Computational Biology - Protein Structure Prediction

Alessandro Bonadeo, Giuseppe Intilla, Sofia Moroni

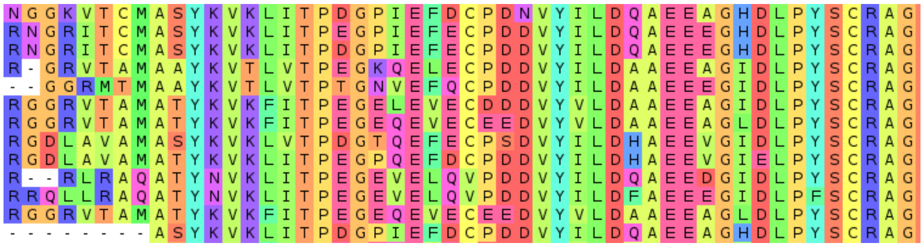
December 5, 2022

1 Introduction

In this project we attempt to predict and model the structure of the protein which, following from the previous project, carries the mutation responsible for being resistance to tetracycline in a Salmonella variant. This protein belong to the TetR family which are proteins playing an important role in conferring antibiotic resistance to large categories of bacterial species. Then, we compare the predicted protein structure to the actual one, which is included in the public Protein Data Bank dataset as a reference quantifying the difference between the data structure.

2 Multiple Sequence Alignment

We use the online tool Pfam, which is a great dataset of protein families, to retrieve the multiple sequence alignment (MSA) corresponding to the given protein sequence. By mean of this comparison one can hope to find a common function between the gene considered and its family of homologous. We can learn about the conserved amino acids at specific sites as well as the relationships between two or more positions MSA of homologous proteins. The number of sequence generated from the alignment is around 13000 which should be enough to proceed with a sufficient accuracy.
We decide to filter the extracted sequences based on the frequency of missing values on the columns: we chose to keep only the columns presenting a number of non missing element equal to half the total number of sequences. In this way we reduce the number of columns from 100 to 46. This final alignment gives us information about the structure of the protein: indeed, if one amino-acid changes, the amino-acid in contact with it must compensate for the change. For instance, if two amino acids, X and Y, have positive and negative charges, respectively, and X changes to a positively charged amino acid, Y must also change to a positively charged amino acid. One can determine the potential connections by aligning several related protein sequences and computing the mutual information between the alignment's positions.



```
N G G K V T C M A S Y K V K L I T P D G P I E F D C P D N V Y I L D Q A E E A G H D L P Y S C R A G
R N G R I T C M A S Y K V K L I T P E G P I E F E C P D D V Y I L D Q A E E E G H D L P Y S C R A G
R N G R I T C M A S Y K V K L I T P D G P I E F E C P D D V Y I L D Q A E E E G H D L P Y S C R A G
R - G R V T A M A A Y K V T L V T P E G K Q E L E C P D D V Y I L D A A E E A G I D L P Y S C R A G
- - G G R M T M A A Y K V T L V T P T G N V E F Q C P D D V Y I L D A A E E E G I D L P Y S C R A G
R G G R V T A M A T Y K V K F I T P E G E L E V E C D D D V Y V L D A A E E A G I D L P Y S C R A G
R G G R V T A M A T Y K V K F I T P E G E Q E V E C E E D V Y V L D A A E E A G L D L P Y S C R A G
R G D L A V A M A S Y K V K L V T P D G T Q E F E C P S D V Y I L D H A E E V G I D L P Y S C R A G
R G D L A V A M A T Y K V K L I T P E G P Q E F D C P D D V Y I L D H A E E V G I E L P Y S C R A G
R - - R L R A Q A T Y N V K L I T P E G E V E L Q V P D D V Y I L D Q A E E D G I D L P Y S C R A G
R R Q L L R A Q A T Y N V K L I T P E G E V E L Q V P D D V Y I L D F A E E E G I D L P F S C R A G
R G G R V T A M A T Y K V K F I T P E G E Q E V E C E E D V Y V L D A A E E A G L D L P Y S C R A G
- - - - - - - - A S Y K V K L I T P D G P I E F D C P D D V Y I L D Q A E E A G H D L P Y S C R A G
```

Figure 1: Example of MSA

3 Mutual Information

The mutual coevolutionary relationship between two positions in a protein family can be estimated using the Mutual Information concept. Mutual Information theory is frequently used to predict the correlations between positions in an MSA that are structurally or functionally significant in a certain fold or protein family. The information found in MSAs may be used to predict residue pairs that are likely to be close to each other in the three-dimensional structure, because the evolutionary mutations in the sequences are constrained by a number of factors, such as the preservation of favorable interactions in direct residue-residue contacts.

So by aligning multiple related protein sequences and computing the mutual information between the position of the alignment, it is possible to quantify how much two columns are correlated and infer the possible contacts. For each possible combination of columns *(i,j)*, we compute the Mutual Information as the sum over each possible pair of amino-acids *(a,b)* of the joint probability to find amino-acid a at position x_i and amino-acid b at position x_j times the logarithm of the ratio between the joint probability and the product of the marginal ones.

$$MI(i,j) = \sum_{a,b} p\left(x_i = a, x_j = b\right) \log \frac{p\left(x_i = a, x_j = b\right)}{p\left(x_i = a\right) p\left(x_j = b\right)},$$

with

$$p\left(x_i = a\right) = \frac{\#\text{sequences having amino acid } \mathbf{a} \text{ in position } \mathbf{i}}{N},$$
$$p\left(x_i = a, x_j = b\right) = \frac{\#\text{sequences having } \mathbf{a} \text{ in } \mathbf{i} \text{ and } \mathbf{b} \text{ in } \mathbf{j}}{N}.$$

If a position is fully conserved, it means that does not carry any additional information about mutations in another position and the corresponding Mutual information will be close to 0. If two positions i and j are independent, then the Mutual Information will be exactly 0 since $p(x_i{=}a,\ x_j{=}b){=}p(x_i{=}a)p(x_j{=}b)$ and the logarithm becomes equal to 0. The sequences that co-evolve together provide the greatest amount of mutual information, ensuring that for every mutation in position i the equivalent mutation occurs in position j: the maximum value of the mutual information is obtained when the joint probability assumes the highest possible value.

4 Algorithm

In order to estimate the mutual information matrix we write a function that iteratively updates the value for each entry. In particular the first step is calculating the marginal probabilities for each amino acid and each position. Iterating through the columns of our MSA we calculate the marginal probability for every amino acid and then use them to calculate the joint probabilities between each couple of positions. Although the algorithm involves different loops it is fast enough. In fact we only need to calculate the upper tridiagonal matrix since the mutual information matrix is symmetric.
After that we try to optimize the performance of the prediction by applying different adjustments to the matrix:

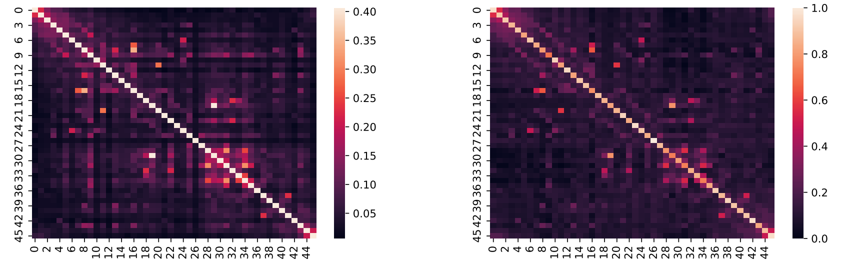
- first of all we put on the diagonal the maximum value of the matrix;
- then we correct every value of the matrix in the following way

$$M(i,j) = M(i,j) - \frac{1}{N}\sum_k (MI(k,j) + MI(i,k));$$

- at last we normalize the matrix as

$$M(i,j) = \frac{M(i,j) - min(M)}{max(M) - min(M)}.$$

In figure 2 we can see the effects of the corrections applied before by visualizing the heatmaps.



(a) Heatmap of the original MI matrix (b) Heatmap of the corrected MI matrix

Figure 2: Heatmaps of the MI matrix

The next step of our algorithm is the creation of the contact map. Here the important parameter τ, i.e the threshold, is involved since the contact map is given by the following rule

$$CM(i,j) = \begin{cases} 1 & \text{if } MI(i,j) >= \tau; \\ 0 & otherwise. \end{cases}$$

Here we report the visualization of a contact map generated by using a fixed threshold.

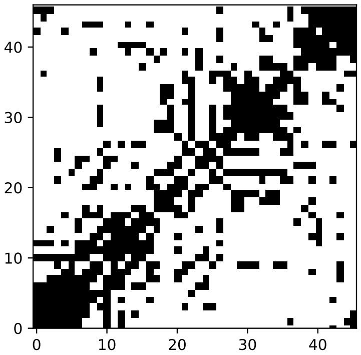


Figure 3: Contact map using $\tau = 0.13$

In order to choose an efficient value for the threshold we decide to write different contact maps varying τ and then select the one that gives us the better value of RMSD. In particular it is defined as

$$RMSD = \sqrt{\frac{1}{N}\sum_i \delta_i^2},$$

where δ_i is the distance between atom i and the reference structure, and it's used as a performance metrics for our prediction. We decide to test the algorithm for 20 values of τ in the range $[0.1, 0.3)$, running the FT-COMAR software and using its output with Pymol to visualize the predicted structure and align it with the reference protein. Now we show a plot to visualize the variation of the RMSD with respect to the different threshold.

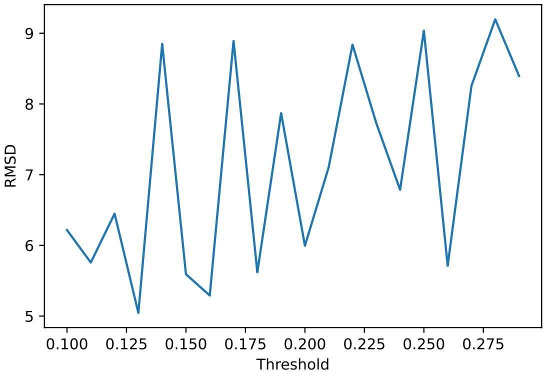


Figure 4: Values of RMSD obtained by changing the threshold

We can see the minimum value at RMSD $= 5$ for $\tau = 0.13$ that we pick as the best threshold. Now we can use this value to obtain the best prediction and visualize the alignment to the original protein.

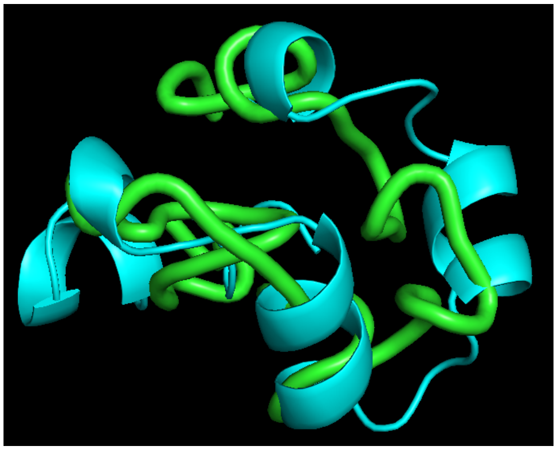


Figure 5: Alignment between the prediction and the reference structure

5 Conclusion

We can state that our algorithm is a quite good tool to predict a molecular structure. Although, it is not error-free since a lot of approximations were assumed. For example we didn't take into account the internal composition of the molecule, which contains a huge quantity of information.