Comparative Study of Hydrophobic-Polar and Miyazawa-Jernigan Energy Functions in Protein Folding on a Cubic Lattice Using Pruned-Enriched Rosenbluth Monte Carlo Algorithm

Timothy Chan
University of British Columbia, Canada
Department of Computer Science
timchan@cs.ubc.ca

Bojana Jankovic
University of British Columbia, Canada
Department of Experimental Medicine
bojana_jankovic@shaw.ca

Viet Le
University of British Columbia, Canada
Department of Computer Science
viet@biovires.com

Igor Navernouk
University of British Columbia, Canada
Department of Computer Science
igor@cs.ubc.ca

March 15, 2004

Abstract

In this analysis of the contact energies guiding the protein folding, the performance of the PERM algorithm on a simple, cubic lattice is examined when Miyazawa-Jernigan (MJ) and Hydrophobic-Polar (HP) energy matrices are applied. Geometric similarity of minimum energy conformations of twenty proteins, generated when HP and MJ are used, is determined by the Root Mean Square Difference (RMSD) and the fraction of Common Contacts (CC). RMSD is a measure of global similarity, while CC indicates local similarity. On average, RMSD is one lattice edge, which is the resolution of the model. More contacts are formed when the MJ energies are used, as opposed to HP energies. This difference is attributed to the greater likelihood of contact formation between two polar residues in the MJ model. The local and global differences in conformations predicted by HP and MJ, increase with protein length, and decrease with hydrophobic fraction. A novel parameter, protein ”foldability”, is introduced. It is a measure of the simplicity of folding a protein. RMSD has a negative correlation with foldability. Conversely, CC demonstrates a positive, linear correlation with foldability. For highly foldable proteins, HP and MJ energy functions predict both globally and locally similar structures, while the opposite is true of low foldability proteins. Moreover, the search for conformation with optimal energy is more efficient for foldable proteins. For proteins with low foldability, the use of known secondary structure facilitates the search for lowest energy structure. For all twenty proteins, our method identified conformations with lower energy states than those previously reported.

Key words: Lattice models, protein folding, Monte Carlo, Miyazawa-Jernigan, Hydrophobic-Polar Model, PERM

1 Introduction

Proteins are modular chemical units that have a wide variety of functions in intra- and extra-cellular environments. They are principal structural components of the extracellular matrix and the cytoplasm, they act as enzymes, and they assume the role of regulators for most metabolic and signaling processes. All of these functions are directly dependent on the protein's structure.

1.1 Background

Investigations of protein folding are motivated by development in many areas of life sciences. In order to obtain precise structural information of sequenced coding regions obtained from the Human Genome Project and to advance structure-based drug design, a reliable method for prediction of the three-dimensional structure of proteins is necessary. Furthermore, elucidation of protein folding mechanisms is essential for developing therapy for health conditions caused by misfolded proteins. Premature degradation and aggregate formation are both consequences of protein misfolding, and result in loss of function. Alzheimer's disease, amyloidosis (deposition of protein materials), osteogenesis imperfecta (disease resulting from a mutation in collagen) and many types of cancer [17] are consequences of protein misfolding.

Answering the question ”How does a protein fold?” will not only elucidate one of the most puzzling phenomena in molecular biology, but it will also enable us to predict protein's sites of interaction and activation domains. As a result, a plausible hypothesis regarding a protein s mechanism of action, as well as its function can be postulated. These results can provide insight into signaling pathways, protein-protein interactions, and drug target sites. Furthemore, protein modeling is a potential tool for identifying functions of a sequenced genome segment.

1.2 Protein Modeling Methods

There are three methods currently used for protein modeling: *ab initio* modeling, homology modeling and threading. *Ab initio* modeling, the only method that somewhat resembles the folding process [19], is the one discussed in this paper. Homology modeling is used when the sequence similarity between a protein of uncharacterized structure and a protein of known structure exceeds thirty percent (reviewed in [5]). Threading is applied when BLASTing the target protein against the Protein Data Bank, the largest structural library, generates very weak or no homologs. This method is used under the assumption that there are a limited number (4000) of native folds. The probability of a particular fold can be estimated from the energy of this fold (reviewed in [5]).

Ab initio modeling approach is based on the assumption that the protein's linear amino acid sequence (LAAS) is sufficient and necessary for determining its three-dimensional conformation. The principal forces that stabilize the three-dimensional protein structure are hydrophobic forces, electrostatic forces, van der Waals interactions, hydrogen bonds, and covalent disulphide bonds formed between cysteine residues. Some of the energy functions utilized contain many approximations of these forces, while others take into account only some forces. As a result, *ab initio* protein modeling has been unsuccessful at predicting the tertiary structure of proteins. In contrast, secondary structure prediction is up to seventy-seven percent accurate [5]. One of the central problems in *ab initio* modeling is developing an energy function whose global minimum occurs when the protein is folded into its native state [11]. The contact energy matrices under investigation in this study are Hydrophobic-Polar (HP) [7], and Miyazawa-Jernigan (MJ) [13, 14] ([14] is used in this study) matrices. We explore their strengths and limitations with varying protein length and hydrophobic composition in order to determine the importance of forces characterized by both of them, and to establish in which instances they render similar predictions.

1.3 Subject of the Paper

To this day, there have not been any comparative studies between the two energy models HP and MJ. Studying the differences and similarities of these models could provide us with valuable insight into the strengths and weaknesses of each model. Furthermore, it may lead to better future protein folding analyses and a step closer to the ultimate goal of accurate protein structure prediction. In this study, we investigate the similarity of these two energy functions and their ability to predict the secondary structure of the native three-dimensional conformation of a polypeptide.

The proteins are folded on a three-dimensional cubic lattice by means of a self-avoiding walk. Using a set of 20 proteins with known structures, we apply HP and MJ energy functions to Pruned-Enriched Rosenbluth Method (PERM) [10, 2] to estimate their threedimensional conformation on a cubic lattice. For each protein, we select the conformation with the lowest energy and extract short-range interactions between residues. Next, we compare the predictions resulting from different energy models and their respective short-range residue interactions. We extend the analysis by comparing the predictions by each model, with and without known secondary structure, which is applied as a special constraint on the self-avoiding walk. Here we report the degree of similarity of structures generated by the two energy functions. Each function stresses different forces, the weight of which is reflected in the results. To determine the accuracy of each function, we selected polypeptides whose structures are known and are available in PDB.

2 Methods

2.1 Energy Functions

2.1.1 Hydrophobic-Polar (HP) Model

Under the assumption that hydrophobic interactions are the guiding force behind a protein's collapse into its globular structure, Dill developed the Hydrophobic-Polar energy model [7]. This energy model approximates the folding of globular structures by maximizing the number of interactions between hydrophobic residues and thereby packing them into a hydrophobic core while placing polar residues on the surface of the protein, where they interact with the polar solvent [7, 20]. The advantage of HP model on a cubic lattice is that it has very few global minima conformations, making it considerably easier to identify the most likely native ones according to their energy [20].

Application of HP model to PERM results in a configuration on the cubic lattice in which the hydrophobic contacts (H-H) are maximized [20]. A single contact between two hydrophobic (H) monomers results in the energy of negative one. Hydrophobic-Polar (HP) contacts and Polar-Polar (P-P) contacts have zero energy. The total energy of the structure is the sum of all H-H contacts, and is linearly proportional to the number of hydrophobic contacts. It follows that the protein s optimal conformation is one with the most hydrophobic contacts, having global energy minimum [18, 21].

Before folding the protein according to this model, the amino acid sequence is converted to hydrophobic polar (HP) sequence using a hydrophobicity table, generated as a consensus of forty existing hydrophobicity scales [22].

2.1.2 Miyazawa-Jernigan (MJ) Model

This model relies on the assumption that the average characteristics of residue-residue contacts formed in a large number of protein crystal structures directly correspond to actual differences in interactions among residues, as if there were no significant contribution from the amino acid sequence of each protein as well as intra-residue and short-range interactions [14].

Assumptions of the MJ Model:

1. Each residue is represented by the centre of its side-chain atom positions.
2. A position on a three-dimensional lattice is filled by a residue or an effective solvent molecule.
3. Contacts among residues and effective solvent molecules are defined to be those pairs within $6.5\mathring{A}$.
4. Nearest neighbors in polypeptide chains are explicitly excluded in counting contacts.
5. Coordination numbers are estimated from the mean volume of each type of residue.
6. Interactions occur only among residues and effective solvent molecules that are in contact with each other.

The model uses the quasi-chemical approximation, which regards contact pair formation as a chemical reaction from which we can obtain formulas relating statistical averages of the numbers of contacts to the contact energies. A 20 x 20 matrix of effective contact energies between all amino acid pairs, generated by Miyazawa and Jernigan in 1996 [14], is used for calculation of the contact energies.

2.1.3 Limitations of the HP and MJ Models on a cubic lattice

1. An effective solvent molecule corresponds to a group of solvent molecules whose total size equals to the average size of residue that is in contact with it.
2. Long-range interactions are ignored (*i.e.*interactions between residues that are not in contact). Although this may not present a problem in most cases, these forces may play an important role in others.
3. Chain connectivity is neglected in determination of relative values of effective contact energies.

2.1.4 Random Function

To compare MJ and HP models thoroughly, a random function was generated. This function assigns energy of negative one to a contact between any pair of amino acids. The optimal conformation is the one in which the number of interactions is maximized, regardless of the amino acid identities in the chain.

2.1.5 Using Secondary Structure Information together with HP and MJ Energy Models

For this special case of folding, when secondary structure information of a protein is known, alpha helix and beta sheet constraints were imposed on the polypeptide growth. In order to mimic the contacts in an alpha helix, two dimensional, repetitive structures, in which each residue i is in contact residue $i + 3$, were forced. To model beta sheets, exclusively forward chain steps were forced. Consensus secondary structures for each protein were obtained from five different secondary structure prediction servers [23, 24, 25, 26, 27].

2.2 Simple Cubic Lattice

The configuration of each protein is the result of a self-avoiding walk of amino acids 3 to N (where N is the number of residues in a protein, and residues one and two are grounded) on the simple cubic lattice. This lattice model is a simplified depiction of the protein's three-dimensional conformation. The amino acids are represented as spheres on the vertices of a simple cubic lattice. The angles between adjacent amino acids can be 90° or 180°, leading to a crude estimate of amino acid relative positions. The self-avoiding walk described in the next section can only step towards unoccupied positions in this lattice, and each step of the walk is one unit (one edge) long. The allowed directions of a single step in polypeptide chain growth are ($\pm$1,0,0), (0, $\pm$1,0), (0,0, $\pm$1).

Another important feature of the simple cubic lattice limits each amino acid's neighbours to at most six. Two of these are consecutive amino acids in the chain (except for the two amino acids at the end). As a result, all but end amino acids have at most four nearest neighbours. This is an underestimate compared to close-range interactions in native structures extracted from PDB, in which amino acids have up to twelve nearest neighbours. Each pair of nearest, non-bonded neighbours is considered a topological contact. The sum of energies of topological contacts within a conformation is the total energy of this conformation.

2.3 Pruned-Enriched Rosenbluth Method (PERM)

The problem of protein folding is NP-complete [3, 15], and is therefore subject to various heuristic methods, many of which are also stochastic. Up to date, algorithms such as Metropolis Monte Carlo Method [8], Core-Directed Chain Growth [4], Lattice Chain Growth [9], Genetic Algorithm [6], Hierarchical Method [19], and Constraint Logic Programming [15] have been used to solve this problem. In this study, we use a Monte Carlo strategy called PERM.

Pruned-Enriched Rosenbluth Method was developed by Grassberger et al. in 1998 [10]. This algorithm samples protein configurations according to the Gibbs-Boltzmann distribution. Its strategy is to build instances according to a biased distribution, but to correct for this by cloning ”good” and killing ”bad” configurations [10].

The most challenging aspect of polymer simulations is that, too often, it leads to entanglement. Metropolis strategies are typically not efficient at avoiding this problem, whereas PERM manages to minimize these occurrences by correcting the bias of the sample by means of assigning a weight to each biased sample.

Let Q be the Boltzmann weight of a configuration C. Then:

$$Q(C) = \prod_i e^{\frac{-E_i}{kT}}$$

Where k is the Boltzmann constant, T is temperature and the product is over all residues i, and

$$E_i = \sum_j E_{ij}$$

where j are the lattice neigbours of i which i is not bonded to, and which precede i in the chain, and

$$p(C_i) = \prod_n \frac{1}{m_n}$$

where $p(C_i)$ is the probability of sampling conformation i, and m_i is the number of free neighbours in the i^{th} step.

The partition function used to normalize the weight is:

$$Z = \frac{1}{M} \sum_i \frac{Q(C_i)}{p(C_i)}$$

where M is the number of conformations.

The weight of conformation i is then defined as:

$$W(C_i) = \left[\frac{Q(C_i)}{p(C_i)} \right] / Z$$

It is the product structure of the weight that leads to efficient sampling, which avoids entanglement: good configurations are sampled with a higher frequency than the bad ones. If the partial weight of a configuration (the weight of the structure before completion of folding) is too large or too small, the following strategies are applied. If weight of conformation C_i, $W(C_i)$, is greater the upper threshold for weight, W_+, enrichment of the structure is applied. During this process, the configuration is cloned to k copies, each carrying weight $W(C_i)/k$. Pruning occurs if weight of conformation C_i is less than the lower threshold for weight, W_-. During pruning, the configuration is killed with a probability of fifty percent, and its weight is doubled with a probability of fifty percent.

$$W(C_i) = \frac{1}{Z} \prod_{n=2}^{N-1} w_n(C_i)$$

where n represents all the residues

$$w_n(C_i) = m_n e^{\frac{-E_n}{kT}} = m_n \beta^{E_n}$$

The value β depends on temperature. Low β corresponds to low temperature (low T), and large β corresponds to high temperature (high T).

Each sequence, composed of either amino acids (for MJ model) or H's and P's (for HP model), was folded using the PERM algorithm with 1000 independent folding simulations. Energies were obtained for each resultant fold, and the structure with the lowest energy was selected as the most favourable onformation.

2.4 Polypeptide Test Set

The test set of polypeptides for this study was selected on the basis of the following criteria: 1) size of each polypeptide is not greater than 100 amino acids, 2) proteins' native structures are known and are in PDB, and 3) structural predictions of these proteins by *ab initio* methods are reported in the literature (reported in: [6, 9, 15, 19]). Consequently, an evaluation of our methods can be performed against the standard modeling methods and the native conformation. The PDB names of the proteins in the set are: 1mhu, 2mhu, 2cro, 1pgb, 1r69, 1rop, 2cro, 1le0, 1le3, 1pg1, 1zdd, 1ed0, 1kvg, 1edp, 1vii, 1e0m, 2gp8, 1enh, 1beo, 1ctf, and 1dkt-a.

The protein set composition is depicted in Figures 1 and 2. Figure 1 shows that the number of hydrophobic residues increases as a linear function of protein length, indicating that longer proteins, on average, have more hydrophobic residues. Furthermore, the hydrophobic fraction settles at approximately thirty five percent for longer proteins, as shown on Figure 2. This is representative of the hydrophobic population of the naturally occurring amino acids. According to the hydrophobicity table used in this study, seven out of twenty naturally occurring amino acids are hydrophobic. It is interesting to note that, for longer proteins, the fraction of hydrophobic residues in the chain converges to this value.

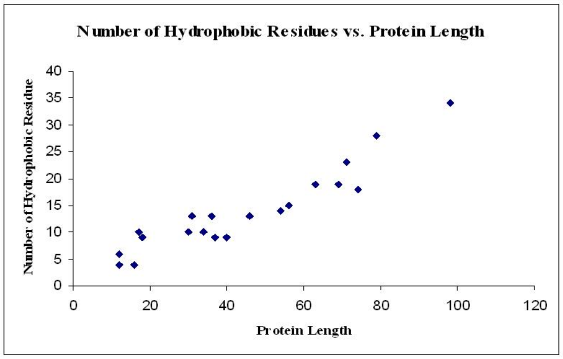

Figure 1: Number of Hydrophobic Residues vs. Protein Length

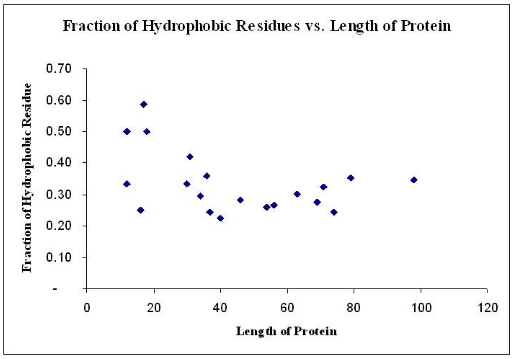

Figure 2: Fraction of Hydrophobic Residues vs. Protein Length

2.5 Analysis of Similarity

Use of quantitative methods is necessary in assessing the geometric similarities between 3D protein structures. Here, we use two such methods: one known as Root Mean Square Difference (RMSD), and another one that we introduce called Common Contacts (CC), to compare the differences between the two energy models.

2.5.1 Mean Square Difference (RMSD)

Today, one of the most accepted quantitative methods used to compare protein-folding algorithms is the Root Mean Square Difference (RMSD) [6, 9, 19]. This value represents the geometric difference between a pair of structures. RMSD in the two conformations is a collection of differences in distances (d_{ij}) between all amino acid pairs of the polypeptide examined. A large RMSD value for two structures signifies a large discrepancy between the pair. Conversely, an RMSD value of zero indicates that the structures are exactly the same. In computation of RMSD, a sum is taken over Euclidean distances (d_{ij}) between all amino acid pairs for each polypeptide, and the root of the sum of their squared differences is taken, as shown below:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

where x, y and z are the coordinates of the amino acid in the cubic lattice.

$$RMSD = \frac{1}{n}\sqrt{\sum_{ij-pairs}(d1_{ij} - d2_{ij})^2}$$

where $d1_{ij}$ is the distance between the i^{th} and the j^{th} residues obtained by the first model, $d2_{ij}$ is the distance between the i^{th} and the j^{th} residues of the same polypeptide obtained by the second model, and n is the number of amino acids in the polypeptide.

The RMSD values reported in this paper are in the units of lattice edges. The length of edge is one unit, which is equivalent to 3.8$\mathring{A}$, the typical distance between connected alpha carbons in a polypeptide.

2.5.2 Common Contacts (CC)

For a robust and complete analysis, we use a complementary method to RMSD to quantify the structural similarity. This method, named Common Contacts (CC), is a measure of local similarities between two structures. The fraction of common contacts signifies the proportion of nearest-neighbour contacts that two structures have in common. Nearest-neighbour contacts are all contacts on the lattice, which are one edge apart, excluding amino acids connected by a peptide bond. The CC algorithm works as follows:

1. Generate two n x n matrices M_1 and M_2 containing the Euclidean distances between all amino acid pairs (as in RMSD) for conformation 1 (Energy Model 1) and conformation 2 (Energy Model 2).

2. Scan the distance matrices M_1 and M_2. If the distance between residues is in the range of 0.9 to 1.2, assign that position (i, j) in the matrix a 1, else assign it a 0. In the case of M_1 and M_2 extracted from PDB files, if the distance between residues is between 0.9 and 1.44, assign the respective position in the matrix a 1. Else, assign it a 0.

3. Calculate CC where $CC = |M_1 \cap M_2|/|M_1 \cup M_2|$. The intersection and union are taken over all of the matrix cells, and $|M|$ denotes ”the number of ones in matrix M”.

The distance range 0.9 to 1.2 is chosen to account for the rounding errors generated when calculating the distances between amino acid pairs. A higher upper bound of 1.44 was chosen for PDB conformations, after the alpha carbon coordinates were normalized by division by 3.8$\AA$ (this is the distance between alpha carbons in a polypeptide). The upper bound was chosen to account for interactions between residues that occupy up to 5.47$\AA$ in the direction of interaction.

2.5.3 Energy

Proteins found in nature reach their native state very quickly during co-translational folding, indicating a ”hidden mechanism” for finding the global minimum of energy [12]. This leads us to the underlying assumption of the Levinthal Paradox: proteins are in their native conformations if and only if they have reached their global energy minimum. The ”native”, on-lattice conformation for each protein in the test set was selected according to this hypothesis. Out of one thousand resultant conformations for each protein, the one with lowest energy was selected as our prediction of the native conformation. Upon normalization of the energies rendered by the HP and MJ models, energy comparisons were performed between these two models, and between each model and models used by other groups.

2.5.4 Analysis of Geometric Similarity in Proteins Folded Using HP or MJ with and without Knowledge of Secondary Structure

Structural analyses were applied between three different groups of models (please see below). Those belonging to group one quantify similarity between HP and MJ with respect to the random function. The ability of HP and MJ energy models to predict secondary structures of proteins is inferred from the second group of results, while the accuracy of prediction of our models is derived from group three.

For group one, we generate the RMSD’s and the CC’s for all twenty proteins. For group two, we generate RMSD’s and CC’s for fifteen proteins, since for five of them, no secondary structure prediction was obtained from the servers used. For group three, only three proteins were used for analysis because folding on a simple cubic lattice does not result in conformations similar to native states, making the comparisons uninformative. This is in part due to the crudeness of the model, and to the nature of the energy functions MJ and HP [9, 20]. In the results section, we show representative plots of these analyses.

```
1. MJ vs. HP:
    1.1 MJ vs. HP
    1.2 MJ vs. Random
    1.3 HP vs. Random
2. Secondary vs. Non-secondary constraints:
    2.1 MJ2 vs. HP2
    2.2 MJ2 vs. MJ
    2.3 HP2 vs. HP
3. MJ and HP vs. Native Folds (obtained from PDB):
    3.1 MJ vs. PDB
    3.2 HP vs. PDB
```

Here, MJ2 refers to the MJ energy model used together with secondary structure constraints to simulate folding; HP2 refers to the HP energy model used together with secondary structure constraints to simulate folding.

3 Results

3.1 Optimizing PERM for MJ Energy Function

For the HP energy model, the optimal value of β, the temperature parameter, is 18. In an attempt to refine the MJ model, optimization of the folding algorithm was tested at different values of β. The dependence of energy of conformations on β was observed as shown in Figure 3.

After exploring the energy landscape using different values of β (ranging from 0.01 to 1000), for all proteins, we arrived at the following solutions. The optimal value of β may vary over two orders of magnitude with respect to the intrinsic characteristics of a polypeptide. In order to maintain the consistency and generality of PERM throughout the folding simulation, we developed an algorithm that generates random values of beta in the range of 10-40 for each independent run. In addition, we increased the number of independent runs to 1000, in order to increase the probability of sampling a conformation whose energy is a global minimum.

3.2 Mean Square Difference RMSD

3.2.1 RMDS(Length)

The polypeptide length is a major determinant of the quality of the algorithm's prediction. The shorter the chain, the smaller the size of the search space, and hence the fewer steps are taken during the random walk. As a result, a larger fraction of possible conformations is sampled, and the probability of accurate prediction is higher. Graphs of RMSD vs. Length comparing various models (HP vs. MJ; HP2 vs. MJ2; HP2 and HP, MJ2 and MJ) indicate that when the length of the protein is less than 55 residues, RMSD is below 1 for the structures obtained by these energy models. Conversely, when the length of the polypeptide is greater than 55 residues, the RMSD values vary considerably (between 1 and 3, on average). Figures 4 and 5 show two

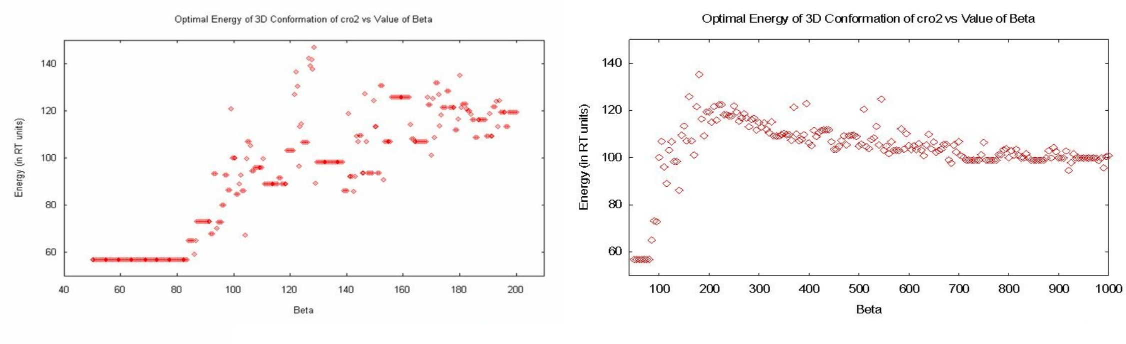

Figure 3: Refining the folding algorithm using the Miyazawa-Jernigan effective contact energies resulted in observation of the following relationship between negative optimal energy of conformation and β for cro2 protein.

a) Varying the value of β from 50 to 200, in increments of 0.5, results in the negative optimal energy values on the graph above obtained by performing 20 independent runs on the protein cro2 at a given β value. The best energy of the 20 runs increases as a function of β for beta greater than 85.

b) Varying the value of β from 50 to 1000 in increments of 5, results in the above negative optimal energy values above obtained by performing 20 independent folding simulations for each value of β. The conformation with optimum energy is found when β is 240.

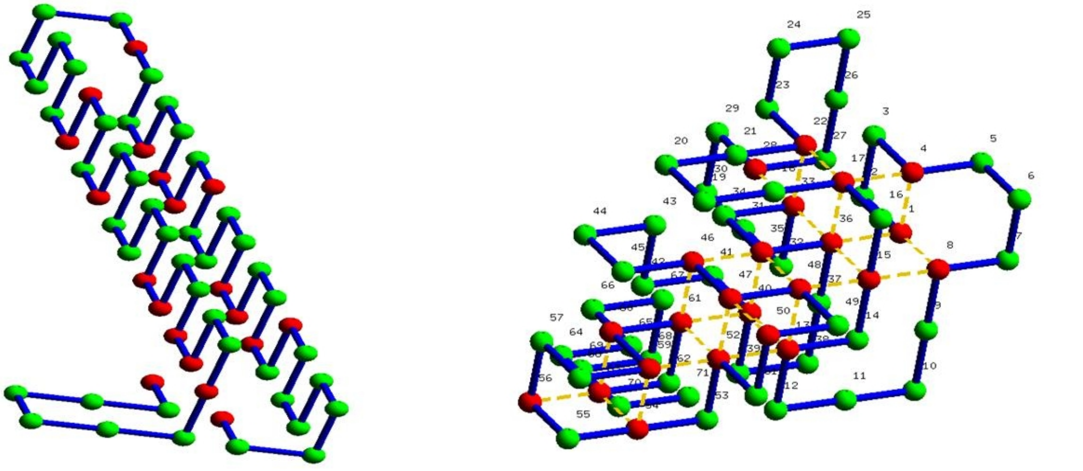

Figure 4: Two dissimilar structures of protein 2cro are produced when HP model is used together with secondary structure information (left) and when HP model is used with no secondary structure information (right). Note the two alpha helices in the structure on the left.

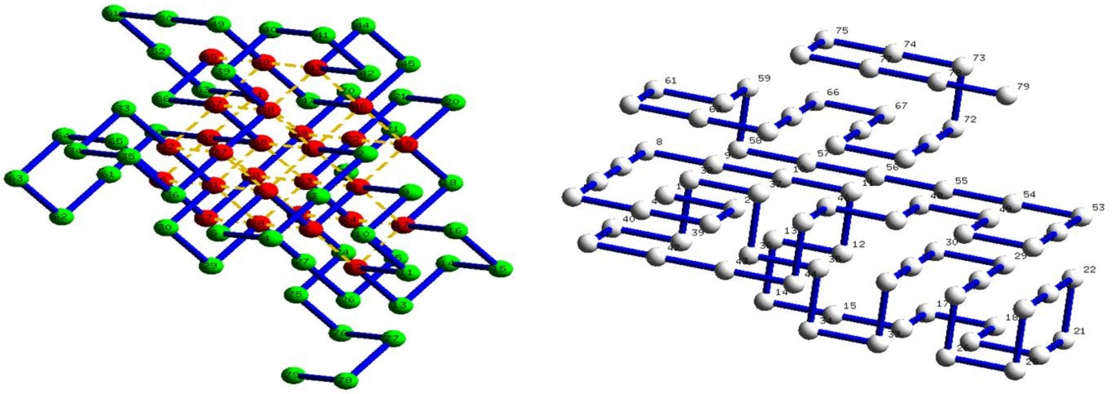

Figure 5: Two structures formed when protein 1DKT-A is folded using HP (left) and MJ (right) energy models. The conformation resulting from HP maximizes the hydrophobic contacts (red spheres), and is more compact than the structure generated by MJ.

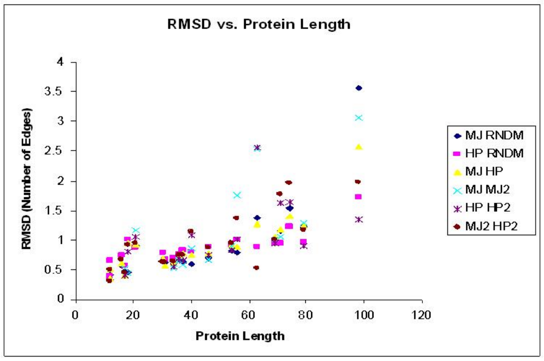

Figure 6: This graph depicts RMSD vs. Protein Length. Note that in comparisons of structures generated by all models, RMSD increases as a function of chain length.

conformations of proteins 2CRO (generated by HP and HP2), whose length is 71 amino acids, and 1DKT-A (generated by HP and MJ), whose length is 79 amino acids.

In Figure 6, an exponential relationship is observed between RMSD's and protein length, for all combinations of energy models. Although the trend is the same for all models when protein length is below 55 residues, RMSD's between certain pairs of models (MJ, RNDM and MJ, MJ2) have a considerably higher rate of change with respect to protein length, than RMSD's between other pairs of models (HP, RNDM). Large proteins are folded more similarly by HP and Random (whose RMSD of all proteins is close to one), than they are by MJ and HP, or MJ and Random (Figure 6). This is likely due to similar assignment of contact energies by HP and Random models. Both functions assign energy of negative one per favourable contact. While Random counts all contacts, Hydrophobic-Polar is specific to hydrophobic-hydrophobic contacts. As a result, the weights of the conformations differ only slightly and are not as diverse for these two models as they are for MJ.

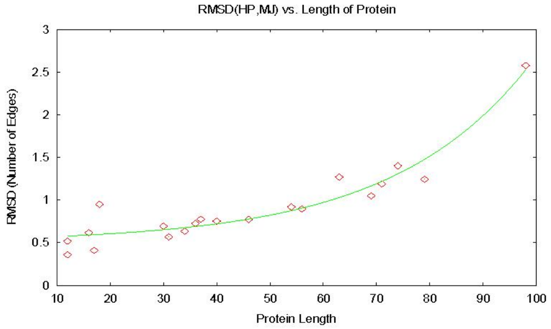

Figure 7: RMSD between MJ and HP as a function of Protein Length. This relationship is fitted to the function: $f(x) = 0.04e^{0.04x+0.51}$.

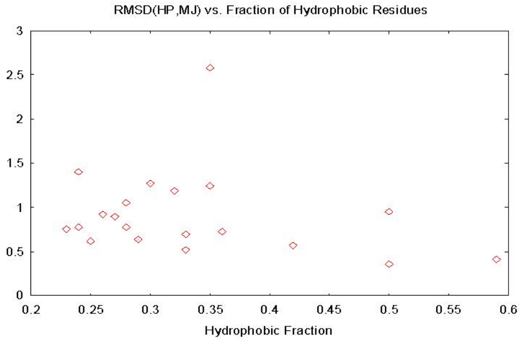

Figure 8: Relationship between RMSD's and the Fraction of Hydrophobic Residues

Figure 7 depicts RMSD dependence on protein length for HP and MJ models. These two energy models assign different contact energies to amino acid pairs. MJ makes a large distinction between each potential contact (the empirical potential energies between pairs of amino acids in the MJ matrix range from -7.37 to 0.14). As observed in Figure 7, it follows that for longer chains, more contacts are formed, and this difference becomes more prominent.

3.2.2 RMSD(Hydrophobicity)

In Figure 8, RMSD is plotted against hydrophobic fraction for HP vs. MJ. From corresponding plots of all the energy model pairs, we define the threshold value of hydrophobic content for which the RMSD value is less than one. The reason this value is of interest is that one edge length is the resolution of our simple cubic lattice, and an RMSD below 1 indicates that the atoms are, on average, taking the same positions in the two models. Interestingly, our analyses show that RMSD for all pairs of models is less than 1 when the proteins hydrophobic fraction is greater than 35%. When, on the other hand, the hydrophobic content is less than 35% it is more difficult to find the optimal energy, since interactions are infrequent, and RMSD values diverge from one. No

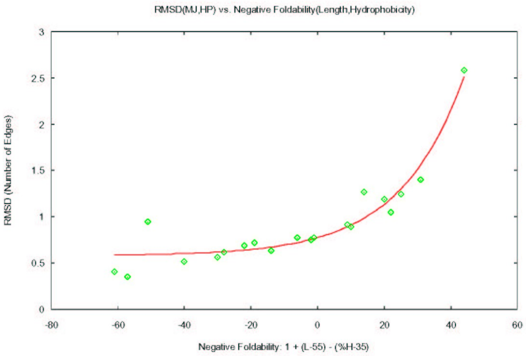

Figure 9: RMSD vs. Negative Foldability indicates an exponential relationship with a positive slope between the two parameters, which is fitted to a function $f(x) = 0.20e^{0.05x} + 0.58$.

relationship is observed between RMSD and hydrophobic residue distribution (data not shown).

3.2.3 RMSD(Length, Hydrophobicity)

Our results show that both the protein length and hydrophobic content influence the folding prediction by the algorithms. For longer proteins, the search space increases exponentially, indicating that a very small fraction of the sample space is explored. Thus, the results may actually be those of a local minimum. Similarly, for proteins that have low hydrophobic content, the energy function does not assign preference to one step in relation to (at most) four other steps. The weights of conformations converge as protein length increases, making it difficult to distinguish between ”good” and ”bad” structures.

Since both length and hydrophobic fraction appear to play important roles in the accuracy of the folding prediction algorithms, combinations of the two parameters were plotted against RMSD (Figure 9). RMSD is exponentially dependent on: $[1 + Length - 55 - (\%H - 35\%)]$. This term, which we named ”Negative Foldability” (hereafter referred to as NF), is negative when the chain is short and has a high hydrophobic fraction (*i.e.*, when the chain is ”Foldable”). NF takes on high positive values when the chain is long and has a low hydrophobic fraction.

3.2.4 Mean RMSD

A histogram of mean RMSD values for each model pair was generated in Figure 10. The average RMSD over twenty proteins is representative of the overall similarity of two models. Mean RMSD varies between 0.87 (for HP, Random) to 1.17 (for MJ, MJ2), indicating that all pairs of models render similar conformation predictions.

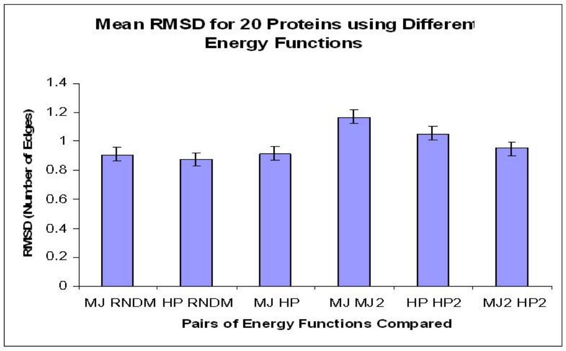

Figure 10: Mean RMSD fluctuates between 0.87 and 1.17 for all models when all 20 proteins are compared (note that only 15 proteins are compared between models taking secondary structure information).

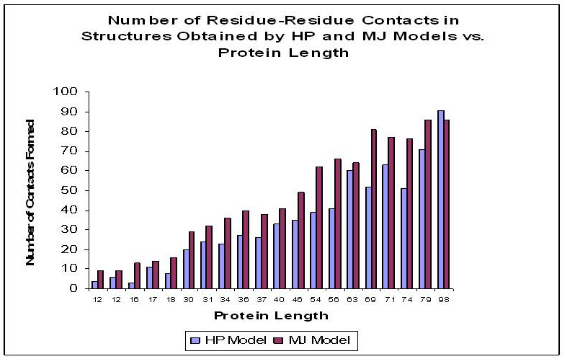

Figure 11: Number of contacts formed in predicted structures of proteins with lengths ranging from 12 amino acids to 98 amino acids. Predictions obtained by MJ and HP models are shown.

3.3 Fraction of Common Contacts (CC)

Figure 11 depicts the number of contacts formed in structures folded using MJ and HP energy functions. For most proteins, MJ outperforms HP in maximizing the number of nearest-neighbour contacts. The Hydrophobic-Polar model forms fewer contacts between amino acid pairs because of its selective properties: it assigns a negative energy only when both amino acids in contact are hydrophobic, and it disregards all other contacts. Surprisingly, HP outperforms MJ for the longest polypeptide (98 residues). This is consistent with the report that HP model has a tendency to over-fold the structure by creating more contacts than are present in the native state of a protein [15].

In Figure 12, the Fraction of Common Contacts is plotted versus protein length for structures obtained by various models. Common contacts help determine similarity in the neighbourhood of each residue, averaged over all residues. For all energy model pairs, an exponential decrease in the fraction of common contacts is observed with increasing length. Common contacts in HP2 vs. HP

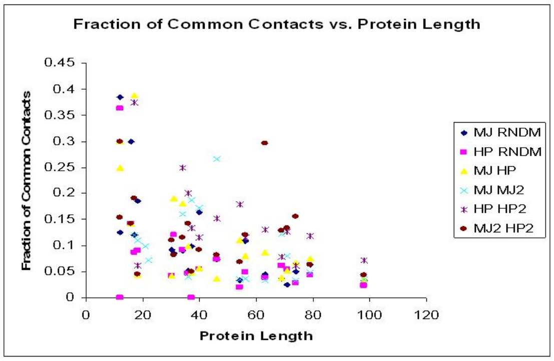

Figure 12: Fraction of Common Contacts with respect to protein length plotted for all pairs of energy models.

structures are the highest, while only a small fraction of common contacts is observed for HP vs. Random model. At first, it may seem contradictory that RMSD indicates a very good correlation between the Random model and HP energy function, while their CC are one of the lowest ones among the pairs. In Discussion, we proceed to explain why this is not contradictory.

3.3.1 CC(Hydrophobicity)

No notable relationship is observed when CC is plotted against hydrophobicity alone (data not shown). Similarly, no relationship is observed between CC and hydrophobic residue distribution (data not shown). For Common Contacts analysis, emphasis is placed on the dependence of CC on Negative Foldability.

3.3.2 CC(Length, Hydrophobicity)

The relationship between the fraction of common contacts and negative foldability is linear, with a negative slope for MJ vs. HP, MJ vs. Random, and HP vs. Random (MJ vs. HP shown in Figure 13). The fraction of common contacts never exceeds 40% for all pairs of models, indicating that the structures generated by different models with and without secondary structure constraints never reach a high structural similarity. In addition, when NF is less than zero (*i.e.*, when the protein is foldable), the fraction of common contacts is highly variable, while it remains below 15% when NF is positive (*i.e.*, when the protein has low foldability) (Figure 13).

The lowest fraction of common contacts is 3.5% for conformations obtained by using HP vs. MJ energy models. This is greater than the fraction of common contacts between MJ or HP vs. randomly generated structures (0%), suggesting that HP and MJ diverge to a lesser extent than either one of these models compared to the Random model.

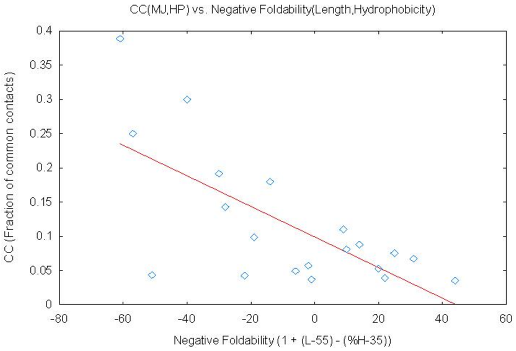

Figure 13: Fraction of Common Contacts between MJ and HP structures vs. Negative Foldability. A negative correlation is observed between CC and NF, and was fitted to the function $f(x) = -0.002x + 0.077$.

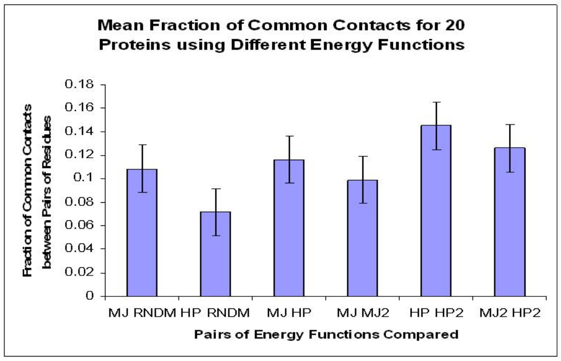

Figure 14: Mean Fraction of Common Contacts averaged over 20 proteins for all pairs of energy models (note that only 15 proteins are compared within models taking secondary structure information).

3.3.3 Mean CC

Figure 14 shows a histogram of the mean CC values. They vary from 7.2% (HP, Random) to 14.5% (HP, HP2). In contrast to mean RMSD, mean common contacts display drastic heterogeneity across different pairs of energy models. This may be due to the nature of the measurement CC provides - it quantifies local similarities, and indicates high variability over all neighbourhoods, while RMSD measures global similarity.

Table 1 shows the fraction of common contacts in structures generated by HP and MJ models with the native structures obtained from PDB. This comparison was performed only on the first three proteins as a preliminary test of structural similarity of our onlattice model to the native structure. As anticipated, a low fraction of common contacts was obtained for all three polypeptides. Gan et al. stated a similar prediction for their onlattice model [9]. The geometry of the lattice imposes certain limitations on the folding and formation of contacts, which are discussed in the next

Fraction of Common Contacts

Protein	HP vs. PDB	MJ vs. PDB
1MHU	7.89%	4.55%
2MHU	8.82%	11.90%
1PGB	7.30%	7.53%

Table 1: Fraction of common contacts between HP and MJ models versus native structures obtained from PDB.

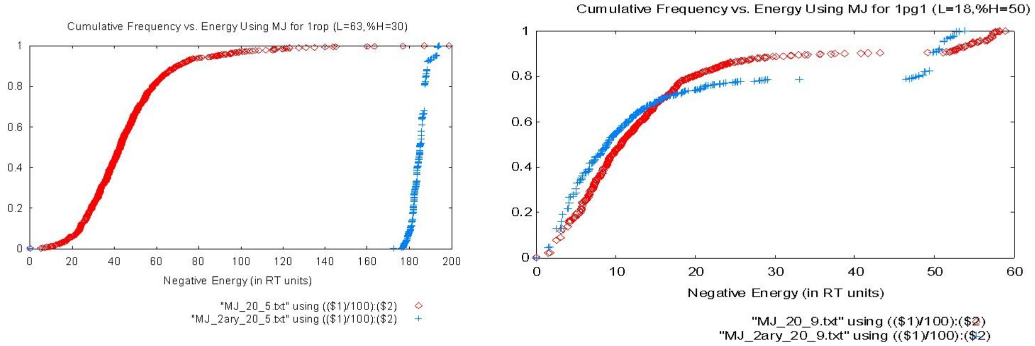

Figure 15: Cumulative Frequency plot of negative energy with and without knowledge of secondary structure for 1rop (left) and 1pg1 (right). Secondary is shown in blue, and non-secondary in red. Note that 1rop is a long polypeptide with average hydrophobic fraction, while 1pg1 is short and has above average hydrophobicity.

section.

3.4 Energy

3.4.1 Energy Distribution among Independent Runs

To study the energy values that appeared most frequently during the execution of the PERM algorithm, we ran the program using the MJ model as described in section 3.1. We then computed the distribution of the optimal conformation energy values and plotted the corresponding cumulative distribution by treating the optimal energy after one independent run as a random variable.

We use the Miyazawa-Jernigan model as it is considered a more realistic representation of the forces guiding the folding process. We then plot the cumulative frequency of energies for each protein, obtained with and without the knowledge of secondary structures. These plots are useful as they shed light on the energies of conformations that are yielded by the algorithm.

For most proteins, one can see a large shift toward lower energies (greater negative energies) in the cumulative frequency plot when the energy function is used with secondary structure information vs. when it is used alone (Figure 15(left)). In contrast, the cumulative frequency plots are the

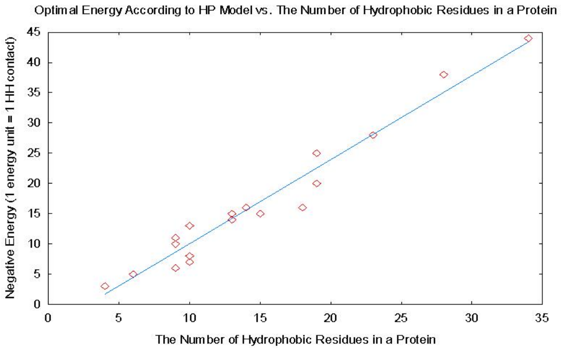

Figure 16: Negative Hydrophobic-Polar Energy vs. the Number of Hydrophobic Residues in a Protein.

same for short proteins with a high hydrophobic fraction (with length less than 18 amino acids, and hydrophobic fraction greater than 50%), when secondary structure is known and when it is not known (Figure 15(right)).

3.4.2 Energy and Protein Characteristics

We used the minimum Miyazawa-Jernigan and Hydrophobic-Polar energies and plotted them against several parameters that define proteins: hydrophobic fraction, the number of hydrophobic residues, and the length of the protein. No correlations were observed with the protein test set other than that of minimum negative Hydrophobic-Polar energy and the number of hydrophobic residues in a protein. This linear relationship, shown in Figure 16, is fitted to the function $f(x) = 1.4x - 4$, suggesting that, on average, for each hydrophobic residue added to the polypeptide, the energy decreases by 1.4 units (*i.e.*, the number of hydrophobic interactions increases by 1.4). This regularity is remarkable, and it would be worthwhile exploring if it is preserved in longer proteins (length of hundreds, or thousands of amino acids). The function, $f(x)$, is a reliable indicator of the efficiency of the algorithm and the suitability of the energy function for longer proteins.

3.4.3 Minimum Energy Comparison

Minimum energy values obtained by our folding simulations were compared to those of other groups that used the same proteins as in our test set. In Table 2, negative energies are shown, which were normalized with respect to the contact potential matrix used. Palu et al. used Amino Acid Empirical Contact Energies matrix [15].

For short proteins, such as 1LE0 (12 residues), 1KVG (12 residues), and 1EDP (17 residues), normalized energies of conformations predicted by Palu et al. [15] are similar to normalized energies of conformations predicted in this study. For longer chains, however, minimum energy is considerably lower for conformations predicted by our group compared to energies of conformations

Negative Normalized Energies (MJ matrix, HP values and AA Empirical Contact Potentials matrix)

Protein	MJ normal	MJ2 normal	HP normal	Palu et al.
1PGB	0.588	0.406	0.306	N/A
1R69	0.751	0.674	0.408	N/A
1LE0	0.077	0.076	0.061	0.052
1LE3	0.112	0.111	0.061	0.078
1PG1	0.172	0.155	0.122	0.069
1ZDD	0.344	0.344	0.265	0.057
1ED0	0.434	0.421	0.286	0.188
1KVG	0.100	0.098	0.102	0.088
1EDP	0.179	0.160	0.143	0.128
1VII	0.416	0.392	0.306	0.145
1E0M	0.317	0.317	0.224	0.159
2GP8	0.364	0.289	0.184	0.086
1ENH	0.547	0.539	0.327	0.119

Table 2: Comparison of negative normalized minimum HP and MJ energies with negative normalized minimum energies obtained by Palu et al. [15] for 11 proteins.

obtained by Palu et al.

To complement the observations noted for minimum MJ energy search obtained for a series of independent runs, normalized MJ and HP, and MJ and MJ2 energies of short proteins with high hydrophobic content show high similarity (1LE0, 1KVG, 1EDP, Table 2). In contrast, normalized, optimum MJ and HP energies of long proteins, or proteins with low hydrophobic content, demonstrate large differences (up to two fold) (1PGB, 1R69, 2GP8 Table 2).

In addition, a more general relationship between normalized, optimum MJ and HP energies was observed (Figure 17). The normalized values of the two energies are almost identical when negative foldability is less than zero (this pertains to proteins which are easy to fold and are usually short and highly hydrophobic) as shown in Figure 18. It follows that for proteins with high values of negative foldability, optimal MJ and HP energies are more variable. This can be attributed to various insensitivities of the HP model, such as the failure to assign preference to polar-polar interaction over hydrophobic-polar interaction.

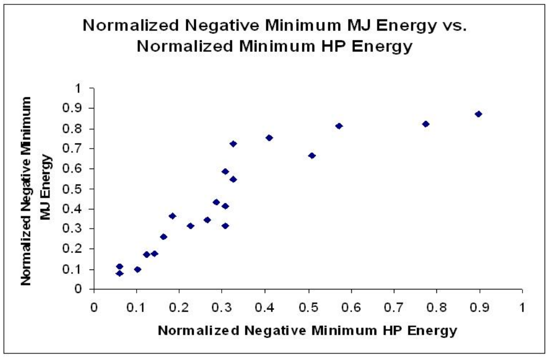

Figure 17: Correlation between normalized optimal MJ and HP energies

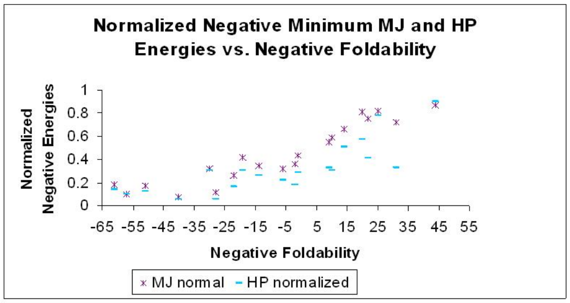

Figure 18: Correlation between normalized minimum energies and negative foldability. Note that negative foldability is a function of length and hydrophobicity.

4 Discussion

4.1 Structural Similarity Analyses and Their Implications

4.1.1 Local Characteristics

The CC metric allows for comparison of the two energy functions because high local similarity is a good indication of the energy functions rewarding same interactions between pairs of residues. Local similarity also helps draw conclusions about the way the protein folding algorithm predicts secondary structure since CC is concerned with amino acids that are in close proximity.

4.1.2 Global Characteristics

RMSD is a global measure of geometric similarity between two conformations. It is a good measure of how similar the two conformations ”look”. Local changes that involve only a few amino acids do not have a significant impact on RMSD. On the other hand, if two structures have a low RMSD, and then they likely have similar residues in the core of the protein and similar looking surfaces. Thus, we would expect two conformations with the same hydrophobic content and the same length to have a low RMSD, especially when HP model is used.

4.1.3 Protein Conformations and Domains

In vivo, both local and global characteristics are important for a protein’s function for the following reasons. Proteins contain one or more domains, which confer specificity for interacting with other proteins, as well as for localization. For instance, SH2 (Src Homology 2) domains have binding specificity to phosphorylated tyrosine residues, while SH3 domains only bind to proline-rich regions of other proteins. Furthermore, specific local and global structural features are necessary for normal protein function. Accurate prediction of both is necessary for determining protein function and sites of interaction.

4.2 RMSD

4.2.1 RMSD(Length)

An increase in RMSD as the length of the sequence increases, observed by our group, is also noted by Gan et al. [9]. They observed a doubling in RMSD values of an ensemble of conformations (compared to native states) as the peptide length doubles. Both observations confirm that the differences in contact potentials have a cumulative effect on the weights assigned at each folding step. More specifically, as the length of the chain increases, so does the heterogeneity of the conformation ensemble generated by the folding algorithm.

4.2.2 RMSD(Hydrophobicity)

There is no consistent dependence of RMSD on the hydrophobic fraction of the protein. However, there is a trend for proteins with a low hydrophobic fraction to fold into conformations whose RMSD is greater than one when different energy functions are used. In addition, secondary structure information is required for accurate folding of proteins whose hydrophobic content is less than 35%. Not surprisingly, using HP or MJ energies on most proteins, without imposing native secondary structure constraints, yields conformations whose secondary structures are different from those of native states.

4.2.3 RMSD(Length, Hydrophobicity)

The exponential relationship between RMSD and NF indicates that the difference in the conformation arises mainly due to length and hydrophobic content. The longer and less hydrophobic the protein, the greater the NF, and the greater the RMSD between conformations generated using the Miyazawa-Jernigan energy potentials and Hydrophobic-Polar energy model.

4.3 Common Contacts

The fraction of common contacts vs. Negative Foldability data are approximated by the same linear function of negative slope for comparisons of pairs of all energy models. When negative foldability is less than zero, the fraction of common contacts is highly variable, while it remains below 15% when NF is positive.

This observation may have an implication in the importance of amino acid sequence in this type of *ab initio* protein modeling. MJ potentials are highly variable between different pairs of amino acids. Consequently, this model is very sensitive to any changes to the identities of amino acids. On the other hand, the Hydrophobic-Polar model, classifies amino acids into two types. Consequently, this model is only sensitive to changes in the identities of the amino acids if the changes lead to a swap in the identity of hydrophobic and polar amino acids. The Random function does not take into account amino acid identity and is not sensitive to any changes in amino acid identity. Its only parameter is the chain length. The poor correlation between conformations obtained using HP and MJ functions, reflected in RMSD and CC results indicates that forces accounted for by MJ and HP are important factors in folding. However, when the proteins with high hydrophobic content and short polypeptide length are modeled, the two energy matrices similarly approximate the three-dimensional protein conformation.

4.4 Energy

Our data shows that in most cases, the conformations sampled during the folding process have considerably lower energies when native secondary structure constraints are imposed. A likely reason for these observations is that the sample space is greatly reduced when the knowledge of secondary structure of a protein is applied. A restriction to a subset of native-like conformations aids the folding algorithm by directing it to sample a smaller number of conformations, most of which have lower energies than a random structure. These results are in parallel with protein folding *in vivo*, where secondary structures are determined first, followed by a number of conformational changes to form tertiary structures.

5 Conclusion

Using HP and MJ energy matrices, we have shown that protein length, as well as the number of hydrophobic residues strongly influence the PERM protein folding algorithm. We have also learned that imposing a secondary structure on a linear amino acid sequence renders a much closer approximation to the natural folds with sequences that have less than 35% of hydrophobic residues. The common contacts analysis, using both MJ and HP models, suggests the importance of the different forces of attraction and repulsion that were not taken into account by the two energy models.

6 Future Work

6.1 Lattice Modification

Various cubic lattices can be used to explore folding under more realistic constraints (*e.g.*, bond angles, steric effects, and higher coordination numbers). Face-centered cubic and 311 lattices are appealing models that have been used extensively due to their high coordination numbers and bond angles which approximate those occurring in nature [15, 9]. Another alternative is the triangular lattice, reported in [1].

6.2 Algorithm Modification

6.2.1 Exploration of Limitations and Strengths of the Algorithm

It would be worthwhile to consider the limitations and the advantages of the current algorithm in more detail and in a more systematic manner. A large number of artificially generated sequences could be used as a test set that is folded by this algorithm. Subsets, composed of sequences that vary in length, hydrophobic fraction, or hydrophobic dispersion, could be used to study algorithm's performance with respect to each of these parameters.

6.2.2 Avoidance of Poorly Configured Conformations

In order to avoid conformations in which a segment of first five to ten amino acids are positioned in a straight line along the lattice, the folding algorithm could be modified in such a way that it re-folds the first protein segment after it has folded into an optimal configuration. The length of the segment and its re-folding constraints should be examined. Re-folding would present a problem if the first segment were buried in the hydrophobic core. In such cases, refolding may not result in a better conformation. Another way to deal with the problem of dangling amino acids may be to add non-native noise to assist folding [16].

6.2.3 Enhancing Folding Efficiency by Further Optimization of β

Folding efficiency optimization should be performed by testing the dependence of folding parameters (such as energy and RMSD) more thoroughly. Large test sets of proteins could be used for this analysis. The goal of this would be to determine if the parameter β can be optimized. If a single value of β does not yield an optimal solution, an optimal way to vary this parameter throughout the folding simulation should be investigated.

6.3 Foldability

Exploration of foldability (a measure of simplicity of folding a protein) is of great importance. Primary amino acid sequences are believed to have evolved in such a way that they easily assume their native fold [12]. An artificially created polypeptide generally performs much worse at folding, than a naturally occurring one. Protein composition, frequency of certain amino acids (such as hydrophobic amino acids, or cysteines), and polypeptide length should be examined in relation to this parameter. The components necessary for a protein to fold easily may provide insight to the ”hidden mechanism” of folding.

6.4 Integration of Protein Modeling Methods

Homology modeling, threading, and *ab initio* protein folding algorithms are usually applied exclusively. It would be of benefit to integrate these methods in order to gain multi-dimensional insight into possible native conformations. Use of libraries of known homologues and native folds would limit the number of plausible configurations a protein assumes by *ab initio*. Consensus folds, obtained by the three methods, could be subjected to more than one energy function as each stresses different forces, and the weights could be assigned according to consensus energies of conformations. Additionally, secondary structure should be incorporated where possible, as it limits the search space and leads to sampling of low energy conformations, as shown in this paper.

6.5 Concluding Remarks

Despite the efforts to improve the search algorithms and energy functions used in protein modeling, we are still far from understanding and modeling the folding mechanisms used by a living cell. Moreover, many crucial factors in protein folding are neglected by our approximations. Crowding effects promote, and chaperone proteins enhance folding *in vivo*. A modeling algorithm must take into account all the intricacies of this highly sophisticated system, in order to be successful. Such an algorithm is not likely to be achieved in the near future, as our knowledge of the processes in the cell is limited.

Acknowledgements. We would like to thank Alena Shmygelska for providing the code for protein modeling on a three-dimensional cubic lattice using HP with the PERM algorithm. Our thanks extend to Dr. Holger Hoos and Dr. Artem Cherkasov for very fruitful discussions.

References

[1] Agarwala R, et al.: ”Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model” *J Comput Biol*, 4(3): 275-96, 1997

[2] Bastolla U et al.: ”Testing a New Monte Carlo Algorithm for Protein Folding” *Proteins: Structure, Function, and Genetics*, 32: 52-66, 1998

[3] Berger B and Leighton T: ”Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete” *J Comput Biol*, 5(1): 27-40, 1998

[4] Beutler TC and Dill KA: ”A fast conformational search strategy for finding low energy structures of model proteins” *Protein Sci*, 5(10): 2037-43, 1996

[5] Casbon J: ”Protein Secondary Structure Prediction with Support Vector Machines” *unpublished*, www.cogs.susx.ac.uk/easy/Publications/Online/MSc2002/jac23.pdf

[6] Cui Y et al.: ”Protein Folding Simulation with Genetic Algorithm and Supersecondary Structure Constraints” *Proteins: Structure, Function, and Genetics*, 31:247-257, 1998

[7] Dill K.: ”Theory for the Folding and Stability of Globular Proteins” *Biochemistry*, 24: 1501-1509, 1985

[8] Galzitskaya OV et al.: ”Folding Nuclei in 3D Protein Structures” *unpublished*

[9] Gan HH et al.: ”Generating folded protein structures with a lattice chain growth algorithm” *Journal of Chemical Physics*, 113(13), 2000

[10] Grassberger, P., Frauenkron H., and Nadler W.: ”PERM: a Monte Carlo Strategy for Simulating Polymers and Other Things” in Monte Carlo Approach to Biopolymers and Protein Folding, eds. P. Grassberger et al. *World Scientific*, Singapore, 1998

[11] Huang ES et al.: ”Recognizing Native Folds by the Arrangement of Hydrophobic and Polar Residues” *J Mol Biol*, 252: 709-720, 1995

[12] Klimov DK and Thirumalai D: ”Linking Rates of Folding in Lattice Models of Proteins with Underlying Thermodynamic Characteristics” *arXiv:cond-mat/9805061*, v1, 5 May, 1998

[13] Miyazawa S. and Jernigan RL.: ”Estimation of Effective Inter-residue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation” *Macromolecules*, 18:534-552, 1985.

[14] Miyazawa S. and 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*, 256: 632-644, 1996

[15] Dal Pal, A., Dovier A., and Fogolari F: ”Protein Folding in CLP(FD) with Empirical Contact Energies” *In proceedings of Joint Annual Workshop of the ERCIM Working Group on Constraints and the CoLogNET area on Constraint and Logic Programming*, Budapest, Hungary 30 June - 2 July, 2003

[16] Plotkin SS: ”Speeding Protein Folding Beyond the GM Model: How a Little Frustration Sometimes Helps” *Proteins: Structure, Function, and Genetics*, 45:337- 345, 2001

[17] Radford, SE: ”What's new in protein folding? EMBO Workshop: Protein folding and misfolding inside and outside the cell” *Fold Des*, 3(3):R59-63 1998

[18] C. T. et al.: ”Mean-Field HP Model, Designability and Alpha-Helices in Protein Structures” *Physical Review Letters*, 84(2):386-389. Jan, 2000

[19] Xia Y et al.: ”Ab Initio Construction of Protein Tertiary Structures Using a Hierarchical Approach” *J Mol Biol*, 300:171-185, 2000

[20] Yue K and Dill KA: ”Forces of tertiary structural organization in globular proteins” *Proc. Natl. Acad.Sci*. USA 92:146-150, 1995

[21] Yue, Kaizhi et al.: ”A test of lattice protein folding algorithm” *Proc. Natl. Acad. Sci*, USA 92:325-329. Jan, 1995

[22] Hydrophobicity Table http://solon.cma.univie.ac.at/∼neum/software/protein/aminoacids.html

[23] BMERC server: http://bmerc-www.bu.edu/psa/index.html

[24] NNPREDICT server: http://www.cmpharm.ucsf.edu/cgi-bin/nnpredict.pl

[25] PSIPRED server: http://bioinf.cs.ucl.ac.uk/psiform.html

[26] PROF system: http://www.aber.ac.uk/∼phiwww/prof/

[27] SAM-T99 server: http://www.cse.ucsc.edu/research/compbio/HMM-apps/T99query.html