A graphical model approach for predicting free energies of association for protein-protein interactions under backbone and side-chain flexibility

Hetunandan Kamisetty[1] **Chris Bailey-Kellogg**[2]

Christopher James Langmead[1,3][1]

December 2008
CMU-CS-08-162

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[1] Computer Science Department, Carnegie Mellon University, Pittsburgh, PA
[2] Department of Computer Science, Dartmouth College, Hanover, NH
[3] Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA

[1] Corresponding Author: cjl@cs.cmu.edu

Keywords: Graphical Models, Free Energy, Protein-Protein Interactions, Belief Propagation

Abstract

Biomolecular systems are governed by changes in free energy, and the ability to predict binding free energies provides both better understanding of biomolecular interactions and the ability to optimize them. We present the first graphical-model based approach, which we call GOBLIN (*Graphical mOdel for BiomoLecular INteractions*), for predicting binding free energies for all-atom models of protein complexes. Our method is physically sound in that *internal energies* are computed using standard molecular-mechanics force fields, and *free energies* are obtained by computing a rigorous approximation to the partition function of the system. Moreover, GOBLIN explicitly models both backbone and side-chain flexibility, and, when desired, employs non-linear regression to optimize force-field parameters. In tests on a benchmark set of more than 700 mutants, we show that our method is fast, running in a few minutes, and accurate, achieving root mean square errors (RMSEs) between predicted and experimental binding free energies of 2.05 kcal/mol. GOBLIN's RMSEs are 0.55 kcal/mol better than the well-known program ROSETTA, despite the fact that we use the ROSETTA force field for computing internal energies. That is, our increase in accuracy is due to our ability to accurately estimate entropic contributions to the free energy. Finally, using our novel algorithm for optimizing force-field parameters on specific protein complexes reduced GOBLIN's RMSE by 0.26 kcal/mol on average.

1 Introduction

Protein-protein interactions are essential to the molecular machinery of the cell; transient or persistent complexes mediate processes including regulation, signaling, transport, and catalysis. While coarse-grained, high-throughput techniques such as yeast two-hybrid (Fields and Song, 1989) are primarily focused on *which* proteins interact, finer-grained techniques based on structural analysis address questions of *how* and *why* these interactions occur. By modeling the physical interactions between constituent atoms, structure-based approaches provide deeper insights into, for example, the specificity of an interaction or its sensitivity to various mutations (e.g., (Weber and Harrison, 1999; Wang and Kollman, 2001)). In addition to answering questions of interest to basic science, such methods are also well-suited to designing variants with improved or novel properties (Kortemme and Baker, 2004; Lilien *et al.*, 2005; Joachimiak *et al.*, 2006).

A fundamental law of thermodynamics states that interactions are governed by *binding free energies*. Free energy should not be confused with *internal energy*. Internal energy is a property of a specific configuration of a system; it accounts for electrostatic and hydrophobic interactions, among others. Free energy, on the other hand, is a property of the *entire* configuration space and consists of both enthalpic and entropic contributions. Enthalpy is the expected internal energy, averaged over a Boltzmann distribution over the entire configuration space. Entropy is the negative expected log probability over the same distribution. We note that entropic contributions are known to be important to biologically relevant phenomena (e.g., (Missimer *et al.*, 2007; Chang *et al.*, 2008)).

There are a variety of techniques for estimating free energies computationally, and each method makes a different trade-off between computational efficiency and fidelity to the underlying physics. The most rigorous, and computationally demanding approaches require extensive sampling or molecular dynamics simulations (e.g., Jarzynski (1997); Srinivasan *et al.* (1998); Åqvist *et al.* (2002); Gohlke *et al.* (2003)). Statistical and coarse-grained methods are a popular alternative (e.g., Böhm (1992); Bahar *et al.* (1997); Verdonk *et al.* (2003); Muegge (2006)), because they are computationally much faster. Unfortunately, in addition to concerns regarding the utility of statistical potentials (e.g., Thomas and Dill (1994)), it has been argued (Leach *et al.*, 2006; Warren *et al.*, 2006) that they do not adequately account for the conformational flexibility and in general do poorly when estimating the change in entropy upon binding.

This paper introduces a method, called GOBLIN (*Graphical mOdel for BiomoLecular INteractions*), that lies between the extremes of detailed molecular dynamics and statistical potentials. GOBLIN models the energy landscape for a protein-protein complex as a probability distribution over an exponentially large number of configurations. This distribution is compactly encoded using an undirected probabilistic graphical model known as a Markov Random Field (MRF). Under this model, internal energies are calculated using standard atomic-resolution force-fields, and rigorous binding free energy calculations are performed using Belief Propagation (Pearl, 1986). Our method runs in a few minutes and this therefore is significantly faster than MD. At the same time, GOBLIN is more rigorous than statistical methods.

The key contributions of GOBLIN are as follows:

- The first graphical model-based approach for modeling protein-protein interactions.

- The first graphical model-based approach for free energy calculations under *both* side-chain and backbone flexibility.

- A novel approach to optimizing force-field parameters by employing non-linear regression.

- An inferential approach to calculating free energies that properly accounts for the discretization of the conformation space.

GOBLIN significantly extends the state of the art of all-atom graphical models of proteins, including our own previous work (Kamisetty *et al.*, 2007, 2008) in this area. Our earlier work focused on computing intra-molecular *folding free energies* for fixed backbone configurations, while the present paper develops and demonstrates a technique for computing inter-molecular *binding free energies* under backbone flexibility. Additionally, in order to account for the particularities of the binding free energies of the system under study, we solve a non-linear optimization problem to fit the force-field parameters to predict the partition function correctly. This optimization problem is novel and distinct from the work by Yanover *et al.* (2007) who learnt force-field parameters to maximize the accuracy of side-chain placement; in the presence of flexible backbones, doing this efficiently is also very challenging.

We demonstrate the utility of GOBLIN by performing binding free energy calculations for more than 700 mutants of protein-protein complexes. It is fast, typically requiring less than 5 minutes per calculation. It is also accurate, with root mean squared errors (RMSE) of 2.05 kcal/mol relative to experimental values. Significantly, our method outperforms the well-known program ROSETTA in terms of accuracy by 0.55 kcal/mol. This result is especially interesting because we implemented ROSETTA's own force field in order to compute internal energies. That is, our improved accuracy can be attributed to our approach to computing free energies. Finally, GOBLIN's RMSEs can be reduced by 0.26 kcal/mol on average, when our learning algorithm is used to optimize force-field parameters.

2 A Markov Random Field Model for Protein Complexes

In this section, we will describe a compact encoding of the equilibrium protein ensemble.

A protein consists of some number of atoms across one more more polypeptide *chains*. A *configuration* of the protein corresponds to the geometry of each of its constituent atoms. While a protein is commonly represented as a single configuration (usually the crystalline form), at room temperature, a more accurate representation of the protein would be as an ensemble of configurations. We will adopt such a representation, by treating the configuration of a protein as a random variable in some configurational space $\mathcal{C}$.

In what follows, random variables are represented using upper case letters, while lower case variables represent specific values that the random variables can assume. Additionally, we will use bold face variables to describe sets or vectors of variables.

If we use $\mathbf{X}$ to refer to the random variable corresponding to the configuration of the entire set of atoms in the protein. Boltzmann's law describes the probability distribution over $\mathcal{C}$ of a physical system at equilibrium; according to it, the probability of a configuration $\mathbf{x_c} \in \mathcal{C}$, $P(\mathbf{X} = \mathbf{x_c})$, with *internal energy* $E_\mathbf{c}$ is

$$P(\mathbf{X} = \mathbf{x_c}) = \frac{1}{Z} \exp\left(\frac{-E_\mathbf{c}}{k_B T}\right) \tag{1}$$

where $Z = \sum_{\mathbf{x_c} \in \mathcal{C}} \exp(-E_\mathbf{c})$ is the *partition function*, k_B is Boltzmann's constant, and T is the absolute temperature in Kelvin.

It is customary to partition the entire set of atoms into two disjoint sets: *backbone* and *side-chain*. *Backbone atoms* refer to those that are common to all 20 amino acid types, while *side-chain* atoms are those that differ among the different kinds of amino acids. We will use the $\mathbf{b}, \mathbf{s}$ suffixes to denote backbone and side-chain variables variables respectively: $\mathbf{x_b} = \{x_{\mathbf{b}_1}, x_{\mathbf{b}_2}, \ldots\}$, is a set of variables, one for each chain in the protein, representing the conformation of the backbone atoms, $\mathbf{x_s} = \{x_{\mathbf{s}_1}, x_{\mathbf{s}_2}, \ldots\}$ where $x_{\mathbf{s}_i}$ represents the conformation of the side-chain atoms of residue i, and $E_\mathbf{b}, E_\mathbf{s}$ represent energies of $\mathbf{x_b}, \mathbf{x_s}$.

Using the fact that $\mathbf{x_c} = \{\mathbf{x_b}, \mathbf{x_s}\}$, $E_\mathbf{c} = E_\mathbf{b} + E_\mathbf{s}$, we can now rewrite the joint distribution and the partition function as:

$$P(\mathbf{X} = \mathbf{x_c}) = P(\mathbf{X_b} = \mathbf{x_b}) P(\mathbf{X_s} = \mathbf{x_s} | \mathbf{X_b} = \mathbf{x_b}) \tag{2}$$

$$Z = \sum_{\mathbf{x_b}} \exp(-\frac{E_\mathbf{b}}{k_B T}) Z_b \tag{3}$$

where $Z_b = \sum_{\mathbf{x_s}} \exp(\frac{-E_\mathbf{s}}{k_B T})$ is the partition function over the side-chain conformational space with a fixed backbone.

Note that this distribution is over the entire configurational space $\mathcal{C}$, a space that is exponentially large in the size of the protein. In what follows, we will attempt to simplify this distribution by exploiting its properties around near-native equilibrium.

First, in order to account for the flexibility in backbones around the native structure, we sample the space of backbones around the wild type backbone by using backrub motions – prevalent but subtle mode of local backbone motions that are observed in native crystal structures – as described in Davis *et al.* (2006). Recent work has described how incorporating such backbone flexibility improves side chain prediction and protein design (Smith and Kortemme, 2008; Friedland *et al.*, 2008; Georgiev *et al.*, 2008). We generate an ensemble of backrub backbone traces ($\mathbf{x_b}$) which we assume efficiently samples the space of local backbones.

Given a specific backbone trace $\mathbf{x_b}$, due to the nature of the physical forces in action, pairs of residues distally located according to trace are expected to exert very little direct influence on one another. In statistical terms, we say that such residues are independent of each other when conditioned on the event $\mathbf{X_b} = \mathbf{x_b}$. We will exploit these conditional independencies present in $P(\mathbf{X_s} = \mathbf{x_s} | \mathbf{X_b} = \mathbf{x_b})$ to compactly encode it as a Markov Random Field(MRF).

An MRF $\mathcal{G}$ is a probability distribution over a graph, and can be represented as a tuple $(\mathbf{X}, \mathcal{E}, \Phi)$, where the set of random variables in the multivariate probability distribution are the set of vertices – $\mathbf{X_s}$ and $\mathbf{X_b}$ in this case – while edges $e \in \mathcal{E}$ join residues that are directly dependent on each other and $\Phi = \{\phi_1, \phi_2, ..., \phi_m\}$ is a set of functions (popularly called factors) over random variables. [1]

The PDB structure is assumed to be one of the micro-states of the system at equilibrium. Atoms are allowed to deviate from their crystal structure coordinates. This is done by discretizing each degree of freedom and grouping atoms together according to side-chain rotamer libraries (e.g., Lovell *et al.* (2000)). Additional simplifying assumptions include fixing bond lengths and bond angles to idealized values or to the observed in the crystal structure. When discrete approximations are used, the MRF efficiently encodes an ensemble of micro-states of size $O(k^n)$, in $O(kn)$ space, where k is the average number of conformations per residue, and n is the number of residues in the protein.

In our model, the functions in $\mathcal{G}$ (i.e., ϕ_i) are defined in terms of a Boltzmann factor. That is, $\phi_i(x_{\phi_i}) = \exp\left(-\frac{E(x_{\phi_i})}{k_B T}\right)$, where x_{ϕ_i} is the set of atoms that serve as arguments to ϕ_i, and $E(x_{\phi_i})$ is the potential energy of those atoms as defined by a molecular force field. In theory, any molecular force field can be used. We specifically use the ROSETTA potential $E_{Rosetta}$ that ROSETTA uses in computing $\Delta\Delta G$(Kortemme and Baker, 2002) which is composed of the following terms:

- E_{ljatr}, E_{ljrep}, the attractive and repulsive parts of a $6-12$ Lennard-Jones potential used to model van der Waals interactions.

- E_{sol}, the Lazardus-Karplus solvation energy that approximates the solvation energy by using an implicit solvent Lazaridis and Karplus (1999).

- E_{hb}, is the Hydrogen bond energy as computed by Kortemme *et al.* (2003)

$E_{Rosetta}$ is a linear combination $\mathbf{w}^T\mathbf{E} = w_{ljatr}E_{ljatr} + w_{ljrep}E_{ljrep} + w_{sol}E_{sol} + w_{hb}E_{hb}$. The vector **w** that defines the linear combination is typically learnt by fitting the energy terms to physical observations like $\Delta\Delta G$(Kortemme and Baker, 2002).

Fig. 1 illustrates the construction of $\mathcal{G}$ using a protein complex in our dataset: Chymotrypsin complexed with the third domain of turkey ovomucoid(OMTKY). Fig. 1-A shows a single configuration configuration $\mathbf{x_c}$ of the protein complex, with the residues in the Chymotrypsin chain and the the OMTKY chain shown in different colors (red,blue respectively) for visual clarity. In contrast, the MRF $\mathcal{G}$ shown in Fig. 1-B models a *distribution* over all possible $\mathbf{x_s}$ for a given backbone trace. The construction of $\mathcal{G}$ is identical irrespective of whether the protein is a single chain, or multiple chains as in the case of a protein complex. What does change is the nature of the physical interactions being captured by the ϕ in $\mathcal{G}$ in each case. The potential terms in $E_{Rosetta}$ capture both the intra-molecular interactions (shown as solid lines), and the inter-molecular interactions(shown as dashed lines).

Given the structure of $\mathcal{G}$ and the potentials Φ as described above, we can rewrite the conditional distribution $P(\mathbf{X_s} = \mathbf{x_s}|\mathbf{X_b} = \mathbf{x_b})$ by using the Hammersley-Clifford theorem Clifford (1990) in the following manner.

$$P(\mathbf{X_s} = \mathbf{x_s}|\mathbf{X_b} = \mathbf{x_b}) = \frac{1}{Z_\mathbf{b}} \prod_{\phi_i \in \Phi} \phi_i(\mathbf{x}_{\phi_i}) \tag{4}$$

[1] Since we use $\mathcal{G}$ to represent a conditional probability distribution $P(\mathbf{X_s}|\mathbf{X_b} = \mathbf{x_b})$, this is also referred as a Conditional Random Field(CRF). Since commonly used CRFs Lafferty *et al.* (2001) are usually chain graphs, we use the more general term, MRF, to avoid confusion.

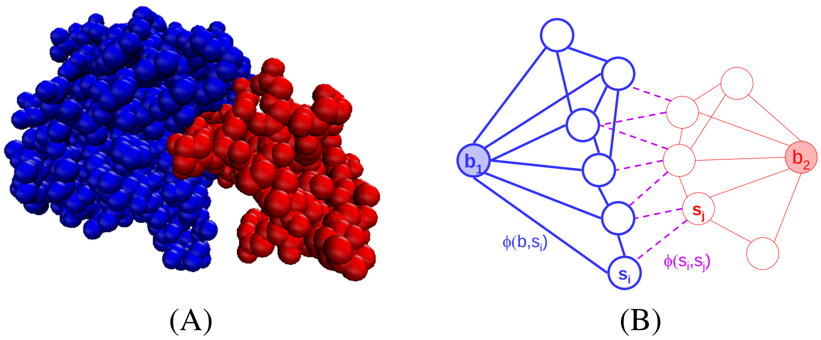


Figure 1: (A) Chymotrypsin complexed with the third domain of turkey ovomucoid (OMTKY). While protein structures are often shown as a single conformation, in reality they occupy ensembles of conformations; our method models both side-chain and backbone ensembles. (B) Part of an MRF encoding the conditional distribution over the ensembles of conformations. Blue nodes with thick lines correspond to Chymotrypsin, red nodes with thin lines correspond to OMTKY. Solid lines refer to intra-molecular interactions, and dashed lines refer to inter-molecular interactions. The nodes labeled $\mathbf{b}_1, \mathbf{b}_2$ represent the conformation of the backbone atoms in the two chains, while the nodes labeled $\mathbf{s}_i$ and $\mathbf{s}_j$ corresponds to conformations of side-chain atoms. (Since all the variables in the graph represent conformations, for visual clarity, we omit the X in their labels.)

where once again we have the partition function $Z_{\mathbf{b}}$ associated with a specific backbone:

$$Z_{\mathbf{b}} = \sum_{\mathbf{X_s}} \prod_{\phi_i \in \Phi} \mathbf{x}_{\phi_i} \tag{5}$$

Notice that the due to the choice of the Boltzmann factor for Φs, this distribution is consistent with the Boltzmann distribution of Eq. 1. To obtain the joint distribution, one needs to simply multiply Eq. 4 with the probability of the particularly backbone conformation $\mathbf{x_b}$ according to Eq. 2.

Thus, the probability of a given state is simply the product of the functions, suitably normalized. Evaluating the product in Eq. 4 is straightforward for any given configuration of the random variables. Computing the partition function in Eq. 5, on the other hand, is computationally intractable in the general case (Dagum and Chavez, 1993) because it involves summing over every state. However, the machine learning community has developed a number of efficient algorithms for performing probabilistic inference over MRFs. One of these algorithms – Belief Propagation – and its relationship to free energy calculations is discussed in section Sec. 3.

3 Probabilistic Inference and Free Energy Calculations

Recall that in solution a protein exists in an ensemble of different configurations. The *free energy* of a protein is a measure on the ensemble and is defined as: $G = H - TS$ where H is the *enthalpy*, or the expected internal energy, S is the entropy of the ensemble and T is the temperature of the system.

The rates of biochemical reactions are determined by *changes* in free energies. In particular, ΔG_{bind}, the *binding free energy*, is the change in free energy when two proteins, A and B, bind:

$G_{AB} - (G_A + G_B)$. It determines the rate of association of A and B. Often, the quantity of interest in tasks such as protein design is $\Delta\Delta G$, the change in the ΔG value upon mutation from wild-type: $\Delta G_{mutant} - \Delta G_{wild-type}$. For example, a beneficial mutation results in a better ΔG for the variant than for the wild-type, and thus a negative $\Delta\Delta G$. Computing G therefore is extremely useful in determining the properties of the protein.

The free energy of a physical system is related to the Boltzmann distribution by way of the partition function: G is simply $-k_B T \log Z$. The MRF model from the previous section provides a compact representation of the Boltzmann distribution that enables us to compute a quick and good approximation to the binding free energy by solving a statistical inference problem to compute Z.

Probabilistic inference in an MRF involves computing marginal distributions over the random variables in the graph. In general, the problem is intractable (Dagum and Chavez, 1993). However, a number of rigorous approximation algorithms have been devised for performing inference in MRFs. Significantly, it has been shown that mathematically, these algorithms are equivalent to performing free-energy approximations (Yedidia *et al.*, 2005). This is not surprising, because inference and free energy calculations both require estimating a partition function. What is surprising, however, is that some existing inference algorithms are mathematically equivalent to specific free-energy approximations introduced by statistical physicists (e.g., Bethe (1935); Kikuchi (1951); Morita (1991); Morita *et al.* (1994)). For example, it is now known that Pearl's *Belief Propagation* (BP) algorithm (Pearl, 1986) is equivalent to the Bethe approximation (Bethe, 1935) of the free energy. We use BP in this paper.

The term 'belief' in both BP refers to the marginal distributions over the random variables in the MRF. Briefly, each node in the graph keeps track of its own marginal probability distribution (i.e., belief). Belief Propagation algorithms start with random initial beliefs, and then use message passing between nodes to converge on a final set of beliefs. Informally, each node updates its own beliefs based on the beliefs of its neighbors in the graph, and the value of the potential function, Φ. When the algorithm converges, the final beliefs can be used to obtain the partition function (or an approximation thereof), and hence a free energy.

If the MRF happens to form a tree (i.e., a graph with no cycles), Belief Propagation is exact and takes $O(|\mathcal{E}|)$ time, where $|\mathcal{E}|$ is the number of edges in the graph. The MRFs considered in this proposal, however, are not trees and have $O(|\mathcal{V}|)$ edges. In this case, we use a closely related algorithm known as Loopy Belief Propagation. Loopy BP is not guaranteed to converge, but has always done so in our experiments.

Using Loopy Belief Propagation on the MRF that encodes the conditional distribution, we can obtain an estimate of the partition function of the conditional distribution $Z_{\mathbf{b}}$ for each backbone configuration $\mathbf{x_b}$ as we previously did for folding free energies (Kamisetty *et al.*, 2008, 2007); Z, the partition function over $\mathbf{x_c}$ can then be computed using Eq. 3.

3.1 Discretization

We now briefly discuss a subtle, yet important issue that we have glossed over so far in our presentation: the effects of discretizing the conformational space $\mathcal{C}$.

The assumption of a discrete rotamer library is fairly well-founded, cf. Canutescu *et al.* (2003); Ponder and Richards (1987); McGregor *et al.* (1987). While a common use of such rotamer libraries is in performing side-chain placement, i.e. finding the single most energetically favorable side-chain conformation $\mathbf{x_s}$ (Yanover and Weiss, 2002; Xu, 2005; Kingsford *et al.*, 2005; Canutescu *et al.*, 2003), these rotamer libraries have also been used in computing free energies and conformational entropies of protein structures (Koehl and Delarue, 1994; Kamisetty *et al.*, 2007, 2008; Lilien *et al.*, 2005).

This approach of using a set of discrete rotameric states to compute the entropy faces a subtle problem. To understand this, let us consider an imaginary protein with exactly one residue whose side-chain atoms are unconstrained in configurational space, i.e. the energy $E_{\mathbf{s}}$ of the side-chain atoms is the same, no matter what configuration they are in.

This protein has a physically measurable amount of entropy $S_{physical}$. Now suppose we discretized the configurational space into n points, each representing an equal fraction of the space. It is clear that in this scenario, the probability of each rotamer will be equal to $1/n$(and indeed, this is what Belief Propagation would predict). The free energy in our discrete model, $H - TS$ equals $< E_{\mathbf{s}} > -T \sum_{i=1}^{n} -\frac{1}{n}\log(\frac{1}{n})$ = $E_{\mathbf{s}} - T\log(n)$. In other words, as the granularity of the discretization increases, the discrete entropy increases and as $n \longrightarrow \infty$, $S \longrightarrow \infty$ and is completely unconnected to $S_{physical}$.

This problem arises in many scenarios, most notably for our purposes, in information-theoretic treatments of statistical physics (Jaynes, 1963, 1968). Fortunately, a solution to this problem is available, which to the best of our knowledge is due to E.T. Jaynes (Jaynes, 1963). By using a measure (i.e. a possibly unnormalized probability distribution) m over the configurational space and replacing the discrete entropy by the relative entropy $S = -\sum_{\mathbf{X_s}} P(\mathbf{X_s}) \log \frac{P(\mathbf{X_s})}{m(\mathbf{s})}$, we now obtain a quantity that behaves correctly in the limit. To use this for our purposes, we point out that the rotamer library we use (Canutescu *et al.*, 2003) provides such a measure m_{dun} for each rotamer which we utilize.

Our earlier treatment of inference can be modified to use the relative entropy instead of the discrete entropy, by observing that

$$S = -\sum_{\mathbf{X_s} \in \mathcal{C}} (P(\mathbf{X_s}) \log P(\mathbf{X_s}) - P(\mathbf{X_s}) \log m(\mathbf{X_s}))$$

and therefore, $G =$

$$\begin{aligned} &\sum_{\mathbf{X_s}} P(\mathbf{X_s}) E_{\mathbf{s}} + \sum_{\mathbf{X_s}} (P(\mathbf{X_s}) \log P(\mathbf{X_s}) - P(\mathbf{X_s}) \log P(m(\mathbf{X_s}))) \\ &= \sum_{\mathbf{X_s}} P(\mathbf{X_s})(E_{\mathbf{s}} - \log m(\mathbf{X_s})) - \sum_{\mathbf{X_s}} P(\mathbf{X_s}) \log P(\mathbf{X_s}) \end{aligned}$$

In other words, the move from the discrete entropy to the discrete relative entropy can be made by adding a $E_{rot} = -\log m(\mathbf{X_s})$ term to the energy function. Furthermore, due to the properties of the measure m_{dun} we actually use, any $(m_{dun})^{w_{rot}}$ is also a valid measure; we can therefore use any linear combination $w_{rot}E_{rot}$ in the energy function leading to E_{goblin} , the "pseudo" energy function being $E_{Rosetta} + w_{rot}E_{rot}$.

A similar problem arises when summing over multiple backbone traces according to Eq. 3: by increasing the number of backbone traces, the value of Z monotonically increases. Again, this can be easily fixed by assigning a fraction of volume of the conformational space to each trace. For our experiments we assume that the conformational space is uniformly sampled by the traces, i.e. each trace represents an equal volume of the conformational space.

4 Learning force fields

Molecular mechanics force-fields are classical approximations of what is fundamentally a quantum mechanical phenomenon. The basic elements for any force field are a) a defined set of atom types, b) a function defining the internal energy of the system, and c) a set of parameters. One of the key challenges when working with force fields is that the due to various approximations, the parameters that work well in one application domain might not work well in another. In practice, force-field parameters are estimated by computing a best fit to large heterogeneous datasets (e.g., Kortemme and Baker (2002)), which may not reflect important interactions that are relevant to a particular protein-protein complex. Therefore, it is not unusual to optimize or develop force-fields for specific complexes (Fong *et al.*, 2004; Lu *et al.*, 2001).

A force-field parameter optimization protocol generally requires a set of *training data*. The parameters of the force field are adjusted to (approximately) minimize a given objective function. Normally, that objective function is simply the square errors in estimates of internal energies in training data. Of course, as shown earlier, GOBLIN considers both the internal energy and the entropy of the system when predicting free energies. This, in turn, argues for the use of a different objective function — the square errors of the free energy estimates in the training data. This is a rather different objective; it replaces a simple linear regression problem with a much more complicated non-linear regression problem. However, as we shall show, it is possible to efficiently, albeit approximately solve this problem. Though our (more accurate) modeling by explicitly accounting for entropic effects introduces a problem that is harder to solve exactly, our results indicate that the loss in accuracy from the approximations in solving this problem are more than outweighed by the gain in accuracy obtained from capturing these entropic differences.

Given a training set that consists of experimentally measured changes in $\Delta\Delta G$, of N mutants of some protein complex, our strategy is optimize the vector of weights $\mathbf{w}$ that define the linear combination E_{goblin} to minimize the mean square error (MSE) in predicting $\Delta\Delta G$.

Let ΔG^{wt}_{obs}, ΔG^{i}_{obs} be the experimentally measured binding free energy for the wild-type (i.e., non-mutated) and the i^{th} mutant protein complex respectively. Let ΔG^{wt}, ΔG^{i} be their predictions obtained using inference on the corresponding wildtype and mutated complexes obtained using Belief Propagation as described in Sec. 3. The quantity $\Delta\Delta G^{i}_{obs}$, the experimentally observed change in the change in binding free energy upon mutation equals $\Delta G^{i}_{obs} - \Delta G^{wt}_{obs}$ and is predicted by $\Delta G^{i} - \Delta G^{wt}$.

The MSE across the dataset then equals $\frac{1}{N}\sum_{i=1}^{N}(\Delta G^{i} - \Delta G^{wt} - \Delta\Delta G^{i}_{obs})^2$

To minimize MSE subject to $\mathbf{w} \succeq 0$ by gradient descent, we need to compute the gradient w.r.t $\mathbf{w}$, $\frac{\partial mse}{\partial \mathbf{w}} =$

$$\frac{1}{N}\sum_{i=1}^{N} 2\left(\Delta G^{i} - \Delta G^{wt} - \Delta\Delta G^{i}_{obs}\right)\left(\frac{\partial \Delta G^{i}}{\partial \mathbf{w}} - \frac{\partial \Delta G^{wt}}{\partial \mathbf{w}}\right) \tag{6}$$

In the case of rigid backbones, we have $G = -k_B T \log Z_{\mathbf{b}}$. The partial derivative $\frac{\partial G}{\partial \mathbf{w}}$ is therefore

$$\frac{\partial G}{\partial \mathbf{w}} = -k_B T \frac{\partial \log Z_{\mathbf{b}}}{\partial \mathbf{w}} = -k_B T \frac{1}{Z_{\mathbf{b}}} \frac{\partial Z_{\mathbf{b}}}{\partial \mathbf{w}} = \langle \mathbf{E} \rangle_{\mathbf{s}} \tag{7}$$

where $\langle \mathbf{E} \rangle_{\mathbf{s}}$ is the expected value of the vector of the individual force-fields over all $\mathbf{x}_{\mathbf{s}}$ with the current value of $\mathbf{w}$.

By using the fact that the difference in derivatives is just the derivative of the differences, we get that Eq. 6 for the model without backbone flexibility is

$$\frac{\partial mse}{\partial \mathbf{w}} = \frac{2}{N} \sum_{i=1}^{N} (\Delta G^i - \Delta G^{wt} - \Delta\Delta G^i_{obs})(\Delta\langle \mathbf{E}^i \rangle - \Delta\langle \mathbf{E}^{wt} \rangle) \tag{8}$$

Now, if we had backbone flexibility, we would have $G = -k_B T \log Z$. Substituting for Z using Eq. 3, the partial derivative would therefore be

$$\frac{\partial G}{\partial \mathbf{w}} = \frac{1}{Z} \sum_{x_{\mathbf{b}}} e^{-E_{\mathbf{b}}/(k_B T)} Z_{\mathbf{b}} \langle \mathbf{E} \rangle_{\mathbf{s}} \tag{9}$$

which we can substitute into Eq. 6 to get the derivative for the model with backbone flexibility. In other words, given the enthalpies and free energies of each backbone trace, we can compute the free energy of the entire distribution and its derivatives.

5 Results

We studied the efficacy of our approach on a database of over 700 single-point mutants from eight large and well studied complexes. For each of these variants, the database contains the $\Delta\Delta G_{obs}$, the experimental change in binding free energy upon mutation. The details of the datasets, along with the PDB ids of the wildtype complexes, are shown in Table 1. Of these, the three largest datasets (wildtype PDB ids: 1sgr, 1cho, 1ppf) are from the Kazal family of protein inhibitors (Lu *et al.*, 2001) while the rest of the interactions are part of an Alanine-scanning database previously used in Kortemme and Baker (2002) and Kortemme *et al.* (2003). We note that the amount of thermodynamic data available for protein-protein interactions is limited, and the database we considered is among the largest of its kind.

We performed *in silico* mutagenesis with a fixed backbone on the wildtype complex to obtain a plausible structure of each mutant. Hydrogen atoms were then added using the REDUCE software program (Word *et al.*, 1999). In order to compute $\Delta\Delta G$, we also need the structures of the individual partners. As is common, we assumed that the native backbone in the complex is a good approximation of both the complexed and apo backbones of the engineered proteins. Thus, at the end of this process, we obtained plausible structures for the complexed and apo partners. The results on fixed backbones were obtained using these structures. We describe how backbone flexibility was added later in this section.

We determined atom types and computed energies using our implementation of the ROSETTA force-field as specified in Kuhlman *et al.* (2003); Kortemme *et al.* (2003); Kortemme and Baker

Wt PDB id	Partner A	Partner B	# mutants in A, B
1sgr	OMTKY	SGP B	150,0
1cho	OMTKY	Chymotrypsin	170,0
1ppf	OMTKY	Human LE	170, 0
3hfm (AS)	HYHEL	HEL	12,13
1gc1 (AS)	CD4	GP120	49,0
1a22 (AS)	HGH	HGHBP	34,29
1dan (AS)	BCF VII-A	TF	20,23
1bxi (AS)	E9 Dnase	IM 9	30,0

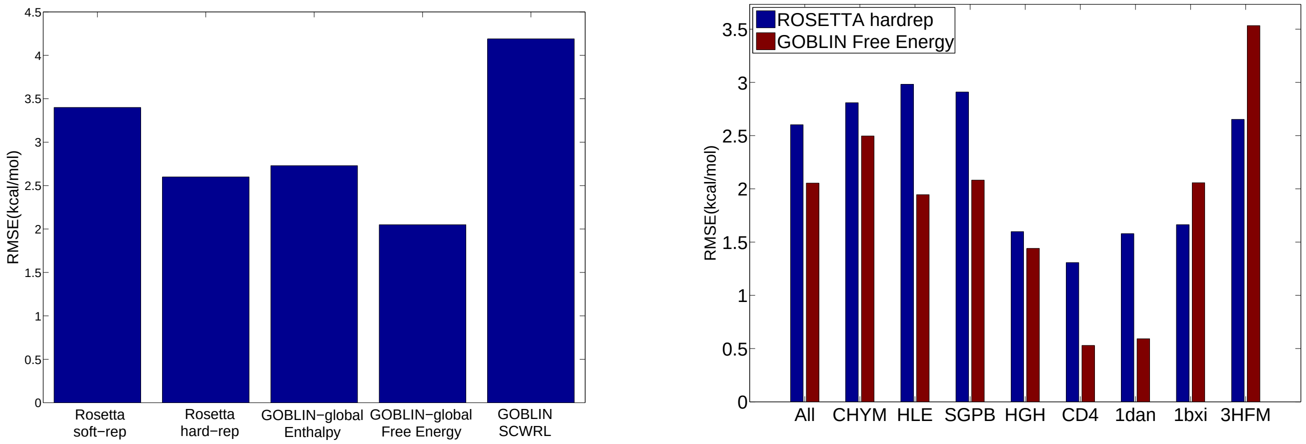
Table 1: Summary of datasets of $\Delta\Delta G$ values for mutant protein-protein complexes. Datasets annotated "AS" are Alanine-scanning experiments

(2002). We used the `soft-rep` atomic radii used in ROSETTA and the `soft-rep` force-field setting since previous studies (Yanover *et al.*, 2007) have indicated that it is better suited for computations with discrete conformations.

Since there are multiple contributions in this paper, we will describe the benefits of each of these by showing the incremental improvement in the accuracy of predicting $\Delta\Delta G$ due to each contribution. To clearly disambiguate between these models, we will refer to the side-chain free energy computation using generic weights as GOBLIN-global, the side-chain free energy optimized for a specific dataset as GOBLIN-spec and the free energy incorporating both backbone and side-chain flexibility with parameters optimized for a specific dataset as GOBLIN-bbflex-spec. To quantify the contribution of the improved force-field, we will also compare our results with our previously published approach (Kamisetty *et al.*, 2007, 2008) which used the SCWRL potential. We will refer to those results as GOBLIN-SCWRL.

Side-chain Flexibility: We first quantify the contributions of using a better force-field and incorporating entropic contributions, by comparing the RMSE in $\Delta\Delta G$ predictions over the entire dataset using GOBLIN-global, GOBLIN-SCWRL, and two different force-field settings of ROSETTA: `hard-rep` (the default) and `soft-rep`. GOBLIN-global uses the default set of `soft-rep` weights along with a weight of 1 for E_{rot}. Note that GOBLIN-global uses the `soft-rep` parameters because, unlike ROSETTA, it models side-chain flexibility while computing $\Delta\Delta G$.

Fig. 2-A illustrates a number of interesting findings from our study. First, GOBLIN-global outperforms ROSETTA-`hard-rep` by approximately 0.55 kcal/mol (second bar vs. fourth bar), and ROSETTA-`soft-rep` by approximately 1.35 kcal/mol (first bar vs. fourth bar). Second, in order to evaluate the amount by which our explicit entropic term reduces RMSE, we also computed RMSE when the entropic term is ignored. The enthalpy-only RMSE is comparable in accuracy to ROSETTA-`hard-rep` (second bar vs. third bar), indicating that the entropic factor reduces RMSE by roughly 0.7 kcal/mol. Third, the RMSE of the enthalpy-only term is roughly 0.65 kcal/mol smaller than ROSETTA-`soft-rep` (first bar vs. third bar). This is noteworthy because, as previously mentioned, both use the same force field parameters. That is, there is benefit to reporting an expected internal energy (i.e., the enthalpy) as opposed to the internal energy of a single structure. Finally, Fig. 2-A also shows the obvious benefit of moving from the simple force-field (GOBLIN-



(A) (B)

Figure 2: (A) Comparison of predictions by GOBLIN-unopt, ROSETTA, GOBLIN-unopt's enthalpy term, and our previously published work (Kamisetty *et al.*, 2007) using the SCWRL force-field instead of the ROSETTA force-field. GOBLIN performs significantly better (0.55 kcal/mol lower RMSE) than ROSETTA. Accounting for the entropic component of the free energy significantly improves GOBLIN-unopt's performance (0.70 kcal/mol lower RMSE) over the enthalpy-only term. (B) Breakdown of the error across the individual datasets for the two best methods in (A): ROSETTA's `hard-rep` and GOBLIN-global. The datasets are ordered according to their size, with the largest dataset appearing to the left.

SCWRL) used in Kamisetty *et al.* (2007, 2008) to a more expressive force-field used in the present study (fourth bar vs. fifth bar).

Fig. 2-B shows the breakdown of the two top methods, GOBLIN-global and ROSETTA-`hard-rep`, across the eight datasets listed in Table 1. The datasets are listed in decreasing order of their sizes. GOBLIN-global's error is much lower than ROSETTA on six out of the eight datasets; the largest improvement occurs in HLE (1.04 kcal/mol). On the two smallest datasets (1bxi and 3hfm), comprising less than 8% of our entire data, GOBLIN-global performs worse than ROSETTA.

On studying the performance of GOBLIN-global as a function of the kind of mutation, we observed that in many cases, mutations to Proline had a large error. This is to be expected since Proline has an atypical backbone and a mutation to it can cause significant structural changes. If we omit all mutations to Proline from the dataset, our global RMSEs decrease from 2.05 kcal/mol down to 1.92 kcal/mol. The most significant decrease in RMSE was observed in SGPB, where the error decreases from approximately 1.8 kcal/mol down to nearly 1.5 kcal/mol.

While the predictions to $\Delta\Delta G$ in mutations to Proline had large errors, they were not the only kinds of mutations to have such errors. For example, in the Chymotrypsin dataset, GOBLIN-global had the largest error on a mutation to Tryptophan (a bulky amino acid). To measure the effect of such outliers on the relative ordering of the various methods, we computed the RMSE for GOBLIN-global and ROSETTA-`hard-rep` by removing the top outliers for their respective predictions. We found that the RMSE of GOBLIN-global ignoring the top $k\%$ of the outliers was better by at least 0.3 kcal/mol for all values of k up to 75%. Thus, while GOBLIN-global performed poorly on some kinds of mutations, it was still consistently better than ROSETTA.

Learning: We believe that these results using GOBLIN-global itself represent a significant improvement over the state-of-the-art method of computing $\Delta\Delta G$. However, while a general-purpose set of force-field parameters is often useful when dealing with a protein-protein complex that hasn't been well studied, studies have shown the utility of considering the use of complex-specific force-field parameters (Lu *et al.*, 2001; Fong *et al.*, 2004). We optimize the force-field parameters for the same reasons, though it should be pointed out that our optimization method is fundamentally different than those used by other researchers owing to the fact that we have an explicit entropic term in our free energy.

To perform this optimization, we first compared the performance of a simple gradient descent algorithm and the constrained L-BGFS-B, a quasi-Newton gradient algorithm (Liu and Nocedal, 1989). Since the performance of L-BFGS-B was far superior both in running time and accuracy (results not shown), we simply used this approach for the rest of our experiments.

To test the efficacy of our learning algorithm, we learned a set of weights for the three largest datasets – the OMTKY datasets – by constructing five random partitions for each of the datasets into two equal sized train/test splits. For each training set, we learned a set of weights over the five force-field terms mentioned earlier, so as to minimize the square error in $\Delta\Delta G$ predictions as described in Sec. 4. We then computed the RMSE in the test set. Fig. 3 shows the error of GOBLIN-global (blue bar) and GOBLIN-spec (green bar). On all three datasets, incorporating learning decreases the average RMSE of predictions by 0.08-0.28 kcal/mol.

Backbone Flexibility: Finally, we incorporate backbone flexibility into our predictions. We used the approach described in Smith and Kortemme (2008) to sample a set of four plausible

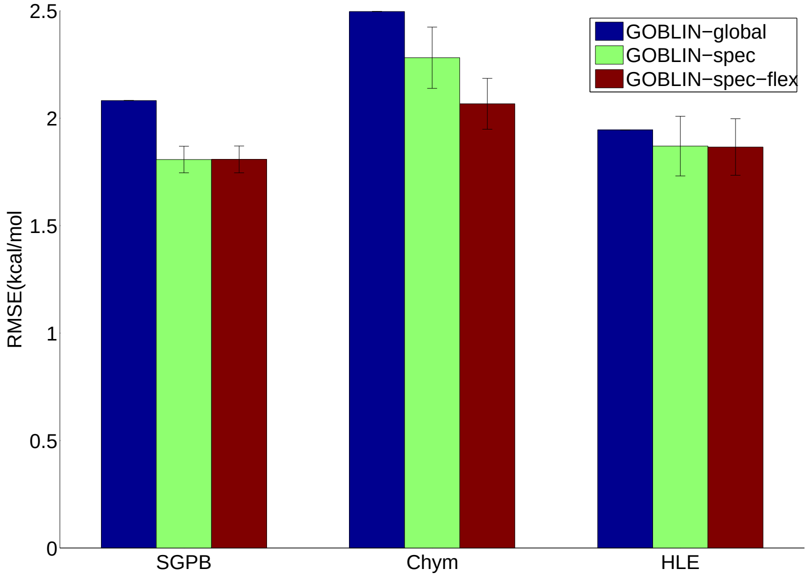


Figure 3: Comparison of $\Delta\Delta G$ prediction accuracy on the three OMTKY datasets, with vs. without backbone flexibility. The backbone ensemble was generated using ROSETTA's backrub mode as described in (Smith and Kortemme, 2008). The error bars on the learned models(GOBLIN-spec ,GOBLIN-bbflex-spec) show the standard deviation of the RMSE across the five test sets.

backbones for each wild-type and mutant structure. Together with the backbone of the crystal structure, this gives us a set of five backbones for each protein. We then repeated our procedure of learning weights to improve accuracy, this time using an ensemble of backbones. On two of the three datasets, incorporating backbone flexibility did not further improve the predictions. However, on the OMTKY-Chymotrypsin complex, incorporating backbone flexibility decreased RMSE by 0.22 kcal/mol to about 2.1 kcal/mol (Fig. 3, red bar). Fig. 4 shows the error of GOBLIN-bbflex-spec on the OMTKY-Chymotrypsin complex as a function of the mutations.

On closer examination, we found that in the OMTKY datasets, the errors were largely similar with and without backbone flexibility for most kinds of mutations which is consistent with studies showing that the structure of the OMTKY complexes is largely stable to mutations (Lu *et al.*, 2001). However, we expected that in cases of disruptive mutations, incorporating backbone flexibility would decrease the error. In the OMTKY-Chymotrypsin dataset, this did happen: the mutation with the worst error, Lysine to Tryptophan which is fairly disruptive owing to the bulky structure of Tryptophan, had lower error once backbone flexibility was incorporated. In SGPB and HLE on the other hand, the worst error was due to a mutation from Arginine to Proline. For this mutation, we found that the energies of the backrub structures were at least 7 kcal/mol worse than the native, resulting in nearly negligible contribution to the partition function (since they are exponentially down-weighted).

To see if this could be caused due to the number of backrub structures we used, we redid our GOBLIN-bbflex-spec experiments in SGPB and HLE with 10 backbone traces (9 backrub + 1 crystal). We found that despite doubling the number of backbone traces, our results did not change measurably (less than a 0.05 kcal/mol change in RMSE). To see if alternate sources of backbone traces would affect results, we generated and used backbone traces from FRODA (Wells *et al.*, 2005) instead of backrub; again, there was no measurable change in RMSE.

Based on these experiments, we concluded that the lack of improvement in SGPB and HLE was probably due to the poor ability of current methods that generate backbone traces to handle (disruptive) mutations to Proline.

6 Discussion and Conclusion

Fast and accurate free energy calculations are essential to a number of significant tasks within Computational Structural Biology, including structure-based protein and drug design. Our probabilistic graphical model-based approach to all-atom free energy calculations strikes a balance between the rigor of physical methods (i.e., molecular dynamics based free energy calculations) and the speed of statistical methods. Our method is physically rigorous in that (i) it uses all-atom force fields when computing internal energies, and (ii) it computes a rigorous approximation of the true partition function of the system. At the same time, our method is competitive with statistical methods, in terms of speed.

GOBLIN represents the first graphical model for protein-protein complexes, as well as the first graphical model simultaneously modeling backbone and side-chain flexibility. Furthermore, it demonstrates how to properly account for discretization of the conformation space in calculating free energies. Finally, it incorporates a novel algorithm for optimizing force fields by minimizing

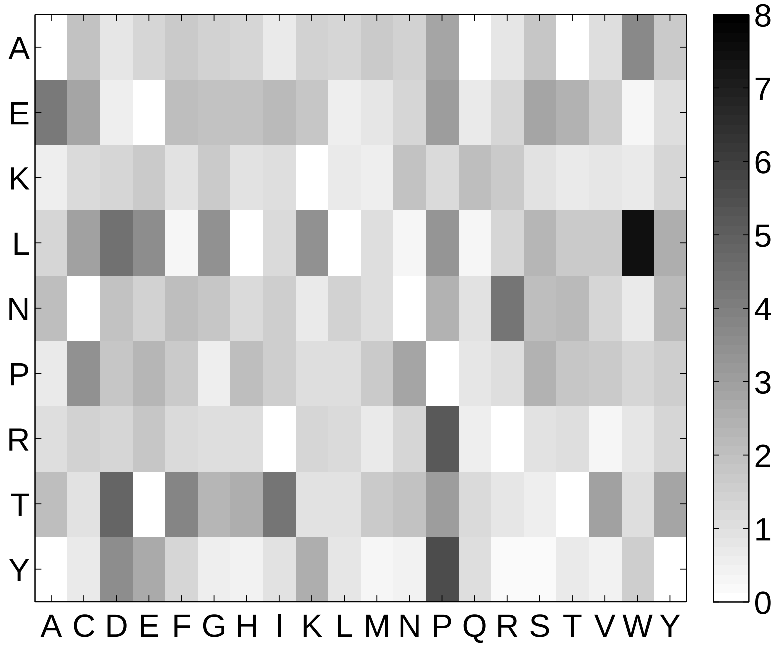


Figure 4: Mean Square Error between predicted and actual $\Delta\Delta G$ values for all mutations across all five test partitions of the Chymotrypsin dataset using GOBLIN-bbflex-spec. The Y-axis shows the amino acid of the wild-type and the X-axis shows that of the mutant. Our predictions were accurate for most types of mutations, with a few outliers. The largest outlier in this dataset was a mutation from Lysine to Tryptophan though the error was lower with GOBLIN-bbflex-spec than with GOBLIN-spec or GOBLIN

differences in free energies. We have shown that GOBLIN is both efficient and accurate on the task of computing the free energy of protein-protein complexes. In particular, it outperforms the gold-standard ROSETTA on a benchmark set of more than 700 mutants by at least 0.55 kcal/mol.

Our results indicate that an explicit incorporation of backbone and side-chain flexibility is feasible. Interestingly, for most complexes in our benchmark set, backbone flexibility did not substantially improve our results relative to the fixed-backbone case. There are two likely reasons for this. First, our chosen benchmark set contains relatively rigid complexes. That is, the backbones are largely stable *in vivo*, as previously suggested by Lu *et al.* (2001). We hope to study the effects of backbone flexibility in other complexes, most notably those where high throughput data are available (Pal *et al.*, 2006). Second, our backrub-generated backbones generally have high internal energies and therefore do not contribute substantially to the total free energy. To address this problem, we are presently investigating the use of energy-minimized backbones, as well as alternative methods for generating backbones.

We have also shown that it is possible to improve the accuracy of free energy predictions by approximately 0.26 kcal/mol on average by learning complex-specific parameters. This is to be expected, as proteins all have specific properties of stability, reactivity, and so forth that are hard to capture with a general set of parameters (Carter Jr *et al.*, 2001). The learning process also supports an iterative approach to protein design, in which an initial graphical model is used to design variants, which are then are then used to improve the model in order to design further improved variants. Such iterative approaches to protein design have been considered before (e.g., Liao *et al.* (2007)), although not using probabilistic graphical models.

In addition to providing a compact encoding of probability distributions supporting powerful and efficient algorithms for inference, another advantage of graphical models is their extendability. In this paper, for example, we have shown that modeling protein-protein complexes can be achieved in a straight-forward fashion by, in effect, composing two single-protein models of the type we previously introduced (Kamisetty *et al.*, 2007, 2008). Notice that our model does not distinguish between intra- and inter-molecular interactions. That is, we use a single set of force-field parameters. Statistical approaches, on the other hand, generally learn separate parameters for intra- and inter molecular interactions. We have also shown that it is also possible to extend the model to account for limited backbone flexibility. Others have shown that graphical models are well suited to such tasks as constructing protein backbone traces (DiMaio *et al.*, 2006) and all-atom models (DiMaio *et al.*, 2007) from electron density maps, modeling residue couplings (Thomas *et al.*, 2008), and protein sequence design (Fromer and Yanover, 2008; Thomas *et al.*, 2009).

Acknowledgments

This work is supported in part by US DOE Career Award (DE-FG02-05ER25696) (CJL), a grant from Microsoft Research (CJL), US NSF grant IIS-0444544 (CBK), and an Alfred P. Sloan Foundation Fellowship (CBK). We thank Menachem Fromer for raising the significant question of the effects of discretization on entropic contributions at ISMB 2008 and Prof. Tanja Kortemme for her kind assistance with some implementation details of Kortemme *et al.* (2003).

References

Åqvist, J., Luzhkov, V., and Brandsdal, B. (2002). Ligand binding affinities from MD simulations. *Acc. Chem. Res.*, **35**(6), 358–365.

Bahar, I., Atilgan, A. R., and Erman, B. (1997). Direct evaluation of thermal fluctuations in protein using a single parameter harmonic potential. *Fold. Des.*, **2**, 173–181.

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

Böhm, H.-J. (1992). The computer program LUDI: A new method for the de novo design of enzyne inhibitors. *J. Comput.-Aided Mol. Des.*, **6**(1), 61–78.

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., LeFebvre, B., Cammer, S., Tropsha, A., and Edgell, M. (2001). Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *Journal of Mol. Bio*, **311**, 625–38.

Chang, C., Mclaughlin, W., Baron, R., Wang, W., and McCammon, A. (2008). Entropic contributions and the influence of the hydrophobic environment in promiscuous protein-protein association. *PNAS*, **105**(21), 7456–7461.

Clifford, P. (1990). Markov random fields in statistics. In G. R. Grimmett and D. J. A. Welsh, editors, *Disorder in Physical Systems. A Volume in Honour of John M. Hammersley*, pages 19–32, Oxford. Clarendon Press.

Dagum, P. and Chavez, R. M. (1993). Approximating probabilistic inference in bayesian belief networks. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **15**(3), 246–255.

Davis, I., Arendall, W., Richardson, D., and Richardson, J. (2006). The backrub motion: How protein backbone shrugs when a sidechain dances. *Structure*, **14**(2), 265–274.

DiMaio, F., Shavlik, J., and Phillips, G. (2006). A probabilistic approach to protein backbone tracing in electron density maps. *Bioinformatics*, **22**(14), e81–e89.

DiMaio, F., Soni, A., Phillips Jr., G., and Shavlik, J. (2007). Creating all-atom protein models from electron-density maps using particle-filtering methods. *Bioinformatics*, **23**, 2851–2858.

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, **340**(6230), 245–246.

Fong, J., Keating, A., and Singh, M. (2004). Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biol.*, **5**(2).

Friedland, G., Linares, A., Smith, C., and Kortemme, T. (2008). A simple model of backbone flexibility improves modeling of side-chain conformational variability. *J. Mol. Biol.*, **380**(4), 757–774.

Fromer, M. and Yanover, C. (2008). A computational framework to empower probabilistic protein design. *Bioinformatics*, **24**(13), i214–222.

Georgiev, I., Keedy, D., Richardson, J. S., Richardson, D. C., and Donald, B. R. (2008). Algorithm for backrub motions in protein design. *Bioinformatics*, **24**(13).

Gohlke, H., Kiel, C., and Case, D. (2003). Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J. Mol. Biol.*, **330**(4), 891–913.

Jarzynski, C. (1997). A nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, **78**, 2690.

Jaynes, E. T. (1963). Information theory and statistical mechanics. *Statistical Physics*, pages 181–218.

Jaynes, E. T. (1968). Prior probabilities. *Systems Science and Cybernetics, IEEE Transactions on*, **4**(3), 227–241.

Joachimiak, L., Kortemme, T., Stoddard, B., and Baker, D. (2006). Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. *J. Mol. Biol.*, **361**, 195–208.

Kamisetty, H., Xing, E., and Langmead, C. (2007). Free Energy Estimates of All-atom Protein Structures Using Generalized Belief Propagation. In *Proc. 7th Ann. Intl. Conf. on Research in Comput. Biol. (RECOMB)*, pages 366–380.

Kamisetty, H., Xing, E. P., and Langmead, C. J. (2008). Free energy estimates of all-atom protein structures using generalized belief propagation. *J. Comp. Biol.*, **15**(7), 755–766.

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

Kingsford, C. L., Chazelle, B., and Singh, M. (2005). Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, **21**, 1028–1036.

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.

Kortemme, T. and Baker, D. (2002). A simple physical model for binding energy hot spots in protein-protein complexes. *PNAS*, **99**(22), 14116–14121.

Kortemme, T. and Baker, D. (2004). Computational design of protein-protein interactions. *Curr. Opin. Chem. Biol.*, **8**(1), 91–97.

Kortemme, T., Morozov, A. V., and Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.*, **326**(4), 1239–1259.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**(5649), 1364–1368.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Lazaridis, T. and Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins*, **35**(2), 133–152.

Leach, A., Shoichet, B., and Peishoff, C. (2006). Prediction of protein-ligand interactions. docking and scoring: successes and gaps. *J. Med. Chem.*, **49**(20), 5851–5855.

Liao, J., Warmuth, M., Govindarajan, S., Ness, J., Wang, R., Gustafsson, C., and Minshull, J. (2007). Engineering proteinase k using machine learning and synthetic genes. *BMC Biotechnology*, **7**(1), 16.

Lilien, R., Stevens, B., Anderson, A., and Donald, B. R. (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. Comp. Biol.*, **12**(6-7), 740–761.

Liu, D. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, **45**(1), 503–528.

Lovell, S., Word, J., Richardson, J., and Richardson, D. (2000). The penultimate rotamer library. *Proteins*, **40**, 389–408.

Lu, S., Lu, W., Qasim, M., others, and Laskowski, Jr., M. (2001). Predicting the reactivity of proteins from their sequence alone: Kazal family of protein inhibitors of serine proteinases. *PNAS*, **98**(4), 1410–1415.

McGregor, M. J., Islam, S. A., and Sternberg, M. J. (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol.*, (2), 295–310.

Missimer, J. H., Steinmetz, M. O., Baron, R., Winkler, F. K., Kammerer, R. A., Daura, X., and van Gunsteren, W. F. (2007). Configurational entropy elucidates the role of salt-bridge networks in protein thermostability. *Protein Sci.*, **16**(7), 1349–1359.

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., and Kaburagi, M. (1994). Foundations and applications of cluster variation method and path probability method. *Prog. Theor. Phys. Supplement*, **115**.

Muegge, I. (2006). PMF scoring revisited. *J. Med. Chem.*, **49**(20), 5895–5902.

Pal, G., Kouadio, J., Artis, R., Kossiakoff, A., and Sidhu, S. (2006). Comprehensive and Quantitative Mapping of Energy Landscapes for Protein-Protein Interactions by Rapid Combinatorial Scanning. *Journal Of Biological Chemistry*, **281**(31), 22378.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artif. Intell.*, **29**(3), 241–288.

Ponder, J. W. and Richards, F. M. (1987). Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, **193**(4), 775–791.

Smith, C. A. and Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.*, **380**(4), 742–756.

Srinivasan, J., Cheatham, T., Cieplak, P., Kollman, P., and Case, D. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *J. Am. Chem. Soc.*, **120**(37), 9401–9409.

Thomas, J., Ramakrishnan, N., and Bailey-Kellogg, C. (2008). Graphical models of residue coupling in protein families. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **5**(2), 183–197.

Thomas, J., Ramakrishnan, N., and Bailey-Kellogg, C. (2009). Protein design by sampling an undirected graphical model of residue constraints. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*. In press.

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

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins*, **52**(4), 609–623.

Wang, W. and Kollman, P. (2001). Computational Study of Protein Specificity: The molecular basis of HIV-1 protease drug resistance. *PNAS*, **98**(26), 14937–14942.

Warren, G., Andrews, C., Capelli, A., Clarke, B., LaLonde, J., Lambert, M., Lindvall, M., Nevins, N., Semus, S., Senger, S., Tedesco, G., Wall, I., Woolven, J., Peishoff, C., and Head, M. (2006). A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, **49**(20), 5912–5931.

Weber, I. T. and Harrison, R. (1999). Molecular mechanics analysis of drug-resistant mutants of HIV protease. *Protein Engineering*, **12**(6), 469–474.

Wells, S., Menor, S., Hespenheide, B., and Thorpe, M. (2005). Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol*, **2**, S127–S136.

Word, J., Lovell, S., Richardson, J., and Richardson, D. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, **285**(4), 1735–1747.

Xu, J. (2005). Rapid protein side-chain packing via tree decomposition. In *Proc. 9th Ann. Intl. Conf. on Comput. Biol. (RECOMB)*, pages 423–439.

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

Yanover, C., Schueler-Furman, O., and Weiss, Y. (2007). Minimizing and learning energy functions for side-chain prediction. In *Proc. 7th Ann. Intl. Conf. on Research in Comput. Biol. (RECOMB)*, pages 381–395.

Yedidia, J., Freeman, W., and Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, **51**, 2282–2312.