Statistical Analysis of Sequence Characteristics of Single Transmembrane Cluster of Differentiation Proteins: A Study of Functional Relevance

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 2011

Ashlie Kiel Patterson

BS, Valdosta State University, 2007

FAIR USE

DUPLICATION

I authorize the Head of Interlibrary Loan or the Head of Archives at the Odum Library at Valdosta State University to arrange for duplication of this thesis for educational or scholarly purposes when so requested by a library user. The duplication shall be at the user's expense.

Signature_____

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

Signature_____

ABSTRACT

Cluster of Differentiation (CD) proteins are human white blood cell markers involved in immune reactions. They are used as targets for immunological studies, diagnostic techniques of disease states as well as utilized in therapeutic applications for cancers and other maladies of the body. Analysis of sequence characteristics of CD proteins has enabled the identification of functional trends that may be useful in understanding their roles within the immune system. Through statistical analysis, Principal components analysis (PCA) and K-means clustering, 36 Type II CD proteins were analyzed by one hundred twenty-four biochemical and biophysical properties derived from amino acid sequences alone. The analysis revealed these 36 CD proteins can effectively be grouped into two distinct known functional groups. In addition, possible unknown functional properties of CD71 have been identified specifically. This research shows the potential for a quantitative approach in proteomics to unlock predictive capabilities. It is believed that upon the basis of this analytical technique, functional properties may be predicted of currently undiscovered CD proteins as well as known, poorly characterized CD proteins.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

To my advisor Dr. Kang, thank you for seeing the potential in me I may not have recognized myself. Thank you for the countless numbers of hours you have dedicated to my success and the words of wisdom you have shared with me over my graduate career.

To Dr. Grove and Dr. Gosnell, thank you for the time and guidance you have given me through my graduate journey and thank you for always having an open door for my impromptu meetings.

To the department and my graduate family, thank you for the support and encouragement these past few years, you have all made my time here memorable; I extend to you my deepest appreciation.

To my closest friends and family, thank you for the sanity you have lent me, for the hours you spent listening to my hopes and fears, and for standing with me still. I love you all.

This thesis, "Statistical Analysis of Sequence Characteristics of Single Transmembrane Cluster of Differentiation Proteins: A Study of Functional Relevance," by Ashlie Kiel Patterson, is approved by:

Major Professor

Jonghoon Kang, Ph.D.
Assistant Professor of Biology

Committee Member

Theresa Grove, Ph.D.
Assistant Professor of Biology

Donna Gosnell, Ph.D.
Associate Professor of Chemistry

Dean of the Graduate School

Alfred F. Fuciarelli, Ph.D.
Professor of Chemistry

Chapter I

INTRODUCTION

Human cell surface proteins are the body's main players in immune function and cell-cell interactions (Zola & Swart, 2003). They play a pivotal role in the immune system's ability to fight foreign invaders, pathogens, and parasites. Research into leukocyte cell surface proteins (white blood cell proteins) has been fundamental in the development of drugs against disease states as well as autoimmune diseases (Zola, 2006). Key to the discovery of surface proteins has been the development of monoclonal antibodies (mAb) that recognize cell surface receptors (or antigens). Several important cell surface proteins are well known and include Cluster of Differentiation (CD)20, a regulator of B cell activation and progression (Zola et al., 2007) and CD4, the central receptor of the acquired immunodeficiency virus HIV (Dalgleish et al., 1984). However, there is expected to be a large number of unknown or undiscovered surface proteins (Zola and Swart, 2003) that would be key to the understanding of disease states and cancers. Not only is the discovery of these unknown cell surface proteins important, the functional significance of these proteins must be understood in order for them to be of assistance in the treatment of immune disorders. Functional relevance of proteins is however not an easy thing to come by; often it is difficult or impossible to determine the active conformation of a protein. This is especially difficult for membrane proteins due to hydrophobic associations with the lipid bilayer. Therefore, alternative inventive techniques must be devised to elucidate the functional relevance of membrane-bound

leukocyte cell-surface proteins. The primary sequence of a protein determines several key characteristics such as probable location within a cell, protein size, and secondary structure can be used to estimate other key characteristics such as average hydrophobicity, helix and sheet propensity (in secondary structure), aliphatic index, and stability index. Based on the amino acid sequence of a protein one can determine several key attributes of a membrane-bound protein without the complication of crystallization (Alberts et al., 1994). Additionally, statistical techniques may be applied to primary sequence characteristics to further predict the functional properties of proteins when in comparison with proteins of known function (Blom et al., 1999). The primary objective of this research is to show that protein function prediction can be obtained through multivariate statistical techniques, specifically principal component analysis and K-means clustering, by way of biochemical properties of a protein primary sequence; (i.e., characteristics determined from amino acid content of a cell surface protein).

Membrane Proteins

The term membrane protein usually refers to any protein associated with the lipid bilayer of a cell. There are many different ways a protein can be associated with the cell membrane which can be grouped into three distinct categories: integral membrane proteins, peripheral membrane proteins, and lipid-anchored proteins (Alberts et al., 1994).

Integral membrane proteins are those which span the entire cell membrane. Integral membrane proteins are characteristically amphipathic having both hydrophilic and hydrophobic regions. Due to the hydrophobic nature of the lipid bilayer of the cell, the membrane spanning portion of the protein tends to be largely hydrophobic resulting

in hydrophobic interactions between the lipid bilayer and the transmembrane region of the protein. The hydrophilic portions of the membrane spanning protein are generally held at the N- and C-terminals of the protein and interact with the extracellular or intracellular environment. There are four types of integral or transmembrane protein classifications, Types I-IV (Alberts et al., 1994). Each typically has hydrophobic regions with affinity for the cell membrane and hydrophilic regions on one or both sides of the lipid bilayer.

Type I integral membrane proteins are single-pass proteins oriented with the N-terminus exposed to the extracellular environment. Their N-terminal signal sequence is cleaved prior to transport to the cell surface. Type II integral membrane proteins are also single-pass with the N-terminus exposed to the cytoplasmic environment. No cleavage of the N-terminus occurs, and these cytoplasmic regions are generally short (Zola et al., 2007).

Type I and Type II membrane proteins have similar membrane associations and geometry and are the most simple in orientation. They consist of hydrophilic regions (cytoplasmic and extracellular) with a hydrophobic region spanning the cell membrane. The cell-membrane spanning region is usually oriented as an alpha helix interacting with the hydrophobic lipid bilayer. The secondary structure of the hydrophilic region is determined by the amino acid sequence and may be fashioned in domains which are combinations of secondary structures (e.g., alpha helixes or beta sheets) or short sequences.

Type III proteins are multipass proteins that have generally between two and twenty hydrophobic membrane spanning regions. All previously mentioned integral

membrane types are structurally similar. Their membrane spanning regions are typically composed of a hydrophobic region, approximately twenty amino acids long arranged in an alpha-helical conformation. This structural conformation is different than that of the final type of integral membrane proteins, Type IV.

Type IV integral membrane proteins are similar to Type III membrane proteins in that the protein spans the membrane several times. However, unlike all previous mentioned types, Type IV proteins span the membrane by arrangement of multiple beta sheets formed by the wrapping of the polypeptides passing through the membrane region and formation of a closed beta sheet termed a beta barrel. Type IV proteins associate with the membrane through hydrophobic interactions of the exterior of the beta barrel with hydrophilic residues inside the beta barrel. The pore through the membrane is often used for large molecule transport across the membrane of the cell. Integral membrane proteins often form oligomers with other integral or peripheral membrane proteins.

Peripheral membrane proteins do not span the membrane of the cell but are attached to the cell surface via different interactions. Because they lack hydrophobic regions, or are hydrophilic in nature, they cannot be inserted into the membrane and are bonded through weak electrostatic interactions or hydrogen bonds with the polar head of the lipid bilayer. Peripheral membrane proteins can also be non-covalently attached to other membrane proteins present on the lipid bilayer.

The final category of membrane proteins are the lipid-anchored membrane proteins. Lipid-anchored proteins are hydrophilic proteins present on either side of the lipid bilayer. They are covalently bound to lipid molecules of the cell membrane. There is a distinct difference in the nature of the covalent interactions of extracellular and

cytoplasmic lipid-anchored proteins. Those bound to the cytoplasmic surface of the cell membrane are typically covalently linked to a fatty acid or an isoprenyl group. Lipid-anchored proteins bound to the extracellular surface of the membrane are covalently linked to a glycosylphosphatidylinositol (GPI) glycolipid.

This research covers the known leukocyte cell surface proteins, specifically designated CD proteins, that span the cell membrane only once, which are exclusively Type II proteins. CD proteins are used as biological markers of a variety of leukocyte cells (Zola et al., 2007) and have been under investigation since the discovery of monoclonal antibodies (mAb) in the early 1980s.

Monoclonal Antibodies

Since the discovery of monoclonal antibodies in 1975 by Köhler and Milstein, the field of biomedical application has expanded rapidly (Köhler and Milstein, 1975). Traditional antibodies that are produced by B cells as an immune response to foreign pathogens are polyclonal and act against several epitopes of the antigen (Sompayrac, 2008). However, monoclonal antibodies (mAb) are developed to be target-specific for a single epitope of an antigen. The conventional method of antibody development produces polyclonal antibodies in reaction to immunization by an antigen typically in a mammalian host; mouse, goat, horse, pig, or rabbit. Development of mAbs begins with the introduction of an antigen to a mammalian host. However, once an immune reaction has taken place (i.e., antibodies have been produced against the foreign antigen) blood serum is drawn and polyclonal antibodies are isolated. These antibodies are specific for the antigen but not epitope-specific. Further testing is conducted to isolate the antibody with the greatest affinity for the antigen, and that antibody is clonally-developed into a mAb,

specific to a unique epitope on the original antigen. In order to prevent the human body from recognizing the mAb as a foreign invader a chimeric antibody is engineered that contains binding regions developed from human deoxyribonucleic acid (DNA).

The development of this technique has greatly improved the accuracy and efficiency of medical research diagnostic and treatment. Monoclonal antibodies have led to improved monitoring of organ graft and transplant rejection; recognition of disease states such as chlamydia, herpes, and hepatitis; and identification of infectious agents such as rapidly mutating viruses like influenza (Rieger, 1987). Additionally, mAbs have led to more sensitive detection of tumors and malignant cells (Rieger, 1987), and have been used in the discovery of leukocyte cell surface proteins (Zola & Swart, 2003). The large number of proteins discovered in this manner has lead to a problematic situation of cataloging while avoiding dual identification because several antibodies may be reactive to the same antigen and misidentification (Wilkins et al., 2006). Creation of large databases to catalog and organize this information has been the answer to this dilemma (Wilkins et al., 2006).

Bioinformatics, Databases, and Proteomics

Bioinformatics is the marriage of science and computers (Spengler, 2000). The use of computers and large databases to hold large collections of biological data is immensely useful in the study and practice of all areas of science and medicine. Databases make available a seemingly never-ending flow of information. Since the establishment of the human genome project (Adams et al., 1991), genomic data has been collected and indexed on Web-accessible servers available to the scientific community

and general public. From these databases, scientists have been able to develop larger databases that include research into gene expression, gene products, and related topics.

With this availability of large amounts of data comes the responsibility to properly organize and keep information current with new discoveries and amendments to details published. Some of the leading bioinformatics databases include GenBank (genes), UniProt (proteins), and PubMed (scholarly publications made in the field of medicine). These three databases are the most popular in the field of science with several smaller databases in existence that are administered by business or industrial organizations. It is often the case that several databases will be in collaboration to provide the most current, up to date, and accurate information available. For example, the Swiss Institute of Bioinformatics (SIB), a bioinformatics and proteins database, works in collaboration with the European Bioinformatics Institute (EBI) to form the Swiss-Prot database. Swiss-Prot in addition works with UniProt to provide a manually annotated and reviewed protein database available free of charge to the public. It is the use of these databases that has led to new discoveries by way of cross reference and comparisons with advances in the understanding of disease states (Tang et al., 2009). It has been proposed that the use of these bioinformatics tools hold the future for development in the field of immunology (Elder, Thompson & Kang, 2011).

Poor quality control of extensive data sets available to the scientific community is a significant hazard. Erroneous cataloging and duplications of entries can create a nightmare for the most thorough of bioinformaticists, resulting in skewed and incorrect data. Scientists who undergo proteomic studies by way of data mining must be extra careful to check and crosscheck the information obtained from the Web, even if it

appears that a database utilizes stringent controls to reduce errors. With this in mind we discuss the purpose and significance of this study.

Purpose and Significance

The primary objective of this research is to show that protein function prediction can be obtained through multivariate statistical techniques. Using data mining from reputable databases, biochemical properties and characteristics related to the primary sequence of Type II single transmembrane CD proteins were collected and recorded.

The significance of this study lies in the predictive capabilities and its application to the field of immunology. Protein function prediction is important particularly for membrane proteins that are not able to be crystallized and structurally studied. It is predicted that the number of known surface proteins is a fraction of the actual number in existence (Zola & Swart, 2003), so it is feasible that the amount of time, money, and manpower needed for the discovery of the functional relevance of newly discovered, not well understood, proteins can be greatly decreased by statistical techniques that strategically compare these less understood proteins to their structurally and functionally counterparts.

It is hypothesized that upon the basis of this analysis technique, functional properties can be predicted of currently undiscovered proteins as well as known, poorly characterized proteins. Specifically, we hypothesize that proper grouping, in terms of known function, may be achieved through PCA and K-means clustering of 36 Type II CD proteins based on 124 sequence characteristics.

A complete list of known human CD molecules was retrieved from HCDM.org and cross-referenced with UniProt.org (The UniProt Consortium, 2011) and (Jain et al.,

2009) for completeness. All 36 Type II single transmembrane proteins were chosen for this study. A complete list of CD numbers with SwissProt annotation can be found in Appendix A.

Chapter II

LITERATURE REVIEW: CLUSTER OF DIFFERENTIATION PROTEINS

The identification of a CD molecule is the responsibility of the Human Cell Differentiation Molecules organization (HCDM). Every three to five years HCDM holds a workshop, the Human Leukocyte Differentiation Antigens (HLDA) workshop, where researchers collaborate over new antibodies, and conduct blind analysis of these proteins. Through this collaborative effort, laboratories test the reactive properties of these antibodies against various cell types. After a surface molecule has been identified by at least two antibodies, a CD designation is assigned to the surface molecule (Zola & Swart, 2005). If only one antibody is responsive a 'w' is assigned with the CD designation describing (e.g., CDw113). The current listing of CD molecules includes numbers 1-363 as of April 2011 (UniProt.org). Protein isoforms are given separate CD designations and listed alpha-numerically (e.g., CD120a, CD120b) increasing the total number of known molecules. There are 389 total defined CD molecules as listed in the UniProt - Swiss-Prot Protein Knowledgebase. CD designation is universally accepted and governed directly by HLDA workshops. The first workshop was held in 1982 and the most recent workshops were held in March of 2010.

Functional Significance

The majority of CD proteins function in signal transduction often involved with complex pathways. Additionally, although all CD proteins are expressed on leukocytes many are co-expressed on several cell types throughout the body. Leukocyte surface

molecules specifically are often used as cell markers and as targets for research, diagnosis, and therapeutic applications (Zola et al., 2007) and, because of their direct role within the immune system, are of special interest in immunology (Abbas & Lichtman, 2009). The functions of CD proteins within the immune system include acting as markers, cell adhesion molecules, receptor proteins, ligand proteins, and enzymes (Zola et al., 2007).

Surface proteins commonly act as cell markers distinguishing them from other cell types; for example, CD4+ or CD4-, designating the presence or absence, respectively, of CD4 on a T lymphocyte. CD proteins can also function as cell adhesion molecules important in the interaction of cell types. For instance, CD34 has been found experimentally to increase the adhesion ability of human-CD34 induced murine cells to bone marrow stromal layers of human origin (Healy et al., 1995), which is believed important in stem cell regulation.

Several CD proteins serve the function of receptor or ligand. For example, CD71 is an important G-protein coupled receptor (7-transmembrane protein that functions in signal transduction through G-protein activation) important in iron regulation that binds transferrin with its associated iron ions within the cell (Kucia et al., 2004). Another example of CD proteins that function as ligand and receptor are CD40 and CD40L present on dendritic cells and helper T cells, respectively, of the adaptive immune system. The interaction of these receptor and ligand proteins present on the different cell types increases immune response and duration of dendritic cell life span and battle of infection (Sompayrac, 2008).

Many CD proteins also have enzymatic activity and play a functional role within the immune system. CD10, also known as neprilysin, has enzymatic properties including cleavage of peptides at hydrophobic residues (Zola et al., 2007). CD10 and has been implicated in Alzheimer's disease through associations in degradation of amyloid beta peptides within the brain (Wang et al., 2006). CD10 plays a role in the normal metabolism of amyloid beta peptides in the brain and was found to promote accumulation of the peptide in neprilysin-gene deficient mice (Iwata et al., 2001).

Structural Aspects

The structures of these white blood cell proteins vary from single transmembrane proteins to multipass transmembrane proteins (such as G-coupled receptors) and lipid-anchored proteins (Langel et al., 2010). Most of the integral membrane CD proteins are glycoproteins that have a carbohydrate attached either through N-linked glycosylation or O-linked glycosylation (Langel et al., 2010). Several potential glycosolation sites exist on individual CD proteins due to variation in tissue type expressed and stages of activation (Zola et al., 2007). Single-pass CD proteins consist of an extracellular region, transmembrane region, and a cytoplasmic region. They may be Type I (N-terminus extracellular) or Type II (C-terminus extracellular) proteins. Type I membrane proteins have had their N-terminal signal sequence cleaved prior to transport to the cell surface. Type II protein sequences remain uncleaved and the cytoplasmic regions (N-terminus) are generally short (Zola et al., 2007). Other CD proteins are associated with the cell membrane through lipid anchors, specifically glycosylphosphatidylinisotol (GPI)-anchors. The C-terminus of a GPI-anchored protein is linked by a phosphodiester bond to the hydrophobic region on the cell membrane.

Protein domains of CD proteins also provide information about function. There have been at least twenty functional domain types identified within CD proteins (Zola et al., 2007). Protein domains have been used to classify groups of proteins into families based on representative domain types. A representative example is Immunoglobulin super-family (IgSF) which several CD proteins belong to. These protein families function in protein-protein and protein-ligand interactions through immunoglobulin(Ig)-like (or antibody-like) domains.

It is estimated that the number of CD molecules currently identified is less than half of the total number in existence (Zola et al., 2007). As experimental techniques advance it is reasoned the number of defined molecules will continue to increase. As of the 9th International Conference on HLDAs, nineteen additional CD molecules were accepted into the HLDA database (HLDA.org).

Application of CD Proteins in Immunology

There are many examples of the functional application of CD proteins in medical practice and human health (Zola, 2006). Most research conducted regarding CD proteins focus on individual proteins with very few collective studies. Extensive research on the CD4 protein and its antigen, the cellular receptor of acquired immune deficiency syndrome (AIDS), is one of many examples with direct implications to human health (Dalgleish et al., 1984). Another illustration of CD protein functional significance is the T-cell surface glycoprotein CD8, which plays an important role along with CD4 in the activation of T cells through co-stimulation of T cell receptors (TCRs), and activation of the major histocompatibility complex (MHC) (Gascoigne et al., 2010). The use of CD molecules and the antigens against them has increased the diagnostic capabilities of

several disease states including cancer (Zola, 2006). An example of identification of disease states through CD proteins is the increased expression of CD184 on cancer cells. CD184 is a chemokine receptor specific for one particular chemokine SDF-1. It has been used in drug targeting due to its increased expression in tumor cells of various types of cancer and stages of cancer (Burger & Kipps, 2006).

Chapter III

METHODOLOGY

Principal Components Analysis

Principal components analysis (PCA) is a statistical method used to reduce the dimensionality of a data set enabling a researcher to draw relevant information from a complex or confusing set of data (Jolliffe, 2002). PCA is a multivariate analysis tool the enables the researcher to examine the relationships between several variables relating to one larger target of investigation and to draw a logical conclusion using a reduced number of new variables termed principal components (PCs). PCA is used to identify patterns in data sets that are too complex at the onset and make noticeable connections of variables.

Key Concepts

PCA explains the maximum amount of variance in the data set with the fewest number of PCs which simplifies interpretation of the data. PCA works through linear algebraic calculations of mean and covariance or correlation of each variable, reducing the calculations to eigenvalues and eigenvectors. Eigenvalues associated with PCs define the amount of variation within the data set expressed by each PC, and the total number of PCs is the same as the original number of variables. For example, if the first three eigenvalues represented 90% of the variation in the data set, the first three PCs would be used to further analyze the data. The use of PCA to analyze a data set will invariably result in a loss of information (in reduction of dimensionality), but that loss is negligent

with respect to the amount of information gained (with respect to the relationships among variables).

PCA can be used on any large or complex data set when there is need for reduction in dimensionality of a data set in any field of study, including business development, socio-economics, ecology, and proteomics. There are three visual representations of PCA produced by the statistical software application defined as a Scree plot, Score plot, and Loadings plot. These three graphical outputs facilitate an understanding of the data set based on several variables.

The Scree plot of the data depicts the eigenvalues across the PCs of the analysis. Each PC has a corresponding eigenvalue, and each eigenvalue can be calculated to reveal the percent variance of the data. The ideal pattern for a scree plot is a steep curve followed by a bend and a straight line. The first few components would ideally be located along the initial steep curve of the scree plot and explain a large percentage of the total variance in the data set.

The Score plot will plot the observations (i.e., proteins) against each other based on the new coordinate system. This x- and y- coordinate system is labeled according to the score matrix that is calculated by the sample matrix multiplied by the eigenvector matrix, which is calculated by the statistical software. Eigenvectors can be calculated from coefficients assigned to each variable of the component (i.e., PC1) which is also given as an output of the statistical software. Coefficients represent the weight of the variable, or its importance within each particular PC. The larger the absolute value of a coefficient, the more important it is in constructing that PC. The Score plot is useful in identifying outliers within a data set.

The Loadings plot reveals the eigenvector value of each characteristic. Eigenvector values are the correlation numbers of each variable to each PC. The eigenvector value can be either positive or negative and can be used to rename PCs based on user-defined data correlations.

PCA in Science

An example of the PCA method applied to biology can be seen in the study conducted by Simmons-Boyce et al. (2009). The dietary effect of *Ascophyllum nodosum*, a common commercially harvested seaweed, was studied on a group of male Sprague-Dawley rats with observance of several variables: physiological effect, dosage levels needed to elicit a response in urinary profile, and possible toxic effects of the seaweed. The complexity of the research lies within the four study groups (control, 5%, 10%, 15% *A. nodosum* diet) at four time points (week 1, week 2, week 3, week 4) with three urine collection time points each day (0-4, 4-8, and 8-24 hr) compounded with expression of multiple metabolites as collected in spectral data of urine samples. The complexity of the data set makes the ability to isolate the metabolites as influenced by the research design (i.e., study group, time point, and urine collections) practically impossible without multivariate statistical analysis. Through application of PCA, the research group was able to distinguish differences in metabolites excreted over the four week period, a difference in diurnal excretion of metabolites (0-4 hr and 4-8 hr as compared to 4+ hr), as well as noted increases and decreases in metabolites of the four separate study groups.

Another example of PCA methods as applied to a biological study is found in research conducted by Koshi and Bruno (1999). In this study the researchers used existing multiple sequence alignments of fourteen transmembrane or cytoplasmic protein

families separated into datasets of secondary structure to determine the most important structural aspects of transmembrane proteins. PCA confirmed that hydrophobicity was the most important secondary structural feature, while β-branching was identified as the second most important structural feature. The complexity of the alignments compounded by each structural feature-type per variable (alpha content, beta content, etc.) as influenced by the fourteen families of transmembrane and cytoplasmic proteins made it unlikely to draw relationships within the data without reduction in the dimensionality through a multivariate data analysis such as PCA.

One final example applies PCA in the field of botany in which Rosen, Hatch and Carter (2007), investigated the taxonomy and nomenclature of a morphologically diverse species of *Eleocharis acutangula* (cyperaceae). Through PCA, six morphological characteristics concerning 198 specimens of *Eleocharis acutangula* were used to separate the widely distributed species into three subspecies. This division was based on three PCs which accounted for 87.3% of the data. The outcome of this research led to the recommendation that this species of plant should be categorized into subspecies as opposed to separate taxa in the face of a large range of distributions and a diverse range of vegetative character.

K-Means Clustering

The basis of cluster analysis is the grouping of similar data points in close proximity or clusters with dissimilar data points belonging to different clusters. K-means clustering (Hartigan, 1975) works on this basic idea with the addition of mathematical calculation of sum of error squared. This statistical technique is used to find correlations within non-hierarchical data set (non-ranked data). It is best utilized when the data set

contains similar observations, grouping is unknown, but an ideal starting point for clustering is known. Standardization of variables (to set all data points equal) must be performed prior to analysis. This will set the centroid of the clusters at zero. In the simplest of explanations the distance of clustering points represents the mean distance from the centroid, which was ideally set at zero after standardization of values. K-means can be applied two ways. If known, the ideal centroid (or centroids) may be defined and input into the software for analysis. Conversely, the researcher can choose the number of centroids desired, and the software will continuously rearrange the data points until the best fit is obtained based on measurement of means.

The relatedness of data points to each other as calculated by the sum of error squared in relation to a decided number of centroids can be found from K-means clustering. Each data point belongs to one centroid only. Error in this technique lies in the choice of centroids. The ideal centroid will correctly cluster the data into distinct groups while improperly assigned centroids will skew the data. Knowing as much as possible about the data set is key to proper K-means cluster analysis.

K-Means in Science

An example of K-means clustering as applied to biology is a study conducted by Birnbaum et al., (2003) in which the expression of over 22,000 genes of the *Arabidopsis* root was mapped to 15 different zones with correspondence to different cell types. Based on PCA, the research group assigned K-means clustering of 5,717 differentially expressed genes into eight groups (number of centroids desired). The resultant clustering of the differentially expressed genes about eight cluster centroids further supported the results previously found for expression patterns.

Retrieval Methods

Sequence Retrieval

Sequence information of single transmembrane proteins was retrieved from UniProt protein knowledge database (UniProtKB). Specifically, only SwissProt manually annotated and non-redundant protein sequences were utilized. Data mining of sequence characteristics was out sourced from SwissProt via links to various databases as indicated.

Retrieval of Region Lengths

Total number of amino acids was collected and compiled for each protein region: extracellular, transmembrane, and cytoplasmic from [Sequence annotation (Features)] resource at UniProt.org. Total lengths recorded included propeptide sequences and excluded signal peptides.

Retrieval of Charges

Total number of charged residues per region was retrieved through ExPAsy Bioinformatics Resource Portal of the SIB Swiss Institute of Bioinformatics (Gasteiger et al., 2003) using the tool ProtParam which computes various physico-chemical properties that can be deduced from a protein sequence (Gasteiger et al., 2005). Absolute charges were calculated manually from sequence retrieval of total positive and negative charged residues per region.

Retrieval of Amino Acid Composition

The standard one letter abbreviations of twenty amino acids were used to record the number of amino acid type per region length. These were recorded from the ProtParam tool of ExPAsy Bioinformatics Resource Portal. Percentage composition of

amino acid was calculated manually from the total number of amino acids per region as divided by number of each individual amino acid group. Amino acid composition was recorded for each region length; extracellular, transmembrane and cytoplasmic with total amino acid composition being the culmination of the three.

Retrieval of Theoretical pI

Theoretical isoelectric point (pI) is the pH at which no charge is carried by an amino acid. Theoretical pI was also retrieved from the ProtParam tool of ExPAsy Bioinformatics Resource Portal and recorded for the following regions: extracellular, transmembrane, and cytoplasmic.

Retrieval of Instability Index

Instability index is a numerical value used to determine if a molecule is stable as determined within a test tube environment. Statistical analysis of 12 unstable and 32 stable proteins has revealed that there are certain dipeptides, the occurence of which is significantly different in the unstable proteins compared with those in the stable ones (Guruprasad et al., 1990). Values less than forty are decidedly stable. The stability index for regions extracellular, transmembrane and cytoplasmic were retrieved from the ProtParam tool of the ExPAsy Bioinformatics Resource Portal.

Retrieval of Aliphatic Index

The aliphatic index is the relative volume of a protein occupied by aliphatic side chains (e.g., alanine, valine, isoleucine, and leucine). These values were calculated using the ProtParam tool of the ExPAsy Bioinformatics Resource Portal.

Retrieval of Hydropathicity

Hydropathicity is a positive or negative value of a protein detailing whether the molecule is soluble in water. Positive values indicate large numbers of hydrophobic residues; negative numbers indicate an abundance of hydrophilic residues. Hydropathicity values were recorded for extracellular, transmembrane and cytoplasmic regions as determined by [ProtParameters] of the ExPAsy Bioinformatics Resource Portal.

Retrieval of the Number of Glycosylation Sites Extracellular

The number of glycosylation sites per extracellular region of the proteins was obtained from the UniProt database. Numerical values of glycosylation residue were then cross-checked to ensure its inclusion within the extracellular region specifically.

Retrieval of the Number of Phosphorylation Sites Cytoplasmic

Phosphorylation sites of the cytoplasmic region of the proteins were retrieved through the ExPAsy tool NetPhos-Prediction of Serine, Threonine and Tyrosine Phosphorylation sites in eukaryotic proteins (Blom et al., 1999)

Retrieval of Helix, Beta Sheet, and Random Coil Content

Secondary structure content, recorded as a percentage, was retrieved from the GOR IV secondary structure prediction web server (Garneir et al., 1996) of the ExPASy: SIB Bioinformatics Resource Portal - Proteomics Tools Online software for protein analysis from the Swiss Institute of Bioinformatics (SIB). Helix content, beta sheet content and random coil content were recorded for extracellular and cytoplasmic regions of each protein.

Retrieval of Dynamic Average

Average dynamics were calculated for each region; extracellular, transmembrane and cytoplasmic for each protein. Disorder tendency was calculated through IUPred, a Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content of each residue (Dosztányi et al., 2005). Manual calculation of total region average was recorded as well as standard deviation. Prediction type: long disorder and Output type: raw data were specified.

Statistical Methods

Statistical analyses included PCA followed by K-means clustering. Both analyses were conducted through MiniTab® (15) statistical software[1].

Principal Component Analysis

PCA was chosen from the [Stat menu] of MiniTab® under the subheading [Multivariate statistics]. All 124 variables were accounted for with the number of components to compute equaling two. Correlation matrix was chosen for the purpose of standardizing the variables. A Scree plot, Score plot, and Loadings plot were chosen for graphical output.

K-means Clustering

Cluster K-means was chosen from the [Stat menu] of MiniTab® under the subheading [Multivariate statistics]. Scores of the PCA were used as the variables for K-means clustering. An initial partition column was created specifying an initial partition of two (CD10 and CD314, respectively). The option to standardized variables was also chosen.

Chapter IV

RESULTS

Our results for PCA and K-means clustering are shown in Figures 1 through 4. PCA utilizes three graphs, a Scree Plot, Loadings Plot, and a Score Plot. K-means is shown as overlaid on the Scree Plot of PCA.

Principal Components Analysis

Two PCs were determined to account for 68.8% of the variation in the data. Eigenvalue calculations for each component are found in Appendix B. Eigenvector values for PC1 and PC2 as pertains to each protein characteristic are found in Appendix C. Scores for PCs are found in Appendix D. Three graphical outputs were generated: Scree Plot, Loadings Plot, and Score Plot. The Scree Plot of protein characteristics displays eigenvalues associated with each component; PC1 eigenvalue equaling 42.24 and PC2 eigenvalue of 12.61 (see Figure 1). The Loadings Plot shows the eigenvector value for each characteristic as pertains to PC1 and is illustrated in Figure 2. The Score Plot of the proteins illustrated in Figure 3 reveals the score of each variable (all 36 proteins) as pertains to PC1and PC2 on a two dimensional plane.

Eigenvalues are shown for each PC (Protein Characteristics). PC1 and PC2 are the first and second marker, respectively. The eigenvalue of PC1 being 42.24 and the eigenvalue of PC2 being 12.61. Eigenvalues represent the variation within the data revealing that PC1 and PC2 account for 44.2% of the variation within the data.

Loadings values of each characteristic illustrated in Figure 2 as pertains to PC1 (y-axis) with all 124 protein characteristics (x-axis).

The Score Plot of protein characteristics as applied to PC1 and PC2 are illustrated in Figure 3. Values indicated are PC scores for each observation under consideration (all 36 proteins) as each pertain to PC1 and PC2.

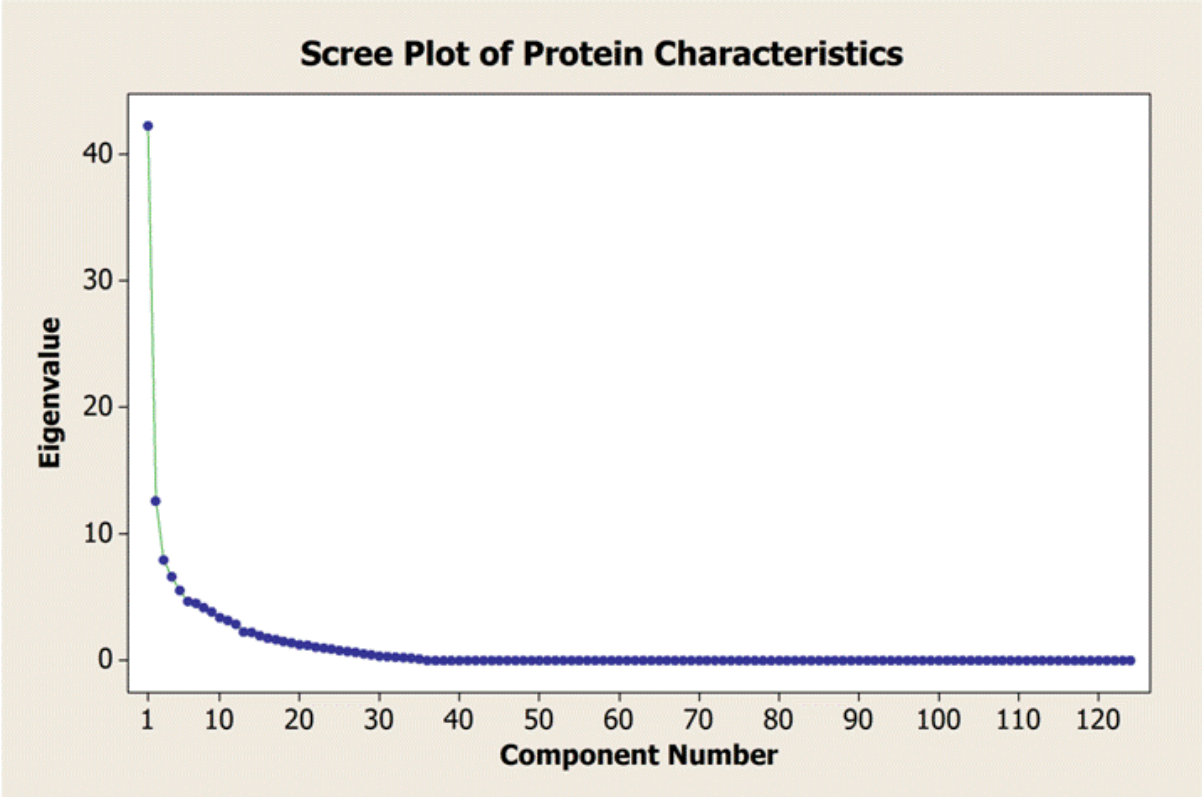

Figure 1. Scree Plot of Protein Characteristics

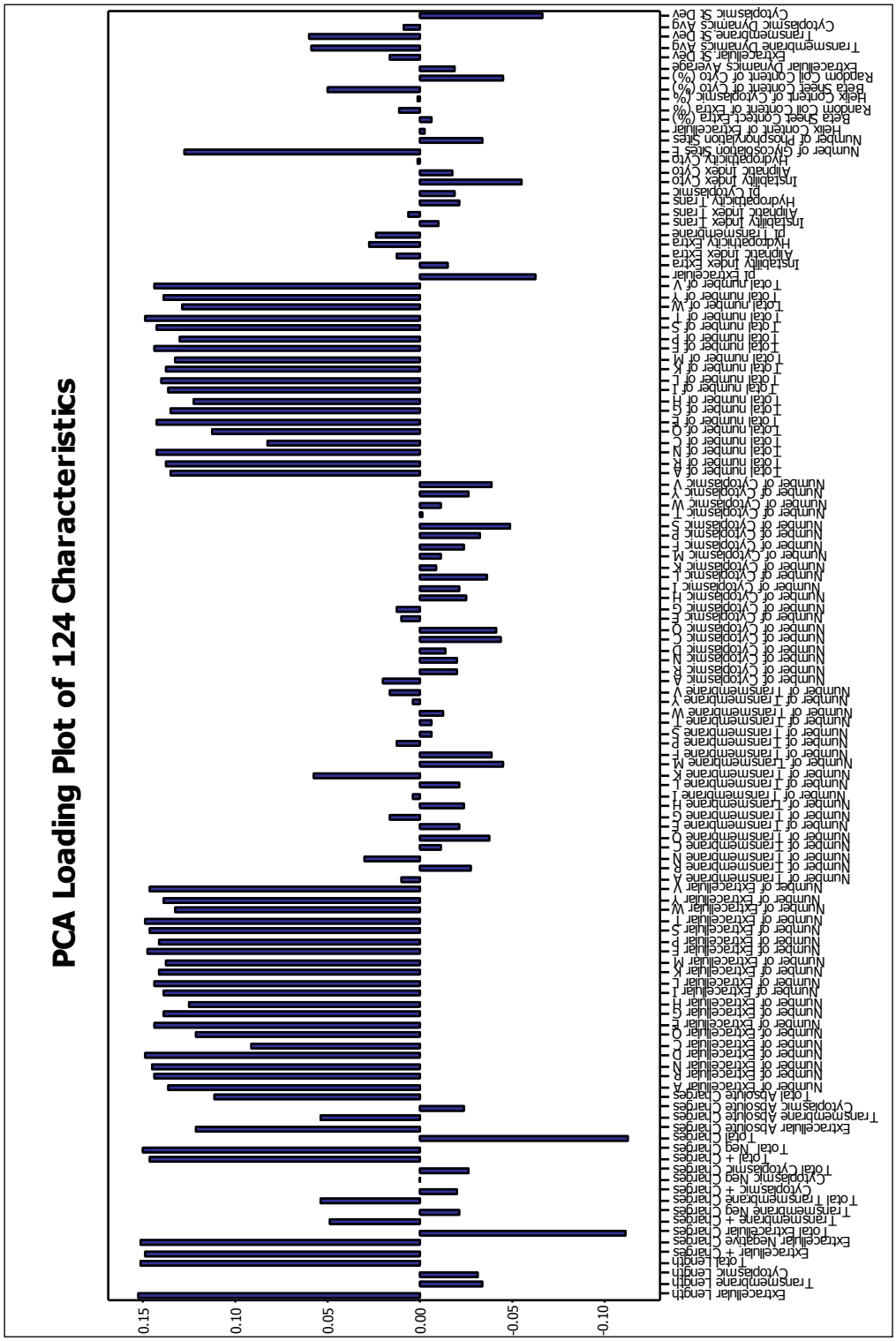

Figure 2. Loadings Plot of Protein Characteristics as applied to PC1

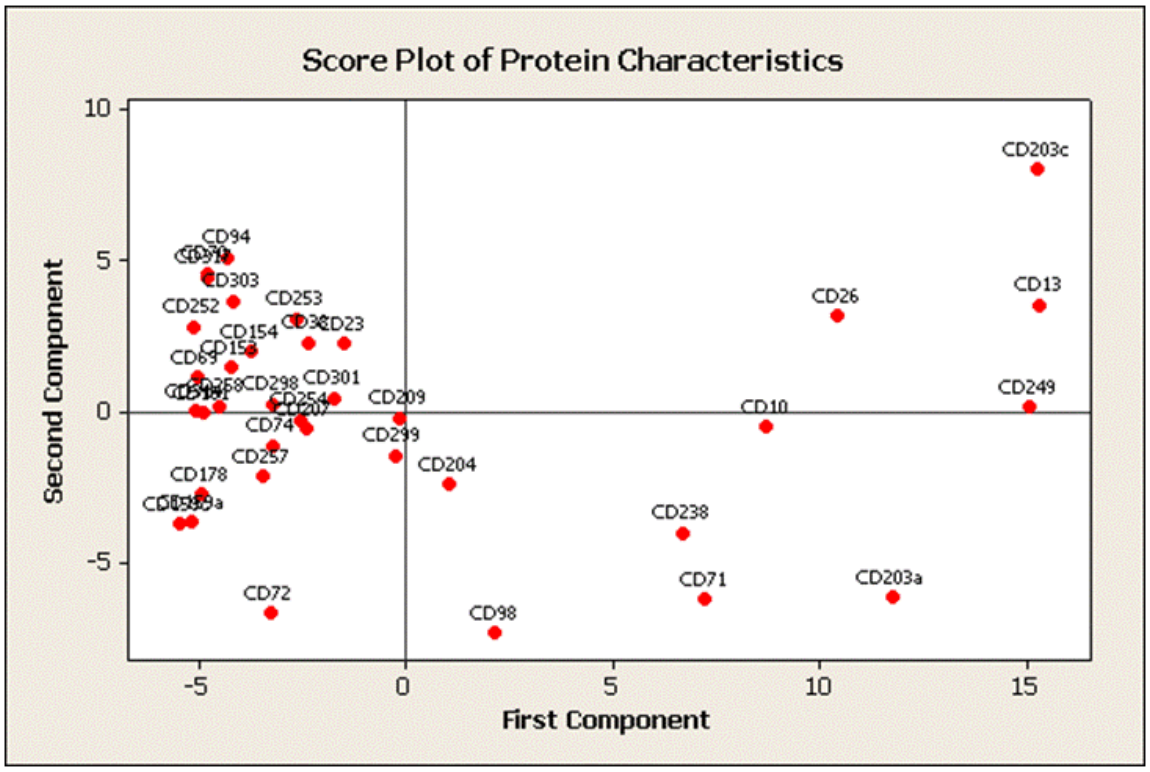

Figure 3. Score Plot of Protein Characteristics as applied to PC1 and PC2

K-Means Clustering

K-means cluster analysis groups the 36 proteins into two clusters using PC scores generated through PCA revealed in Figure 4. The functional relevance of the proteins in group one is binding such as ligand or receptor. The functional relevance of group two is enzymatic function.

K-means clustering grouped the 36 proteins/variables/observations into two distinct clusters based on score plot values generated through PCA. As indicated by the triangle (group 1) and the oval (group 2), 27 observations belong to group one containing proteins with binding properties and 9 observations belong to group two containing proteins with enzymatic properties.

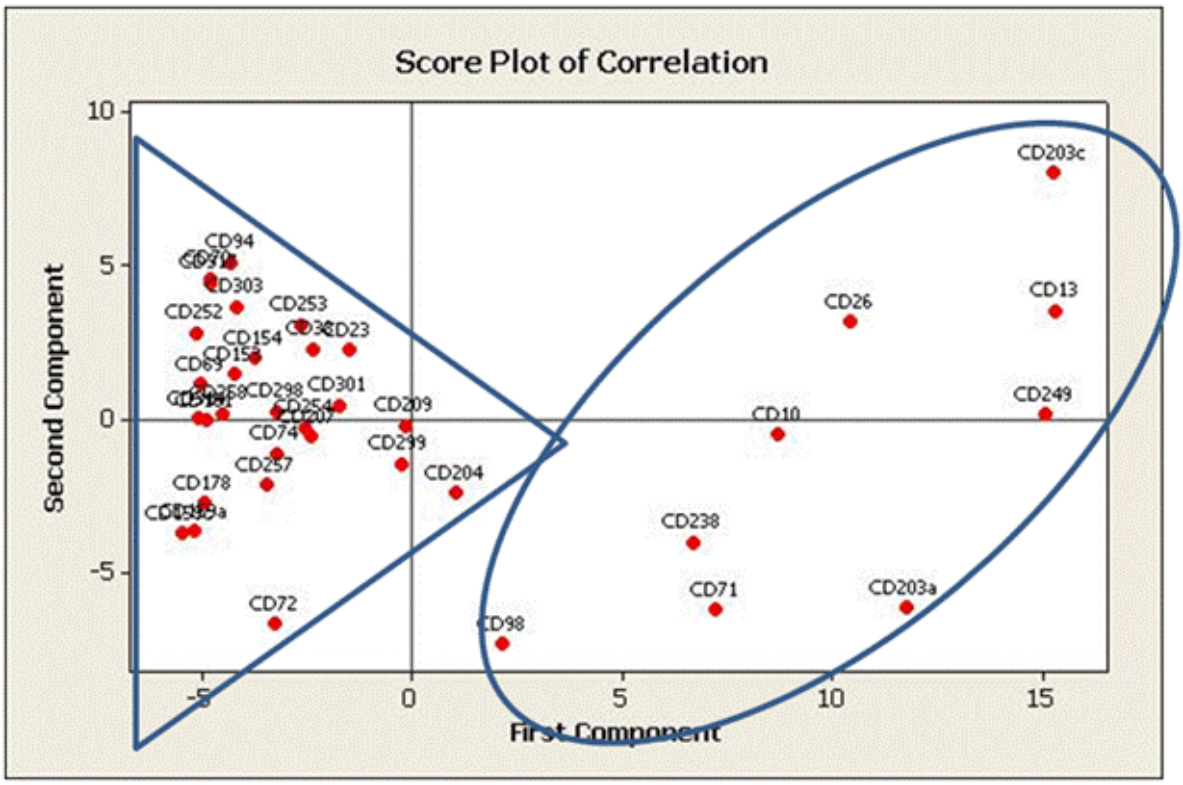

Figure 4. K-means clustering as applied to the PCA Score Plot of protein characteristics

.

Chapter V

DISCUSSION

PCA is a statistical technique useful in the analysis of multivariable data sets. PCA can adequately reduce the dimensionality of a data set by displaying correlation between observations on a two dimensional plot. The reduction in the dimensions of the data allows the researcher to draw information that was before unseen.

The complexity of our data set is defined by 124 protein characteristics as applied to thirty six Type II CD proteins. Analysis by PCA reduced that dimensionality to two main principal components that represent 44.2% of the variance in the data set (PC1 eigenvalue equaling 42.24 and PC2 eigenvalue of 12.61). An interesting discovery was found within the functional grouping of the proteins through K-means clustering. K-means clustering was divided into two groups based on two centroid points designated through MiniTab®.

The twenty-seven observations, or proteins, included within group one (see Figure 4) are known for binding function such as ligand or receptor. This diverse group of proteins have properties that include cell-cell signalling such as induction of apoptosis, activation and differentiation of immune cells, and inhibition of cytotoxicity. The nine observations, or proteins, included within group two (see Figure 4) are known for enzymatic functions that include endopeptidases, metalloproteases, exopeptidases, phosphatases, phosphodiesterases, endopeptidases, and ectoenzymes. The exception to

our finding is CD 71, which is assigned to group two but is known for its function as a receptor.

CD71, also known as transferrin receptor protein, is located on all proliferating cells but specifically on erythrocytes and leukocytes and functions in the cellular uptake of iron. Uptake of iron occurs through endocytosis of its ligand transferrin protein when it is bound to iron or IgA. Through receptor-mediated endocytosis transferrin receptor is internalized by the cell through coated pits of the cell membrane. Endocytic vesicles are formed which contain the transferrin receptor, transferrin protein, and transferrin bound iron ions. The pH of the endosome is relatively acidic (pH 5-5.5) and induces transferrin to release the bound iron within the cell. Iron deplete transferrin proteins (apotransferrin) remains bound to the transferrin receptor and recycled back to the cell membrane where the neutral pH of the extracellular environment causes the release of apotransferrin from the transferrin receptor allowing the receptor to bind another transferrin-iron complex and repeat the cycle. The entire transferrin receptor-transferrin cycle takes about 15-20 minutes**.**

CD71 plays multiple roles within the immune system. CD71 is a marker of immature proliferating T cells in the immune system as was discovered by Brekelmans et al., (1994), through the analysis of CD71 expression on fetal, neonatal, and adult thymocytes in correlation with cell size and cell cycle status. Additional down regulation of CD71 in adult T lymphocytes confirmed the hypothesis of CD71 as a marker for immature T lymphocytes.

CD71 has also been implicated the transferrin receptor in the abnormal immune response to gluten-derived peptides in celiac disease. The overexpression of CD71 in

patients with active celiac disease is believed to be a factor in triggering the immune response to gliadin peptides through IgA-CD71 mediated transport of peptides across the intestinal lamina (Matysiak-Budnik et al. 2008). As well, CD71 has been implicated in the activation of the adaptive immune system through the ability to bind to MHCII (major histocompatibility complex two) of helper T cells (Sompayrac, 2008). We hypothesize that CD71 has an enzymatic function not yet discovered since CD71 falls into the second cluster consisting of enzymatic proteins.

This research exhibits the potential for a quantitative approach in proteomics to unlock predictive capabilities. Time intensive protein studies could be supplemented by computational applications consequently reducing the amount of time invested in function prediction. It is hypothesized that upon the basis of this analysis technique, functional properties may be predicted of currently undiscovered CD proteins as well as known, poorly characterized CD proteins.

This method of predicting CD protein function through statistical analysis of sequence characteristics, specifically through the dual application of PCA and K-means clustering, has been shown dependable through the grouping of thirty-six Type II CD proteins into two distinct groups that exhibit known functional properties as based on 124 sequence characteristics. In addition to the clustering into two functional groups, a new hypothesis was developed regarding the functional properties of CD71. Future use of this methodology will hopefully enhance the understanding of functionality of newly discovered CD proteins.

REFERENCES

Abbas, A. K., Lichtman, A. H. (2009). Basic immunology: Functions and Disorders of the Immune System, (3rd ed.). Philadelphia:Saunders Elsevier.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science, 252,* 1651-1656.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J. D. (1994). Molecular Biology of the Cell, (3rd ed.). New York:Garland Publishing.

Birnbaum, K., Shasha, D. E., Wang, J. Y., Jung, J. W., Lambert, G. M., Galbraith, D. W., Benfey, P. N. (2003). A Gene Expression Map of the Arabidopsis Root. *Science, 302*, 1956-1960.

Blom, N., Gammeltoft, S., Brunak, S. (1999). Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol., 294*, 1351-1362.

Brekelmans, P., van Soest, P., Voerman, J., Platenburg, P. P., Leenen, P. J., van Ewijk, W. P. (1994).Transferrin receptor expression as a marker of immature cycling thymocytes in the mouse. *Cell Immunol., 159*, 331-339.

Burger, J. & Kipps, (2006). T. CXCR4: a key receptor in the crosstalk between tumor cells and their microenvironment. *Blood, 107*, 1761-1767.

Dalgleish, A. G., Beverley, P. C., Clapham, P. R., Crawford, D. H., Greaves, M. F., Weiss, R. A. (1984). The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature, 312*, 763-767.

Dosztányi, Z., Csizmók, 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*.

Dosztányi, Z., Csizmók, V., Tompa, P., Simon, I. (2005). The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol., 347*, 827-839.

Elder, J. F., Thompson, S. M., Kang, J. (2011). The Molecular Biology/Immunology Paradigm Extended to Bioinformatics. *J Clin Cell Immunology, 2*, 111.

Gascoigne, N. R., Zal, T., Yachi, P. P., Hoerter, J. A. (2010). Co-receptors and recognition of self at the immunological synapse. *Curr Top Microbiol Immunol., 340*, 171-189.

Garnier, J., Gibrat, J. F., Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol., 266*, 540-53.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res., 31*, 3784-3788.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server; (In) The Proteomics Protocols Handbook, (John M. Walker ed.). Humana Press pp. 571-607.

Guruprasad, K., Reddy, B. V. B., & Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng., 4*, 155-161.

Hartigan, J. A. (1975) Clustering Algorithms. John Wiley & Sons.

Healy, L., May, G., Gale, K., Grosveld, F., Greaves, M., Enver, T. (1995). The stem cell antigen CD34 functions as a regulator of hemopoietic cell adhesion. *Proc. Natl. Acad. Sci., 92*, 12240-12244.

Iwata, N., Tsubuki, S., Takaki, Y., Shirotani, K., Lu, B., Gerard, N. P., Gerard, C., Hama, E., Lee, H. J., Saido, T. C. (2001). Metabolic Regulation of Brain Aβ by Neprilysin. *Science, 292*, 1550-1552.

Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B. E., Martin, M. J., McGarvey, P., Gasteiger, E. (2009). Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics, 10*, 136.

Jolliffe, I. T. (2002). Principal Component Analysis, (2nd ed.). New Jersey:Springer

Köhler, G. & Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature, 256*, 495-497.

Kucia, M., Jankowski, K., Reca, R., Wysoczynski, M., Bandura, L., Allendorf, D. J., Zhang, J.,Ratajczak, J., Ratajczak, M. Z. (2004). CXCR4-SDF-1 signaling, locomotion, chemotaxis and adhesion. *Journal of Molecular Histology, 35*, 233-245.

Koshi, J. & Bruno, W. (1999). Major Structural Determinants of Transmembrane Proteins Identified by Principal Components Analysis. *Proteins, 34*, 333-340.

Langel, Ü., Cravatt, B. F., Gräslund, A., Heijne, G., Land, T., Niessen, S., Zorko, M. (2010). Introduction to Peptides and Proteins, Florida:CRC Press.

Matysiak-Budnik, T., Moura, I. C., Arcos-Fajardo, M., Lebreton, C., Ménard, S., Candalh, C., Ben-Khalifa, K., Dugave, C., Tamouza, H., van Niel, G., Bouhnik,

Y., Lamarque, D., Chaussade, S., Malamut, G., Cellier, C., Cerf-Bensussan, N., Monteiro, R. C., Heyman, M. (2008). Secretory IgA mediates retrotranscytosis of intact gliadin peptides via the transferrin receptor in celiac disease. *JEM, 205*, 143-154.

Rieger, P. T. (1987). Monoclonal Antibodies. *AJN, 87*, 469-473.

Rosen, D. J., Hatch, S. L., Carter, R. (2007). Infraspecific Taxonomy and Nomenclature of *Eleocharis Actutangula* (cyperaceae). *J. Bot. Res. Inst. Texas, 1*, 875-888.

Simmons-Boyce, J. L., Purcell, S. L., Nelson, C. M., MacKinnon, S. L. (2009). Dietary Ascophyllum nodosum Increases Excretion of Tricarboxylic Acid Cycle Intermediates in Male Sprague-Dawley Rats. *J. Nutr., 139*, 1487-1494.

Sompayrac, L. (2008). How the Immune System Works, (3rd ed). Blackwell Publishing

Spengler, S. J. (2000). Bioinformatics in the Information Age. *Science, 287,* 1221-1223.

Tang, J., Tan, C. H., Oresic, M., Vidal-Puig, A. (2009). Integrating post-genomic approaches as a strategy to advance our understanding of health and disease. *Genome Medicine, 1*, 35.

The UniProt Consortium. (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res., 39*, D214-D219.

Wang, D. S., Dickson, D. W., Malter, J. S. (2006). β-Amyloid Degradation and Alzheimer's Disease. *Journal of Biomedicine and Biotechnology*, pp. 1-12.

Wilkins, M. R., Appel,R. D., Van Eyk, J. E., Chung, M. C. M., Görg, A., Hecker, M., Huber, L. A., Langen, H., Link, A. J., Paik, Y., Patterson, S. D., Pennington, S. R., Rabilloud, T., Simpson, R. J., Weiss, W., Dunn, M. (2006). Guidelines for the next 10 years of proteomics. *Proteomics, 6*, 4-8.

Zola, H. (2006). Medical Applications of Leukocyte Surface Molecules-the CD molecules. *Mol. Med., 12*, 312-316.

Zola, H. & Swart, B. (2003). Human leukocyte differentiation antigens. *Trend Immunol., 24*, 353-354.

Zola, H. & Swart, B. (2005). The human leukocytes differentiation antigens (HLDA) workshops: the evolving role of antibodies in research, diagnosis and therapy. *Cell Research, 15*, 691-694.

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

APPENDIX A:

List of CD Proteins with SwissProt Annotation and Function

APPENDIX A: List of CD Proteins with SwissProt Annotation and Function

Molecule (CD Number)	Accession Number	Function
CD10	P08473	Neutral endopeptidase cleaves biologically active peptides at hydrophobic amino acids.
CD13	P15144	Zinc metalloproteinase that trims peptides bound to MHC class II.
CD23	P06734	Involved in negative feedback regulation of IgE synthesis. Thought to play a role in endocytosis and/or phagocytosis
CD26	P27487	Serine-type exopeptidase, cleaves dipeptides (X-proline or X-alanine) from the N-termini of proteins
CD38	P28907	Can act as an ectoenzyme and cytosolic enzyme, adhesion molecule and has regulatory functions.
CD69	Q07108	In vitro: appears to have signaling functions and is involved in early stage activation of lymphocytes, monoctyes, and platelets.
CD70	P32970	Plays a role in T-cell activation, inducing proliferation of costimulated T cells and enhancing the generation of cytolytic T cells.
CD71	P02786	Mediates cellular iron uptake via internalization and recycling transferrin.
CD72	P21854	Putative ligand for CD100 and for CD5. Plays a role in B-cell proliferation and differentiation.
CD74	P04233	Regulates loading of exogenous-derived peptides onto MHC class II heterodimers.
CD94	Q13241	Functions as an inhibitory receptor.
CD98	P08195	Functions as a chaperone for amino acid transporters closely associated with actin.
CD153	P32971	Proliferation of peripheral blood T cell is costimulated by CD153. Enhances antigen-induced proliferation and cytokine production by Th0 and Th2 clones.
CD154	P29965	CD154/CD40 complex signals to CD40+ B-cells in germinal centers and is for germinal center formation as well as Ig class switching.
CD159a	P26715	In association with CD94 can bind MHC class I antigen on target cells and inhibit NK cell-mediated cytotoxicity.
CD159c	P26717	In association with CD94 can bind MHC class I antigen on target cells and inhibit NK cell-mediated cytotoxicity. CD159c is absent in 4% of healthy individuals as a result of a homozygous deletion mutant.
CD161	Q12918	Monoclonal antibodies against CD161 have been reported to either augment or inhibit NK cell-mediated cytotoxicity against certain Fc-receptor bearing targets and induce thymocyte proliferation

CD178	P48023	Induces apoptosis in cells expressing CD95.
CD203a	P22413	Involved primarily in ATP hydrolysis at the plasma membrane
CD203c	O14638	Capable of cleaving a variety of phosphodiester and phosphosulfate bonds present in deoxynucleotides, nucleotide sugars, and NAD.
CD204	P21757	Mediates the uptake of a wide variety of negatively charged macromolecules by mononuclear phagocytes.
CD207	Q9UJ71	High level expression of CD207 induces membrane superimposition and zippering, which form pentilaminar organelles known as Birbeck granules (BG). Ligand binding results in rapid internalization of CD207.
CD209	Q9NNX6	Binds the mannose-like carbohydrate of ICAM-2 (CD102) and ICAM-3 (CD50).
CD238	P23276	Acts as a zinc endopeptidase that cleaves a large intermediate precursor of endothelin-3 to its bioactive form.
CD249	Q07075	Ectoenzyme that catalyses the release of N-terminal glutamate (and to a lesser extent aspartate) from a peptide.
CD252	P23510	Cytokine that can act as a costimulator through interaction with its ligand on T-cells.
CD253	P50591	Principal activity seems to be to induce apoptosis. Knockout studies suggest plays a role in tumor surveillance by immune cells.
CD254	O14788	Principal factor stimulating osteoclast activation and differentiation.
CD257	Q9Y275	Primarily a B-cell survival, activation, and differentiation factor.
CD258	O43557	Can trigger apoptosis as well as cell activation, by signaling through TRAF3.
CD298	P54709	N/K ATPase involved in Na/K transport.
CD299	Q9H2X3	Probably a pathogen recognition receptor involved in peripheral immune surveillance in the liver.
CD301	Q8IUN9	On immature dendritic cells is involved in receptor-mediated endocytosis of glycosylated proteins.
CD303	Q8WTT0	C-type lectin of plasmacytoid dendritic cells. Invitro studies show has dual function of both capturing and targeting antigen for processing and presentation to T cells.
CD314	P26718	Recognizes transformed or virus-infected cells expressing CD314 ligands, activating cell-mediated killing.
CD317	Q10589	Suggested that plays role in interaction between lymphocytes and bone stromal cells.

as adapted from Zola et al., 2007.

APPENDIX B:

Definition of Protein Characteristics Used

Appendix B: Definition of Protein Characteristics Used

Region Lengths	Total count of amino acids per region: extracellular, transmembrane, cytoplasmic, and total length.
Charges	Calculated charges based on amino acid properties for each region and total length.
Absolute Charge	Absolute value of each region charge and total charge.
Number of Amino Acids	Amino acid content per region and total length. Standard abbreviations used.
Isoelectric point (pI)	pH at which a protein carries no charge. Calculated for extracellular, transmembrane and cytoplasmic regions.
Instability Index	Numerical value used to determine if a molecule is stable in a test tube environment [<40 (stable)] Calculated for extracellular, transmembrane and cytoplasmic regions.
Aliphatic Index	Relative volume of a protein occupied by aliphatic side chains (e.g. alanine, valine, leucine, and isoleucine). Calculated for extracellular, transmembrane and cytoplasmic regions.
Hydropathicity	Solubility of a protein (+ values hydrophobic; - values hydrophilic). Calculated for extracellular, transmembrane and cytoplasmic regions.
Glycosolation sites	Calculated for extracellular region only.
Phosphorylation sites	Calculated for cytoplasmic region only.
Secondary structure content	Helix, beta sheet, and random coil content calculated for extracellular, transmembrane and cytoplasmic regions.
Dynamic Average	Also known as degree of disorder. Prediction of intrinsically unstructured regions of proteins based on estimated energy content of each residue. Calculated for extracellular, transmembrane and cytoplasmic regions.

APPENDIX C:

Eigenvalues of First 10 Components i.e. Protein Characteristic

APPENDIX C: Eigenvalues of first 10 components i.e. protein characteristic

	Eigenvalue	Porportion	Cumulative
PC1	42.241	0.341	0.341
PC2	12.614	0.102	0.442
PC3	7.945	0.064	0.506
PC4	6.619	0.053	0.560
PC5	5.504	0.044	0.604
PC6	4.646	0.037	0.642
PC7	4.522	0.036	0.678
PC8	4.177	0.034	0.712
PC9	3.826	0.031	0.743
PC10	3.391	0.027	0.770

APPENDIX D:

Eigenvector Values for PC1 and PC2 for Each Protein Characteristic

APPENDIX D: Eigenvector values for PC1 and PC2 as pertains to each protein characteristic

Variable	PC1	PC2
Extracellular Length	0.153	-0.020
Transmembrane Length	-0.034	0.047
Cytoplasmic Length	-0.032	-0.255
Total Length	0.152	-0.044
Extracellular + Charges	0.150	-0.037
Extracellular Negative Charges	0.151	-0.011
Total Extracellular Charges	-0.113	-0.071
Transmembrane + Charges	0.049	0.132
Transmembrane Neg Charges	-0.022	-0.050
Total Transmembrane Charges	0.053	0.142
Cytoplasmic + Charges	-0.021	-0.219
Cytoplasmic Neg Charges	-0.001	-0.230
Total Cytoplasmic Charges	-0.027	0.020
Total + Charges	0.147	-0.063
Total Neg Charges	0.150	-0.038
Total Charges	-0.114	-0.049
Extracellular Absolute Charges	0.121	0.044
Transmembrane Absolute Charges	0.053	0.142
Cytoplasmic Absolute Charges	-0.025	-0.060
Total Absolute Charges	0.111	0.007
Number of Extracellular A	0.136	-0.037
Number of Extracellular R	0.144	-0.030
Number of Extracellular N	0.145	0.013
Number of Extracellular D	0.149	-0.002
Number of Extracellular C	0.091	0.010
Number of Extracellular Q	0.121	-0.042
Number of Extracellular E	0.145	-0.018
Number of Extracellular G	0.139	-0.052
Number of Extracellular H	0.125	-0.008
Number of Extracellular I	0.139	0.006
Number of Extracellular L	0.144	-0.058
Number of Extracellular K	0.141	-0.040
Number of Extracellular M	0.138	0.024
Number of Extracellular F	0.147	-0.024
Number of Extracellular P	0.141	-0.018
Number of Extracellular S	0.146	-0.039
Number of Extracellular T	0.149	-0.001
Number of Extracellular W	0.133	0.014
Number of Extracellular Y	0.139	0.018

Number of Extracellular V	0.147	0.001
Number of Transmembrane A	0.010	0.033
Number of Transmembrane R	-0.028	0.045
Number of Transmembrane N	0.030	0.089
Number of Transmembrane C	-0.012	-0.002
Number of Transmembrane Q	-0.039	0.032
Number of Transmembrane E	-0.022	-0.050
Number of Transmembrane G	0.016	-0.069
Number of Transmembrane H	-0.024	0.058
Number of Transmembrane I	0.003	0.022
Number of Transmembrane L	-0.022	-0.041
Number of Transmembrane K	0.057	0.122
Number of Transmembrane M	-0.046	-0.047
Number of Transmembrane F	-0.039	-0.041
Number of Transmembrane P	0.012	0.115
Number of Transmembrane S	-0.007	0.043
Number of Transmembrane T	-0.007	0.006
Number of Transmembrane W	-0.013	-0.013
Number of Transmembrane Y	0.003	-0.017
Number of Transmembrane V	0.016	0.031
Number of Cytoplasmic A	0.020	-0.188
Number of Cytoplasmic R	-0.020	-0.140
Number of Cytoplasmic N	-0.021	-0.148
Number of Cytoplasmic D	-0.015	-0.208
Number of Cytoplasmic C	-0.044	-0.043
Number of Cytoplasmic Q	-0.042	-0.139
Number of Cytoplasmic E	0.009	-0.184
Number of Cytoplasmic G	0.013	-0.181
Number of Cytoplasmic H	-0.026	-0.064
Number of Cytoplasmic I	-0.022	-0.137
Number of Cytoplasmic L	-0.037	-0.191
Number of Cytoplasmic K	-0.010	-0.182
Number of Cytoplasmic M	-0.012	-0.105
Number of Cytoplasmic F	-0.024	-0.080
Number of Cytoplasmic P	-0.033	-0.103
Number of Cytoplasmic S	-0.049	-0.191
Number of Cytoplasmic T	-0.001	-0.136
Number of Cytoplasmic W	-0.012	-0.062
Number of Cytoplasmic Y	-0.027	-0.079
Number of Cytoplasmic V	-0.040	-0.148
Total number of A	0.136	-0.072
Total number of R	0.138	-0.056
Total number of N	0.143	-0.001
Total number of C	0.082	0.004

Total number of Q	0.113	-0.066
Total number of E	0.143	-0.045
Total number of G	0.135	-0.087
Total number of H	0.123	-0.014
Total number of I	0.136	-0.002
Total number of L	0.140	-0.080
Total number of K	0.138	-0.061
Total number of M	0.133	0.004
Total number of F	0.144	-0.035
Total number of P	0.130	-0.050
Total number of S	0.143	-0.067
Total number of T	0.148	-0.013
Total number of W	0.129	0.005
Total number of Y	0.138	0.012
Total number of V	0.144	-0.014
pI Extracelular	-0.064	-0.053
Instability Index Extra	-0.016	0.012
Aliphatic Index Extra	0.012	-0.035
Hydropathicity Extra	0.026	-0.025
pI Transmembrane	0.023	0.156
Instability Index Trans	-0.011	0.066
Aliphatic Index Trans	0.006	-0.031
Hydropathicity Trans	-0.022	-0.098
pI Cytoplasmic	-0.019	0.003
Instability Index Cyto	-0.056	0.017
Aliphatic Index Cyto	-0.018	-0.040
Hydropathicity Cyto	0.000	0.060
Number of Glycosolation Sites E	0.128	0.013
Number of Phosphorylation Sites	-0.035	-0.170
Helix Content of Extracellular	-0.004	-0.060
Beta Sheet Contect Extra (%)	-0.006	0.067
Random Coil Content of Extra (%)	0.011	0.038
Helix Content of Cytoplasmic (%)	0.001	-0.185
Beta Sheet Content of Cyto (%)	0.049	0.188
Random Coil Content of Cyto (%)	-0.046	0.012
Extracellular Dynamics Average	-0.019	-0.011
Extracellular St Dev	0.015	-0.075
Transmembrane Dynamics Avg	0.058	0.116
Transmembrane St Dev	0.059	0.113
Cytoplasmic Dynamics Avg	0.008	-0.091
Cytoplasmic St Dev	-0.067	-0.135

APPENDIX E:

Principal Component Scores

Appendix E: Principal Component Scores

CD Molecule	PC1	PC2
CD10	8.5597	-0.48569
CD13	15.1776	3.53865
CD23	-1.5207	2.27423
CD26	10.2670	3.18327
CD38	-2.3981	2.32396
CD69	-4.9752	1.16782
CD70	-4.7395	4.58819
CD71	7.0419	-6.18392
CD72	-3.2600	-6.63813
CD74	-3.2297	-1.11401
CD94	-4.2354	5.11938
CD98	2.0124	-7.29751
CD153	-4.2313	1.48472
CD154	-3.6592	1.99856
CD159a	-5.1408	-3.61898
CD159c	-5.4251	-3.68640
CD161	-4.9006	-0.01895
CD178	-4.8762	-2.70428
CD203a	11.5556	-6.16234
CD203c	15.1619	8.00036

CD204	1.0287	-2.41090
CD207	-2.4162	-0.54761
CD209	-0.0520	-0.21244
CD238	6.7074	-4.02782
CD249	14.8764	0.15797
CD252	-5.0901	2.83325
CD253	-2.5958	3.08578
CD254	-2.5923	-0.23664
CD257	-3.4443	-2.11916
CD258	-4.4856	0.19700
CD298	-3.2138	0.27057
CD299	-0.2550	-1.43651
CD301	-1.7071	0.47265
CD303	-4.1242	3.67934
CD314	-5.0395	0.05507
CD317	-4.7807	4.47050