JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 7, 2008
© Mary Ann Liebert, Inc.
Pp. 755–766
DOI: 10.1089/cmb.2007.0131

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

HETUNANDAN KAMISETTY, ERIC P. XING, and CHRISTOPHER J. LANGMEAD

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 useful in distinguishing 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 immunoglobulin folds. The remaining 12.5% of native structures are ranked among the top four of all structures. Second, we show that our estimates of $\Delta\Delta G$ upon mutation upon mutation for three different data sets have linear correlations of 0.63–0.70 with experimental measurements and statistically significant *p*-values. Together, these results suggest 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.

Key words: free energy inference, belief propagation, protein structure.

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 (Yedidia et al., 2005). We will show that these free energy estimates are accurate enough to perform non-trivial tasks

Computer Science Department, Carnegie Mellon University, Pittsburgh, Pennsylvania.

within structural biology. In particular, we use GBP to **(a)** identify native immunoglobulin 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 0.70 to laboratory measurements.

The free energy is defined as $G = E - TH$, where E is the enthalpy of the system, T is the absolute temperature, and H is the entropy of the system.[1] There are numerous 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). 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 (Betancourt and Thirumalai, 1999; Tobi and Elber, 2000) 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. This is in agreement with the findings of others who have demonstrated the practical benefits of including entropy in energy calculations (Lilien et al., 2005).

Numerous investigators have observed and attempted to address the limitations of pairwise energy functions. Multi-body statistical potentials are a common alternative (Carter et al., 2001; Summa et al., 2005). Such potentials do not model the physics directly, but instead use statistics mined from the Protein Data Bank (Berman et al., 2000) under the assumption that these statistics encode both the entropy and the internal energy. Carter et al. (2001), 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 (Thomas and Dill, 1994).

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 ith residue and x_b^i represents a particular value that the backbone of the ith 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 three-dimensional (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.[2] The probability of a particular conformation $\mathbf{x}$ can be written as

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

or more compactly,

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

[1] To avoid confusion, we will use the symbol "H" to represent entropy, instead of the more traditional "S," which we will use to denote side chain atoms.

[2] 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.

where $\mathbf{\Theta}$ represents any parameters used to describe this model, including sequence information, temperature etc.

Note that $P(\mathbf{X})$ is a distribution over the *entire* phase space of the molecule. In this paper, we make three common simplifying assumptions: *First*, we assume that bond-lengths are fixed. This assumption is justified based on the fact that bond lengths vary over extremely fast time-scales. The kinds of motion considered in this paper occur at much slower rates, so it is reasonable to fix the length of each bond at some expected value. These lengths are obtained from a given crystal structure. *Second*, while each X_s^i is, in principle, a continuous random variable, due to steric clashes not all dihedral angles are energetically favorable. We therefore allow side-chains to vary over a discrete set of favorable conformations called *rotamers*. These rotameric configurations can be obtained from any of a number of publicly-available rotamer libraries (Lovell et al., 2000). *Finally*, it is frequently assumed that the backbone configuration is known (e.g., from a crystal-structure), and rigid. Under this assumption, $P(\mathbf{X_b} = \mathbf{x_b}) = \mathbf{1}$, for some particular given $\mathbf{x_b}$. The distribution of interest then becomes, $P(\mathbf{X_s}|\mathbf{X_b} = \mathbf{x_b}, \mathbf{\Theta})$. That is, the probability distribution over the side chain configuration, given the configuration of the backbone and the parameters of the model.

This distribution $P(\mathbf{X_s}|\mathbf{X_b} = \mathbf{x_b}, \mathbf{\Theta})$ can be simplified further. Specifically, it is possible to list out conditional independencies that this 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 these two residues are not close to each other, their direct influence on each other is negligible. Also, if $\mathbf{X}_s^*$ is the set of residues which strongly influence either X_s^i or X_s^j, then X_s^i and X_s^j become conditionally independent given $\mathbf{X}_s^*$. That is, $X_s^i \perp X_s^j|\mathbf{X}_s^*$. 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).

A MRF is a compact encoding of $P(\mathbf{X})$ as an undirected graph and a set of potential functions. More formally, a MRF, M, is a pair $M = (\mathcal{G}, \Psi)$ where $\mathcal{G} = (V, E)$ is an undirected graph, and Ψ is a set of potential functions over the maximal cliques in $\mathcal{G}$, $C(\mathcal{G})$. The graph's vertex set $V = \{V_1, V_2, \ldots, V_n\}$ is isomorphic to the set of variables (i.e., $\mathbf{X}$) and we will make no distinction between the ith vertex and the ith random variable. Each edge $e = \{u, v\} \in E$, represents a dependency between random variables $u \in V$ and $v \in V$. Each potential, ψ_c, is a mapping from the possible joint assignments of the elements of $c \in C(\mathcal{G})$ to the positive reals.

For example, consider a particular backbone conformation $\mathbf{x_b}$ of Lysozyme(pdb id: 2lyz) shown in Figure 1a with a few residues highlighted. Figure 1b shows that part of the MRF 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. This MRF thus represents the probability distribution of the side chain atoms of a protein with a given backbone.

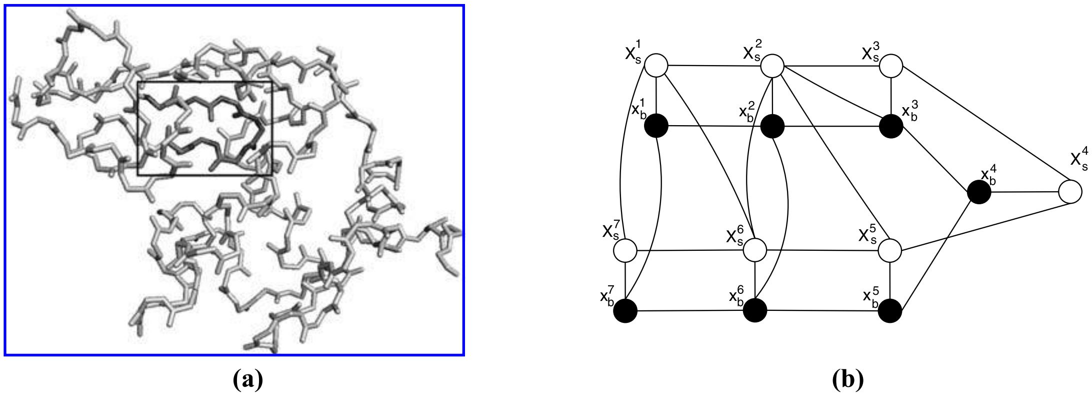


FIG. 1. **(a)** Structure of the backbone atoms of lysozyme (pdb id: 2lyz) with a few residues outlined. **(b)** Part of the random field induced by the outlined residues: X_s^i's are the hidden variables representing the rotameric state, the visible variables are the backbone atoms in conformations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

In general, an MRF 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

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

where 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., $\psi(x_b^i, x_b^j) = 1$.

This undirected graphical model, characterized by the variables $\mathbf{X}$, the edges between the variables and the potential ψ can also be represented more conveniently, as a bipartite graph $(\mathbf{X}, F)$, called a *factor graph*. Here, F is the set of factors which represent the interaction between variables. For example, 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 Figure 1b is shown in Figure 2. Each edge between side chain

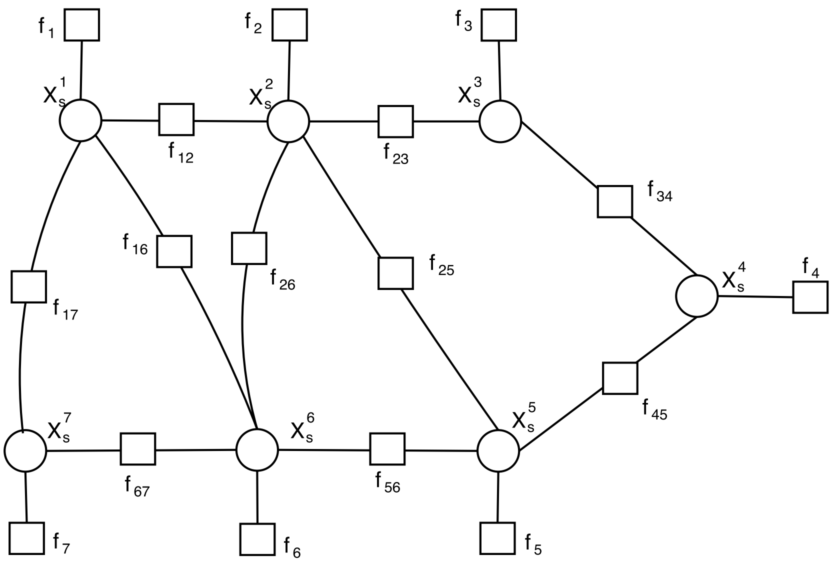


FIG. 2. Factor graph representation for the graph shown in Figure 1b. The observed variables corresponding to the backbone atoms can be replaced by a factor, f_i, at each side chain variable.

variables has been replaced by edges to a factor, f_i, representing the interaction between these variables. Also, 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 free energy of a system is given by

$$G = E - TH$$

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

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

where the first term and second terms on the right are the enthalpic and entropic contributions respectively and the summation is over all possible $\mathbf{x}$. 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 (Bethe, 1935; Kikuchi, 1951; Morita, 1991; Morita et al., 1994). 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 w_{R_i} to the contribution of every node in region $R_i \in R$, 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 (1935) approximation is a region graph with each region containing at most 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 (Kikuchi, 1951; Yedidia et al., 2005).

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. For a recent survey of previously used methods and their performance relative to GBP, see Pelizzola (2005). 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[3] (Yedidia et al., 2000, 2005).

[3] The free energy is often referred to as the "energy functional" in this literature.

3.1. 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* (Yedidia et al., 2005) 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 $\mathcal{P}(R)$ and children $\mathcal{O}(R)$

$$n^0_{R\to P}(\mathbf{x_r}) = \tilde{f}_R(\mathbf{x_R}) \prod_{P'\in\mathcal{P}(R)\backslash P} m_{P'\to R}(\mathbf{x_r}) \prod_{O\in\mathcal{O}(R)} n_{O\to R}(\mathbf{x_O})$$

$$m^0_{R\to O}(\mathbf{x_O}) = \sum_{x_R\backslash x_O} \tilde{f}_R(x_R) \prod_{P\in\mathcal{P}(R)} m_{P\to R}(\mathbf{x_R}) \prod_{O'\in\mathcal{O}(R)\backslash O} n_{O'\to R}(\mathbf{x_{O'}}),$$

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

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

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

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

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

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 (Yanover and Weiss, 2002; Yedidia et al., 2002).

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 (Chu et al., 2004). Applications of graphical models to tertiary structure are generally limited to applications of Hidden Markov Models (HMMs) (Karplus et al., 2003). 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 (Liu et al., 2005). 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 (Xu, 2005) 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 Yanover and Weiss (2002). 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.

The pioneering work of Lee and Levitt (1991) and Lee (1994) computed estimates using a sampling scheme which can be computationally expensive, while Koehl and Delarue (1994) attempted to solve the same problem using a mean-field approach. Mean-field techniques have also been used to perform energy calculations, such as those pioneered by Lee (1992) for predicting binding free energies, and protein mutant energetics (Lee and Levitt, 1991; Lee, 1994). Recent work (Minka, 2005) has shown that most message passing inference 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. Mean field methods minimize the Kullback-Leibler Divergence while GBP (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, GBP 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 (Moreland et al., 2005). 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 Å to 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 Canutescu et al. (2003) and a linear approximation to the repulsive van der Waals force used by Canutescu et al. (2003) and Yanover and Weiss (2002). 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 300 K, 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 Figure 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. Aji and McEliece (2003) studied 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 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 (Samudrala and Levitt, 2000). We used the immunoglobulin datasets from the "multiple decoy sets." Each such dataset consisted of multiple decoy structures along with the native structure of a protein. We selected immunoglobulin because it had a large number of decoys close to the native structure and has been used extensively to test methods for decoy detection (Summa et al., 2005).

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 Å to the native structure, from each dataset. We then removed any dataset that ended up with fewer than five decoys so that we didn't end up with too few decoys in a dataset. We also removed three datasets which had missing backbone atoms. At the end of this pruning, there were 48 datasets left with an average of around 35 decoys per data set.

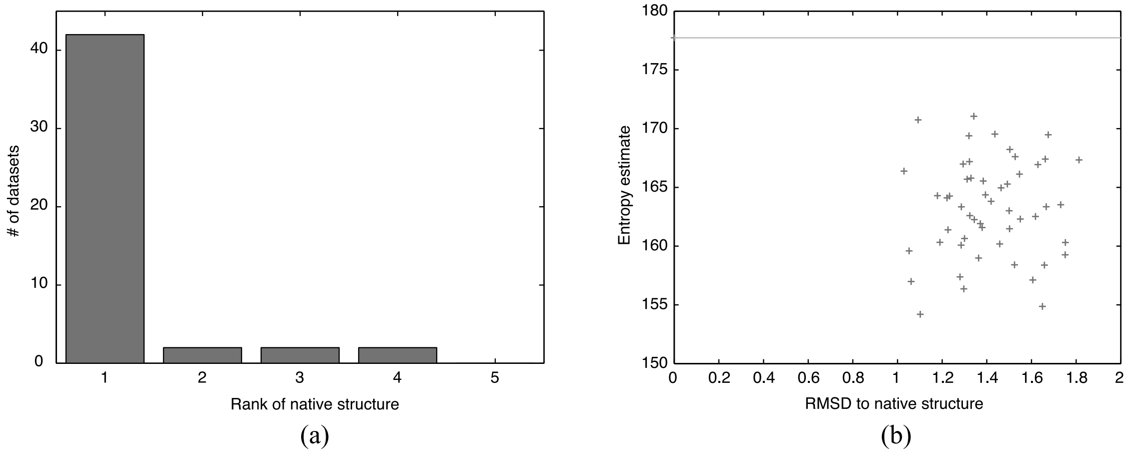
Figure 3 shows our results on the immunoglobulin dataset. When we ranked the structures in the decreasing order of their entropy, the native structure ended up at the top in 42 of the 48 datasets (87.5%). In no dataset was the native structure ranked higher than 4. Figure 3b shows the scatter plot of the entropy estimates for a dataset where the native structure(3hfm) has the highest entropy.

To study the structures further, we ran PROCHECK (Laskowski et al., 1993)—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 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 two such angles, among the rest of the structures. However, 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 repeated our experiments on the entire immunoglobulin 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.

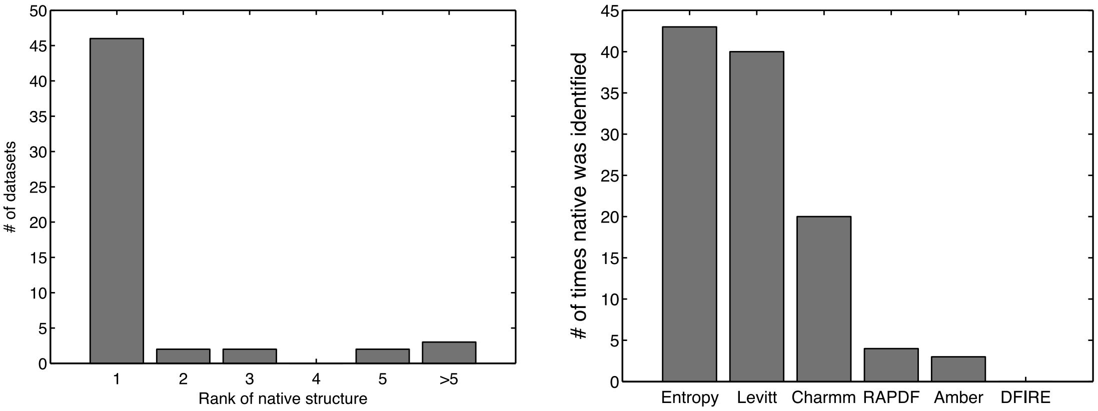
Figure 4a shows the results of our experiments on the entire immunoglobulin dataset. As can be seen, despite the addition of the dissimilar backbones, the ranking of the native structure isn't affected much—in



(a)

(b)

FIG. 3. **(a)** Histogram shows the distribution of the rank of the native structure, when ranked in decreasing order of entropy for the culled immunoglobulin decoy dataset. Over this dataset, the native structure has the highest entropy 87.5% of the time. **(b)** Entropy estimates for 3hfm 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.



(a) (b)

FIG. 4. **(a)** Histogram showing the distribution of the rank of the native structure. **(b)** Comparison of Results using various energy functions as reported in Summa et al. (2005), along with rankings based on our entropy estimates. These results are on the 51 immunoglobulin datasets for which data was available, including decoys with RMSD greater than 2.0 Å. Overall, the entropy estimates outperform all energy functions.

84% of the datasets, the native structure has the highest entropy. We then compare our results to the following different energy functions as reported in Summa et al. (2005): a four body statistical potential ("4body") developed in Summa et al. (2005), the coulombic part of the CHARMM19 forcefield (Brooks et al., 1983), "RAPDF" (Samudrala and Moult, 1998), "DFIRE" (Zhou and Zhou, 2002), and the sum of van der Waal and coulombic terms of the AMBER force field (Weiner et al., 1984). These energy functions are described in detail in Summa et al. (2005).

It must be noted that "4body" has a distance parameter; the numbers reported are the best results obtained across different values of this parameter. In contrast, the temperature T, which is the only tunable parameter in our approach, was set to room temperature. Yet, our entropy estimates calculated using a simple linear potential function, marginally outperforms "4body" and significantly outperforms all the other pairwise energy terms 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 their multiple single point mutants selected from the ProTherm database (Kumar et al., 2006), with corresponding estimates obtained using GBP. Only mutations in buried positions were considered in order to minimize the effects of the solvent. All the $\Delta\Delta G$ experiments in a single dataset were conducted at the same pH value.

Since these mutants have different sequences, the free energy 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. Figure 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 of $1.5{*}10^{-5}$ to 0.0063. This compares favorably with the estimates—correlations between 0.7 and 0.94—obtained using the four body potential of Carter et al. (2001) over all their (much smaller) datasets. This gives evidence that our estimates predict the relative magnitude of $\Delta\Delta G$ with reasonable accuracy.

5. CONCLUSION

We have shown that free energy calculations for all-atom models of protein structures can be computed efficiently using GBP. Moreover, these estimates are sufficiently accurate to perform non-trivial tasks.

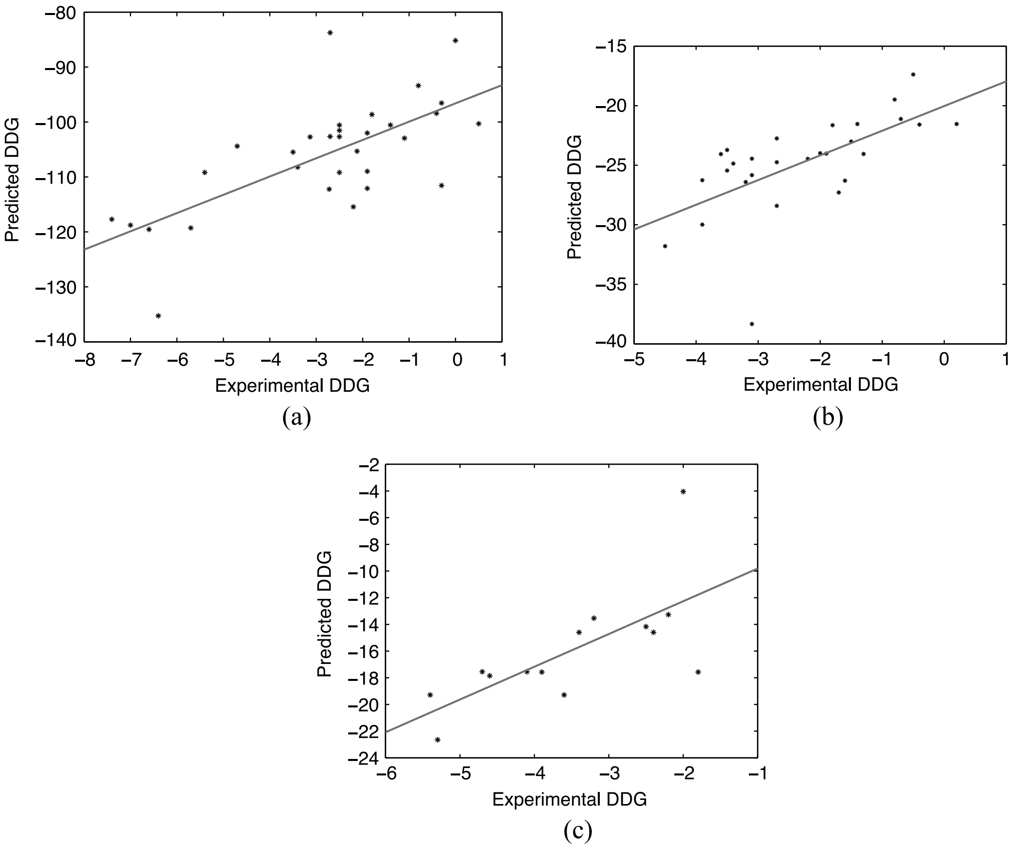


FIG. 5. Plots showing variation of experimental $\Delta\Delta G$ (on the x-axis) with computed estimates of $\Delta\Delta G$, along with a least squares fit for 31 mutants of barnase (pdb id: 1BNI), $R = 0.70$, $p = 1.5 * 10^{-5}$ **(a)**, 28 mutants of T4 Lysozyme(pdb id:1L63), $R = 0.63$, $p = 3.0 * 10^{-4}$ **(b)**, and 14 mutants of staphylococcal nuclease (pdb id:1STN), $R = 0.69$, $p = 0.0063$ **(c)**.

We first demonstrated that it is possible to identify native immunoglobulin 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 are not only in the right relative order, but also have approximately the right relative magnitudes.

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 (Rohl et al., 2004). The accuracy of existing homology modeling methods is acknowledged to be an important issue in structural biology (Marti-Renom et al., 2000; Protein Structure Initiative, 2003). 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 backbone flexibility and a somewhat more sophisticated energy term 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 the anonymous reviewers for their valuable comments. We also thank K. Arun and Jianhua Shao of the Langmead lab and all members of the CBML group for their suggestions. 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.). E.P.X. is supported by the Pennsylvania Department of Health's Health Research Program (grant no. 2001NF), Cancer Health Research (grant ME-01-739) and by a NSF CAREER Award (grant no. DBI-0546594).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Aji, S.M., and McEliece, R.J. 2003. The generalized distributive law and free energy minimization. *Proc. 39th Allerton Conf. Commun. Control Comput.* 459–467.

Berman, H.M., Westbrook, J., Feng, Z., et al. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242.

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

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

Brooks, B.R., Bruccoleri, B.D., Olafson, D.J., et al. 1983. CHARMM: a program for macromolecular energy minimization and dynamics calculations. *J. Comp. Chem.* 4, 187–217.

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

Carter, Jr., C.W., LeFebvre, B.C., Cammer, S.A., et al. 2001. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J. Mol. Biol.* 311, 625–638.

Chu, W., Ghahramani, Z., and Wild, D. 2004. A graphical model for protein secondary structure prediction. *Proc. 21st Ann. ICML*, pgs. 161–168.

Karplus, K., Karchin, R., Draper, J., et al. 2003. Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction. *Proteins* 53, 491–496.

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

Koehl, P., and Delarue, M. 1994. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* 239, 249–275.

Kumar, M.D., Bava, K.A., Gromiha, M.M., et al. 2006. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* 34, D204–D206.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., et al. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* 26, 283–291.

Lee, C., and Levitt, M. 1991. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature* 352, 448–451.

Lee, C. 1992. Calculating binding energies. *Curr. Open. Struct. Biol.* 2, 217–222.

Lee, C. 1994. Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.* 236, 918–939.

Lilien, R., Stevens, B., Anderson, A., et al. 2005. 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. Comput. Biol.* 12, 740–761.

Liu, Y., Carbonell, J., Weigele, P., et al. 2005. Segmentation conditional random fields (SCRFs): a new approach for protein fold recognition. *RECOMB 2005*, 408–422.

Lovell, S.C., Word, J.M., Richardson, J.S., et al. 2000. The penultimate rotamer library. *Struct. Funct. Genet.* 40, 389–408.

Marti-Renom, M.A., Stuart, A., Fiser, A., et al. 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29, 291–325.

Minka, T. 2005. *Divergence Measures and Message Passing*. Microsoft Technical Report (MSR-TR-2005-173). Microsoft Research, Cambridge, UK.

Moreland, J.L., Gramada, A., Buzko, O.V., et al. The Molecular Biology Toolkit (MBT): A modular platform for developing molecular visualization applications. *BMC Bioinform.* 6, 21.

Morita, T. 1991. Cluster variation method for non-uniform Ising and Heisenberg models and spin-pair correlation function. *Prog. Theor. Phys.* 85, 243–255.

Morita, T., Suzuki, T.M., Wada, K., et al., eds. 1994. Foundations and Applications of Cluster Variation Method and Path Probability Method. *Prog. Theor. Phys.* 115, Suppl, Entire Issue.

Pelizzola, A. 2005. Cluster variation method in statistical physics and probabilistic graphical models. *J. Phys. A* R309–R339.

Protein Structure Initiative. 2003. Report on the NIGMS Workshop on High Accuracy Comparative Modeling. Available at: *http://archive.nigms.nih.gov/psi/reports/comparative_modeling.html.* Accessed July 6, 2008.

Rohl, C.A., Strauss, C.E., Misura, K.M., et al. 2004. Protein structure prediction using Rosetta. *Methods Enzymol.* 383, 66–93.

Samudrala, R., and Moult, J. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* 275, 895–916.

Samudrala, R., and Levitt, M. 2000. Decoys “R” Us: a database of incorrect protein conformations to improve protein structure prediction. *Protein Sci.* 9, 1399–1401.

Summa, C.M., Levitt, M., and Degrado, W.F. 2005. An atomic environment potential for use in protein structure prediction. *J. Mol. Biol.* 352, 986–1001.

Thomas, P.D., and Dill, K.A. 1994. Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* 257, 457–469.

Tobi, R., and Elber, D. 2000. Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins* 41, 40–46.

Weiner, S.J., Kollman, P.A., Case, D.A., et al. 1984. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* 106, 765–794.

Xu, J. 2005. Rapid protein side-chain packing via tree decomposition. *Lect. Notes Comput. Sci.* 3500, 423–439.

Yanover, C., and Weiss, Y. 2002. Approximate inference and protein folding. *Proc. NIPS* 84–86.

Yedidia, J.S., Freeman, W.T., and Weiss, Y. 2000. Generalized belief propagation. *Adv. NIPS* 13, 689–695.

Yedidia, J.S., Freeman, W.T., and Weiss, Y. 2002. Characterizing belief propagation and its generalizations. Available at: *www.merl.com/reports/TR2002-35/.* Accessed July 6, 2008.

Yedidia, J.S., Freeman, W.T., and Weiss, Y. 2005. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inform. Theory* 51, 2282–2312.

Zhou, H., and Zhou, Y. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* 11, 2714–2726.

Address reprint requests to:
Dr. Christopher J. Langmead
Computer Science Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

E-mail: cjl@cs.cmu.edu