Université Grenoble Alpes

PROTEIN STRUCTURE PREDICTION

Hanning Yang
Lynn Farhat
Niloufar Zarghampour

December 6, 2022
Computational Biology

Introduction

Our goal for this project was to predict a protein structure from its sequence. This specific protein carries the mutation responsible for high resistance to tetracyclin of a Salmonella variant, which have dealt with in our previous project.
Proteins differ from each other from their sequence of amino acids, which results in different structures and different functionalities. Nonetheless, it has been shown that proteins that have high similarities in their amino acid sequences actually have similarities in their structures and functionalities. Such similar sequences are called "homologous protein sequences" and their similarity is usually an indication that they share a common evolutionary origin.
Thus, based on that information, we sought to predict the structure of our given protein by doing comparisons to other similar sequences.

1 About TetR

To get some more information about our sequence, we uploaded to the Pfam database of protein families.
In fact, our protein belongs to the tetR-family of transcriptional regulators (TFTRs). TFTR protein family members are mostly transcriptional repressors, meaning that they prevent the expression of certain genes at the DNA level. These proteins can act on genes with various functions including antibiotic resistance, biosynthesis and metabolism, bacterial pathogenesis, and response to cell stress.
In our case, we know that our given protein carries the mutation responsible for high resistance to tetracyclin of a Salmonella variant. When tetracycline is present, it is bound by tetR, causing a conformational change such that TetR can no longer bind DNA.
Furthermore, our protein belongs to the set CL0123 in the Pfam database, which contains a diverse range of mostly DNA-binding domains that contain a helix-turn-helix motif, which is a major structural motif capable of binding DNA.

2 Implementation

2.1 Fetching Multiple Sequence Alignment

As previously mentioned, amino acid sequences having similarities actually implies similarities in the structure of the proteins that they correspond to. That is why we fetched a Multiple Sequence Alignment (MSA) from the Pfam database which gave us an exhaustive list of over 150,000 sequences with similar lengths. We selected 1,000 of those sequences at random to implement in our algorithm for computational purposes. Figure 1 shows the first 24 sequences aligned.

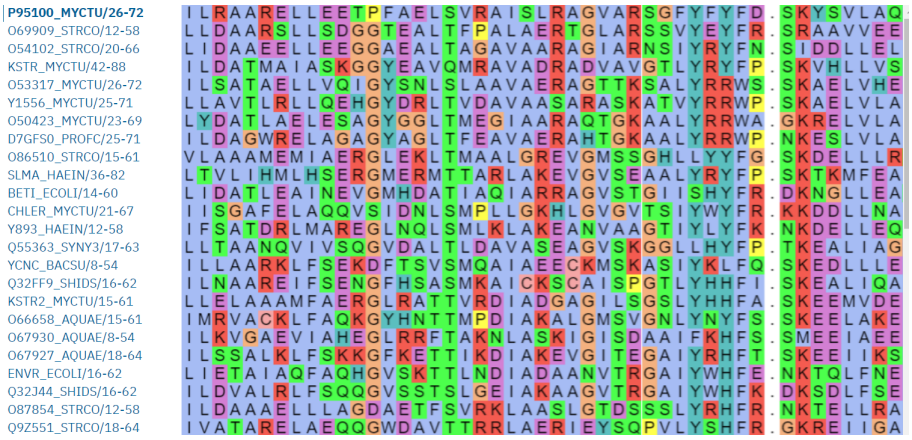


Figure 1: Multiple Sequence Alignments for TetR

The goal of MSA is to align a given set of sequences in a way that will either reflect their evolutionary, functional or structural relationship. This means that it will insert gaps between sequences in order to allow the homologous parts of the sequences to be aligned. An example of homologous parts is reflected in Figure 1 whereby they are generally the same color. The MSA is done column-wise.

2.2 Predicting Contact Map

After implementing the MSA, our next step was to predict the contact map of our protein. We did so by calculating the Mutual Information between two positions of different aligned sequences.

Mutual Information is a measurement of the uncertainty reduction for a MSA of homologous proteins. The MI between two positions (two columns in the MSA) reflects the extent to which knowing the amino acid at one position allows us to predict the amino acid identity at the other position.

The formula for mutual information between two positions i and j is given as:

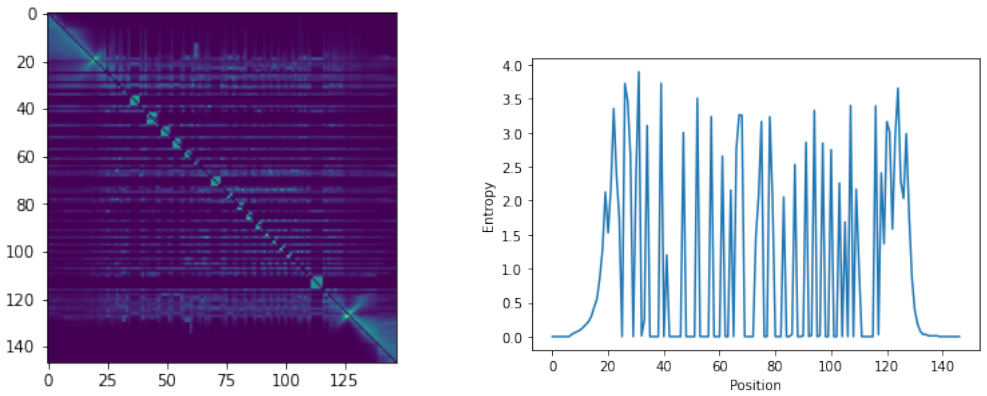
$$MI(i,j) = \sum_{a,b} P(a_i, b_i) log(\frac{P(a_i, b_i)}{P(a_i)P(b_i)}$$

where $P(a_i)$ is the frequency of amino acid a at position i $P(b_j)$ is the frequency of amino acid b at position j, and $P(a_i, b_i)$ is the frequency of both occurring simultaneously.

Mutual information is always nonnegative and achieves its maximum value if there is complete covariation. The minimum value of 0 is obtained either when i and j vary completely independently or when there is no variation. The mutual information (in bits) is 1 when two parties (statistically) share one bit of information. However, they can share a arbitrary large data. In particular, if they share 2 bits, then it is 2. If i and j are identical then all information conveyed by i is shared with j: knowing i reveals nothing new about j and vice versa, therefore the mutual information is the same as the information conveyed by i (or j) alone, namely the entropy of i.

The construction of the MSA is crucial and decisive to the calculation of Mutual Information

(MI). Collecting sequences and aligning them can be a demanding task. For optimal performance the MSA should be large and diverse. MSAs with $>$ 400 sequences (or 400 clusters of sequences) $<$ 62 $\%$ identical show good predictive performance values (AUC $>$ 0.75). These values are usually achieved by having 2000 sequences in the alignment. Thus, if a group of proteins in MSA from a relatively recent common ancestor, the identical proportion will be higher. Our prediction will be less accurate. We can correct that by randomly select sequences throughout the full MSA.
After calculating the mutual information matrix, we were able to select a certain threshold (refer to figure 3) in order to identify the contacts. If the MI entry is greater than the threshold, its contact matrix entry is set to 1; otherwise, it is set to 0.
Figure 2 shows the heatmap of the normalized mutual information next to the plot of the shannon entropy of each position of the MSA. Shannon's entropy is a quantitative measure of uncertainty in a data set.8u78



(a) Normalized MI Heatmap

(b) Entropy of MSA

Figure 2

2.3 Getting the Protein Structure

From the contact map, we used the given FT-COMAR tool which took as input our contact map and gave as output our predicted protien structure.

2.4 Optimizing the Threshold

The choice of distance threshold defines the number of contacts in a protein. At lower distance thresholds, a protein has fewer contacts. The procedure to retrieve the contacts relies heavily on the threshold, we implemented the optimization of the threshold in two ways :

- **Comparison to the MI plot**:
  The goal here, is to compare the predicted contact map with the true mutual information calculated among the different positions of the MSA. Starting with an initial guess for the threshold, then increasing/decreasing this value to get closer to the MI plot as much as possible. Refer to the normalized mutual information plot from Figure

2

Depicted below are the results for the predicted contact maps based on different thresholds.

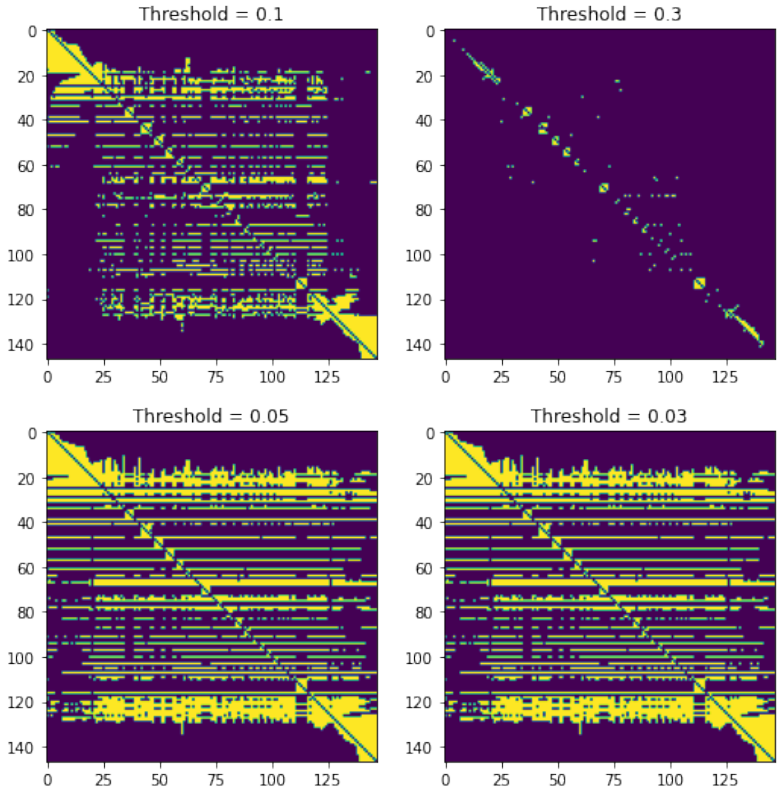


Figure 3: Predicted Contact Maps for Different Threshold

The drawback of the practice is that we are comparing these plots only visually, so it is very much possible to contain error, since the human eye is not a precise measure of comparison. We might get close but not precise. For example, comparing the different figures in 3 with the one in **??**, we can see the optimized threshold is some value between 0.03 and 0.05 but the difference is not visible, and we can not conclude one specific value for it based on this. The advantage of this method is that we can find an interval of values for the candidate thresholds to minimize the RMSD.

- **Minimizing the RMSD**

  For this method, we use the .cmap files based on different thresholds and use the FT-COMAR to compute the pdb files, and with the help of PyMol, visualize the structures and calculate the RMSD regarding the candidate thresholds. The most optimized model was achieved using the threshold **0.045** with an RMSD of 6.93. Here is a visualization of the known tetR structure and the predicted one, using a threshold of 0.045 :

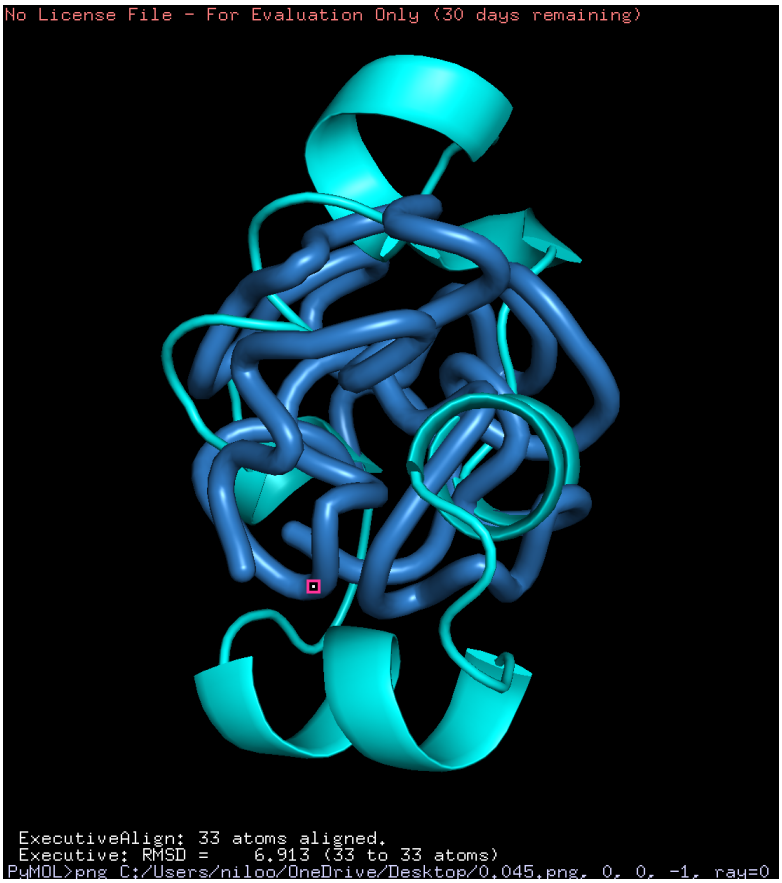


Figure 4: (Light blue) The given TetR structure vs. the predicted structure (dark blue)

The disadvantage in both of these methods is that the created structure is based on the threshold, which is being chosen by trial and error. However, in the last method, having RMSD as a measure of comparison is much more stable than the first method.

2.5 Learning the Weights

The basic Idea here is to calculate the frequency count $f_i(A)$ for a single MSA column i, characterizing the relative frequency of finding amino acid A in this column, and the frequency count $f_{ij}(A, B)$ for pairs of MSA columns i and j, characterizing the frequency that amino acids A and B coappear in the same protein sequence in MSA columns i and j. The raw statistical correlation mentioned, suffers from a sampling bias, resulting from phylogeny, multiple-strain sequencing, and a biased selection of sequenced species. The problem has been discussed extensively in the literature (Burger & Van Nimwegen, 2010) (Wollenberg & Atchley, 2000) (Tillier & Lui, 2003). Here, we implemented a simple sampling correction, by counting sequences with more than 80 percent identity and reweighting them in the frequency counts. All the frequency calculations and results are obtained using this sampling correction; the number of nonredundant sequences is measured as the effective sequence number M_{eff} after reweighting. In our code, we have a function called, "Compute-frequencies" which calculates the reweighted frequencies and then computes the contacts based on that. The calculations in this function are based on these formulas:

$$f_i(A) = \frac{1}{M_{\text{eff}}} \left(\frac{1}{q} + \sum_{a=1}^{M} \frac{1}{m^a} \delta_{A,A_i^a} \right)$$
$$f_{ij}(A,B) = \frac{1}{M_{\text{eff}}} \left(\frac{1}{q^2} + \sum_{a=1}^{M} \frac{1}{m^a} \delta_{A,A_i^a} \delta_{B,A_j^a} \right)$$

Where $\delta_{A,B}$ denotes the Kronecker symbol, which equals one if $A = B$, and zero otherwise. Furthermore, we have defined $q = 45$ for the number of different amino acids (also counting the gap). The factor $1/m^a$ aims at correcting for the sampling bias. It is determined by :

$$m^a = \left| \left\{ b \in \{1, \ldots, M\} \mid \text{seqid}\left(A^a, A^b\right) > 80\% \right\} \right|$$

Also, M_{eff} is the sum over all sequence weights after the sampling correction, meaning $M_{eff} = \sum \frac{1}{m^a}$. This allows us to update our Mutual Information using the reweighted frequencies as :

$$\text{MI}_{ij} = \sum_{A,B} f_{ij}(A,B) \ln \frac{f_{ij}(A,B)}{f_i(A) f_j(B)}$$

The idea used in this section is from the paper (Oliveira & Pedersen, 2007), which we tried to implement the method.

3 Detecting the Mutation

As previously mentioned, the tetR proteins contain a helix-turn-helix (HTH) motif that is the DNA-binding domain.
Mutations in a protein sequence may result in a change the original protein structure, which in turn may cause a change in its functionality.
In our case, tetR is known for its resistance to tetracycline. From our previous project, we were able to detect the mutation in the sequences by comparing a wild type sequence and a resistant one to tetracycline.

4 Conclusion

Eventually, the goal of our project was achieved by predicting the structure of our protein. We learned that the tetR protein has a mutation apparent in its structure and derived from its sequence. This mutation is a result of evolutionary causes or simply of errors in the copying of the DNA, and it is characterized by its resistance to tetracycline which is primarily found in Salmonella antibiotics.

References

Burger, L., & Van Nimwegen, E. (2010). Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology*, *6*(1), e1000633.

Oliveira, R. G., & Pedersen, A. G. (2007). Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms for molecular biology*, *2*(12).

Tillier, E. R., & Lui, T. W. (2003). Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, *19*(6), 750–755.

Wollenberg, K. R., & Atchley, W. R. (2000). Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proceedings of the National Academy of Sciences*, *97*(7), 3288–3291.