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

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 **??**.

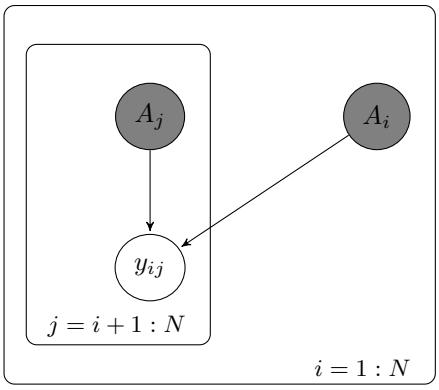

Figure 1: Simplest graphical model. A_i, A_j represent observed amino acids in the sequence, y_{ij} is an indicator variable that determines whether the two are close enough to interact.

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 **??**.

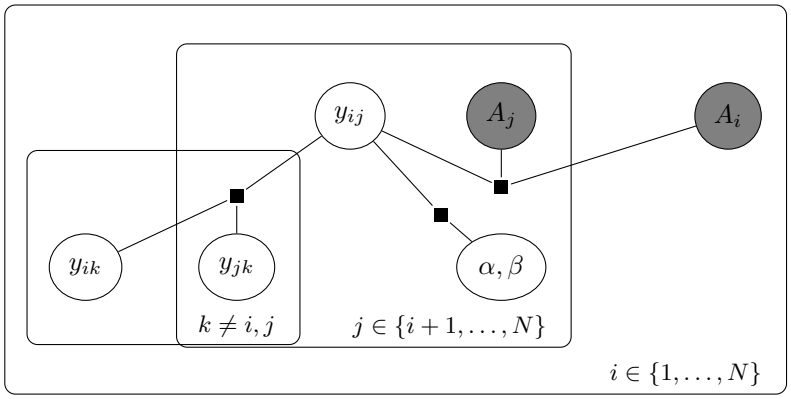

Figure 2: Potential graphical model with multiple additions. y_{ij}, y_{jk}, y_{ik} have factor increasing the probability of the third if the other two equal 1. α, β are hyperparameters on a prior on the distribution of y_{ij}.

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 Evaluation

Some methods for protein contact prediction already exist, such as CoinFold. CoinFold 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 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.

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.

6 Timeline