Algorithms in Bioinformatics
(Exercise 6 Report)
CMPUT 606
Winter 2020
ccid: afia

The main objective of this exercise is to experiment with the secondary structure prediction of proteins utilizing their tertiary structure. In order to solve this exercise, we have collected 100 protein structure FASTA sequences and their corresponding (.pdb) file in a way such that the proteins are globular or water-soluble. The Protein Data Bank is used as the resource to collect all these sequences.

Each of the collected protein structure contains a pdb file which has the information about the helix and sheet positions for a particular sequence. These values are extracted with the help of a python program and were used as the ground truth for calculating the prediction accuracies. The code is attached to the exercise directory.

Additionally, we used three other secondary structure prediction programs such as Chou and Fasman Secondary Structure Prediction (CFSSP), Self-Optimized Prediction Method with Alignment (SOPMA), and PROTEUS. All these three programs are able to predict the three-state secondary structure (Helix (H), Sheet (E) and Coil (C)) for a protein given it's protein sequence in FASTA format. These results were collected and compiled in three different csv files appended in the exercise directory.

The prediction program CFSSP is based on the Chou-Fasman algorithm. This algorithm analyzes the relative frequencies of each amino acid in alpha helices, beta sheets, and turns. The frequency is calculated on the basis of X-ray crystallography of known protein structures. The program SOPMA predicts all the sequences of a set of aligned proteins for a given protein sequence. On the other hand, Proteus is a high-performing prediction program which is basically a concatenation of different expert prediction approaches such as, PSI-BLAST, PSI-PRED, JNET and homologous searches [3].

The protein secondary structure prediction accuracy is measured by Q_3 which is a ratio of the total number of correctly predicted residues and the number of residues. The prediction accuracy for CFSSP mentioned in their official paper is around 56-60% [1], SOPMA can correctly predict 69.5% of amino acids for a three-state of the secondary structure [2] , while for water-soluble protein, Proteus is able to

achieve an accuracy of 88% and above. Whenever a protein shows no similarity to any known structure, even at that point PROTEUS is able to achieve a Q3 score of 79%. The formula for Q_3 is mentioned below:

$$Q_3 = \frac{\text{correctly predicted residues}}{\text{number of residues}}$$

At this point, we check the number of correctly predicted residue, using the ground truth and the predicted structure. For each prediction program, we calculate Q3 for each sequence. All such q3 values are then averaged to get the overall prediction accuracy for a program. Table 1 shows the overall prediction accuracies of the three utilized prediction programs.

Table 1: Overall prediction accuracy of three prediction programs.

Programs	Q_3 Accuracy
CFSSP	55.214
SOPMA	68.637
PROTEUS	87.357

We further calculate the true positives (TP) , false positives (FP), true negatives (TN), and false negatives (FN) for each of the prediction program depicting all the three states of the secondary structure in a single table. The confusion matrices are shown below:

17256 (TP)	4102 (FP)
7409 (FN)	2533 (TN)

Fig. 1: Confusion matrix for CFSSP

25793 (TP)	1564 (FP)
2624 (FN)	1522 (TN)

Fig. 2: Confusion matrix for SOPMA

30541 (TP)	1239 (FP)
2166 (FN)	855 (TN)

Fig. 3: Confusion matrix for Proteus

From the above experiments and measures (Table1), it is found that the prediction programs achieves the claimed accuracy scores. Analyzing the prediction result, we can say that, CFSSP performs poorly in predicting the Helices, while it does a good job in predicting the Sheets. SOPMA performs satisfactorily on Helix prediction but performs poorly for Sheets prediction, while, Proteus seems to perform well on predicting sheet and coil states

References:

1. Ashok Kumar, T. (2013). CFSSP: Chou and Fasman Secondary Structure Prediction server. WIDE SPECTRUM: Research Journal. 1(9):15-19.
2. Geourjon C, Deleage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Bioinformatics. 1995 Dec 1;11(6):681-4.
3. Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC bioinformatics. 2006 Dec;7(1):301.