GOR method for the prediction of protein secondary structure from sequence

Sahar Heidari

October 2022

Garnier, Osguthorpe, and Robson have proposed a method in 1978 that predicts protein secondary structure from amino acid sequences. This method has been continuously improved and modified, and the most crucial change in the algorithm was the inclusion of evolutionary information using sequence profiles obtained by converting multiple sequence alignments (MSA) to Position-Specific Substitution Matrix (PSSM). The method is based on information theory and the statistical propensities of residues for secondary structure conformations. The information function I calculates the probability of a residue R to be in the conformation S, such that:

$$I(S;R) = log\frac{P(S|R)}{P(S)}$$

Where R is one of 20 possible amino acids and S is one of three secondary structure classes: H, E or C. $P(S|R)$ is the conditional probability of observing residue R given conformation S, and P(S) is the probability of observing conformation S. Using Bayes's theorem we get:

$$I(S;R) = log\frac{P(R,S)}{P(S)P(R)}$$

We also need to take into consideration the influence of neighbors on the structural conformation of each residue R:

$$I(S,R_{-d},...,R_d) = log\frac{P(S|R_{-d},...,R_d)}{P(S)P(R_{-d},...,R_d)}$$

Where d is the number of residues taken into consideration immediately before and after the central residue. This is computationally expensive as it would be an exponential number of possible configurations and a very large sequence database to estimate reliable distributions. To overcome this problem, we can assume statistical independence of all residues within the window, so that:

$$I(S,R_{-d},...,R_d) \approx \sum_{k=-d}^{d} I(S;R_k)$$

The GOR method training phase consists of the computation of a matrix containing the information function for each position in the sliding window of profile positions around it, considering all kinds of secondary structure conformation and residue. This matrix will represent the secondary structure conformation in the first axis, the position along the sliding window in the second axis, and the residue type in the third axis. For the prediction phase, the information matrix and sequence profiles are used to predict the secondary structure of a test set of protein sequences. Each residue position of all the query sequences is analyzed, associating it with the conformation S characterized by the highest value in the information matrix:

$$S = argmax_s \sum_{k=-d}^{d} I(S; R_k)$$