Face-centered cubic (FCC) lattice models for protein folding: energy function inference and biplane packing

Author: Allan Stewart

Adviser: Sorin Istrail, Second Reader: Franco P. Preparata

1 Abstract

Protein Structure Prediction (PSP) is a grand challenge of bioinformatics with broad implications for combinatoric optimization, computational geometry, and physical energy models. Under most accepted computational models, PSP is NP-hard, even when the chemical and physical properties of proteins are greatly simplified. The objective of PSP (also known as protein folding) is to select the molecule conformation which minimizes the energetic potential. We study real biological proteins from the Protein Data Bank in order to infer general energy functions for the protein folding problem. While the general problem is intractable, we seek to find methods of folding and mathematical principles that might suggest how to optimize *sufficiently realistic* energy functions. We will study the established HP model on the face-centered cubic lattice, conjectured by Kepler and proven by Hales to provide the "tightest" packing of identical spheres. The objective of analyzing Protein Data Bank structures is two-fold. We firstly seek the best possible energy function for this model. Secondly, we will investigate biplanar and octahedral structures in the FCC lattice which may yield novel algorithms for structure prediction.

Contents

2 Introduction

Proteins are the primary functional units of the living cell. The challenge of inferring a protein's three-dimensional structure from its sequence is known as the Protein Structure Prediction (PSP) problem. PSP is of crucial importance to the biological community, which has collected protein structures in databases since the 1960s. Chemical methods of measuring protein conformations (such as X-ray crystallography) have improved over time and contributed to the founding of large protein structure databases such as the Protein Data Bank (PDB). Nevertheless, there exist classes of proteins for which 3D structure reconstruction is impossible using these methods. In addition, scientists wish to discover first principles of protein folding ("folding" and "structure prediction" are used interchangeably throughout this paper, although folding refers to a larger body of protein-related problems). Solving the protein folding problem would enhance understanding of diseases which are caused by misfolding and allow scientists to speculate on unknown proteins (ie. drug design). For these reasons, PSP is a long-standing and elusive problem in computational biology.

Protein structure is diverse, but by no means arbitrary. Biologists classify proteins as polypeptide polymers which form a backbone chain by the dehydration of their carboxyl group. Each amino acid carries a sidechain functional group, except for the compact amino acid glycine. The *primary structure* of the protein is the sequence of these amino acid residues from amino terminus to carboxyl terminus. The primary structure is thus a string over the twenty-letter alphabet of amino acid types. The *secondary structure* of the protein is defined by domains which exhibit certain structural motifs that stabilize the protein's conformation. The most frequent of these are the right-handed α-helix and the β-sheet, a planar motif in which the backbone threads back around itself. The goal of PSP is to predict the *tertiary structure* of proteins, which is the three-dimensional shape of the folded protein. Of biological relevance, but beyond the scope of this paper, is *quaternary structure*, in which several polypeptide subunits interact to form a stable conformation together.[3]

The dogma of PSP is that biology has selected the lowest possible energy conformation of the polypeptide as native. The problem is thus to infer the lowest energy conformation from the energy landscape.

Molecular biology has accumulated a corpus of knowledge on the commonalities and eccentricities of

biological proteins (ie. those present in cells). Several secondary structural motifs are known. The aforementioned α-helix is a right-handed helical structure spaced by 3.6 residues per turn, and in which the structure requires certain sidechains and hydrogen bonding to be present. The β-sheet consists of multiple strands which are attracted by hydrogen bonds in either a parallel or antiparallel conformation. Less common, but prominent, formations such as the greek-key and β-hairpin are observed. For membrane-spanning proteins (transmembrane proteins) and the TIM barrel, both α and β domains interact. The secondary structure motifs play a significant role in the tertiary structure, for they define the dihedral angles between adjacent amino acid residues. The polypeptide backbone adopts two torsional angles for each residue: ϕ for the rotation between the α-carbon and the amine nitrogen atom, and ψ between the α-carbon and the adjacent carbon$'$ ("carbon-prime", in the carbonyl group).[3] Ramachandran showed that the joint distribution of these two angles clusters at three disparate domains, each corresponding to its own secondary structural element.[29] Frequently, scientists use this prior knowledge in order to inform models of protein folding. Nevertheless, PSP is a grand problem which is difficult even when secondary structure information is known beforehand. In fact, Levinthal estimated that under restriction of ϕ and ψ angles to realistic values, a polypeptide of length 100 has over 10^{143} possible conformations.[20] It is an understatement that naïve enumeration is no solution for protein folding.

One approach to PSP, which we focus on, is *ab initio* protein folding. *Ab initio* requires that no prior knowledge of the protein is known except for its primary structure. In practice, these efforts have been less fruitful than which exploit *a priori* knowledge such as the protein type; for example, the globular protein hemoglobin may be better targeted using a special method. *Ab initio* methods seek to discover general characteristics of protein structure and are the subject of decades of work. A landmark study by Anfinsen in the 1960s established that a protein denatured (unfolded) by interfering agents reconstituted its native confirmation when those agents were removed.[1] This experiment develops the intuition that protein structure is dependent only on the primary structure and justifies the *ab initio* effort.

2.1 Hardness results for protein folding

Computationally, it is generally understood that PSP is an NP-hard problem. Arguably, the most general formulation (Ngo and Marks) postulates a protein in which each atom is identified, and their respective connectivities are enumerated. Given this information alone, and a generic energy function similar to all-atom physical models, U:

$$U = \sum_b K_b^{bond}(l_b - l_b^0)^2 + \sum_a K_a^{angle}(\theta_a - \theta_a^0)^2 + \sum_t K_t^{torsion}(1 - \cos\left[n_t(\phi_t - \phi_t^0)\right]) + \sum_{i>j} K_{ij}^{nonlocal} f(r_{ij}/r_{ij}^0) \quad (1)$$

Ngo and Marks showed that the minimization of U is NP-hard for any dimensionless function $f(x)$ with a unique minimum at $x = 1$.[23] The proof of NP-hardness supposes a polypeptide chain in which a variably folded region (one with several possible torsions) exists, and the global minimum is only achievable when some scaffold region comes in contact with the variable region. The scaffold region must bend around a fixed domain in order to make this contact. Ngo and Marks arrive at a reduction to the partition problem: the energy is minimized if and only if the algorithm answers the Partition problem.[23] The Partition problem asks whether a set of integers may be partitioned into two subsets of equal sum, and is necessary, under the conditions described, for the geometric conformation that is energetically minimum. This formulation corresponds to the general chemical and geometric characteristics of the problem and is respected as a proof of hardness; indeed, there is yet to be an popular formulation which is computationally "easy". For instance, discrete models of protein folding are also NP-hard by reduction to bin packing (the minimization of bins into which objects are placed) or Hamiltonian path problems.[16]

Furthermore, PSP is a problem for which there is no unique formulation. There are several models of protein folding, and they generally fall into one of two categories: off-lattice and on-lattice. Off-lattice models allow the protein's components to move, free-floating, in a continuous space. These models are popular in biology since they agree with the observed flexibility of proteins and the science of X-ray crystallography. On-lattice models map the protein's components to points on a discrete lattice. The goal of on-lattice models is to confine the size of the energy landscape and reduce the protein folding problem to its simplest

formulation.

2.2 Bridging the lattice debate

PSP research is sharply divided between the off-lattice and on-lattice factions. On-lattice protein folding algorithms have mathematically provable guarantees on performance, but these methods are not currently popular among biological researchers; most frequently, structural prediction in the twenty-first century has favored off-lattice models and heuristic algorithms. This study seeks to bridge the divide by projecting biological proteins from their continuous representations onto a reasonable lattice model, from which one may gain intuition for solving the discrete PSP problem. We adopt, and will formalize in Section 3.2, a model derived from the well-studied hydrophobic-polar (HP) model proposed by Kenneth Dill. Dill's survey of proteins identified the interactions between hydrophobic residues to be the dominant force in protein folding.[19] The model generalizes amino acids by partitioning them to two sets: hydrophilic (liking water) and hydrophobic (repelled by water).

A hallmark of correct protein folding is the early incidence of hydrophobic collapse, in which internal hydrophobic residues bury themselves into the core of the molecule. The HP model exploits the dominance of hydrophobic-hydrophobic contact in the folding event in order to drastically reduce the problem complexity. Furthermore, we adopt a sidechain model in which the sidechain portion of the amino acid residue is localized to a lattice point adjacent to, but separate from, the α-carbon's position. Intuitively, this formulation is more faithful to polypeptide structure. In addition, Bromberg and Dill showed that the sidechain model has beneficial properties for minimization of protein energy, particularly by the increase of degrees of freedom of conformation, the distinct and sharp minimum in entropy when a sidechain-folded protein arrives at the native formation, and an empirical distribution of residue-residue contacts adhering more closely to PDB structures.[4]

The as-of-yet unanswered questions for on-lattice models pertain to the *appropriateness* and *effectiveness* of discretization in regards to real proteins. Does there exist a lattice which expresses known elements of protein structure? What is a reasonable and predictive lattice energy function for proteins? How can we model the hydrophobic collapse – the "key ingredient" of folding – in a way that aids algorithmic design?

2.3 Method of Protein Chain Lattice Fitting

The Protein Chain Lattice Fitting (PCLF) problem consists of choosing the optimal superposition of a lattice structure such that it minimizes the root mean squared distortion between the subject protein's (continuous) coordinates and the lattice coordinates. Solving the PCLF problem requires choosing the optimum scaling for the lattice (generally these parameters are known in the case of proteins) and then the optimum translation and rotation pair. Gaur describes a reduction of PCLF to VLP-3-SAT (var-linked planar 3-SAT), which is an instance of 3-SAT admitting a linear ordering of the variables such that the implication graph is planar.[6] The problem of RMSD minimization for PCLF is therefore NP-complete. In order to plot proteins from the PDB repository onto an FCC-sidechain lattice, we use the University of Freiburg's LatFit software and its default parameters.[22]

LatFit uses a technique first described by Park and Levitt[25] for a greedy solution to the PCLF problem. The LatFit algorithm first assigns coordinates for each of the protein residues and sidechains. The backbone coordinate is that of the residue's α-carbon, and the sidechain is coordinate is the center of mass for all atoms in that sidechain (except for the amino acid glycine, which is sidechain-less; LatFit maps the "sidechain" coordinate for this amino acid to the α-carbon).

LatFit assumes a fixed lattice distance of $3.8\mathring{A}$, which is the distance between adjacent α-carbons in the polypeptide structure. Additionally, it is approximately equal to the average distance from the α-carbon to the center of mass of the sidechain ($3.6\mathring{A}$).[21]

LatFit builds the lattice-fit polypeptide residue-by-residue starting from the amino terminus. The algorithm stores an array of the k best fittings (by cRMSD, see Eq. 2). To fit residue i, LatFit attempts all possible elongations of the k best self-avoiding fittings up to residue $i-1$. The k best of these are then retained for the next iteration. One challenge for the algorithm is to compute the optimally rotated lattice which fits the PDB structure; this is accomplished by dividing the entire rotation range for each of the three dimensions r^X, r^Y, r^Z into s equal parts and repeating the aforementioned fitting procedure a total of s^3 times. The optimal rotation of these s^3 instances is then subjected to further refinement over a rotation range of $m_r\pi$ for s_r intervals to produce the final output.[21] For fewer than 50 of 1000 of the tested 1198 proteins the algorithm fails because at some stage, none of the k best fits can be extended in a self-avoiding

fashion.

$$cRMSD \quad = \quad \sqrt{\frac{\sum_{i=1}^{l}(|\hat{P}_i^b - L_i^b|)^2 + (|\hat{P}_i^s - L_i^s|)^2}{2 \cdot l}} \tag{2}$$

LatFit is a greedy heuristic, but uses a respected method which roughly covers the search space. LatFit achieves a mean 1.84$\mathring{A}$ cRMSD over the protein corpus for fitting to an FCC-sidechain model (3.29$\mathring{A}$ for the plain cubic lattice), and only 11 protein instances exceed a cRMSD of 2.00$\mathring{A}$. For reference, the X-ray crystallography resolution of these structures is roughly 1.5$\mathring{A}$, and recall that the distance between α-carbons is $\approx 3.8\mathring{A}$. Thus quantization introduces some penalty, but it is not of an order of magnitude different than error observed in other fitting processes.[22] LatFit is currently the only software which fits arbitrary PDB files to sidechain lattice models. In principle, fitting of lattice models could be improved by an Expectation Maximization formulation which descends to a local minimum cRMSD, or by a comprehensive study which analytically probed the asymptotic limit of cRMSD for biological proteins. In 1996, Huang et al. performed an analysis of a contact potential model which achieved a mean cRMSD of 1.52$\mathring{A}$ for native structures at 298K. This result is not necessarily generalizable due to the small sample of five proteins used, and the continuous space of the contact potential. However, the study establishes that a cRMSD of this magnitude is strong amongst the available fitting methods.[15]

It should be noted that there exist artificial lattices with far higher coordination numbers than the FCC, which has regularly spaced connectivity and is thus classified as a regular lattice. In fact, researchers have developed lattices that span the continuum between the FCC's kissing number (12) and the (specially designed for proteins) CABS (Cα-Cβ-Side groups) model with 800 allowable vectors. We favor the FCC because it is the unique solution to Kepler's conjecture for the tightest packing of equally-sized spheres.[8] Thus the FCC vectors not only express paths between α-carbons and sidechain, but also define a closest "kissing" for all 12 adjacent spheres. The FCC thus has the geometrical properties desired for tight packing of hydrophobic residues, and we choose it as the simplest and most natural lattice for folding. The high coordination lattices do offer much stronger average RMSD (for example, on the order of .35$\mathring{A}$ for the CABS model described above), but arguably that is a result of "overfitting" to biological structures. It would be a great advance to prove that enough information is captured in the FCC model, which my analysis will

answer to some degree.

3D models of the lattice-fit proteins used for the subsequent analysis are accessible via the Internet at `http://www.brown.edu/Research/IstrailLab/pdbView.php`. The corpus extracted from the PDB repository consists of 1198 proteins selected on the basis of having a length within 50-300 amino acids and no greater than $1.5\mathring{A}$ RMSD from error in X-ray crystallography (the technology of fitting crystallography experiments to Euclidean space too has error!). The restriction on X-ray crystallography resolution prohibits analysis on proteins which may not discretize well due to random deviations, or which were measured on older hardware. In the online 3D structures, a Jmol applet is used to visualize the FCC on-lattice structure with the backbone colored gray, hydrophobic sidechains red, and hydrophilic sidechains green.

Character	Δx	Δy	Δz
0	+1	+1	0
1	+1	-1	0
2	+1	0	+1
3	+1	0	-1
4	0	+1	+1
5	0	+1	-1
6	0	-1	+1
7	0	-1	-1
8	-1	+1	0
9	-1	-1	0
A	-1	0	+1
B	-1	0	-1

Figure 1: Vectors in the Face-centered Cubic Lattice

3 The Protein Energy Potential Function Problem

3.1 Introduction to the energy function inference problem

In motivating the NP-hardness of PSP, recall that we cited a generic energy function U for polymers of atoms. Physical energy functions of this variety are used in order to produce all-atom computer simulations of the protein folding process, or as a "brute-force" method for solving protein structural prediction problems (most notably, Folding@Home). The advantage of deriving energies from thermodynamics (typically free

energy) is that in these cases the potential is a function of known quantities (e.g. electrostatic repulsion).

Most commonly, energy functions used for PSP are summative quantities of pairwise interactions between residues. Statistical energy functions which describe a Boltzmann distribution of conformations are typical for PSP problems. Statistical mechanics dictates that the probability p_N of the native structure is given by

$$p_N = \frac{e^{-E_N/(kT)}}{\sum_m e^{-E_m/(kT)}} \tag{3}$$

where E_N is the energy of the native structure, k is the Boltzmann constant, T is the absolute temperature, and $m \in M$, the set of all possible conformations. p_N is the greatest probability in the sum of the denominator.[9] Taking the natural logarithm of this above equation, we achieve

$$max(kT \ln p_N) = max(-E_N + F) \tag{4}$$

where F is the free energy quantity, and is equal to the average energy minus an entropy term TS. Using the above equation we can solve for the energy of the native structure in terms of p_N and the free energy. We can substitute the definition of free energy (Eq. 6) and decompose our formula into an energy term for a single pairwise interaction (Eq. 7).

$$E_N = -kT \ln p_N - F \tag{5}$$

$$F = \langle E \rangle - TS = -kT \ln p_{XX} - TS = -kT \ln p^o_{XX} \tag{6}$$

$$\forall A, B \in \{amino\ acids\}, E_{AB}(r) = -kT \ln[p_{AB}(r + \delta r)/p^o_{XX}(r + \delta r)] \tag{7}$$

Equation 7 gives the calculation for the energy of the interaction of two amino acid types A and B with respect to a reference state denoted by the "generic" amino acid pair XX. The function $p_{AB}(r \pm \delta r)$ is the probability of observing residues A and B separated by the distance $r \pm \delta r$. $p_X X^o(r \pm \delta r)$ gives a reference probability for residues separated in distance by $r \pm \delta r$; as implied by Equation 6 this constant corresponds to an average probability adjusted by the entropy term TS.[9] Using an interaction matrix of probabilities p_{AB} with the correct empirical properties, it is reasonable that a protein folded such that the

pairwise distance between residues approximates the "native" distribution will have the lowest energy state. Physical chemistry has developed methods for constructing the potential functions $p_{AB}(x)$ and $p_{XX}(x)$ in order to make structural prediction.[31]

In off-lattice models, Function 7 is minimized over the length of the protein by assuming empirical probability parameters and computing the maximum likelihood structure; however, as we seek a model for on-lattice proteins and particularly one that is *as fundamental* as possible, we use Dill's HP model as a starting point for energy function inference. The HP model proposes to simplify the function above by summing only pairwise hydrophobic interactions and assigning each of these contacts an energy of -1.[7] Formally, a contact is made between two residues which are adjacent on the lattice, but not adjacent in the protein backbone. (The argument for this definition is that local interactions dominate at a close-range scale, whereas the fold energy is derived from "long-range" interactions of atoms which are brought to close proximity in the native.)

3.2 Construction of the HP Model

The classical HP model classifies each amino acid sidechain as either 'H' (hydrophobic) or 'P' (hydrophilic). Of the twenty amino acids, eight are defined as hydrophobic: alanine, cysteine, isoleucine, leucine, phenylalanine, proline, tryptophan, and valine. Arginine, asparagine, aspartate, glutamine, histadine, lysine, serine, threonine, and tyrosine are hydrophilic residues. Glycine can act like either of the two, due to its small size, and the sulfuric amino acid methionine is also typically specified as "special"; in the HP model, we will ignore this complication and declare them both hydrophobic amino acids.[7][19]

The energy potential in Dill's HP model is

$$E = \sum_{i,j} c_{i,j} \tag{8}$$

$$c_{i,j} = \begin{cases} -1 & \text{if } j > i+1 \text{ and } S_i, S_j = H \\ 0 & otherwise \end{cases} \tag{9}$$

E is thus the sum of pairwise contacts on lattice. More generally, this may be extended to a symmetric interaction matrix indexed by the nominal values assigned to the sidechain residues. More sophisticated

	H	P
H	-1	0
P	0	0

Figure 2: The most frequently used protein energy function in HP model algorithms for PSP.

models also adopt a solvent term. The biochemistry of proteins dictates that protein structure is in large part determined by interactions with the solvent environment. Hydrophobic residues are observed to bury themselves into the protein and form compact structures in the presence from solvent (as suggested by the hydrophobic collapse). The "hydrophobic effect" maximizes the entropy of the solvent, without which the energy landscape is disrupted and no "funnel" develops for fast energy minimization.[3] Hydrophilic residues are expected to tolerate the aqueous environment without frustration of the energetically favorable state. (Note: Beware that this is by no means a comprehensive view on the importance of solvent! It would be wrong to claim that hydrophilic residues "like water" or that hydrophobic residues do not normally come in contact with solvent. In fact, we observe that there is a complex interaction between aqueous solvents and the polar or "hydrophilic" residues, and depends partly on dissolved substances in the solvent. Moreover, hydrophobic residues are observed on the outer regions of biological proteins. The above statements refer to that which is *energetically and entropically favorable*.)

Solvent is introduced into the model by adding a new category to the interaction matrix. Shown in Figure 3 are three commonly used matrices in the HP model.[10] For the purposes of my model, I define

HP	H	P	S
H	-1	0	0
P	0	0	0
S	0	0	0

SHP	H	P	S
H	-2	0	1
P	0	0	0
S	1	0	0

ASA	H	P	S
H	0	0	1
P	0	0	0
S	1	0	0

Figure 3: The interaction matrix depends on which parameters are believed to be significant in the folding process, or otherwise derived by some fitting process. HP is the standard hydrophobic-polar model, and is equivalent to the previous figure. SHP is the solvent-exposed hydrophobic-polar, wherein a penalty is made for interactions between solvent and hydrophobic residues. ASA is the accessible surface area, in which the collapse is explained by entropic effects alone, driven by the hydrophobic effect.[10]

the solvent interaction as occurring between all pairs of residues which are not adjacent to any backbone or sidechain coordinate, except for the α-carbon to which it is attached. This identifies sidechains which are not "buried" inside the molecule. Note that there is a dependency on the number of contact interactions;

since most of the residues which do not form any contacts are declared "solvent-interacting", then there is a strong negative correlation. Furthermore, the number of solvent interactions are proportionally smaller: whereas a single hydrophobic residue could in theory make up to twelve contacts, a hydrophobic-solvent interaction is only counted once per residue. Nevertheless, both the number of hydrophobic-hydrophobic contacts and the number of hydrophobic-solvent interactions both grow asymptotically linearly [not proven here] so this does not pose a grand problem.

Thomas and Dill showed that an optimum three-parameter model ('H-H', 'H-P', 'P-P') with values $E_{HH} = -5$, $E_{HP} = -1$, $E_{PP} = -2$ fit all 2040 PDB models for 14-mer chains on a simple cubic lattice. When reversing the method – to extract from these proteins values for the three parameters by statistical fitting of these 2040 structures under the Boltzmann function described above, Thomas and Dill observed empirical parameters $e_{HH} = -5, e_{HP} = -1.1, e_{PP} = -2.1$.[31] Their result is restricted in scope to those very short polypeptides, and in general, findings about on-lattice models only hold for their underlying lattice model. However, the success of the statistical method inspires this paper's analysis for longer, more realistic proteins fit to FCC-sidechain lattice.

3.3 Model fitting in general

Typically, simple symmetric interaction matrices of size no greater than 3×3 are applied to the HP model. However, the field is open for science to propose and score new candidate functions for protein folding.. The goal of the model fitting problem is to construct a model M and the optimal parameters θ_1, θ_2, etc. Define $\Theta = (M, \theta_1, \theta_2, ...)$. Then the optimum model minimizes some metric of the distance between structures predicted under that model and the respective PDB structures. The aforementioned cRMSD is frequently applied to this purpose, although at times dRMSD is chosen as metric because it is tolerant of transformation or reflection in the predicted structure.

$$optimum\ model\ \Theta^* \quad = \quad argmin_\Theta RMSD_\Theta(predicted, native) \tag{10}$$

For example, Huang et al. exploit the geometric properties of hydrophobic cores in order to develop a unique energy function. Their "hydrophobic fitness score" is designed to reach a maximum when the

hydrophobics are (1) optimally packed and (2) buried into the exact center of the protein. They define the hydrophobic fitness score F [15]

$$F = \quad -\frac{(\sum_i B_i)(\sum_i (H_i - H_i^o))}{n^2} \quad 1 \leq i \leq n = \#\ hydrophobic\ residues \tag{11}$$

where B_i is the hydrophobic burial function, H_i is the number of hydrophobic contacts, and H_i^o is the number of hydrophobic contacts by expectation under a random model. In Huang et al, B_i is defined as the number of sidechain coordinates within $10\mathring{A}$, which may be adapted to the discrete formulation (lattice distance $\approx 3.8\mathring{A}$) as within 3 lattice units.

For our particular model fitting problem, we choose a discrete formulation of the problem which achieves a similar goal to Equation 10: the "predicted" structures used to score our models will be decoy structures. "Decoy" proteins are suboptimal structures with large similarity to the native fold, such that they might appear – to a poor energy function – to be the biological native for that protein.

Decoy Generation

Decoy generation is a challenging problem that requires close attention to mimicing both the mathematical and biological properties of the native. Principally, decoys must meet the hard constraints of connectedness and self-avoidance; lattice decoys must also place all α-carbons and sidechains on the lattice. It is important that the decoys could not be separated from native folds by any trivial means (for example, dihedral angles which never occur in biological proteins). Park and Levitt describe four desiderata for decoy proteins:[26]

1. decoy sets must include structures "close to" native structures (by $RMSD$)
2. "be native-like in all properties of the real polypeptide chain." For example, secondary structures must be retained.
3. "be diverse" so they are representative of other likely folds of low energy
4. be of sufficient quantity. This is a challenge because frequently researchers generating decoy sets have worked on small subsets of the PDB. Levitt's Decoys R' Us, for example, has thousands of decoys

per protein, but covers a subset of size ≈ 200 of all proteins. Given that computing near-optimal folds is difficult – and to do so respecting these four desiderata is particularly difficult – computing a comprehensive decoy set would be of interest to protein research.

Decoys used to discriminate models were generated from the LatFit dataset. The dataset, generally derived from newer, high-resolution PDB models and longer sequences than the Levitt data, had no overlap with any known decoy databases. I generated for each of the proteins in the dataset some number of decoys averaging between 15 and 25. Care was taken to respect connectedness, self-avoidance, and the above desiderata, with one caveat: the decoys do not entirely replicate all properties of the native. In some instances, secondary structures such as α-helices may have been disrupted by the method. The algorithm applied was as follows:

Algorithm DecoyMaker(P, p, R)

Input: P = set of PDB proteins
Input: p = probability of perturbation
Input: R = number of repeats
Output: return a set $\{D_i\}$ of decoys for each $P_i \in P$
for $R \in \{1...R\}$ **do**
  for *amino acid* $k \in P_i$ **do**
    if *randomUniform* $\in \{0, 1\} < p$ **then**
      m = Choose a move from $\Delta x, \Delta y, \Delta z$
      if c*onnected and self-avoiding after move* **then**
        Δ^0 = Find the change m (eg. Δx) from amino acid $k-1$ *to* k
        Δ' = Find the change m from amino acid k *to* $k+1$
        $acid_k = acid_k - \Delta^0 + \Delta'$
        $acid_{k+1} = acid_{k+1} - \Delta' + \Delta^0$
        Repeat for sidechain
      end
    end
  end
end

The above algorithm makes randomly chosen perturbations in the protein structure such that the chain is still connected and self-avoiding. 25 decoy generation attempts are made for each protein, with 2 possible modes of failure: insufficient difference from native (due to too many proposed moves being non-self-avoiding) or program error (if it returns an incorrect, crossed chain, or if too many hydrophobic interactions were dropped in the decoy). Failures were purged from the decoy sets.

Any proposed decoy method deserves criticism, and if this work were to be extended, there are a variety of different decoy methods that are to be explored. As noted above, this method may distort realistic protein structures which are known to form on these sequences, such as secondary structures. In addition, there are compact domains in the proteins which are difficult to perturb in a manner which is self-avoiding. Empirically, the algorithm can in some cases attempt to misfold to a molten globule state (which might account for the high number of apparent hydrophobic contacts on these decoys).

The decoy-making algorithm changed an average of 16.1% residues over the PDB set and for the given parameters ($p = .2, R = 20$). In rare cases, moves in later iterations of the main loop may cancel earlier moves, thus the total displaced residues are a little less than 15%. It should be said that these decoys are fairly close to the native conformation, as the above algorithm is unlikely to greatly change the shape of the protein; rather, it alters the chain residue-by-residue. While it was beneficial to have decoys within such a close distance of the native, a number of other techniques can be attempted. For example, Park and Levitt use a method which retains the structure in small blocks along the protein and slightly modify the "hinges" between these domains. In addition, it is generally good practice to relax the protein energy by filtering out interactions which are chemically unstable due to steric hindrance or unrealistic bond angles. Frequently perturbed structures may be returned to a more native-like fold by a variety of finishing steps which use statistical data (for example, the expected number of interactions of a given type) to restore properties of biological proteins.[26]

3.4 Model inference for the FCC-fit protein dataset

We attempted to fit an energy function to the observed intrachain interactions in our lattice-fit PDB set. That is, we sought to find an equation which is a function of the long-range interactions within the protein chain. The objective function for scoring said function was its capacity to discriminate true natives from decoys; in other words, the biological native is scored with a lower energy than all of its decoys. Since this may not be possible, we choose to optimize an objective function maximizing the the number of decoys which are scored at a higher energy than their native. For the following objective function, let P be the set of PDB natives (indexed by i), f be the energy function, and I be an indicator variable that valued at 1 when the

predicate is satified (0 otherwise).

$$S = \sum_i \sum_{d \in \{ \text{ decoys for } P_i \}} I[f(d) > f(P_i)] \tag{12}$$

Our decoy set consisted of 19,492 decoys, so S is valued between 0 (all false classifications) and 19,492 (perfect classication). A random guess of the native would yield a S approximately one-half times the size of the PDB set (≈ 9726).

Before assessing a more complicated energy function, we attempted the simple HP energy function described by interaction matrix HP (Figure 3). The score of the objective function function for this model was 12,206, achieving a 63% success rate. We then allowed all of HH, HP, and PP to be free variables, and this model was optimized by the parameter vector $(-.14, .01, -.04)$ with a score of 13,541. Using the full five-variable model that we have before described, we add solvent terms for each of the two types of residues. The optimum linear model was thus

opt-linear	H	P	S
H	-.13	0	-.05
P		-.04	-.06
S			X

Figure 4: The best linear model as scored by the objective function S. Note that the hydrophobic-hydrophobic interaction is strongest in magnitude, and that hydrophilic-hydrophilic interactions are also favorable. Hydrophobic-hydrophilic interactions are found to have no effect on protein energy. Solvent interactions are have an effect on energy with the same magnitude as hydrophilic-hydrophilic interactions, yielding evidence to a body of protein research studying the utility of exposed residues.

The optimum set of parameters was first found using gradient descent, and later confirmed by exhaustively enumerating a $10 \times 10 \times 10 \times 10 \times 10$ hypercube centered at the origin. The exhaustive enumeration yielded a solution which converged onto a multiple of that found in Figure 4 after a refinement step. The optimum parameterization had a score S of 13,997 (72% success rate). In assessing the model, it should be noted that the inclusion of solvent terms only increases the success rate by about 2.5%, and therefore it may be that the additional parametrization is not useful.It is not so surprising that the solvent terms do not add greatly to the predictive capability of the model, as they are partly dependent on the number of residue-residue interactions (as interactions in the molecule increase, the number of solvent interactions decreases under our

definition).

The model has interesting mathematical and biological consequences for understanding protein kinetics. As is the expectation of Ken Dill's HP model, hydrophobic-hydrophobic interactions are highly favorable for biological proteins. Hydrophilics also appear to interact favorably with each other, to a lesser extent. The solvent interactions are more difficult to understand, but they agree with recent findings showing that the solvent helps guide the protein to a native conformation. The dissatisfying dillemma with the model is that it only achieves 72% success in discriminating decoys from native, which is not greatly more powerful than random classification (50%). Our goal is to develop a model that universally describes optimality for protein energy, with (perfect or) near 100% accuracy. However, decoys generated under our method did not necessarily have fewer hydrophobic-hydrophobic contacts than the native, and a nontrivial portion had greater numbers of hydrophobic-hydrophobic contacts. This may be the result, for example, of the tendency of the decoy generating method to alter the molecule to a more "molten globule" conformation. The parameters we have chosen appear to satisfy biological expectations for proper folding, but the energy function does not satisfy our requirement of selecting the native from all possible conformations.

The functions we have so far explored are linear in the arguments, and furthermore, are functions exclusively of pairwise interactions. We attempted fitting models that are non-linear (but nevertheless used data from the same pairwise data), for example, by squaring or multiplying different terms and evaluating the value of S for that model. None of the non-linear models tested outperformed the 5-parameter linear model. It may be desirable, for the purposes of folding algorithms, to make the energy function as minimal as possible. The two-parameter linear model $\alpha_1(\text{H-H } \textit{contacts}) + \alpha_2(\text{P-P } \textit{contacts})$ with parametrization $\alpha = (-.15, -.04)$ achieves a score $S = 13365$ (69%), which was superior to any non-linear model tested which multiplied these terms or powers thereof.

More sophisticated energy functions incorporate variables that represent different quantities than pairwise interactions. In fact, pioneering work on protein decoys and energy function inference by Michael Levitt [15][26] suggest that functions which mix a variety of known properties for good folds tend to perform better in practical applications. Park and Levitt state that a "combination function" of pairwise interactions, van de Waals forces, surface energy, and residue burial yields the best results for proteins for the 200 or so

proteins they extracted from the PDB. Levitt explains that this type of function excels because each of the interaction measures we use are highly distance-dependent, and therefore, combining multiple functions of the data reduces the error in any one measure.[26] If this work is to be continued, experimentation with incorporating these terms into the objective function might yield better results. (Note, on the other hand, that we desire a potential that is simple enough to be minimized, so the addition of variables must be weighed against the increased complexity of the function.)

In an attempt to test Levitt's hypothesis, we assessed each of the natives and the decoys using the Hydrophobic Fitness (HF) score in Equation 11. Clearly, terms represent a quantity containing information more complex than the pairwise interactions. Using the PDB set, we computed an expected number of hydrophobic contacts for a given length of protein H_i^o. Then we computed an HF score for each of the PDB proteins and its decoys, and calculated S where the HF score of the native structure exceeded that of the chosen decoy. Although Huang et al. report that this scheme separates better than the standard HP interaction matrix, it was not the case in our analysis, where we computed $S = 11222$ (58%). The cause of the failure of this function is as yet unknown, but probably is a result of differences in the discrete vs. continuous models and/or different methods of decoy generation.

We wish to do better than the 72% classifier, which was the optimum linear model with hydrophobic/hydrophilic interactions and solvent terms. The difficulty in fitting a function of this form to the dataset suggests that there is no possible energy function which incorporates only pairwise interactions between residues. In collaboration with Warren Schudy and Sorin Istrail, we conjecture impossibility for two problems in energy function inference: the Unique Native Linear Pairwise Energy problem and the General Optimal Linear Pairwise Energy Problem which are to be described in the following subsection.

3.5 Pairwise Function Impossibility Conjecture [2]

We have reason to suspect that the difficulty in modeling true proteins is inherent to the laws of physics governing biochemical energies. Our potential function belongs to the general set of functions where pairwise interactions between residues have some additive score, and thus the potential is calculated by a sum once the contacts are specified. This function is convenient, but in all likelihood the true potential has terms

which consider multiple interacting factors at once – and such a potential is more difficult to optimize. We conjecture that no pairwise linear function sufficiently approximates the native potential of the set of PDB proteins. We propose to demonstrate this via a generalized linear programming technique where we show that a linear model with the desired properties of a proper energy function is *infeasible* (no solution to the parametrization exists).

Define: Let P be a set of proteins from the Protein Data Bank (PDB). Define a distance d: d is the maximum acceptable Euclidean distance between the center of mass of any two sidechains for which a pair may be called "in contact". Then, for each P_i, those pairs of sidechains which are "in contact" compose a set $S_i(d)$ of contacts. Each "contact" $s \in S_i$ is a tuple (l, j) of the two amino acid types which are "in contact." Let M be a matrix of real-valued numbers indexed by the twenty types of amino acids. M is a symmetric matrix with up to 190 unique elements, and $M_{(x,y)}$ is the potential assigned to an amino acid x being "in contact" with amino acid y.

where P_i is the native structure of an arbitrary protein from the PDB and D is a reasonable domain for the distance between an interaction classified "in contact". That is, any nontrivial function f will fail to assign to all native PDB proteins at a minimum energy state. We must show that this is also invariant on the distance d chosen to distinguish contacts from non-contacts, within a reasonable domain D.

As you can see, the conjecture is proven true under the infeasibility of a linear model which is to be described below. Such a proof is a "hard" problem under the requirement of $\forall p$. Our computer-assisted proof must be able to propose protein folds p which challenge the native structure P_i. It is possible that the computational method may fail to generate a p which proves infeasibility of the function f (due to the NP-hardness of protein structure prediction). Therefore, a positive result for f is not a rigorous solution, nor is it necessarily a "true" energy function. However, in the case that the computational method gives an infeasibility result we can definitively conclude that no linear function of this form admits a solution. Thus, the "true" energy function must either be non-linear and/or must be a function of k-tuples of sidechain residues where $k > 2$.

Conjecture: We conjecture that there exists no protein energy function $E_p = f(p, d, M)$ for which $f(p, d, M)$ is a linear combination of the elements of M. In other words, we will let the distance d define the

number of sidechains k, ℓ in contact $n_{k,\ell}(P, d)$ for the protein P. All nontrivial functions f of the form

$$f(p, d, M) \quad = \sum_{k,\ell} M_{k,\ell} n_{k,\ell}(P_i, d), \quad k, \ell \in \text{amino-acids} \tag{13}$$

will not satisfy

$$\forall i, \forall p, \forall d \in D, \quad f(p, d, M) \geq f(P_i, d, M) \tag{14}$$

Formulation: In order to efficiently prove an infeasibility result, we propose a linear program. Such a linear program will attempt to find the optimum linear sum of pairwise potential values (parameters) M. Under the assumption that protein structures in the PDB are at a minimum energy ("native" structure), we will attempt to minimize the energy function for all PDB structures. Therefore, we define our objective function as the minimization of the sum of the potential of all PDB structures:[32]

$$Objective : \min_{M} \sum_{i} E_{P_i} = \quad \min_M \sum_i f(P_i, d, M) = \sum_i \sum_{k,\ell} M_{k,\ell} n_{k,\ell}(P_i, d) \tag{15}$$

If there exists an "true" energy function E_p, then the M^* which is argmin to the above function would be the true parametrization: it gives a minimum energy for all PDB proteins. Recall that due to the hardness of the problem, we have no means of proving that any output M^* corresponds to the energy function for proteins in nature; it might be that we have simply not enumerated a p for which the energy function is less than that of a PDB structure.

There are are at least $N \times (\#\ decoys/structure)$ constraints, too many to enumerate in their entirety. Therefore, we plan to implement a separation oracle which is called by the LP solver. The oracle as input the set of proteins P, the distance d, and a matrix M and returns "true" if M is feasible, "false" otherwise. The crucial constraint is as follows:

$$\forall d \in D, \forall i, \forall p \neq P_i \quad f(p, d, M) \geq f(P_i, d, M) + \epsilon \quad Unique\ Native \tag{16}$$

$$\forall d \in D, \forall i, \forall p \qquad f(p, d, M) \geq f(P_i, d, M) \qquad Generalized\ Optimum \tag{17}$$

The constraint is defined by our conjecture and it must hold for a feasible solution under the conditions of an energy function. We require that the energy function distinguishes each decoy from its native, for each protein in the PDB. There are two hypotheses to test: Equation 16 is true under the hypothesis that the biological native structure is uniquely minimum in the energy landscape. A weaker constraint (Equation 17 does not suppose that all sequences have a single global minimum conformation. Some natural proteins have a small ensemble of energetically favorable states, and furthermore, we also wish to solve the problem on discrete lattices. We desire to show impossiblility for the latter, although the computation may be more difficult.

*S*ummary of **Pairwise Function Impossibility Conjecture**

Purpose: **show impossibility of a pairwise energy function of the following form**
Let P be the set of proteins in the PDB repository and d be a distance within a appropriate domain D of maximum contact distance

$$E_{P_i} = f(P_i, d, M) = \sum_{k,\ell} M_{k,\ell} n_{k,\ell}(P_i, d), \quad k, \ell \in \text{amino-acids}$$

Unique Native Linear Pairwise Energy

$$\sum_{1 \le i \le \#\text{ of decoys}} (F(native\ for\ decoy\ i) > F(decoy\ i))$$

Problem
Let $\vec{m}$ be the vector representation of parameter matrix M, indexed by an amino acid pair (i, j).

$$\begin{array}{ll} & \min_{\vec{m}} \sum_{i=1}^{N} \vec{m} \cdot n(P_i, d) \\ st. \forall d \in D, \forall i, \forall p \neq P_i & E_p \geq E_{P_i} + \epsilon \end{array}$$

General Optimal Linear Pairwise Energy Problem (GOLPE)
Let $n^b_{k,\ell}$ be an additional parameter vector: $n^b_{k,\ell}(P, d)$ is a estimate for the number contacts we expect between residues k, ℓ in the database P (the sum of expectations over each P_i's decoy set).

$$\begin{array}{lll} & \min_{\vec{m}} \left[\sum_{i=1}^{N} \vec{m} \cdot n(P_i, d)\right] & \\ st. \forall d \in D, \forall i, \forall p & E_p \geq E_{P_i} & \\ \forall k, \ell & -M_{k,\ell} \geq \epsilon & \text{if } \sum_{i=1}^{N} n_{k,\ell}(P_i, d) > n^b_{k,\ell}(P, d) \\ \forall k, \ell & M_{k,\ell} \geq \epsilon & \text{if } \sum_{i=1}^{N} n_{k,\ell}(P_i, d) < n^b_{k,\ell}(P, d) \end{array}$$

Specifically, without any additional constraints to Equation 17, the LP admits a pathological "solution": if M equals the all-zeroes matrix, then the constraint is trivially satisfied (the inequality is not strict). We must place an additional constraint to enforce good properties of an energy function. One method is to require, $\forall k, \ell, \sum_{k,\ell} M_{k,\ell} = c$. However, this heuristic is ad hoc because there is no biochemical basis for inferring a net positive, net negative, or net zero potential over all possible residue pairings (additionally it could be dependent on d. Under linearity of the objective function, a feasible solution is easily scaled to sum to c, but it supposes a choice of positive, negative, or zero which is possibly restrictive of a solution. We suggest an alternative workaround for this difficulty by enforcing a "parameters sign constraint": suppose that the decoy sets for all proteins in P comprise a background distribution of proteins. Then, compute a number $n^b_{k,\ell}(P, d)$ of contacts which represents a reference number of contacts for the $k \leftrightarrow \ell$ interaction over the entire set P. For those k, ℓ pairs which are enriched in the natives (P) set a constraint that $M_{k,\ell}$ is negative, and for those pairs which are deficient, set a positive constraint. This constraint makes $f(P_i, d, M)$ behave like the Boltzmann distribution described earlier, in which we reward residues which have more contacts in the native structures. We conjecture further that for any non-zero feasible solution of M, this additional constraint is also satisfied (by linearity of f). There shall be a greater number of those contacts in the natives, and thus the negative value is chosen under minimization of the objective.

As the function is non-linear in the choice of d (d defines the discrete inclusion or exclusion of residue pairs from S_i), we propose to prove infeasibility for several values of d in some domain D. We define the domain D by the underlying biochemistry of long-range interactions in proteins. The minimum value of $d \in D$ should be some value such that there exist many contacts in the protein structure. For completeness, the minimum d could be the minimum distance between the center of mass for any two sidechains in the the entire PDB repository. The maximum d value might be chosen to be a value such that not too many residue pairs are in contact; perhaps we could choose a d such that fewer than βn residue pairs are in contact when there are n possible residue pairs over the entirety of the proteins in the PDB repository.

We must take care when choosing the input and algorithm used for computing the LP result. There exist conditions under which an LP solver might declare infeasibility but the result might not be valid. First, an appropriate LP solver must be chosen. There exist dual simplex algorithms and ellipsoid solvers which may

make use of a separation oracle with linear constraints. In addition, there are biological complications for the input set P of PDB proteins. The PDB must be filtered to prevent the introduction of proteins for which other factors like ligands or obscure solvents influence the conformation of the protein recorded under X-ray models. For example, we must exclude multiple-chain proteins; in these proteins, multiple chains form quaternary structures and the interaction between chains might interfere with the calculation of an energy function. In addition, many PDB structures are made for the purposes of demonstrating alternate conformations of the protein under binding or dimerization scenarios and these must also be excluded. Lastly, a non-redundant, high-resolution reduction of the PDB repository should be used. For several polypeptides with the same sequence, the PDB gives multiple (divergent) structures. We may not assume that these proteins have equal energies. In all likelihood, they too depict alternative conditions for the protein's environment. If we were to allow a redundancy in the database, then it is likely that even under a reasonable energy function f that equal-sequence proteins $P_i = p'$ and $P_j = p''$ would not satisfy the constraint $f(p'', d, M) \geq f(P_i, d, M)$. We suggest that the set of proteins P be derived from a curated non-redundant (but *representative*) repository derived from PDB such as Pisces[33], and furthermore, there might be additional curation performed to remove multiple-chain proteins.

The separation oracle will need to either (1) test the energy function against a database of decoys for each PDB entry P_i or (2) produce such folds by Monte Carlo simulation. If we are to use the sign constraints on parameters $M_{k,\ell}$ then the database is especially important for generating the background distribution. It seems likely that decoy construction methods will come to bear in this problem: not only do we require a comprehensive decoy database (covering all P_i with thousands of decoys each) but the decoy database must compile decoys constructed by a variety of different algorithms. Even stochastically generated decoys might have some substructure which make them unsuitable for proving impossibility, if similar methods are used for each decoy. To prove impossibility in any reasonable amount of time, we must prepare by "stocking" the decoy database well.

Under an infeasible model, the LP algorithm returns an IIS and an MIS. The IIS, or Irreducible Infeasible Subset, is a set of constraints which are simultaneously unsatisfiable. This corresponds to a subset of P (the PDB repository), $P^{(infeasible)}$. No model defined by f is feasible for the set $P^{(infeasible)}$,

but removing any single PDB structure from $P^{(infeasible)}$ **is** feasible. Ideally, $P^{(infeasible)}$ would consist of several classes of proteins such that we may infer that no pairwise contact function can fit the diverse landscape of proteins. The MIS, a minimum-cardinality IIS set cover, is the smallest set which must be removed from P such that the objective function is feasible. The MIS consists of one element from each IIS such that $P \backslash MIS$ is feasible. The minimum-cardinality IIS set cover problem is NP-complete, but many LP programs solve for the MIS in reasonable time using heuristics. We might demand that the MIS is not *too small*. It is well-known that there are some PDB structures in error (and historically several have been withdrawn). An infeasibility result with a very small MIS may not be satisfactory because a small MIS is likely to consist of those PDB structures which exhibit high error or poor experimental procedure.

To address the problem of testing for d at several discrete points in some domain D, we may replace the single value $M_{x,y}$ for the pairing of amino acid x to amino acid y by a function $M_{x,y}(r)$ which varies in the distance r which separates the two sidechains. Some suggested functions for this are the Boltzmann distribution (a logarithmic, non-linear function) or the Lennard-Jones potential (a function linear in the its parameters but which does not deviate greatly from "true" energy curves).[32] Lennard-Jones potential has the form

$$M_{x,y}(r) = \quad \frac{m_{x,y,0}}{r^6} - \frac{m_{x,y,1}}{r^{12}} \tag{18}$$

and thus under this new model we would need to prove that there exist no 380 parameters ($M_{x,y,z}, x, y \in$ *amino acids*$, z \in \{0,1\}$) which satisfy the constraints which we have defined. This may provide more conclusive evidence that there is no "potential-like function" of pairwise residue potentials contacts which governs the folds of biological proteins.

4 Modelling Hydrophobic Collapse

4.1 The State of the Art

Since the proposed energy function of the standard HP model only accounts for the pairwise contacts of hydrophobic residues, a large body of work in the literature confronts the protein folding problem by

prioritizing the hydrophobic collapse. Under the interaction matrix HP (Figure 3), an optimum packing of hydrophobic sidechains optimizes the fold energy. Standard HP algorithms attempt principally to form a molten globule structure in the center of the fold, as are observed in the globular class of proteins.

A frequently explored method is to make well-chosen folding points such that the hydrophobic residues in one block are adjacent to another block of hydrophobic residues. That is, these algorithms form a partition of the residues into several contiguous blocks and match those with similar hydropathy. We can arrive at this intuition by observing that in any lattice L there exists a linear bound on the number of contacts for fold F defined by the number of hydrophobic residues and their positions in the sequence S (and the lattice coordination number):[16]

$$\#Contacts(F(S), L) \leq g(S, L) \tag{19}$$

For example, in the 2D square lattice (a plane with residues at 90 degree angles to one another), any two residues may only come in contact if and only if there is a path of even length which brings the downstream residue to an adjacent position; therefore, any odd amino acid can only create a contact with some other odd amino acid. This defines a function $g(\mathcal{O}(S), \mathcal{E}(S)) = 2\min(\mathcal{O}(S), \mathcal{E}(S))$ which bounds the number of contacts.

The algorithmic procedure for hydrophobic-block-based folding chooses the partition into x and y blocks such that x and y blocks alternate, as in $x_1y_1x_2y_2$ and so forth. We choose the partitioning such that the number of hydrophobics in x blocks exceeds that of y blocks. Then a fold in which x blocks are aligned and in contact (with turns at the x-y boundaries) is produced. A crucial feature of these Hart-Istrail partitions is that they are computable in linear time (per the aforementioned partition function) and provide guaranteed approximation of the optimum value of the contact potential. Once the partition is made, x blocks are aligned to maximize contacts and the length of the y blocks are used to bridge between the x-x pairs. In this 2D square lattice, the "worst-case" scenario in which $\#_x$ *hydrophobics* $= \#_y$ *hydrophobics* for every adjacent x and y block dictates that only half of all even or odd hydrophobics (whichever is minimum) pairs may come in contact. Therefore, a lower bound on the algorithm's performance is $\frac{1}{2}\min(\mathcal{O}(S), \mathcal{E}(S))$. It

follows that the approximation guarantee of this algorithm is

$$G_{2D} = \frac{\frac{1}{2}\min(\mathcal{O}(S), \mathcal{E}(S))}{2\min(\mathcal{O}(S), \mathcal{E}(S))} = \frac{1}{4} \tag{20}$$

As the coordination number increases and the combinatorics allow an increased number of ways to join pairs of amino acids on-lattice, then the Hart-Istrail improves in performance.[12][13]

Methods like Hart-Istrail belong to the "approximation algorithms" class, by which we mean that the algorithm always produces a result within some fraction of the optimum (native) fold's energy. As noted above, it has some beneficial characteristics: linear runtime and determinism (the result does not vary randomly). In protein folding, approximation algorithms are by no means the most popular. The most frequently used algorithms have advanced heuristics coupled to an analytical optimization method. The Rosetta program for off-lattice inference uses heuristics and a highly-parallelized linear programming algorithm to optimize computation speed. Classically, Monte Carlo algorithms have been applied to folding in which "moves" (various geometric transformations of the protein which retain connectedness and self-avoidingness) are chosen randomly with some defined probability.

Novel techniques such as Genetic Algorithms (GA) and Constraint Programming (CP) algorithms are increasingly popular in the twenty-first century. A great many GA approaches have been applied to protein folding; these algorithms tend not to require biological intuition, but rather, mate randomly generated solutions so as to descend in conformational energy. They are reported to have good results, but are not as respected as more analytical methods. CP algorithms are also gaining momentum because their performance has improved given the increase in computational power and wider knowledge of optimization techniques such as branch-and-bound. Advanced CP algorithms such as Palú et al's Constraint Logic Programming over finite domains (CLP(FD)) solution use a large library of biological knowledge as constraints to guide the algorithm.[24] GA and CP favor practicality over the geometric and combinatorial dilemmas of protein folding. At the same time, they introduce a great number of assumptions to the folding problem which make analysis difficult. In exploring PSP, this paper will embrace more traditional approximation algorithms, but with an eye towards the practical goal of making a performant general-purpose algorithm. I use quantitative data from the PDB repository to make empirical judgments on the quality of the approximation algorithm

and will try to drive at novel intuition.

4.2 Evaluating the biplanar model for hydrophobic collapse

Hart and Istrail propose a model in which the hydrophobic mass of the protein is localized to a biplane in the center of the protein conformation. More specifically, the algorithm $\mathcal{C}$ which they describe constructs a 4×2 biplane in the center of the protein. That is, eight large columns (long hydrophobic blocks are brought into contact with each other; these columns are orthogonal with the planes of the FCC lattice. The intuition of the algorithm is that packing all hydrophobics in the center maximizes the energy when only hydrophobic interactions are considered. Then, when a fold is made according to algorithm $\mathcal{C}$ with the correct height of columns, then the 4x2 biplane forms at least 14 contacts between H residues on this plane. At most the other 8 adjacent positions (which are inhabited by other residues on the chain) could have been hydrophobic residues for each plane. Therefore, Hart and Istrail arrive at an approximation which covers the core in at least 14 of the 22 accessible columns. Then the energy of the algorithm for h hydrophobic residues arranged with width B in a $B \times 2$ biplane is

$$E \leq -(B+2)\frac{31}{8}h - \frac{14}{8}h + Z, \quad Z > 0 \tag{21}$$

For the choice $B = 4$ the approximation ratio is 86% on the FCC-sidechain lattice, which is as of yet unbeaten for PSP approximation algorithms in the HP model.

In order to better develop the relationship between real, continuous proteins and the rigorous lattice model, we developed an experiment to validate the observation of biplanes in the PDB dataset used for the energy function energy problem. We formulated the following *statistical* question: can we show that, in general, PDB proteins (or a subset thereof) may be described as biplanar structures with slight deviations by random error? Among the various methods that could be used to answer this question, we formulated the problem in terms of the relative hydrophobicity of the central plane to the entire protein. Can we account for a majority of the magnitude of the native energy by the sidechain interactions formed in some biplayer superimposed on the discretized PDB structure? If so, we conclude that that native belongs to a class of biplane-like proteins, and otherwise, the evidence contradicts the model.

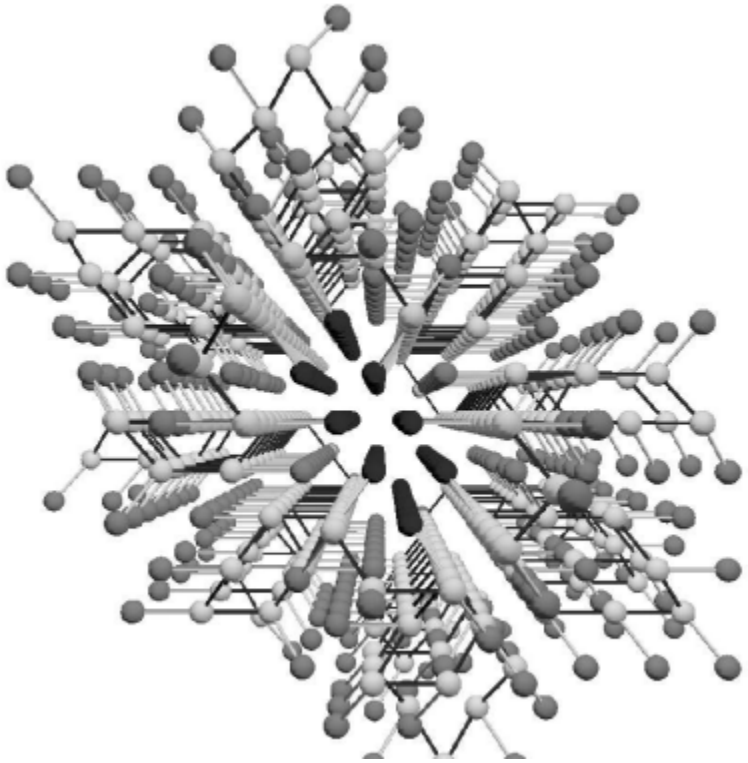


Figure 5: A 4×2 biplane conformation as folded by the Hart-Istrail algorithm on the FCC-sidechain lattice. The algorithm guarantees a conformation within 86% of optimal. The authors showed that this biplanar fold achieves a 98% approximation of the native energy. From Hart and Newman[11].

We developed a method to fit the best possible cutting biplane on the protein structure. Operating on the set of FCC-sidechain lattice-fit structures, we sampled 100 random planes in three-dimensional space (which are each determined by three points in the space) which intersected the center of mass for each protein. We discard a plane as a failure if and only if more than 2/3 of the residues fell on one side of some "center axis" (ie. to enforce cutting along the central axis). We then created a threshold two lattice units orthogonal to the center axis in both directions, and identified each of the hydrophobic sidechains contained within.[1]

Our hypothesis was as follows: if PDB proteins are biased towards a biplanar conformation, then we would expect to see a significant amount of hydrophobicity contained within this small window. If the hypothesis were false, then very little of the total hydrophobicity of the polypeptide could be located to

[1]Thresholds with "interplane" width k were scored for each of $k \in \{1, 2, 3, 4, 5\}$. We chose $k = 2$ because it describes a biplane and one additional unit of distance to account for errors in crystallography and lattice fitting.

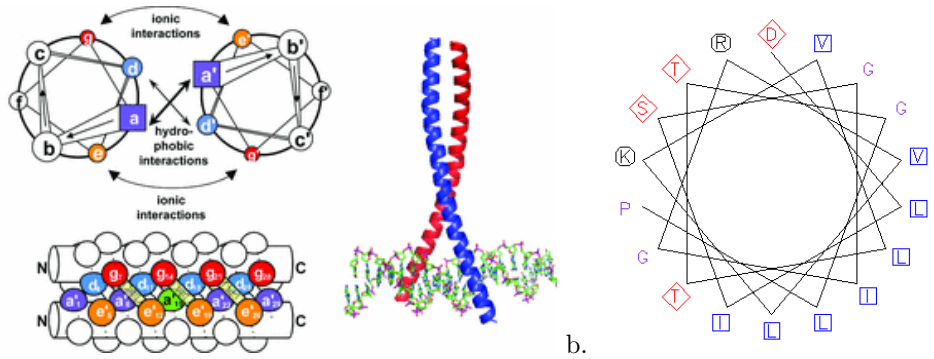
this region. The method used to score this was as follows: using the hydrophobic residues annotated in the "biplanar region" we computed the hydrophobic contacts which arise from only this subset of hydrophobic contacts, and partitioned our set of PDB proteins by the classifier (22):

$$\textit{biplanar} \iff \tau \times (\#\ \text{H-H}\ \textit{contacts}) < (\#\ \textit{biplane}\ \text{H-H}\ \textit{contacts}) \tag{22}$$

For $\tau = \frac{1}{3}$, classifying a subset of the PDB proteins yielded a ratio of 1 *biplanar* : 1 *not biplanar* (29:29). We infer that approximately half of the PDB proteins can be characterized as "biplanar" and the other half is "not biplanar". The observation is not striking, as certain classes of proteins are known to be nearly biplanar; for example, a globular protein has a strongly hydrophobic core which would be likely to have high hydrophobicity along a plane).

A striking trend in the data is that proteins tended to be strongly biplanar or strongly non-biplanar. Qualitatively, we observed that the distribution of hydrophobic contacts in the biplanar region was bimodal. To quantify this, we normalized the hydrophobicity within the biplane region to the total number of hydrophobic interactions in the entire protein and took a logarithmic ratio (all values were negative as the biplane hydrophobicity cannot exceed the total). The sample variance of the natural log ratios was 1.26, indicating a wide distribution of ratios which included proteins for which biplanar hydrophobics $\approx$ total hydrophobicity and some in which the optimal biplane had very few hydrophobic residues.

The tables in Appendix 7.3 give the accession numbers and descriptions for PDB proteins which scored as biplanar and those which we classify as not biplanar. We found that the types of proteins in each set were diverse, having no clear bias in their SCOP or CATH annotations (which classify proteins by structure and function). However, there were general characteristics in common for each set. The biplanar proteins were of similar length (mean 119 residues biplanar vs. 123 residues not biplanar). Using the SSpro secondary structure prediction software on the SCRATCH server [5] we also analyzed the secondary structures on these proteins. Biplanar proteins contained fewer residues in α-helices (30% vs. 37%) and were slightly more unstructured (49% vs. 44%). The proportion of β structures was approximately equal (19% vs. 20%), as was solvent accessibility (46% vs. 47% of residues with more than 25% solvent accessibility). We conclude that the biplanar approximation is applicable to a great number of cases in the PDB, and that it has favorable



a. b.

Figure 6: Two α-helix motifs with specific patterns of hydrophobicity. In the coiled coil (a), both hydrophobic and hydrophilic interactions occur at the intersection of two α-helices. Hydrophobic residues and ionic interactions occur diagonally rather than at right angles to one another, as expected in a biplane. In the helical wheel (b), hydrophobic residues (red) cluster on the left side of the structure, whereas hydrophilic residues (blue) form a hydrogen-bonding domain on the right side.

properties for optimizing protein energy. However, the equally large number of counterexamples shows that conformations similar to biplanes do not characterize realistic proteins in a universal sense, and that the α-helical secondary structure does not conform to the approximation.

The α-helix may reduce the conformational energy of the protein in a more complicated manner than supposed by the energy functions we have so far discussed. The α-helix forms an electrostatic bipole in which a net charge points down the barrel of the helix. The helix places demands on the sequence of amino acids such that polar atoms on the hydrophilic sidechains point parallel to the axis of the helix so as to come in align to hydrogens on nearby residues (hydrogen bonding). A consequence of this arrangement is that that hydrophobic residues in the α-helix sequence will align on one side of the helix, whereas hydrophilics inhabit the opposite side.[3] Therefore, α-helices tend to create their own domains of hydrophobicity which do not conform to the biplane model. (See Figure 6) That is, a protein with several α domains is likely to have several hydrophobic regions in its 3D conformation but no such "biplane". On the other hand, a protein with significant lengths of β-sheets or unstructured regions is more likely to have an axis along which hydrophobic residues pair and form energetically favorable biplanes. While the biplane model is not "wrong", we must develop additional structures to account for the interruption in biplanarity caused by α-helices and which have flexibility for the multiple, secondary-structure-specific domains of hydrophobicity which might arise in

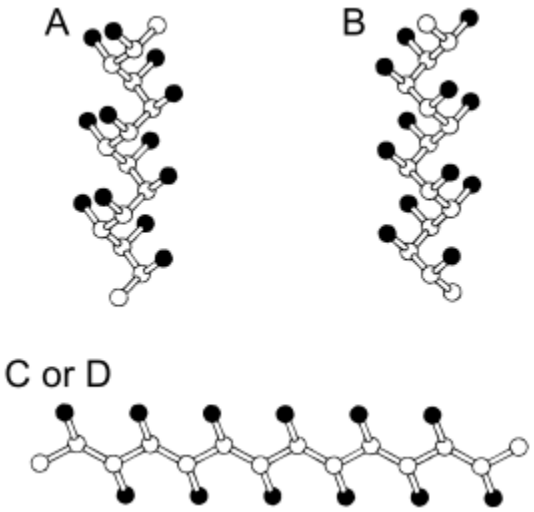


Figure 7: Rotamer conformations in the FCC lattice which correspond to the main features of protein secondary structure. A gives a right-handed α-helix. B is a left-handed α-helix, which does not occur naturally. C/D gives an approximation of the antiparallel or parallel β-sheet.

a complex protein.

4.3 α-helices and polyhedra in the FCC lattice

Previous studies suggest that the face-centered cubic lattice and its capacity for describing octahedrons are applicable to the protein folding problem.[17][18][14] In Toma and Toma 1999, the authors describe Simulated Annealing algorithms for folding proteins into cuboctahedrons. A cuboctahedron is the medial of an octahedron and is the analog of cube for the FCC lattice in that it describes a cube rotated about a central sphere. Therefore, it can be inferred that a near-optimal algorithm for folding on the FCC lattice will form one or several octahedral conformations threaded along the protein backbone. The authors describe other properties of the FCC which may be exploited to form known protein structures; Toma and Toma describe conformations (Figure 7) that agree roughly with right-handed (1) and left-handed α-helices, and β-sheets. In real proteins the average right-handed α-helix rotamer angles are 55^o and -55^o whereas they are 50^o and -56^o respectively in real proteins. Similarly the torsional angles for a β-sheet are 180^o and 72^o in the model compared to 180^o and 56^o in biological proteins.[17]

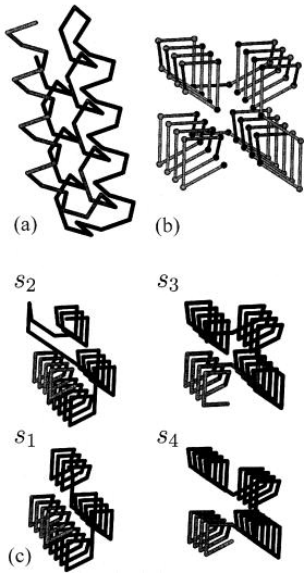


Figure 8: Three models for on-lattice interactions of four α-helices in the FCC. (b) simulates a hydrophobic core, whereas (a) and (c) represent neighboring four-helix bundles. From Pokarowski et al.[27]

Pokarowski et al. [27][28] report a series of four vectors in the FCC lattice which correspond roughly to the torsion of the true α-helix, visualized in Figure 8. The sequence of vectors may be repeated for a longer helix, or two vector sequences may be placed on opposite sides of a polypeptide "turn" in order to form a bundle. Their study finds that the FCC lattice is a minimal protein model for expressing α-helical motifs.

So far, we have shown that the biplanar model does not allow the hydrophobic signature of highly α-helical proteins which we observe in nature. However, we conjecture – and multiple sources of evidence agree – that the FCC model is capable of approximating the geometric constraints of helices. Furthermore, the vectors in FCC space which describe this helix resemble the outline of a path around the surface of an octahedron, with sidechain residues pointing into and closing the octahedral shape. From this intuition, we have analyzed near-octahedral shapes in the FCC as a potential partner to the biplane for folding.

To compare the optimality of a biplanar structure against an octahedron on the FCC lattice, a small example of twenty hydrophobic residues was constructed for both structure types. (20 is chosen, as you will see, because this number of spheres coincides with a perfect near-octahedron fold which is known as a locally optimal value. See Figure 13 where $i = 6$, $T_i = 20$.) Spheres were plotted onto an FCC lattice

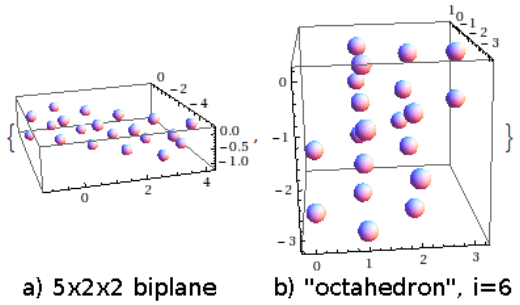


Figure 9: Favorable structures in the FCC: We compare 20 hydrophobic residues packed in (a) a 5×2 biplane conformation and a (b) near-octahedral conformation to be described in the next subsection ($i = 6, T_i = 20$). Using the HP contact function (Equation 8) (a) has 69 hydrophobic contacts and (b) has 73.

in a Mathematica notebook. For this example (Figure 9) the $5 \times 2 \times 2$ biplane achieves 69 contacts and the near-octahedron only achieves 73. Mathematica notebook for this example can be downloaded from `http://www.brown.edu/Research/Istrail_Lab/Allanthesis/fcc20_example.nb` or viewed as an XML page for those without Mathematica software. The octahedron – analogous to a cube – is favorable on the FCC lattice. However, the caveats are clear: octahedral conformations in FCC space may have favorable properties yet not be *optimal* under all circumstances. Hart and Istrail's algorithm within 86% of optimal suggests to us the conjecture that the biplane is the single universal structure that achieves near-optimality for a variety of HP sequences. The optimality of the octahedron is dependent on the number of hydrophobic residues falling within a confined set of values, whereas the biplane can be constructed for any even number of residues and is abundant in contacts. When a comparison like the above (Figure 9) was constructed for 18 residues, a $3 \times 3 \times 2$ biplane outscored the near-octahedron 53 to 52. It follows that a future algorithm for PSP on face-centered cubic lattices ought to incorporate multiple substructures into its folding routine. The biplane is a single best approximation (by conjecture), but not sufficient as a solution for PSP.

In exploring novel models of hydrophobic collapse, we will scrutinize these octahedral structures to derive conjectures and mathematical results. We report these results in terms of their relative optimality under the HP model, even though it is of interest to propose a model where these conformations are more favorable. The octahedron is a regular polyhedron well-suited for the FCC lattice because the lattice angles match the face angles of 120 deg. In addition, the octahedron is compact with a surface area to volume ratio of 5.72.

While there exist a multitude of folding schemes which could incorporate small polyhedra along the chain, we will attempt to characterize a method which assembles masses of hydrophobicity into one octahedral "core" within the protein. This process is known as self-assembly, wherein the hydrophobic residues select an ideal packing which is highly energetically favorable.

In order to further develop the octahedron idea for FCC lattices, we will further simplify the protein to a bipole model. Suppose that we no longer enforce the backbone – the "string" on which residues lie – as a constraint on the folding problem. Thus, we decompose each protein into its constituent "α-carbon with attached sidechain" fragments. Each fragment is a *bipole* with strong hydrophobicity on one side (as is the case for a hydrophobic sidechain) and nearly neutral hydropathy on the other. Then, we can ask: how can one arrange the bipole fragments so as to minimize the energy? I will propose a bipole packing method in the HP model with approximation ratio of greater than 44% (developed in the following subsection and proven (sketch) in Appendix 7.1) which develops on the intuition of octahedrons in the FCC model.

4.4 Optimal Bipole Packing: folding's paralog

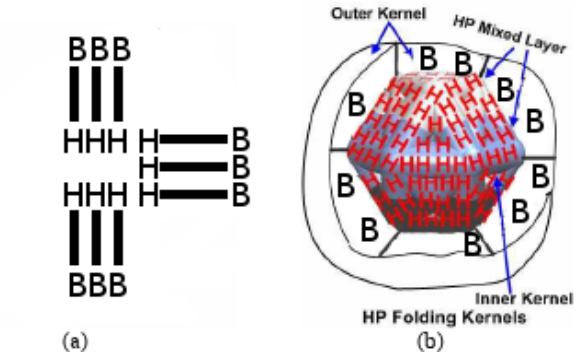


Figure 10: (a) The OBP problem models each residue as a unit length segment with an "H" sphere on one end and "B" sphere on the other end. (b) The general goal of the Stewart algorithm for bipole packing on FCC is to place bipoles on the borders of an octahedron, with "H" ends facing in. (b) is from Hoque et al.[14]

First, we must give a formal definition of bipole packing. Istrail and Lam define the Optimal Bipole Packing (OBP) problem[16] as follows: Given a fixed number of bipoles, find a non-overlapping conformation of bipoles and an assignment for the bipole labels such that

1. "each bipole has one endpoint labeled H (representing the hydrophobic side chain) and one endpoint labeled B (the relatively neutral backbone)

2. the number of contacts between endpoints labeled H is maximized over all possible assignments" (ie. energy is minimized)

To construct an approximation algorithm for OBP, I take the approach of forming an near-octahedral hydrophobic core. That is, we thread the protein such that the hydrophobic sidechains face into the center of a surface formed by two tetrahedrons reflected over a center axis. Although there may be unused area in the center which is inaccessible in the case of a large octahedron, this conformation maximizes the hydrophobic contacts which are possible by forming a large octahedral surface.

For certain quantities of bipoles, we are capable of forming an "optimal-octahedral packing". Define an optimal-octahedral packing as one in which the hydrophobic sides of bipoles are laid on the layers of the FCC lattice in triangles of sequentially increasing size towards some largest center plane. From this center plane, the size of triangles tapers off symmetrically to the other side. The triangle layer comes in contact with the layers above and below. One acceptable conformation is to have a singleton below a three-unit triangle and a singleton above. The next possible conformation has a singleton with two three-unit layers in the center, then a singleton at the top, and so on. (See Figure 13)

Triangle index j	Number of residues on border t_j
1	1
2	3
3	6
4	12
5	24
6	48
7	96
$n \geq 2$	$6 \times 2^{n-2}$

Figure 11: Triangles which are allowable under the model of the Stewart algorithm. Triangle j admits t_j residues on its border. An optimal packing on the FCC-sidechain lattice places all of its hydrophobic residues on the boundary of these triangles, forming a hydrophobic molten core through which the backbone (or its hydrophilic residues) does not pass.

The optimal-octahedral packing is achievable when the number of hydrophobic residues equals one of the T_is for an acceptable conformation in (Figure 13). In this case, the packing is perfect and admits a maximum

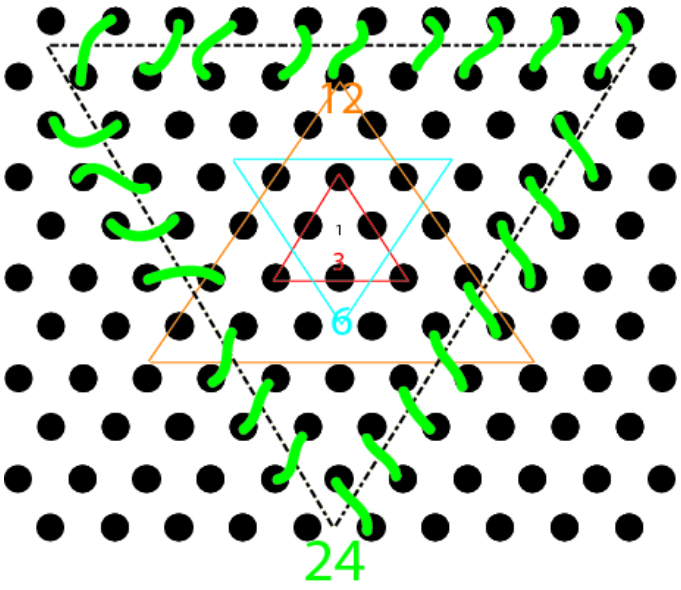


Figure 12: Overlapping inverted triangles form a well-packed, tetrahedron-like shape on the FCC lattice. The FCC lattice consists of many layers of the above pattern, where the odd-numbered layers are shifted in one position. We overlay triangles which are reflected relative to the ones above and below, each of increasing size, until the size reaches a maximum in the middle plane. Here we show a layer with 24 border bipoles, signified by the green bonds. The hydrophobic side of the bipoles face into the center of the core, forming the border of each triangle for every plane in the conformation. The adjacent layers are shown in color, where odd-numbered level layers are offset in accordance to the FCC basis vectors. The borders of adjacent layers create contact points, with hydrophobicity concentrated at corners.

number of hydrophobic contacts. However, as the series of values T_i increases rapidly, it is rare that any T_i equals the number of hydrophobics in a given protein *exactly*. The proposed Stewart FCC-sidechain algorithm is as follows: thread the protein such that the hydrophobic residues point into the symmetric mass as described above ("optimal packing"), supposing that it is possible to do so in a self-avoiding manner. Suppose there are h bipoles, and let T^* be the T_i for which $T^* \geq h$. Build the protein with the optimal-octahedral conformation for T^*, but omit residues in the core $T^* - h$ times. (For a true protein, a backbone or hydrophilic residue might inhabit this location.)

In order to prove an approximation ratio for this folding method, we must show that the omission of hydrophobics in the "optimal-octahedral packing" (due to restriction on the number of bipoles) reduces the number of contacts by some quantity which we may bound. This amounts to finding the "worst-case"

approximation ratio when some number of bipoles in between two of the T_i (Figure 13) is selected. Supposing an HP model in which only hydrophobic-hydrophobic contacts reduce energy and the hydrophilic ends of bipoles have no effect on the energy potential, an approximation ratio of 44% is shown for this bipole packing method in Appendix 7.1.

Size index i	Hydrophobic residues T_i	Hydrophobic residues on all planes t_j (Figure 11)
1	1	1
2	2	1 + 1
3	5	1 + 3 + 1
4	8	1 + 3 + 3 + 1
5	14	1+3+6+3+1
6	20	1+3+6+6+3+1
7	32	1+3+6+12+6+3+1
8	44	1+3+6+12+12+6+3+1
9	68	1+3+6+12+24+12+6+3+1
10	92	1+3+6+12+24+24+12+6+3+1
11	140	1+3+6+12+24+48+24+12+6+3+1
12	188	1+3+6+12+24+48+48+24+12+6+3+1
13	284	1+3+6+12+24+48+96+48+24+12+6+3+1
14	380	1+3+6+12+24+48+96+96+48+24+12+6+3+1
15	572	1+3+6+12+24+48+96+192+96+48+24+12+6+3+1
16	764	1+3+6+12+24+48+96+192+192+96+48+24+12+6+3+1
17	1148	1+3+6+12+24+48+96+192+384+192+96+48+24+12+6+3+1
18	1532	1+3+6+12+24+48+96+192+384+384+192+96+48+24+12+6+3+1

Figure 13: Optimal hydrophobic cores using the octahedral planes model on an FCC sidechain lattice. Each core consists of triangles increasing in size towards a large central triangle in the center plane.

4.5 Discussion and caveats: Algorithms on discrete lattice

Even the best protein folding algorithms raise several issues, the two most problematic being (1) inadequacy of the energy function and (2) in the case of deterministic algorithms, a deliberately confined set of possible conformations. The first problem is simple to state: the energy function domains which these algorithms require are typically missing terms which we know to be thermodynamically crucial to the stable structure of proteins. The approximation algorithms described above have proven optimality for the standard HP interaction matrix. Once additional solvent terms are added, or even more challenging quantities – non-pairwise interactions for example – then the proof must be restructured. Biologically, it is not entirely

concrete that hydrophobic-hydrophobic interactions are the single most important forces for folding; several experts in protein folding argue that the repulsion of hydrophobics *away from* solvent drives the protein fold.

Furthermore, our energy functions overlook the significance of thermodynamic entropy. Recall that in adapting Boltzmann's distribution to describe the fold energy, we found that

$$E_N = -kT \ln p_N - F \tag{23}$$

$$F = \langle E \rangle - TS \tag{24}$$

F, the reference free energy, is a function of the average energy and the entropy S, scaled by the absolute temperature T. The entropy term is quite important: if it were not for entropy, the air you are breathing right now would lazily lie on the floor! When developing a universal protein energy function, we are computing the expectation of F over all proteins and for all temperatures. In the environment of the cell, and given the residue-specific determinants of the reference free energy F, these proteins may reach entirely different energy minima than we would expect from observing the "average" protein. (Not surprisingly, for example, the native can be very sensitive to changes in temperature.) Therefore, it is not entirely clear whether the conditions assumed by most protein folding problems are widely applicable to all possible proteins. *But we forge ahead*, looking for "tricks" that will fold a number of biological proteins correctly.

When we fold with a deterministic algorithm, we restrict the space of solutions in a way which might not be appropriate. This is related to the first concern; proteins have evolved in and exist in a random environment, so we cannot presuppose the likelihood that they attain a steady-state that rigidly conforming to the global optimum. In this paper, we celebrate the approximation algorithm for shedding light on general geometric and statistical mechanics principles. These algorithms are provably good, but we must also look for good randomized (Monte Carlo) algorithms which might converge on structures the algorithm designer never foresaw. It is clear from observing PDB structures that these proteins take on forms far and beyond the biplane or octahedron. In addition, heuristics may be applicable, such as the Bounded Block Fails (BBF) method used in CP algorithms.[24] BBF divides the protein into several blocks and folds, sequentially, a random permutation of the blocks. A "fail" statistic tracks when the current block being folded does not reach a certain threshold expectation, upon which the algorithm backtracks to a previous block and "fixes"

the previously made error. The heuristic works well when there are wide enough valleys in the energy landscape to allow descent towards the minimum. Coupling this with other optimization techniques allows researchers to explore the regions of the energy function most likely to lead to minima.

Educated scholars of protein folding implicitly disagree about the efficacy of *ab initio* and informed methods in PSP. For practical purposes, algorithms which incorporate secondary structure information into the objective function have faster runtimes and more accurate prediction. These methods can rely on repositories of secondary structure data which are mature and reliable. For example, we can align the sequences of short secondary structure domains to X-ray crystallography models in order to infer the tertiary structure of the protein at multiple segments along the chain. In particular, the Ramachandran aligner uses protein-protein alignment and known Ramachandran trajectories in order to select rotamer bond angles which are energetically favorable.[30] See that these methods incorporate statistically known properties which we observe in the true biology – with solvent and all – and not just the geometric intuition of contacts. *Ab initio* is the ideal target for PSP, but is it too idealistic? Does biological data distort the physical principles governing PSP and bias it towards what is known, or is it necessary for successful prediction? The answer is yet unknown.

5 Acknowledgements

I thank Austin Huang for enlightening discussions on proteins, Warren Schudy for his linear programming expertise and mathematical insights, Derek Aguiar for encouragement, and Sorin Istrail for three years of regaling me with protein lore. I would also like to honor several current and former members of the Istrail Lab group: Fumei Lam, Ryan Tarpine, Kyle Schutter, Chris Maloney, David Moskowitz, and Erin Klopfenstein. Thanks also to Martin Mann of the University of Freiburg group for his assistance and for writing the LatFit software which made this research possible.

6 Bibliography

References

[1] Anfinsen CB, Haber E, Sela M, and White FH. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 47:1309–1314, September 1961.

[2] Stewart AP, Schudy W, and Istrail SC. Pairwise impossibility for protein energy functions (in progress). 2010.

[3] Carl-Ivar Branden and John Tooze. *Introduction to Protein Structure: Second Edition*. Garland Publishing, second edition, January 1999.

[4] Sarina Bromberg and Kenneth Dill. Side-chain entropy and packing in proteins. *Protein Science*, 3:997–1009, 1994.

[5] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi. SCRATCH: a protein structure and structural feature prediction server. *Nucl. Acids Res.*, 33:W72–76, 2005.

[6] Daya Ram Gaur. Fitting protein chains to cubic lattice is np-complete. In *The Proceedings of 2007 APBC, Full Text*, 2007.

[7] Kenneth A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 6:1501–1509, 1985.

[8] Thomas C. Hales. A proof of the Kepler conjecture. *Ann. of Math. (2)*, 162(3):1065–1185, 2005.

[9] Ming-Hong Hao and Harold A. Scheraga. Optimizing potential functions for protein folding. *J. Phys. Chem.*

[10] William Hart and Sorin Istrail. An algorithmic analysis of lattice and off-lattice protein folding models with repulsion and attraction (unpublished work).

[11] William Hart and Alantha Newman. *Protein Structure Prediction with Lattice Models*. Chapman and Hall, 2005.

[12] William E. Hart and Sorin Istrail. Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. *Journal of Computational Biology, Spring*, 3:157–168, 1996.

[13] William E. Hart and Sorin Istrail. Lattice and off-lattice side chain models of protein folding: linear time structure prediction better than 86 *J. Comput. Biol*, 4:241–259, 1997.

[14] Madhu Hoque, Tamidul Chetty and Abdul Sattar. Protein folding prediction in 3d fcc hp lattice using genetic algorithm. pages 4138–4145, 2007.

[15] Enoch S. Huang, S. Subbiah, Jerry Tsai, and Michael Levitt. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *Journal of Molecular Biology*, 257(3):716 – 725, 1996.

[16] Sorin Istrail and Fumei Lam. Combinatorial algorithms for protein folding in lattice models: A survey of mathematical results. *Commun. Inf. Syst.*, 9:303–346, 2009.

[17] Toma L and Toma S. Contact interactions method: a new algorithm for protein folding simulations. *Protein Sci*.

[18] Toma L and Toma S. Folding simulation of protein models on the structure-based cubo-octahedral lattice with the contact interactions algorithm. *PRS*, 8(01):196–202, 1999.

[19] Kit Fun Lau and Kenneth A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 10(22):3986–3997, 1989.

[20] Cyrus Levinthal. Mossbauer spectroscopy in biological systems. In J. T. P. DeBrunner and E. Munck, editor, *A meeting held at Allerton House, Monticello, Illinois*.

[21] Martin Mann. Latfit manual. 2008.

[22] Rolf Backofen Martin Mann, Rhodri Saunders and Charlotte Deane. Latfit – producing high accuracy lattice models from protein atomic coordinates including side chains. 2008.

[23] JT Ngo, J Marks, and M Karplus. Computational complexity: Protein structure prediction and the levinthal paradox. *Protein Engineering*, 5:313–321, 1992.

[24] Ro Dal Palú, Agostino Dovier, Enrico Pontelli, and Universitá Di Udine. A new constraint solver for 3d lattices and its application to the protein folding problem. In *In Geoff Sutcliffe and Andrei Voronkov, editors, Logic for Programming, Artificial Intelligence, and Reasoning, 12th International Conference, LPAR 2005, Montego*, pages 48–63. Springer, 2005.

[25] B. H. Park and M. Levitt. The complexity and accuracy of discrete state models of protein structure. *Journal of molecular biology*, 249(2):493–507, June 1995.

[26] Britt Park and Michael Levitt. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *Journal of Molecular Biology*, 258(2):367 – 392, 1996.

[27] Piotr Pokarowski, Karol Droste, and Andrzej Kolinski. A minimal proteinlike lattice model: An alpha-helix motif. *The Journal of Chemical Physics*, 122(21):214915, 2005.

[28] Piotr Pokarowski, Andrzej Kolinski, and Jeffrey Skolnick. A minimal physically realistic protein-like lattice model: Designing an energy landscape that ensures all-or-none folding to a unique native state. *Biophysical Journal*, 84(3):1518 – 1526, 2003.

[29] Simon C. Lovell et al. Structure validation by cα geometry: psi, phi, and cβ deviation. 50.

[30] Yash Thakore. Ramachandran aligner, 2008.

[31] Paul D. Thomas and Kenneth A. Dill. Statistical potentials extracted from protein structures: How accurate are they? *Journal of Molecular Biology*, 257(2):457 – 469, 1996.

[32] Michael Wagner, Jaroslaw Meller, and Ron Elber. Large-scale linear programming techniques for the design of protein folding potentials. *Math. Program.*, 101(2):301–318, 2004.

[33] Guoli Wang and Dunbrack, Roland L. Jr. PISCES: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

7 Appendices

7.1 Proof sketch of approximation ratio for the Stewart FCC-sidechain algorithm

Suppose that the algorithm is input protein P with h bipoles. If there exists a T_i for which $T^* = h$, then there exists an optimal packing in the octahedral conformation as shown by Toma and Toma[18]. Therefore, the number of hydrophobic-hydrophobic contacts is optimal: let $\sum_{i,j} c_{i,j} = OPT(P)$. In the more likely case, h falls in between two T_i values: let the lesser be T^- and the greater be T^*. We may compute the number of contacts in the folded algorithm by supposing a model in which the octahedral conformation for T^* is used, but some contacts are lost due to the absence of hydrophobic residues at this position.

Since this is a worst case analysis, the computation for the approximation ratio will be the minimum over all values of h lying between T^- and T^*:

$$R^* = \lim_{h \to \infty} min_{h \in \{T^- + 1 \ldots T^*\}} \frac{\sum_{(i,j)} c_{i,j}}{OPT(P)} \tag{25}$$

Since the output of the Stewart algorithm gives a conformation which omits some of the hydrophobic ends from the T^* core (the larger core with T^* 'H's), we can compute for any h that $\sum_{i,j} c_{i,j} = c_{i,j}^{T^*} - (\#\ max\ contacts\ of\ (T^* - h)\ hydrophobics)$.We will create a bound for $\sum_{(i,j)} c_{i,j}$ by bounding the contribution of the $T^* - h$ hydrophobics.

In the worst case, we remove $T^* - h$ hydrophobics from the border of the largest plane(s) in the conformation. These hydrophobics form the greatest number of contacts. In fact, since the volume within the core is hollow, we can bound the number of contacts of any one of these hydrophobics to less than 12. The number of contacts formed by a hydrophobic residue removed from the "optimal packing" is 7, since at the boundary of the shape the residue makes contacts with at most two neighbors in its plane, two neighbors in the smaller of the two adjacent planes, and 3 neighbors in the plane of larger size (or possibly equal size in the same of T_i for i even).

To compute the contribution of the lost hydrophobics we must also bound $T^* - h$. The recursion for the

T_is (number of hydrophobics at boundary of the core) is

$$T_n = T_{n-1} + t_{\lceil n/2 \rceil} \tag{26}$$

Substituting T^* for T_n we get

$$T^* = T^- + t_{\lceil n/2 \rceil} \tag{27}$$

$$T^* - h = T^- + t_{\lceil n/2 \rceil} - h \tag{28}$$

$$T^* - h < t_{\lceil n/2 \rceil}, \qquad h > T^- \tag{29}$$

$$T^* \geq t_1 + t_2 + t_3 ... + t_{\lceil n/2 \rceil} > 2 * \textstyle\sum_{i=2}^{\lfloor n/2 \rfloor} 6 * 2^{i-2} \tag{30}$$

$$T^* \geq \epsilon + 1 + t_{\lceil n/2 \rceil} + 2 * \textstyle\sum_{i=2}^{\lfloor n/2 \rfloor} 6 * 2^{i-2}, \qquad \epsilon \geq 0 \;\; when\ n > 2 \tag{31}$$

$$T^* > 2(6 * 2^{(\lfloor n/2 \rfloor - 2)}) + 6 * 2^{(\lceil n/2 \rceil - 2)} \tag{32}$$

$$T^* > 6 * (2^{(\lfloor n/2 \rfloor - 1) + (\lceil n/2 \rceil - 2)}) \tag{33}$$

$$T^* > 2t_{\lceil n/2 \rceil} \tag{34}$$

$$\circ\!\!\circ\, T^* - h < 1/2T^* \tag{35}$$

$$h > 1/2T^* \tag{36}$$

Using the knowledge that we lose no more than $1/2T^*$ hydrophobics (and at most 7 contacts for each), we can bound the number of contacts that the algorithm outputs for protein P:

$$\sum_{i,j} c_{i,j} \geq c_{total}^{T^*} - 7(1/2T^*) \tag{37}$$

and this is true for any value of h by Equation 36. Then it follows that the approximation ratio is the number of hydrophobic-hydrophobic interactions divided by the upper bound on contacts $OPT(P)$ for the set of bipoles P. Recall that the bound depends on the worst possible value in $T^- + 1 ... T^*$, so we minimize

this ratio (equivalent to maximizing the energy).

$$R^* \geq \lim_{h \to \infty} min_{h \in \{T^- + 1 \ldots T^*\}} \frac{c_{total}^{T^*} - 7(1/2T^*)}{OPT(P)} \tag{38}$$

We then use the result from Hart and Istrail that, in the limit as number of hydrophobics n_S in the protein S increases to infinty, the number of optimal contacts $OPT(S) \leq \frac{9n_S}{2}$.[13]. Now we use the fact that we can precompute a perfect packing for T^* bipoles in order to substitute $c_{total}^{T^*}$ with $OPT(T^*)$.

$$R^* \geq \lim_{h \to \infty} min_{h \in \{T^- + 1 \ldots T^*\}} \frac{OPT(T^*) - (7/2)(T^* - h)}{OPT(P)} \tag{39}$$

$$R^* \geq \lim_{h \to \infty} min_{h \in \{T^- + 1 \ldots T^*\}} \frac{(9/2)T^* - (7/2)(T^* - h)}{(9/2)(T^* - h)} \tag{40}$$

The minimum of the function occurs when h is at its minimum, but we have proven (Equation 36) that $T^* - h < 1/2T^*$ so

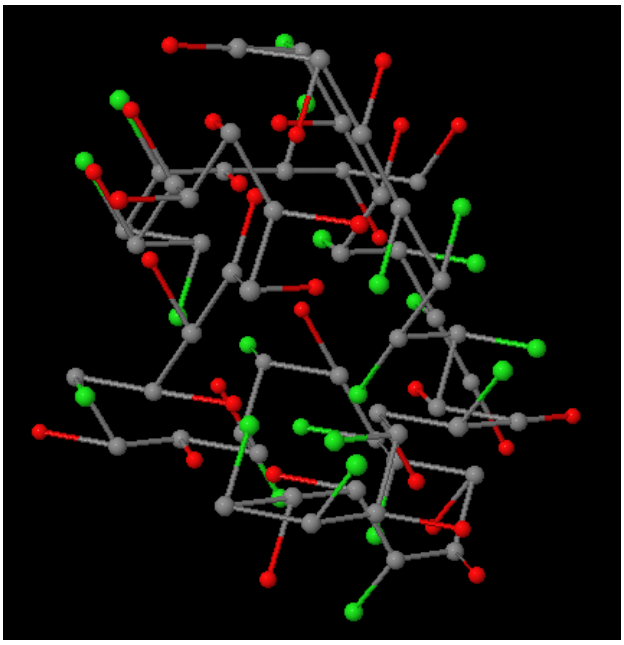
$$R^* \geq \frac{(9/2)T^* - (7/2)T^*}{(9/2)(1/2)T^*} \tag{41}$$

$$R^* \geq \frac{1}{2.25} = \frac{4}{9} = .4\bar{4} \tag{42}$$

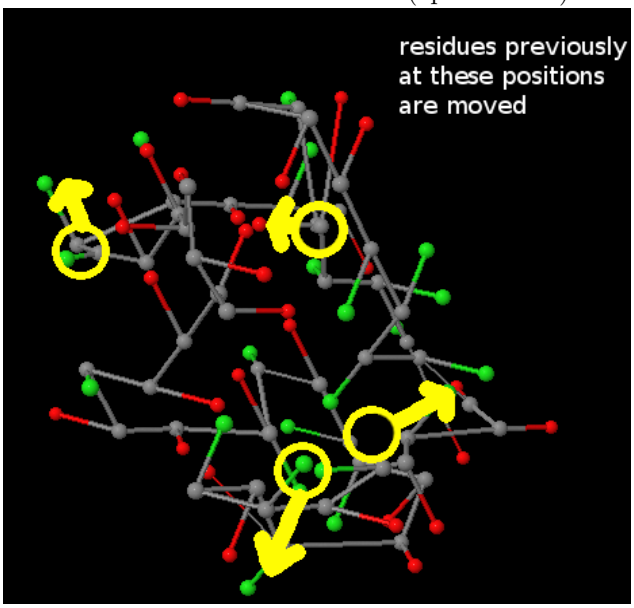
Therefore, we show evidence that the performance of the algorithm (supposing that the construction of the imperfect polyhedron is feasible) returns a solution with at which minimizes the energy function to within 44% of the optimal value for any generic protein. As has been reiterated frequently, the algorithm is not likely to generate a biologically reasonable solution; only globular proteins are characterized entirely by their hydrophobic cores. However, given the previous arguments that a model composed of octahedrons leads to favorable folds which are expressive of protein structure, it is forseeable that applying these principles could develop better algorithms in the future. For example, an algorithm which folds blocks of the protein into their optimal octahedrons and threads the backbone through the domains might fold proteins to their biologically favorable conformations. In addition, it could be adapted to a more complex model than the standard HP interaction matrix (valued at -1 for hydrophobic contacts and 0 otherwise).

7.2 A protein and its decoy: 8RXN

8RXN is a short (52 residue) rubredoxin protein in the bacterium Desulfovibrio vulgaris, with high resolution ($1.0\mathring{A}$ rms).



The following illustrates residues which are moved ("perturbed") to construct a decoy.



7.3 PDB IDs Associated with Biplanar and Non-Biplanar Structure

The tables which follow describe in Table 14 subset of PDB proteins for which the number of hydrophobic residues located on a biplane contributes to at greater than $\tau = 33\%$ of the hydrophobic interactions observed in the lattice-fit molecule and in Table 15 the subset of PDB proteins for which the number of hydrophobic residues located on a biplane contributes to no greater than 33% of the hydrophobic interactions. A larger set of proteins is to be published to `http://www.brown.edu/Research/Istrail_Lab/Allanthesis/biplane.html`.

PDB ID	Title	Resolution	Keyword
1P6O	The crystal structure of yeast cytosine deaminase bound to 4(r)-hydroxyl-3	1.14	
1P7W	Crystal structure of the complex of proteinase k with a designed heptapeptide inhibitor pro-ala-pro-phe-ala-ser-ala at atomic resolution	1.02	Hydrolase
1PB7	Crystal structure of the nr1 ligand binding core in complex with glycine at 1.35 angstroms resolution	1.35	Ligand binding protein
1PIN	Pin1 peptidyl-prolyl cis-trans isomerase from homo sapiens	1.35	Isomerase
1PP0	Volvatoxin a2 in monoclinic crystal	1.42	Toxin
1PSR	Human psoriasin (s100a7)	1.05	Ef-hand protein
1PZ4	The structural determination of an insect (mosquito) sterol carrier protein-2 with a ligand bound c16 fatty acid at 1.35 a resolution	1.35	Lipid binding protein
1WBE	X-ray structure of bovine gltp	1.36	Lipid transport
1WDD	Crystal structure of activated rice rubisco complexed with 2-carboxyarabinitol-1	1.35	
1WM3	Crystal structure of human sumo-2 protein	1.20	Protein transport
1WMH	Crystal structure of a pb1 domain complex of protein kinase c iota and par6 alpha	1.50	Transferase/cell cycle
1WMS	High resolution crystal structure of human rab9 gtpase: a novel antiviral drug target	1.25	Protein transport
1WZD	Crystal structure of an artificial metalloprotein: fe(10-ch2ch2cooh-salophen)/wild type heme oxygenase	1.35	Oxidoreductase
1X6Z	Structure 1: cryocooled crystal structure of the truncated pak pilin from pseudomonas aeruginosa at 0.78a resolution	0.78	Structural protein
1X8Q	Crystal structure of nitrophorin 4 from rhodnius prolixus in complex with water at ph 5.6	0.85	Ligand binding protein
1XAW	Crystal structure of the cytoplasmic distal c-terminal domain of occludin	1.45	Cell adhesion
1YBK	Rhcc cocrystallized with capb	1.45	Protein binding
2CAR	Crystal structure of human inosine triphosphatase	1.09	Hydrolase
2CCV	Structure of helix pomatia agglutinin with zinc and n-acetyl-alpha-d-galactoseamine (galnac)	1.30	Lectin
2CF7	Asp74ala mutant crystal structure for dps-like peroxide resistance protein dpr from streptococcus suis.	1.50	Peroxide resistance
2CFE	The 1.5 a crystal structure of the malassezia sympodialis mala s 6 allergen	1.50	
2CS7	1.2 a crystal structure of the s. pneumoniae phta histidine triad domain a novel zinc binding fold	1.20	Structural genomics
2CVD	Crystal structure analysis of human hematopoietic prostaglandin d synthase complexed with hql-79	1.45	Isomerase
3BLN	Crystal structure of acetyltransferase gnat family from bacillus cereus	1.31	Transferase
3BOI	Snow flea antifreeze protein racemate	1.00	Antifreeze protein
3CJS	Minimal recognition complex between prma and ribosomal protein l11	1.37	Transferase/ribosomal protein
3CX2	Crystal structure of the c1 domain of cardiac isoform of myosin binding protein-c at 1.3a	1.30	Contractile protein
3F5V	C2 crystal form of mite allergen der p 1	1.36	Hydrolase
3SEB	Staphylococcal enterotoxin b	1.48	Toxin

Figure 14: PDB proteins which fit a biplanar characterization per the method described in *Evaluating the biplanar model for hydrophobic collapse.*

PDB ID	Title	Resolution	Keyword
1PKO	Myelin oligodendrocyte glycoprotein (mog)	1.45	Immune system
1PQ7	Trypsin at 0.8 a	0.80	
1PVM	Crystal structure of a conserved cbs domain protein ta0289 of unknown function from thermoplasma acidophilum	1.50	Structural genomics
1PWA	Crystal structure of fibroblast growth factor 19	1.30	Hormone/growth factor
1WL8	Crystal structure of ph1346 protein from pyrococcus horikoshii	1.45	Ligase
1WLU	Crystal structure of tt0310 protein from thermus thermophilus hb8	1.45	Structural genomics
1WVQ	Structure of conserved hypothetical protein pae2307 from pyrobaculum aerophilum	1.45	Structural genomics
1WXC	Crystal structure of the copper-free streptomyces castaneoglobisporus tyrosinase complexed with a caddie protein	1.20	Oxidoreductase/metal transport
1X1Z	Orotidine 5'-monophosphate decarboxylase (odcase) complexed with bmp (produced from 6-cyanoump)	1.45	Lyase
1YBI	Crystal structure of ha33a		1.50
1YCC	High-resolution refinement of yeast iso-1-cytochrome c and comparisons with other eukaryotic cytochromes c	1.23	Electron transport (cytochrome)
2BK9	Drosophila melanogaster globin	1.20	Oxygen transport
2BWQ	Crystal structure of the rim2 c2a-domain at 1.4 angstrom resolution	1.41	Transport protein
2CBZ	Structure of the human multidrug resistance protein 1 nucleotide binding domain 1	1.50	Transport
2CC6	Complexes of dodecin with flavin and flavin-like ligands	1.27	Flavoprotein
2CE0	Structure of oxidized arabidopsis thaliana cytochrome 6a	1.24	Electron transfer
2CG7	Second and third fibronectin type i module pair (crystal form ii).	1.20	Signaling protein
2CIA	Human nck2 sh2-domain in complex with a decaphosphopeptide from translocated intimin receptor (tir) of epec	1.45	Sh2-domain
2CIO	The high resolution x-ray structure of papain complexed with fragments of the trypanosoma brucei cysteine protease inhibitor icp.	1.50	Hydrolase/inhibitor
2CM5	Crystal structure of the c2b domain of rabphilin	1.28	Protein transport
2CVI	Crystal structure of hypothetical protein phs023 from pyrococcus horikoshii	1.50	Structural genomics
2CWS	Crystal structure at 1.0 a of alginate lyase a1-ii	1.00	
3BHD	Crystal structure of human thiamine triphosphatase (thtpa)	1.50	Hydrolase
3BHW	Crystal structure of an uncharacterized protein from magnetospirillum magneticum	1.50	Structural genomics
3BOG	Snow flea antifreeze protein quasi-racemate	1.20	Antifreeze protein
3BPV	Crystal structure of marr	1.40	Transcription regulator
3CI3	Structure of human-type corrinoid adenosyltransferase from lactobacillus reuteri	1.11	Transferase
3CJW	Crystal structure of the human coup-tfii ligand binding domain	1.48	Transcription
3CWR	Crystal structure of transcriptional regulator of tetr family from rhodospirillum rubrum	1.50	Transcription regulator

Figure 15: PDB proteins which are not characterizable as "mostly biplanar". On average, 37% of the residues were part of α-helices while 30% of those residues in the mostly biplanar set were classified as α-helix.