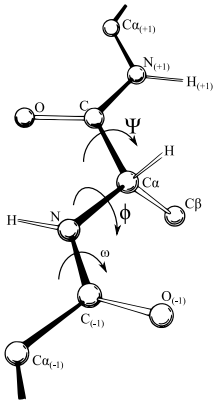
An ML Search Approach to Protein Structure Prediction

Michael Untereiner

1 Introduction

Protein structure prediction (PSP) is a foundational problem in bioinformatics. PSP is defined as follows: given the amino acid sequence of a protein, predict the native 3-dimensional configuration of the folded protein. The goal of this project is to develop a hierarchical Bayesian model to infer the structure of a protein from its primary sequence.

The structures of thousands of proteins have been experimentally validated using x-ray crystallography. A major challenge for building ML models with these sequences is that they display a large variation in length, from hundreds to thousands of amino acids. To get around this problem, I will develop a Bayesian learner which treats each amino acid in the protein chain as an independent variable, but which samples dihedral angles for each amino acid according to a conditional distribution determined by surrounding amino acids in the sequence. I expect the performance of this model to improve once conditioned on longer sequences.



2 Preprocessing

Data was acquired for 100,000 proteins from the PDB database, available at:

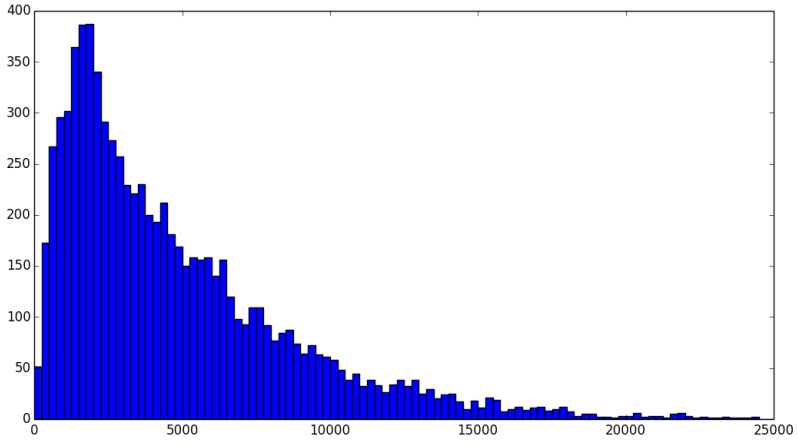
ftp://ftp.wwpdb.org/pub/pdb/data/structures/divided/mmCIF

Dihedral angles were compiled into a conditional distribution, where for a given x, an amino acid at the middle position in the n-letter sequence s, $p(x|s)$ is a probability distribution over the possible dihedral angles of x, calculated from the relative frequencies of dihedrals observed in the data. Dihedral angles were calculated from atomic coordinates using the geometric relationship $\cos\theta = \frac{|n_A \cdot n_B|}{|n_A||n_B|}$, where n_A and n_B are the normal vectors to the two nitrogen-central carbon-carbon planes of adjacent amino acids. Finally, angles were rounded to the nearest degree.

Sequence length	Present in data	Total possible	Percentage
3	8000	8000	100
5	1.6 million	3.2 million	50
7	6 million	1.2 billion	5

The table above compares the numbers of distinct n-length sequences in the dataset with the numbers of possible sequences. To our surprise, the proportion of possible sequences utilized in biological proteins is very small for $n > 3$. It appears that the size of our conditional distribution will scale linearly with the length n, in which case we could add conditionals for $n > 7$.

This histogram shows the number of distributions by how many observations were found in the data. The average distribution has approximately 2000 observations. N-means clustering with a value of N=10 was repeatedly applied to each distribution until the number of sample points fell below the mean value of 2000. This was to reduce the memory size of the distributions and to create a smaller search space for sampling.



3 Bayesian Model

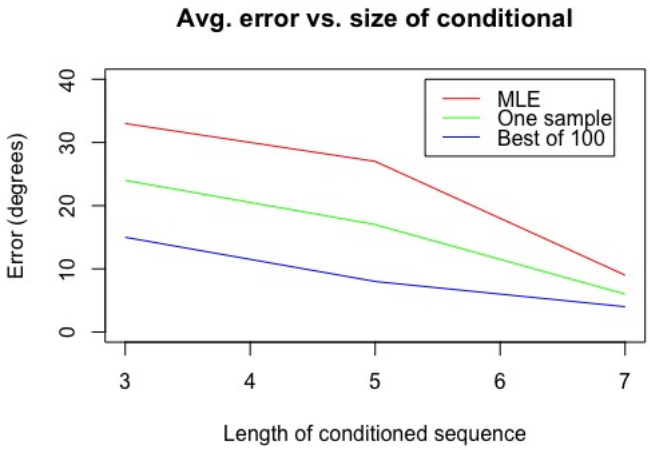
Under our Bayesian model, estimates of protein structure are produced by iterating over the amino acids of the chain and repeatedly sampling from the conditional distributions.

Major improvements in performance can be achieved by manipulating the distributions, using more sophisticated markov chain sampling methods, or both. The distributions themselves can be modified with n-means clustering to clump dihedrals which are distinct when rounded to the nearest degree, yet still very close to one another, into a single dihedral with greater probability mass, and therefore greater likelihood of being sampled. In addition, markov chain monte carlo methods may be able to refine sampling to regions of conformational space which are very close to the lowest energy configuration by analyzing previous samples using a scoring function which converts backbone dihedral conformations back to atomic coordinates, establishes bonds between constituent amino acids, and uses these bond energies to score the configuration.

The energy function was approximated by converting dihedral representations back into 3D coordinates and detecting contacts between amino acids using the distance formula. The goal of sampling will be to maximize this function.

4 Results

Average error was calculated for prediction on a set of 500 proteins with an average length of 220.



The MLE estimate takes the highest probability dihedral for each amino in the protein chain. In this sense, it maximizes the likelihood under the conditionals. But, a more sophisticated estimator would sample many possible conformations from the conditional and choose the best using some score function. Interestingly, taking a random sample from the conditional outperforms the MLE, suggesting that the best estimates are some distance from maximum likelihood but likely enough to perform well with single sampling. If we take 100 samples and choose the one with lowest error, we significantly outperform the two other estimates. This suggests that we can search through the sample space to greatly improve our error and, hopefully, reduce it close to the rounding error of one degree.

The best result was acheived by the markov chain sampler, which first independently samples conformations for a protein, then chooses the best as a starting state, and iteratively changes one dihedral angle at a time, ultimately picking the conformation with the highest score in the sequence. Using this approach, the error was brought down to 5 percent in the best case.

5 References

Wei L, Zou Q. Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition. Int J Mol Sci. 2016;17(12):2118. Published 2016 Dec 16. doi:10.3390/ijms17122118

Wang J, Cao H, Zhang J, Qi Y. Computational Protein Design with Deep Learning Neural Networks. Scientific Reports. 2018;8(1). https://arxiv.org/pdf/1801.07130.pdf

Godzik A, Kolinski A, Skolnick J. Topology fingerprint approach to the inverse protein folding problem. Journal of Molecular Biology. 1992;227(1):227-238. doi:10.1016/0022-2836(92)90693-E.