Free Energy Estimates of All-atom Protein Structures Using Generalized Belief Propagation

H. Kamisetty*, E. P. Xing†, C. J. Langmead*

November 2006
CMU-CS-06-160

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

*Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA 15213.
† Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA 15213

E-mail: cjl@cs.cmu.edu

CJL is supported by a Young Pioneer Award from the Pittsburgh Lifesciences Greenhouse and a CAREER award from the U.S. Department of Energy.

Keywords: Protein Structure, Decoy Detection, Free Energy, Probabilistic Graphical Models

Abstract

We present a technique for approximating the free energy of protein structures using Generalized Belief Propagation (GBP). The accuracy and utility of these estimates are then demonstrated in two different application domains. First, we show that the entropy component of our free energy estimates can be used to distinguish native protein structures from *decoys* — structures with similar internal energy to that of the native structure, but otherwise incorrect. Our method is able to correctly identify the native fold from among a set of decoys with 87.5% accuracy over a total of 48 different immunoglobin folds. The remaining 12.5% of native structures are ranked among the top 4 of all structures. Second, we show that our estimates of $\Delta\Delta G$ upon mutation for three different data sets have linear correlations between 0.63-0.70 with experimental values and statistically significant p-values. Together, these results suggests that GBP is an effective means for computing free energy in all-atom models of protein structures. GBP is also efficient, taking a few minutes to run on a typical sized protein, further suggesting that GBP may be an attractive alternative to more costly molecular dynamic simulations for some tasks.

1 Introduction

This paper describes a technique for modeling protein structures as complex probability distributions over a set of torsion angles, represented by a set of rotamers. Specifically, we model protein structures using probabilistic graphical models. Our representation is complete in that it models every atom in the protein. A probabilistic representation confers several advantages including that it provides a framework for predicting changes in *free energy* in response to internal or external changes. For example, structural changes due to changes in temperature, pH, ligand binding, and mutation, can all be cast as inference problems over the model. Recent advances in inference algorithms for graphical models, such as Generalized Belief Propagation (GBP), can then be used to efficiently solve these problems. This is significant because GBP is a rigorous approximation to the free-energy of the system [30]. We will show that these free energy estimates are accurate enough to perform non-trivial tasks within structural biology. In particular, we use GBP to a) identify native immunoglobin structures from amongst a set of decoys with 87.5% accuracy, and b) compute changes in free energy after mutation that have a linear correlation of upto 0.69 to laboratory measurements.

The Gibbs Free energy is defined as $G = E - TS$, where E is the internal energy of the system, T is the absolute temperature, and S is the entropy of the system. There are numerous internal energy functions (i.e., E) from which to choose. These functions often model inter- and intra molecular interactions (e.g., van der Waals, electrostatic, solvent, etc.). Unfortunately, entropy estimates can be difficult to compute because they involve sums over an exponential number of states. For this reason, the entropy term is often ignored altogether, under the assumption that it does not contribute significantly to the free energy. This is equivalent to modeling the system at 0 Kelvin. Not surprisingly, this simplification can sometimes limit the accuracy, and thus the utility, of the energy calculations. For example, it has been conjectured [25, 3] that energy functions comprising sums of pairwise interactions cannot distinguish a protein's native structure from decoy structures within about 1 Å RMSD. If true, one likely explanation is that entropy contributions become significant when structures are similar. Our findings are consistent with this hypothesis. In particular, we find that the native structure is usually the one with the highest entropy.

Numerous investigators have observed and attempted to address the limitations of energy functions. Statistical potentials are common alternative (e.g., [6, 23]). Such potentials do not model the physics directly, but instead use statistics mined from the Protein Data Bank [2] under the assumption that these statistics encode both the entropy and the internal energy. Carter and coworkers [6], for example, have developed a 4-body statistical potential that predicts $\Delta\Delta G$s upon mutations with significant accuracy. There are, however, those that doubt the ultimate utility of statistical potentials (e.g., [24]). Our $\Delta\Delta G$ predictions achieve a high linear correlation (0.71) with experimentally measured quantities. This is consistent with the findings of others who have have demonstrated the practical benefits of including entropy in energy calculations (e.g., [13]).

We note that the contributions of this paper do not lie in the suggestion that a protein's structure be treated as a probability distribution — clearly this is the very essence of statistical physics. Rather, our contribution lies in the demonstration that an inference-based approach to free energy calculations is sufficiently accurate to perform non-trivial tasks. Additionally, our technique is efficient and runs in minutes on typical-sized proteins, suggesting it is well-suited for large-scale

proteomic studies.

2 A Markov Random Field Model for Protein Structure

In what follows, random variables are represented using upper case variables, sets of random variables appear in bold face while lower case variables represent specific values that the random variables can take. Thus, the random variables representing the position of all the backbone atoms is denoted by $\mathbf{X_b}$, those representing all the side chain atoms, by $\mathbf{X_s}$, X_s^i is the random variable representing the side chain conformation of the i^{th} residue and x_b^i represents a particular value that the backbone of the i^{th} residue takes.

Let $\mathbf{X} = \{\mathbf{X_b}, \mathbf{X_s}\}$ be the random variables representing the entire protein structure. $\mathbf{X_b}$ can be represented by a set of 3-d coordinates of the backbone atoms, or equivalently, by a sequence of bond lengths and dihedral angles. Thus, X_b is typically a continuous random variable. Each X_s^i, is usually represented by a set of dihedral angles [1] . While this too is a continuous random variable, due to steric clashes not all dihedral angles are energetically favorable, allowing a discretization of this state space into a set of discrete favorable conformations called *rotamers*.

The probability of a particular conformation $\mathbf{x}$ can be written as

$$p(\mathbf{X} = \mathbf{x}|\Theta) = p(\mathbf{X_b} = \mathbf{x_b})p(\mathbf{X_s} = \mathbf{x_s}|\mathbf{X_b}, \Theta)$$

or more compactly,

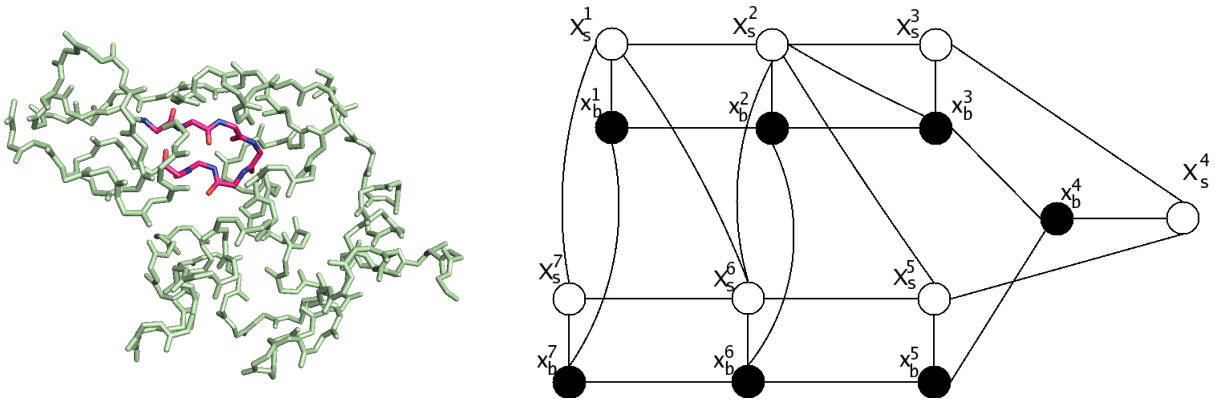
$$p(\mathbf{X}|\Theta) = p(\mathbf{X_b})p(\mathbf{X_s}|\mathbf{X_b}, \mathbf{\Theta})$$

where Θ represents any parameters used to describe this model, including sequence information, temperature etc. Frequently the backbone is assumed to be rigid with a known conformation. Therefore $\mathbf{X_b} = \mathbf{x_b}$ for some particular $\mathbf{x_b}$. The term of interest then becomes, $p(\mathbf{X_s}|\mathbf{X_b} = \mathbf{x_b}, \mathbf{\Theta})$.

This can be further simplified. Specifically, it is possible to list out conditional independencies that the above probability distribution must satisfy. Consider the random variables X_s^i, X_s^j representing the side chain conformations of residues i, j. Due to the underlying physics, if the residues are not close to each other, their direct influence on each other is negligible. Also, if the residues that directly influence these residues are in specific conformations, X_s^i, X_s^j become conditionally independent of each other. Similar independencies can be listed between side chain variables and backbone variables. These conditional independencies can be compactly encoded using an undirected probabilistic graphical model, also called a Markov Random Field(MRF).

For example, consider a particular backbone conformation $\mathbf{x_b}$ of Lysozyme(pdb id: 2lyz) shown in Fig. 1(a) with a few residues highlighted. Fig. 1(b) shows that part of the markov random field that is induced by the highlighted set of residues. Two variables share an edge if they are closer than a threshold distance. Edges can thus be present between backbone atoms, between backbone and side chain atoms and between side chain atoms.

[1] This is a slight abuse of notation, since it is actually the differences $X_b^i - X_b^{i-1}$ and $X_s^i - X_b^i$ that are represented using bond lengths and angles.



(a) (b)

Figure 1: (a) Structure of lysozyme (pdb id: 2lyz) with a few residues highlighted (b) Part of the random field induced by the highligted residues: X_s^i's are the hidden variables representing the rotameric state, the visible variables are the backbone atoms in conformations x_b^i

In general, a markov random field encodes the following conditional independencies for each vertex X_i and for any set of vertices $\mathbf{X}'$ not containing X_i.

$$p(X_i|\mathbf{X}', Neighbors(X_i)) = p(X_i|Neighbors(X_i))$$

That is, a random variable X_i is conditionally independent of every other set of nodes in the graph, given its immediate neighbors in the graph.

Given this representation, the probability of a particular side chain conformation $\mathbf{x_s}$ given the backbone conformation $\mathbf{x_b}$ can be expressed as

$$p(\mathbf{X_s} = \mathbf{x_s}|\mathbf{X_b} = \mathbf{x_b}) = \frac{\mathbf{1}}{\mathbf{Z}} \prod_{\mathbf{c} \in \mathbf{C(G)}} \psi_{\mathbf{c}}(\mathbf{x_s^c}, \mathbf{x_b^c})$$

where $C(G)$ is the set of all cliques in G, ψ is a potential defined over the variables, and Z is the so called partition function.

To completely characterize the MRF, it is necessary to define the potential function ψ. A common simplifying assumption is that of a pair-wise potential. We use the Boltzmann Distribution to define a pairwise potential function in the following manner:

$$\psi(X_s^{i_p}, X_s^{j_q}) = exp(-E(x_s^{i_p}, x_s^{j_q})/k_B T)$$

where E_{i_p,j_q} is the energy of interaction between rotamer state p of residue X_s^i and rotamer state q of residue X_s^j and k_B is the Boltzmann constant. Similarly, we can define the potential function between a side chain random variable X_s^i and a backbone random variable X_b^j which is in an observed state x_b^j

$$\psi(X_s^{i_p}, X_b^j) = exp(-E(x_s^{i_p}, x_b^j)/k_B T)$$

Finally, we define the potential function between two backbone random variables to have the trivial value of 1, since both are observed, i.e. $\psi(X_b^i, X_b^j) = 1$.

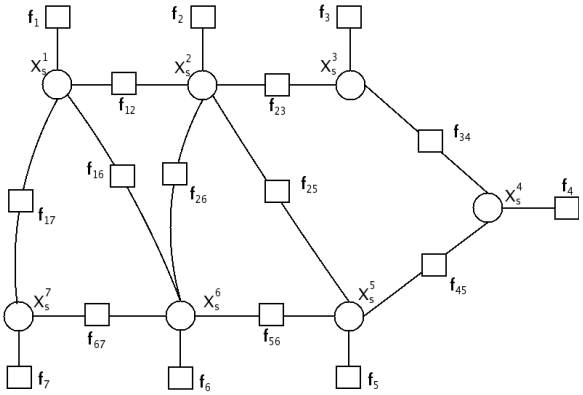


Figure 2: Factor graph representation for the graph shown in Fig. 1(b). The observed variables corresponding to the backbone atoms can be replaced by a factor at each side chain variable

This undirected graphical model, characterized by $(\mathbf{X}, E, \psi)$ can also be represented more conveniently, as a bipartite graph $(\mathbf{X}, F)$, called a factor graph. If we restrict ourselves to pairwise potentials, as we have done already by our form of potential function, the equivalent factor graph for the MRF of Fig. 1(b) is shown in Fig. 2. Each edge between side chain variables has been replaced by edges to a factor representing the interaction between these variables. Also, it can be shown that the observed backbone variables can be dropped from the factor graph by replacing their interactions with each side chain variable by a factor. The probability of a particular conformation can then be expressed using the factor notation, as

$$p(\mathbf{x_s}) = \frac{1}{Z} \prod_{f_a \in F} f_a(\mathbf{x_s^a})$$

where $\mathbf{X_s^a}$ is the set of variables connected to factor f_a in the factor graph.

3 Approximating Free Energy

A corollary of the second law of thermodynamics is that a physical system seeks to minimize its free energy. Thus, the most accurate entropy estimates are obtained when the system has the least free energy. Under the assumption of constant temperature, the Gibbs free energy of a system is given by

$$G = E - TS$$

where E is the enthalpy of the system, T the temperature and S, the entropy. If we associate a belief $b(\mathbf{x})$ with state $\mathbf{x}$, this can be rewritten as

$$G = \sum_{\mathbf{x} \in S} b(\mathbf{x}) E(\mathbf{x}) + T \sum_{\mathbf{x} \in S} b(\mathbf{x}) ln(b(\mathbf{x}))$$

where the first term and second terms on the right are the enthalpic and entropic contributions respectively. Intuitively, the enthalpic term corresponds to the energy of the system. However, the second law of thermodynamics dictates that not all energy can be used to do work. The free energy is the energy left to be used to do work after deducting the energy that is "lost" which is the entropic deduction mentioned above.

There has been a considerable amount of work by physicists at developing approximations to estimate these terms [4, 10, 18, 19]. The popular methods are based on approximating the free energy using a *region based* free energy. Intuitively, the idea is to break up the factor graph into a set of regions R, each containing multiple factors $\mathbf{f_R}$ and variables $\mathbf{X_R}$, compute the free energy over the region using estimates of the marginal probability over $\mathbf{X_R}$, and then approximate the total free energy by the sum of the free energies over these regions. Since the regions could overlap, contributions of nodes – factors or variables – which appear in multiple regions have to be subtracted out, so that each node is counted exactly once. This can be done by associating weights c_R to the contribution of every node in region R, in such a way that the sum of weights of the regions that the node appears in, sums to one.

This region graph formalism is fairly general and one can create approximations of varying degrees. For example, the Bethe approximation[4] is a region graph with each region containining atmost one factor, while the Kikuchi approximation is a region graph where the regions are created using the so-called *cluster variational approach* that allows regions to contain more than one factor, and is therefore a better approximation[10, 30].

While the Kikuchi approximation has been extensively studied, until recently, there was a dearth of algorithms that could compute such region graph based approximations efficiently. See [20] for a recent survey of previously used methods and their performance relative to GBP. In fact, even computing exact marginals for the purpose of computing these approximations is NP-Hard, if the graph, like the MRF described above, has cycles. The Junction Tree algorithm for exact inference has a running time that is exponential in the tree width of the graph, which can be prohibitively expensive in large graphs. However, recent advances within the Machine Learning community on approximate algorithms for inference now allow efficient computation of these approximations [30, 28].

3.1 Generalized Belief Propagation

Generalized Belief Propagation(GBP) is a message passing based algorithm that approximates the true marginals. As the name suggests, it is a generalization of the famous Belief Propagation(BP) algorithm, due to Pearl, and differs from the latter in the size of its regions that estimate the Free Energy. While BP attempts to find a fixed point of the Bethe approximation to the free energy mentioned above, GBP computes fixed points of the more general region based free energy.

There are many variants of GBP; we focus on the so called *Two-Way* [30] algorithm since it naturally extends BP. The algorithm can be viewed as running BP on the region graph, with one crucial difference in the messages – since the same node can appear in multiple regions, its contribution to each region must be weighed in such a way as to ensure it is counted only once. This is done, by first defining the "pseudo" messages for a region R with parents $P(R)$ and children

$C(R)$

$$n^0_{R\to P}(\mathbf{x_r}) = \tilde{f}_R(\mathbf{x_R}) \prod_{P'\in P(R)\setminus P} m_{P'\to R}(\mathbf{x_r}) \prod_{C\in C(R)} n_{C\to R}(\mathbf{x_C})$$

$$m^0_{R\to C}(\mathbf{x_C}) = \sum_{x_R\setminus x_C} \tilde{f}_R(x_R) \prod_{P\in P(R)} m_{P\to R}(\mathbf{x_R}) \prod_{C'\in C(R)\setminus C} n_{C'\to R}(\mathbf{x_{C'}}),$$

where $\tilde{f}_R(\mathbf{x_R}) = (\prod_{a\in A_r} f_a(\mathbf{x_a}))^{c_R}$ and then compensating for overcounting by defining the actual messages as

$$n_{R\to P}(\mathbf{x_r}) = (n^0_{R\to P}(\mathbf{x_r}))^{\beta_R}(m^0_{R\to C}(\mathbf{x_C}))^{\beta_R - 1}$$

$$m_{P\to R}(\mathbf{x_r}) = (n^0_{R\to P}(\mathbf{x_r}))^{\beta_R - 1}(m^0_{R\to C}(\mathbf{x_C}))^{\beta_R}$$

where c_R is the weight given to region R, p_R the number of parents of region R, and $\beta_R = p_R/(2p_R + c_R - 1)$. The beliefs at R, are then given by

$$b_R(x_R) = \tilde{f}_R(\mathbf{x_R}) \prod_{C\in C(R)} n_{C\to R}(\mathbf{x_C}) \prod_{P\in P(R)} m_{P\to R}(\mathbf{x_P})$$

Note that if $\beta_R = 1$, this algorithm becomes equivalent to running BP directly on the region graph.

The algorithm is typically started with randomly initialized messages and run until the beliefs converge. If it does converge, GBP is guaranteed to find a fixed point of the region based free energy. While convergence isn't guaranteed, in practice, it has been found to converge successfully in many cases, even when BP doesn't [27, 29] .

3.2 Related Work

Probabilistic graphical models have been used to address a number of problems in structural biology, primarily in the area of secondary structure prediction (e.g., [7]). Applications of graphical models to tertiary structure are generally limited to applications of Hidden Markov Models (HMMs) (e.g., [9]). HMMs make severe independence assumptions to allow for efficient learning and inference, the result of which is that long-range interactions cannot be modeled. Long-range interactions are, of course, found in all protein structures. Our method models these long range interactions. Graphical models have also been used in the area of fold recognition/threading [14]. An important difference between threading and our work is that we model every atom in the structure, while threading is generally performed over reduced representations.

We focussed on the problem of computing entropy using marginal probabilities for the unobserved variables, $\mathbf{X_s}$. This however isn't the only interesting inference problem. If our task was to find the *single* most likely structure, the problem reduces to Side Chain Placement. Indeed, one of the recent approaches to this problem of placing side chains [26] can be viewed as a variant of the Junction Tree algorithm for computing the most likely estimate.

It must be noted that our model is essentially similar to that of [27]. While they use it in a study to evaluate inference algorithms and perform Side Chain Placement, our task is to use it to obtain entropy and free energy estimates.

Recent work [16] has shown that most message passing algorithms can be viewed as minimizing the divergence between the actual probability distribution and a family of distributions suitably parametrized. The different algorithms differ in their choice of the divergence measure and their parametrization of the family of distributions. [11] attempts to solve the same problem using a mean-field approach. Mean field methods minimize the Kullback-Leibler Divergence while Generalized Belief Propagation (and BP) minimize an "inclusive" divergence. While the former is more accurate at capturing the zeros of the actual distribution, the latter performs better at predicting marginals. As we have shown in this section, marginal probabilities allow us to compute estimates of the entropy and free energy of the distribution. Thus, Generalized Belief Propagation is more suitable for the problem at hand.

4 Implementation and Results

We implemented the *Two-way* GBP algorithm described earlier, to compute region graph estimates of free energy and entropy. We parsed the pdb files using the pdb parser in the Molecular Biology Toolkit [17]. We then created the factor graph by computing interatomic distances and creating a factor between residues if the C_α distance between them was lesser than a threshold value. This threshold is largely dictated by the sensitivity of the energy function. For the energy terms we used, we found a threshold of 8.0 Åto be adequate. In the few datasets that we tested, our results were not affected by small changes in this threshold. We used the backbone dependent library provided by [5] and a linear approximation to the repulsive van der Waals force used by [5, 27]. Each rotamer in the library also had an associated apriori probability which we incorporated into the factor as a prior. We set the Temperature of the system to be 300K, which corresponds to normal room temperature.

We used a region graph construction which created two levels of regions. The top level contained "big" regions – regions with more than one variable – while the lower level contained regions representing single variables. Since we expect the interaction between residues closest in sequence to be very strong, we placed all factors and nodes between residues within two sequence positions of each other in one region. Each of the rest of the factors, representing edges between residues connected in space, formed "big" regions with two nodes in them. Thus, in the example shown in Fig. 2, $(X_s^1, X_S^2, X_S^3, f_1, f_2, f_3, f_{12}, f_{23})$, $(X_s^2, X_S^3, X_S^4, f_2, f_3, f_4, f_{23}, f_{34})$ and (X_s^1, X_s^7, f_{17}) would be examples of big regions which appear in the top level, while (X_s^1) would be an example of a small region in the lower level. Finally, we add edges from "big" regions to all small regions that contain a strict subset of the "big" region's nodes. In our example, the region encompassing X_s^1, X_s^2, X_s^3 would thus be connected to the small regions corresponding to each of X_s^1,X_s^2, and X_s^3.

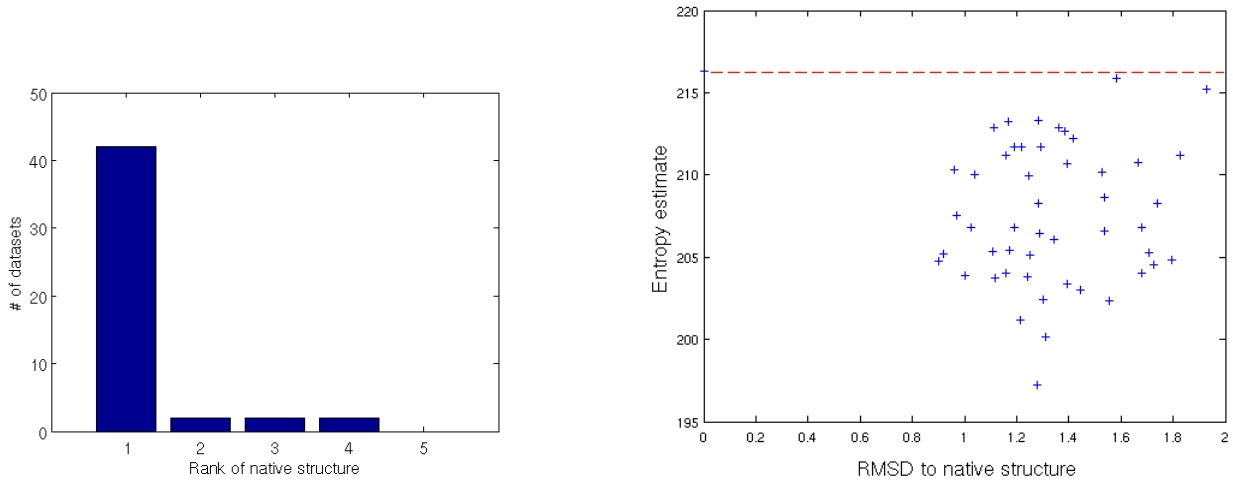
Since the region graph formalism is very flexible, other equally valid alternatives for creating the graph exist. The best choice of regions will largely depend on the application at hand and the computational constraints. Our choice of regions reflects a balance between accuracy and running time by focussing on residues which are expected to be closely coupled together and placing them in bigger regions. [1] studies this class of region graphs in more detail.

We initialized the GBP messages to random starting points and ran until beliefs converged or

a maximum number of iterations(100) was reached. It must be noted that we did not have any problems with convergence: the beliefs converged in all cases.

We ran our program on datasets obtained from the "Decoys R Us" database[22]. We used the immunoglobin datasets from the "multiple decoy sets". Each such dataset consisted of multiple decoy structures along with the native structure of a protein. We selected immunoglobin because it had a large number of decoys close to the native structure and has been used extensively to test methods for decoy detection[23].

Under our assumption of a rigid backbone, our estimates of entropy of different structures will be comparable only when the other sources of entropy are largely similar. Thus, our estimates will be most relevant only when the structures have largely similar backbones. To ensure that we didn't have backbones very different from the native structure among our decoys, we removed all decoys with a C_α RMSD greater than 2.0 Åto the native structure, from each dataset. We then removed any dataset that ended up with less than 5 decoys so that we didn't end up with too few decoys in a dataset. We also removed three datasets where our program crashed on the native structure due to missing backbone atoms. Since this happened on very few small fraction of the cases, we donot expect this to affect our results. At the end of this pruning, there were 48 datasets left with an average of around 35 decoys per data set.



(a) (b)

Figure 3: (a) Histogram shows the distribution of the rank of the native structure, when ranked in decreasing order of entropy for the culled immunoglobin decoy dataset. Over this dataset, the native structure has the highest entropy 87.5% of the time(b) Entropy estimates for 1mlb and its decoys with the value of the entropy along the Y-axis and the rmsd to native structure along the X-axis. The horizontal line indicates the value of the entropy of the native structure; all other structures have a lower entropy in this dataset

Fig. 3 shows our results on the immunoglobin dataset. When we ranked the structures in the decreasing order of their entropy, the native structure ended up at the top in 42 of the 48

datasets(87.5%). In no dataset was the native structure ranked higher than 4. Fig. 3(b) shows the scatter plot of the entropy estimates for a dataset where the native structure(1mlb) has the highest entropy.

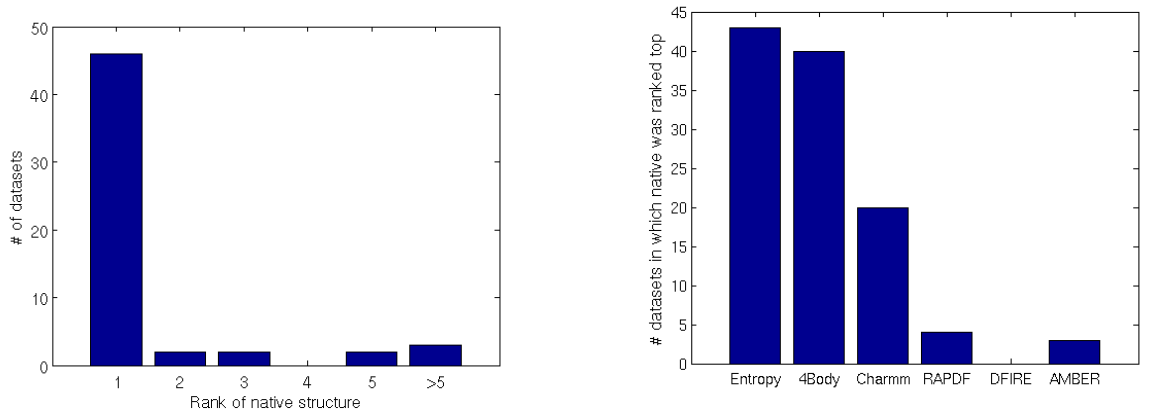
To study the structures further, we ran PROCHECK[12] – a program for structure validation that runs a suite of structural tests. PROCHECK reported a very high number of main chain bond angles (nearly 13 angles on an average) as "off graph" – bond angles so far away from the mean that they don't show up on the output plots of PROCHECK – for the four native structures which have a rank three or four. For example, a total of 27 angles were determined to be "off graph" for 1igc. In contrast, there were an average of around 2 such angles, among the rest of the structures. It must be noted that, not all datasets in which the native structure had bad main chain bond angles had a decoy as the best rank. 1jel, for example, had 21 main chain bond angles "off graph" and yet had the best rank among its dataset. This is not unexpected, since the rank of the native structure is not only determined by its quality, but also by the quality of the decoys. Thus, our results seem to be affected, but not solely determined, by unusual main chain conformations.

Since the structures have very similar backbones, we expect that the entropic contributions from the backbone atoms and our entropy estimates to be most meaningful in relative *order* and *magnitude*. However, in order to test the efficacy of these estimates in decoy detection, we re-run our experiments on the entire immunoglobin dataset. Our hope is that while the magnitudes of the entropy estimates might not be meaningful, the relative order of the native structure will still be useful.

Figure Fig. 4(a) shows the results of our experiments on the entire immunoglobin dataset. As can be seen, despite the addition of the dissimilar backbones, the ranking of the native structure isn't affected much – in 84% of the datasets, the native structure has the highest entropy. We then compare our results to various different energy functions as reported in [23]. Again, our entropy estimate, calculated using a simple linear potential function outperforms all other methods on this dataset.

Thus these results show that our entropy estimates are very successful in detecting the native structure from a set of decoys. However, they do not provide any evidence about the relative magnitude of these estimates. To test this, we perform a different experiment. We compare experimentally determined values of difference in the free energy between the native structures of Barnase, T4 Lysozyme and Staphylococcal Nuclease(pdb ids: 1BNI, 1L63 and 1STN respectively) and multiple single point mutants of them, selected from the ProTherm database[**?**]. Only mutations in buried positions were considered in order to minimize the effects of the solvent. Care was taken to ensure that the $\Delta\Delta G$ experiments were conducted at similar pH values.

Since these mutants have different sequences, the entropy of the denatured state has to be estimated along with that of the crystal structure, in order to estimate $\Delta\Delta G$ values. We estimate the free energy of the denatured state by computing the free energy of the system before inference. Fig. 5 shows our results on the three datasets. The correlation coefficient between our estimates of $\Delta\Delta G$ and the experimentally determined values varied from 0.63 to 0.70 with p values between $1.5*10^{-5}$ to 0.0063. This compares favorably with the estimates – correlations between 0.7 and 0.94 – obtained using the four body potential of [6] over all their (smaller) datasets. This gives evidence that our estimates predict both the sign and the magnitude of $\Delta\Delta G$ with reasonable



(a) (b)

Figure 4: (a) Histogram showing the distribution of the rank of the native structure. (b) Comparison of Results using various energy functions as reported in [23], along with rankings based on our Entropy estimates. 4body refers to the four body atomic potential described in [23]. The other energy functions are described in detail in [23]. These results are on the 51 Immunoglobin datasets for which data was available, including decoys with RMSD greater than 2.0 Å. Overall, the entropy estimates outperform all energy functions.

accuracy.

5 Conclusions

We have shown that free energy calculations for all-atom models of protein structures can be computed efficiently using Generalized Belief Propagation. Moreover, these estimates are sufficiently accurate to perform non-trivial tasks. We first demonstrated that it is possible to identify native immunoglobin structure from a set of decoys, with high accuracy, by comparing the computed entropies. We then demonstrated that our $\Delta\Delta G$ predictions for a set of mutations achieved high linear correlations with experimentally measured quantities. This suggests that our predictions for the change in entropy are not only in the right general direction, but are approximately the right order of magnitude.

Our results have implications for a number of problem domains. First, we believe that our method could be used in the contexts of protein structure prediction and comparative modeling. Our decoy-detection results suggest that our method could be used in conjunction with protein structure prediction programs that produce multiple putative folds, like ROSETTA [21]. The accuracy of existing homology modeling methods is acknowledged to be an important issue in structural biology (e.g., [15, 8]). We are presently extending our technique to allow backbone flexibility. This would facilitate refining of homology models towards a lower free-energy configuration, and potentially higher accuracy. Second, we note that one of the advantages of a graphical model is that it

is easily extended. For example, we could enhance our edge potentials to incorporate experimental measurements from X-ray crystallography, Nuclear Magnetic Resonance, or Cryogenic Electron microscopy. These enhancements could be very beneficial in the context of structure determination experiments where the data are sparse or low-resolution. Third, we can also extend our model to include ligands by adding nodes to our graph. This, plus a combination of a backbone flexibility and a somewhat more sophisticated energy potential may lead to more accurate $\Delta\Delta G$ calculations which, in turn, may be useful in the context of ligand binding and docking studies. Finally, while our experiments assumed a known protein sequence, it is possible to simultaneously perform inference over the sequence and structure, leading to new techniques for performing protein design. We are actively pursuing these goals as part of ongoing research into the application of graphical models to protein structures.

Acknowledgments

We thank K. Arun and Jianhua Shao of the Langmead lab and all members of the CBML group for their valuable comments. This research was supported by a U.S. Department of Energy Career Award (DE-FG02-05ER25696), and a Pittsburgh Life-Sciences Greenhouse Young Pioneer Award to C.J.L.

References

[1] S.M. Aji and R. J. McEliece. The generalized distributive law and free energy minimization. *Proceedings of the 39th Allerton Conference on Communication, Control and Computing*, pages 459–467, 2003.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucl. Acids Res.*, 28:235–242, 2000.

[3] M. R. Betancourt and D. Thirumalai. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Science*, 8:361–369, 1999.

[4] H. A. Bethe. Statistical theory of superlattices. *Proc. Roy. Soc. London A*, 150:552–575, 1935.

[5] A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack Jr. A graph theory algorithm for protein side-chain prediction. *Protein Science*, 12:2001–2014, 2003.

[6] C.W. Carter Jr, B.C. LeFebvre, S.A. Cammer, A. Tropsha, and M.H. Edgell. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *Journal of Mol. Bio*, 311:625–38, 2001.

[7] W. Chu, Z. Ghahramani, and D. Wild. A graphical model for protein secondary structure prediction. *Proc. 21st Ann. Intl. Conf. on Machine Learning (ICML) Banff, Canada)*, 2004.

[8] Protein Structure Initiative. Report on the nigms workshop on high accuracy comparative modeling. http://archive.nigms.nih.gov/psi/reports/comparative_modeling.html, 2003.

[9] K. Karplus, R. Karchin, J. Draper, J. Casper, Y. (Mandel-Gutfreund), M. Diekhans, and R. Hughey. Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction. *Proteins*, 53:491–6, 2003.

[10] R. Kikuchi. A theory of cooperative phenomena. *Phys. Rev*, 81:988–1003, 1951.

[11] P. Koehl and M. Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *Journal of Mol. Bio.*, pages 249–275, 1994.

[12] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. Procheck: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26:283–291, 1993.

[13] R. Lilien, B. Stevens, A. Anderson, and B. R. Donald. A Novel Ensemble-Based Scoring and Search Algorithm for Protein Redesign, and its Application to Modify the Substrate Specificity of the Gramicidin Synthetase A Phenylalanine Adenylation Enzyme. *J. Comp Biol*, 12(6-7):740–761, 2005.

[14] Y. Liu, J. Carbonell, P. Weigele, and V. Gopalakrishna. Segmentation conditional random fields (scrfs): A new approach for protein fold recognition. *Proc. of the 9th Ann. Intl. Conf. on Comput. Biol. (RECOMB) Boston, MA, May 14-18)*, pages 408–422, 2005.

[15] M.A. Marti-Renom, A. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative Protien Structure Modeling of Genes and Genomes. *Ann Rev Biophys Biomol Struct*, 29:291–325, 2000.

[16] T. Minka. Divergence measures and message passing. *Microsoft Technical Report*, 2005.

[17] J.L. Moreland, A.Gramada, O.V. Buzko, Qing Zhang, and P.E. Bourne. The molecular biology toolkit (mbt): A modular platform for developing molecular visualization applications. *BMC Bioinformatics*, 6, 2005.

[18] T. Morita. Cluster variation method for non-uniform ising and heisenberg models and spin-pair correlation function. *Prog. Theor. Phys.*, 85:243 – 255, 1991.

[19] T. Morita, T. M. Suzuki, K. Wada, and M. Kaburagi. Foundations and applications of cluster variation method and path probability method. *Prog. Theor. Phys. Supplement*, 115, 1994.

[20] A. Pelizzola. Cluster variation method in statistical physics and probabilistic graphical models. *J. phys. A : math. gen*, 38:R309–R339, 2005.

[21] C.A. Rohl, C.E. Strauss, K.M. Misura, and D. Baker. Protein structure prediction using Rosetta. *Methods Enzymol*, 383:66–93, 2004.

[22] R. Samudrala. Decoys 'R' Us. http://dd.compbio.washington.edu/.

[23] C.M. Summa, M. Levitt, and W.F. Degrado. An atomic environment potential for use in protein structure prediction. *Journal of Mol. Bio.*, 352:986–1001, 2005.

[24] P. D. Thomas and K. A. Dill. Statistical potentials extracted from protein structures: How accurate are they? *Journal of Mol. Bio.*, 257:457–469, 1994.

[25] R. Tobi, D. Elber. Distance-dependent, pair potential for protein folding: Results from linear optimization. *Proteins: Structure, Function and Genetics*, 41:40–46, 2000.

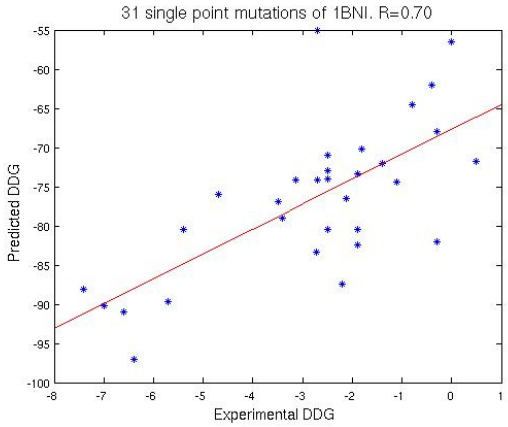
[26] J. Xu. Rapid protein side-chain packing via tree decomposition. *Research in Computational Molecular Biology, Lecture Notes in Computer Science*, 3500:423–439, 2005.

[27] C. Yanover and Y. Weiss. Approximate inference and protein folding. *Proceedings of NIPS 2002*, pages 84–86, 2002.

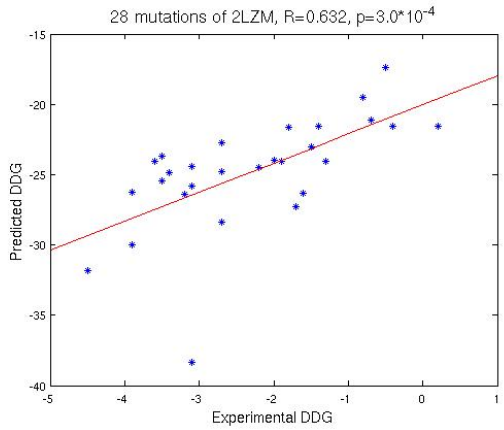
[28] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Generalized belief propagation. *Advances in Neural Information Processing Systems (NIPS)*, 13:689–695, 2000.

[29] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Characterizing belief propagation and its generalizations. http://www.merl.com/reports/TR2002-35/, 2002.

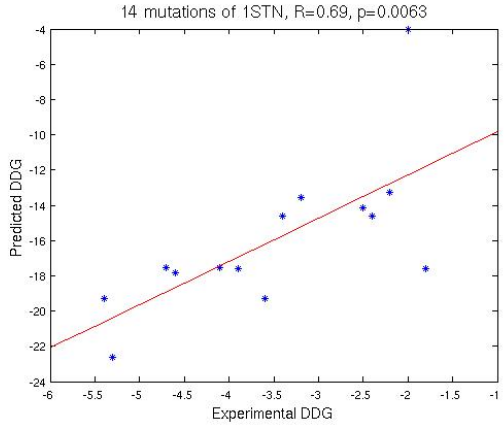
[30] J.S. Yedidia, W.T. Freeman, and Weiss Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51:2282–2312, 2005.



(a)



(b)



(c)

Figure 5: Plots showing variation of experimental $\Delta\Delta G$ (on the X-axis) with computed estimates of $\Delta\Delta G$, along with a least squares fit for (a) thirty one mutants of barnase (pdb id: 1BNI) (b) twenty eight mutants of T4 Lysozyme(pdb id:1L63) and (c) fourteen mutants of staphylococcal nuclease(pdb id:1STN)