Principal Component Analysis and Clustering Analysis of Motif Vectors

The Final Report for the Senior Design Project: Automated Annotation of Amino Acid Sequences Using Hierarchical Motif Vectors

Student Name, Number: Mert Ekren, 200206014

Objective

To finalize the project, principal component analysis is applied to process the obtained motif vectors faster. Then, clustering analysis applied for annotation of clustered motif vectors.

The progress

To start, the set of extracted proteins has been altered from all-human proteins to functioning as DNA-binding in human proteins in the database UniProt Knowledge- base/ Swiss-Prot. Same previous procedure is followed to find the motif vectors of proteins. Information about the set of protein will be given in extras.

In the third phase of the project, Principal Component Analysis has been applied in order to process data faster while reducing feature dimension without much information loss. PCA seeks to find uncorrelated components to represent large data using eigenvalue decomposition. First, the covariance matrix of the feature matrix that is subjected to wavelet decomposition is constructed. Explicitly, each feature's covariance is calculated with respect to all 955 features. Here, covariance matrix is constructed with finding covariances of each protein and weighing by the length of sequence of proteins. Resulting information is a matrix 955 by 955. Secondly, the eigenvalue decomposition is applied to covariance matrix. Largest eigenvalues are responsible for most principal eigenvectors. Eigenvectors are selected so that the corresponding largest eigenvalues sums up to at least %95 of the total sum of eigenvalues of the matrix. 17 components are found to be sufficient for representing 955 features with loss near only %5.

After PCA, aa sequences of human proteins including DNA-binding regions are ready to be clustered. In start, a peculiar approach is followed due to the enormity of the motif vectors of total 335913 amino acid. k-means

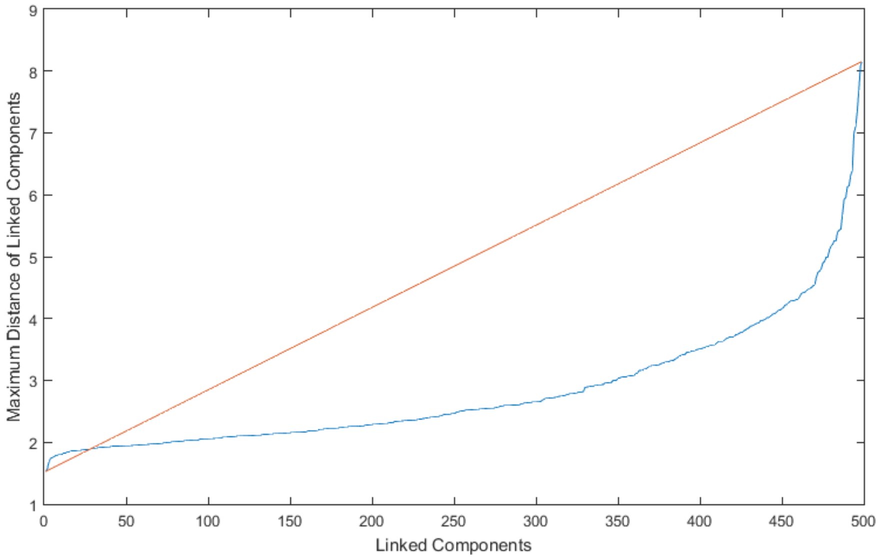

Figure 1: Elbow Method for Optimal Number of Clusters (k=500)

clustering is applied to motif vectors in order to reduce data for further analysis of hierarchical clustering. k-means clustering is a popular algorithm aims to partition data into 'k' number of groups based on finding nearest centroid to each data, and recalculation of k number of centroids. Here, the idea is to obtain a reduced size of cluster so that we can hierarchically link data further based on k-means centroids. Elbow method is followed for determining the number of cluster for linked centroids, a similar technique we applied in constructing motif vectors. This procedure shown in Figure 1.

After obtaining clusters of aa sequences for all proteins, a match is sought which is in between the DNA-binding regions in sequences and whether these regions are clustered in our data. InterPro and UniProt/Swiss-Prot are the two online databases we referenced to check the matches.