

Incorporating Protein Flexibility in Predicting Free Energies of Association for Protein-protein Interactions

Hetunandan Kamisetty¹, Chris Bailey-Kellogg² and Christopher James Langmead^{1,3,4}

1 A Markov Random Field Model for Protein Complexes

Protein-protein interactions are essential to the molecular machinery of the cell. A fundamental law of thermodynamics states that interactions are governed by *binding free energies*, or equivalently, the change in the log partition function of a statistical ensemble. This paper introduces a method, called GOBLIN (*Graphical mOdel for BiomoLecular INteractions*), that predicts binding free energies via an approximation of the partition function Z of the equilibrium Boltzmann distribution.

Boltzmann’s law describes the probability distribution over the configurational space \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_c is $P(\mathbf{X} = \mathbf{x}_c) = \frac{1}{Z} \exp\left(-\frac{E_c}{k_B T}\right)$ where $Z = \sum_{\mathbf{x}_c \in \mathcal{C}} \exp(-E_c)$ is the *partition function*, k_B is Boltzmann’s constant, and T is the absolute temperature in Kelvin. The free energy is then simply the log of partition function, $\log Z$.

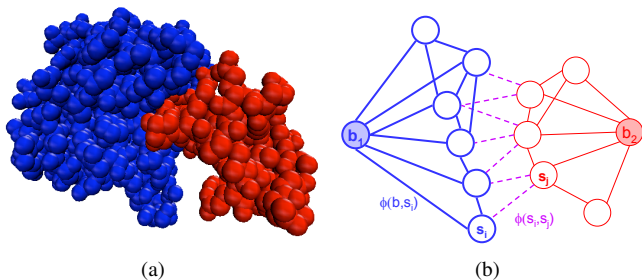


Figure 1: (a) Chymotrypsin complexed with the third domain of turkey ovomucoid (OMTKY). While protein structures are often shown as a single conformation, in reality they occupy ensembles of conformations; our method models both side-chain and backbone ensembles. (b) Part of an MRF encoding the conditional distribution over the ensembles of conformations. Solid lines refer to intra-molecular interactions, and dashed lines refer to inter-molecular interactions.

as a Markov Random Field(MRF) as shown in Fig. 1.

We can then rewrite the joint distribution and the partition function as $P(\mathbf{X} = \mathbf{x}_c) = P(\mathbf{X}_b = \mathbf{x}_b)P(\mathbf{X}_s|\mathbf{X}_b = \mathbf{x}_b)$ and $Z = \sum_{\mathbf{x}_b} \exp\left(-\frac{E_b}{k_B T}\right)Z_b$ where $Z_b = \sum_{\mathbf{x}_s} \exp\left(-\frac{E_s}{k_B T}\right)$ is the partition function over the side-chain conformational space with a fixed backbone. By computing the partition function for each backbone Z_b , we can thus compute the partition function Z over all backbones.

2 Probabilistic Inference and Free Energy Calculations

Probabilistic inference in an MRF involves computing marginal distributions over the random variables in the graph. In general, the problem is intractable. 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 [4]. We use Pearl’s *Belief Propagation* (BP) algorithm [3] that has been shown to minimize the Bethe approximation of the free energy, to approximate Z_b and therefore Z .

Our approach uses the ROSETTA force field to compute the energy of a configuration. As an optional step to optimize the parameters of the force-field for specific protein-protein complexes, we develop a novel algorithm that determine parameters that minimize the mean square error (MSE) in predicting $\Delta\Delta G$ of a training set that consists of experimentally measured changes in $\Delta\Delta G$. A more detailed presentation of this work is available in [1].

¹Computer Science Department, Carnegie Mellon University, Pittsburgh, PA.

²Department of Computer Science, Dartmouth College, Hanover, NH.

³Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA

⁴This research was sponsored in part by a grant from Microsoft Research to CJL.

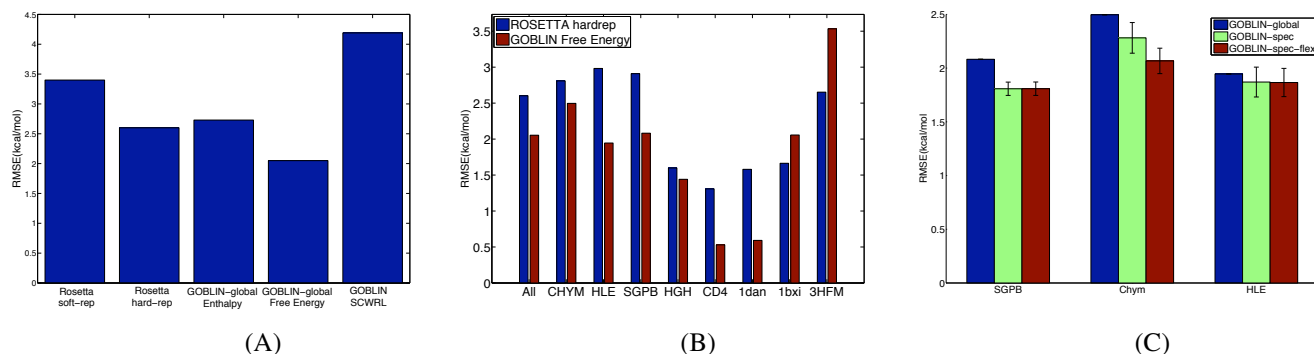


Figure 2: (A) Comparison of predictions by GOBLIN-uno,pt, ROSETTA, GOBLIN-uno,pt’s enthalpy term, and GOBLIN-SCWRL. Accounting for the entropic component of the free energy significantly improves GOBLIN-uno,pt’s performance (0.70 kcal/mol lower RMSE) over the enthalpy-only term. (B) Breakdown of the 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. (C) Comparison of $\Delta\Delta G$ prediction accuracy on the three largest datasets, with vs. without backbone flexibility. The error bars on the learned models(GOBLIN-spec,GOBLIN-spec-flex) show the standard deviation of the RMSE across the five test sets.

3 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 which experimentally measured $\Delta\Delta G$ are available. We compare the accuracy of various settings of GOBLIN : side-chain flexibility only(GOBLIN-global); side-chain flexibility with parameters optimized for a specific dataset(GOBLIN-spec); backbone and side-chain flexibility with parameters optimized for a specific dataset(GOBLIN-spec-flex). We also compare our results with our previously published approach [2] which used a simple Lennard-Jones potential (GOBLIN-SCWRL).

Our results(Fig. 2) show that GOBLIN is 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 incorporating protein flexibility. 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.

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, typically requiring less than 5 minutes per calculation.

References

- [1] H. Kamisetty, C. Bailey-Kellogg, and C. J. Langmead. A graphical model approach for predicting free energies of association for protein-protein interactions under backbone and side-chain flexibility. Technical Report CMU-CS-08-162, Carnegie Mellon University, 2008.
- [2] H. Kamisetty, E.P. Xing, and C.J. Langmead. 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, 2007.
- [3] J. Pearl. Fusion, propagation, and structuring in belief networks. *Artif. Intell.*, 29(3):241–288, 1986.
- [4] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51:2282–2312, 2005.