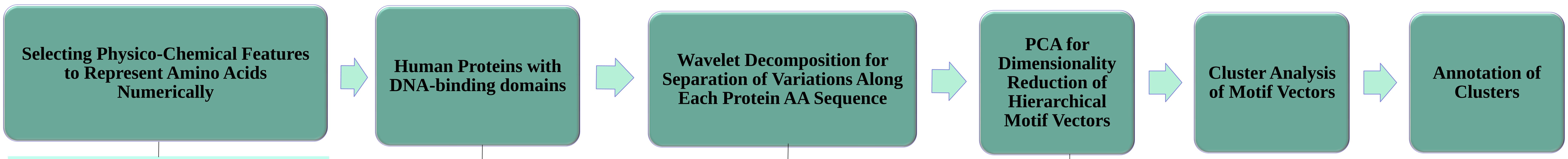
Annotation of Amino Acid Sequences Using Hierarchical Motif Vectors

Mert Ekren
Advisor : Prof. Dr. Bilge Karaçalı
Izmir Institute of Technology - Department of Electronics & Communication Engineering


Selecting Physico-Chemical Features to Represent Amino Acids Numerically
Human Proteins with DNA-binding domains
Wavelet Decomposition for Separation of Variations Along Each Protein AA Sequence
PCA for Dimensionality Reduction of Hierarchical Motif Vectors
Cluster Analysis of Motif Vectors
Annotation of Clusters

20 Naturally Occuring Amino Acids

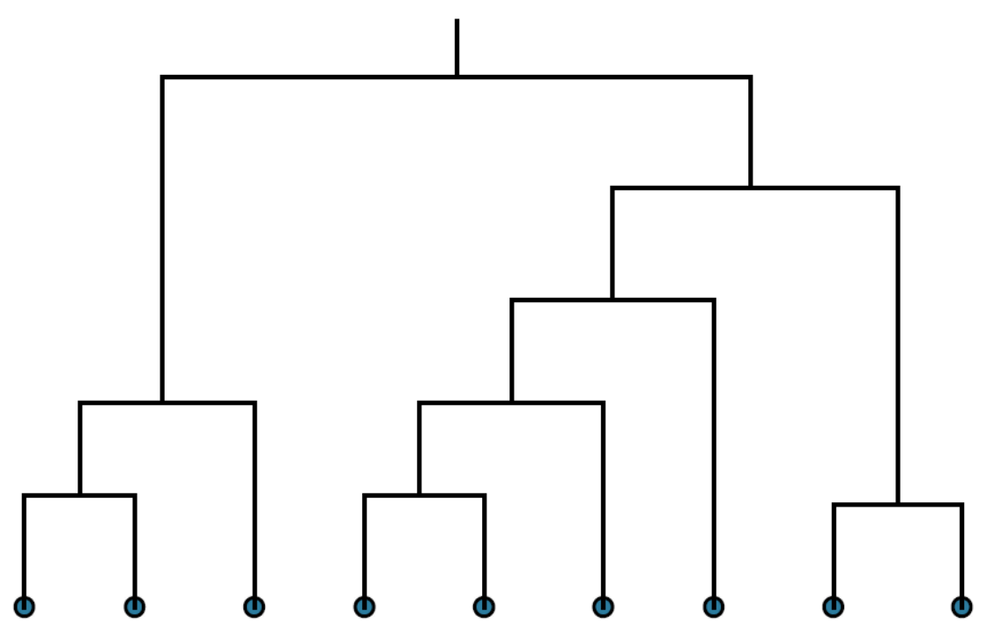
Physico-Chemical Features	θ_1^A	θ_1^R	…	θ_1^Y	θ_1^V
	θ_2^A	θ_2^R	…	θ_2^Y	θ_2^V
	⋮	⋮	…	⋮	⋮
	θ_{190}^A	θ_{190}^R	…	θ_{190}^Y	θ_{190}^V
	θ_{191}^A	θ_{191}^R	…	θ_{191}^Y	θ_{191}^V

Sequence:’AMGWCNR’

Motif Vector :
‘θ^A θ^M θ^G θ^W θ^C θ^A θ^N θ^R’

	Level	Sequences
Wavelet Approximation Sequences	0	‘θ^A θ^M θ^G θ^W θ^C θ^A θ^N θ^R’
	1	‘$\theta^A(1)\theta^M(1)\theta^G(1)\theta^W(1)\theta^C(1)\theta^A(1)\theta^N(1)\theta^R(1)$’
	2	‘$\theta^A(2)\theta^M(2)\theta^G(2)\theta^W(2)\theta^C(2)\theta^A(2)\theta^N(2)\theta^R(2)$’
	3	‘$\theta^A(3)\theta^M(3)\theta^G(3)\theta^W(3)\theta^C(3)\theta^A(3)\theta^N(3)\theta^R(3)$’
	4	‘$\theta^A(4)\theta^M(4)\theta^G(4)\theta^W(4)\theta^C(4)\theta^A(4)\theta^N(4)\theta^R(4)$’

	Dimension	
Raw Data	335913	955
PCA Data	335913	57



Abstract

Hierarchical motif vectors can numerically represent the organization of physico-chemical properties along amino acid sequences in proteins. In a family group of proteins, there are common functional regions along sequences. Clustering analysis is useful for grouping set of hierarchical motif vectors that are nearby. This project aims to accomplish identifying clusters of hierarchical motif vectors from the DNA-binding human proteins and labeling the responsible clusters with regards to the obtained motif vectors.

Procedure

1. Selection of Physico-chemical Features for Numerical Amino Acid Reprentation

Firstly, the project has been started from researching the physico-chemical properties that can numerically describe each 20 naturally occurring amino acids(aa). GenomeNet(1) provides 566 indices of different physico-chemical and biological features of aa under the database "AAindex: Amino Acid Index Database". This data is parsed to the MATLAB environment with extracting 13 indices which lack of not having numerical data for each of 20 aa.

Moving on, similarity has been sought among features in order to obtain more compact motif vectors. The correlation coefficients are implemented for the similarity parameter. A matrix of correlation coefficients among numerical properties is constructed which shows the correlation of features ith and jth. The color-mapped figure of correlation coefficients is shown in Figure 1. Noting that, yellowish colors represent the high correlations whereas blueish colors signify the low correlations among the matrix.

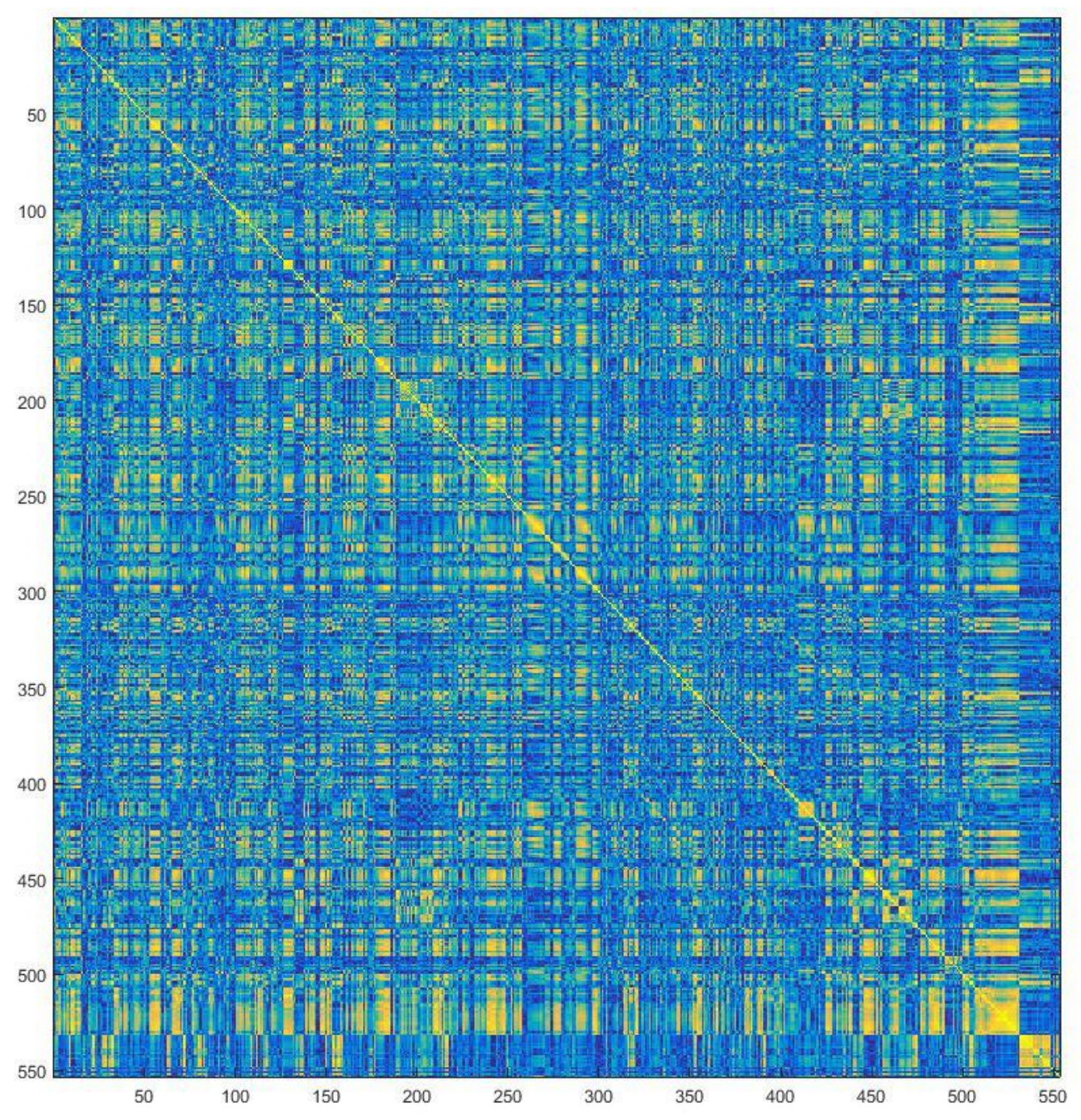


Figure 1. Colored figure of Correlation Coefficients of Features

Then, hierarchical clustering is operated to cluster similar physico-chemical features for condensed motifs. In this manner, hierarchical clustering is applied. After linking 533 features into clusters, the similarity between clustered features has been grouped with elbow method. Then, for each cluster, a feature is sought to represent the its corresponding cluster. This problem is handled by similar to finding the centroid feature of each cluster. Correlation coefficients are implemented and for each feature in a cluster, the minimum correlation is found. Then, finding the maximum among this values has revealed the best possible feature for representing that cluster.

2. Wavelet Decomposition for Separation of Variations in Features

Second, the sequence data is downloaded from UniProt Knowledgebase/ Swiss-Prot (2) that is a hub for collecting functional knowledge about proteins with annotations obtained from literature and curator-evaluated analysis. Moving on, human-related with DNA-binding domain proteins are obtained which is the interest for this project. Set of 606 proteins with sequences and annotations are parsed to the MATLAB environment. Obtained motif vectors replaced the aa sequences for each protein.

Then, wavelet decomposition was subjected to data sequence. It is applied to amino acid sequences for each 191 representative numerical features for all 606 proteins. Discrete wavelet decomposition is an efficient way to separate the variations in a property. Also, decomposing the signal into 4 multiple levels demonstrates variations on ranges that exponentially varies that is the fundamental reasoning of hierarchical motif vectors. Daubechies 4 is used which is a popular wavelet with quadrature mirror filter. After the sequences are subjected to transform, 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 shows an example of wavelet approximation signals for a selected feature.

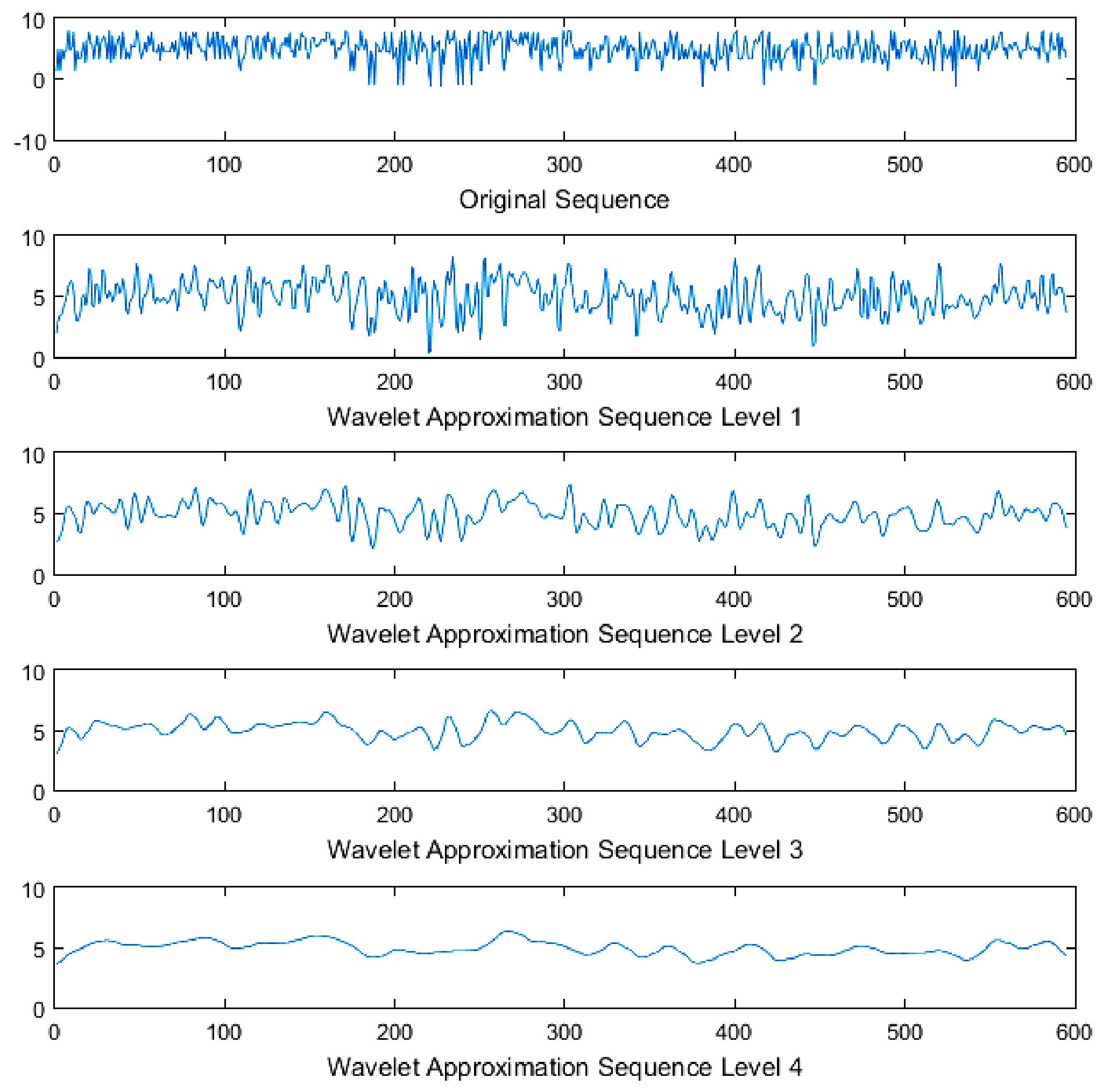


Figure 2. Wavelet Approximation Signal for Protein Estrogen receptor for selected feature ‘Side chain interaction parameter’ (Krigbaum-Rubin, 1971)

3. Principal Component Analysis on Hierarchical Motif Vectors

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. 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. 57 components are found to be sufficient for representing 955 features with loss near only %5.

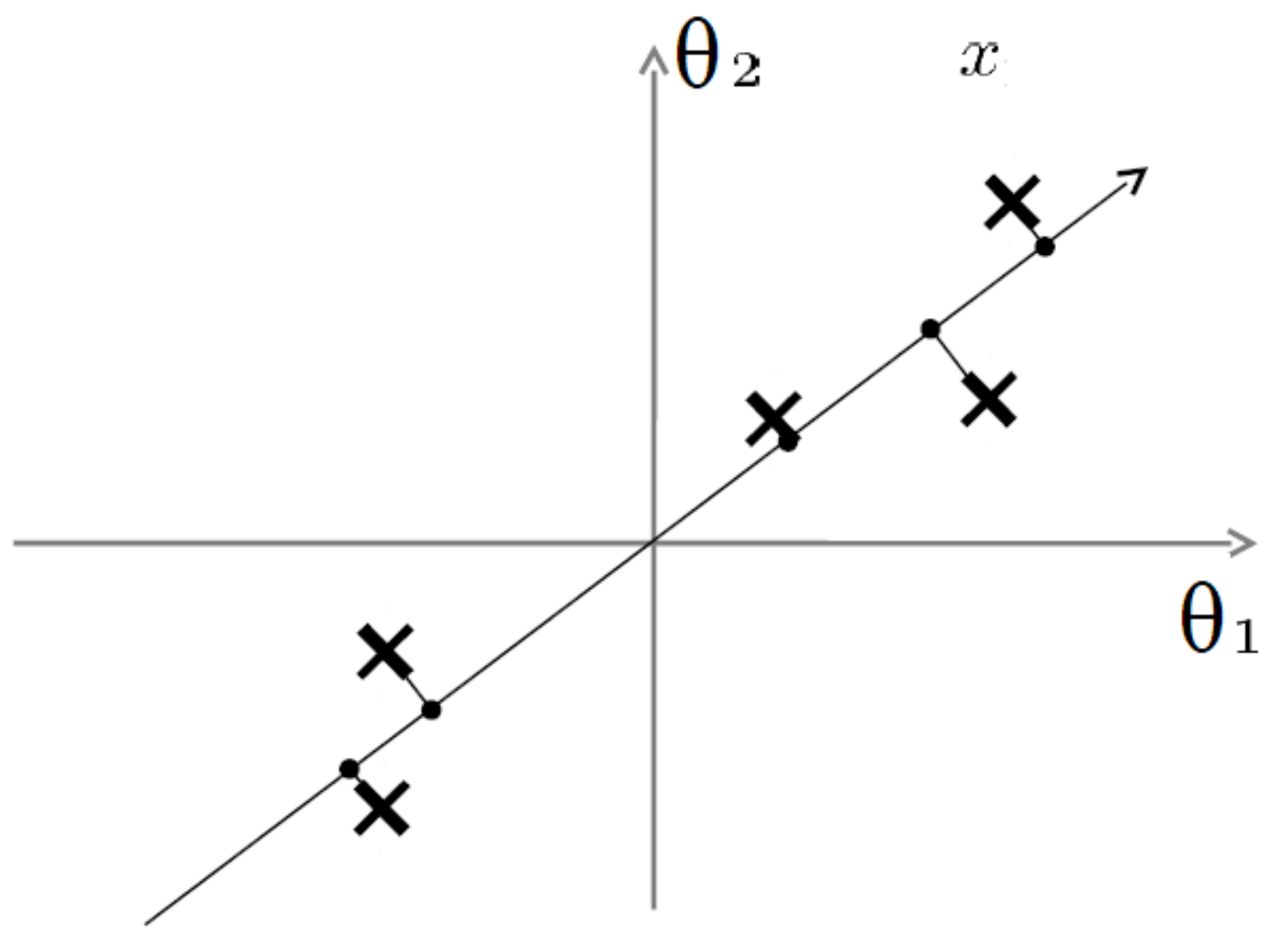


Figure 3. Principal Component Analysis on a data with 2 features

4. Cluster Analysis

After PCA, the data is normalized by dividing standard deviations along features in order to avoid any dominant feautre effects. Then, amino acid 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 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.

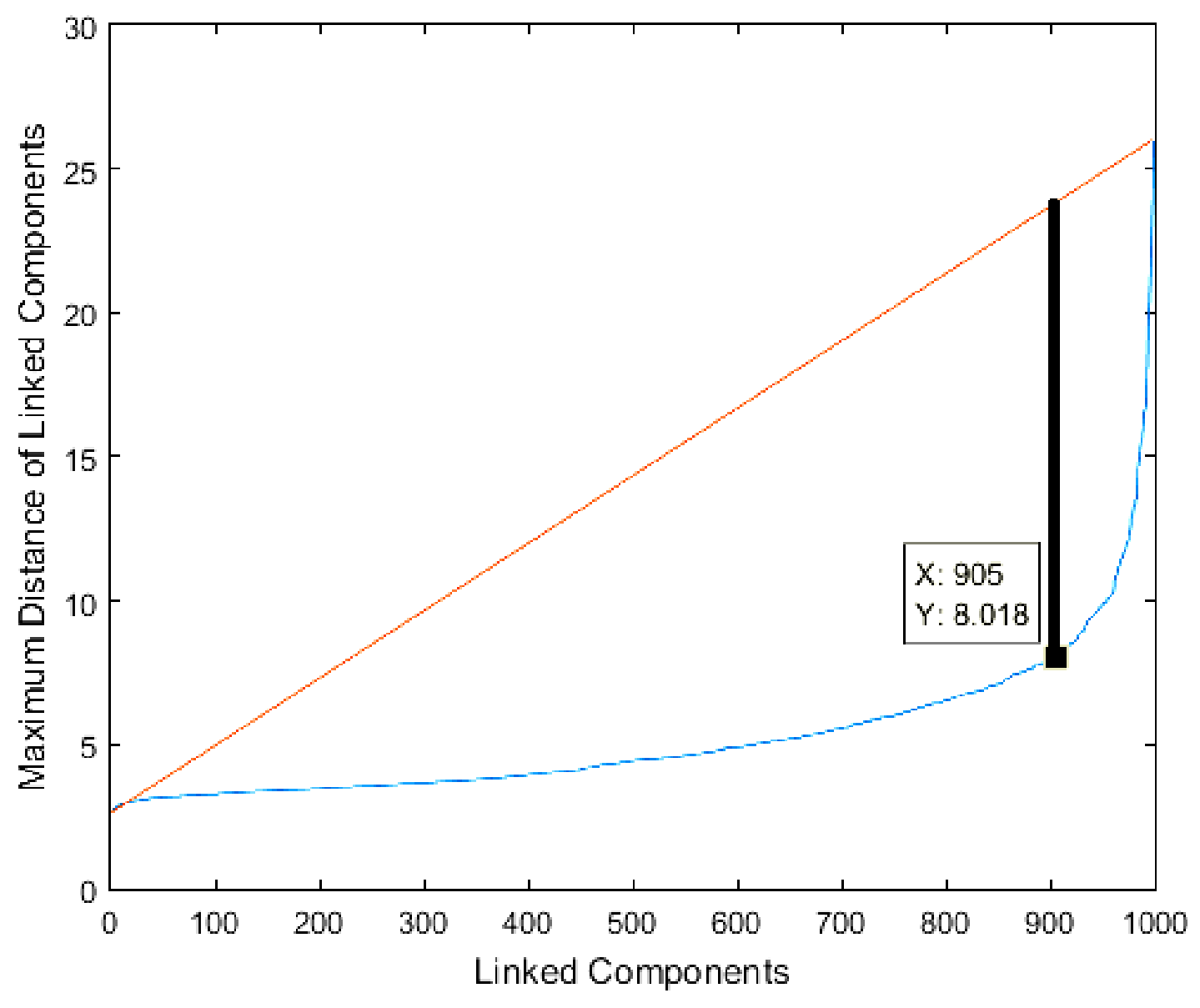


Figure 5. Demonstration of Elbow Method for Hierarchical Clustering

Results

- We have 191 physico-chemical properties that can represent each amino acid numerically.
- AA sequences of human proteins including DNA-binding regions are reconstructed with motif vectors. Our raw data includes 606 proteins with total 335913 aa.
- AA sequences of proteins are represented with 5 approximation signal for each select features, thus leaving us with 955 features.
- PCA reduced the heavy work of raw data by %94 without much information loss. The PCA matrix can be processed much more faster.
- Cluster analysis helped us with grouping segments of proteins. So, we compared the database of experimentally found domains of proteins with the clusters. Here, databases are InterPro (3) and UniProt.

Elbow method suggests to cluster the linkage at 95, but it may not show the optimum answer. The observations at protein’s amino acid clusters occurs in different groups consecutively. This may signify the excessive number of different clusters.

- Research is continued with lower number of clusters. The number 35 found to be sufficient, due to lower number of clustering tend to cluster excessive amount of data in one group.
- Looking the cluster distribution of receptor proteins such as *ERR1_HUMAN*, *ERR2_HUMAN*, *GCR_HUMAN*, *ANDR_HUMAN*, *RARA_HUMAN*, *NR1H3_HUMA*, *VDR_HUMAN*, *MCR_HUMAN* and *RORA_HUMAN* show that they possess cluster 27 with probability 0.38, 0.33, 0.32, 0.39, 0.35, 0.40, 0.34, 0.38 and 0.38 in their DNA-binding domain, even though this cluster occurs at 0.04 rate in total of amino acid sequences. This shows that it’s 10 times likely to occur in DNA-binding segments of these proteins.
- Also, some of the proteins such as BARH1_HUMAN BARX2_HUMAN, ARX_HUMAN and ASH1L have cluster number 23 with likelihood 0.18, 0.13, 0.21 and 0.69. The occurance rate of cluster 23 is 0.04 in all amino-acids.

Even though, this project aimed to achieve higher performance rates, results are not implausible. We have the chance to observe and estimate regions of such domains in unreviewed proteins and the probabilities are persuasive. There were several drawbacks, but they are mainly computer capacity. Handling big data require much bigger processing power. In addition, reliability to databases are still an issue, because the technology for examining such small molecules.

References

1. A n online database of amino acid physico-chemical properties
ftp://ftp.genome.jp/pub/db/community/aaindex/
2. An online hub for collecting functional knowledge about proteins with annotations obtained from literature and curator-evaluated analysis
http://www.uniprot.org/
3. An online website on protein sequence analysis & classification
https://www.ebi.ac.uk/interpro/