group B. The correlation of the test results on two different data bases is 0.88. So this classification doesn't have bias on data base. We happily found that applying seven of the ten CPs in group A will contribute noticeable and desired improvement in scoring function.

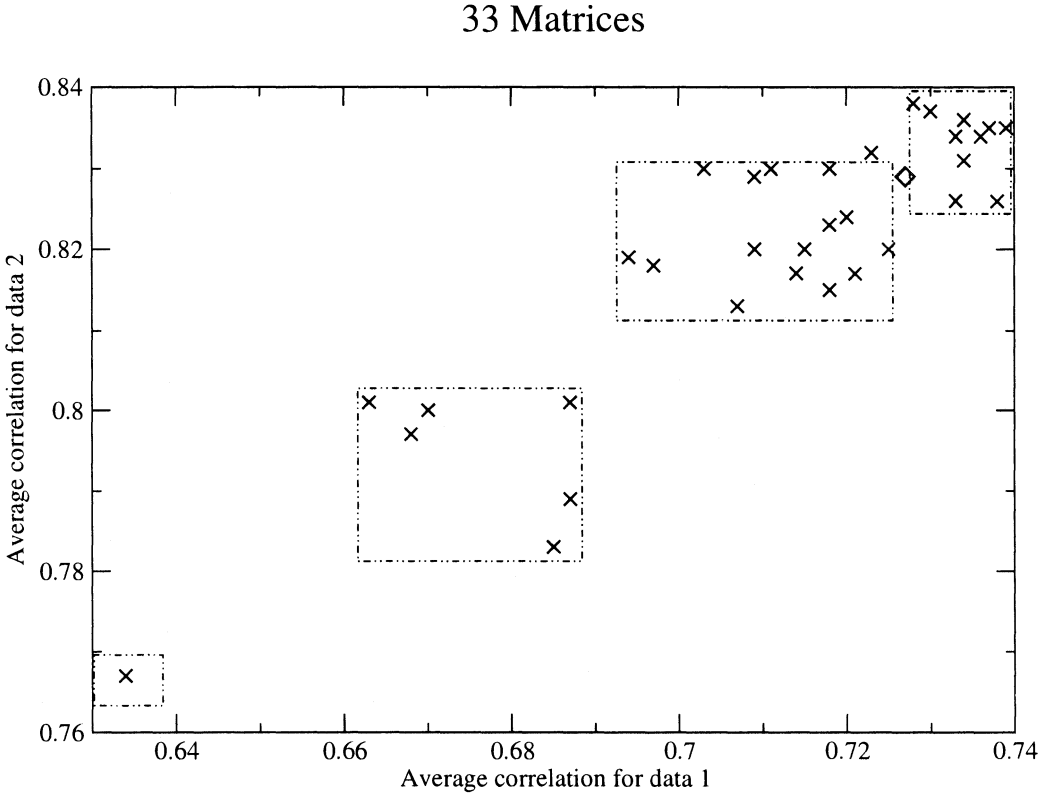

Figure 3.6 Correlations of each matrix for two databases

The 29 Cps and four matrices are listed and abbreviated as follows:

◇ TEl, TEs — Pairwise interaction contact potentials generated by Tobi, Elber et al(14). TEl and TEs potentials are obtained for large and small sets of decoys, respectively.

◇ MJPL, HLPL — Potentials developed by Park and Levitt with generated decoys(15). MJPL is refined from MJ1h, and HLPL is improved from an earlier potential of Hinds and Levitt.

◇ SJKG, SKOa, SKOb — Potentials derived by Skolnick et al with quasichemical approximation based on protein structural database(16)(17).

◇ BFKV — Optimized potential tested by PDB database, also a modified version of VD(18).

◇ MJ1, MJ1h, MJ2, MJ2h, MJ3, MJ3h — Potentials derived by Miyazawa Jernigan and published in 1985(19),1996(20), and 1999(21). Each article contain a derivation of two potentials. CPs marked with the suffix "h" include energy of transfer of amino acids from water to the protein environment. Matrice N.MJ2 is the number of contacts and Matrices IN.MJ2 is the logarithms of the number of contacts.

◇ TS, N.TS, IN.TS — TS is the statistical potential including the role of the solvent (water) published in 1976 (22). Matrices N.TS = (N_{ij}) and IN.TS = $[\log(N_{ig})]$ are also evaluated, where N_{ij} is the number of contacts between amino acids i and j.

◇ TD — Effective interresidue interaction "potentials" derived by Thomas PD, Dill KA, from protein structures in the Protein Data Bank(PDB) (23).

◇ Qa, Qm, Qp — The statistical potentials for the side chain interactions are orientation dependent. These new quasi-chemical potentials are developed by Kolinski et al(24).

◇ BL — distance-dependent statistical potential proposed by Bryant and Lawrence (25).

◇ BT — refined potential of MJ2h, derived by Betancourt and Thirumalai using lattice models of proteins(26).

◇ VD —effective potential proposed by Vendruscolo and Domany(27).

◇ B1,...,B5 —the latest version of quasi-chemical potential derived by the research group of Baker. The earlier versions were presented by Simons et al(28),(29). The potential is

distance dependent and are part of ROSETTA.

◇ RO —the potential developed by Robson and Osguthorpe (30).

◇ MS — a pairwise potential obtained by optimization procedure, which simutaneously maximizes the energy gap for all proteins in the database(31).

◇ MSBM — potential derived by optimize interactions, developed by Micheletti et al(32).

◇ GKS — quasi-chemical statistical potential developed by Godzik et al(33).

Table 3.2: Average Correlation for Matrices

	Data Base I	Data Base II	
Matrix	Ave. Corr.	Ave. Corr.	Group
TEl	0.739	0.835	A
MJPL	0.738	0.826	A
SJKG	0.737	0.835	A
BFKV	0.736	0.834	A
MJ2h	0.734	0.831	A
SKOa	0.734	0.836	A
TS	0.733	0.826	A
TD	0.733	0.834	A
Qa	0.730	0.837	A
MJ3h	0.728	0.838	A
TEs	0.723	0.832	A/B
BL	0.725	0.820	B
IN.TS	0.721	0.817	B
MJ3	0.720	0.824	B
BT	0.718	0.823	B
HLPL	0.718	0.830	B
VD	0.718	0.815	B
B1	0.715	0.820	B

Table 3.2: Average Correlation for Matrices (Continued)

Matrix	Data Base I Ave. Corr.	Data Base II Ave. Corr.	Group
RO	0.714	0.817	B
SKOb	0.711	0.830	B
Qm	0.709	0.829	B
MS	0.709	0.820	B
IN.MJ2	0.707	0.813	B
Qp	0.703	0.830	B
B5	0.697	0.818	B
B2	0.694	0.819	B
N.MJ2	0.687	0.789	C
MSBM	0.687	0.801	C
N.TS	0.685	0.783	C
GKS	0.670	0.800	C
B4	0.668	0.797	C
B3	0.663	0.801	C
MJ1	0.634	0.767	D

CHAPTER 4. Addition of Blosum Matrix in Scoring Function

4.1 Introduction of Blosum Matrix

It is well known that certain amino acids can substitute for one another in related proteins, because they have similar physiochemical properties. One example of this "conservative substitutions" is isoleucine for valine. Both of them are small and hydrophobic amino acids. Another example is serine for threonine, and both are polar. When calculating the sequence alignment scores, identical amino acids should be greater than substitutions and conservative substitutions should be greater than nonconservative cases. These relationships are explicitly represented by BLOSUM substitution matrices, derived by Henikoff,S and Henikoff, J. G(34).

The BLOSUM substitution matrices have been constructed similarly with the point-accepted-mutation (PAM) model of evolutionary generated by Dayhoff et al.(35), but different strategy was applied for estimating the target frequencies. In Dayhoff model, one PAM is a unit of evolutionary divergence with the 1% of amino acids being changed, but 100 PAM doesn't lead to a totally different amino acid sequence. Because in some positions, the amino acids change multiple times, and even turn back to the original amino acids; in other positions the amino acids don't change at all. The protein sequences are at least 85% identical. Taking the changes to be completely random, the substitution frequencies can be determined by the frequencies of the different amino acids (background frequencies). While the substitution frequencies in the related proteins (target frequencies) are only composed of the substitutions that maintain the protein function, so they can carry on in the evolution. The scores for each substitution pair are proportional to the natural log of the ratio of target frequencies to background frequencies. The frequencies for derive BLOSUM matrices are from a data base of

blocks, which are generated by An automated system, PROTOMAT(36). The method allow the substitution frequencies to be bigger and sequence identity to be less. BLOSUM62 require sequences having at least 62% identity. We choose this matrix adding in our scoring function instead of BLOSUM30 for highly divergent sequences and BLOSUM90 for very similar sequences.

4.2 Modification of Scoring Function

In our scoring function, the contact energy score and secondary structure score are both considered. The secondary structure score is also a discriminative score. We use a 'global fitness' factor f to represent the matches between the predicted secondary structure of the target sequence and the secondary structure of template structure in PDB. The fitness factor f is defined as

$$f = (N_+ - N_-)/N_s, \tag{4.1}$$

where N_+ is the total number of matches , N_- is the total number of mismatches, and N_s is the total number of residues in threaded structure selected from the alignment. The secondary structure prediction for the query sequence is obtain from the consensus of three secondary structure predictors, PSIPRED, PROF and SAM. The contributions of contact energy score and secondary structure score are calculated as following:

$$E_{tot} = (1 + \alpha f)E_c, \tag{4.2}$$

Where E_c is the contact energy score defined by equation 3.1; α is a constant number determining the relative weight of these two scores for the total energy score. The decisive energy score is the 'relative score' E^{rel} defined by:

$$E^{rel} = E^{raw} - E^{ave}, \tag{4.3}$$

Where E^{raw} is the result of the query sequence threaded on the template calculated by equation 4.2. E^{ave} is the average score obtained by randomly shuffling the query sequence, threading on the template structure calculating.

Sequence information was add by modifying equation 4.2 as follows:

$$E_{tot} = (1 + \beta b) \times (1 + \alpha f) E_c, \tag{4.4}$$

Factor b is defined by:

$$b = \frac{\sum_{k=1}^{m} B_{i(k)j(k)}}{\sum_{k=1}^{m} B_{i(k)i(k)}}, \tag{4.5}$$

Where m is the length of the aligned sequence, so m is less than or equal to the length of the query sequence. k is the index of the aligned sequence. $i(k)$ is the index of the residue in the query sequence which is at the position k in the aligned sequence. $j(k)$ is the index of the template sequence which is the k^{th} aligned residue in the template sequence. So $B_{i(k)j(k)}$ is the elements in the BLOSUM62 matrix for residue i and residue j. the $B_{i(k)i(k)}$ is the sum of the blosum score when the query sequence aligned to itself. The range of factor b is $-1 < b \leq 1$. Because modulus of the off diagonal elements in the BLOSUM62 matrix are less than or equal to the diagonal elements, factor b can't reach -1. β is the relative weight constant.

4.3 Results

We select eleven CM targets from data base I and repeat the procedure of prediction with the original code and the new code with addition of blosum score. The weight constant β is set as 0.2. Nine out of eleven targets have enhanced average correlations.

Target	Original	Blosum	Enhance
T0196	0.388	0.412	√
T0205	0.479	0.590	√
T0211	0.906	0.922	√
T0231	0.913	0.935	√
T0234	0.864	0.871	√
T0240	0.761	0.768	√
T0265	0.739	0.708	
T0267	0.868	0.878	√
T0271	0.773	0.652	
T0275	0.913	0.925	√
T0277	0.640	0.714	√

In the above test, the decoy structures have nearly the same sequence with the query sequence, actually the diagonal elements are used in the calculation. In order to test the performance of the whole BLOSUM62 matrix, we do the whole database, containing 13391 protein structures, search for the query sequence. This is exactly what we did in the CASP season. Here we repeat the whole database search three times: one is for original code; two are for scoring function with blosum score while the constant β are 0.2 and 0.7 respectively. In the following table the results are compared. The constant *beta* as 0.7 outperforms the value of 0.2, so we only list the results when *beta* is 0.7. The Z score is the evaluation score provided by DALI (Distance Matrix Alignment) which was used to evaluate the predictions in CASP5. The LGA score was applied to evaluate the CASP6 results by Michael et al. (37). Column two and six are the PDB template provided by Dali and LGA scores respectively. Column four and five are the ranks for each protein template for original scoring function and new scoring function with sequence information. It is obvious that the blosum score does help to improve the prediction.

Table 4.1: Comparing original score with new score

Target	PDB	Z Score	Original	blosum 0.7	LGA	%ID	LGA score
t196			209,237	344,682	1skqA	32	85.5
	1sywA	11.6				22	80.1
	1g7rA	11.6		3123		15	77.9
	1efcA	11.6	69, 83	7,8,16		28	78.2
t205			4,8,14	1,3,5,6-12	1h0yA	24	76.5
	1h0xA	10.5	5,255	2,12,32		26	76.2
	1tig	7.6	236	568		11	52.5
t211	1eut	17.7		8,10,12,13	1eut	22	84.8
	1gof	16.8				14	84.6
	1o59A	16.3	22,27	34,38,46		18	69.3
	1jhjA	15.4	2,4,5	1,2,3,4,5		12	73.1
	1cztA	14.0		338,453,483		11	76.7
	1xnaA	13.5				12	63.4
t231	1v6fA	22.2	1,2,3...	1,2,3...10	1v6fA	80	95.2
	1cof	19.4		25,30,31		15	88.8
	1m4jA	15.7		52,60		18	76.8
t234			569,587	431,455	1g76A	16	63.4
	2arzA	17.1		352		26	86.3
	1rfeA	14.0		1,2,3,4,5,9		18	71.2
t240			581,623	2,5-17	1ihrAB	94	61.5
	1lr0A	5.8	138,350	686		15	59.7
t265			28,72,74	6,31,43	1ku9B	29	67.9
	1mkmA	9.4	4559	5202		21	61.4

Table 4.1: Continued

Target	PDB	Z Score	Original	blosum 0.7	LGA	%ID	LGA score
	1qbjA	9.3	2313C	913		18	58.0
	1bjaA	9.3	1,2,3...	1,2,4,7,16		17	60.6
	1ku9A	9.2	6,16	9,12,14		28	69.6
t267			10,11,12	16,20,37	1j4jB	20	76.1
	1tiqA	18.2	3,4,5...	1,2,3,6,7,9		18	75.5
	1vhsA	16.5	1,2	4,5,8,12,14		23	62.8
	1s3zA	16.2				18	63.4
t271			1-5,8,12	1-10	1rlhA	41	80.0
	1psdA	6.3				12	34.0
	1ygyA	5.7				8	32.3
	1m1gA	5.5				18	30.6
t275	1mjh	16.5	1,2,5...	1,2,3...	1mjhA	29	74.5
	1jmvA	11.2	3,9...	18,21,28,30		24	57.8
t277			86,93	14,20,30	1jogD	29	92.5
	1wwpA	18.5	12,45,99	1-8,10,11		40	94.5
	1jogA	12.9	176,236	9,15,19,25		29	91.8

CHAPTER 5. Summary and Future Work

5.1 Summary

We found two schemes that can effectively improve the scoring function. We replace the LTW q vector with the a CP matrix in group A. Assessment of blosum score can also improve scoring function to select out the right template for the native structure.

5.2 Future Work

Our threading approach gives the scores based on the knowledge of contact potential, secondary structure prediction and spatial compactness of the alignment. The secondary structure score is also a discriminative score. The secondary structure prediction for the query sequence is obtain from the consensus of three secondary structure predictors, PSIPRED, PROF and SAM. Although the predictors provide about 80% correct predictions, the 20% of unknown or not correct secondary structure prediction still affects the secondary structure score. So we want to use backbone potential to replace the secondary structure score to yield better performance without the limitation of the secondary structure prediction.

Finally, we will combine all the enhanced score schemes in the scoring function, such as contact potential matrix, blosum score and backbone potential (if desired). Then we will do different kinds of testing for CM targets, FR targets and NF targets to adjust the parameters for the relative weight of these scores and ensure the new scoring scheme do yield obvious improvement.

BIBLIOGRAPHY

[1] Essential cell biology: an introduction to the molecular biology of the cell [Bruce] Alberts ... [et al]

[2] Anfinsen C. Science 1973;181:223.

[3] Nozaki, Y; Tanford, C. *J.Biol.Chem.* 1971,246,2211-2217.

[4] Finney,J.L.; Gellatly,B.J.; Golton, I.C.; Goodfellow, *J.Biol.phys.J.* 1980,32,17-33.

[5] Sanzo Miyazawa; Robert L. Jernigan. *Macromoecules* 1985,18,534-552.

[6] Miyazawa S, Jernigan RL.J Mol Biol 1996;256:623-44

[7] Miyazawa,S. *Biopolymers* 1983,22,2253-2271.

[8] Hao Li,Chao Tang, and Ned S. Wingreen. *PhysicalReviewLetters* 1997;79:765-8.

[9] Lathrop RH, Smith TF. *JMolBiol* 1996;255:641-65.

[10] Haibo Cao, Yungok Ihm, Cai-Zhuang Wang, James R. Morris, Mehmet Su, Drena Dobbs, Kai-Ming Ho. (2004). *T*hree-dimensional threading approach to protein structure recognition. *P*olymer *45*, 687–697.

[11] Bryant SH, Altschul SF. Curr Opin Struct Biol 1995;5:236-44

[12] Meller J, Elber R. Proteins 2001;45:241-61.

[13] Piotr Pokarowski, Andrzej Kloczkowski, Robert L. Jernigan, Neha S. Kothari, Maria Pokarowska, and Andrzej Kolinski. Proteins 2005;59:49-57.

[14] Tobi D, Shafran G, Linial N, Elber R. On teh design and analysis of protein folding potentials. Proteins 2000;40;71-85.

[15] Park B, Levitt M. Energy functions that discriminate the X-ray and near-native folds from well-constructed decoys. J Mol Biol 1996; 258: 367-392.

[16] Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding: When is the quasichemical approximation correct ? Protein Sci 1997;6:676-688.

[17] Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3-16.

[18] Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most representative structures in the Protein Data Bank. Proteins 2001;44:79-96.

[19] Myazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures—quasi-chmical approximation. Macromolecules 1985;18:534-552.

[20] Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;2556:623-644.

[21] Miyazawa S, Jernigan RL. Self-consistent estimation of interresidue protein contact energies based on an equilibrium mixture approximation of residues. Proteins 1999;34:49-68.

[22] Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules1976;9:945-950.

[23] Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci USA 1996;93:11628-11633.

[24] Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A. Protein fragment reconstruction using various modeling techniques. J Comput Aided Mol Des 2003;17:725-738.

[25] Byrant SH, Lawrence CE. An empirical energy function for threading protein-sequence through the folding motif. Proteins 1993;16:92-112.

[26] Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;8:361-369.

[27] Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;109:11101-11108.

[28] Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209-225.

[29] Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82-95.

[30] Robson B, Osguthorpe DJ. Refined models for computer-simulation of Protein folding—applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin-inhibitor. J Mol Biol 1979;132:19-51.

[31] Mirny LA, Shakhnovich EI. How to derive a protein folding potential?: a new approach to an old problem. J Mol Biol 1996;264:1164-1179.

[32] Micheletti C, Seno F, Banavar JR, Maritan A. Learning effective amino acid interactions through iterative stochastic techniques. Proteins 2001;42:422-431.

[33] Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids?: analysis of energy parameter sets. Protein Sci 1995;4:2107-2117.

[34] Henikoff, S. Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Siciences of United States of America 89, 10915-10919.

[35] Dayhoff M.O., Schwartz R. Orcutt B.C. (1978) Atlas of Protein Sequence and Structure. Vol.5. Suppl. 3, 345-358.

[36] Henikoff,S. Henikoff, J. G. (1991) Nucleic Acids Res. 19, 6565-6572.

[37] Michael Tress, Chin-Hsien Tai, Guoli Wang, Jakes Ezkurdia, Gonzalo Lopez, Alfonso Valencia, Byungkook Lee, Roland L. Dunbrack, Jr. Domain Definition and Target Classification for CASP6. Proteins 2005 Accepted.

ACKNOWLEDGMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, I am indebted to the major professor in my graduate career: Dr. Kai-Ming Ho and Dr. Cai-Zhuang Wang. I am grateful to them for their guidance, patience and support throughout this research and the writing of this thesis. I learned many things from both of them, whose vision, wisdom and impeccable insights have been and will continue to be an inspiration to me. They believe in me and guided me through the most difficult times throughout this work.

I would also like to thank my committee member, Dr. Drena L. Dobbs, Dr. E. Walter Anderson and Dr. Edward Yu, for their valuable comments and timely flexibility. I would additionally like to thank for Tzu-Liang Chen and Ming Li, their kind help and suggestions. Without them, this project would not have been completed as smoothly and successfully as it has been.