Analysis of Amino Acid Sequence Characteristics of Type I Cluster of Differentiation (CD) Proteins Using Multivariate Statistics to Determine Their Functional Class

A Thesis submitted
to the Graduate School
Valdosta State University

in partial fulfillment of requirements
for the degree of

MASTER OF SCIENCE

in Biology

in the Department of Biology
of the College of Arts and Sciences

December 2017

Pramir K.C.

BS, Southeastern Louisiana University, 2013

This thesis, "Analysis of Amino Acid Sequence Characteristics of Type I Cluster of Differentiation (CD) Proteins Using Multivariate Statistics in Order to Determine Their Functional Class," by Pramir K.C., is approved by:

Thesis Committee Chair

Jonghoon Kang, Ph.D.
Professor of Biology

Corey Andr

Committee Member

Corey Anderson, Ph.D.
Associate Professor of Biology

Donna Gosnell

Donna Gosnell, Ph.D.
Associate Professor of Chemistry

J T. LaPlat

Dean of the Graduate School

James T. LaPlant, Ph.D.
Professor of Political Science

FAIR USE

DUPLICATION

Signature ____

I refuse permission for this thesis to be duplicated in whole or in part.

Signature ____

ABSTRACT

Cluster of Differentiation (CD) proteins are proteins found in the cell membranes of leukocytes. These proteins are important because they are cell surface markers for many immune cells and can be used as therapeutic and diagnostic targets. Biophysical methods like X-ray crystallography and nuclear magnetic resonance (NMR) are commonly used to determine the function of proteins through the generation of their three-dimensional structures. However, applications of these experimental methods do not work very well in order to determine the function of membrane proteins because of their high flexibility and instability, their partial hydrophobic surface, and the requirement of highly specific detergents for their extraction from phospholipids membranes. In order to address this problem, we devised a theoretical approach where type I CD proteins can be classified into two different functional groups (enzyme and non-enzyme) by using physicochemical parameters related to the primary sequence of the individual CD proteins. Principal component analysis (PCA) was used to analyze 126 parameters of 244 type I CD proteins. Two different clusters of type I CD proteins with enzymatic activity and non-enzymatic activity were found on the score plot, and the separation of those clusters was found to be statistically significant. Cytoplasmic amino acid count was found to be the most important variable for separating enzymes and non-enzymes. The continuous probability densities of CD proteins with enzymatic activity and non-enzymatic activity were then approximated by kernel density estimation (KDE) of cytoplasmic amino acid count. This is the first time this method of determining type I CD proteins functional classes has been employed and appears quite promising. In the future, this statistical approach could be very useful in determining the functional class of newly discovered or poorly characterized type I CD proteins.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

My journey of graduate school started with much excitement, and along the way, I went through many ups and downs. Finally, I am here about to graduate with a M.S. in Biology. This experience has been a very special one, and it has taught me the importance of work ethics, industriousness, collaborations and time management.

This massive endeavor would not have been possible without the help and support from a myriad of people. I would like to acknowledge Dr. Jonghoon Kang, who provided me with the opportunity to work on this research project. Together, we shared many exciting moments in our research lab and outside lab (in pubs). I would also like to acknowledge my thesis advisors, Dr. Corey Anderson and Dr. Donna Gosnell, for their invaluable help and support throughout my time in graduate school. Special thanks to Dr. William J. Loughry, Dr. Robert Gannon, Dr. Teresa Doscher, Dr. Leslie Jones, and Mrs. Regina L. Ogden who helped me feel at home at Valdosta State University.

I would like to thank my friends: Namrata Bhandari, Josephine Shieh, Ellen Vedas, Keisha Flukas, Brandi Griffin, and Tan Nguyen who made my life a lot more enjoyable in Valdosta. Finally, I would like to thank my family. Without their love and support, I would have been stuck in a tiny town of Tulsipur in Nepal, and I never would have experienced so many amazing and crappy moments. I believe that because of the summation of those experiences, I am the person I am today! Peace!

Chapter I

INTRODUCTION

Proteins are important macromolecules found in living organisms. According to Alberts et al. (2002), proteins exhibit a wide variety of functions, such as: catalysts, signal receptors, switches and motors. This huge diversity in protein function is attributed to their ability to bind with specific molecules. There are two broad classifications of proteins: water-soluble proteins and membrane proteins. Water-soluble proteins are found in the aqueous medium, and they fold into globular structure, because the amino acids found in their interior are mostly hydrophobic while the amino acids found on the surface are hydrophilic (Berg, Tymoczko, and Stryer, 2002). Membrane proteins consist of integral membrane proteins and peripheral membrane proteins. While peripheral membrane proteins are found on the surface of the phospholipid bilayer, integral membrane proteins have one or more transmembrane domains embedded in the hydrophobic phospholipid bilayer core and have mostly hydrophobic amino acids. Cytosolic and exoplasmic domains are found in aqueous mediums, and have mostly hydrophilic amino acids (Lodish et al., 2000). The study of membrane proteins presents major challenges, because they can only be studied in vitro, outside of their native membranes where they are often flexible and unstable, and specific detergents are required to keep them functionally stable (Carpenter, Beis, Cameron, and Iwata, 2008).

Seddon, Curnow, and Booth (2004) state three major challenges while studying the three-dimensional structures of membrane proteins. The first challenge is acquiring the desired protein types. A wide variety of membrane protein types are present in the membranes of most cells and are usually present in low quantity. To enhance the yield of desired protein, heterologous expression is used. This works well for water-soluble proteins. However, for membrane proteins, the expressed proteins aggregate in the cytoplasm of the heterologous host. Similarly, for mammalian proteins, post-transcriptional modifications are required to generate functional proteins, but heterologous hosts are devoid of those mechanisms. The second major challenge arises from the fact that the phospholipid membrane provides a complex, heterogeneous, and dynamic environment to membrane proteins. Thus, to study the structure of membrane proteins using standard biophysical methods (NMR and X-ray crystallography) samples need to be prepared in vitro using a detergent/lipid medium. Unless a highly appropriate detergent/lipid is selected, there is a high probability that erroneous results might be obtained due to the spectral contributions from the lipid/detergents. The third major challenge to studying membrane proteins is the preparation of a synthetic system that emulates the behavior of the protein of interest in its native environment. Creating the lipid/detergent environment where the isolated membrane proteins retain their native structure and function is very challenging.

One of the most important steps in the reconstruction of membrane proteins during in vitro studies of three-dimensional structure is solubilization. Solubilization is necessary because if the transmembrane proteins are separated from the membrane, their hydrophobic portions will interact with one other, causing them to precipitate out of the solution (Lodish et al., 2000). Moreover, difficulty exists in other areas including: expression, purification, crystallization, data collection and structure solution for membrane proteins. According to Carpenter et al. (2008), the selection of an appropriate detergent is critical for biochemical methods used to determine three-dimensional structure of proteins (i.e., isolation, purification, solubilization and recrystallization). This problem is so prevalent that among all the membrane proteins currently available in the PDB database, less than 1% of them have been successfully crystallized (Parker & Newstead, 2016). Hence, there is a great need for alternative theoretical methods that can accurately predict the functions of membrane proteins.

To address this problem, an alternative theoretical approach was developed, where bioinformatics and multivariate statistics tools were used to categorize type I CD proteins into two distinct groups, based on their functional relevance, simply by analyzing their physicochemical characteristics derived from their amino acid sequence. According to Zola, Swart, Nicholson, and Elena (2007), the amino acid sequence of a protein determines its three-dimensional structure, which then dictates the function of the protein. Based on this fact, we formed a scientific hypothesis: by analyzing 126 different physicochemical properties (directly related to their primary sequence) of 244 type I CD proteins, they can be classified into two broad categories based on function: Enzymes and Non-enzymes.

The purposes of this thesis were: 1) to classify the type I CD proteins into distinct groups based on function (enzymes or non-enzymes); 2) to use principal components analysis to identify the variable (out of 126 different variables) that contributed the most in separating enzyme and non-enzyme; 3) to predict the likelihood of a novel type I CD protein being an enzyme or non-enzyme based on the probability distribution of currently available type I CD proteins using the highest contributing variable. We found out that, just by using bioinformatics datasets and multivariate statistics tools (and few others statistical tools), type I CD proteins can be classified into two groups: enzymes and non-enzymes. Using the kernel density estimation of the most important variables, the likelihood of a newly discovered type I CD protein being enzyme or non-enzyme can be predicted.

Chapter II

LITERATURE REVIEW

Membrane Proteins

There are two broad categories of membrane proteins: integral membrane proteins and peripheral membrane proteins. Integral membrane proteins span through the plasma membrane, while peripheral membrane proteins are found on the surface of the membrane (Lodish et al., 2000). Furthermore, integral membrane proteins can be grouped into different classes based on the orientation of N-terminus and C-terminus of the polypeptide chain. For type I integral membrane proteins, the N-terminus lies outside of the cell, while the C-terminus lies in the cytoplasm. The signal sequence from the protein is cleaved before it is transported to the cell membrane. For type II integral membrane proteins, the C-terminus lies outside of the cell membrane, while the N-terminus of the polypeptide lies in the cytoplasm. In type II integral membrane proteins, the N-terminus cytoplasmic sequence is short and no cleavage of the protein occurs. Type III and type IV proteins are both multipass proteins. Type V protein is attached to the membrane by its C-terminus through glycosyl-phosphatidylinositol (GPI linked) (Zola et al., 2007). Our research focuses on the type I class of integral membrane proteins.

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a multivariate statistics tool which analyzes the observations from multiple correlated variables such that important information can be obtained. The purpose of applying PCA is to reduce the dimensionality of potential predictor variables by identifying which variables best explain the variance in the data set. PCA computes the new set of orthogonal variables called principle components (PCs). The number of PCs obtained after principal components analysis is the same as the number of variables for the original dataset. The first PC (PC1) explains the highest amount of the variance from the original dataset. The second PC (PC2) is orthogonal to PC1 and is responsible for the second highest amount of variance. The remaining PCs are computed in a similar way. Each observation has score values for each principal component, which can be plotted to observe the distribution of the whole dataset based on the new dataset. When the most important PCs are retained (and the PCs carrying noise are excluded) and their score values are plotted, a clear and concise picture of the distribution of the data can be observed (Abdi & Williams, 2010).

Parallel Analysis

Parallel analysis is the most effective statistical method to determine the optimal number of principal components to be retained (Dinno, 2009). Most of the methods used to determine the optimal number of principal components are subjective. This results in either the loss of information in the case of under-extraction or inclusion of noise, which affects subsequent analysis in the case of over-extraction (Franklin et al., 1995). Parallel analysis uses Monte Carlo simulation to generate a random dataset equal to the original data set in terms of number of observations and variables. Horn's parallel analysis performs PCA on random datasets (uncorrelated variables) and the original dataset in order to compare the eigenvalues between them and produce the principal components that are adjusted for sampling error-induced inflation. Only those principal components whose estimated bias (difference of unadjusted eigenvalues and adjusted eigenvalues) are greater than one are retained (Dinno, 2009).

Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a non-parametric method, which is used to predict the probability density function from a set of discontinuous measurements of a random variable. When plotting a histogram, data are binned into discrete classes and the bar represents the relative frequency of occurrence in that bin, whereas during KDE, a continuous distribution (of a specified shape and bandwidth) is centered on each data point and then all of the kernels are added to obtain the KDE (Deng & Wickham, 2011; Zambon & Dias, 2012). The most commonly used kernel weighting function is the Gaussian distribution, but other commonly used kernel weighting functions are Epanechnikov, Uniform, and Triweight. Once a kernel function is chosen, one must also select a bandwidth. When the bandwidth is too large, an overly smooth probability density estimate is generated, which will obscure important characteristics of the distribution. On the other hand, when the bandwidth is small, the distribution becomes noisy (Zambon & Dias, 2012).

Chapter III

MATERIALS AND METHODS

Data Retrieval

Retrieval of Type I Protein List

A CD protein list was obtained from www.hcdm.org (Engel et al., 2015). Out of 371 CD protein molecules found on hcdm.org, 244 of them were identified as type I CD proteins using a bioinformatics protein database called UniProt Knowledgebase (UniProtKB) (Breuza et al., 2016). The type of CD protein was accessed from the "Subcellular Location" panel of UniProtKB for each type I CD protein from the list. The correct UniProtKB accession number was obtained for each type I CD protein. In order to verify if the CD proteins obtained from hcdm.org were the same as the proteins that were used in UniProtKB, alternative names were matched from both websites.

For each type I CD protein, 126 different sequence parameters were selected. These parameters were primary and secondary physicochemical characteristics, which were obtained from the primary amino acid sequence of each type I CD proteins using various bioinformatics tools.

Retrieval of Protein Sequence

In order to obtain the sequence for each type I CD protein, the UniProtKB accession number was entered in the UniProtKB search toolbar. The "Feature Table" displayed protein features such as: topology, molecular processing, secondary structure etc. The sequence for each topological domain: extracellular, transmembrane, and cytoplasmic were retrieved from the topology tab under “Feature Table.”

Retrieval of Regional Amino Acid Count

The regional amino acid count (extracellular amino acid count, transmembrane amino acid count, and cytoplasmic amino acid count) were extracted from the subcellular location panel of UniProtKB. The number of amino acids in each topological domain were recorded in Excel for each type I CD protein.

Total Amino Acid Count Calculation

The total amino acid count for each type I CD protein was obtained by summing up each regional amino acid count. The signal peptide sequence was excluded from the total amino acid count calculation.

Retrieval of Charges

The positive and negative charges for each topological domain: extracellular, transmembrane, and cytoplasmic were extracted by using ExPASy ProtParam tool under the “Sequence” feature of UniProtKB (Gasteiger et al., 2005). The total positive charge was calculated by summing the positive charges from each topological region. Similarly, the total negative charge was calculated by summing the negative charges for each topological region. The total charge was calculated by subtracting total negative charge from total positive charge. Similarly, the absolute charges were calculated by taking the absolute value of total extracellular charges, total transmembrane charges and total cytoplasmic charges, respectively in Excel. Similarly, total absolute charges were calculated by taking absolute values of total charges.

Retrieval of Number of Amino Acids

Retrieval of Number of Individual Amino Acid Type

For each type I CD protein, the number of amino acids from each of the 20 individual amino acids types were recorded separately using the ExPASy ProtParam tool. The numbers were recorded for each topological domain.

Retrieval of Total Number of Each Individual Amino Acid Type

The total number of each individual amino acid was recorded from the chain sequence tab of ExPASy ProtParam tool. The chain sequence tab excluded the signal peptide and included only the total amino acid sequence. To avoid mistakes, while recording the data, the entire set of amino acid numbers was copied directly into Microsoft Excel file from the ExPASy ProtParam tool's screen.

Retrieval of Theoretical Isoelectric Point (pI)

The theoretical isoelectric point (pI) was retrieved from the amino acid sequence for each topological domain (extracellular, transmembrane and cytoplasmic) using the ExPASy ProtParam tool.

Retrieval of Instability Index

The instability index was retrieved from sequence fragments for each topological domain (extracellular, transmembrane and cytoplasmic) using the ExPASy ProtParam tool.

Retrieval of Aliphatic Index

The aliphatic index was retrieved from sequence fragments for each topological domain (extracellular, transmembrane and cytoplasmic) using the ExPASy ProtParam tool.

Retrieval of Grand Average of Hydropathicity

The grand average of hydropathicity was retrieved from sequence fragments for each topological domain (extracellular, transmembrane and cytoplasmic) using the ExPASy ProtParam tool.

Retrieval of Glycosylation Site for Extracellular Region

The glycosylation site for extracellular region were retrieved using the PTM/Processing tab of UniProtKB. The total number of glycosylation position(s) located only in the extracellular domain was counted and recorded.

Retrieval of Phosphorylation Site for Cytoplasmic Domain

The amino acid sequences from the cytoplasmic domain for each type I CD protein were extracted from UniProtKB. The phosphorylation sites were computed by running the sequence in ExPASy: SIB Bioinformatics Resource Portal tool Netphos 2.0 Server (Artimo et al. 2012; Blom, Gammeltoft, & Brunak, 1999). The total number of Serene, Threonine and Tyrosine residues predicted by the Netphos 2.0 neural network was recorded. The number of phosphorylation sites with less than 15 and more than 4000 amino acid residues could not be computed by Netphos 2.0 server. For the cytoplasmic sequence with less than 15 amino acid residues, the number of phosphorylation sites was recorded as zero.

Retrieval of Secondary Structure Content

The amino acid sequences for the extracellular and cytoplasmic domain were extracted from UniProtKB. The secondary structure content was obtained by using the GOR IV tool of the ExPASy: SIB Bioinformatics Resource Portal (Sen, Jernigan, Garnier, & Kloczkowski, 2005).

Alpha Helix Content

For each type I CD protein, alpha helix content (%) for extracellular and cytoplasmic domains were retrieved.

Beta Sheet Content

For each type I CD protein, beta sheet content (%) for extracellular and cytoplasmic domains were retrieved.

Random Coil Content

For each type I CD proteins, random coil content (%) for extracellular and cytoplasmic domain were retrieved.

Retrieval of Disorder Average and Standard Deviation

Disorder average and disorder standard deviation was calculated for each topological domain of type I CD proteins. The bioinformatics tool IUPred: Prediction of Intrinsically Unstructured Proteins (Dosztányi, Csizmok, Tompa, and Simon, 2005) was used to predict the disorder tendency for each amino acid residue. The average and standard deviation of disorder tendency was computed in MS Excel using the disorder tendency of individual residue, for each topological domain. While selecting the prediction criteria for IUPred, long disorder and raw data were used as prediction type and output type, respectively.

Determination of Function (Enzyme or Non-Enzyme)

To assess the functional category of each type I CD protein (enzyme or non-enzyme), the information under the “Function” tab of UniProtKB was examined. Based on the information found under the “Function” tab, type I CD proteins with enzymatic activity were assigned as enzymes, and those with non-enzymatic activity (i.e., binding, signal anchor) were assigned as non-enzymes. For those CD proteins that could not be clearly identified as having enzymatic activity or non-enzymatic activity, further investigation into scientific publications was made to clarify their functional class.

Statistical Analysis

Principal Component Analysis (PCA)

MiniTab 17 statistical software (2010) was used to perform principal component analysis (PCA) on the Type I CD protein dataset, with 244 observations and 126 different variables. While performing the PCA, the correlation matrix option was selected because the units and range of the values of the variables differed.

Parallel Analysis to Determine the Number of Principal Components to be Retained

Horn's parallel analysis was performed to determine the number of principal components to be retained for the analysis. R's 'paran' package was used to perform Horn's parallel analysis (Dinno, 2012).

Assessment of Statistical Significance of Separation of Enzymes and Non-Enzymes from PCA Data

Applying the methods used by Goodpaster and Kennedy (2011), the statistical significance of separation of enzymes and non-enzymes from the principal component analysis data was tested. The centroids for Enzymes and Non-Enzymes clusters were calculated from the PCA score value. The centroid values are the average score values for each principal component, from PC1 to PC10 for Enzymes and Non-Enzymes, respectively. The Mahalanobis distance (D_M) was calculated as:

$$D_M = \sqrt{d' C_w}^{-1} d$$

where d = 1×2 Euclidian difference vector between the centroids of enzymes and non-enzymes, calculated as

$$d = \left[\overline{\mathrm{PC1}}(\mathrm{NE}) - \ \overline{\mathrm{PC1}}(\mathrm{E}), \overline{\mathrm{PC2}}(\mathrm{NE}) - \ \overline{\mathrm{PC2(E)}}, \ldots\ldots\ldots\ldots.., \overline{\mathrm{PC10}}(\mathrm{NE}) - \ \overline{\mathrm{PC10(E)}}\right]$$

and C_w^{-1} = Inverse of the pooled variance-covariance matrix between enzymes (E) and non-enzymes (NE).

Here, V is the pooled variance-covariance matrix between enzyme (E) and non-enzyme (NE).

Hoteling's T^2 was calculated using the formula:

$$T^2 = \frac{\mathrm{n1\,.\ n2}}{\mathrm{n1+n2}}\ d' C_w{}^{-1} d$$

where n1 = 26 (number of Enzymes) and n2= 218 (number of Non-Enzymes).

The Hoteling's T^2 can be converted into F-statistics by using the following formula:

$$F = \frac{\mathrm{n1+\ n2-p-1}}{\mathrm{p(n1+n2-2)}} T^2$$

In this case, p is the number of discriminator variables. Since, 10 principal components (PC1 to PC10) are being evaluated in this case, p = 10.

The F-value is the ratio of between group variance to that of within group variance between enzyme and non-enzyme.

$$\text{Critical } F\text{-value} = \ \mathrm{F(p, n1 + n2 - p - 1)}$$

The critical F-value was calculated using http://www.danielsoper.com/statcalc/calculator.aspx?id=4 (Soper, 2017). The server required the number of degrees of freedom in the numerator, which is equal to two. The number of degrees of freedom in the denominator is given by the formula $(\mathrm{n1 + n2 -}$

$\mathrm{p} - 1$), which equaled 241. The F-critical value was calculated at a significance level (α) of 0.05.

To determine if the separation of type I CD proteins with enzymatic activity (as represented by blue dots in Figure 2) and with non-enzymatic activity (as represented by red dots in Figure 2) was statistically significant, the F-value computed using Hoteling's T^2 value and F-critical value was compared.

Matrix Plot of Scores Values

To observe the pattern of distribution of enzymes and non-enzymes in a two-dimensional scatter plot, a matrix plot was created where PCs (PC1 to PC10) were paired with one another. Forty-five different two-dimensional scatter plots were obtained in the matrix plot. Mahalanobis distance was computed between the centroid of enzymes and non-enzyme clusters for each individual scatter plot. By using the Mahalanobis distances, Hoteling's T^2 test and F-statistics were assessed. Out of 45 scatter plots, the PC that had the highest Mahalanobis distances and their corresponding T^2 values and F-statistics values was be the best PC for separating enzymes and non-enzymes.

Wilcoxon Ranked-Sum Test and Kernel Density Estimation (KDE) Plot

The Wilcoxon ranked-sum test was performed to if the variable that had highest loading value for the most important PC had the same distribution for enzyme and non-enzyme data. R 3.4.2 was used to perform the analysis. Kernel Density Estimation (KDE) was used to plot the probability density estimation function for the non-parametric data of cytoplasmic amino acid count for enzymes and non-enzymes. The density function from the 'stats' package was used to perform KDE analysis (Bowman & Azzalini, 2014). Gaussian kernel and direct plug-in (dpi) from bw.SJ functions were used to calculate the kernel type and bandwidth used. In order to integrate the kernel density function over different cytoplasmic AA count, R package sfsmisc was used (Maechler et al., 2017).

Chapter IV

RESULTS

Principal Component Analysis

Scree Plot

Based on the scree plot (Figure 1), 126 total principal components (PCs) were obtained. Among those 126 PCs, principal component 1 (PC1) and principal component 2 (PC2) had eigenvalues of 42.504 and 20.967 respectively. The first ten PCs accounted for 71.8 % of the total variance of the data.

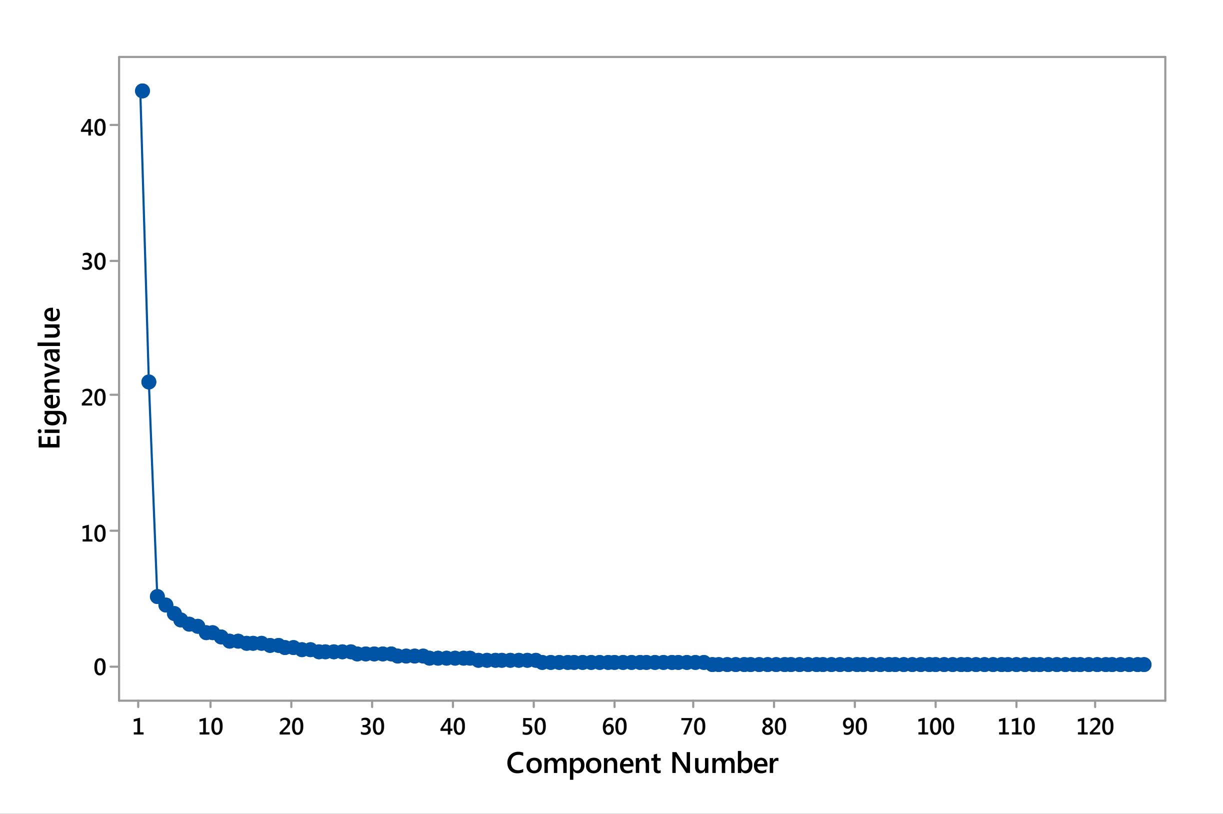

Figure 1. Scree plot of physicochemical characteristics of Type I CD proteins.

Horn's Parallel Analysis

Horn's parallel analysis was performed to determine the number of principal components to be retained. Based on Horn's parallel analysis (Figure 2, Table 1), the first 10 PCs were retained.

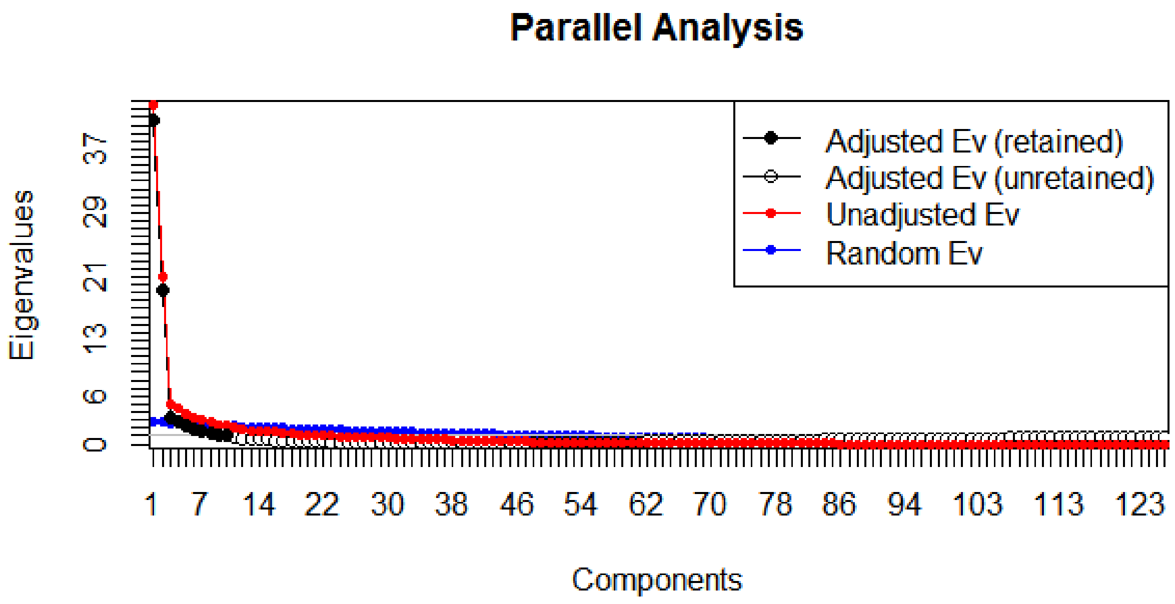

Figure 2. Horn’s Parallel Analysis plot to determine the optimal number of PCs retained. PCs with adjusted EV greater than 1 were retained.

Table 1. Components Retention of First 10 PCs (3780 Iterations).

Component	Adjusted Eigenvalue	Unadjusted Eigenvalue	Estimated Bias
1	40.5995	42.45059	1.851084
2	19.25349	20.98743	1.733934
3	3.331805	4.979356	1.647551
4	2.852459	4.42674	1.574281
5	2.255807	3.76413	1.508323
6	1.881135	3.329175	1.44804
7	1.626716	3.018798	1.392082
8	1.445181	2.785131	1.339949
9	1.087669	2.377974	1.290304
10	1.118029	2.36047	1.242441

Matrix Plot for First 10 PCs Score Values

After plotting the matrix plot of the first 10 PCs score values with each other, 45 different combinations of unique two-dimensional scatter plots were obtained (Figure 3). It is clearly visible that all the scatter plots including PC2 had a higher separation of enzymes and non-enzymes cluster compared to the other 36 scatter plots where PC2 score values were not used. This clearly indicates that among the first 10 PCs retained from Horn’s parallel analysis, PC2 was the most effective in separating enzymes and non-enzymes into two different clusters.

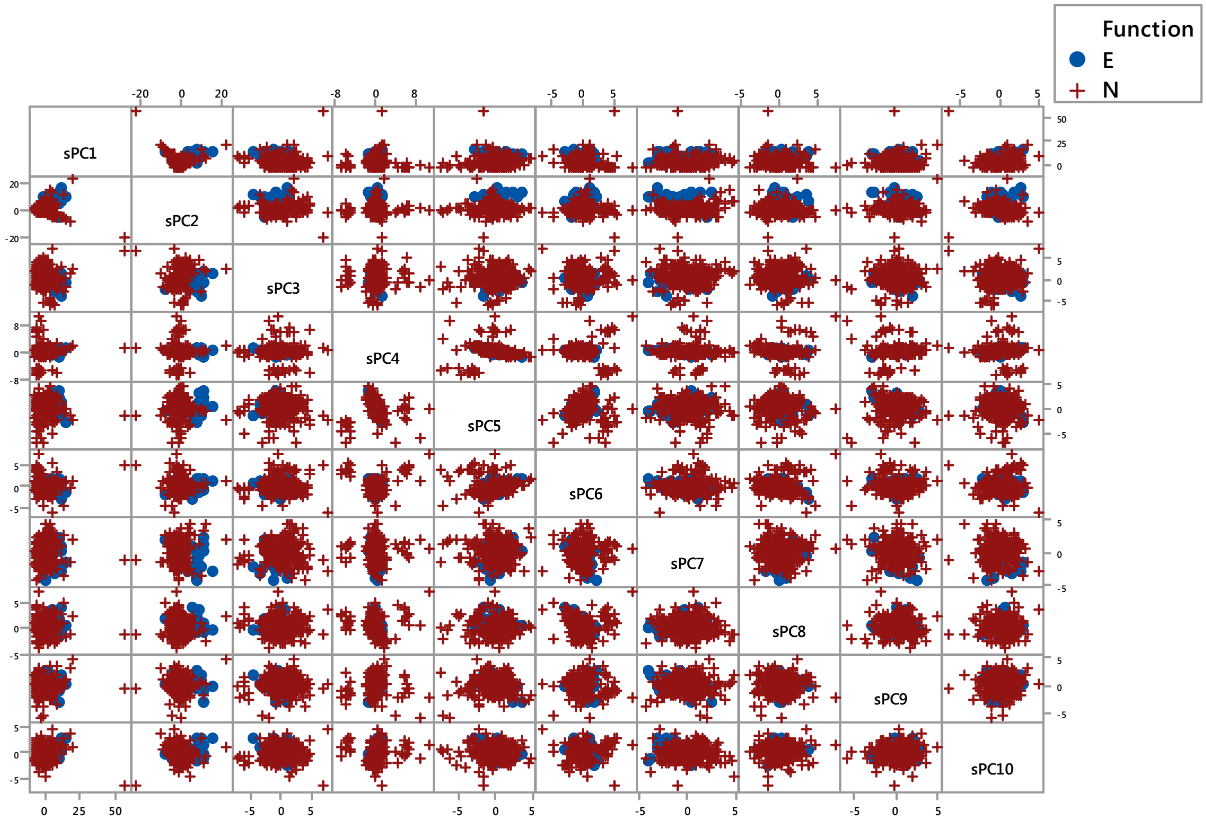

Figure 3. Matrix plot of score values for PC1 to PC10. All the score plots that involved PC2 yielded the greatest amount of separation between enzymes and non-enzymes. Separation of Enzyme (E) and Non-enzyme (NE) clusters.

Hoteling's T^2 and F-statistics for the enzyme group and the non-enzyme group were calculated using the score values for the first 10 PCs to find out if the separation of enzymes and non-enzymes clusters obtained after performing PCA was statistically significant (Table 2). Mahalanobis distance between the enzyme and non-enzyme groups was 7.014. The Hoteling's T^2 value obtained for enzyme and non-enzyme clusters was 1142.824. The F-statistic calculated using Hoteling's T^2 value was 110.0322 and the critical value of $F_{0.05,10,233}$ was 1.87. This signifies that when PCA was used for the type

I CD protein data, there was a statistically significant separation of enzyme and non-enzyme clusters.

Table 2. Mahalanobis Distance (Dm), Hoteling's T^2 and F-statistics between Enzyme and Non-enzyme Based on the First 10 PCs.

	For Enzymes and Non-enzymes (10 PCs)
Mahalanobis Distance (Dm)	7.014
Dm^2	49.197
Hotelling's T^2	1142.82
F-statistics	110.032
F-critical	1.87

From the matrix plot, it was apparent that the two-dimensional score plots containing PC2 had the highest separation of enzyme and non-enzyme clusters. To test this observation objectively, the Mahalanobis distance and the corresponding T^2 and F-statistics for each individual score plot involving the first ten PCs were calculated (Appendix A).

Loadings Plot for PC1

Based on the loading values for PC1 (Appendix B, Figure 4), total amino acid count displayed the highest contribution to PC1, with a loadings value of 0.153. Similarly, other variables that have relatively high loadings values (in descending order) are total negative charges, total positive charges, total number of Glutamic acid (E), total number of Aspartic Acid(D), total number of Serine (S), total number of Arginine (R), total number of Glycine(G), total number of Tyrosine (Y), and total number of Valine (V), respectively. The variables that are related to the entire type I protein (not just variables for particular topological domains) contribute the most to PC1 (Figure 4, Appendix B).

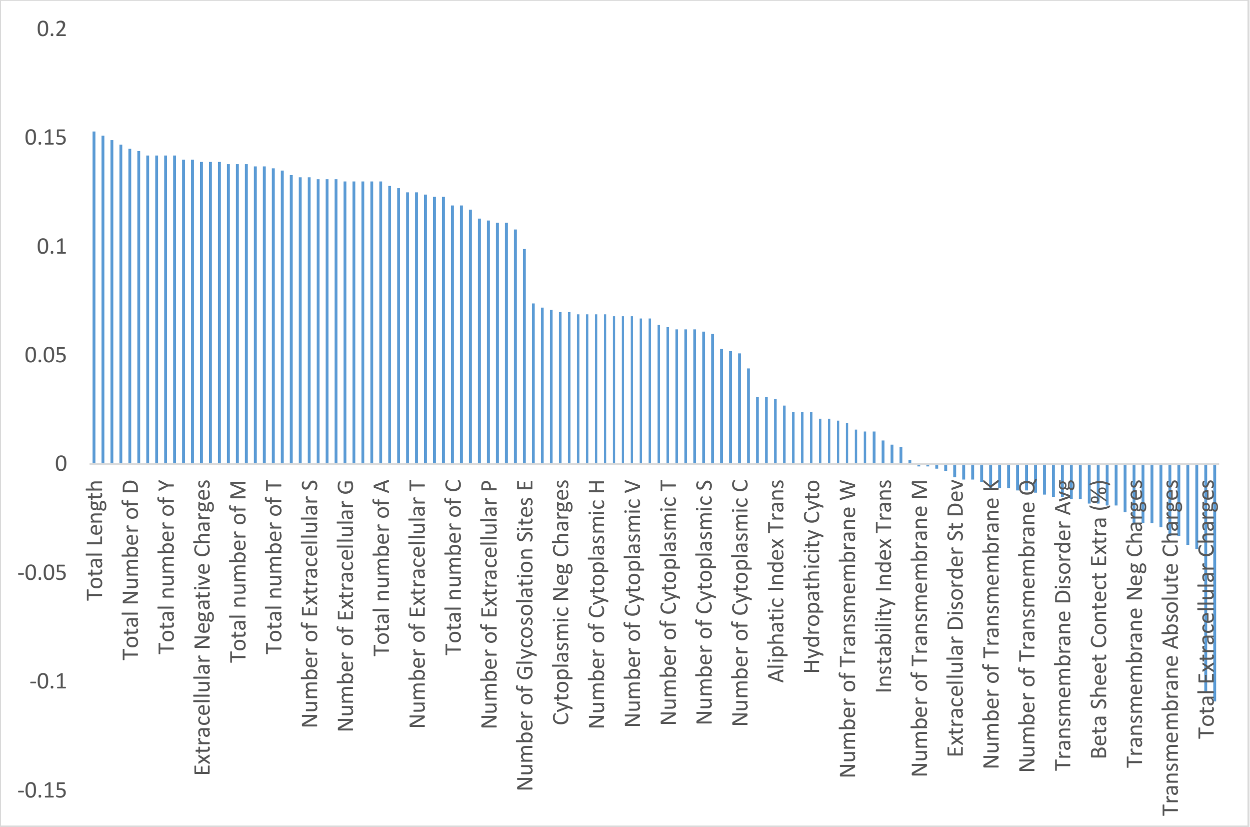

Figure 4. Loadings plot for 126 protein characteristics using PC1.

Loadings Plot for PC2

Since PC2 was established as the most important PC to separate enzyme and non-enzyme clusters, PC2 loading values were analyzed to identify the variable that showed the highest contribution to PC2. The loading values were plotted in MS Excel to obtain the loadings plot for PC2 (Figure 5). Based on the PC2 loading values (Table 2, Figure 5), cytoplasmic amino acid count had the highest loading value of 0.193. Similarly, some of the other parameters with high loadings value were number of Cytoplasmic Negative

Charges, number of Cytoplasmic Glutamic Acid (E), number of Cytoplasmic Leucine (L), number of Cytoplasmic Valine (V), number of Cytoplasmic Serine (S), Cytoplasmic positive Charges, Number of Phosphorylation Sites, number of Cytoplasmic Alanine(A), number of Cytoplasmic Aspartic Acid (D), number of Cytoplasmic Glutamate (Q), number of Cytoplasmic Threonine (T) and number of Cytoplasmic Arginine (R). As indicated in (Figure 5 and Appendix B), the variables that are related to cytoplasmic topological domain contribute the most to PC2.

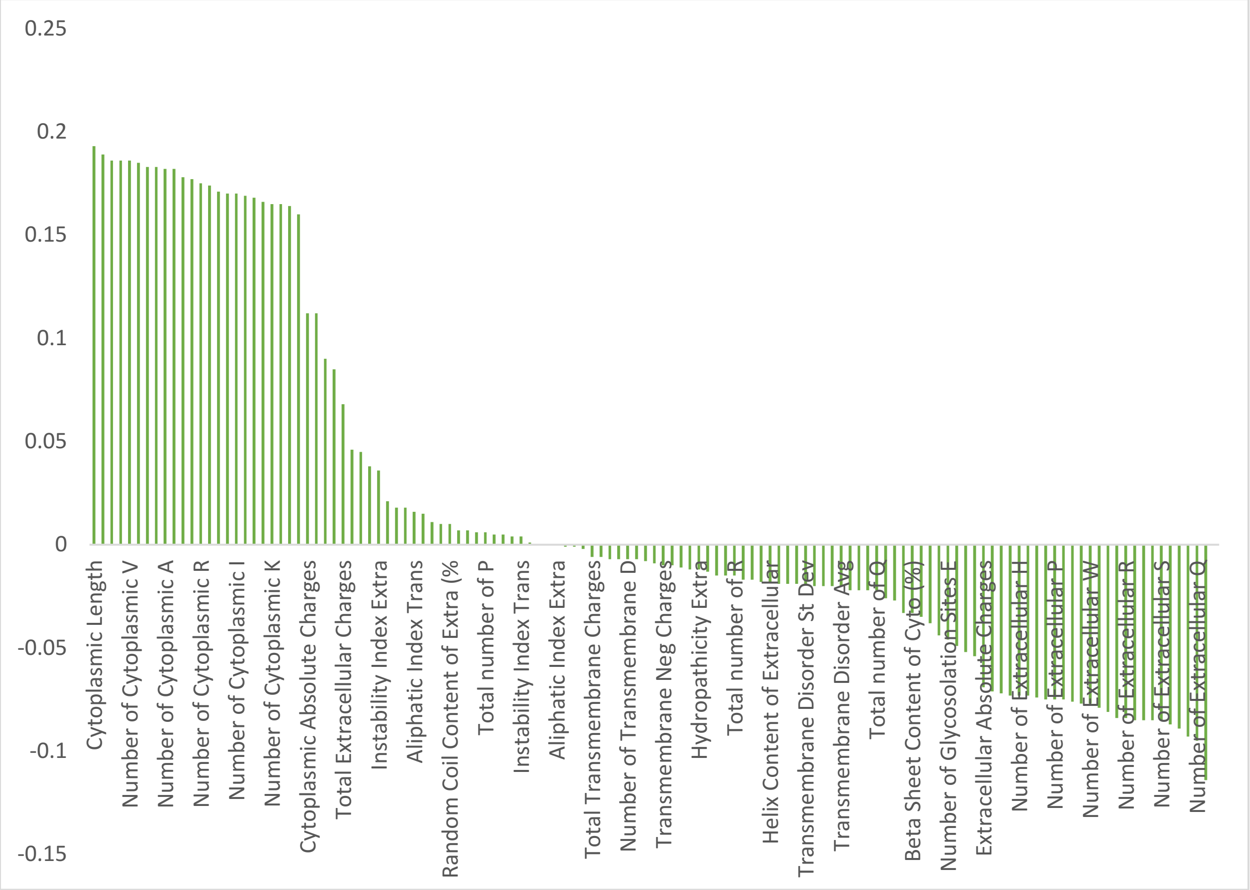

Figure 5. Loadings plot for 126 variables using PC1.

Statistical Difference and Kernel Density Estimation (KDE)

Wilcoxon ranked-sum test (equivalent to the Mann-Whitney U test) was performed to determine if there was a significant difference in the median of cytoplasmic amino acid count between enzymes and non-enzymes. The number of type I CD proteins that were identified as enzymes and non-enzymes were 218 and 26, respectively (Appendix C). The median cytoplasmic amino acid count between enzyme and non-enzyme was 424 and 60 respectively. The p-value was found to be 2.054e^{-10} which is much smaller than the significance level of 0.05. This result indicated that there was a significant difference in the distribution of cytoplasmic amino acid counts between enzymes and non-enzymes of Type I CD proteins.

When the probability density functions were plotted for the cytoplasmic amino acid count for enzymes and non-enzymes, the probability of obtaining non-enzyme was extremely high in the 0 to 280 range (approximate point where enzyme and non-enzyme kernel density estimation curve met). However, when the cytoplasmic amino acid count exceeded 280, the probability of obtaining non-enzymes was quite small. At a cytoplasmic AA count higher than 280, the probability of obtaining enzymes was substantially higher. When the KDE for enzyme and non-enzyme was integrated within the range of 0-280, the probability of obtaining enzyme and non-enzyme within that range was 0.18 and 0.93, respectively. However, when the KDE for enzyme and non-enzyme were integrated in the range of 280-1089, the respective probabilities of obtaining enzyme and non-enzyme were 0.79 and 0.04, respectively.

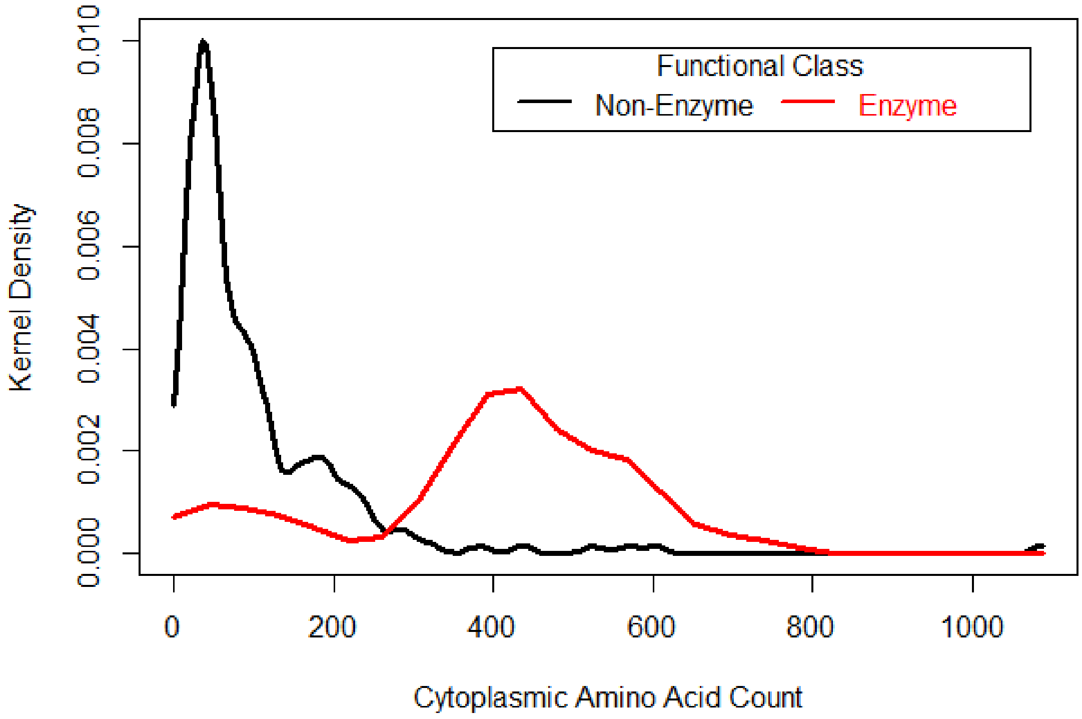

Figure 6. Kernel density estimation for enzymes and non-enzymes.

Chapter V

DISCUSSION

PCA was found as an effective tool for separating enzyme and non-enzyme clusters for type I CD proteins. Previously, Patterson and Kang (2011) successfully separated enzyme and non-enzyme clusters for type II CD proteins using principal component analysis. Thus, the results from these two experiments help to establish principal component analysis as an important statistical tool for the separation and classification of all single-pass CD proteins based on their function. Since type I and type II CD proteins are a subclass of single-pass transmembrane proteins, it is proposed that PCA can be applied to all single-pass transmembrane proteins for the separation and classification of enzymes and non-enzymes.

Prediction of Type I CD Proteins Functional Class Based on Their Cytoplasmic Amino Acid Counts

Results from this experiment indicated that the probability of a randomly selected type I CD protein being an enzyme or non-enzyme can be predicted based on its cytoplasmic amino acid count. Out of 126 protein characteristics, the 30 most important protein characteristics were associated with cytoplasmic domain. This finding suggested that the cytoplasmic domain was important for determining the enzymatic activity of a type I CD protein. When we assessed the catalytic function of each enzyme for type I CD proteins, we found that the majority of the enzymes (21 out of 26) were protein kinases/phosphatases kinases/phosphatases, which were involved in signal transduction, and their catalytic domain was present in cytoplasmic region. Moreover, those enzymes involved in signal transduction had their catalytic domain present in the cytoplasmic region and their cytoplasmic amino acid count was substantially larger than their extracellular amino acid count. This clearly indicates that for a type I CD enzyme that participate in signal transduction, cytoplasmic domain plays a key role, which was consistent with our observation. Five out of 26 enzymes were involved in peptidase and oxidase activity and their catalytic domain were present in the extracellular region. This finding suggests that for type I CD enzymes, nature is more biased towards production of signal transduction enzymes and prefers cytoplasmic domain for their catalytic activity. However, this bias is not absolute because type I CD enzymes could be involved in oxidation or cleavage of other proteins, and in that case the catalytic domain is present in the extracellular region. Also, it was found found that if a type I CD protein exists that have enzymatic activity other than signal transduction (peptidase and oxidase activity), its cytoplasmic amino acid count should be subastantly smaller compared to its extracellular amino acid count. In addition, the catalytic domain must be present in the extracellular region. Patterson and Kang (2011) found that type II CD proteins with enzymatic activity had diverse function such as endopeptidases, metalloproteases, ectoenzymes, phosphodiesterases, phosphatases and exopeptidases. All the enzymes for type II CD proteins had their catalytic domains in the extracellular region. They also found that the extracellular amino acid count was the most important protein characteristic for determination of enzymatic activity for a type II CD protein. Since the C-terminus is in the extracellular domain for type II CD proteins, the C-terminus is very important for the catalytic activity of the majority of single-pass CD enzymes.

Since the advent of genome sequencing, the number of novel protein sequences submitted to the UniProt database has grown in an unprecedented way (Mills, Beuning, & Ondrechen, 2015). This method of predicting protein function is broadly applicable because it is a simple tool compared to currently available tools for function prediction. Most of the currently available tools for predicting function depend on the comparison of the uncharacterized protein with a protein having similar structure or function. They use complex algorithms, and many times the functions are incorrectly annotated (Mills et al., 2015). The prediction tool that has been developed from this experiment is quite simple and does not require comparison of sequence or structure to proteins with experimentally verified function. This method employs a well-established statistical method (principal component analysis) to identify the most important protein characteristics responsible for separation of enzymes and non-enzymes. Based on the kernel density estimation (KDE) of the protein characteristics (cytoplasmic amino acid count for type I CD proteins), anyone can predict whether a newly discovered protein or poorly characterized protein will be an enzyme or non-enzyme.

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433-459.

al. MMe (2017). _sfsmisc: Utilities from 'Seminar fuer Statistik' ETH Zurich_. R package

Alberts, B., Johnson, A., Lewis, J., Walter, P., Raff, M., & Roberts, K. (2002). *Molecular Biology of the Cell* 4th Edition: New York: Garland Science. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK21054/

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Stockinger, H. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research*, *40*(Web Server issue), W597–W603. http://doi.org/10.1093/nar/gks400

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2002). *Biochemistry*. 5th edition. New York: W H Freeman. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK21154/

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2008). *Biochemistry (Loose-Leaf)*. Macmillan.

Blom, N., Gammeltoft, S., & Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology*, *294*(5), 1351-1362.

Bowman, AW., and Azzalini, A. (2014). _R package \texttt{sm}: nonparametric smoothing methods (version 2.2-5.4). <URL: URL http://www.stats.gla.ac.uk/~adrian/sm,

Breuza, L., Poux, S., Estreicher, A., Famiglietti, M. L., Magrane, M., Tognolli, M., & UniProt Consortium. (2016). The UniProtKB guide to the human proteome. *Database*, *2016*, bav120.

Carpenter, E. P., Beis, K., Cameron, A. D., & Iwata, S. (2008). Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, *18*(5), 581-586.

Carpenter, E. P., Beis, K., Cameron, A. D., & Iwata, S. (2008). Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, *18*(5), 581–586. http://doi.org/10.1016/j.sbi.2008.07.001

Deng, H., & Wickham, H. (2011). Density estimation in R. *Electronic Publication*.

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, *44*(3), 362-388.

Dinno, A. (2012). paran: Horn’s test of principal components/factors. R package version 1.5 1.

Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, *21*(16), 3433-3434.

Engel, P., Boumsell, L., Balderas, R., Bensussan, A., Gattei, V., Horejsi, V., & Stockinger, H. (2015). CD Nomenclature 2015: Human Leukocyte Differentiation Antigen Workshops as a Driving Force in Immunology. *The Journal of Immunology*, 195(10), 4555-4563.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S. E., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). *Protein Identification and Analysis Tools on the ExPASy Server* (pp. 571-607). Humana Press.

Goodpaster, A. M., & Kennedy, M. A. (2011). Quantification and statistical significance analysis of group separation in NMR-based metabonomics studies. *Chemometrics and Intelligent Laboratory Systems: An International Journal Sponsored by the Chemometrics Society*, *109*(2), 162–170. http://doi.org/10.1016/j.chemolab.2011.08.009

Goodpaster, A. M., & Kennedy, M. A. (2011). Quantification and statistical significance analysis of group separation in NMR-based metabonomics studies. *Chemometrics and Intelligent Laboratory Systems*, *109*(2), 162-170.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141-151.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2000). *Molecular Cell Biology* 4th edition. New York: W. H. Freeman. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK21475/

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2000). *Molecular Cell Biology* 4th edition. New York: W. H. Freeman. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK21475/

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2000). Molecular Cell Biology 4th edition. New York: W. H. Freeman. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK21475/

Mills, C. L., Beuning, P. J., & Ondrechen, M. J. (2015). Biochemical functional predictions for protein structures of unknown or uncertain function. *Computational and Structural Biotechnology journal*, *13*, 182-191.

Mills, C. L., Beuning, P. J., & Ondrechen, M. J. (2015). Biochemical functional predictions for protein structures of unknown or uncertain function. *Computational and Structural Biotechnology Journal*, *13*, 182-191.

Minitab 17 Statistical Software (2010). [Computer software]. State College, PA: Minitab, Inc. Retrieved from: (www.minitab.com)

Parker, J. L., & Newstead, S. (2016). Membrane protein crystallization: Current trends and future perspectives. *Advances in Experimental Medicine and Biology*, *922*, 61–72. http://doi.org/10.1007/978-3-319-35072-1_5

Patterson, A., Kang, J. (2011). *Statistical Analysis of Sequence Characteristics of Single Transmembrane Cluster of Differentiation Proteins: A Study of Functional Relevance.* (Unpublished Master's Dissertation). Valdosta State University, Valdosta, Georgia.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved From:URL https://www.R-project.org/.

Seddon, A. M., Curnow, P., & Booth, P. J. (2004). Membrane proteins, lipids and detergents: not just a soap opera. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1666*(1), 105-117.

Sen, T. Z., Jernigan, R. L., Garnier, J., & Kloczkowski, A. (2005). GOR V server for protein secondary structure prediction. *Bioinformatics*, *21*(11), 2787-2788.

Soper, D. (n.d.). Calculator: Critical F-value. Retrieved September 05, 2017, from http://www.danielsoper.com/statcalc/calculator.aspx?id=4

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, *25*(1), 1-28.

version 1.1-1, <URL: https://CRAN.R-project.org/package=sfsmisc>.

Zambom, A. Z., & Dias, R. (2012). A review of kernel density estimation with applications to econometrics. *arXiv preprint arXiv:1212.2812*

Zola, H., Swart, B., Nicholson, I., & Voss, E. (2007). *Leukocyte and Stromal Cell Molecules: the CD Markers*. New Jersey: John Wiley & Sons.

APPENDIX A:

Mahalnobis Distance, Hotelling's T-squared Values and F-statistics for each Combinations of Principle Components for PC1 to PC10.

Appendix A: Mahalanobis distance, Hotelling's T-squared values and F-statistics for each combinations of principle components for PC1 to PC10. As highlighted in bold, the highest F-statistics values corresponds to the score plots having one of the component as PC2.

PCs Used	Dm	Dm ^{2}	Hot T^2	F-stat	F-critical
PC1-PC2	**5.65**	**31.89**	**740.82**	**368.88**	3.03
PC1-PC3	1.51	2.29	53.28	26.53	
PC1-PC4	1.34	1.79	41.49	20.66	
PC1-PC5	1.34	1.78	41.43	20.63	
PC1-PC6	1.35	1.81	42.03	20.93	
PC1-PC7	1.79	3.20	74.34	37.02	
PC1-PC8	1.36	1.85	43.09	21.46	
PC1-PC9	1.34	1.79	41.53	20.68	
PC1-PC10	1.36	1.86	43.12	21.47	
PC2-PC3	**3.51**	**12.29**	**285.60**	**142.21**	
PC2-PC4	**3.26**	**10.65**	**247.42**	**123.20**	
PC2-PC5	**3.26**	**10.64**	**247.24**	**123.11**	
PC2-PC6	**3.27**	**10.72**	**249.13**	**124.05**	
PC2-PC7	**3.88**	**15.02**	**348.87**	**173.72**	
PC2-PC8	**3.30**	**10.88**	**252.71**	**125.83**	
PC2-PC9	**3.26**	**10.66**	**247.58**	**123.28**	
PC2-PC10	**3.30**	**10.88**	**252.78**	**125.87**	
PC3-PC4	0.14	0.02	0.47	0.23	

PC3-PC5	0.14	0.02	0.46	0.23
PC3-PC6	0.15	0.02	0.51	0.26
PC3-PC7	0.50	0.25	5.91	2.94
PC3-PC8	0.16	0.03	0.61	0.30
PC3-PC9	0.14	0.02	0.47	0.23
PC3-PC10	0.16	0.03	0.62	0.31
PC4-PC5	0.00	0.00	0.00	0.00
PC4-PC6	0.01	0.00	0.00	0.00
PC4-PC7	0.35	0.13	2.92	1.46
PC4-PC8	0.02	0.00	0.01	0.01
PC4-PC9	0.00	0.00	0.00	0.00
PC4-PC10	0.02	0.00	0.01	0.01
PC5-PC6	0.01	0.00	0.00	0.00
PC5-PC7	0.35	0.13	2.91	1.45
PC5-PC8	0.02	0.00	0.01	0.01
PC5-PC9	0.00	0.00	0.00	0.00
PC5-PC10	0.02	0.00	0.01	0.01
PC6-PC7	0.36	0.13	3.04	1.52
PC6-PC8	0.03	0.00	0.02	0.01
PC6-PC9	0.01	0.00	0.00	0.00
PC6-PC10	0.03	0.00	0.02	0.01
PC7-PC8	0.38	0.14	3.29	1.64
PC7-PC9	0.36	0.13	2.93	1.46

PC7-PC10	0.38	0.14	3.30	1.64
PC8-PC9	0.02	0.00	0.01	0.01
PC8-PC10	0.04	0.00	0.04	0.02
PC9-PC10	0.02	0.00	0.01	0.01

APPENDIX B:

Loadings Values of PC1 to PC10 for 126 Different Physicochemical Properties.

Appendix B: Loadings values of PC1 to PC10 for 126 different physicochemical properties.

Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Extracellular Length	0.140	-0.089	0.002	0.005	-0.016	-0.019	-0.006	0.013	0.003	0.022
Transmembrane Length	0.015	-0.044	0.076	0.052	-0.065	0.053	-0.060	-0.167	0.124	0.070
Cytoplasmic Length	0.071	0.193	0.005	0.012	-0.020	0.006	0.010	-0.002	0.015	0.021
Total Length	0.153	-0.015	0.004	0.009	-0.021	-0.015	-0.003	0.011	0.008	0.027
Extracellular + Charges	0.138	-0.084	-0.020	0.006	-0.014	0.020	-0.030	-0.015	0.019	0.025
Extracellular Negative Charges	0.139	-0.085	0.018	0.001	0.010	0.044	-0.018	-0.013	-0.019	-0.054
Total Extracellular Charges	-0.105	0.068	-0.111	0.014	-0.069	-0.095	-0.016	0.005	0.111	0.243
Transmembrane + Charges	-0.018	-0.020	0.020	-0.303	-0.250	0.259	0.028	0.087	-0.108	0.133
Transmembrane Neg Charges	-0.027	-0.010	0.015	0.331	-0.042	0.266	0.054	0.224	-0.116	0.045
Total Transmembrane Charges	0.008	-0.006	0.002	-0.438	-0.136	-0.020	-0.020	-0.103	0.012	0.056
Cytoplasmic + Charges	0.069	0.183	-0.036	0.032	-0.001	0.022	-0.080	0.025	0.042	0.060
Cytoplasmic Neg Charges	0.070	0.189	0.027	0.004	0.013	0.032	0.030	-0.023	-0.020	0.023
Total Cytoplasmic Charges	-0.039	-0.114	-0.143	0.056	-0.036	-0.036	-0.239	0.110	0.136	0.062

Total + Charges	0.149	-0.007	-0.031	0.016	-0.014	0.027	-0.056	-0.004	0.032	0.045
Total Neg Charges	0.151	-0.007	0.026	0.003	0.014	0.053	-0.005	-0.020	-0.025	-0.040
Total Charges	-0.109	0.007	-0.162	0.030	-0.079	-0.099	-0.124	0.053	0.159	0.240
Extracellular Absolute Charges	0.111	-0.069	0.104	-0.019	0.048	0.068	-0.001	0.000	-0.077	-0.222
Transmembrane Absolute Charges	-0.033	-0.022	0.025	0.039	-0.206	0.380	0.060	0.229	-0.162	0.126
Cytoplasmic Absolute Charges	0.024	0.112	0.103	0.001	0.033	0.060	0.233	-0.106	-0.077	-0.052
Total Absolute Charges	0.111	-0.011	0.138	-0.029	0.059	0.078	0.122	-0.055	-0.123	-0.223
Number of Extracellular A	0.117	-0.085	0.062	0.006	-0.058	-0.121	0.034	0.096	-0.036	0.057
Number of Extracellular R	0.132	-0.084	0.036	0.015	-0.063	-0.043	0.019	0.051	0.006	0.002
Number of Extracellular N	0.131	-0.079	-0.058	-0.006	0.028	0.069	-0.055	-0.067	0.011	-0.025
Number of Extracellular D	0.133	-0.087	0.041	0.003	0.001	0.057	-0.027	-0.009	-0.009	-0.088
Number of Extracellular C	0.113	-0.073	0.090	0.036	-0.059	0.067	-0.087	-0.053	0.075	-0.138
Number of Extracellular Q	0.125	-0.095	-0.008	-0.008	-0.001	-0.057	0.068	0.042	-0.003	0.027
Number of Extracellular E	0.137	-0.077	-0.012	-0.001	0.023	0.024	-0.006	-0.018	-0.028	-0.009
Number of Extracellular G	0.130	-0.093	0.056	0.017	-0.039	-0.035	-0.016	0.042	0.013	0.023

Number of Extracellular H	0.130	-0.073	0.017	0.001	-0.055	-0.055	-0.001	0.063	0.028	0.016
Number of Extracellular I	0.130	-0.072	-0.073	-0.017	0.056	0.084	-0.022	-0.074	-0.009	0.021
Number of Extracellular L	0.123	-0.075	-0.118	-0.004	-0.032	-0.093	0.073	0.090	-0.017	0.004
Number of Extracellular K	0.124	-0.071	-0.083	-0.005	0.046	0.091	-0.083	-0.090	0.032	0.048
Number of Extracellular M	0.128	-0.076	-0.038	-0.004	0.007	0.042	0.018	0.006	-0.003	-0.013
Number of Extracellular F	0.127	-0.085	-0.094	-0.012	0.000	0.004	0.021	0.002	0.019	0.034
Number of Extracellular P	0.112	-0.075	0.105	0.023	-0.064	-0.127	-0.030	0.076	0.028	0.100
Number of Extracellular S	0.132	-0.085	0.000	0.007	-0.028	-0.059	-0.017	0.020	0.013	0.094
Number of Extracellular T	0.125	-0.075	0.052	-0.003	0.014	-0.045	-0.054	0.009	-0.031	0.078
Number of Extracellular W	0.108	-0.077	0.019	0.034	-0.021	-0.005	0.037	-0.036	0.030	-0.036
Number of Extracellular Y	0.130	-0.081	-0.032	0.007	0.025	0.057	0.003	-0.050	0.024	0.040
Number of Extracellular V	0.131	-0.074	-0.021	-0.005	0.013	-0.028	0.049	0.005	-0.045	0.099
Number of Transmembrane A	-0.003	-0.022	0.024	0.029	-0.038	-0.137	-0.135	-0.058	-0.216	0.047
Number of Transmembrane R	-0.016	-0.006	0.006	-0.150	-0.169	0.100	0.013	0.013	-0.027	0.037
Number of Transmembrane N	-0.012	0.001	-0.028	0.028	-0.043	-0.010	0.067	-0.125	0.114	-0.097
Number of Transmembrane D	-0.022	-0.007	0.010	0.262	-0.025	0.218	0.023	0.191	-0.057	0.025

Number of Transmembrane C	-0.007	0.046	-0.038	0.047	0.006	0.009	-0.089	-0.078	0.146	-0.102
Number of Transmembrane Q	-0.012	-0.017	-0.027	0.029	-0.039	-0.057	-0.022	0.099	0.092	-0.061
Number of Transmembrane E	-0.019	-0.009	0.013	0.239	-0.041	0.184	0.063	0.144	-0.124	0.045
Number of Transmembrane G	-0.013	-0.054	-0.005	0.083	-0.072	-0.160	0.054	-0.140	-0.125	0.120
Number of Transmembrane H	0.015	-0.015	-0.007	0.012	-0.105	-0.002	0.082	0.021	0.106	-0.218
Number of Transmembrane I	0.016	0.018	0.017	-0.103	0.230	0.098	-0.070	-0.160	-0.055	0.034
Number of Transmembrane L	0.027	-0.022	0.091	-0.032	-0.026	-0.066	0.198	0.303	0.203	0.005
Number of Transmembrane K	-0.011	-0.019	0.019	-0.259	-0.188	0.238	0.025	0.092	-0.108	0.131
Number of Transmembrane M	-0.001	0.000	-0.128	0.037	-0.009	0.079	-0.093	0.041	0.136	0.000
Number of Transmembrane F	-0.014	-0.001	-0.031	-0.019	-0.045	0.081	0.043	-0.059	0.221	-0.020
Number of Transmembrane P	-0.001	-0.012	0.045	0.055	-0.086	0.099	0.050	-0.075	0.160	-0.094
Number of Transmembrane S	-0.007	0.006	0.009	-0.016	-0.167	0.101	0.032	-0.095	0.036	-0.076
Number of Transmembrane T	-0.002	0.010	-0.130	0.068	-0.163	-0.039	-0.062	-0.180	-0.057	-0.006
Number of Transmembrane W	0.019	-0.017	0.004	0.073	0.014	0.123	0.067	0.033	0.246	-0.063
Number of Transmembrane Y	0.009	0.000	-0.046	0.067	-0.039	0.042	-0.053	-0.115	0.008	0.068
Number of Transmembrane V	-0.011	0.007	0.059	-0.001	0.086	-0.007	-0.157	-0.045	-0.219	0.066

Number of Cytoplasmic A	0.062	0.182	0.058	0.019	-0.026	-0.043	0.014	0.047	0.006	0.015
Number of Cytoplasmic R	0.062	0.175	0.002	0.048	-0.037	-0.019	-0.016	0.085	0.068	0.032
Number of Cytoplasmic N	0.072	0.160	-0.044	0.010	0.010	0.059	-0.087	-0.032	-0.010	0.072
Number of Cytoplasmic D	0.069	0.182	0.029	-0.002	0.014	0.046	0.028	-0.015	-0.017	0.017
Number of Cytoplasmic C	0.051	0.170	-0.016	-0.006	-0.081	-0.016	0.042	-0.034	0.042	-0.049
Number of Cytoplasmic Q	0.060	0.178	0.044	0.022	-0.037	-0.011	0.047	0.009	0.051	0.045
Number of Cytoplasmic E	0.068	0.186	0.024	0.008	0.012	0.019	0.030	-0.029	-0.021	0.027
Number of Cytoplasmic G	0.064	0.174	0.067	0.024	-0.034	-0.010	0.090	-0.014	0.003	-0.014
Number of Cytoplasmic H	0.069	0.166	-0.022	0.019	-0.024	-0.021	-0.024	-0.001	0.041	0.061
Number of Cytoplasmic I	0.068	0.170	-0.085	-0.008	0.011	0.045	-0.113	0.001	-0.003	0.036
Number of Cytoplasmic L	0.067	0.186	-0.021	0.006	-0.046	-0.002	-0.008	0.045	0.003	-0.037
Number of Cytoplasmic K	0.067	0.165	-0.068	0.012	0.034	0.059	-0.132	-0.038	0.011	0.079
Number of Cytoplasmic M	0.070	0.165	-0.021	0.010	0.005	0.024	-0.062	0.004	-0.032	0.038
Number of Cytoplasmic F	0.074	0.171	-0.083	-0.013	-0.015	0.016	-0.021	-0.005	0.017	0.019
Number of Cytoplasmic P	0.053	0.168	0.090	0.010	-0.044	-0.033	0.146	-0.034	0.028	0.023
Number of Cytoplasmic S	0.061	0.185	0.024	0.006	-0.022	0.004	0.067	-0.049	0.025	0.023

Number of Cytoplasmic T	0.063	0.177	0.029	0.014	-0.022	0.008	-0.010	0.002	0.082	0.037
Number of Cytoplasmic W	0.052	0.164	-0.090	-0.015	-0.079	-0.014	0.002	0.054	0.000	-0.088
Number of Cytoplasmic Y	0.069	0.169	-0.063	0.016	0.026	0.018	-0.052	0.004	-0.042	0.033
Number of Cytoplasmic V	0.068	0.186	-0.010	0.009	-0.026	0.002	-0.017	0.015	0.024	0.013
Total number of A	0.130	-0.020	0.078	0.014	-0.065	-0.134	0.029	0.102	-0.043	0.061
Total number of R	0.142	-0.015	0.034	0.030	-0.071	-0.046	0.012	0.077	0.029	0.013
Total number of N	0.139	-0.033	-0.064	-0.003	0.028	0.079	-0.072	-0.071	0.008	-0.006
Total Number of D	0.145	-0.026	0.047	0.004	0.005	0.068	-0.017	-0.012	-0.014	-0.077
Total number of C	0.119	-0.049	0.086	0.036	-0.069	0.065	-0.083	-0.060	0.085	-0.147
Total number of Q	0.140	-0.022	0.009	0.001	-0.016	-0.058	0.081	0.044	0.018	0.041
Total number of E	0.147	0.015	0.000	0.004	0.025	0.030	0.009	-0.027	-0.034	0.005
Total number of G	0.142	-0.035	0.073	0.026	-0.050	-0.042	0.015	0.030	0.008	0.022
Total number of H	0.139	-0.018	0.009	0.006	-0.059	-0.056	-0.007	0.058	0.039	0.028
Total number of I	0.137	-0.010	-0.089	-0.026	0.072	0.096	-0.060	-0.079	-0.014	0.032
Total number of L	0.135	0.005	-0.106	-0.003	-0.048	-0.087	0.073	0.114	-0.002	-0.011
Total number of K	0.131	0.004	-0.097	-0.001	0.051	0.102	-0.122	-0.091	0.031	0.073

Total number of M	0.138	0.021	-0.050	0.005	0.008	0.052	-0.024	0.010	-0.008	0.009
Total number of F	0.138	-0.019	-0.115	-0.016	-0.008	0.015	0.015	-0.004	0.039	0.035
Total number of P	0.123	0.006	0.134	0.026	-0.077	-0.125	0.038	0.052	0.040	0.097
Total number of S	0.144	-0.008	0.010	0.009	-0.038	-0.050	0.011	-0.003	0.023	0.093
Total number of T	0.136	-0.019	0.053	0.004	0.002	-0.042	-0.056	0.003	-0.007	0.084
Total number of W	0.119	-0.038	-0.002	0.035	-0.038	0.000	0.041	-0.021	0.045	-0.059
Total number of Y	0.142	-0.007	-0.055	0.015	0.031	0.059	-0.020	-0.048	0.006	0.051
Total number of V	0.142	-0.002	-0.019	-0.002	0.009	-0.025	0.028	0.007	-0.047	0.099
pI Extracellular	-0.027	0.005	-0.059	-0.024	-0.139	-0.104	-0.024	0.049	0.151	0.210
Instability Index Extra	-0.015	0.036	0.155	0.088	-0.151	-0.106	-0.070	0.084	0.072	-0.134
Aliphatic Index Extra	0.024	0.000	-0.328	-0.018	0.080	-0.052	0.234	0.052	-0.103	0.025
Hydropathicity Extra	0.021	-0.012	-0.241	0.007	0.003	-0.122	0.232	0.074	-0.043	-0.006
pI Transmembrane	0.002	-0.020	0.015	-0.384	-0.225	0.121	0.030	-0.015	-0.002	0.021
Instability Index Trans	0.011	0.004	0.005	-0.007	0.054	0.139	0.098	0.078	0.246	-0.145
Aliphatic Index Trans	0.030	0.016	0.142	-0.157	0.278	-0.042	0.063	0.245	-0.065	0.060
Hydropathicity Trans	0.020	0.038	0.108	-0.169	0.357	-0.064	-0.054	0.153	-0.055	0.065

pI Cytoplasmic	-0.037	-0.073	-0.125	0.050	-0.032	0.023	-0.149	0.113	0.151	0.063
Instability Index Cyto	-0.027	-0.013	0.043	-0.024	-0.037	-0.020	0.205	0.073	0.096	0.081
Aliphatic Index Cyto	0.021	0.085	-0.103	-0.071	-0.065	-0.045	-0.168	0.135	-0.096	-0.277
Hydropathicity Cyto	0.024	0.090	-0.082	-0.117	-0.112	-0.110	-0.065	0.089	-0.106	-0.268
Number of Glycosolation Sites E	0.099	-0.046	-0.025	-0.013	0.050	0.015	-0.053	-0.120	0.013	0.054
Number of Phosphorylation Sites	0.062	0.183	0.033	0.013	0.006	-0.001	0.033	-0.042	0.015	0.057
Helix Content of Extracellular	0.031	-0.018	-0.259	-0.072	0.038	-0.119	0.202	0.076	-0.073	-0.058
Beta Sheet Contect Extra (%)	-0.018	0.011	-0.012	0.085	0.129	0.215	0.050	-0.110	0.007	0.126
Random Coil Content of Extra (%)	-0.019	0.010	0.290	0.010	-0.145	-0.044	-0.259	0.006	0.074	-0.038
Helix Content of Cytoplasmic (%)	0.031	0.045	-0.034	0.035	0.066	0.001	-0.202	0.193	-0.157	-0.050
Beta Sheet Content of Cyto (%)	-0.008	-0.033	-0.142	-0.088	0.006	0.062	-0.011	0.077	0.159	-0.068

Random Coil Content of Cyto (%)	-0.029	-0.027	0.145	0.025	-0.080	-0.047	0.238	-0.277	0.059	0.109
Extracellular Disorder Average	-0.033	-0.001	0.218	-0.013	0.001	-0.056	-0.118	0.059	-0.153	0.152
Extracellular Disorder St Dev	-0.006	0.018	0.186	-0.029	-0.082	-0.098	-0.169	0.142	0.106	-0.021
Transmembrane Disorder Avg	-0.015	-0.020	-0.067	0.166	-0.282	-0.105	-0.003	-0.119	-0.245	-0.052
Transmembrane Disorder St Dev	-0.011	-0.019	-0.080	0.166	-0.265	-0.105	0.024	-0.108	-0.233	-0.020
Cytoplasmic Disorder Avg	-0.016	-0.052	0.226	0.104	0.056	0.017	0.162	-0.103	0.078	0.241
Cytoplasmic Disorder St Dev	0.044	0.112	0.103	-0.021	-0.012	-0.094	0.089	-0.020	-0.064	-0.081

Appendix C:

Wilcoxon Ranked-Sum Test for Cytoplasmic Amino Acid Count for Enzymes and Non-enzymes for Type I CD Proteins

Appendix C: Wilcoxon Ranked-Sum Test for Cytoplasmic Amino Acid Count for enzymes and non-enzymes for Type I CD proteins. The two populations are significantly different as the p-value is less than critical value of 0.05.

	Enzyme	Non-enzymes
Number (n)	26	218
Median Cytoplasmic Count	424	60
W-value	671	
p-value	2.054e^{-10}	