Graph-Based Representations of the Protein Design Problem

Adi Mukund Jennifer Zou

1 Context

The protein design problem can be described as follows. Given a set of backbone coordinates $\mathbf{c} = (\vec{c_1}, \vec{c_2}, \ldots, \vec{c_n})$ and rotamer library R of length r, identify the optimal rotamer assignment sequence $\vec{r} = (r_1, r_2, \ldots, r_n)$, $r_i \in R$, $1 \leq i \leq n$ according to an energy function $E(\vec{r})$. This assignment sequence is known as the global minimum energy conformation (GMEC). This problem has been proven to be NP-hard [5], and algorithms such as DEE [1] have been created to allow for combinatorial pruning of the residue search space and make the design problem computationally tractable. However, such algorithms cannot guarantee a time complexity less than the worst-case $O(nr^n)$.

Recent graph based algorithms such as BWM* [2] use sparse residue interaction graphs in order to more efficiently compute functions over the residue space and identify optimal assignments more rapidly. Such graph-based algorithms have been able to achieve combinatorial speedups while maintaining provable accuracy and returning ensemble of minimum energy conformations.

Probabilistic models of protein design assign a probability distribution for rotamers in each position of the protein sequence. In algorithms such as belief propagation, the beliefs or the approximate marginal probabilities of each rotamer are computed iteratively. These approximations are computationally less intense, but they are not always provably accurate [3]. Furthermore, such algorithms are only guaranteed to converge on tree graphs, which are not commonly observed in natural proteins [4].

2 Goals

2.1 Provable Convergence

The project will attempt to develop provably accurate algorithms utilizing belief propagation and graph cuts on sparse residue interaction graphs to solve the protein design problem. The major obstacle will be maintaining provable convergence in the presence of loops in the sparse graph. We will attempt to do this by:

1. Using graph cuts and dynamic programming in order to treat loops within the residue graphs as subproblems and maintain provable accuracy.

2. Continually pruning edges until the graph is sufficiently sparse, maintaining an error bound in order to see the maximum difference ϵ in energy between the full GMEC and sparse GMEC, and enumerating all low- energy conformations produced by the sparse interaction graph within ϵ of the sparse GMEC.

3. Attempting to find generalized belief propagation algorithms that provably converge on graphs that contain loops, and applying those algorithms to the protein design problem.

4. Investigating other properties of the sparse residue interaction graph such as planarity and seeing if these allow for provably accurate algorithms.

2.2 Machine Learning Models of Protein Stability

We hope to train energy functions in order to improve the speed of residue-assignment algorithms. Current energy functions include pairwise terms and are frequently slow enough that running them on a large number of possible conformations is impossible; machine learning models could yield significant speed improvements.

Machine learning methods may also increase accuracy of residue-assignment algorithms. When creating sparse residue interaction graphs, assumptions must be made about which interactions to include. These assumptions are usually made with distance or energy cutoffs, which may be biased. Furthermore, many energy functions do not include terms for entropy, which can decrease the accuracy of residue assignment algorithms [6]. By training directly with free energy data, these issues may be addressed.

Using protein sequence and thermodynamic data from the Protein Data Bank and ProTherm datavases, we hope to train energy functions to predict energies of novel sequences. A model such as a neural network may be ideal due to its flexibility and ability to capture nonlinear relationships.

2.3 Test Accuracy

Once we have implemented our approximation algorithms, we would like to test them for accuracy and compare the results to known algorithms. We would also like to test the effectiveness of the algorithms different protein families and for different regions (surface, boundary, and core) of proteins. This will allow us to assess whether certain algorithms are better for specific types of structures.

References

[1] Dahiyat, B. I. De Novo Protein Design: Fully Automated Sequence Selection. Science 278, 82-87 (1997)

[2] Jou, J.D. Jain, S. Georgiev, I. Donald, B.R. BWM*: A Novel, Provable, Ensemble-based Dynamic Programming Algorithm for Sparse Approximations of Computational Protein Design. RECOMB (2015) Warsaw, Poland. April 12, 2015 (In Press)

[3] Kamisetty H1, Xing EP, Langmead CJ. Free energy estimates of all-atom protein structures using generalized belief propagation. J Comput Biol. 2008 Sep;15(7):755-66. doi: 10.1089/cmb.2007.0131.

[4] M. Fromer, C. Yanover, and M. Linial. Design of multispecific protein sequences using probabilistic graphical modeling. Proteins 75;3(2009 May 15):682-705.

[5] Pierce, Niles A. and Winfree, Erik. Protein Design is NP-hard Protein Eng. (2002) 15 (10): 779-782 doi:10.1093/protein/15.10.779

[6] Silver, N.W. King, B.M. Nalam, M.N. Cao, H. Ali, A. Kiran Kumar Reddy, G.S. Rana, T.M. Schiffer, C.A. Tidor, B. Efficient Computation of Small-Molecule Configurational Binding Entropy and Free Energy Changes by Ensemble Enumeration. J Chem Theory Comput 9, 5098-5115 (2013).