ASSIGNMENT REPORT

Statistical analysis of protein families

Agnese Ilaria Curatolo - Christopher Joseph Luna - Riccardo Giuseppe Margiotta

Introduction The aim of the exercise is to analyse three datasets of protein domains given by a multiple-sequence alignment (MSA). We have performed this analysis in three steps:

1. Finding conserved positions in the sequences, i.e. positions where there is a high probability of finding one amino acid, and small probabilities for the others.
2. Evaluation of new sequences: the comparison is made between a dataset of known sequences and a dataset of unknown sequences by the method of the log-odds ratio, that assumes no correlation between different positions in the sequences.
3. Detection of amino acid co-variance between different MSA columns, and relation to the 3D protein structure.

1 Conserved positions

The three files contain 1000 sequences of 53 amino acids. They are expressed by 20 relevant letters (“A”, “C”, ..., “Y”) and few unrelevant letters (“B”, “X”, “-”, ect.). In order to analyse the data, we have converted the relevant letters into numbers from 1 to 20 and the unrelevant letters into the number 21. We have then used the data of the file “train.faa” that contains the known sequences to construct our statistics.

After computing the probability $\omega_\ell(a)$ of the occurrence of an amino acid a in position ℓ of the sequences, we select the maximum $\omega_\ell(a^*) = \omega_{max}$ for each position and plot it as a function of ℓ (see fig. 1.1 *(Left)*).

The positions where $\omega_{max} > 0.5$ are considered to be *conserved*, i.e. the same amino acid is present along more than 50% of the sequences.

2 Family belonging

In the previous part we defined the PSWM matrix $\omega_\ell(a)$ which characterizes the family of sequences of the dataset “train.faa”. The aim of this part is now to determine which dataset between “test1.faa” and “test2.faa” belongs to the same family of “train.faa”. In order to do this we use the method of the log-odds ratio, based on the assumption that amino acids in different positions have no correlations between each other (see fig 1.1 *(Right)*). When the log-odds ratios of one set of sequence are positive (negative), it does (does not) belong to the same family of the set in “train.faa”. It is clear that the file that contains sequences that belong to the same family of the ones in “train.faa” is “test2.faa”. What we observe is that the histogram on the left (belonging to another family) is very peaked and symmetric, which means that the sequences are coherent within the same set. On the other hand, the histogram on the right is less peaked and has a longer tail going towards zero: it means that the sequences in that set belong to the same family but are more different between each other.

3 Co-evolution of contact residues

Co-evolution between two positions, which are in contact in the folded 3D structure, induces correlations in the amino acid occurrences in these positions. The aim of the last part is to detect these correlations and check if they correspond to a contact in the 3D structure by comparing them with the values in the file “distances.txt”.

In the figure 3.1 the two colors correspond to the different datasets that were investigated. From the graphs we can see that for the first 2^2 pairs the fraction of true-positives is 1, which can be expected since the mutual information of these pairs is the highest. Beyond 2^5, the trend in the fraction of true-positives is strictly decereasing, which corresponds well with the fact that these pairs are less correlated as their rank decereases. What we do not expect is the dip around 2^4 number of pairs, which we can attribute to the presence of false-positives. This is comfirmed by the fact that in both datasets we see that the fraction slightly recovers at 2^5.

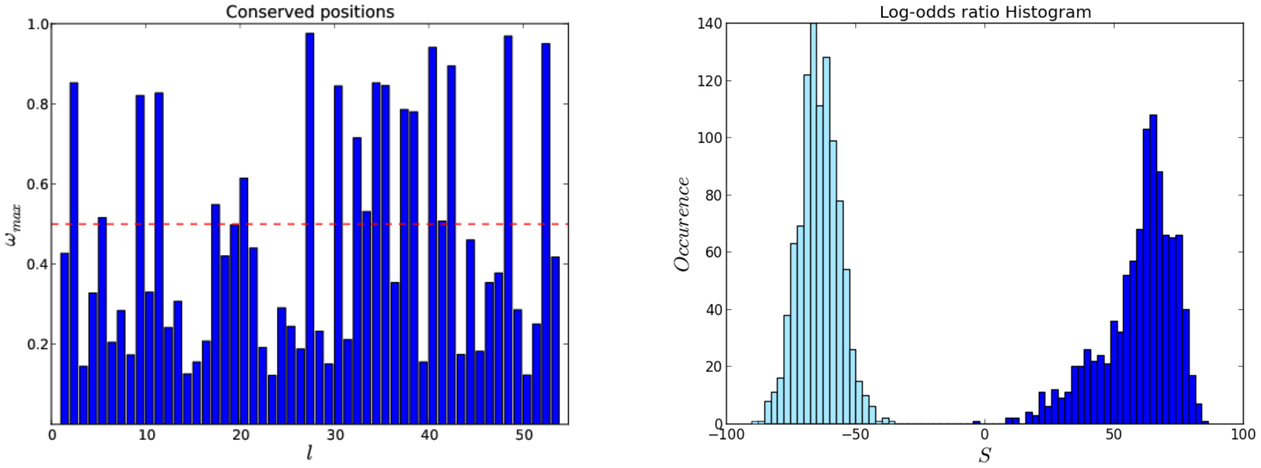

Figure 1.1: (Left)The dotted red line corresponds to $\omega = 0.5$. The positions whose bars are above 0.5 are the conserved positions. (Right) The histogram on the left shows the log-odds ratios for the dataset in the file "test1.faa", while the one on the right corresponds to the log-odds ratios of the dataset in the file "test2.faa".

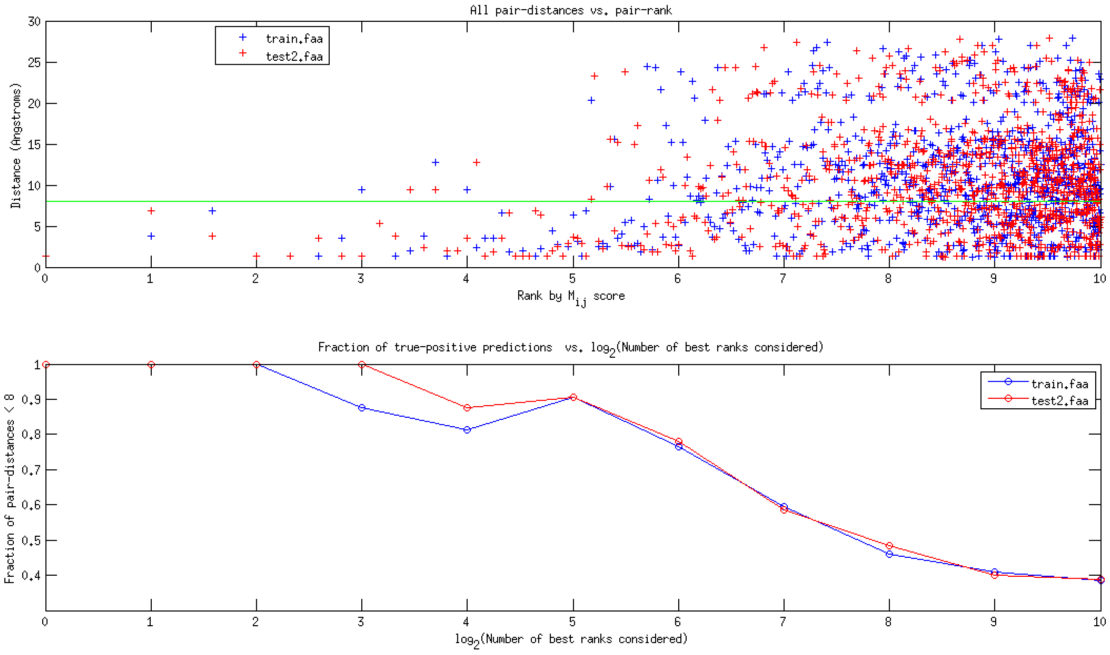

Figure 3.1: *(Top)* All of the distances are matched with their respective pairs and plotted by their rank depending on their score in M_{ij}. The green line denotes the boundary for distances less than 8 Angstroms. *(Bottom)* The fraction of pair-distances succesfully matched plotted against the number of pairs considered as increasing powers of 2.