1 Introduction

Protein structure prediction is a seminal problem in computational biology. The goal of protein structure prediction is to infer the 3D structure of a protein from its amino acid sequence. Previous work in this field generally uses simulated annealing to determine protein structure, however Pacheco et. al. are currently attempting to apply the D-PMP algorithm, which had success in the sub-problem of protein side chain prediction, to full structure prediction. In order to obtain efficient inference for D-PMP it is necessary to provide a sparse graphical model. Currently, D-PMP is being provided with the true graph structure, obtained using knowledge of the correct protein structure. Obviously, this is impossible when attempting to predict protein structure without prior knowledge of the correct structure. We propose to develop a method for estimating this sparse graphical model of a protein from its amino acid sequence (a subproblem known as protein contact prediction). If successful, this would allow the application of D-PMP to non-training data, which (if D-PMP is successful) would be a significant advancement in computational biology. More accurate methods of structure prediction would improve the speed of biochemical research and drug design.

A protein can be represented with a Pairwise Markov Random Field by assigning a variable to each amino acid to keep track of its position and orientation, and adding an edge between variables if the corresponding amino acids are close enough to interact. Various forces affect atoms in a protein; they are bonded to other atoms in the protein, and so must stay close to them, but are also repelled by some atoms, so try to move further from those. Atoms that are strongly under the influence of some attractive or repulsive force will have high energy. Note that the complete graph is always a valid representation of a protein. Inter-atomic energies never vanish, but they have an inverse square relationship with distance. Therefore it is computationally beneficial to remove edges between amino acids that are beyond some specified interaction distance, and thus have inter-atomic energy near zero. On the other hand, removing edges between amino acids that are within that interaction distance can significantly impact energy evaluation and cause inference to fail. Seen in this manner, the problem of predicting a sparse graph structure is equivalent to determining the probability that two amino acids will be within the interaction distance in the final protein structure, with a goal of limiting false negatives.

2 Related Work

Predicting amino acid contacts is an extremely active area of research in the field of computational protein structure prediction. An accurate set of protein contacts can be used to initialize a protein structure prediction algorithm with a restricted set of edges, reducing the computational load of the molecular interaction calculations [1]. As well, efforts to predict a 3D protein structure *de novo* by using amino acid proximity estimates to constrain structure have proven partially successful [2]. Rather than attempt to predict amino acid contacts from solely a sequenced protein, modern contact prediction algorithms have relied on relationships between homologous protein domains, which are subdivisions of proteins that are evolutionarily related [1] [2][3]. Ekeberg et al. [3] use sequence-aligned protein domains to fit a Potts model to protein domain families. In this model, amino acid positions in a sequence are likely to be in contact if their values are correlated across many proteins domains in the sequence's family. Put another way, amino acids in contact with each other are likely to evolve in parallel, or risk damaging the greater protein structure. Other approaches, such as that of Ma et al. [4] model protein families as Gaussian Graphical Models, with some correlations existing between families. Protein domain family data is plentiful, with 14,000 families listed in the Pfam database [3]. However, even with this large amount of data, predicting amino acid contacts in a protein requires a Potts model for each of the different domains within said protein. Perhaps more relevant to our application is the tendency of these models, such as the Potts model presented by Ekeberg et al., to produce many false negatives in its final predictions [3]. As mentioned above, false negatives are much more detrimental to D-PMP inference than false positives. Combining evolutionary correlation models with sequence-level features fed into a supervised learning system, like the CoinFold contact prediction system presented by Wang et al. [5], may produce a model more uniquely suited to initializing a physical protein structure prediction algorithm.

3 Model

Let $A_1, A_2, \ldots, A_N$ represent the amino acid sequence of a protein. Each variable A_i can be represented as a tuple $(i, type)$ where i denotes the position in the sequence and *type* denotes the amino acid type. Our goal is to predict a set of binary interaction variable $y_{ij}, i \neq j$ which indicate whether the amino acids A_i, A_j are close enough to interact. Since distance is symmetric, $y_{ij} = y_{ji}$ for every i, j. Let

$$p(y_{ij} \mid A_i, A_j) \sim Ber(y_{ij} \mid \phi(A_i, A_j))$$

where ϕ is a function on two amino acids and outputs a number in $[0, 1]$. The graphical model corresponding to this is given by Figure 1.

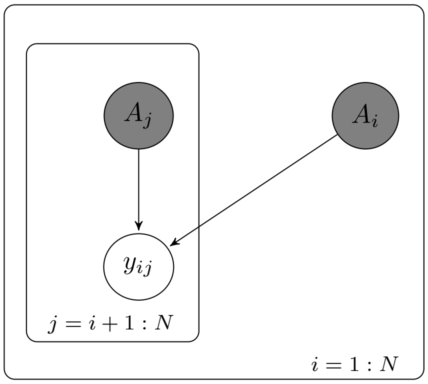

Figure 1: Simplest graphical model. A_i, A_j represent observed amino acids in the sequence, y_{ij} is an indicator variable that determines whether the two are close enough to interact.

This is a very simple model of a protein. Note that given the observed amino acid sequence, each interaction potential is conditionally independent of every other potential. This reduces the problem to predicting estimates for a series of independent unary factors, which is both unlikely to be successful and not particularly interesting. However, it could provide a baseline against which to test other models.

Here, we suggest several possible model adjustments that could improve performance. An obvious improvement would be to place a prior on the distribution of each y_{ij}, so that

$$p(y_{ij} \mid A_i, A_j) \sim Ber(y_{ij} \mid \phi(A_i, A_j))Beta(\alpha, \beta)$$

Adjusting α and β could help encourage or discourage sparseness in final prediction. Another possible change is adding an interaction term between trios of factors y_{ij}, y_{jk}, and y_{ik} such that the presence of an edge between amino acids i and j, and of an edge between amino acids j and k, increases the probability of an edge between amino acids i and k. Including all these terms would result in the graphical model predicted in Figure 2.

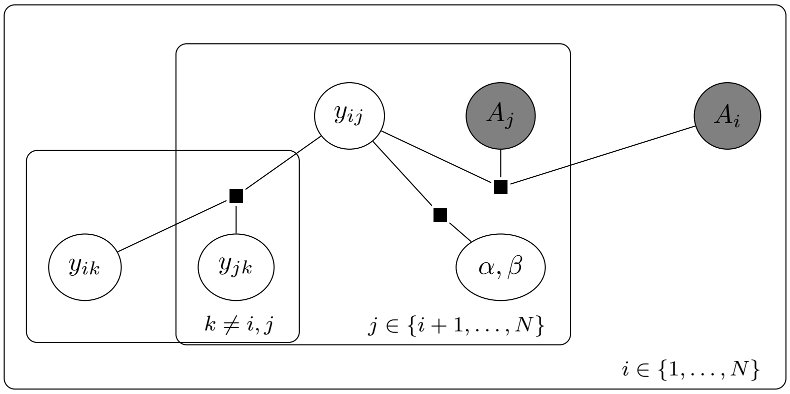

Figure 2: Potential graphical model with multiple additions. y_{ij}, y_{jk}, y_{ik} have factor increasing the probability of the third if the other two equal 1. α, β are hyperparameters on a prior on the distribution of y_{ij}.

4 Challenges

There are two challenges to overcome. First, we must learn feature functions ϕ for the probability of y_{ij} given A_i, A_j and for the joint distribution $p(y_{ij}, y_{jk}, y_{ik})$. We can obtain the true graph structures of proteins using data from the Protein Data Bank, then use these graph structures, along with the corresponding amino acid sequences, to determine these distributions. We are also looking into using evolutionary data to determine if certain types of amino acids are more or less likely to interact. Second, we must perform inference on the graphical model. This may be possible using loopy belief propagation (although with the additional factor $p(y_{ij}, y_{jk}, y_{ik})$ the graph will be very loopy).

Additionally, we hope to further refine the estimated structure using D-PMP inference. We plan to initialize inference using the estimated (potentially flawed) graph structure. Then, after performing inference for some number of iterations, we will use the result to obtain a new graph structure. It is unlikely that inference will find the exact positions of amino acids using a flawed graph structure, however it may be able to find approximate positions and thus fix missing links or remove unnecessary ones in the graph structure.

5 Data

To train and test our model, we are using protein data from the Protein Data Bank [6]. PDB structure factor files contain x, y, z coordinates for each atom in each amino amino acid for a sequenced proteins. From these coordinates, distances between atoms can be computed and converted into a graph, where edges exist between amino acids if they lie within a threshold distance d of each other, with $d = 10$ for our initial data processing. We are using an 1000-protein

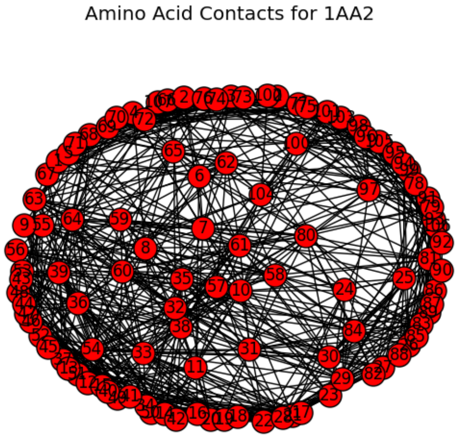

Figure 3: Amino Acid Contact graph for Protein 1AA2.

subset of the 100732 structure for our initial model training and testing, with 750 allocated for training and 250 for test. If our model proves reasonably successful, we plan to scale up the size of our training and test data, restricted only by our available storage and computational resources. These graphs will serve two primary purposes: training and testing our factor potentials, as well as providing ground-truth data for our edge prediction model. Fig. 1 contains example of a graph produced from a processed PBD file:

We observe that for this protein, most edges are highly localized among adjacent amino acids. However, many edges have connections across the protein, which is in and of itself an indicator of protein fold structure. We also observe that this graph is relatively sparse (compared to the complete graph), with 882 edges connecting 108 amino acids.

6 Evaluation

As mentioned above, methods for protein contact prediction already exist, at least two of which are freely available on the internet.[3] [5]. The CoinFold model described by Wang et al. uses both the amino acid sequence and evolutionary data to predict contact between amino acids in a protein. It is not a direct parallel, because CoinFold relies on a secondary and tertiary prediction system to perform contact prediction in addition to inferring contacts from sequence and evolutionary data. As well, CoinFold attempts to minimize total error, while we seek to minimize false negatives. Despite this, CoinFold would likely provide a good baseline against which to evaluate. The Ekeberg et al. [3] Potts model is more limited in its application, due to its reliance on domain-specific information, but could also provide a valuable baseline for our model. A key concept used for the evaluation of both systems is the differential evaluation of short, medium, and long-range contacts, which provides a more granular knowledge of model strengths and weaknesses. Predicting contacts between amino acids that are close in the sequence is intuitively simpler than predicting long-range contacts, which may implicitly represent complex secondary and tertiary protein structure.

We also plan to evaluate different models against each other. For example, the model with joint $p(y_{ij}, y_{jk}, y_{ik})$ can be evaluated against the naive model involving only unary factors (assuming A_i, A_j are observed). To evaluate the use of our model combined with D-PMP inference, we plan to initialize D-PMP inference with different graph structures including a complete graph, a random graph, and a graph with connections only along the backbone (i.e. between amino acids A_i and A_{i+1}). This will help determine what improvements come from our model versus from D-PMP inference.

7 Timeline

References

[1] Qiqige Wuyun, Wei Zheng, Zhenling Peng, and Jianyi Yang. November 1, 2016. A large-scale comparative assessment of methods for residue?residue contact prediction. Brief Bioinform (2016) doi:10.1093/bib/bbw106

[2] Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, Chris Sander. 25 October 2011. 3D Protein Structure Predicted from Sequence. `arXiv:1110.5091 [q-bio.BM]`

[3] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, Erik Aurell. 12 January 2013. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. `arXiv:1211.1281 [q-bio.QM]`

[4] Ma J., Wang S., Wang Z., Xu J. August 14, 2015. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. Bioinformatics 2015:btv472

[5] Sheng Wang, Wei Li, Renyu Zhang, Shiwang Liu and Jinbo Xu. April 12, 2016. CoinFold: a web server for protein contact prediction and contact-assisted protein folding. Nucl. Acids Res. (2016) doi: 10.1093/nar/gkw307

[6] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. 2000. The Protein Data Bank Nucleic Acids Research, 28: 235-242.