Wavelet Decomposition for Separation of Variations in Features

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

Student Name, Number: Mert Ekren, 200206014

Objective

The vector of properties that can numerically represent naturally occurring amino acids are put to aa sequences from human proteins. Then, wavelet decomposition is applied to the proteins to extract the information of variations for multiple scales in properties.

The progress

Firstly, distance metric to the hierarchical clustering of amino acid representing features has been altered as the maximum distance. Even though this method resulted as increase in the number of clusters from 131 to 191, each cluster have become more compact in terms of similarity among features in a cluster. Exact previous approach is continued for finding representative feature of a cluster, and thus resulted in different set of features for feature vectors. Information about which features are clustered, and which feature is used for the vector is provided in extras.

Second, the sequence data were downloaded from UniProt Knowledgebase/ Swiss-Prot that is a hub for collecting functional knowledge about proteins with annotations obtained from literature and curator-evaluated analysis. Moving on, human-related proteins are obtained which is the interest for this project. Set of 9606 proteins with sequences and annotations are parsed to the MATLAB environment. Seven proteins which containing amino acid that has not a representation among the 20 letters of naturally occurring amino acids are excluded from the data sequence. Plus, short and long sequences are also excluded. This is determined by finding the mean of the logarithm base 2 of the 9599 proteins's lengths and trying to obtain a distribution similar to log normal distribution. Short limit for the length is 32. Then, the longest limit was determined by obtaining symmetry in distance between shortest and longest limits with respect to calculated mean.

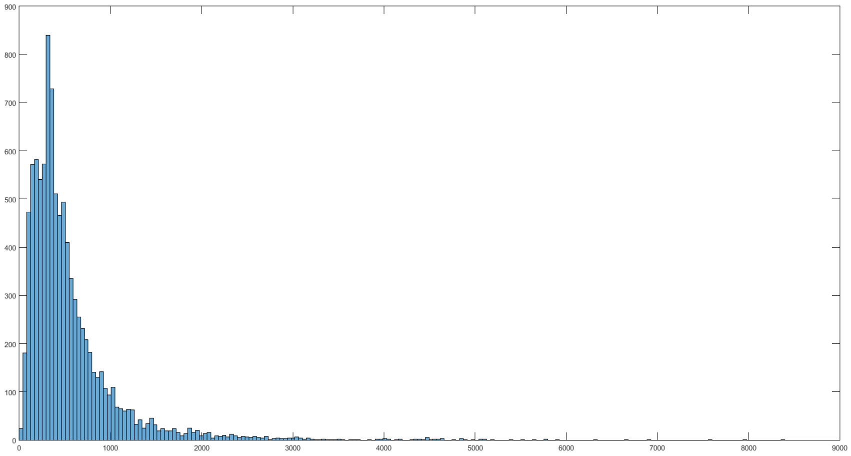

Figure 1: Distributions of Sequence Lengths

The longest limit is determined as 5149 and thus remains 9569 proteins . The original distributions of sequences's lengths is as shown above which has the mean length 557.

Thirdly, wavelet decomposition was subjected to data sequence. It is applied to amino acid sequences for each 191 representative numerical features for all 9600 proteins. Discrete wavelet decomposition is an efficient way to separate the variations in a property (Figure 1). Also, decomposing the signal into 4 multiple levels demonstrates variations on ranges that exponentially varies that is fundamental reason for hierarchical motif vectors. Daubechies 4 is used which is a popular wavelet with quadrature mirror filter. Also,'wavedec' function in MATLAB successfully outputs wavelet coefficients. Also, the function gives bookkeeping vector, which stores the length for corresponding levels. This is necessary because function does not seek circular symmetry which results in varying lengths of sequences at each level.

After the function separated the approximation coefficient(level-4 approximation) and details coefficients(level-4,level-3, level-2 and level-1 details), the approximation signals(level-4, level-3, level-2, level-1 and level-0 approximation sequences) are reconstructed and calculated. Since the linearity holds, it's easy to move to lower approximation sequences as long as the detail sequences is present for corresponding level. In addition, level-0 approximation sequence is the original sequence that is subjected to decomposition. Five approximation sequences are stored in order to further manipulate variations in the sequences.

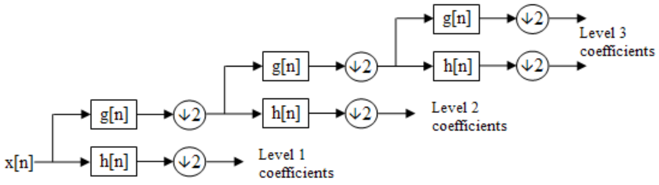

Figure 2: 3-Level Discrete Wavelet Decomposition Implementation