# Characterisation and Classification of Protein Sequences by Using Enhanced Amino Acid Indices and Signal Processing-Based Methods

Charalambos Chrysostomou

Faculty of Technology

De Montfort University

A thesis submitted for the degree of

*Doctor of Philosophy*

May 2013

I would like to dedicate this thesis to my family, for their support throughout the course of this thesis.

# Acknowledgements

The author wishes to express his gratitude to his first supervisor, Dr Huseyin Seker, who was abundantly helpful and offered invaluable assistance, support and guidance. Deepest gratitudes are also due to the members of the supervisory committee, Dr Ruta Furmonaviciene, and Professor Robert John, without whose knowledge and assistance this study would not have been successful. The author would also like to thank his family for the support and encouragement they provided through the years. In conclusion, the author recognises that this research project would not have been possible without the financial assistance of De Montfort University, which fully funded this project.

# Declaration

I declare that the work described in my thesis is original work undertaken by me for the degree of Doctor of Philosophy, at The Centre for Computational Intelligence (CCI), at De Montfort University, United Kingdom. No part of the material described in this thesis has been submitted for the award of any other degree or qualification.

# Publications Produced From The Ph.D. Thesis

## Published Papers

1. C. Chrysostomou, H. Seker, N. Aydin, and P. Haris, "Complex resonant recognition model in analysing influenza a virus subtype protein sequences," in Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine, (Corfu, Greece), pp. 1-4, November 2010.
   *C. Chrysostomou developed the methodology, collected and analysed the data, evaluated the results produced and wrote the paper.*
2. C. Chrysostomou, H. Seker, and N. Aydin, "Effects of windowing and zero-padding on complex resonant recognition model for protein sequence analysis," in Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (Boston, USA), pp. 4955-8, August 2011.
   *C. Chrysostomou developed the methodology, collected and analysed the data, evaluated the results produced and wrote the paper.*
3. C. Chrysostomou, H. Seker, and N. Aydin, "Investigation into the effects of an individual amino acid on protein function by means of a resonant recognition model," in Proceedings of the 5th International Conference on Convergence and Hybrid Information Technology, (Korea), pp. 229-236, Springer-Verlag, 2011.
   *C. Chrysostomou developed the methodology, collected and analysed the data, evaluated the results produced and wrote the paper.*
4. C. Chrysostomou and H. Seker, "Construction of Protein Distance Matrix Based on Discrete Fourier Transform", in Proceedings of the 35rd Annual International Conference of the IEEE Engineering in Medicine

and Biology Society, (Osaka, Japan), July 2013
*C. Chrysostomou developed the methodology, collected and analysed the data, evaluated the results produced and wrote the paper.*

5. C. Chrysostomou and H. Seker, "Signal-processing based bioinformatics approach for Classification of Protein Sequences", in Proceedings of the 35rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (Osaka, Japan), July 2013
*C. Chrysostomou developed the methodology, collected and analysed the data, evaluated the results produced and wrote the paper.*

6. C.J. Carmona, C. Chrysostomou, H. Seker, M.J. del Jesus. Fuzzy Rules for Describing Subgroups from Influenza A Virus Using a Multi-objective Evolutionary Algorithm. Applied Soft Computing (2013)
*C. Chrysostomou collected the data used in the analysis, evaluate the results produced and wrote part of the paper.*

# Abstract

Protein sequencing has produced overwhelming amount of protein sequences, especially in the last decade. Nevertheless, the majority of the proteins' functional and structural classes are still unknown, and experimental methods currently used to determine these properties are very expensive, laborious and time consuming. Therefore, automated computational methods are urgently required to accurately and reliably predict functional and structural classes of the proteins. Several bioinformatics methods have been developed to determine such properties of the proteins directly from their sequence information. Such methods that involve signal processing methods have recently become popular in the bioinformatics area and been investigated for the analysis of DNA and protein sequences and shown to be useful and generally help better characterise the sequences. However, there are various technical issues that need to be addressed in order to overcome problems associated with the signal processing methods for the analysis of the proteins sequences.

Amino acid indices that are used to transform the protein sequences into signals have various applications and can represent diverse features of the protein sequences and amino acids. As the majority of indices have similar features, this project proposes a new set of computationally derived indices that better represent the original group of indices. A study is also carried out that resulted in finding a unique and universal set of best discriminating amino acid indices for the characterisation of allergenic proteins. This analysis extracts features directly from the protein sequences by using Discrete Fourier Transform (DFT) to build a classification model based on Support Vector Machines (SVM) for the allergenic proteins. The proposed predictive model yields a higher and more reliable accuracy than those of the existing methods.

A new method is proposed for performing a multiple sequence alignment. For this method, DFT-based method is used to construct a new distance matrix in combination with multiple amino acid indices that were used to encode protein sequences into numerical sequences. Additionally, a new type of substitution matrix is proposed where the physicochemical similarities between any given amino acids is calculated. These similarities were calculated based on the 25 amino acids indices selected, where each one represents a unique biological protein feature. The proposed multiple sequence alignment method yields a better and more reliable alignment than the existing methods.

In order to evaluate complex information that is generated as a result of DFT, Com-

plex Informational Spectrum Analysis (CISA) is developed and presented. As the results show, when protein classes present similarities or differences according to the Common Frequency Peak (CFP) in specific amino acid indices, then it is probable that these classes are related to the protein feature that the specific amino acid represents. By using only the absolute spectrum in the analysis of protein sequences using the informational spectrum analysis is proven to be insufficient, as biologically related features can appear individually either in the real or the imaginary spectrum. This is successfully demonstrated over the analysis of influenza neuraminidase protein sequences.

Upon identification of a new protein, it is important to single out amino acid responsible for the structural and functional classification of the protein, as well as the amino acids contributing to the protein's specific biological characterisation. In this work, a novel approach is presented to identify and quantify the relationship between individual amino acids and the protein. This is successfully demonstrated over the analysis of influenza neuraminidase protein sequences.

Characterisation and identification problem of the Influenza A virus protein sequences is tackled through a Subgroup Discovery (SD) algorithm, which can provide ancillary knowledge to the experts. The main objective of the case study was to derive interpretable knowledge for the influenza A virus problem and to consequently better describe the relationships between subtypes of this virus. Finally, by using DFT-based sequence-driven features a Support Vector Machine (SVM)-based classification model was built and tested, that yields higher predictive accuracy than that of SD.

The methods developed and presented in this study yield promising results and can be easily applied to proteomic fields.

# Contents

# List of Figures

## LIST OF FIGURES

# List of Tables

## LIST OF TABLES

## LIST OF TABLES

# Nomenclature

| | |
|---|---|
| 3-D | Three-Dimensional |
| AAID | Amino Acid Index Database |
| AR | Autoregressive |
| ARP | Allergen Representative Peptide |
| BLAST | Basic Local Alignment Search Tool |
| BLOSUM | Blocks of Amino Acid Substitution |
| CD4 | Cluster of Differentiation 4 |
| CFP | Characteristic Frequency Peak |
| CISA | Complex Informational Spectrum Analysis |
| CISAPS | Complex Informational Spectrum for Analysis of Protein Sequences |
| DFT | Discrete Fourier Transform |
| EFS | Evolutionary Fuzzy System |
| EIIP | Electro-ion Interaction Potential |
| EMBL-EBI | European Molecular Biology Laboratory - European Bioinformatics Institute |
| FFT | Fast Fourier Transform |
| FGF | Fibroblast Growth Factors |

**Nomenclature**

| | |
|---|---|
| FN | False Negatives |
| FP | False Positives |
| G-mean | Geometric Mean |
| HA | Hemagglutinin |
| HIV | Human Immunodeficiency Virus |
| IgE | Immunoglobulin E |
| ISA | Informational Spectrum Analysis |
| kNN | k-Nearest-Neighbour |
| LPS | Lipopolysaccharide |
| M | Matrix |
| MAFFT | Multiple Alignment Using Fast Fourier Transform |
| MCC | Matthews Correlation Coefficient |
| MSA | Multiple Sequence Alignment |
| NA | Neuraminidase |
| NMEEF-SD | Non-dominated Multiobjective Evolutionary Algorithm for Extracting Fuzzy Rules in Subgroup Discovery |
| NP | Nucleoprotein |
| NS | Non-Structural |
| NSGA-II | Nondominated Sorting Genetic Algorithm II |
| NW | Needleman-Wunsch |
| PAM | Point Accepted Mutation |
| PCA | Principal Component Analysis |

| | |
|---|---|
| PK-C | Protein Kinase C |
| PRL | Prolactin |
| PROFEAT | Protein Feature Server |
| PseAA | Pseudo Amino Acid Composition |
| RBF | Radial Basis Function |
| RNA | Ribonucleic Acid |
| RRM | Resonant Recognition Model |
| SD | Subgroup Discovery |
| SE | Sensitivity |
| SFA | Space-Frequency Analysis |
| SFR | Space-Frequency Representation |
| SP | Specificity |
| SSFT | Short-Space Fourier Transform |
| STFT | Short-Time Fourier Transform |
| SVM | Support Vector Machines |
| T-Coffee | Tree-based Consistency Objective Function For alignment Evaluation |
| TACC | Total Accuracy |
| TFA | Time-Frequency Analysis |
| TFR | Time-Frequency Representation |
| TN | True Negatives |
| TP | True Positives |
| UniProt | Universal Protein Resource |

| | |
|---|---|
| UPGMA | Unweighted Pair Group Method With Arithmetic Mean |
| WT | Wavelet Transform |

# Chapter 1

# Introduction

In the recent years, biology has been revolutionised from being a completely lab-based science to becoming an information science. Bioinformatics is the area of science in which computer science, information technology and biology are combined to establish a new discipline. In this chapter, an introduction to Bioinformatics will be given as well as protein sequences and amino acids that are the main focus of the thesis. Furthermore, signal processing as a bioinformatics method will be briefly introduced, and its applicability in bioinformatics will be discussed. Finally, the thesis aims and contributions will also be provided.

## 1.1 What Is Bioinformatics?

The main aim of the bioinformatics research area is to enable the discovery of new biological insights, as well as to create a global set of rules that govern biology [1; 2]. Originally, this area was associated with the creation and maintenance of databases for storing biological information such as DNA [3] and protein sequences [4]. These types of databases would usually be associated with complex user-interfaces that would enable researchers to access and update as well as append biological information. As the number and quality of biological data gathered in these databases increased, this information could be used to create a complete description of cellular activities and their associations with distinct disorder conditions. Therefore, the field of bioinformatics has evolved [1; 2] to analyse various types of data including DNA and protein sequences, protein domains, and protein struc-

tures. The research area in which biological data are analysed and interpreted is referred to as computational biology. Some of the significant fields of study within bioinformatics and computational biology as follows

- The development of tools and databases for managing biologically related information, such as
  - Swiss-Prot & TrEMBL [5], high quality annotated and non-redundant protein sequence databases.
  - UniProt [4] (Universal Protein Resource), an information database on protein sequences.
- The development of statistical and mathematical algorithms to analyse biological databases like the prediction of structural and functional class of protein sequences [6] and the protein sequences clustering into groups of associated sequences [7].

One of the aims of bioinformatics is to identify, understand and model the driving forces of the biological processes. In order to accomplish this goal this research area uses computational methods such as machine learning, data mining and pattern recognition techniques. To name research areas where these computational methods are successfully implemented are drug design and discovery [8], DNA and protein sequence alignment [9; 10], protein structure alignment [11] and gene prediction and identification [12].

Bioinformatics is considered to be an important research area as it can be used to help improve our understanding of the biological processes where the collection of further information is difficult or challenging. For example, in a specific biological process that can cause a disease in humans and the collection of further information is ethically challenging [13; 14] or forbidden by the current laws [13; 14], various models of the same process within other organisms can be used to gain further knowledge.

The reason for applying computational methods in order to understand different biological processes is that by using these methods a generalised model of the biological processes can be constructed. Additionally, by using these models, a hypothesis can be constructed regarding a specific biological process, which can be validated with lab-based experiments. By using this practice, money and time can be saved as lab-based experiments are expensive and highly time-consuming [15; 16].

Bioinformatic-based methods have been developed, studied and applied for various areas. As the work presented in this research project focuses mainly upon the analysis of protein sequences a brief description of proteins will be given in the following section.

## 1.2 Proteins

A protein is a biomodule that generally consists of 20 smaller components called amino acids, which are linked together [17]. They are connected with strong bonds creating a one-dimensional chain, in a similar way of DNA single strands. By using the 20 amino acid representation, each protein can be represented mathematically by a character string. The length of this string is relatively small, a few hundreds or thousands, relative to the DNA representation which is usually in millions or hundred of millions [17]. Protein functions are usually determined by their complex three-dimensional (3-D) structures in which they usually tend to fold. The 3-D structure determines where a protein can bind with other modules in a process that resemble a hand fitting into a glove [18]. Although all proteins in living cells are determined by character sequences, there is not yet a method to predict the 3-D structures.

The genetic code controls the protein synthesis by mapping each of the 64 possible triplets of DNA characters into the 20 amino acids [17], as shown in Table 1.1. A particular codon, the M amino acid (methionine), serves as a START codon but also appears in other locations in the sequence. Additionally, there are also three amino acids that serve as STOP codons, as Table 1.1 shows. There are two ways to code the nucleotides into amino acids depending on the direction the DNA strands (forward or reverse) [19].

One of the most important and unsolved problems in bioinformatics is to automatically identify which proteins serve a particular function. The sequence is coded in nucleotides of three (codons), so the sequence needs to be a multiple of three, starting with a START codon and ending with a STOP codon. The STOP codon cannot appear in any other location in the sequence, contrary to the START codon that can appear as the M amino acid. In the next section a more detailed description of the amino acids will be given.

Table 1.1: Amino Acids

| Name | Abbreviation | Codons |
|---|---|---|
| Alanine | Ala/A | GCU, GCC, GCA, GCG |
| Arginine | Arg/R | CGU, CGC, CGA, CGG, AGA, AGG |
| Asparagine | Asn/N | AAU, AAC |
| Aspartic Acid | Asp/D | GAU, GAC |
| Cysteine | Cys/C | UGU, UGC |
| Glutamic Acid | Gln/Q | CAA, CAG |
| Glutamine | Glu/E | GAA, GAG |
| Glycine | Gly/G | GGU, GGC, GGA, GGG |
| Histidine | His/H | CAU, CAC |
| Isoleucine | Ile/I | AUU, AUC, AUA |
| START | | AUG |
| STOP | | UAG, UGA, UAA |
| Leucine | Leu/L | UUA, UUG, CUU, CUC, CUA, CUG |
| Lysine | Lys/K | AAA, AAG |
| Methionine | Met/M | AUG |
| Phenylalanine | Phe/F | UUU, UUC |
| Proline | Pro/P | CCU, CCC, CCA, CCG |
| Serine | Ser/S | UCU, UCC, UCA, UCG, AGU, AGC |
| Threonine | Thr/T | ACU, ACC, ACA, ACG |
| Tryptophan | Trp/W | UGG |
| Tyrosine | Tyr/Y | UAU, UAC |
| Valine | Val/V | GUU, GUC, GUA, GUG |

## 1.3 Amino acids

An amino acid has an average molecular weight of 135 daltons [20], and is one of the smallest bio-molecules that exist in nature. These bio-molecules exist naturally in a zwitterion state [20] (zwitterion-state exist when an amino acid has zero net charge but carries positive and negative charges on different atoms) where the carboxylic acid part is ionized and the basic amino group is protonated [20]. The common structure of the class of amino acid is an organic carboxylic acid group with an amino group attached to the saturated carbon atom [20]. The simplest amino acid is glycine, which is the only one that is optically inactive as its saturated carbon atom is unsubstituted. The remaining of the 20 most common

amino acids used are optically active in both dextrorotary and levorotary stereoisomers [20]. These amino acids are listed in Table 1.2 along with relevant biological information. In nature, the groups of proteins that are studied, are levorotary isomers. On the saturated carbon atom of the amino acid various substituents can connect from lower alkyl groups to aromatic amines and, alcohols and there are also acidic and basic side chains as well as thiol chains that can be oxidized to dithiol linkages between two similar amino acids [20].

All proteins and enzymes that exist in nature are the product of amino acid combinations and so they are considered to be principal building blocks. Ribonucleic Acid (RNA) unites these amino acids into proteins according to the genetic code, while ribosomes decode messenger RNA. The content of the amino acids that are used to assemble the protein or enzyme determines the spatial and biochemical properties of this protein or enzyme. The primary sequence of the protein may be determined by the amino acid backbone, and the nature of this chain can specify the properties of the protein. These chains can be polar, non-polar or even neutral [21]. The polar side chains of the protein usually appear on the surface, where they can interact with the water based environment found in cells [22]. The non-polar side chains tend to appear in the center of the protein, where they can interact with other non-polar amino acids [22]. These interactions can create a hydrophobic region in an enzyme in a non-polar atmosphere. Additionally, substituents can exist in the active side of enzymes, where they can provide a polar region in which to be able to conduct biochemical synthesis [23].

The twenty amino acids are obtained from different sources and used differently as listed in Table 1.2 [20; 22].

## 1.4 Amino Acid Indices

An amino acid index is a fixed length vector containing twenty numerical values representing a protein's physiochemical or biochemical property. Amino acid indices have been developed and extensively investigated since the early sixties and derived by means of laboratory experiments on biological specimens (e.g., hydrophobicity [24], polarity [25], size [26] or volume [25]) and/or computational experiments on laboratory-derived indices (e.g., principal components extracted from relations between chemical structure and biological activity in peptides [27]).

The largest database of amino acid indices is the AAindex1 database [28] that is located

Table 1.2: Biological Information About Amino Acids

| Amino Acid | Biological Information |
|---|---|
| Alanine | Used in most of the proteins and it is the second simplest amino acid |
| Arginine | Usually used at the active sites of enzymes |
| Asparagine | Obtained from aspartic acid |
| Aspartic Acid | Intermediate in the citric acid cycle |
| Cysteine | Involved in active sites and protein tertiary structure determination |
| Glutamic Acid | Negatively charged and found on the surface of proteins |
| Glutamine | Can easily cross the barrier between blood and brain tissue |
| Glycine | Simplest amino acid - acts as neurotransmitter antagonist |
| Histidine | Responsible for histamine biosynthesis |
| Isoleucine | Exclusively used in protein and enzyme construction |
| Leucine | Exclusively used in protein and enzyme construction also |
| Lysine | Essential with a positive charge on the aliphatic side chain |
| Methionine | Essential and initiate protein synthesis |
| Phenylalanine | Most common aromatic amino acid in proteins |
| Proline | Used in the synthesis of collagen |
| Serine | Found in the active site of serine proteases |
| Threonine | Involved in porphyrin metabolism |
| Tryptophan | Used the least frequently in proteins |
| Tyrosine | Used to build neurotransmitters and hormones |
| Valine | Used to hold proteins together |

at the GenomeNet. The latest update of this database was in March 2008 and contained 544 amino acid indices collected from various published literatures dating between 1964 and 2005.

Amino acid indices represent distinctive physicochemical and biochemical properties of a protein, and can be used in a variety of bioinformatics problems where different protein characteristics play a key role. In the literature, various studies have utilised amino acid indices in identification of protein structural class [29; 30; 31], protein subcellular location [32], secondary structure [33; 34], transmembrane sequences [35], predicting chemical structure and biological function [36], surface prediction [37; 38] and prediction of disordered regions [39]. As amino acid indices are assigned to amino acids in a protein sequence the protein can then become a discrete sequence or signal by which signal processing methods can then be used to analyse.

## 1.5 Signal Processing Methods

In recent years, signal-processing techniques have been used in bioinformatics to extract information that is expected to reveal a protein's biological function [40; 41; 42]. Signal processing can be defined as the research area that deals with the analysis of discrete and continuous signals. A signal can be described as a mathematical relation that shows how the signal magnitude varies over time. Signal processing is applied in many research areas, such as engineering and applied mathematics and the involved signals can be in the format of images, sound and any type of measurement from sensors. One of the most common signal analysis techniques is the Fourier transform, in which a signal can be converted into a series of frequencies. This expresses a range over which some measurements produced by a physical phenomenon, such as sound, electromagnetic radiation, or the mass of specific kinds of particles, can vary [43].

There are various research areas where signal processing is successfully applied such as audio signal processing [44], image processing [45], speech processing [46; 47], digital communications [48], and biomedicine [49]. For a variety of applications [50; 51] it is important to characterise a signal for the frequency and time domain at the same time. In application to biology, such signal processing methods can point to the existence of important characteristics in the data that the signal represents. Regarding the analysis of protein sequences, representing them as numerical signals should be expected to reveal specific biological characteristics such as the binding sites.

## 1.6 Thesis Aim

Protein sequencing has produced overwhelming amounts of protein sequences especially in the last decade [52; 53; 54; 55]. Nevertheless, functional and structural classes of the majority of the protein sequences are still unknown, and experimental methods currently used to determine these properties are very expensive, laborious and time consuming [15]. Therefore, automated computational methods are urgently required to accurately and reliably predict functional and structural classes of the proteins. Several bioinformatics methods have been developed to determine such properties of proteins directly from their sequence information [56]. Such methods involving signal processing methods, which are discussed and analysed in the literature review, have recently become popular throughout

the bioinformatics area and have been investigated for the analysis of DNA and protein sequences and are shown to be useful and generally help better characterise the sequences [41; 57]. The following research question arises: **How can signal processing techniques be utilised in order to improve already existing bioinformatics approaches, or used to extract novel features that can be used for characterisation and classification of protein sequences?**

In order to address the research question, the outcome of this study will be a series of bioinformatics systems that consider signal-processing techniques for the analysis of the protein sequences as a signal and hence are expected to be capable of better characterising the proteins. This will also be further improved by fusing multiple protein characteristics and pattern recognition methods. This research study will bring a novel concept to characterising the proteins, and will help predict structural and functional classes of the proteins from primary sequence information with greater reliability and accuracy. In addition, this will be the most comprehensive and consistent study that considers the use of amino acid indices. It is also expected that this work will provide guidelines to other areas of study in proteomics, in which automated prediction of a protein characteristic can achieved by using only the primary structure of this protein. It is anticipated that the outcome of the research has great potential in commercialisation and patent application.

## 1.7 Thesis Contributions

The main contributions of this thesis are

- As different groups of amino acid indices included in the database have similar features, a hybrid computational method is developed by using hierarchical clustering and principal component analysis. The hybrid method helped to summarise the groups of the indices and derive a small representative set of the amino acid indices. A new Database is also developed and presented in the Amino Acid Index Database (AAID) web-server (http://cisaps.com/indices), containing all the latest published amino acid indices as well as new indices.

- A comprehensive study is carried out that resulted in finding a unique and universal set of best discriminating amino acid indices for the characterisation of the allergenic proteins. This was achieved by extracting features from protein sequences by

using Discrete Fourier Transform (DFT) and building a classification model based on Support Vector Machines (SVM), for allergenic proteins.

- A new method is proposed for performing a multiple sequence alignment. For this method, Discrete Fourier Transform-based signal processing method is used to construct the distance matrix in combination with multiple amino acid indices that were used to encode protein sequences into numerical sequences. Additionally in this work, a new type of substitution matrix is proposed where the physicochemical similarities between any given amino acids is calculated. These similarities were calculated based on the 25 amino acids indices selected, where each one represents a unique biological protein feature.

- Complex Informational Spectrum Analysis (CISA) is developed to analyse protein sequences using the information captured in the real and imaginary spectrum in addition to the absolute spectrum. A web-based server is also developed and presented, named CISAPS, which provides CISA for the analysis of protein sequences.

- A novel approach is developed and presented to identify and quantify the relationship between amino acids and proteins by using signal processing methods. Two methods are presented in this analysis based on DFT to calculate the effect of the individual amino acid to the protein sequence. The first takes into consideration of the Common Frequency Peak (CFP) that is calculated from the Informational Spectrum Analysis (ISA), and the second considers the entire absolute spectrum generated by DFT.

- Characterisation and classification problem of the influenza A virus is tackled through a Sub-group Discovery (SD) algorithm and compared with Support Vector Machines (SVM), in order to accurately predict the subtype class.

## 1.8 Thesis Organisation

Having given a brief introduction to the research study, its contributions and achievements, they are explained in detail in the following sections of the thesis that has been organised as:

- Chapter 2 reviews the literature regarding the sequence-driven features that can be extracted from protein sequences as well as signal processing methods used in the analysis of proteins.
- Chapter 3 presents the Amino Acid Index Database web-server, which contains the largest up to date database with all the latest published amino acid indices. Additionally, this chapter introduces a new set of computationally derived indices that better represent the original group of indices.
- Chapter 4 presents a study that tries to discover a unique and universal set of best discriminating amino acid indices for the characterisation and prediction of allergenic proteins.
- In Chapter 5, a new method is proposed for performing multiple sequence alignment based on Discrete Fourier Transform.
- In Chapter 6, the Complex Resonant Recognition Model, based on the Discrete Fourier Transform, is developed and its robustness is demonstrated by applying in the influenza A subtypes Neuraminidase gene. Additionally, the Complex Informational Spectrum Analysis of Protein Sequences (CISAPS) web-server is presented.
- In Chapter 7, a novel approach is presented for identifying and quantifying the relationship between amino acid and protein by firstly considering the characteristic frequency peak that is obtained from the Resonant Recognition Model, and secondly by considering the entire absolute frequency spectrum.
- In Chapter 8, characterisation and classification of the influenza A virus is tackled through a Subgroup Discovery algorithm, namely Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery (NMEEF-SD), and SVM that resulted in higher predictive accuracy than that of NMEEF-SD .
- In Chapter 9 concludes the thesis by providing future works.

# Chapter 2

# Literature Review

If it is considered that proteins' biological functions are controlled by the selective ability of the protein to interact with specific elements in the environment an argument can arise: How is this selected ability achieved and what is its physical basis? Several attempts have been made to decode the rules that drive biological functions of the proteins, which mainly deal with secondary and tertiary structures, but not directly from the amino acid sequences.

In recent years, different models have been developed for prediction of biological properties (e.g. functional or structural classes) of proteins or peptides. These models usually use physicochemical properties and sequence-derived structural information that can be extracted directly from protein sequences' primary structures. These sequence-driven features are generally found to be capable of representing and classifying protein sequences with distinctive functional and structural classes. They are further combined with statistical and computational learning methods in order to classify protein sequences regardless of their sequence similarity [58; 59].

## 2.1 Sequence-Driven Features for Proteins

A list of features that can be extracted directly from the primary structure of protein sequences, and are widely accepted and used in the literature can be found in Table 2.1.

One of the first and most widely used features of protein sequences is Amino Acid Composition (AAC) that can be represented using a vector of 20 dimensions. The percentage of representation of all the 20 natural amino acids can be calculated by using Equation

Table 2.1: List of the Main Sets of Sequence-Driven Features

| Descriptors | Number of Subsets of the Descriptors | Number of Descriptors Values | Reference |
|---|---|---|---|
| Amino acid composition | 1 | 20 | [60] |
| Pseudo-amino acid composition | *a* | *a* | [61] |
| Dipeptide composition | 1 | 400 | [62] |
| Moreau-Broto autocorrelation | 8 | 240 | [63] |
| Moran autocorrelation | 8 | 240 | [63] |
| Geary autocorrelation | 8 | 240 | [63] |
| Sequence-order-coupling number | 2 | 60 | [64] |
| Quasi-sequence-order | 2 | 100 | [65] |
| Amphiphilic pseudo-amino acid composition | *a* | *a* | [66] |
| Topological atom model | - | 405 | [58] |
| Total amino acid properties | *a* | *a* | [58] |

*a* - The number of features depends on the $\lambda$ parameter and the length of the protein sequences

2.1.

$$AAC(i) = \frac{total\ number\ of\ amino\ acids\ i}{total\ number\ of\ amino\ acids} \tag{2.1}$$

where *i* is a specific amino acid in a protein. From the literature, AAC has been successfully used in a variety of applications such as predicting protein structural classes [67] and subcellular localisation of proteins [68].

Another important set of the features is the pseudo amino acid composition (PseAA) [61]. This feature was proposed in order to take in consideration of sequence-order information that would otherwise be lost with alternative approaches [61]. PseAA composition consists of a set of 20 or greater discrete factors where the first 20 factors are the conventional composition amino acid indices; the supplemental factors are calculated with additional algorithms using sequence-order information. These additional factors can be any combination of other factors, including the traditional 20 amino acid indices or a series of rank-different correlation factors throughout a protein chain. Hence the significance of PseAA composition is to include the amino acid composition, and maintain all the infor-

mation beyond the AA composition that might be lost in using alternative procedures, and reveal the characteristics of a protein sequence through a discrete model. Many different models exist that can formulate a PseAA composition. The algorithm for the PseAA [61] is as follows:

$$P = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7 \cdots \mathrm{R}_L \tag{2.2}$$

$$\mathbf{P} = \begin{bmatrix} f_1 & f_2 & \cdots & f_{20} \end{bmatrix}^{\mathbf{T}} \tag{2.3}$$

$$P = [p_1, p_2, \cdots, p_{20}, p_{20+1}, \cdots, p_{20+\lambda}]^{\mathbf{T}}, (\lambda < L) \tag{2.4}$$

where P represents the protein sequence, L the amino acid residues, $f_u (u = 1, 2, ...20)$ the normalised occurrence frequency of the 20 amino acids and T the transposing operator. The $20 + \lambda$ components are given by

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i \ + \ w \sum_{k=1}^{\lambda} \tau_k}, & (1 \leq u \leq 20) \\ \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i \ + \ w \sum_{k=1}^{\lambda} \tau_k}, & (20 + 1 \leq u \leq 20 + \lambda) \end{cases} \tag{2.5}$$

where the weight factor is $w$, and $k$ is the $k$-th correlation factor. The sequence order between all the $k$-th correlation factors reveal the most immediate residues as devised by:

$$\tau_k = \frac{1}{L-k} \sum_{i=1}^{L-k} \mathrm{J}_{i,i+k}, \ \ (k < L) \tag{2.6}$$

with

$$\mathrm{J}_{i,i+k} = \frac{1}{\Gamma} \sum_{g=1}^{\Gamma} \left[\Phi_\xi \left(\mathrm{R}_{i+k}\right) - \Phi_\xi \left(\mathrm{R}_i\right)\right]^2 \tag{2.7}$$

where $\Phi_\xi \left(\mathrm{R}_i\right)$ is the $\xi - th$ function of the AA $\mathrm{R}_i$ and $\Gamma$ the total number of the functions taken into consideration.

Physical characteristics of the protein sequences which can be used in the PseAA composition, $\Phi_1 \left(\mathrm{R}_i\right), \Psi_2 \left(\mathrm{R}_i\right)$ and $\Psi_3 \left(\mathrm{R}_i\right)$ can be the hydrophobicity, hydrophilicity, and side

chain mass values respectively for the amino acid $R_i$ [61]. The total number of functions considered for this example is $\Gamma = 3$. From Equation 2.4, the first 20 components $(p_1, \cdots, p_{20})$ show the conversional amino acid composition of the protein while the remaining components $(p_{20+1}, \cdots, p_{20+\lambda})$ are the correlation factors. These additional correlation factors are thought to hold information about the protein is sequence order. By changing the integer parameter $\lambda$ in Equation 2.4, the PseAA composition will lead to a dimension-difference. Other models can be used in addition to the example given (Equation 2.7) such as physicochemical distance or amphiphilic pattern mode to calculate different types of PseAA compositions.

In line with the development of such sequence-driven features Protein Feature Server (PROFEAT) [56; 58] has been developed for researchers to be able to calculate the aforementioned physicochemical and structural features of the amino acid sequences. PROFEAT can be accessed and used freely at: http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi. Applications areas where the above-mentioned features can be used for predicting are given in Table 2.2.

Table 2.2: Examples of the Predictive Models Developed Using the Sequence-Driven Features

| Application Area | Reference |
|---|---|
| Protein structural classes | [69; 70] |
| Functional families | [71] |
| Sub-nuclear location of proteins | [72] |
| Protein-protein interactions | [73; 74] |
| Sub-cellular locations | [75; 76] |
| Outer membrane proteins | [77] |
| Transmembrane regions in protein | [78] |
| Lipase types | [79] |
| Protein folding | [80; 81] |
| Protease types | [82] |
| DNA-binding proteins | [83; 84] |
| Protein secondary structural contents | [85] |

Regardless of the various uses and applications of non-signal processing-based techniques, currently the area of signal processing within bioinformatics has became highly motivational in the recent years while studying protein sequences, with diverse applications emerging. One of the main use of signal processing-based techniques is to be able to

capture sequence order information, which will be discussed in the next section.

## 2.2 Signal Processing Methods

By using signal processing techniques the biological function of each sequence should be able to be extracted without the secondary and tertiary structures being known [18; 86]. Every macro-module, amino acid or nucleotide, can be represented by a corresponding value. This value can be any of the biochemical properties like electron-ion interaction potential [87; 88], hydrophobicity [88; 89], solubility [88; 89] or molecular weight [88; 89]. By applying existing signal processing techniques it is possible to extract information that will match sequence biological functions. A very common method used for analysing macro-module sequences to extract biological functions is based on the search for similarities in the arrangements between the groups of sequences. To be able to proceed with current signal processing techniques a set of numerical values needs to be assigned to nucleotides or amino acids [17; 87; 88; 89]. These values should be, by nature, representative of some of the characteristics of the macro-modules with which they pair, and be relevant to the biological activity of each module. By using this technique a module in the sequence is represented by the same number despite its position. The final goal of the use of these discrete signals is to determine biological functions of amino acid or nucleotide sequences by extracting related parameters. So, any signal processing method can then be applied to separate these unique parameters from the sequences.

For each group of proteins analysed [87; 89], there is a group of proteins that corresponds to specific frequency in the spectrum. Every biological function corresponds to one unique frequency or a set of unique frequencies. The importance of this conclusion is that specific biological functions can be recognised from macro-modules by using signal processing methods by extracting significant features of the frequencies, which are not found in unrelated frequencies. There are various types of signal processing techniques that can be used in the analysis of protein sequences, like Frequency Analysis (Discrete Fourier Transform) and Space-Frequency Analysis (short-space Fourier Transform and Wavelet). These types of transform will be discussed in more detail in the following sections.

### 2.2.1 Frequency Analysis Using Discrete Fourier Transform

The Discrete Fourier Transform (DFT) is defined as follows:

$$X(n) = \sum_{m=0}^{N-1} x(m) e^{-j(2\pi/N)nm} \quad n = 0, 1, ..., N-1 \tag{2.8}$$

where $x(m)$ is the $m$th member of the numerical series, N is the total number of points in the series, and $X(n)$ are coefficients of the DFT. As the DFT coefficients consisted of two mirror parts, only the first half of the series $(N/2)$ points) will be hereafter considered. The following formula determines the maximal frequency in the spectrum

$$F = \frac{1}{2d} \tag{2.9}$$

where $F$ is the maximal frequency of all signals and $d$ is the distance between points of the sequence. If it is assumed that distance $d = 1$ then the maximum frequency in the spectrum can be found as $F = 1/2(1) = 0.5$. The output of DFT is a complex sequence and can be characterized as follows

$$X(n) = (R(n) + jI(n)), \quad n = 0, 1, ..., (N-1)/2 \tag{2.10}$$

where $R(n)$ is the Real part of the sequence and $I(n)$ the Imaginary part.
The absolute spectrum can be formulated as follows

$$S_{(n)} = X(n)X^*(n) = |X(n)|^2, \quad n = 0, 1, ..., (N-1)/2 \tag{2.11}$$

where $S_{(n)}$ is the absolute spectrum for a specific protein, $X(n)$ are the DFT coefficients of the series $x(n)$ and $X^*(n)$ are the complex conjugate.

### 2.2.2 Informational Spectrum Analysis

The Informational Spectrum Analysis (ISA) [40; 90; 91; 92] is a physicomathematical model that analyses the interaction of a protein and its target by using Discrete Fourier Transform. One application of this model involves prediction of a protein's biological function [31; 93]. In this technique, Discrete Fourier transform is applied to a numerical

representation of a protein sequence, and a frequency is determined for a protein's particular function. The aim of ISA is therefore to determine a Characteristic Frequency Peak (CFP) in the absolute spectrum that correlates with a biological function expressed by a set of protein sequences by using the informational spectrum.

Informational Spectrum can be defined as

$$C_{(n)} = \prod_{m=1}^{M} S_{(n)}(m) \tag{2.12}$$

where $C_{(n)}$ is the informational spectrum and M is the number of protein sequences used for a specific class. Equation 2.13 is used to scale Informational Spectrum

$$V = \frac{\sqrt{\sum_{n=0}^{N/2} C(n)}}{N/2} \tag{2.13}$$

where $L$ is the number of points in the Informational Spectrum ($C$).

After obtaining the Informational Spectrum the CFP can then be determined by selecting a frequency, or set of frequencies that present higher values of peaks. CFP pursuant to the absolute informational spectrum analysis can be used for characterising and distinguishing the proteins. However, the following conditions should be fulfilled for the CFP to be related to a biological function:

1. For diverse biological functions, CFP is expected to be dissimilar.
2. For biologically dissimilar protein sequences, CFP should not exist.
3. For a collection of protein sequences that allocate the same biological function a single CFP should exist.

A special case of ISA is Resonant Recognition Model (RRM) [40; 91; 92] where the Electro-ion Interaction Potential (EIIP) index [87] is used to encode protein sequences to numerical sequences. The EIIP index values can be found in Table 2.3. In order to show how the method can be applied in practice, an example is given below [94]. This is a case study where Acid and Basic fibroblast growth factor (FGF) protein sequences were analysed. Their protein sequences are

**Acid bovine FGF**

FNLPLGNYKKPKLLYCSNGGYFLRILPDGTVDGTKDRSDQHIQLQLCAESIG
EVYIKSTETGQFLAMDTDGLLYGSQTPNEECLFLERLEENHYNTYISKKHAE
KHWFVGLKKNGRSKLGPRTHFGQKAILFLPLPVSSD

**Basic bovine FGF**

PALPEDGGSGAFPPGHFKDPKRLYCKNGGFFLRIHPDGRVDGVREKSDPHIKL
QLQAEERGVVSIKGVCANRYLAMKEDGRLLASKCVTDECFFFERLESNNYN
TYRSRKYSSWYVALKRTGQYKLGPKTGPGQKAILFLPMSAKS

Table 2.3: EIIP Values

| Amino acid | EIIP | Amino acid | EIIP |
|---|---|---|---|
| Leu | 0.0000 | Tyr | 0.0516 |
| Ile | 0.0000 | Trp | 0.0548 |
| Asn | 0.0036 | Gln | 0.0761 |
| Gly | 0.0050 | Met | 0.0823 |
| Glu | 0.0057 | Ser | 0.0829 |
| Val | 0.0058 | Cys | 0.0829 |
| Pro | 0.0198 | Thr | 0.0941 |
| His | 0.0242 | Phe | 0.0946 |
| Lys | 0.0371 | Arg | 0.0959 |
| Ala | 0.0373 | Asp | 0.1263 |

These sequences can be converted into numerical sequences using the EIIP index. They are presented in signal in Figures 2.1 and 2.2, respectively. The next step in RRM process is to apply DFT to the numerical representations of FGF protein sequences. The frequency spectrum's results are shown in Figures 2.3 and 2.4, respectively. Finally, the informational spectrum of the two bovine FGF proteins is calculated as shown in Figure 2.5.

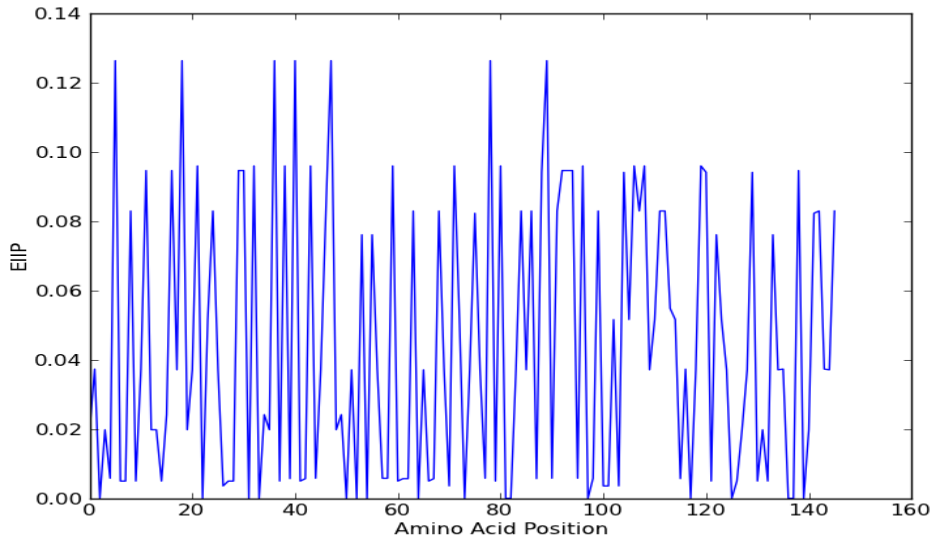



Figure 2.1: Acidic Bovine FGF With EIIP Index Values

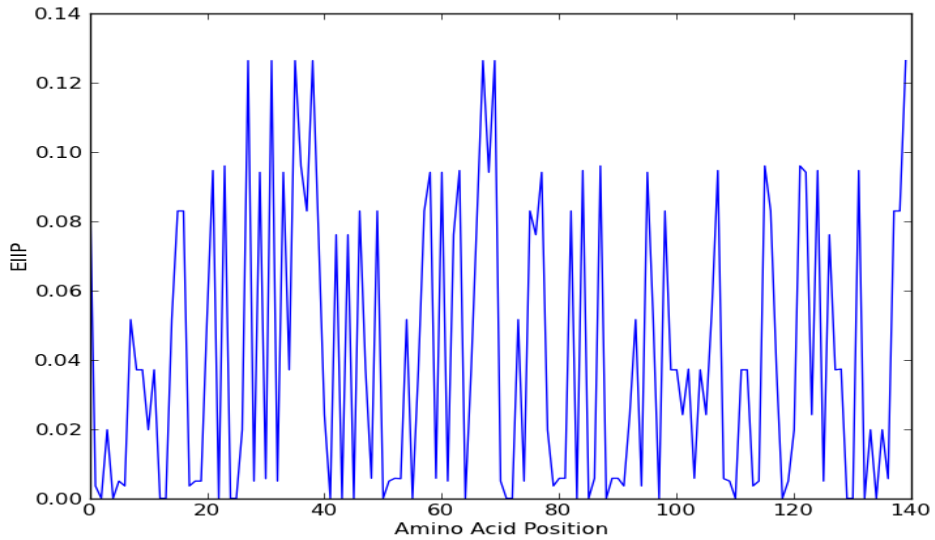



Figure 2.2: Basic Bovine FGF With EIIP Index Values

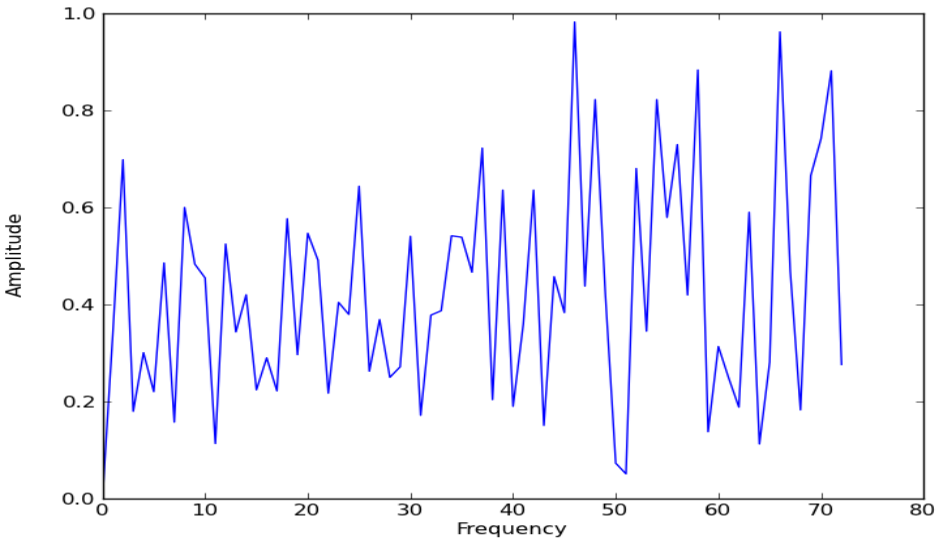



Figure 2.3: Absolute Frequency Spectrum of Acidic Bovine FGF

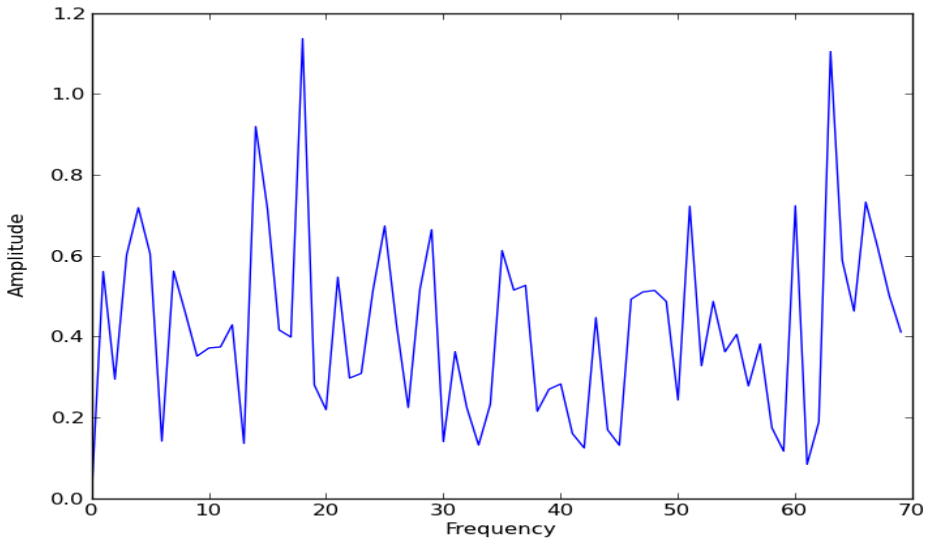



Figure 2.4: Absolute Frequency Spectrum of Basic Bovine FGF

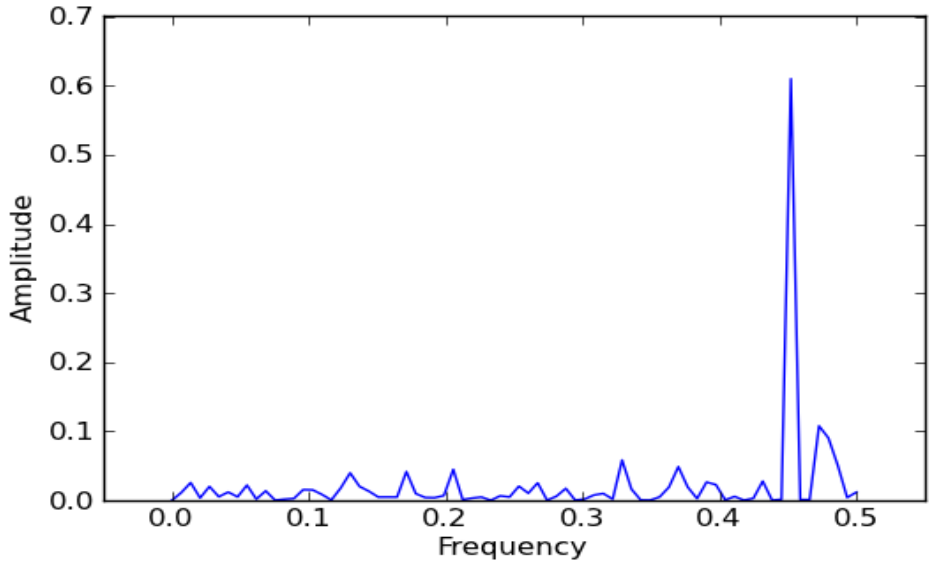



Figure 2.5: Absolute Informational Spectrum of the Two Bovine FGF Proteins Shown in Figures 2.3 and 2.4

### 2.2.3 Space-Frequency Analysis

Time-Frequency analysis (TFA) [95] is the simultaneous analysis of signals in both the time and frequency domains, by using time-frequency representation (TFR). For the analysis of protein sequences as signals, the concept of time is not applicable. Thus, space-frequency representation (SFR) is used a signal is represented over both space and frequency. By using SFR, non-stationary signals can be examined, as this representation can show how the frequency component of a signal changes over space. Additionally, using frequency analysis methods only with non-stationary signals can be inadequate for providing useful information. Space-Frequency Analysis (SFA) studies a two-dimensional signal, instead of analysing a single dimensional signal using a particular transformation (like Fourier Transform). Furthermore, a space-frequency transform is used to analyse this signal with the domain of a two-dimensional real plane acquired from the signal. Some of the applications where SFA is used are prediction of Hydrophobic Cores of Proteins [96], predicting allergenic proteins using wavelet transform [97], prediction of protein structural classes [70], identification of drug binding sites [98] and predicting protein coding regions [99]. In this

section two popular methods, the short-space Fourier transform and wavelet transform will be discussed.

#### 2.2.3.1 Short-Space Fourier Transform

The short-time Fourier Transform (STFT) [95], is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. As mentioned in the previous section the concept of analysis of protein sequences over time and frequency is not applicable. Therefore in this Chapter the concept of time is replaced by space. The concept of short-time Fourier Transform is described as short-Space Fourier Transform (SSFF). In the discrete case, which is useful in the analysis of molecules, the signal that is going to be transformed can be divided into windows (chunks). These windows usually overlap to reduce errors at the boundary. Thereafter, in the next step of the transform each window of the divided data is transformed and the result, which is complex numbers, is added to the matrix. This matrix stores information about the magnitude and phase for every point of the input sequence in space and frequency. This can be defined as:

$$SSFT\left\{x[n]\right\} \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{2.14}$$

where the $x[n]$ is the input signal and $w[n]$ is the window. In the above equation, $m$ is discrete and $\omega$ is continuous. In most typical applications, including bioinformatics, the short-space Fourier transform is performed using the Discrete Fourier Transform, so both variables are discrete and quantised.

The magnitudes squared of the short-space Fourier transform generate the spectrogram as the following function shows:

$$Spectrogram = \{x(t)\} \equiv |X(\tau,\omega)|^2 \tag{2.15}$$

An example of short-space Fourier transformation using the Acid Bovine FGF and Bovine FGF is presented in this chapter. Different window sizes and overlap sizes were tested for this analysis and presented in Figures 2.6 and 2.7 for window = $28\%$ and overlap = $99\%$. In order to show the effect of these parameters, further experiments were carried out and their results are presented in Appendix A as Figures A.1 and A.2 for window = $10\%$

and overlap = 25%, Figures A.3 and A.4 for window = 10% and overlap = 50%, Figures A.5 and A.6 for window = 10% and overlap = 99%, Figures A.7 and A.8 for window = 28% and overlap = 50%, Figures A.9 and A.10 for window = 40% and overlap = 50%, and Figures A.11 and A.12 for window = 40% and overlap = 99% for Acid bovine FGF and Basic bovine FGF protein sequences, respectively. Furthermore, Figures 2.6 and 2.7 with the parameter set to $window = 28\%$ of the protein sequence and $overlap = 99\%$ of the window present a clearer indication regarding the hotspot locations. As the results show, a hotspot is detected at 0.48 and 0.49 for Acid bovine FGF and Basic bovine FGF protein sequences, respectively.

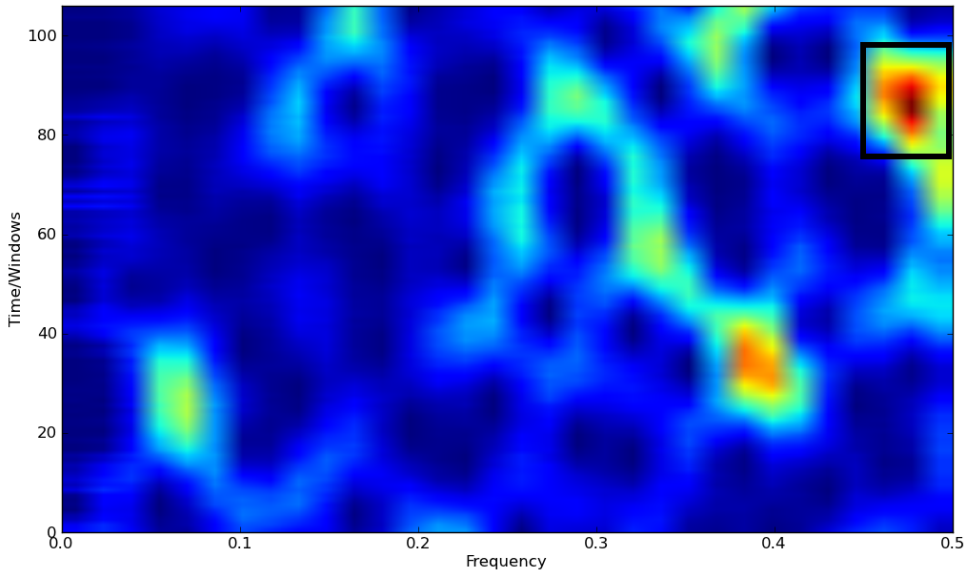



Figure 2.6: Short-Space Fourier Transform of Acid Bovine FGF Protein (Window: 28% - Overlap: 99%)

#### 2.2.3.2 Wavelet Transform

As discussed in the previous section, Fourier transform can only derive spectral information in contrast to the wavelet transform (WT), which can acquire both spectral and temporal information. The main disadvantage of Fourier transform is that Fourier's coefficients simply include universal average space-domain information resulting from the location

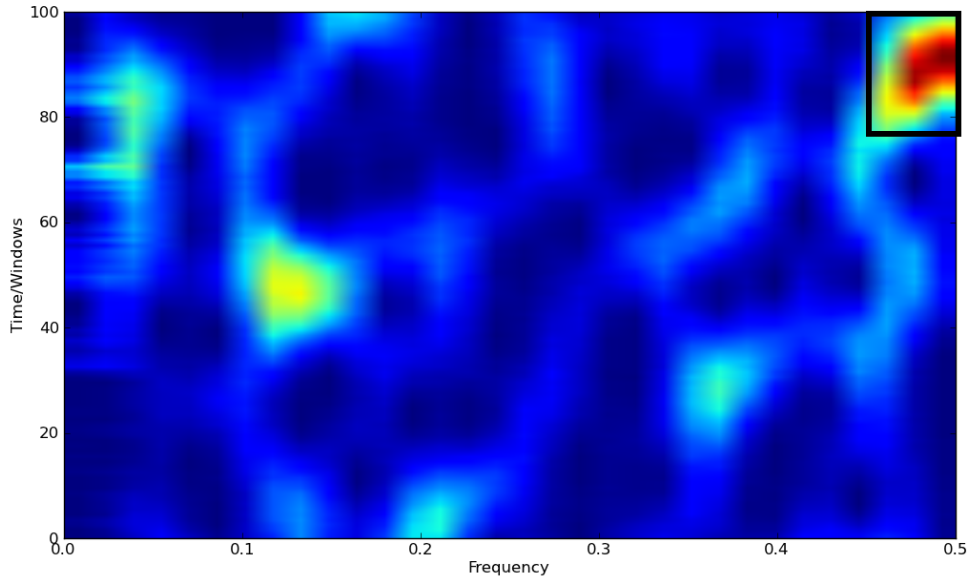



Figure 2.7: Short-Space Fourier Transform of Basic Bovine FGF Protein (Window: 28% - Overlap: 99%)

specific features being lost, in contrast to the wavelet transform.

A continuous wavelet transform can be characterised as the projection of a signal $f(t)$ on the wavelet function $\psi$ [100]. The projection of a signal $x$ onto the subspace of scale $a$ can be defined as

$$x_a(t) = \int_R WT_\psi\{x\}(a,b) \cdot \psi_{a,b}(t)\, db \tag{2.16}$$

where $a$ determine the positive scale and $b$ defines the shift and it can be any real number; $(a, b)$ describes a point in the right half-plane $R_+ \times R$. The subspace of scale a (where $[1/a,\ 2/a]$) can be found from the following equations:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \tag{2.17}$$

the wavelet coefficients ban is described as:

$$WT_\psi\{x\}(a,b) = \langle x, \psi_{a,b}\rangle = \int_R x(t)\psi_{a,b}(t)\, dt \tag{2.18}$$

The wavelet coefficients can be congregate into a scaleogram, for the analysis of a given signal.

From a computational perspective it is impossible to calculate all wavelet coefficient for the analysis of a signal, but it is adequate to select a discrete subset in order to reconstruct a signal from the corresponding wavelet coefficient. The corresponding discrete subset of consists of all the points ($a^m$, $na^m b$) with $m$, $n$ in $Z$. The corresponding sub-wavelets are now given as

$$\psi_{m,n}(t) = a^{-m/2}\psi(a^{-m}t - nb), \tag{2.19}$$

For reconstruction of any given signal x by

$$x(t) = \sum_{m\in\mathbb{Z}}\sum_{m\in\mathbb{Z}}\langle x, \psi_{m,n}\rangle\dot{\psi}_{m,n}(t) \tag{2.20}$$

subject to $(\psi_{m,n} : m, n \in \mathbb{Z})$.

For the analysis of protein sequences using the wavelet transform, three main steps need to be followed:

- Convert protein sequences to numerical sequences
- Selection of wavelet function
- Decompose the sequences by wavelet transform
- Use cross-correlation analysis to identify similar sequences

A method for protein analysis proposed by Fang and Cosic [50] uses a WT to analyse a protein sequence that was converted into a signal using the EIIP values. In addition, as the WT provides the same time-space resolution for each scale, the WT can be chosen to localise individual events, such as active site identification. The amino acids that comprise the active sites are identified as the set of local extrema of the coefficients in the wavelet transform domain. The energy concentrated in the regional extrema represent the locations of sharp variation points of the EIIP, which are proposed as the most critical locations of a protein's biological function [50].

In the experiment with Bovine FGF protein sequences, and by using the Morlet WT [51] the potential cell attachment sites are identified between the residues 46-48 and 88-90

as shown in Figures 2.8 and 2.9 for acid and basic bovine FGF protein, respectively. In this example, it can be observed that there are two bright regions in the spectrogram that correspond to the amino acids at the active sites. As the paper [50] proposes, WTs show promise for identifying amino acids at (potential) biologically active sites when used, but do not reveal the characteristic frequency component of the Resonant Recognition Model. Furthermore, it can be very difficult to explain the spectrogram of the continuous WT, and different space-frequency transforms can be used.

In order to show the effect of the use of different wavelet function, further experiments were carried out and their results are presented in Appendix A as Figures A.13 and A.14 shows the Paul WT [101], Figures A.15 and A.16 shows the Mexican Hat WT [101], Figures A.17 and A.18 shows the Derivative of Gaussian WT [101], and Figures A.19 and A.20 shows the Haar WT [101] for Acid and Basic Bovine FGF protein sequences, respectively.

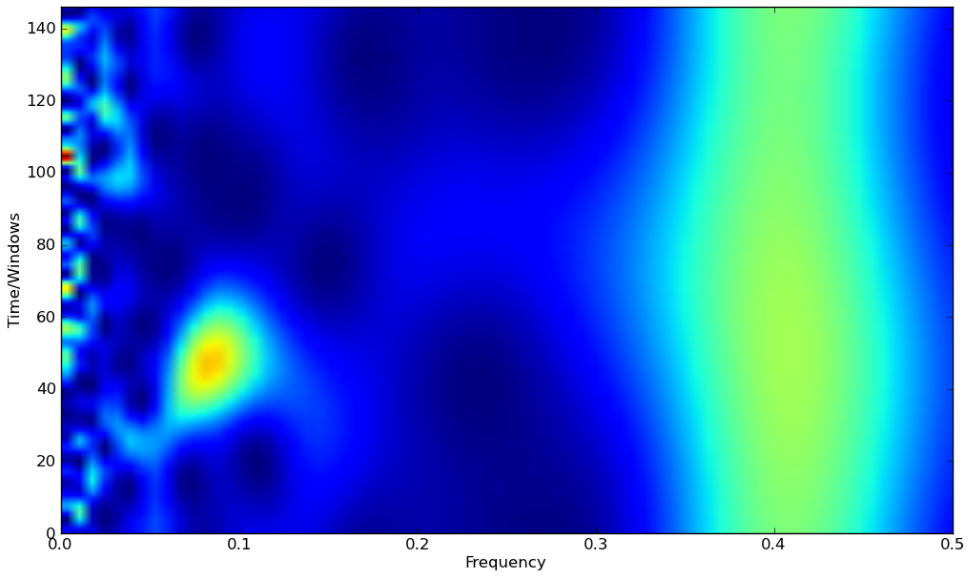



Figure 2.8: Morlet Wavelet Transform of Acid Bovine FGF Protein

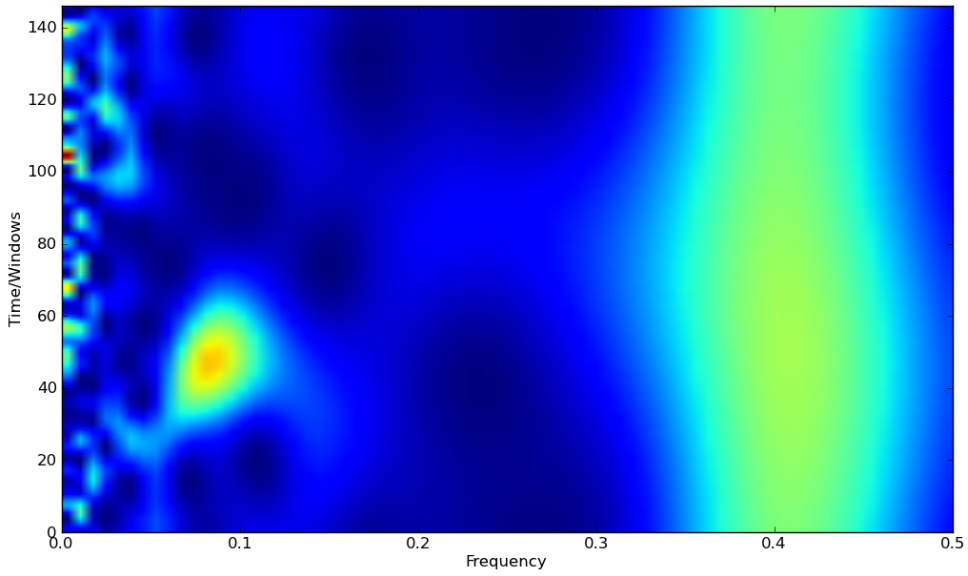



Figure 2.9: Morlet Wavelet Transform of Basic Bovine FGF Protein

## 2.3 Summary

In this Chapter, a literature review regarding existing sequence-driven features and signal processing techniques that are used are presented. For the first section, two main commonly used techniques, AAC and PseAA, are described. For the second section various signal processing methods used throughout the literature are given. For the first part, Frequency Analysis and DFT are presented. For the second part, Space-Frequency analysis, with SSFT and Wavelet Transform analysis are described.

In the next chapter, a description of amino acid indices is given which they represent unique physiochemical protein features. These amino acid indices are used to convert protein sequences to numerical sequences for further analysis. Furthermore, different methods are covered such as hierarchical clustering and principal component analysis used for the analysis.

## 2. LITERATURE REVIEW

# Chapter 3

# Description of Amino Acid Indices

## 3.1 Introduction

In the literature, a protein sequence contains important information regarding many protein properties such as protein structural classes [102; 103], protein-protein interactions [104] and protein function [105] can be derived from the amino acid composition [60] and further improved by including amino acid index values that can be assigned to each amino acid [61]. An amino acid index is a fixed length vector containing twenty values representing a protein's physiochemical or biochemical property. Amino acid indices have been developed and extensively investigated since the early sixties and derived by mainly laboratory experiments such as hydrophobicity [24], polarity [25], size [26] or volume [25] on biological specimens as well as computational experiments, ( e.g. electron-ion interaction potential (EIIP) [87] and Amino Acid Composition [106] (AAC)) on laboratory-derived indices. The largest database of amino acid indices is the AAindex1 database which is located at the GenomeNet [28]. The latest update of this database was in March 2008 and contained 544 amino acid indices from various published literature's dating between 1964 and 2005.

Amino acid indices represent distinctive physicochemical and biochemical properties of a protein, and can be used in a variety of bioinformatics problems where different protein characteristics play a key role. In the literature, various studies have utilised amino acid indices in identification of protein structural classes [29; 30; 31], protein subcellular location [32], secondary structure [33; 34], transmembrane sequences [35], predicting

chemical structure and biological function [36], surface prediction [37; 38] and prediction of disordered regions [39].

Closer investigation of the AAindex database reveals that many of these deposited indices have similar or identical sets of index values. In order to increase the robustness of this dataset these indices need to be removed or merged. In addition, the areas of bioinformatics and proteomics are highly active and new amino acid indices that represent novel protein features or characteristics are bound to be constantly generated. As the AAindex database was last updated [28] in 2008 any amino acid indices generated beyond this date are not included.

Recent studies [107; 108] have used the AAindex database and did not consider removing or combining the duplicate or highly similar entries within the dataset or to include any newly published indices in order to expand the dataset. The aim of the research to be presented in this chapter is therefore to add recently presented amino acid indices in the literature, remove any duplicate indices from the existing databases and to reduce the redundancy of the remaining indices in regard to the features they represent. In order to reduce the redundancy of groups of highly similar indices, which represent the same or related protein features they need to be combined into a computationally generated index. This new amino acid index can then be used as a representative of the entire group from which it was generated. As the similarity of the entire set of amino acids is low, a method is needed to group indices with high similarity. In order to achieve it, hierarchical clustering is used to create clusters of amino acid indices that represent similar features. Using principal component analysis to generate one single index that can represent these indices can then combine the clustered indices.

The chapter covers the methodology used for this analysis and it is organised as follows: Section 3.2 presents the methods (hierarchical clustering and principal component analysis) and materials (amino acid indices) used in this chapter for the analysis. Section 3.3 presents the results obtained from this analysis. Finally, concluding remarks are outlined in Section 3.4.

## 3.2 Methods and Materials

### 3.2.1 Amino Acid Indices

One of the largest sets of amino acid indices were published in the amino acid index database AAIndex [28]. The database consists of 544 indices, each of which is assigned a unique identification code. From this database, 13 amino acid indices were found to have missing index values. The identification codes of these indices are: AVBF000101, AVBF000102, AVBF000103, AVBF000104, AVBF000105, AVBF000106, AVBF000107, AVBF000108, AVBF000109, YANJ020101, GUYH850103, ROSM880104 and ROSM880105. In addition, three indices, namely RICJ880102, PRAM900102 and LEVM780102, have identical index values with RICJ880101, LEVM780101 and PRAM900101, respectively, but come under a unique accession ID.

From the literature, 83 recently published amino acid indices [29; 31; 34; 109; 110; 111; 112; 113] were discovered and added in the database and consequently brought the total number of unique indices to 611, which is one of the largest collection of non-redundant and unique datasets of amino acid indices. These are all listed in Appendix B.

### 3.2.2 Normalisation

As amino acid indices originated from different sources each amino acid index is normalised using z-score [114], as shown in Equation 3.1.

$$E' = \frac{E - \mu(E)}{\sigma(E)} \tag{3.1}$$

where $E$, $\mu$ and $\sigma$ correspond to index value, mean value and standard deviation for a particular amino acid index, respectively. By using z-score the comparison two or more amino acid indices is possible from different normal distributions.

### 3.2.3 Hierarchical Clustering Analysis

Cluster analysis is the method for allocating a set of objects into groups, called clusters. The included objects of each cluster would be more similar, in a measurable way, to each other than objects that belong to a different cluster. Hierarchical clustering is a statistical

method that tries to build clustered structures based on distances. In this chapter, the agglomerative hierarchical clustering method is used where each member of the dataset is assigned to a cluster. This method joins clusters pairwise as it moves one step in the hierarchy. The distance between clusters is measured between data points using the Euclidean distance as shown in Equation 3.2.

$$d = \sqrt{(A1_2 - A1_1)^2 + \cdots + (A20_2 - A20_1)^2} \tag{3.2}$$

where A1-A20 represents values of the 20 amino acids for each amino acid index.

The advantages of using hierarchical clustering in the analysis of amino acid indices are: 1) non-parametric method, 2) calculates a comprehensive hierarchy of clusters and 3) does not require for the number of clusters to be identified before the analysis. For this analysis three types of agglomerative hierarchical clustering, namely single, complete and average linkage, are considered and described below

1. Single linkage: The smallest distance (D) in objects (x) in clusters (r and s) is used.

$$D(r, s) = min(dist(x_{ri}, x_{sj})) \tag{3.3}$$

   where $i \in (i, \cdots, n_r)$, $j \in (1, \cdots, n_s)$, $n_r$ and $s_n$ are the number of objects in cluster $r$ and $s$, respectively. $x_{ri}$ is the $i$th object in cluster $r$ and $x_{sj}$ is the $j$th object in cluster $s$.

2. Complete linkage: the largest distance (D) in objects (x) in clusters (r and s) is used.

$$D(r, s) = max(dist(x_{ri}, x_{sj})) \tag{3.4}$$

   where $i \in (i, \cdots, n_r)$, $j \in (1, \cdots, n_s)$, $n_r$ and $s_n$ are the number of objects in cluster $r$ and $s$, respectively. $x_{ri}$ is the $i$th object in cluster $r$ and $x_{sj}$ is the $j$th object in cluster $s$.

3. Average linkage: the average distance (D) in objects (x) in clusters (r and s) is used.

$$D(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{rj}) \tag{3.5}$$

where $i \in (i, \cdots, n_r)$, $j \in (1, \cdots, n_s)$, $n_r$ and $s_n$ are the number of objects in cluster $r$ and $s$, respectively. $x_{ri}$ is the $i$th object in cluster $r$ and $x_{sj}$ is the $j$th object in cluster $s$.

The main difference between single and complete linkage is that the single linkage approach takes the minimum distance value whereas the complete linkage method takes the maximum distance value. The average linkage takes the average between all the pairs of objects between clusters. Furthermore, in order for the hierarchical clustering to be used a threshold (cut-off point) needs to be used to determine the minimum distance clusters should have from each other. For the analysis presented in this chapter, an optimal cut-off point will be determined by observing the dendrograms produced by the hierarchical cluster method.

### 3.2.4 Principal Component Analysis

Principal component analysis (PCA) [115] was originally introduced in 1901 by Karl Pearson [116]. PCA is a mathematical process that converts a set of related measurements into principal components, which is a set of linearly uncorrelated values, by utilising the orthogonal transformation. The amount of principal components produced by the PCA are equal to or less than the amount of the original measurements.

In this chapter, PCA is used to reduce the high dimensionality of the amino acid indices dataset to develop a smaller set of computational indices. These indices will inherit the biological features of the majority of the indices from the original set of amino acid indices. Using the hierarchical clustering techniques, the amino acid indices will be arranged into clusters, where their members will exhibit high similarity. PCA is then applied to each cluster of amino acids separately. For each cluster, the number of principal components extracted is equal to the number of variables being analysed where they represent a certain variance of the original set. The first principal component represents the highest variance in percentage and thus only this principal component is retained.

The set of values represented in the first component can be considered as a new computationally generated amino acid index that represents the original set of indices that exist in the cluster under investigation. By using a threshold value for the variance a set of computationally generated indices will be selected. An example of the process is the following:

- Based on the hierarchical clustering techniques a cluster of N amino acid indices is selected.
- By using these amino acid indices a matrix of $20xN$ is created.
- PCA is applied to the $20xN$ matrix, which will be converted into principal components.
- The first principal component of the generated matrix will contain a set of 20 values, which are the linear combination for the $N$ given amino acid indices, and the variance value which represents the percentage of representation.
- If the variance is equal to or higher than a pre-set threshold, this computationally generated index is selected.

## 3.3 Results

Results obtained through the analyses of amino acid indices by Hierarchical clustering and principal component analysis are presented. In addition, in order to show robustness of the newly discovered AAI results of the hybrid approach obtained through a case study is presented and discussed.

### 3.3.1 Hierarchical Clustering Analysis

By using the hierarchical clustering methods described in section 3.2.3, five different datasets of amino acid indices were produced. By the applying of a cut-off point to the hierarchical procedure, it can be defined how similar or dissimilar the amino acid indices would be in the generated clusters. The optimal cut-off point was derived by observing the dendrograms produced by the hierarchical clustering method. The optimal cut-off points for single and complete linkage were found to be 1.0 and 0.65, respectively; any higher cut-off point resulted in the majority of indices gathered in a single cluster, and any lower cut-off point a high percentage of clusters created contained only one amino acid index. Furthermore, the optimal cut-off point for average linkage was 1.0. For this case, the use of 0.65 as a cut-off point was considered but the clusters created were identical to the 1.0 cut-off point. In this case, the number of clusters produced are 107, 181 and 155 for single,

complete and average linkage, respectively. In the case where the cut-off point equals 0.65, 134 and 216 clusters were created for single and complete linkage, respectively. Figures 3.4, 3.5 and 3.6 shows the dendrograms produced by the hierarchical clustering techniques.

In this study, by looking into the clusters of amino acid indices, it can be observed that the arrangements within the clusters match the cross-correlation values between these indices; where two or more indices are clustered together the correlation between them is high. For example, amino acid indices 1 and 17 were clustered together using single linkage (1.0), single linkage (0.65) and complete linkage (0.65). Both index 1 and 17 relate to alpha-CH chemical shifts and have a correlation coefficient of 0.949. The maximum correlation coefficient between two amino acid indices is 1.

### 3.3.2 Principal Component Analysis Results

PCA was applied to create a computationally generated index to summarise each cluster. To evaluate if the computationally generated index retains the biological information from the individual the variance is utilised. By using the variance, which is calculated by PCA, the value of representation can be determined for the computationally generated index to the original set of indices in that cluster. The acceptance threshold for the variance is set to 99%, anything lower is discarded. The results in Table 3.1 show the number of computationally generated indices with variance $\geq$0.99. Furthermore, the discussion of the results will focus on the number of computationally generated indices where variance $\geq$0.99. The full set of results are available in the Amino Acid Index Database (AAID) web server as discussed in section 3.3.3.

### 3.3.3 Web-Server Access

In order to make the amino acid indices publicly available the Amino Acid Index Database (AAID) web-server is developed and publicly made available at
http://cisaps.com/indices/. The user can search throughout the database using a unique ID assigned to each amino acid index and search for specific amino acid indices. Additionally, the user can enter a feature-related keyword to find all the related indices. An example can be observed in Figures 3.1, 3.2 and 3.3 for stability amino acid indices. Figure 3.1 shows the initialised web-server. Figure 3.2 shows an example where the user requires amino indices for feature stability. In this case, the search term ”Stability” needs to be entered to

Table 3.1: PCA General Results

| Linkage | Cluster Threshold | No. of Computationally Generated Indices $\geq$ 0.99 Variance | No. of Computationally Generated Indices $\leq$ 0.99 Variance |
|---|---|---|---|
| Single | 1.0 | 64 | 43 |
| | 0.65 | 133 | 1 |
| Complete | 1.0 | 81 | 100 |
| | 0.65 | 216 | 0 |
| Average | 1.0 | 63 | 92 |
| | 0.4 | 63 | 92 |

the web-server as a query. As Figure 3.2 shows the AAID server returns three amino acid indices with identification numbers 441, 497 and 498. If a more detailed description of a selected amino acid index is required the user can search the amino acid index database using the identification number provided. An example is given in Figure 3.3 where a more comprehensive description is provided for the 441 amino acid index that represents side-chain contribution to protein stability (kJ/mol). Another feature of the web-server is that the user can retrieve the computationally generated indices from this investigation for each clustering method. Results for single linkage hierarchical clustering (Figures B.2 and B.3), complete linkage hierarchical clustering (Figures B.4 and B.5) and average linkage hierarchical clustering (Figure B.6) can be found in Appendix B.

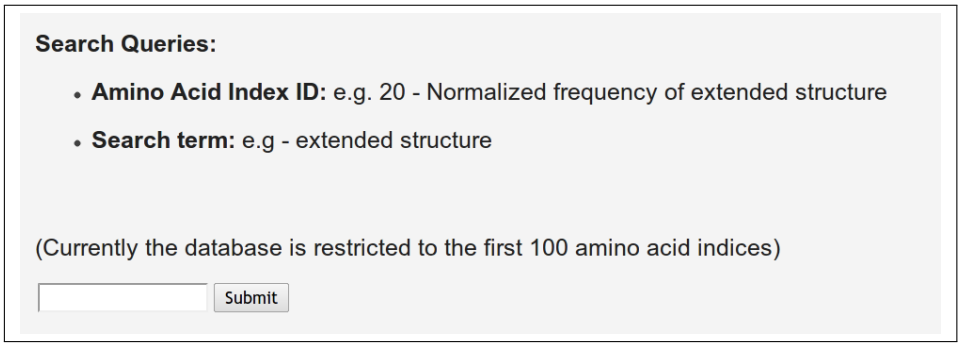
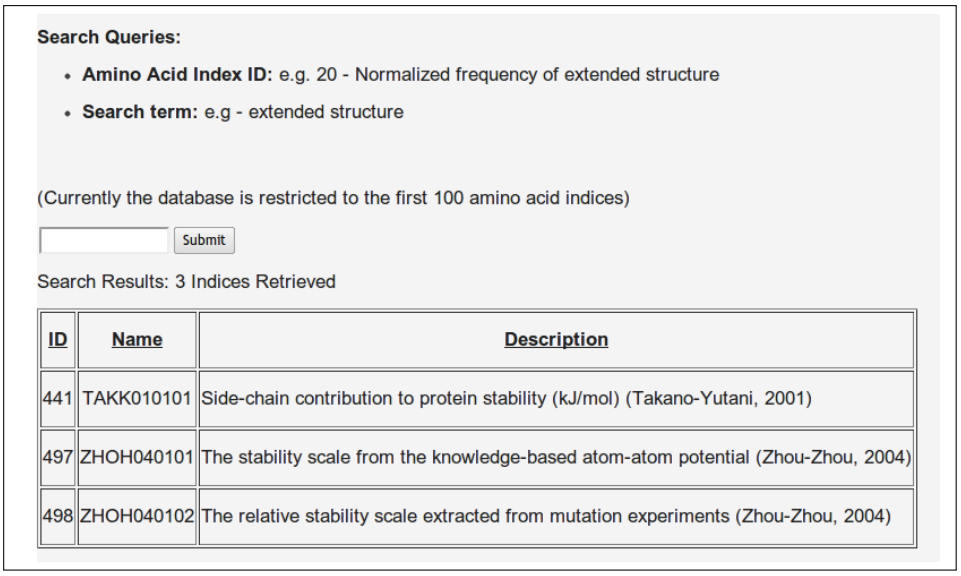



Figure 3.1: Amino Acid Index Database (AAID) Web Server



**Search Queries:**

- **Amino Acid Index ID:** e.g. 20 - Normalized frequency of extended structure
- **Search term:** e.g - extended structure

(Currently the database is restricted to the first 100 amino acid indices)

Submit

Search Results: 3 Indices Retrieved

| ID | Name | Description |
|---|---|---|
| 441 | TAKK010101 | Side-chain contribution to protein stability (kJ/mol) (Takano-Yutani, 2001) |
| 497 | ZHOH040101 | The stability scale from the knowledge-based atom-atom potential (Zhou-Zhou, 2004) |
| 498 | ZHOH040102 | The relative stability scale extracted from mutation experiments (Zhou-Zhou, 2004) |

Figure 3.2: AAID Web Server Search for ”Stability”

## 3.4 Conclusions

Amino acid indices have various applications and can represent diverse features of the protein sequences and amino acids. In this chapter, the largest up to date database is developed, which contains all the latest published amino acid indices and presented in the AAID web-server.

As the majority of indices included in the database have similar features, this chapter proposes a set of computationally derived indices that better represent the original group of indices. For this analysis hierarchical clustering and PCA is used. By using hierarchical clustering methods the amino acid indices with similar features are clustered together. The next step is to use PCA on these clusters to computationally derive an amino acid index that would be able to represent the original amino acid indices included in the cluster. In the original AAID database, 611 AAI exist. By using the computational generated AAI the search space can be reduced without losing any valuable information. The search space can be reduced to 542, 478, 521, 395 and 544 by using single linkage hierarchical clustering (1.0 and 0.65), complete linkage hierarchical clustering (1.0, 0.65) and average linkage hierarchical clustering (1.0 and 0.4) linkage, respectively.

**Search Queries:**

- **Amino Acid Index ID:** e.g. 20 - Normalized frequency of extended structure
- **Search term:** e.g - extended structure

(Currently the database is restricted to the first 100 amino acid indices)

Submit

Amino Acid Scale ID: **441**

Amino Acid Index Name: **TAKK010101 (Ref)**

Amino Acid Index Description: **Side-chain contribution to protein stability (kJ/mol) (Takano-Yutani, 2001)**

| Amino Acid | Value | Amino Acid | Value | Amino Acid | Value | Amino Acid | Value |
|---|---|---|---|---|---|---|---|
| Alanine (A) | 9.8 | Arginine (R) | 7.3 | Leucine (L) | 17.0 | lysine (K) | 10.5 |
| Asparagine (N) | 3.6 | Aspartic acid (D) | 4.9 | Methionine (M) | 11.9 | Phenylalanine (F) | 23.0 |
| Cysteine (C) | 3.0 | Glutamine (Q) | 2.4 | Proline (P) | 15.0 | Serine (S) | 2.6 |
| Glutamic acid (E) | 4.4 | Glycine (G) | 0.0 | Threonine (T) | 6.9 | Tryptophan (W) | 24.2 |
| Histidine (H) | 11.9 | Isoleucine (I) | 17.2 | Tyrosine (Y) | 17.2 | Valine (V) | 15.3 |

Figure 3.3: AAID Web Server Search for Amino Acid Index ID 411

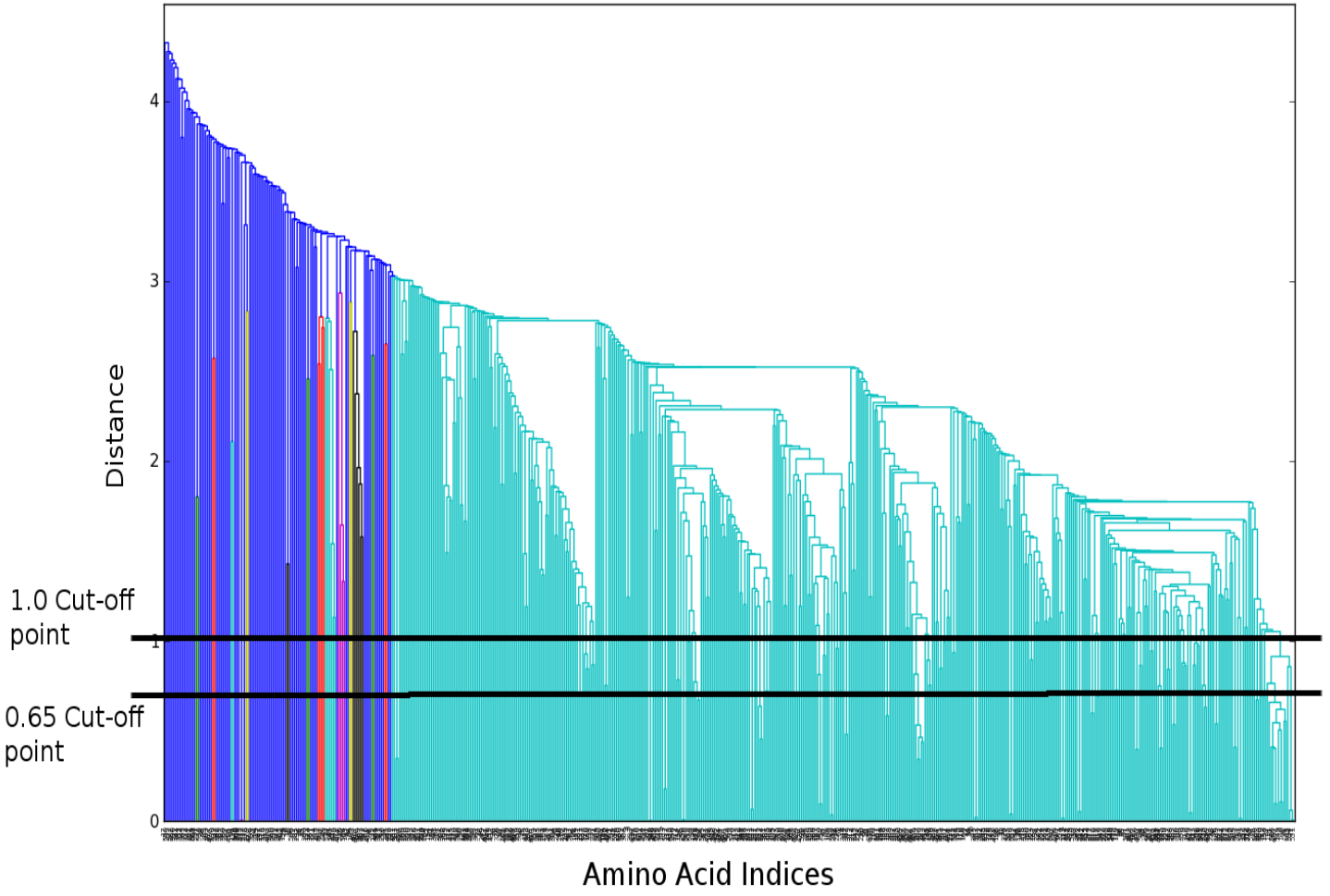



Figure 3.4: Clustering of Amino Acid Indices by using Single Linkage Hierarchical Clustering

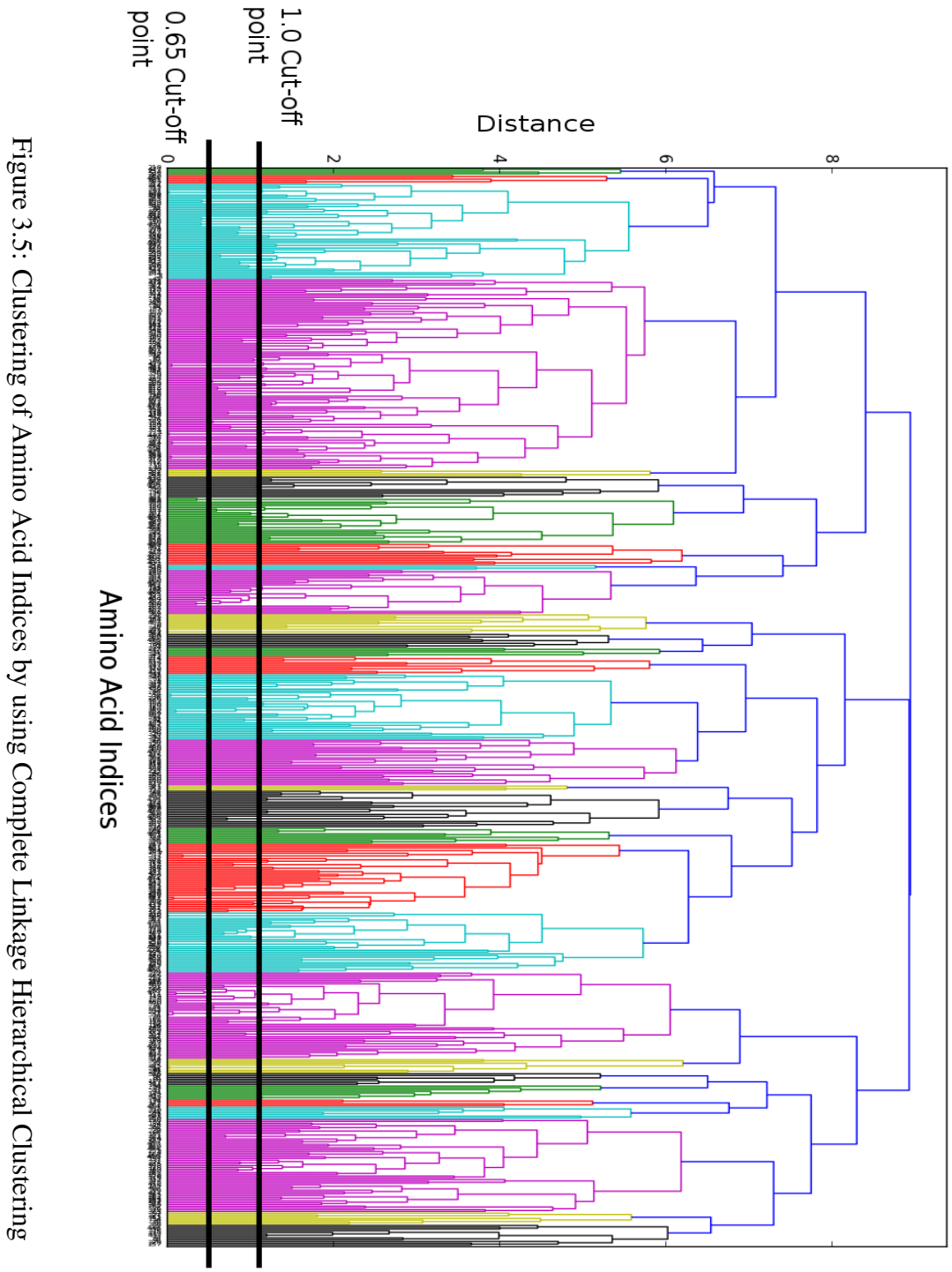



Figure 3.5: Clustering of Amino Acid Indices by using Complete Linkage Hierarchical Clustering

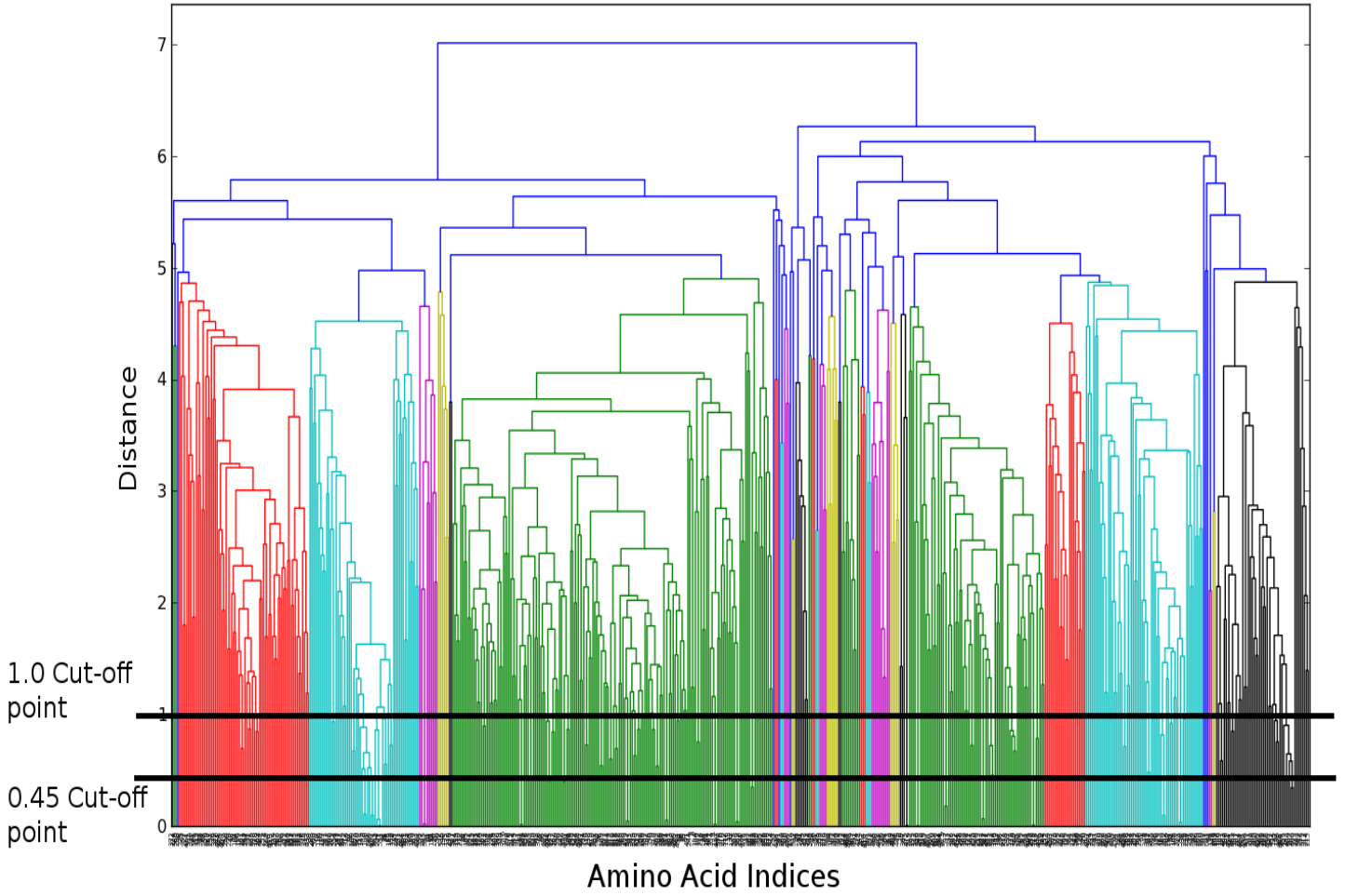



Figure 3.6: Clustering of Amino Acid Indices by using Average Linkage Hierarchical Clustering

# Chapter 4

# Signal Processing-based Bioinformatics Approach to Predict Protein Allergenicity.

## 4.1 Introduction

An allergy is a disorder, where a persons immune system reacts to substances from the environment that normally is considered as harmless [117]. An allergy is one of the four types of immune system hypersensitivity called Type I. In an allergic reaction, excessive activation of white blood cells are activated by an antibody called Immunoglobulin E (IgE). This reaction results in an inflammatory response from mild to life-threading symptoms.

Our current knowledge suggests that allergenic proteins constitute a very limited set of protein families (130 protein families of 9318 total protein sequences) inside the vastly diverse protein kingdom [118]. However, unifying features responsible for the allergenicity of such proteins have not been clearly determined yet [119; 120] despite numerous attempts dating back to the early 1970s [121]. Most recent studies aim to decipher the code of allergenicity, narrowing down to characteristic molecular features such as conserved physicochemical property motifs [118], enzymatic activity [119], IgE-binding sites [118; 122; 123; 124], T cell epitopes [125], receptor binding sites [126; 127; 128] or unique carbohydrates on the surface of allergenic proteins [129; 130; 131].

In an earlier study [132] it was proposed that hydrophobicity could be a characteristic

feature required to unleash the allergenic potential of a protein molecule, but the conclusions were drawn from a very limited set of allergenic and non-allergenic proteins, where there were only 37 allergens and 46 non-allergens in the 3NUL group, 9 allergens and 162 non-allergens in the 2ACT group, 27 allergens and 19 non-allergens in the 1BV1 group, and 22 allergens and 51 non-allergens in the 1QNX group, limiting the study to just 95 allergens from four protein families in total. They reported that study limitations were both due to a small set of known allergen sequences and homologous structures, as well as some technical limitations such as manual sequence input into the server and the computational capabilities of the ConSurf server [133].

In recent years, intelligent tools have been developed for detecting patterns and extracting features that can be used to describe allergenic protein sequences. Some of these tools use methods such as sequence similarity search [134; 135], allergen representative peptides (ARPs) [136], wavelet transforms [97], Support Vector Machines (SVM) [134; 137], k-Nearest-Neighbor (kNN) classifiers [138] and Gaussian classifiers [139]. These methods perform very accurately for high homology allergen sequences, whereas their performance is considerably weaker for low homology allergen sequences. Furthermore, additional studies showed that high homology between proteins is not directly linked to cross-reactivity [140; 141; 142] and to distinguish allergens from non allergen sequences that have the characteristics along with high homology remain challenging. Therefore, a new homology independent method is needed to be developed to determine if a protein is an allergen or not.

The aim of this study is therefore to differentiate sets of allergenic and non-allergenic proteins using a signal-processing based bioinformatics approach. In order to consider variations over the databases, three different databases of allergenic proteins, namely allergenonline [143], AllerHunter [134], and UniProt [4] are used to retrieve the allergen and non-allergen sequences. More importantly, unique sets of amino acid indices are identified that discriminate allergenic proteins from the non-allergenic ones by comparing diverse characteristics of constituting amino acids. In this analysis, it has been determined that relative partition energies [144], mean fraction area loss [145], and hydrophobicity [112] amino acid indices, can be used to discriminate allergenic proteins from the non-allergenic ones more accurately. Additionally, this study confirmed the results presented in previous studies [132], that suggest that hydrophobicity might be important for the overall allergenicity of a protein. The findings suggest novel characteristics of allergenic proteins, in

addition to the method of investigation that offers the advantages of supreme computational simplicity and greater computational power of evaluation in comparison with previous studies in this field.

## 4.2 Materials and Methods

In this chapter, a novel method is presented which uses Discrete Fourier Transform to extract information from protein sequences and Support Vector machines as a predictive tool for allergenic protein sequences. Additionally, the collection of allergenic and non-allergenic protein sequences from three online databases (allergenonline [143], AllerHunter [134], and UniProt [4]), will be discussed.

### 4.2.1 Discrete Fourier Transform

The Discrete Fourier Transform (DFT) is defined as follows

$$X(n) = \sum_{m=0}^{N-1} x(m) e^{-j(2\pi/N)nm} \quad n = 0, 1, ..., N-1 \tag{4.1}$$

where $x(m)$ is the $m$th member of the numerical series, N is the total number of points in the series, and $X(n)$ are coefficients of the DFT. As the DFT coefficients consisted of two mirror parts, only the first half of the series $(N/2)$ points) will be hereafter considered. The following formula determines the maximal frequency $(F)$ in the spectrum

$$F = \frac{1}{2d} \tag{4.2}$$

where $d$ is the distance between points of the sequence. If it is assumed that distance $d = 1$ then the maximum frequency in the spectrum can be found as $F = 1/2(1) = 0.5$.

The output of DFT is a complex sequence and can be characterised as

$$X(n) = (R(n) + jI(n)), \quad n = 0, 1, ..., (N-1)/2 \tag{4.3}$$

where $R(n)$ and $I(n)$ are the Real and Imaginary parts of the sequence, respectively.
The absolute spectrum can be formulated as

$$S_{(n)} = X(n)X^*(n) = |X(n)|^2, \quad n = 0, 1, ..., (N-1)/2 \tag{4.4}$$

where $S_{(n)}$ is the absolute spectrum for a specific protein, $X(n)$ are the DFT coefficients of the series $x(n)$ and $X*(n)$ are the complex conjugates.

To be able to apply DFT, protein sequences need to be encoded into numerical sequences by using an amino acid index as described in Chapter 3. By using the amino acid indices, the protein sequences can be encoded into numerical sequences in order for DFT to be applied. DFT coefficients can then be used to represent feature characteristics of the allergen and non-allergen protein sequences.

### 4.2.2 Preprocessing the Protein Sequences

Studies have shown that preprocessing of the signals by using zero-padding and windowing is generally necessary and can influence the features extracted from signal processing techniques [146]. Before applying DFT to the protein sequences, these two techniques (zero-padding and windowing) used in signal processing therefore need to be considered.

The first technique is the windowing where the encoded numerical sequences are multiplied by a pre-calculated window to reduce spectral leakage. In this case, the Hamming window [147] is used, and can be calculated using Equation 4.5.

$$w = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) \quad n = 0 \leq n \leq N-1 \tag{4.5}$$

The second technique is the zero-padding where a specified number of zero elements is added to the end of each sequence to increase signal length. This technique is essential as the given protein sequences may not be of the same length.

### 4.2.3 Support Vector Machine as a Predictive Tool

In the literature, Support Vector Machines (SVM) have been shown to be a valuable predictive tool in the analysis of protein sequences. Some examples of research areas that SVM was used with promising predictive results are the prediction of protein secondary structure [85], prediction of protein-protein interactions [148], prediction of RNA-binding proteins from primary sequence [149], prediction of protein subcellular localisation [150],

classification of enzyme family [151], and classification of G-protein coupled receptors [152].

A support vector machine (SVM) [153; 154], is a supervised statistical learning method that analyses data and recognises patterns for classification, which makes SVM a non-probabilistic linear classifier [153; 154]. The SVM finds the solutions for the optimisation problem as defined in Equation 4.6 by constructing hyperplanes into a higher or infinite dimensional space by the function $\phi$.

$$\begin{aligned} minimise\ w, b, \xi\ \ in \qquad & \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\ subject\ to \qquad & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad i = 1, ..., l \end{aligned} \tag{4.6}$$

where $x_i$ is a training set which is associated with labels $y_i$ where $x_i \in R^n$, $y \in \{-1, 1\}^l$ and C is the penalisation constant of the error term.

The goal of SVM is to find a linear separating hyperplane, maximum-margin hyperplane, that has the greater distance to the closest training point, the maximal margin, of all the given classes in this higher dimensional space. To construct a nonlinear classifier using SVM, a kernel function [155] needs to be applied to maximum-margin hyperplanes to fit in a transformed feature space. For this analysis, radial basis function (RBF) based kernel function was used as shown in Equation 4.7.

$$K(x_i, x_j) = e^{-\gamma ||x_i - x_y||^2} \quad \gamma > 0 \tag{4.7}$$

where $\gamma$, $r$ and $d$ are kernel parameters. For this analysis, LIBSVM [156] is used to construct the SVM-based classifier and grid search is applied to find the optimal values of the kernel $C$ and $\gamma$ parameters.

This type of kernel function is used because in many applications, RBF has been shown to be the simplest to adapt and the most generally applicable [157; 158].

### 4.2.4 Evaluating the Performance of the Predictive Models

For this analysis, the K-fold cross-validation technique [159] is used for measuring the performance of the allergen classifier as cross-validation is important for independently testing and validating different theories on existing data, where collecting additional data is impossible, costly or time consuming. This technique usually is used to approximate how these predictive models will behave and perform in practice. The first step of the cross-validation procedure involves dividing randomly the data into subsets [159]. The second step involves performing the analysis on one subset called the training set [159], and validate the analysis on the remaining subset(s) called the testing set. To decrease inconsistency the cross-validation technique is performed for multiple times using different partitions of the data for training and testing subsets. The final results are then calculated using the averaged performance of all the testing subsets [159].

In the literature, different cross validation methods exist. Three of the most common and widely used methods are K-fold cross-validation [160], repeated random sub-sampling validation [159], and leave-one-out cross-validation [159]. In this analysis, K-fold cross-validation is going to be used.

In K-fold cross-validation [160], the collected samples are randomly partitioned into K subsets, each one called a ”fold”. After partitioning the samples one subset is used as the testing set for validating the predictive model and the remaining (K-1) subsets are used as training sets. This process is repeated for K times until each fold is tested. The predictive accuracy for the K-fold cross-validation is calculated from the average result of the folds.

For this analysis, two different K-fold cross validation analyses were taken into consideration. For the first analysis, a 2-fold cross validation is performed for all amino acid indices that exist in the database, to be able to determine the most related features in classification of allergenic protein sequences. For the second analysis, for the highest rated feature an additional 5-fold cross-validation is performed, in order to determine accuracy of the SVM classifier .

The performance of the allergen classifier was evaluated based on sensitivity (SE), specificity (SP), geometric mean (G-mean) [161], F-measure [162], Matthews correlation coefficient (MCC) [163] and total accuracy (TACC). SP and SE correspond to the percentage of the correctly classified allergen and non-allergen proteins, respectively and TACC of the classifier represents the percentage of the correctly classified protein sequences. They

can be calculated by using the following equations, respectively

$$SP = \frac{TN}{TN + FP} \tag{4.8}$$

$$SE = \frac{TP}{TP + FN} \tag{4.9}$$

$$TACC = \frac{TP + TN}{TP + TN + FP + FN} * 100 \tag{4.10}$$

where true positives (TP) and true negatives (TN) represent the correctly identified allergen and non-allergen protein sequences, respectively. In addition, false negatives (FN) and false positives (FP) represents the misidentified allergen and non-allergen protein sequences, respectively.

Assessment of the performance of the classifiers by using SE, SP and TACC is not reliable due to the nature of imbalanced data sets [164]. Therefore, MCC, G-Means and F-measure are used additionally, which have been shown to provide more reliable comparison.

The MCC [163] is a measurement used in machine learning to validate a binary classifier and can successfully be used in validating imbalanced data sets. It can be calculated by using Equation 4.11

$$MCC = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TN + FN)(TP + FN)(TN + FP)(TP + FP)}} \tag{4.11}$$

The output of MCC is a value between [-1,1] where 1 represents 100% correct classification and -1 represents 100% misclassification. and 0 indicates random classification [163].

G-mean [161], another metric that uses SE and SP as given in Equation 4.12

$$GMean = \sqrt{SE \ * \ SP} \tag{4.12}$$

Finally, another measurement used to test the accuracy of the classifier is F-measure [162], as shown in Equation 4.13

$$F - Measure = \frac{2 * TP}{2 * TP + FN + FP} \tag{4.13}$$

The output of the F-measure is a value between [0,1] where 1 represents 100% correct classification and 0 represents 100% misclassification.

### 4.2.5 Allergenic Protein Databases

For this analysis, in order to address and investigate diversity over allergenic proteins data were collected from three online databases, allergenonline [143] (http://www.allergenonline.org), AllerHunter [134] (http://tiger.dbs.nus.edu.sg/AllerHunter), and UniProt [4] (http://www.uniprot.org). From UniProt, only the verified Allergens were considered as opposed to the allergenonline and AllerHunter database where all allergenic proteins available were considered. Table 4.1 lists the number of allergen and non-allergen proteins collected from each database. For the allergen dataset, 857, 1489 and 1416 protein sequences were retrieved from Uniprot, allergenonline and AllerHunter, respectively. For the non-allergen dataset, 1000 and 12474 protein sequences were retrieved from Uniprot and AllerHunter, respectively. In addition, Table 4.1 shows the maximum, minimum, and average length of the protein sequences for each dataset.

Table 4.1: Allergen and Non-Allergen Online Databases used in this study

| | **Allergen Proteins** | | | **Non Allergen Proteins** | |
|---|---|---|---|---|---|
| | UniProt | AllergenOnline | AllerHunter | UniProt | AllerHunter |
| No. Proteins | 857 | 1489 | 1416 | 1000 | 12474 |
| Max. Length | 1558 | 1662 | 1662 | 5890 | 7390 |
| Min. Length | 5 | 8 | 3 | 78 | 6 |
| Avg. Length | 235.65 | 227.06 | 221.67 | 752.56 | 380.52 |

Closer investigation was carried out to examine how these databases differ and thus to reveal diversity over different databases. The following list shows the number of overlapping protein sequences pairwise for all the Allergen databases as also shown in Figure 4.1. The complete list of protein sequences used in this study can be found in Appendix E.

- allergenonline versus AllerHunter: 886 common protein sequences
- allergenonline versus UniProt: 416 common protein sequences
- AllerHunter versus UniProt: 411 common protein sequences
- The common proteins for all three databases are 360.

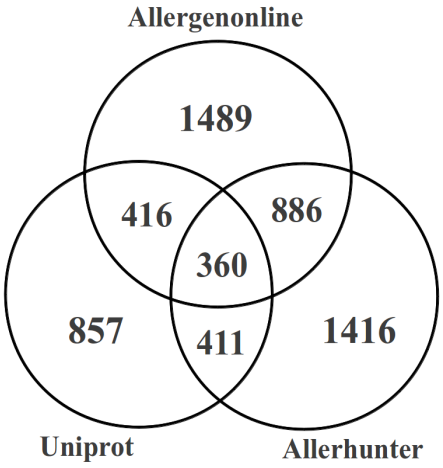



Figure 4.1: Allergen Databases and Number of Proteins

This result clearly shows the diversity of allergenic protein sequences and the inconsistency of allergenic databases. Therefore, the use of one database to construct a universal predictive model that predicts allergenic protein sequences may not be reliable. From the latest literature, AllerHunter is the most precise available online tool for classification of Allergenic protein sequences [134].

## 4.3 Results and Discussion

In this chapter, a study is performed in order to differentiate sets of allergenic and non-allergenic proteins using a signal-processing based bioinformatics approach. For this analysis, DFT coefficients were used to characterise protein sequences as well as multiple amino acid indices in order to encode protein sequences to numerical sequences. Finally, support vector machines were utilised as a predictive tool. In order to test how different databases affect the prediction of allergenic proteins, the following two main analyses were carried out.

Analysis 1: Allergenonline database that consists of 1404 allergen protein was used for training. For this analysis, 85 allergen proteins were found to overlap with the proteins in AllerHunter as an independent test set and therefore removed. Additionally, from the AllerHunter database, 9979 non-allergen proteins were used for training.

Analysis 2: UniProt database that consist of 817 allergen proteins was used for training. Additionally, from the UniProt database, 996 non-allergen proteins were used for training. For this analysis, 40 allergen proteins and 4 non-allergen proteins were found to overlap with validation proteins in the AllerHunter were used as an independent test and therefore removed.

To be able to have a truly accurate comparison between AllerHunter and the method proposed in this chapter, allergen and non-allergen protein sequences given by AllerHunter were used as an independent set. For this analysis, 139 allergen protein sequences and 1245 non allergen protein sequences provided by the AllerHunter web-server [134] were taken into consideration.

By using SVM and the 2-fold cross validation, ten amino acid indices that scored higher in relation to the classification accuracy are identified, as shown presented in Table 4.2. As the results show, hydrophobicity related amino acid indices appears multiple times in these higher related features. Although the hydrophobicity feature has been shown to be related to allergenic proteins [132; 165; 166; 167] which is further supported by this study, other protein features that are represented in amino acid indices, such as the relative partition energies and fractional area loss have equal or greater importance in classification of allergens. For the comprehensive analysis of allergenic protein sequences the highest rated amino acid index (optimised relative partition energies - method C) additional analysis by using SVM and 5-fold cross-validation was performed.

As Table 4.2 suggests that allergenic proteins seem to have distinct amino acid profiles, namely relative partition energies [144], mean fraction area loss [145], hydrophobicity [112; 112; 169], hydrophathy [168], relative amino acid closeness [29], medium-range nonbonded energy [111] and relative connectivity [29]. The results also reveal a novel list of amino acid indices that better discriminate allergenic proteins.

The results suggests that allergenic proteins are best characterised by distinct pattern of relative partition energies, i.e. allergens seem to have distinct conformational energy at the residue level determined by the amino acid indices [144], where inter-residue protein contact energies were estimated based on an equilibrium mixture approximation of residues. Pairwise contact energies for 20 types of residues have been determined from the observed frequencies of contacts with regression coefficients that were obtained by comparing ”input” and predicted values with the Bethe approximation [170] for the equilibrium mixtures of interacting residues. Optimised relative partition energies - method B involved deter-

mination of contact energies only, and Optimised relative partition energies - method C included the estimation of repulsive interaction energies in addition to contact energies.

The next best set of amino acid indices for discriminating allergens is found to be the mean fraction area loss [145]. For this protein feature, proteins of known structures were used to measure the average area that each residue buries upon protein folding; this area buried is correlated with residue hydrophobicity. For medium-range non-bonded energy [111] amino acid indices, studies were based on correlations of free energy change with sequence information and amino acid properties, including hydrophobicity [171], by using Bayesian-regularised genetic neural networks for creating predictive models for the conformational stability, (i.e. protein stability prediction). For relative connectivity amino acid indices exploits new residue networks [29] which have been constructed from the PDB structures of 640 representative proteins; the new indices have been derived from the amino acids in residue networks and were related to hydrophobicity and beta propensity. The hydrophathy index reposted by [168] was used for prediction of protein surface accessibility, based on information obtained from a single amino acid position or pair-information for a window of seventeen amino acids around the test residue. The hydrophobic free energies are directly related to the accessible surface area of both polar and nonpolar groups.

For this case study, two sets of independent test sets were created. For the first set, 262 allergen protein sequences were used. These sequences were obtained by adding the testing sequences of Analysis 1 (fold 5), and Analysis 2 (fold 2). Furthermore, any sequences existing in AllerHunter training data and the remaining folds of Analysis 1 and 2 were removed along with any duplicate proteins. For non-allergen protein sequences in the testing protein sequences of Analysis 1 (fold 5) was used by removing any duplicate protein sequences existing in AllerHunter, totaling 193 protein sequences. For the second independent test set, AllerHunter set was used.

For Analysis 1, using 5-folds cross validation and by using the independent test, the best results were obtained in fold 5. The results are, 0.9534, 0.9424, 0.9479, 0.7988, 0.7842 and 95.23% for sensitivity, specificity, G-Mean, F-measure, MCC and total accuracy, respectively. The average results of Analysis 1 are, 0.9436 $\pm$ 0.0080, 0.9381 $\pm$ 0.0064, 0.9409 $\pm$ 0.0048, 0.7685 $\pm$ 0.0224, 0.7535 $\pm$ 0.0223 and 94.31 $\pm$ 0.71%
for sensitivity, specificity, G-Mean, F-measure, MCC and total accuracy, respectively. For Analysis 2, the best results were obtained in fold 2. The results are, 0.9065, 0.9534, 0.9296, 0.7802, 0.7613 and 94.87% for sensitivity, specificity, G-Mean, F-measure, MCC and total

accuracy, respectively. The average results of Analysis 2 are, 0.8734 $\pm$ 0.0207, 0.9541 $\pm$ 0.0038, 0.9128 $\pm$ 0.0107, 0.7645 $\pm$ 0.0147, 0.7420 $\pm$ 0.0166 and 94.60 $\pm$ 0.38% for sensitivity, specificity, G-Mean, F-measure, MCC and total accuracy, respectively. The complete set of the results for Analysis 1 and 2 can be found in Tables 4.3 and 4.4, respectively. The results obtained from AllerHunter are, 0.8561, 0.9446, 0.8993, 0.7278, 0.7024 and 93.57% for sensitivity, specificity, G-Mean, F-measure, MCC and total accuracy, respectively. A summary of the first independent test set results, can be found at Figure 4.5.

The results using AllerHunter independent set are the following: for AllerHunter, 0.8264, 1.0, 0.9091, 0.9050, 0.8169 and 89.96 %, for Analysis 1, 0.8943, 0.9482, 0.99, 0.9258, 0.8346 and 91.70% and for Analysis 2, 0.9170, 0.9741, 0.9451, 0.9474, 0.8830 and 94.10% obtained for sensitivity, specificity, G-Mean, F-measure, MCC and total accuracy, respectively. The results can also be obtained from Table 4.6. As the results show Analysis 2 performs significantly better than AllerHunter and Analysis 1.

This analysis extracts features from DFT to build a classification model for allergenic proteins, and performs better in comparison to the AllerHunter classification model. As the results show, AllerHunter is more biased to non-allergen protein sequences as sensitivity, specificity and MCC values show. This method of classification of allergenic and non-allergenic protein sequence is more balanced, and it is better for creating a generalised classification model.

## 4.4 Conclusions

Although some homologous regions explained Immunoglobulin E (IgE) cross-reactivity in groups of allergens [165], no universal molecular structure could be associated with allergenicity as reported in previous studies [132; 165; 166; 167]. The study resulted in finding a unique and universal set of best discriminating amino acid indices for the characterisation of allergenic proteins. In summary, all amino acid indices identified from the study characterises different aspects of hydrophobicity of allergenic proteins (Table 4.2) with high sensitivity and specificity.

The results presented in this chapter support previous study [132], and earlier reports by others [165; 166; 167]. Some allergens have been reported previously as hydrophobic proteins, e.g. hydrophobic allergens from *Hevea brasiliensis* [166], soybean hydrophobic

protein [167] and certain hydrophobic lipid-binding lipocalins [165].

Hydrophobic substances have the tendency to form aggregates. Many allergens have been reported to have hydrophobic epitopes responsible for polymerisation, e.g. Ara h 1 was shown to form a highly stable homotrimer [172]. Interestingly, the majority of the IgE-binding epitopes are also located in the same hydrophobic regions at the distal ends of the three-dimensional structure where monomer-monomer contacts occur. This may suggest that further studies should be carried out to explore the role of the hydrophobic epitopes in polymerisation and the importance of quaternary protein structure in the overall allergenicity.

Mite allergen Der p 2, has structural homology with MD-2 protein, which is the lipopolysaccharide (LPS)-binding component of the Toll-like receptor 4 signalling complex [173]. The data further supports the suggestion of the study that allergens may possess intrinsic adjuvant activity due to their hydrophobicity and hence the ability to bind lipid entities, which could be a general functional property responsible for allergenicity.

Bioinformatics provides a powerful means of exploring protein structure and function from the data derived from topological properties of amino acids. Protein primary structure-based methods are less computationally intense and do not require X-ray crystal structure of proteins to be available, which makes such tools very suitable for the exploration of allergenic proteins with a still limited number of structures accessible.

Table 4.2: Top Amino Acid Indices in Classification of Protein Sequences

| Description of Amino Acid Indices | Reference | Sensitivity | Specificity | G-Mean | F-measure | MCC | Total Accuracy |
|---|---|---|---|---|---|---|---|
| Optimised relative partition energies - method C | [144] | 0.8458 | 0.896 | 0.8705 | 0.8836 | 0.7439 | 87.28% |
| Optimised relative partition energies - method B | [144] | 0.8411 | 0.896 | 0.8681 | 0.8819 | 0.7396 | 87.07% |
| Mean fractional area loss | [145] | 0.8364 | 0.89 | 0.8628 | 0.8768 | 0.7287 | 86.53% |
| Hydrophobicity | [112] | 0.8341 | 0.89 | 0.8616 | 0.876 | 0.7265 | 86.42% |
| Hydropathy index based on self-information values in the two-state model | [168] | 0.8271 | 0.882 | 0.8541 | 0.869 | 0.7113 | 85.67% |
| Relative Closeness | [29] | 0.8271 | 0.88 | 0.8531 | 0.8679 | 0.7091 | 85.56% |
| Hydrophobicity index | [169] | 0.8271 | 0.88 | 0.8531 | 0.8679 | 0.7091 | 85.56% |
| Medium-range nonbonded energy | [111] | 0.8248 | 0.876 | 0.85 | 0.8648 | 0.7026 | 85.24% |
| Hydrophobicity indices ph 7.5 | [112] | 0.8248 | 0.87 | 0.8471 | 0.8614 | 0.6961 | 84.91% |
| Relative connectivity | [29] | 0.8248 | 0.864 | 0.8442 | 0.858 | 0.6897 | 84.59% |

Table 4.3: Results of The Analysis 1 (With The Protein Sequences Obtained From UniProt)

| | | Sensitivity | Specificity | G-Mean | F-measure | MCC | Total Accuracy |
|---|---|---|---|---|---|---|---|
| Fold 1 | Testing Fold | 0.8191 | 0.9325 | 0.8740 | 0.8661 | 0.7484 | 87.02% |
| | Independent Test | 0.9502 | 0.9281 | 0.9391 | 0.7818 | 0.7654 | 94.80% |
| Fold 2 | Testing Fold | 0.8492 | 0.9571 | 0.9015 | 0.8940 | 0.8026 | 89.78% |
| | Independent Test | 0.9414 | 0.9424 | 0.9419 | 0.7638 | 0.7494 | 94.15% |
| Fold 3 | Testing Fold | 0.8945 | 0.8834 | 0.8889 | 0.8780 | 0.7771 | 88.95% |
| | Independent Test | 0.9341 | 0.9353 | 0.9347 | 0.7407 | 0.7256 | 93.42% |
| Fold 4 | Testing Fold | 0.8141 | 0.8957 | 0.8539 | 0.8439 | 0.7063 | 85.08% |
| | Independent Test | 0.9390 | 0.9424 | 0.9407 | 0.7572 | 0.7428 | 93.93% |
| Fold 5 | Testing Fold | 0.9394 | 0.9300 | 0.9347 | 0.9281 | 0.8678 | 93.42% |
| | Independent Test | 0.9534 | 0.9424 | 0.9479 | 0.7988 | 0.7842 | 95.23% |
| AVERAGE | Testing Fold | $0.8584 \pm 0.0450$ | $0.9216 \pm 0.0309$ | $0.8891 \pm 0.0277$ | $0.8804 \pm 0.0287$ | $0.7773 \pm 0.0549$ | $88.69 \pm 2.84$ % |
| $\pm$ std | Independent Test | $0.9436 \pm 0.0080$ | $0.9381 \pm 0.0064$ | $0.9409 \pm 0.0048$ | $0.7685 \pm 0.0224$ | $0.7535 \pm 0.0223$ | $94.31 \pm 0.71$% |

Table 4.4: Results of The Analysis 2 (With The Protein Sequences Obtained From Allergen Online)

| | | Sensitivity | Specificity | G-Mean | F-measure | MCC | Total Accuracy |
|---|---|---|---|---|---|---|---|
| Fold 1 | Testing Fold | 0.9181 | 0.9133 | 0.9157 | 0.7247 | 0.6981 | 91.39% |
| | Independent Test | 0.8777 | 0.9518 | 0.9140 | 0.7601 | 0.7378 | 94.44% |
| Fold 2 | Testing Fold | 0.9324 | 0.9674 | 0.9497 | 0.8618 | 0.8439 | 96.31% |
| | Independent Test | 0.9065 | 0.9534 | 0.9296 | 0.7802 | 0.7613 | 94.87% |
| Fold 3 | Testing Fold | 0.8968 | 0.9539 | 0.9249 | 0.8064 | 0.7813 | 94.69% |
| | Independent Test | 0.8561 | 0.9494 | 0.9016 | 0.7414 | 0.7164 | 94.00% |
| Fold 4 | Testing Fold | 0.8683 | 0.9369 | 0.9019 | 0.7496 | 0.7179 | 92.84% |
| | Independent Test | 0.8705 | 0.9566 | 0.9125 | 0.7707 | 0.7481 | 94.80% |
| Fold 5 | Testing Fold | 0.9214 | 0.9469 | 0.9341 | 0.8012 | 0.7781 | 94.37% |
| | Independent Test | 0.8561 | 0.9590 | 0.9061 | 0.7702 | 0.7464 | 94.87% |
| AVERAGE | Testing Fold | $0.9074 \pm 0.0254$ | $0.9437 \pm 0.0203$ | $0.9253 \pm 0.0181$ | $0.7888 \pm 0.0535$ | $0.7639 \pm 0.0578$ | $93.92 \pm 1.87\%$ |
| $\pm$ std | Independent Test | $0.8734 \pm 0.0207$ | $0.9541 \pm 0.0038$ | $0.9128 \pm 0.0107$ | $0.7645 \pm 0.0147$ | $0.7420 \pm 0.0166$ | $94.60 \pm 0.38\%$ |

Table 4.5: Predictive Accuracy Results Based on Independent Dataset 1

| Analysis | Sensitivity | Specificity | G-Mean | F-measure | MCC | Total Accuracy |
|---|---|---|---|---|---|---|
| AllerHunter | 0.8561 | 0.9446 | 0.8993 | 0.7278 | 0.7024 | 93.57% |
| Analysis 1 | 0.9534 | 0.9424 | 0.9479 | 0.7988 | 0.7842 | 95.23% |
| Analysis 2 | 0.9065 | 0.9534 | 0.9296 | 0.7802 | 0.7613 | 94.87% |

Table 4.6: Predictive Accuracy Results Based on Independent Dataset 2

| Analysis | Sensitivity | Specificity | G-Mean | F-measure | MCC | Total Accuracy |
|---|---|---|---|---|---|---|
| AllerHunter | 0.8264 | 1.0000 | 0.9091 | 0.9050 | 0.8169 | 89.96 % |
| Analysis 1 | 0.8943 | 0.9482 | 0.9209 | 0.9258 | 0.8346 | 91.70% |
| Analysis 2 | 0.9170 | 0.9741 | 0.9451 | 0.9474 | 0.8830 | 94.10% |

# 4. SIGNAL PROCESSING-BASED BIOINFORMATICS APPROACH TO PREDICT PROTEIN ALLERGENICITY.

# Chapter 5

# Multiple Protein Sequence Alignment Based on Multiple Amino Acid Indices and Discrete Fourier Transform

## 5.1 Introduction

In bioinformatics, a protein sequence alignment is a procedure that tries to arrange the sequences of proteins in order to identify high similarity regions that may be attributed to functional, structural, or evolutionary relationships between the sequences [16].

In recent years, Multiple Sequence Alignment (MSA) in protein sequences has become an important tool in bioinformatics and has successfully been applied in various fields such as pattern identification [174], domain identification [175], secondary structure prediction [11] and phylogenetic analysis [176]. Most of the algorithms developed to perform MSA generally use progressive alignment based on the Feng and Doolittle algorithm [177]. This type of algorithm is based on dynamic programming [178] to perform the sequence alignment.

Although there are various methods and tools that have been developed to perform MSA the dynamic programming is one of the most common methods [179; 180; 181; 182; 183; 184; 185; 186] for pairwise protein sequence alignments [178]. By using the dynamic programming, the optimal alignment is guaranteed. However, the use of dynamic programming for MSA is limited to a few small protein sequences [187; 188] due

to present limitations of processing power. To overcome these limitations, current methods use heuristics to process large sets of protein sequences [189]. Such methods are called progressive alignment algorithms and have been shown to perform MSA with adequate speed and higher accuracy [16; 190] in comparison to other MSA methods such as iterative methods [191] or Hidden Markov models [192]. By making the assumption that high homology between two protein sequences indicates that these sequences are evolutionarily related, MSA can be constructed using a series of pairwise alignments. The order of these alignments can be established by constructing and following a phylogenetic tree [177]. This algorithm initially aligns high related protein sequences using the dynamic programming and progressively includes less homologically related proteins. Progressive algorithms can perform better with protein sequence sets that have high overall similarity; an example is the alignment of protein sequences with known structure [11] to retrieve their related domains. However, for each MSA performed using the progressive methods, two issues need to be addressed to in order to improve the quality of the results:

- By using the progressive methods for MSA, there is no guarantee that the global optimal alignment will be found [193; 194]. Commonly, this issue is either raised when the relationships of protein sequences in the phylogenetic tree are not in the correct order or if errors occur in the early pairwise alignments and by adding less related proteins, these errors will increase. There is a close relationship between the homology percentage of sequence sets and these alignment errors. The higher the homology, the lower the MSA error congregated will be.

- For the progressive methods, substitution matrix and gap penalties are important parameters that play an important role [195]. In high homology protein sets, the use of these parameters will find the optimal or close to desirable alignment. Contrarily, with low homology protein sets, these parameters are very important as they will affect critically the MSA result. In order to obtain the highest accurate result a wide range of these parameters needs to be tested.

In order to overcome these two algorithmic problems a novel approach that utilises signal-processing techniques and multiple amino acid indices will be described in this chapter. The chapter is organised as follows: Section 5.2 presents briefly the existing methods that implement MSA along with their main differences. Additionally, section 5.2

describes the most commonly used substitution matrices along with the techniques used to construct them. Section 5.3 presents the methods and materials developed and used in this chapter, while Section 5.4 presents the results obtained. Finally, concluding remarks are outlined in Section 5.5.

## 5.2 Background

This section describes existing methods in the literature that implement progressive multiple sequence alignment techniques while focusing on their main differences. The methods described are FASTA [183], Clustal [184; 196], MAFFT [186] and T-Coffee [185].

### 5.2.1 Progressive Alignment Method

In order to use progressive alignment, a series of multiple pairwise alignments needs to be performed. The basic process of progressive alignment is the following:

- Step 1: Select two protein sequences by using a guide tree.
- Step 2: Align the two selected protein sequences by using a pairwise alignment method.
- Step 3: By using the guide tree, select another protein sequence and align it to the aligned sequence generated in Step 2.
- Step 4: Repeat Step 3 until all the protein sequences in the set are covered.

In the literature, their are various progressive alignment schemes, such as FASTA [183], Clustal [184; 196] and T-coffee [185]. These methods can differ in the way the guide tree is constructed, which gives the order of the pairwise alignments.

Progressive alignment method is a heuristic method and has the following characteristics:

- The function used to rate each pairwise alignment is the same as the function used for optimisation of MSA.

- The global optimal score of MSA is indirectly calculated only after a series of pairwise alignment score calculations.
- It does not directly optimise any global scoring function of alignment correctness.
- It can produce accurate results comparatively faster [16; 190] than other MSA methods like iterative methods [191] or Hidden Markov Models [192].

For the progressive alignment methods, one of the most essential heuristic properties is the selection and alignment of protein sequences with high homology first. These initial alignments produce the most reliable results. One of the most common algorithms used for progressive alignment is the Feng-Doolittle algorithm [177].

### 5.2.2 Multiple Sequence Alignment Methods

#### 5.2.2.1 FASTA

FASTA is one of the earliest and simplest protein and DNA multiple sequence alignment algorithm. This method was originally developed in 1985 by David Lipman and William Pearson [183] in the FASTP software package. In order to calculate similarity scores between protein sequences, FASTA uses the following steps to perform the protein sequence alignment:

- Step 1, the mutual regions with two-consecutive identities or regions with high concentration of single residue identities will be determined from pairwise protein sequences.
- Step 2, by using the Point accepted mutation (PAM250) substitution matrix [197], the shared regions identified in Step 1 will be re-evaluated. The region with the best score will be saved for future reference in a library.
- Step 3, the re-evaluated regions from Step 2 can be combined by using gaps. The aligned gapped protein sequences can be used to calculate a similarity score.
- Step 4, by using the Smith-Waterman algorithm [198] an optimum alignment score of the individual protein sequence with the library sequence will be constructed.

- Step 5, the algorithm returns the pairwise alignment and the optimal alignment score for each protein sequence.

#### 5.2.2.2 Clustal

Initially, in Clustal [184; 196] alignment algorithm, fast approximate methods [199] were used to calculate the pairwise distances and align a large number of proteins. The similarity scores for each protein sequence under investigation are calculated as the number of k-tuple matches. This is usually performed by calculating the identical amino acids from the best pairwise alignment between proteins and subtraction of a pre-determined penalty value for each gap.

In later years, more sophisticated versions of the program were developed. They are ClustalW [200], ClustalX [201] and ClustalΩ [202] that have been shown to calculate similarity scores for protein sequences more accurately. These programs utilise dynamic programming for protein sequence alignments. In addition, two separate gap penalties were used, for opening and extending gaps, and a full amino acid weight matrix. These similarity scores are estimated as the number of identities based on the best alignment divided by the number of residues compared and excluding the gap positions. These values are in the format of the similarity score's in percentage and can be converted into distances by using Equation 5.1. The distance can be used to construct the guide tree.

$$distance = 1 - (similarity\ score/100) \tag{5.1}$$

#### 5.2.2.3 T-Coffee

The T-Coffee (Tree-based Consistency Objective Function For alignment Evaluation) [185] algorithm has two main features for performing multiple sequence alignments; the first is the use of integrated data sources, and the second is the optimisation method. For the first attribute, the data sources used by T-Coffee are generated by a series of pairwise alignments. These data sources can contain both local and global pairwise alignments. For the second attribute, which is the optimisation method, attempts to discover the multiple alignment that fits better with the pairwise alignments in the data sources used.

This scheme implemented by T-Coffee resembles the progressive alignments methods used by other methods like ClustalW, which generally have the advantage of speed relative

to other methods (iterative methods or Hidden Markov models), and the results are comparatively robust [16; 190]. The main difference of T-Coffee with the other progressive alignment methods is that it does not follow a pre-constructed guide-tree to perform the pairwise alignments. T-Coffee utilises the information presented in the data sources to proceed with progressive alignment in a way that all the alignments between all the pairs of protein sequences are considered in each step of the multiple alignment. This process combines the advantages that progressive alignment methods have to offer, which is simplicity and speed, but with smaller probability to make errors.

#### 5.2.2.4 MAFFT

The MAFFT (Multiple Alignment using Fast Fourier Transform) method [186] performs multiple sequence alignments by using Fast Fourier transform (FFT). For this algorithm, the scoring system, which includes substitution matrix and gap penalties, was modified to increase accuracy. In a previous study [203] it is suggested that the Needleman-Wunsch (NW) algorithm [178], as used in ClustalW [200] and T-Coffee [185] performs well with all-positive distance matrices. In addition, when the efficiency of such a matrix was examined, only proteins with similar lengths were considered. For the alignment problems for protein sequences with different lengths, it is still not clear if all-positive substitution matrices are suitable. For MAFFT application, a new substitution matrix was used so that similarities can be represented by using both positive and negative values and can be found using Equation 5.2

$$\hat{M}_{ab} = \frac{M_{ab} + \sum_{a,b} f_a f_b M_{ab}}{(\sum_a f_a M_{aa} + \sum_{a,b} f_a f_b M_{ab})} + S^a \tag{5.2}$$

where $a$ and $b$ represent amino acids, $M_{ab}$ is the original all-positive substitution matrix, $f_a$ is the rate of occurrence of amino acid $a$, and $S^a$ is the gap extension penalty parameter. By using this new $M_{ab}$ substitution matrix, the value between two identical protein sequences is $1.0 + S^a$ whereas the similarity score between two unrelated protein sequences is $S^a$.

MAFFT performs multiple sequence alignments by using progressive alignment techniques. In order to improve the speed of the progressing algorithm in relation to other methods it uses Fast Fourier Transform to identify homologous regions in protein sequences. If the protein sequences under study have homologous regions, these regions will appear in several peaks in the correlation. As these peaks will only indicate the existence of homol-

ogous regions and not their location, a sliding window is used. The level of similarity of each region is measured based on the 20 highest peaks in the correlation. If consecutive homologous regions are identified, they are jointed into one region.

### 5.2.3 Substitution Matrix

In bioinformatics, a substitution matrix [197] expresses the rate at which one member in a sequence changes to another over time. Substitution matrices are generally used in protein or DNA sequence alignments, where the similarity between sequences relies upon their mutation rates as characterised in the matrix. There are various types of substitution matrices. Some of widely used and accepted matrices such as, PAM, BLOSUM, and GONNET are described below.

#### 5.2.3.1 Point Accepted Mutation (PAM)

Originally, Point Accepted Mutation (PAM) matrices [197] were introduced in the late 1970s and used in protein sequence alignments. In order to create the PAM matrix of protein mutation [204; 205], a Markov chain model [206] was utilised. These original PAM matrices were calculated based on 1572 measurements on mutation of 71 families with high similarity protein sequences.

For practical uses and to be able to compare different PAM matrices extracted from protein sequences with various lengths, PAM matrices are normalised. Therefore, the PAM matrix that is calculated from protein sequences with only one mutation occurring for every 100 amino acids will be called PAM1. Another example of the PAM matrix, PAM30, which is given in Table 5.1, is normally used in the literature for a protein sequence's alignment [197]. The PAM30 matrix supplies substitution probabilities for sequences where 30 mutations occur for every 100 amino acids. A substitution matrix can also be created by using the 20 standard amino acid indices, generating a 20 x 20 matrix where each position represents the probability of a given amino acid to be substituted by one of the remaining.

#### 5.2.3.2 Blocks of Amino Acid Substitution (BLOSUM)

The Blocks of Amino Acid Substitution (BLOSUM) matrix that is a substitution matrix used for protein sequence alignments was introduced in the early 90s [207]. In contrast

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | J | Z | X | * |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 6 | | | | | | | | | | | | | | | | | | | | | | | | |
| R | -7 | 8 | | | | | | | | | | | | | | | | | | | | | | | |
| N | -4 | -6 | 8 | | | | | | | | | | | | | | | | | | | | | | |
| D | -3 | -10 | 2 | 8 | | | | | | | | | | | | | | | | | | | | | |
| C | -6 | -8 | -11 | -14 | 10 | | | | | | | | | | | | | | | | | | | | |
| Q | -4 | -2 | -3 | -2 | -14 | 8 | | | | | | | | | | | | | | | | | | | |
| E | -2 | -9 | -2 | 2 | -14 | 1 | 8 | | | | | | | | | | | | | | | | | | |
| G | -2 | -9 | -3 | -3 | -9 | -7 | -4 | 6 | | | | | | | | | | | | | | | | | |
| H | -7 | -2 | 0 | -4 | -7 | 1 | -5 | -9 | 9 | | | | | | | | | | | | | | | | |
| I | -5 | -5 | -5 | -7 | -6 | -8 | -5 | -11 | -9 | 8 | | | | | | | | | | | | | | | |
| L | -6 | -8 | -7 | -12 | -15 | -5 | -9 | -10 | -6 | -1 | 7 | | | | | | | | | | | | | | |
| K | -7 | 0 | -1 | -4 | -14 | -3 | -4 | -7 | -6 | -6 | -8 | 7 | | | | | | | | | | | | | |
| M | -5 | -4 | -9 | -11 | -13 | -4 | -7 | -8 | -10 | -1 | 1 | -2 | 11 | | | | | | | | | | | | |
| F | -8 | -9 | -9 | -15 | -13 | -13 | -14 | -9 | -6 | -2 | -3 | -14 | -4 | 9 | | | | | | | | | | | |
| P | -2 | -4 | -6 | -8 | -8 | -3 | -5 | -6 | -4 | -8 | -7 | -6 | -8 | -10 | 8 | | | | | | | | | | |
| S | 0 | -3 | 0 | -4 | -3 | -5 | -4 | -2 | -6 | -7 | -8 | -4 | -5 | -6 | -2 | 6 | | | | | | | | | |
| T | -1 | -6 | -2 | -5 | -8 | -5 | -6 | -6 | -7 | -2 | -7 | -3 | -4 | -9 | -4 | 0 | 7 | | | | | | | | |
| W | -13 | -2 | -8 | -15 | -15 | -13 | -17 | -15 | -7 | -14 | -6 | -12 | -13 | -4 | -14 | -5 | -13 | 13 | | | | | | | |
| Y | -8 | -10 | -4 | -11 | -4 | -12 | -8 | -14 | -3 | -6 | -7 | -9 | -11 | 2 | -13 | -7 | -6 | -5 | 10 | | | | | | |
| V | -2 | -8 | -8 | -8 | -6 | -7 | -6 | -5 | -6 | 2 | -2 | -9 | -1 | -8 | -6 | -6 | -3 | -15 | -7 | 7 | | | | | |
| B | -3 | -7 | 6 | 6 | -12 | -3 | 1 | -3 | -1 | -6 | -9 | -2 | -10 | -10 | -7 | -1 | -3 | -10 | -6 | -8 | 6 | | | | |
| J | -6 | -7 | -6 | -10 | -9 | -5 | -7 | -10 | -7 | 5 | 6 | -7 | 0 | -2 | -7 | -8 | -5 | -7 | -7 | 0 | -8 | 6 | | | |
| Z | -3 | -4 | -3 | 1 | -14 | 6 | 6 | -5 | -1 | -6 | -7 | -4 | -5 | -13 | -4 | -5 | -6 | -14 | -9 | -6 | 0 | -6 | 6 | | |
| X | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | |
| * | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | -17 | 1 |

Table 5.1: PAM30 Substitution Matrix

to other substitution matrices like PAM, which are generated from comparisons of high similarity protein sequences, BLOSUM matrices are supported on observed alignments. In the literature, BLOSUM matrices are commonly used in multiple alignments between biological diverging protein sequences [208].

For building BLOSUM matrices, the BLOCKS database [209] was used to extract protein families with preserved regions. By using these protein sequences the relative frequencies of amino acids and their substitution probabilities were calculated. Furthermore, for the 210 possible substitutions of the 20 standard amino acids, the log-odds score was measured. A BLOSUM matrix can be constructed using Equation 5.3

$$S_{ij} = (\frac{1}{\lambda})log(\frac{p_{ij}}{q_i * q_j}) \tag{5.3}$$

where $i$ and $j$ are any two given amino acids, $p_{ij}$ is the probability of these two amino acids substituting each other in a sequence and $\lambda$ is a scaling factor. Finally, $q_i$ and $q_j$ represent the probability of arbitrarily finding amino acids $i$ and $j$, respectively, in any protein sequence.

Various BLOSUM matrices were constructed by using different protein families in recent years, and a numeric system based on protein similarity is used to differentiate these matrices. A high number attached to BLOSUM matrix, like BLOSUM80 will indicate that this matrix was designed for aligning protein sequences with high similarity, in contrast to a low number, like BLOSUM45 that will indicate being designed for aligning low similarity protein sequences. Table 5.2 shows a commonly used matrix, BLOSUM62 [210] as it has been shown to give a better representation, and improve the performance of protein multiple sequence alignments [211].

#### 5.2.3.3 GONNET

The GONNET matrix was introduced in 1992 by Gonnet, Cohen and Benner [212]. This type of matrix calculates the differences between amino acids by using exhaustive protein pairwise alignments. The first step to derive the GONNET matrix is to align the given protein sequences by using other substitution matrices like PAM or BLOSUM. The next step is to estimate the distance matrix by using the alignment, and interactively refine the alignment to calculate a new distance matrix. All the resulting matrices are normalised to 250 PAMs.

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X | * |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | | | | | | | | | | | | | | | | | | | | | | | |
| R | -1 | 5 | | | | | | | | | | | | | | | | | | | | | | |
| N | -2 | 0 | 6 | | | | | | | | | | | | | | | | | | | | | |
| D | -2 | -2 | 1 | 6 | | | | | | | | | | | | | | | | | | | | |
| C | 0 | -3 | -3 | -3 | 9 | | | | | | | | | | | | | | | | | | | |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | | | | | | | | | | | | | | | | | | |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | | | | | | | | | | | | | | | | | |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | | | | | | | | | | | | | | | | |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | | | | | | | | | | | | | | | |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | | | | | | | | | | | | | | |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | | | | | | | | | | | | | |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | | | | | | | | | | | | |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | | | | | | | | | | | |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | | | | | | | | | | |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | | | | | | | | | |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | | | | | | | | |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | | | | | | | |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | | | | | | |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | | | | | |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 | | | | |
| B | -2 | -1 | 3 | 4 | -3 | 0 | 1 | -1 | 0 | -3 | -4 | 0 | -3 | -3 | -2 | 0 | -1 | -4 | -3 | -3 | 4 | | | |
| Z | -1 | 0 | 0 | 1 | -3 | 3 | 4 | -2 | 0 | -3 | -3 | 1 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | | |
| X | 0 | -1 | -1 | -1 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -2 | 0 | 0 | -2 | -1 | -1 | -1 | -1 | -1 | |
| * | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | 1 |

Table 5.2: BLOSUM62 Substitution Matrix

As the authors indicated in the original description of the algorithm [212] the resulted matrix is affected by the homology of the proteins used. For this reason it is proposed for the initial alignment, the PAM250 substitution matrix to be used, and for the iterative alignment refinements, a PAM matrix to be used that is appropriate to the homology of the protein sequences used.

## 5.3 Methods and Materials

In this chapter a novel method is presented, which uses signal processing techniques, and more specifically Discrete Fourier Transform (DFT) along with multiple amino acid indices to perform MSA in protein sequences.

### 5.3.1 Feng-Doolittle Algorithm

In this chapter the Feng-Doolittle algorithm will be used in order to perform MSA. In order to apply this algorithm the following steps needs to be completed:

- By using DFT, as described in Section 5.3.3, a matrix N(N-1)/2 can be calculated which represents all the pairwise similarity scores between protein sequences.
- By using these distance values a guide tree can be constructed using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering method [213] as shown in Section 5.3.4.
- By using the guide tree the child nodes that present higher similarity are aligned first. By adding protein sequences to the alignment in the order they were appended to the guide tree a final alignment that covers all sequences can be obtained.

Each pairwise alignment is performed using dynamic programming, specifically Needleman - Wunsch (NW) algorithm [178]. A pseudo-code of the NW algorithm is given in the following algorithm.

```
1  Alignment_A = ''
2  Alignment_B = ''
3  size_A = length(A)
4  size_B = length(B)
```

```
5
6   while (size_A > 0 and size_B > 0)
7
8     Score = Score_function(size_A,size_B)
9     Diagonal_Score = Score_function(size_A-1, size_B-1)
10    Upper_Score = Score_function(size_A, size_B-1)
11    Left_Score = Score_function(size_A-1, size_B)
12
13    if (Score == Diagonal_Score + S(A(size_A), B(size_B)))
14      Alignment_A = A(size_A) + Alignment_A
15      Alignment_B = B(size_B) + Alignment_B
16      size_A = size_A-1
17      size_B = size_B-1
18
19    elif (Score == Left_Score + d)
20      Alignment_A = A(size_A) + Alignment_A
21      Alignment_B = '-' + Alignment_B
22      size_A = size_A-1
23
24    elif (Score == Upper_Score + d)
25      Alignment_A = '-' + Alignment_A
26      Alignment_B = B(size_B) + Alignment_B
27      size_B = size_B-1
28
29  while (size_A > 0)
30    Alignment_A = A(size_A) + Alignment_A
31    Alignment_B = '-' + Alignment_B
32    size_A = size_A-1
33
34  while (size_B > 0)
35    Alignment_A = '-' + Alignment_A
36    Alignment_B = B(size_B) + Alignment_B
37    size_B = size_B-1
```

After each pairwise alignment, each gap in the aligned sequences, which is denoted with the symbol '-' is replaced with the amino acid symbol 'X'. This action is performed to distinguish gaps inserted in previews alignments with current alignments and to ensure consistency. The algorithm time complexity is $O(nk)$ [214] for $k$ sequence alignments, where each sequence length is $n$.

### 5.3.2 Amino Acid Indices

In order to encode protein sequences to numerical sequences, amino acid indices need to be selected. In the literature, 611 amino acid indices exist, each one representing a unique biological protein feature. For this study, 25 amino acid indices were selected as shown in Table 5.3. These amino acid indices represent general and widely accepted features [22; 215; 216; 217] of the amino acids, like size [26], volume [25], molecular weight [24] and hydrophobicity [24; 218; 219; 220]. The complete list of the amino acid indices used for this analysis is presented in Table 5.3 and Table 5.4.

Table 5.3: Amino Acid Indices Used For The Alignment

| ID | Name | Description | Reference |
|---|---|---|---|
| 1 | ZIMJ680102 | Bulkiness | [221] |
| 2 | ZIMJ680104 | Isoelectric point | [221] |
| 3 | HUTJ700102 | Absolute entropy | [222] |
| 4 | DAWD720101 | Size | [26] |
| 5 | GRAR740102 | Polarity | [25] |
| 6 | GRAR740103 | Volume | [25] |
| 7 | FASG760101 | Molecular weight | [24] |
| 8 | FASG760102 | Melting point | [24] |
| 9 | FASG890101 | Hydrophobicity index | [24] |
| 10 | ZHOH040101 | The stability scale from the knowledge-based atom-atom potential | [223] |
| 11 | OOBM770103 | Long range non-bonded energy per atom | [224] |
| 12 | MANP780101 | Average surrounding hydrophobicity | [218] |
| 13 | WOLR790101 | Hydrophobicity index | [219] |
| 14 | FAUJ880101 | Hydration potential | [225] |
| 15 | FAUJ880102 | Smoothed upsilon steric parameter | [226] |
| 16 | ARGP820101 | Hydrophobicity index | [220] |
| 17 | VELV850101 | Electron-ion interaction potential | [87] |
| 18 | FAUJ880111 | Positive charge | [226] |
| 19 | FAUJ880112 | Negative charge | [226] |
| 20 | FAUJ880109 | Number of hydrogen bond donors | [226] |
| 21 | KYTJ820101 | Hydropathy index | [227] |
| 22 | BHAR880101 | Average flexibility indices | [228] |
| 23 | Proscale_4 | Recognition factors | [112] |
| 24 | Nl | Long-range contacts | [111] |
| 25 | Rk | Relative connectivity | [29] |

Table 5.4: Amino Acid Indices

| ID | Name | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ZIMJ680102 | 11.5 | 14.28 | 12.82 | 11.68 | 13.46 | 14.45 | 13.57 | 3.4 | 13.69 | 21.4 | 21.4 | 15.71 | 16.25 | 19.8 | 17.43 | 9.47 | 15.77 | 21.67 | 18.03 | 21.57 |
| 2 | ZIMJ680104 | 6 | 10.76 | 5.41 | 2.77 | 5.05 | 5.65 | 3.22 | 5.97 | 7.59 | 6.02 | 5.98 | 9.74 | 5.74 | 5.48 | 6.3 | 5.68 | 5.66 | 5.89 | 5.66 | 5.96 |
| 3 | HUTJ700102 | 30.88 | 68.43 | 41.7 | 40.66 | 53.83 | 46.62 | 44.98 | 24.74 | 65.99 | 49.71 | 50.62 | 63.21 | 55.32 | 51.06 | 39.21 | 35.65 | 36.5 | 60 | 51.15 | 42.75 |
| 4 | DAWD720101 | 2.5 | 7.5 | 5 | 2.5 | 3 | 6 | 5 | 0.5 | 6 | 5.5 | 5.5 | 7 | 6 | 6.5 | 5.5 | 3 | 5 | 7 | 7 | 5 |
| 5 | GRAR740102 | 8.1 | 10.5 | 11.6 | 13 | 5.5 | 10.5 | 12.3 | 9 | 10.4 | 5.2 | 4.9 | 11.3 | 5.7 | 5.2 | 8 | 9.2 | 8.6 | 5.4 | 6.2 | 5.9 |
| 6 | GRAR740103 | 31 | 124 | 56 | 54 | 55 | 85 | 83 | 3 | 96 | 111 | 111 | 119 | 105 | 132 | 32.5 | 32 | 61 | 170 | 136 | 84 |
| 7 | FASG760101 | 89.09 | 174.2 | 132.12 | 133.1 | 121.15 | 146.15 | 147.13 | 75.07 | 155.16 | 131.17 | 131.17 | 146.19 | 149.21 | 165.19 | 115.13 | 105.09 | 119.12 | 204.24 | 181.19 | 117.15 |
| 8 | FASG760102 | 297 | 238 | 236 | 270 | 178 | 185 | 249 | 290 | 277 | 284 | 337 | 224 | 283 | 284 | 222 | 228 | 253 | 282 | 344 | 293 |
| 9 | FASG890101 | -0.21 | 2.11 | 0.96 | 1.36 | -6.04 | 1.52 | 2.3 | 0 | -1.23 | -4.81 | -4.68 | 3.88 | -3.66 | -4.65 | 0.75 | 1.74 | 0.78 | -3.32 | -1.01 | -3.5 |
| 10 | ZHOH040101 | 2.18 | 2.71 | 1.85 | 1.75 | 3.89 | 2.16 | 1.89 | 1.17 | 2.51 | 4.5 | 4.71 | 2.12 | 3.63 | 5.88 | 2.09 | 1.66 | 2.18 | 6.46 | 5.01 | 3.77 |
| 11 | OOBM770103 | -0.491 | -0.554 | -0.382 | -0.356 | -0.67 | -0.405 | -0.371 | -0.534 | -0.54 | -0.762 | -0.65 | -0.3 | -0.659 | -0.729 | -0.463 | -0.455 | -0.515 | -0.839 | -0.656 | -0.728 |
| 12 | MANP780101 | 12.97 | 11.72 | 11.42 | 10.85 | 14.63 | 11.76 | 11.89 | 12.43 | 12.16 | 15.67 | 14.9 | 11.36 | 14.39 | 14 | 11.37 | 11.23 | 11.69 | 13.93 | 13.42 | 15.71 |
| 13 | WOLR790101 | 1.12 | -2.55 | -0.83 | -0.83 | 0.59 | -0.78 | -0.92 | 1.2 | -0.93 | 1.16 | 1.18 | -0.8 | 0.55 | 0.67 | 0.54 | -0.05 | -0.02 | -0.19 | -0.23 | 1.13 |
| 14 | FAUJ880101 | 1.28 | 2.34 | 1.6 | 1.6 | 1.77 | 1.56 | 1.56 | 0 | 2.99 | 4.19 | 2.59 | 1.89 | 2.35 | 2.94 | 2.67 | 1.31 | 3.03 | 3.21 | 2.94 | 3.67 |
| 15 | FAUJ880102 | 0.53 | 0.69 | 0.58 | 0.59 | 0.66 | 0.71 | 0.72 | 0 | 0.64 | 0.96 | 0.92 | 0.78 | 0.77 | 0.71 | 0 | 0.55 | 0.63 | 0.84 | 0.71 | 0.89 |
| 16 | ARGP820101 | 0.61 | 0.6 | 0.06 | 0.46 | 1.07 | 0 | 0.47 | 0.07 | 0.61 | 2.22 | 1.53 | 1.15 | 1.18 | 2.02 | 1.95 | 0.05 | 0.05 | 2.65 | 1.88 | 1.32 |
| 17 | VELV850101 | 0.037 | 0.096 | 0.004 | 0.126 | 0.083 | 0.076 | 0.006 | 0.005 | 0.024 | 0 | 0 | 0.037 | 0.082 | 0.095 | 0.019 | 0.083 | 0.094 | 0.055 | 0.052 | 0.006 |
| 18 | FAUJ880111 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | FAUJ880112 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | FAUJ880109 | 0 | 4 | 2 | 1 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 21 | KYTJ820101 | 1.8 | -4.5 | -3.5 | -3.5 | 2.5 | -3.5 | -3.5 | -0.4 | -3.2 | 4.5 | 3.8 | -3.9 | 1.9 | 2.8 | -1.6 | -0.8 | -0.7 | -0.9 | -1.3 | 4.2 |
| 22 | BHAR880101 | 0.357 | 0.529 | 0.463 | 0.511 | 0.346 | 0.493 | 0.497 | 0.544 | 0.323 | 0.462 | 0.365 | 0.466 | 0.295 | 0.314 | 0.509 | 0.507 | 0.444 | 0.305 | 0.42 | 0.386 |
| 23 | Proscale.4 | 78 | 95 | 94 | 81 | 89 | 87 | 78 | 84 | 84 | 88 | 85 | 87 | 80 | 81 | 91 | 107 | 93 | 104 | 84 | 89 |
| 24 | NI | 3.92 | 3.78 | 3.64 | 2.85 | 5.55 | 3.06 | 2.72 | 4.31 | 3.77 | 5.58 | 4.59 | 2.79 | 4.14 | 4.53 | 3.57 | 3.75 | 4.09 | 4.83 | 4.93 | 5.43 |
| 25 | Rk | 1.05 | 0.94 | 0.93 | 0.88 | 1.17 | 0.93 | 0.85 | 0.99 | 0.99 | 1.11 | 1.07 | 0.88 | 1.04 | 1.07 | 0.92 | 0.96 | 0.99 | 1.05 | 1.05 | 1.12 |

### 5.3.3 Substitution Matrix

In literature, as described in section 5.2.3, three substitution matrices are commonly used, PAM, BLOSUM and GONNET.

Important limitations have been observe in recent works [109; 229; 230; 231] such as

- PAM Matrix Disadvantages
  - Assumes uniform distribution of all mutation types.
  - Uses high homology proteins in order to deduce relationships in diverse proteins.
- BLOSUM Matrix Disadvantages
  - Limited to a subset of conserved domains.
  - Ignores the closeness of relationship between proteins.

In order for GONNET matrix to be constructed, PAM and BLOSUM matrices are used along with exhaustive pairwise sequence alignments, thus it will inherit the same general disadvantages as the substitution matrix it uses.

In recent years, numerous substitution matrices have been developed that include physical characteristics of proteins, such as local sequence-structure information [232], alpha-helix information, secondary structure information [233] and solvent accessibility states [234; 235]. These substitution matrices are shown to have generally improved protein sequence alignments.

In this chapter, a novel similarity (or substitution) matrix constructed and presented. This substitution matrix is not considered based on the mutation rate, like PAM [197] or BLOSUM [207] matrices, but on the physicochemical properties of each amino acid. In order to calculate the substitution matrix, the amino acids need to be converted to numerical values. These values can be derive from the amino acid indices.

As detailed in Chapter 3, 611 amino acid indices exist in the literature, each representing a unique biological feature, which can encode amino acids. By using the numerical representation of the amino acids the pairwise Euclidean distance between all the amino acids can be calculated. The Euclidean distance between two amino acids x and y where

$x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ and n is the number of features, can be calculated by Equation 5.4.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \tag{5.4}$$

Further information regarding the amino acid indices used to calculate the substitution matrix can be found in Section 5.3.2.

By using the proposed substitution matrix the following advantages can be obtained compared to the classical substitution matrices such as PAM, BLOSUM and GONNET:

- The proposed substitution matrix is not biased to specific groups of protein sequences [236] as the values are calculated from the amino acid indices, and not from the protein sequences.
- By using the classical substitution matrix, the use of a different matrix can have a major impact on the alignment [236]. For the proposed substitution matrix, the same matrix can be considered regardless of the protein sequence's homology to be aligned or the mutation rate presented.
- A correlation to the physical characterisations of the amino acids that the substitution matrix derived from can be achieved.
- Different similarity matrices can be generated when different physical characterisations of amino acids are considered. These characteristics are represented by the amino acid indices.

### 5.3.4 Construction of Dendrogram

The first step in constructing a dendrogram, which is used as a guide tree for MSA, is to calculate the distance matrix between all the protein sequences. The following steps need to be completed in order to calculate the distance matrix.

- Each protein sequence in the dataset is converted to 25 numerical sequences using the amino acid indices shown in Table 5.3.

- As different protein sequences are likely to have distinct lengths, each numerical sequence is zero-padded to match the length of the longest protein sequence in the dataset. This step is essential for comparing multiple proteins with different lengths.

- By using DFT as described in Equations 5.5 and 5.6, the absolute frequency spectra is calculated for each of the 25 numerical sequences for all the protein sequences.

- For each protein sequence, the 25 absolute frequency spectra is combined into one vector. By using the correlation distance, as Equation 5.7 shows, the distance matrix for all the protein sequences can be calculated.

$$X(n) = \sum_{m=0}^{N-1} x(m) e^{-j(2\pi/N)nm} \quad n = 0, 1, ..., N-1 \tag{5.5}$$

where $x(m)$ is the $m$th member of the numerical series, N is the total number of points in the series, and $X(n)$ are coefficients of the DFT. As the DFT coefficients consisted of two mirror parts, only the first half of the series $(N/2)$ points) will be hereafter considered. The following formula determines the absolute frequency spectrum

$$S_{(n)} = X(n)X^*(n) = |X(n)|^2, \quad n = 0, 1, ..., (N-1)/2 \tag{5.6}$$

where $S_{(n)}$ is the absolute spectrum for a specific protein, $X(n)$ are the DFT coefficients of the series $x(n)$ and $X^*(n)$ are the complex conjugate.
The correlation distance can be calculated as follows

$$D(X, Y) = 1 - \frac{(X - \bar{X}) \cdot (Y - \bar{Y})}{\| (X - \bar{X}) \|_2 \| (Y - \bar{Y}) \|_2} \tag{5.7}$$

where $\bar{X}$ and $\bar{Y}$ represent the mean values of vectors $X$ and $Y$, respectively.

After calculating the distance matrix, the guide tree can be constructed. The method used to build the guide tree is the Unweighted Pair Group Method with Arithmetic Mean (UPGMA). Moreover, UPGMA is a hierarchical clustering method created by Sokal and Michener [213] and used in bioinformatics for the creation of phenetic trees. Initially, the UPGMA method was used in protein electrophoresis studies, but currently it is a common practice to use it in complex protein or DNA phylogenetic analyses, or in MSA to built the

guide trees. A detailed description of the procedure that the UPGMA algorithm follows to create a rooted dendrogram is given below:

1. Calculate the distance matrix $D$ (Eq. 5.7) for all the protein sequences.
2. Initialise the dendrogram by creating one leaf node for each protein sequence in the dataset.
3. Find X and Y protein sequences that have the smallest distance $D_{X,Y}$ between them.
4. Create a new node XY.
5. Connect X and Y, and give the two branches connecting X and Y to XY with the length $D_{X,Y}/2$.
6. Compute the distance from the new node to all the remaining nodes of the dendrogram as a weighted average using Equation 5.8.

$$D_{XY,Z} = (\frac{n_X}{n_X + n_Y})D_{X,Z} + (\frac{n_Y}{n_X + n_Y})D_{Y,Z} \tag{5.8}$$

7. Append the distances calculated above in D matrix and delete the distances that correspond to the nodes X and Y.
8. Repeat Steps 3-7 until all the protein sequences are covered.

### 5.3.5 Case Study: Multiple Sequence Alignment of Cluster of Differentiation 4 Proteins

The Cluster of Differentiation 4 (CD4) [237] is a glycoprotein and was discovered in late 1970. CD4 is expressed on the surface of T helper cells, monocytes, macrophages, and dendritic cells. The main function of CD4 is to act as a co-receptor that supports the T-cell receptor with an antigen-presenting cell. In recent years, CD4 has become subject to an intense research towards finding a cure for Human immunodeficiency virus (HIV). Protein sequence alignment is important as it can identify regions of similarity that can be considered significant in regards to the functional, structural or evolutionary relationships between the protein sequences. HIV-1 uses CD4 to infect a host T-cell and accomplishes

this by binding gp120, a known viral envelope protein with CD4 [238]. In this study, 32 CD4 protein sequences, as listed in Table 5.5, will be used. These protein sequences were collected from UniProt [4] for various animal species.

Table 5.5: CD4 Proteins

| ID | Uniprot ID | Organism | Protein Length |
|---|---|---|---|
| 1 | P01730 | Human | 458 |
| 2 | P16004 | Chimpanzee | 458 |
| 3 | P79185 | Crab-eating Macaque | 458 |
| 4 | P79184 | Japanese Macaque | 458 |
| 5 | P16003 | Rhesus Macaque | 458 |
| 6 | Q08340 | Pig-tailed Macaque | 458 |
| 7 | Q29037 | Common Squirrel Monkey | 457 |
| 8 | Q08338 | Green Monkey | 458 |
| 9 | P06332 | Mouse | 457 |
| 10 | P46630 | Rabbit | 459 |
| 11 | P05540 | Rat | 457 |
| 12 | Q6R3N3 | Pig | 417 |
| 13 | A7YY52 | Bovine | 395 |
| 14 | NP_001123374 | Sheep | 455 |
| 15 | ACG76115 | Goat | 455 |
| 16 | Q8HZT8 | White-tufted-ear Marmoset | 457 |
| 17 | P33705 | Dog | 463 |
| 18 | AAB24450 | Cat | 474 |
| 19 | Q9XS78 | Beluga Whale | 455 |
| 20 | Q71QE2 | Bottle-nosed Dolphin | 455 |
| 21 | NP_001092760 | Gray short-tailed Opossum | 461 |
| 22 | ABR22561 | Tammar Wallaby | 464 |
| 23 | AAS67020 | Chicken | 487 |
| 24 | CAP04927 | Turkey | 487 |
| 25 | B8YEL3 | Domestic Duck | 482 |
| 26 | AAW63065 | Muscovy Duck | 482 |
| 27 | BAD37153 | Fugu Rubripes | 463 |
| 28 | B5D5T7 | European Seabass | 480 |
| 29 | ADM47441 | Orange-spotted Grouper | 469 |
| 30 | ADV78594 | Mandarin Fish | 549 |
| 31 | ABU95653 | Spotted-green Pufferfish | 466 |
| 32 | ACM50926 | Atlantic Halibut | 461 |

## 5.4 Results and Discussions

Two sets of results are obtained and presented. The first set of results presents the generated dendrograms for the protein sequences. These dendrograms (figures 5.1, 5.2 and 5.3) are essential in current MSA methods as they can be used as guide trees for the order in which protein sequences are to be aligned. The second set of results refers to application of MSA for protein sequences using the modified Feng-Doolittle algorithm as described in section 5.2.1.

### 5.4.1 Dendrogram

Figures 5.1, 5.2 and 5.3 show the dendrograms generated by using MAFFT, ClustalW2 and T-Coffee, respectively. The online European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI) webserver [239] was used to construct the dendrograms for MAFFT, ClustalW2 and T-Coffee. The dendrogram generated by using UPGMA method presented in this chapter is given in Figure 5.4.

By using the algorithm as described in Section 5.3.4, distance matrices and dendrogram can be generated by using any given set of amino acid indices, each one of which represents a unique biological feature of protein sequences. In this section, the results are based on the combination of 25 widely accepted amino acid indices [22; 215; 216; 217], which produced the best results, according to the biological relationships between proteins, for constructing dendrograms and performing MSA for protein sequences. These amino acid indices can be used to construct a dendrogram based on a specific individual protein feature. An example of these dendrograms can be seen in Appendix C.

By using Table 5.6, which represents the pairwise percent identity of the protein sequences used, and considering the assumption that a high-quality guide tree joins the first sequences that present the highest similarity, the algorithm presented in this chapter generated the best dendrogram according to the biological relationships between proteins. A good example can be seen with CD4 protein sequence extracted from rabbit, which belongs to glires. Glires is a clade consisting of rodents (mice and rat) and lagomorphs (rabbits and hares). As Table 5.6 shows, by using only the percent identity, the biological similarity of two protein sequences is not directly observable. By using the percent identity, the CD4 protein with higher similarity for rabbits is extracted from pigs. The pairwise percent iden-

Table 5.6: Pairwise Percent Identity of CD4 Proteins

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 98 | 60 | 90 | 61 | 91 | 91 | 91 | 91 | 54 | 61 | 54 | 80 | 60 | 23 | 23 | 16 | 18 | 81 | 24 | 60 | 57 | 60 | 61 | 22 | 60 | 41 | 15 | 39 | 19 | 21 | 18 |
| 2 | | 60 | 90 | 61 | 91 | 91 | 91 | 90 | 54 | 62 | 53 | 80 | 61 | 23 | 21 | 18 | 18 | 80 | 24 | 60 | 57 | 60 | 61 | 22 | 59 | 40 | 16 | 39 | 19 | 17 | 18 |
| 3 | | | 59 | 63 | 59 | 60 | 59 | 59 | 49 | 58 | 46 | 56 | 63 | 18 | 20 | 14 | 18 | 56 | 21 | 61 | 56 | 62 | 71 | 20 | 61 | 39 | 15 | 39 | 14 | 17 | 21 |
| 4 | | | | 60 | 95 | 95 | 95 | 94 | 54 | 61 | 53 | 79 | 60 | 23 | 24 | 20 | 18 | 80 | 23 | 59 | 57 | 59 | 62 | 21 | 58 | 39 | 15 | 38 | 15 | 18 | 18 |
| 5 | | | | | 61 | 60 | 61 | 60 | 50 | 60 | 50 | 59 | 97 | 19 | 19 | 17 | 18 | 59 | 23 | 73 | 67 | 73 | 63 | 24 | 72 | 40 | 18 | 40 | 16 | 17 | 19 |
| 6 | | | | | | 99 | 99 | 98 | 53 | 61 | 53 | 79 | 60 | 24 | 23 | 16 | 17 | 80 | 22 | 60 | 58 | 60 | 62 | 21 | 60 | 40 | 15 | 38 | 18 | 16 | 20 |
| 7 | | | | | | | 99 | 98 | 54 | 61 | 53 | 79 | 60 | 24 | 23 | 16 | 18 | 80 | 22 | 60 | 58 | 60 | 62 | 22 | 59 | 40 | 15 | 38 | 18 | 17 | 18 |
| 8 | | | | | | | | 98 | 53 | 61 | 53 | 79 | 60 | 24 | 23 | 16 | 18 | 80 | 22 | 60 | 58 | 60 | 62 | 21 | 59 | 40 | 15 | 38 | 18 | 17 | 18 |
| 9 | | | | | | | | | 54 | 60 | 53 | 79 | 60 | 24 | 23 | 15 | 18 | 79 | 21 | 60 | 58 | 60 | 62 | 21 | 59 | 39 | 15 | 39 | 17 | 17 | 18 |
| 10 | | | | | | | | | | 52 | 74 | 54 | 49 | 22 | 22 | 16 | 14 | 53 | 23 | 49 | 46 | 50 | 51 | 23 | 50 | 37 | 14 | 39 | 16 | 16 | 16 |
| 11 | | | | | | | | | | | 52 | 59 | 60 | 22 | 22 | 19 | 17 | 59 | 24 | 59 | 55 | 60 | 59 | 24 | 60 | 39 | 15 | 38 | 12 | 9 | 19 |
| 12 | | | | | | | | | | | | 53 | 49 | 18 | 21 | 17 | 11 | 53 | 21 | 49 | 46 | 50 | 50 | 21 | 49 | 38 | 18 | 37 | 17 | 13 | 19 |
| 13 | | | | | | | | | | | | | 58 | 23 | 24 | 15 | 17 | 90 | 22 | 58 | 55 | 57 | 59 | 22 | 57 | 41 | 20 | 38 | 18 | 14 | 16 |
| 14 | | | | | | | | | | | | | | 19 | 23 | 17 | 18 | 58 | 21 | 72 | 68 | 73 | 64 | 23 | 72 | 40 | 20 | 40 | 18 | 17 | 19 |
| 15 | | | | | | | | | | | | | | | 91 | 15 | 17 | 22 | 62 | 24 | 23 | 24 | 22 | 62 | 24 | 22 | 18 | 24 | 16 | 17 | 11 |
| 16 | | | | | | | | | | | | | | | | 14 | 18 | 23 | 64 | 24 | 25 | 25 | 22 | 63 | 25 | 24 | 19 | 23 | 15 | 15 | 10 |
| 17 | | | | | | | | | | | | | | | | | 47 | 15 | 14 | 18 | 18 | 14 | 16 | 16 | 14 | 20 | 50 | 17 | 49 | 52 | 62 |
| 18 | | | | | | | | | | | | | | | | | | 17 | 18 | 17 | 20 | 18 | 13 | 17 | 19 | 15 | 45 | 17 | 56 | 55 | 46 |
| 19 | | | | | | | | | | | | | | | | | | | 21 | 58 | 56 | 57 | 59 | 22 | 56 | 40 | 20 | 36 | 20 | 16 | 20 |
| 20 | | | | | | | | | | | | | | | | | | | | 23 | 22 | 23 | 21 | 83 | 23 | 18 | 17 | 21 | 15 | 17 | 13 |
| 21 | | | | | | | | | | | | | | | | | | | | | 63 | 68 | 60 | 22 | 68 | 40 | 16 | 40 | 13 | 16 | 17 |
| 22 | | | | | | | | | | | | | | | | | | | | | | 84 | 58 | 23 | 83 | 34 | 20 | 36 | 17 | 12 | 17 |
| 23 | | | | | | | | | | | | | | | | | | | | | | | 62 | 23 | 97 | 39 | 21 | 41 | 16 | 14 | 17 |
| 24 | | | | | | | | | | | | | | | | | | | | | | | | 22 | 62 | 39 | 16 | 38 | 17 | 14 | 19 |
| 25 | | | | | | | | | | | | | | | | | | | | | | | | | 22 | 20 | 16 | 20 | 15 | 15 | 17 |
| 26 | | | | | | | | | | | | | | | | | | | | | | | | | | 39 | 21 | 40 | 19 | 12 | 17 |
| 27 | | | | | | | | | | | | | | | | | | | | | | | | | | | 16 | 60 | 17 | 17 | 18 |
| 28 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 14 | 50 | 55 | 48 |
| 29 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 15 | 15 | 18 |
| 30 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 55 | 48 |
| 31 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 49 |

Table 5.7: Distance Matrix of CD4 Proteins

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.01 | 0.78 | 0.11 | 0.81 | 0.12 | 0.11 | 0.11 | 0.12 | 0.81 | 0.8 | 0.8 | 0.59 | 0.8 | 0.9 | 0.9 | 0.95 | 0.98 | 0.58 | 0.94 | 0.76 | 0.75 | 0.79 | 0.76 | 0.92 | 0.79 | 0.94 | 0.98 | 0.9 | 1 | 0.99 | 0.96 |
| 2 | | 0.78 | 0.11 | 0.81 | 0.13 | 0.12 | 0.12 | 0.12 | 0.81 | 0.8 | 0.8 | 0.59 | 0.8 | 0.9 | 0.9 | 0.95 | 0.98 | 0.58 | 0.94 | 0.76 | 0.76 | 0.8 | 0.76 | 0.92 | 0.8 | 0.94 | 0.98 | 0.9 | 1 | 1 | 0.96 |
| 3 | | | 0.76 | 0.78 | 0.76 | 0.74 | 0.75 | 0.75 | 0.75 | 0.67 | 0.76 | 0.88 | 0.79 | 1.01 | 1 | 0.92 | 0.96 | 0.87 | 0.94 | 0.78 | 0.75 | 0.8 | 0.49 | 0.95 | 0.8 | 0.91 | 0.93 | 0.85 | 0.98 | 1.03 | 0.96 |
| 4 | | | | 0.81 | 0.07 | 0.06 | 0.06 | 0.06 | 0.82 | 0.8 | 0.81 | 0.63 | 0.8 | 0.93 | 0.93 | 0.96 | 0.98 | 0.61 | 0.93 | 0.77 | 0.75 | 0.8 | 0.74 | 0.92 | 0.81 | 0.93 | 0.98 | 0.88 | 1 | 0.99 | 0.97 |
| 5 | | | | | 0.81 | 0.8 | 0.8 | 0.81 | 0.8 | 0.73 | 0.84 | 0.77 | 0.03 | 0.96 | 0.95 | 0.97 | 0.97 | 0.78 | 1.02 | 0.45 | 0.4 | 0.35 | 0.81 | 1.01 | 0.35 | 0.88 | 0.95 | 0.98 | 0.99 | 1.01 | 0.98 |
| 6 | | | | | | 0.02 | 0.01 | 0.02 | 0.84 | 0.82 | 0.83 | 0.63 | 0.79 | 0.95 | 0.95 | 0.95 | 0.98 | 0.63 | 0.94 | 0.77 | 0.75 | 0.79 | 0.74 | 0.93 | 0.8 | 0.94 | 0.98 | 0.89 | 1 | 0.99 | 0.98 |
| 7 | | | | | | | 0 | 0.01 | 0.83 | 0.81 | 0.82 | 0.63 | 0.79 | 0.94 | 0.94 | 0.95 | 0.98 | 0.62 | 0.94 | 0.77 | 0.74 | 0.78 | 0.73 | 0.93 | 0.79 | 0.93 | 0.98 | 0.88 | 1 | 1 | 0.98 |
| 8 | | | | | | | | 0.01 | 0.83 | 0.81 | 0.82 | 0.63 | 0.79 | 0.95 | 0.95 | 0.96 | 0.98 | 0.62 | 0.94 | 0.76 | 0.74 | 0.78 | 0.73 | 0.93 | 0.79 | 0.94 | 0.98 | 0.89 | 1 | 1 | 0.98 |
| 9 | | | | | | | | | 0.83 | 0.81 | 0.82 | 0.63 | 0.8 | 0.95 | 0.94 | 0.95 | 0.98 | 0.62 | 0.94 | 0.76 | 0.74 | 0.78 | 0.73 | 0.93 | 0.79 | 0.94 | 0.98 | 0.88 | 1 | 0.99 | 0.98 |
| 10 | | | | | | | | | | 0.68 | 0.49 | 0.83 | 0.79 | 0.95 | 0.95 | 0.99 | 0.94 | 0.81 | 0.99 | 0.78 | 0.8 | 0.78 | 0.71 | 1.01 | 0.78 | 0.95 | 0.98 | 0.95 | 0.97 | 1 | 0.94 |
| 11 | | | | | | | | | | | 0.65 | 0.73 | 0.73 | 0.96 | 0.94 | 0.92 | 0.88 | 0.7 | 0.98 | 0.74 | 0.79 | 0.8 | 0.71 | 0.96 | 0.79 | 0.94 | 0.96 | 0.93 | 1.02 | 1.03 | 0.91 |
| 12 | | | | | | | | | | | | 0.8 | 0.84 | 0.92 | 0.92 | 0.94 | 0.91 | 0.8 | 0.97 | 0.82 | 0.89 | 0.85 | 0.71 | 0.99 | 0.85 | 0.95 | 0.99 | 0.89 | 0.99 | 1.01 | 0.96 |
| 13 | | | | | | | | | | | | | 0.77 | 0.93 | 0.92 | 0.93 | 0.94 | 0.15 | 1 | 0.81 | 0.79 | 0.78 | 0.85 | 1.01 | 0.79 | 0.93 | 1 | 0.97 | 0.98 | 1 | 0.92 |
| 14 | | | | | | | | | | | | | | 0.96 | 0.95 | 0.97 | 0.96 | 0.78 | 1.02 | 0.44 | 0.38 | 0.34 | 0.81 | 1.02 | 0.34 | 0.88 | 0.96 | 0.98 | 1 | 1.01 | 0.98 |
| 15 | | | | | | | | | | | | | | | 0.12 | 0.94 | 0.96 | 0.92 | 0.76 | 0.96 | 0.95 | 0.95 | 1.02 | 0.76 | 0.95 | 0.9 | 0.93 | 1 | 0.96 | 0.99 | 0.94 |
| 16 | | | | | | | | | | | | | | | | 0.95 | 0.94 | 0.91 | 0.76 | 0.94 | 0.93 | 0.94 | 1.01 | 0.76 | 0.93 | 0.9 | 0.93 | 0.99 | 0.97 | 0.99 | 0.95 |
| 17 | | | | | | | | | | | | | | | | | 0.81 | 0.91 | 0.97 | 0.97 | 0.97 | 0.99 | 0.99 | 0.96 | 1 | 0.98 | 0.79 | 0.98 | 0.83 | 0.9 | 0.69 |
| 18 | | | | | | | | | | | | | | | | | | 0.93 | 0.96 | 1 | 0.94 | 0.95 | 0.97 | 0.99 | 0.96 | 0.96 | 0.86 | 0.99 | 0.8 | 0.91 | 0.74 |
| 19 | | | | | | | | | | | | | | | | | | | 1 | 0.79 | 0.79 | 0.79 | 0.86 | 1 | 0.8 | 0.92 | 0.99 | 0.96 | 0.99 | 1 | 0.91 |
| 20 | | | | | | | | | | | | | | | | | | | | 1.03 | 0.98 | 1.01 | 0.96 | 0.23 | 1.01 | 0.89 | 0.99 | 0.92 | 0.95 | 0.96 | 0.94 |
| 21 | | | | | | | | | | | | | | | | | | | | | 0.48 | 0.48 | 0.76 | 1.01 | 0.46 | 0.89 | 0.95 | 0.94 | 0.99 | 1 | 0.97 |
| 22 | | | | | | | | | | | | | | | | | | | | | | 0.18 | 0.76 | 0.98 | 0.2 | 0.83 | 0.96 | 0.94 | 0.95 | 0.99 | 0.96 |
| 23 | | | | | | | | | | | | | | | | | | | | | | | 0.79 | 1.01 | 0.04 | 0.85 | 0.97 | 0.97 | 0.95 | 0.98 | 0.97 |
| 24 | | | | | | | | | | | | | | | | | | | | | | | | 0.97 | 0.79 | 0.95 | 0.98 | 0.89 | 0.99 | 1.02 | 0.97 |
| 25 | | | | | | | | | | | | | | | | | | | | | | | | | 1.01 | 0.91 | 1 | 0.92 | 0.98 | 0.99 | 0.95 |
| 26 | | | | | | | | | | | | | | | | | | | | | | | | | | 0.83 | 0.95 | 0.95 | 0.95 | 0.96 | 0.98 |
| 27 | | | | | | | | | | | | | | | | | | | | | | | | | | | 0.95 | 0.69 | 0.99 | 0.95 | 0.93 |
| 28 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0.97 | 0.86 | 0.84 | 0.79 |
| 29 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1.02 | 0.94 | 1 |
| 30 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0.86 | 0.83 |
| 31 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0.93 |

tity is calculated to be 74%, 54% and 52% between rabbit - pig, rabbit - mouse and rabbit - rat CD4 protein sequences, respectively. In addition, by using the dendrograms generated from MAFFT 5.1, ClustalW 5.2, and T-Coffee 5.3 methods, there is no direct association connecting all three CD4 proteins (rabbit, mouse and rat). By using the method proposed in this chapter, there is a clearer indication as seen in Figure 5.4, associating rabbit, mouse and rat CD4 proteins. As the dendrogram show, for the proposed method mouse and rat CD4 protein sequences will be aligned first, and thereafter, the aligned protein sequence will be aligned to rabbit CD4 protein sequences.

As stated in the previous sections, by making the assumption that high similarity between protein sequences indicates that these sequences are evolutionarily related, MSA can be constructed by using a series of pairwise alignments. For illustration and validation purposes, two sections of the guide tree are selected to be aligned in which all the protein sequence present high similarity according to the distance matrix. The first subgroup selected as seen in Figure 5.5, represents all the CD4 protein sequences extracted from primates [240]. The second group selected as shown in Figure 5.6 represent all the CD4 protein sequences extracted from animals that belong to the Artiodactyla order [240].

### 5.4.2 Multiple Sequence Alignment

The first step in MSA, after constructing the guide tree, is to calculate the substitution matrix. In order to calculate the substitution matrix, the amino acids need to be converted to numerical sequences. By using Table 5.3 that lists the selected 25 amino acid indices, each amino acid was converted into 25 numerical values. By using the numerical representation of the amino acids, the pairwise Euclidean distance between all amino acids was calculated. Table 5.8 shows the generated substitution matrix. For the gap penalty, the value -16 was used, and it was calculated by $4 * min(SM)$, where SM represents the substitution matrix values, given in Table 5.8. The formula for the gap penalty value was obtained throughout extensive experimental testing that was performed for this case study.

By using the Feng-Doolittle algorithm as described in section 5.3.1, MSA was performed for the two subgroups of protein sequences that were selected. By using the EMBL-EBI webserver, MSA for these selected subgroups was performed by using MAFFT, ClustalW2 and T-Coffee methods with their default parameters, which produced the MSA with the higher percent identity. The results for the first MSA, which includes the protein sequences

## 5. MULTIPLE PROTEIN SEQUENCE ALIGNMENT BASED ON MULTIPLE AMINO ACID INDICES AND DISCRETE FOURIER TRANSFORM

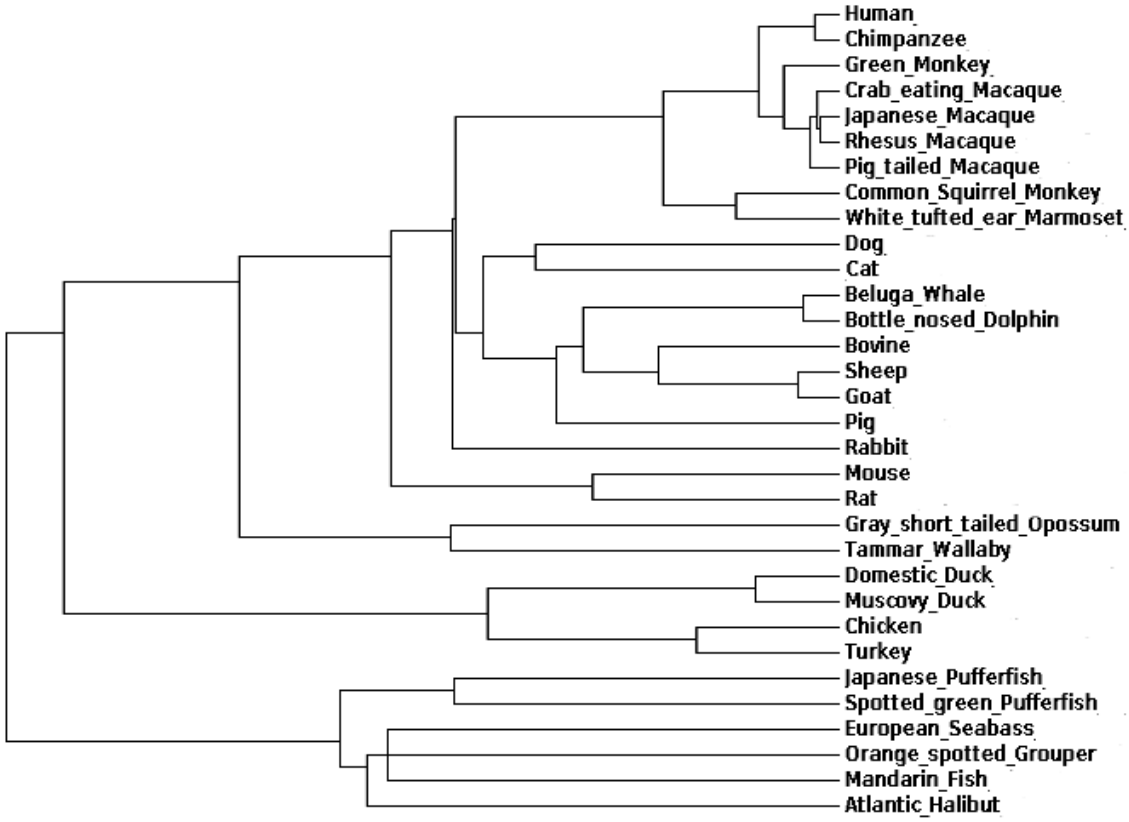



Figure 5.1: Dendrogram Constructed By Using MAFFT Method

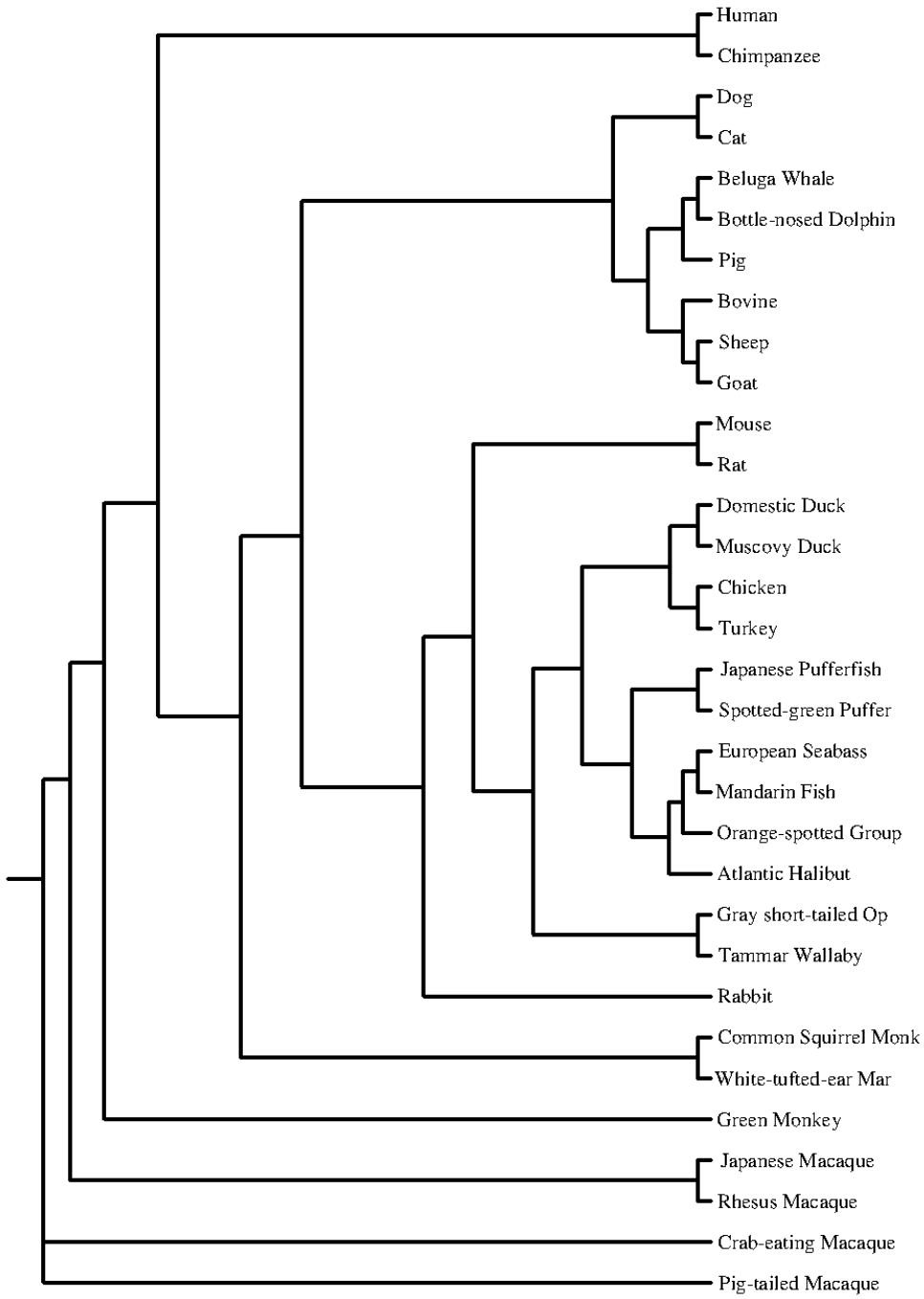



Figure 5.2: Dendrogram Constructed By Using ClustalW2 Method

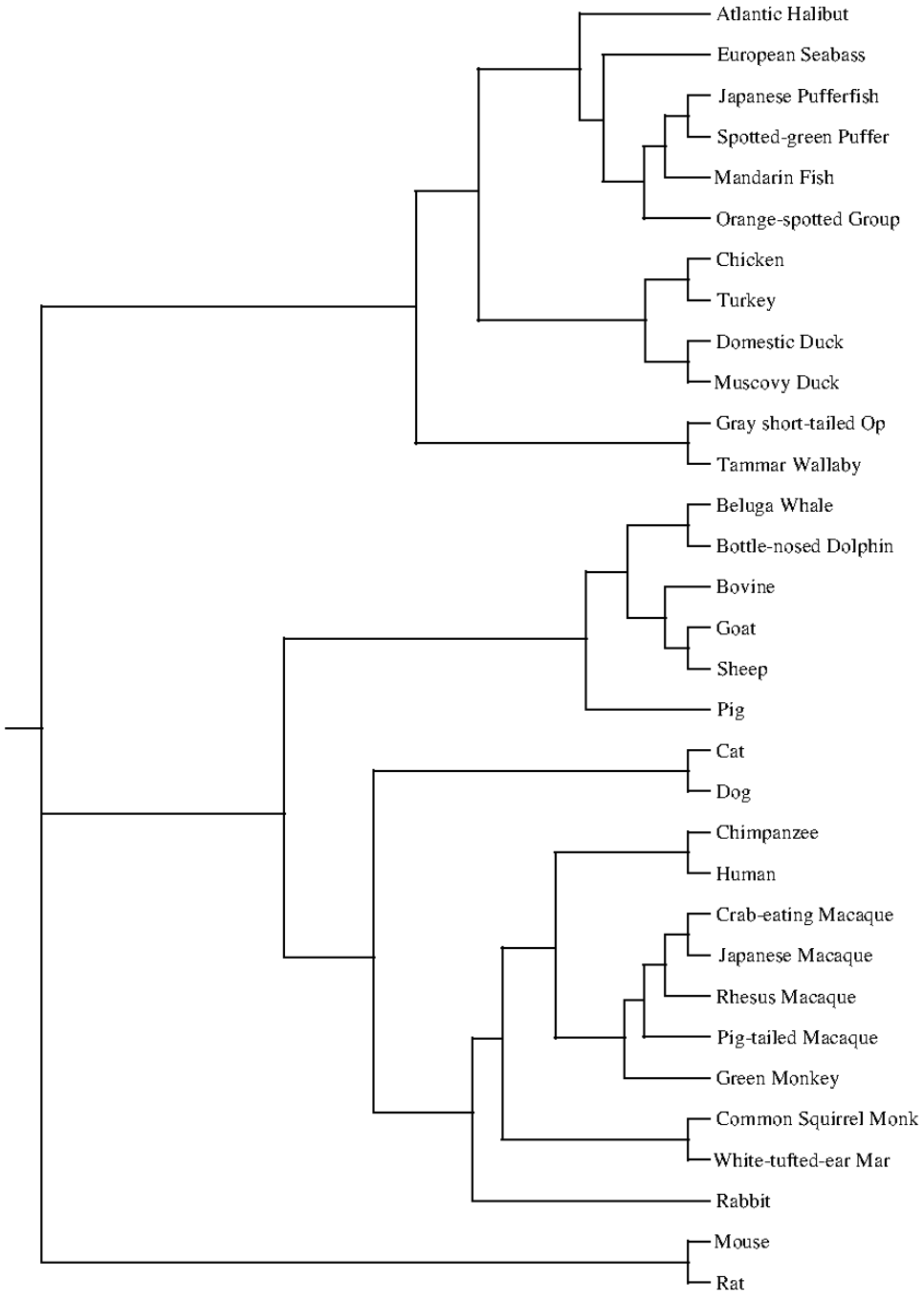



Figure 5.3: Dendrogram Constructed By Using T-Coffee Method

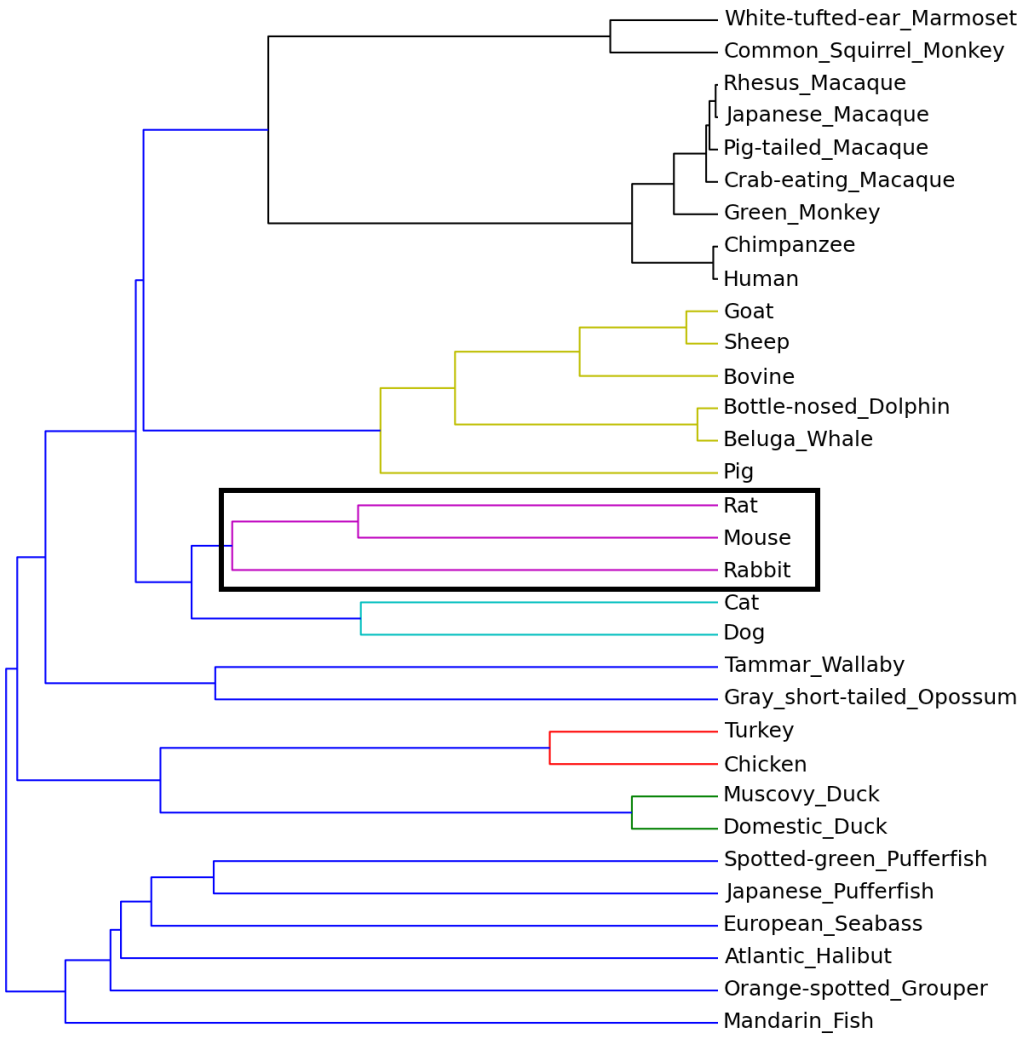



Figure 5.4: Dendrogram Constructed By Using The DFT and 25 Amino Acid Indices

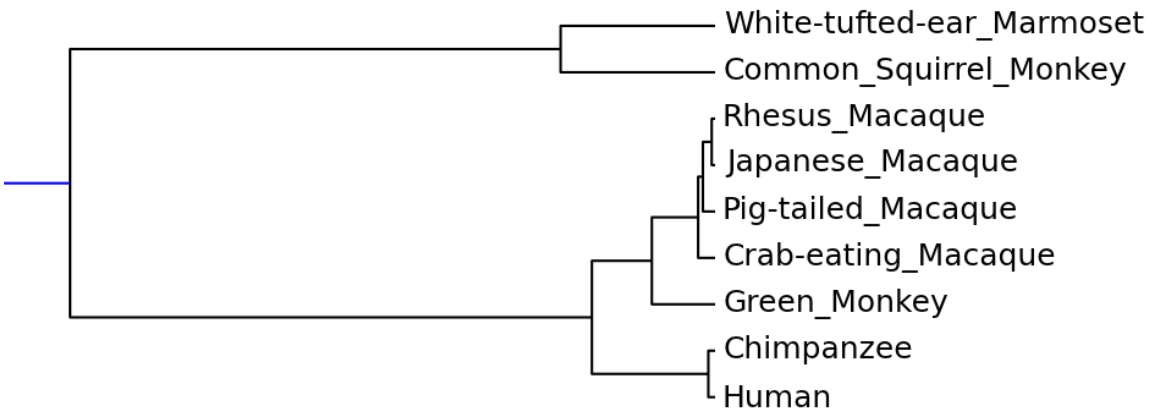



Figure 5.5: Guide-Tree Selected for Multiple Sequence Alignment 1

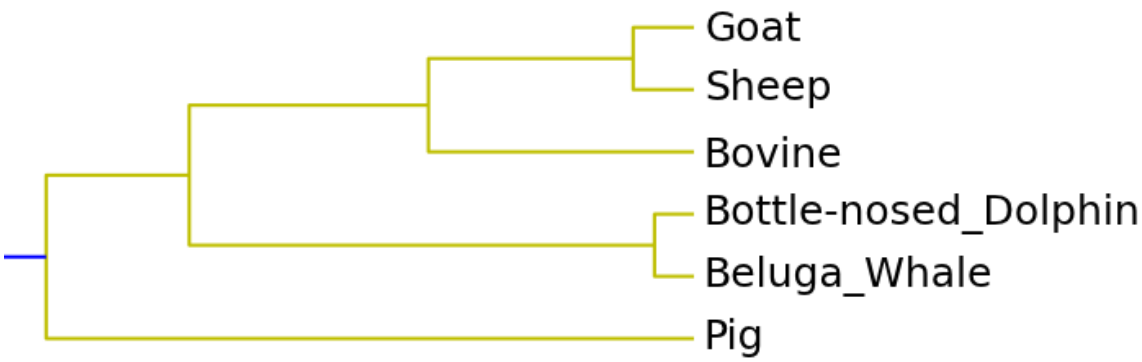



Figure 5.6: Guide-Tree Selected For Multiple Sequence Alignment 2

from primates, are given in Figure 5.4.2. For this analysis, the method described in this chapter and the methods used for comparison (MAFFT, ClustalW2 and T-Coffee), produced identical alignments.

The results for the second MSA, which includes the protein sequences from the order Artiodactyla, are given in Figures 5.4.2, 5.4.2, 5.4.2 and 5.4.2, which correspond to the results obtained from the proposed algorithm, ClustalW2, MAFFT and T-Coffee methods, respectively. For this analysis, the pairwise percent identity of all the protein sequences used was calculated. As the results show, the proposed method performs better compared to the other three MSA methods studied. Table 5.9 presents an overview of the results obtained. By using the percent identity of two aligned protein sequences, a single value can be obtained that represents the percentage of identical residues in relation to the length of the alignment.

For the results presented in Table 5.9, each pairwise percent identity of the proposed method is compared to the Clustal, MAFFT and T-Coffee results. The results are highlighted with red and yellow colours, where red indicates that the proposed method performs better than the percent identity, whereas the results that performed worse are highlighted in yellow. The unhighlighted results indicate identical results to the proposed method.

For the first comparison with Clustal, the results show that for the alignment of pig and bovine CD4 protein sequences, the proposed method performed better than the pairwise percent identity. For the remaining of the results, the two methods produced identical results. For the second and third comparisons with MAFFT and T-Coffee, the proposed method performed better in all the pairwise percent identity comparisons except the comparisons between pig - whale CD4 and pig CD4 - dolphin CD4 protein sequences.

For the average percentage identity of all the protein sequences used is calculated and presented in Table 5.9. As the results show the proposed method has the highest average percent identity of 70.830% in comparison to 70.29%, 70.443% and 70.443% for Clustal, MAFFT and T-Coffee, respectively.

Table 5.8: Similarity Matrix Generated Using the 25 Amino Acid Indices

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -2.6 | 0.3 | -0.52 | 0.22 | -0.19 | -0.58 | 0.99 | -0.36 | -0.84 | 0.11 | -1.38 | 0.58 | -0.48 | 0.38 | 0.43 | 0.97 | -2.19 | -0.48 | 0.07 | -16 |
| R | - | 4 | -0.46 | -1.92 | -2.58 | 0.05 | -1.63 | -3.45 | 0.31 | -2.86 | -2.61 | 1.14 | -1.75 | -2.23 | -1.68 | -1.3 | -0.78 | -1.83 | -1.12 | -2.75 | -16 |
| N | - | - | 4 | 0.29 | -0.96 | 2.12 | 0.92 | -0.35 | 0.51 | -1.77 | -1.15 | 0.48 | -0.53 | -1.46 | 0.81 | 1.51 | 1.53 | -1.83 | -0.51 | -1.19 | -16 |
| D | - | - | - | 4 | -1.95 | 0.49 | 1.55 | -1.14 | -1.03 | -3.02 | -2.39 | -1.08 | -1.36 | -2.15 | -0.63 | 0.07 | 0.22 | -3.12 | -1.58 | -2.55 | -16 |
| C | - | - | - | - | 4 | -0.85 | -2.03 | -1.39 | -0.76 | 0.26 | 0.29 | -2.24 | 1.08 | 0.59 | -0.76 | -0.58 | 0.1 | -0.74 | -0.34 | 0.72 | -16 |
| Q | - | - | - | - | - | 4 | 0.84 | -1.21 | 0.4 | -1.89 | -1.37 | 0.88 | -0.24 | -1.12 | 0.4 | 1.02 | 1.54 | -1.71 | -0.47 | -1.46 | -16 |
| E | - | - | - | - | - | - | 4 | -1.39 | -0.41 | -2.44 | -1.76 | -0.32 | -1.08 | -1.93 | -0.33 | -0.42 | -0.01 | -2.63 | -1.09 | -2.05 | -16 |
| G | - | - | - | - | - | - | - | 4 | -1.93 | -2.73 | -2.11 | -2.5 | -1.79 | -2.73 | -0.23 | 0.23 | -0.54 | -4 | -2.32 | -2.01 | -16 |
| H | - | - | - | - | - | - | - | - | 4 | -1.04 | -0.38 | 1.08 | 0.52 | -0.32 | -0.28 | -0.66 | 0.47 | -0.76 | 0.38 | -0.67 | -16 |
| I | - | - | - | - | - | - | - | - | - | 4 | 2.01 | -2.43 | 0.78 | 1.12 | -1.12 | -2.01 | -0.63 | 0.28 | 0.76 | 2.64 | -16 |
| L | - | - | - | - | - | - | - | - | - | - | 4 | -1.87 | 1.74 | 1.64 | -0.85 | -1.56 | -0.21 | 0.32 | 1.3 | 2.32 | -16 |
| K | - | - | - | - | - | - | - | - | - | - | - | 4 | -1.12 | -1.85 | -0.47 | -0.69 | -0.2 | -2.01 | -0.96 | -2.21 | -16 |
| M | - | - | - | - | - | - | - | - | - | - | - | - | 4 | 2.34 | -0.39 | -0.85 | 0.66 | 0.37 | 1.49 | 1.21 | -16 |
| F | - | - | - | - | - | - | - | - | - | - | - | - | - | 4 | -0.91 | -1.74 | -0.14 | 1.14 | 1.71 | 1.1 | -16 |
| P | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 4 | 0.58 | 1.02 | -1.73 | -0.49 | -0.75 | -16 |
| S | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 4 | 1.69 | -2.25 | -1.24 | -1.35 | -16 |
| T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 4 | -1.06 | 0.31 | 0.05 | -16 |
| W | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 4 | 1.2 | -0.03 | -16 |
| Y | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 4 | 0.73 | -16 |
| V | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 4 | -16 |
| X | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | -16 |

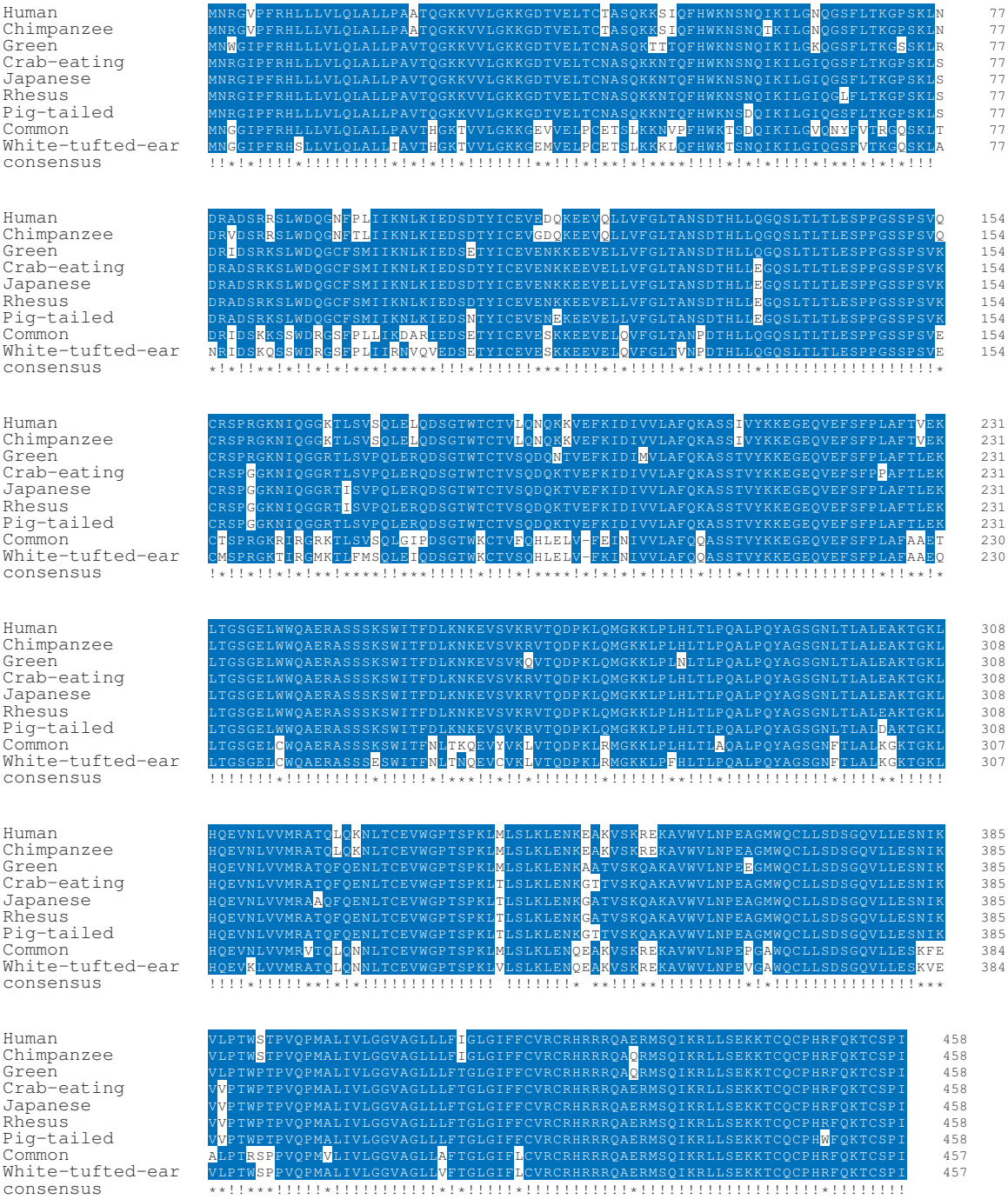



Figure 5.7: Results for the Proposed MSA Method, Clustalw2, MAFFT and T-COFFEE for Alignment 1

## 5. MULTIPLE PROTEIN SEQUENCE ALIGNMENT BASED ON MULTIPLE AMINO ACID INDICES AND DISCRETE FOURIER TRANSFORM

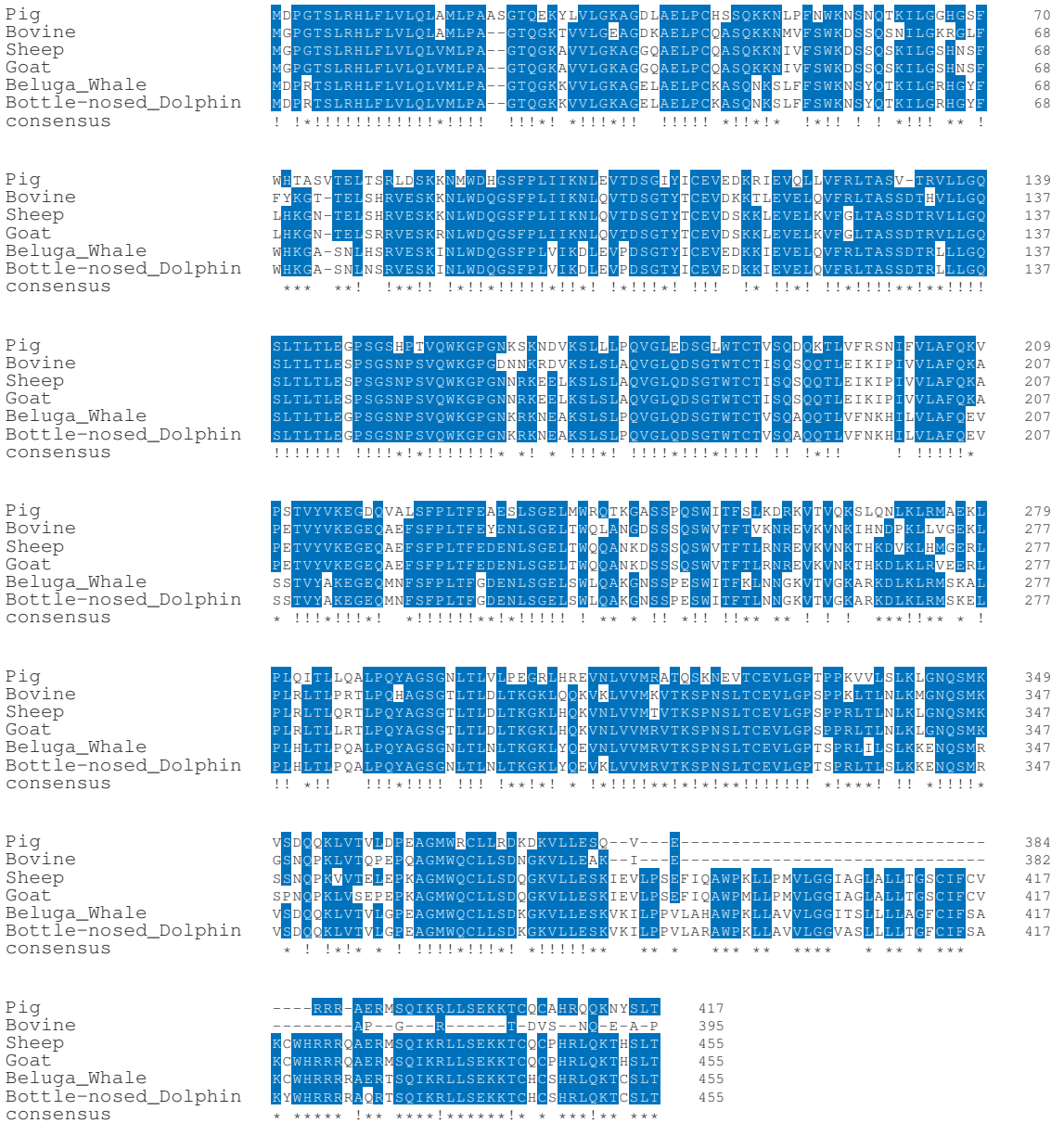



Figure 5.8: Results for the Proposed MSA Method for Alignment 2

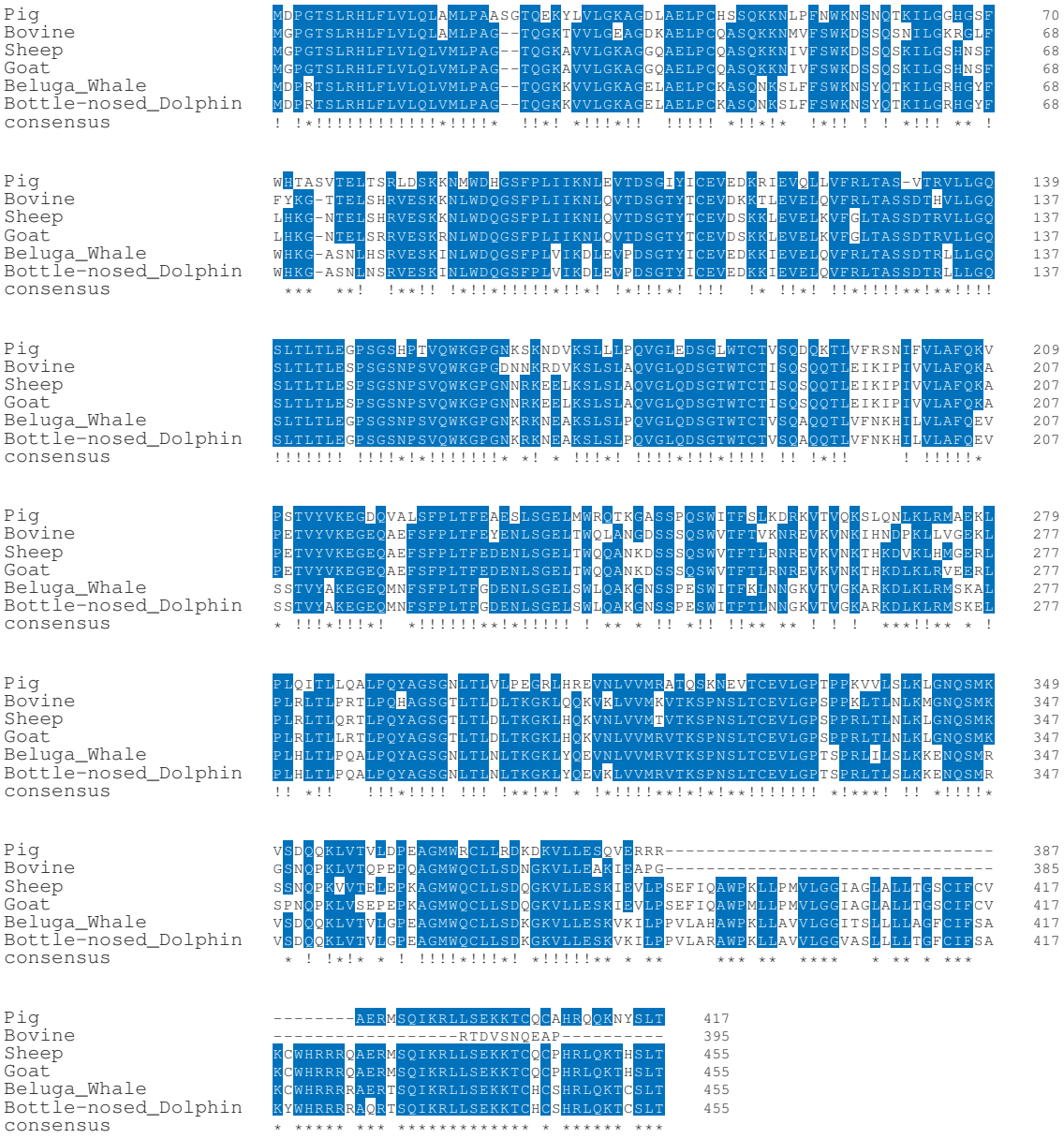



Figure 5.9: Results for the ClustalW2 MSA Method for Alignment 2

## 5. MULTIPLE PROTEIN SEQUENCE ALIGNMENT BASED ON MULTIPLE AMINO ACID INDICES AND DISCRETE FOURIER TRANSFORM

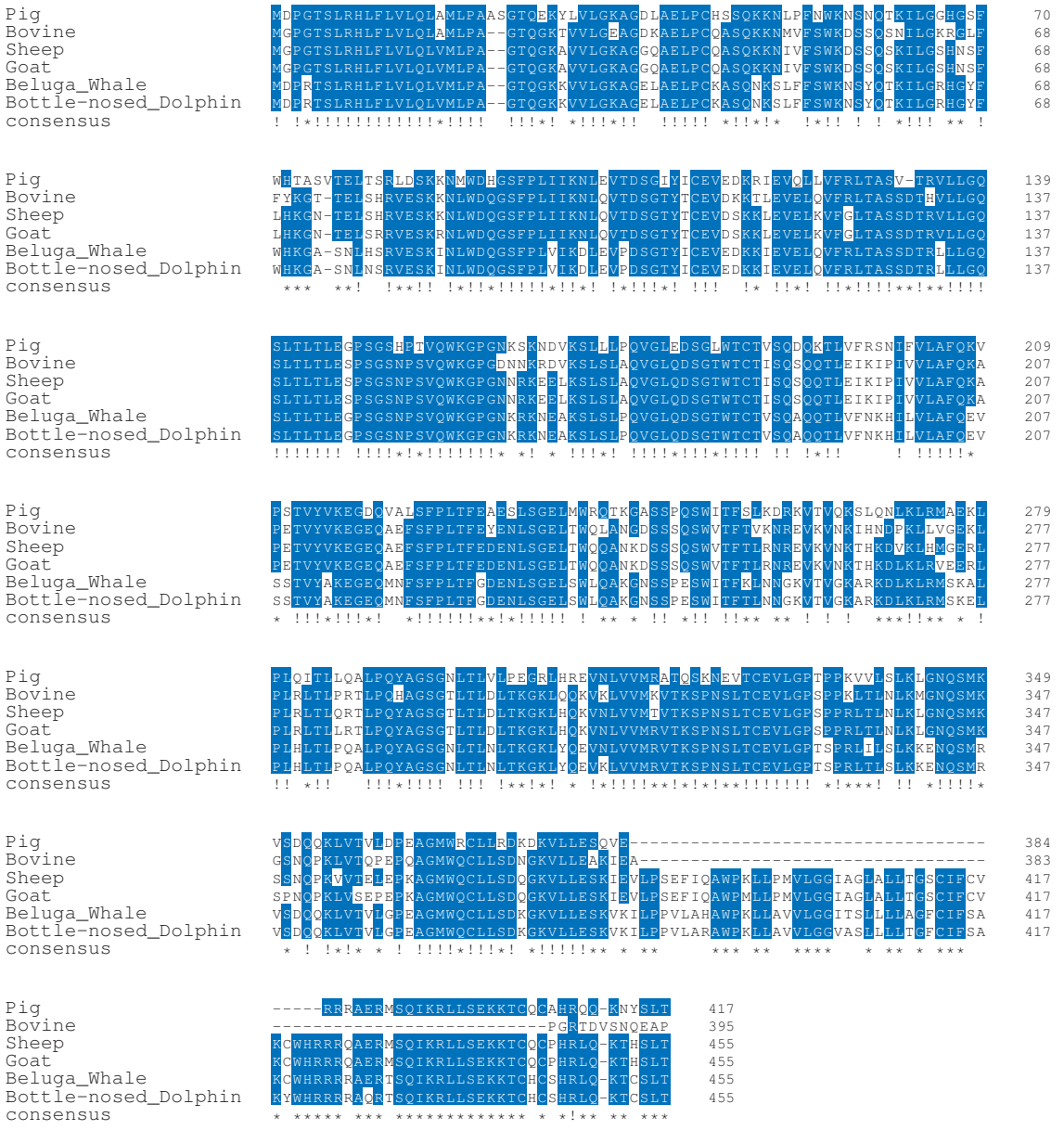



Figure 5.10: Results for the MAFFT MSA Method for Alignment 2

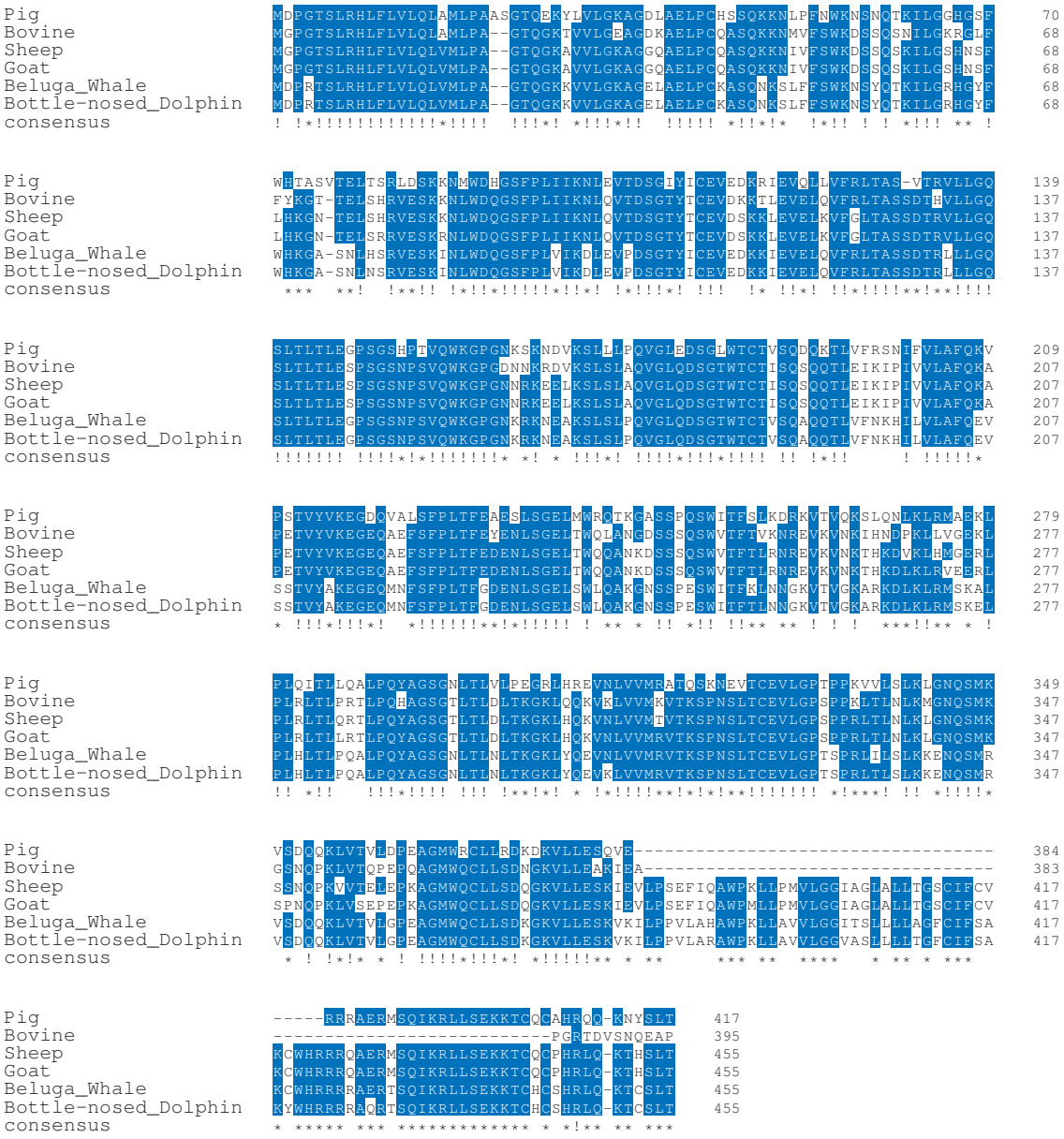



Figure 5.11: Results for the T-COFFEE MSA Method for Alignment 2

Table 5.9: Pairwise Percent Identity of CD4 Aligned Protein Sequences For Analysis 2

Proposed Algorithm

| | Bovine | Sheep | Goat | Whale | Dolphin | API* |
|---|---|---|---|---|---|---|
| Pig | 55.677% | 62.445% | 62.445% | 66.376% | 65.939% | |
| Bovine | | 73.362% | 72.926% | 59.389% | 60.044% | |
| Sheep | | | 96.507% | 72.707% | 72.707% | 70.830% |
| Goat | | | | 72.489% | 72.489% | |
| Beluga Whale | | | | | 96.943% | |

Clustal

| | Bovine | Sheep | Goat | Whale | Dolphin | API* |
|---|---|---|---|---|---|---|
| Pig | 54.585% | 61.354% | 61.354% | 65.721% | 65.284% | |
| Bovine | | 72.926% | 72.489% | 58.079% | 58.734% | |
| Sheep | | | 96.507% | 72.707% | 72.707% | 70.29% |
| Goat | | | | 72.489% | 72.489% | |
| Beluga Whale | | | | | 96.943% | |

MAFFT

| | Bovine | Sheep | Goat | Whale | Dolphin | API* |
|---|---|---|---|---|---|---|
| Pig | 55.120% | 61.874% | 61.874% | 66.449% | 66.013% | |
| Bovine | | 72.985% | 72.549% | 58.170% | 58.824% | |
| Sheep | | | 96.296% | 72.549% | 72.549% | 70.443% |
| Goat | | | | 72.331% | 72.331% | |
| Beluga Whale | | | | | 96.732% | |

T-Coffee

| | Bovine | Sheep | Goat | Whale | Dolphin | API* |
|---|---|---|---|---|---|---|
| Pig | 55.120% | 61.874% | 61.874% | 66.449% | 66.013% | |
| Bovine | | 72.985% | 72.549% | 58.170% | 58.824% | |
| Sheep | | | 96.296% | 72.549% | 72.549% | 70.443% |
| Goat | | | | 72.331% | 72.331% | |
| Beluga Whale | | | | | 96.732% | |

Red indicates lower values in comparison to the proposed method.
Yellow indicates higher values in comparison to the proposed method.
*API - Average Percent Identity

## 5.5 Conclusions

In this chapter, a new method was proposed for performing protein multiple sequence alignment. For this method Discrete Fourier Transform, was used to construct the distance matrix in combination with the multiple amino acid indices that were used to encode

protein sequences into numerical sequences. The distance matrix is important as it can be used to construct a dendrogram that will act as a guide for MSA in which the global alignment is estimated by a series of pairwise alignments. The amino acid indices selected for this analysis are based on general and widely accepted features of the amino acids. Additionally, these indices are used to construct a substitution matrix. In the literature, this similarity or substitution matrix states the rate at which one amino acid is replaced by another over time. In this chapter, a new type of substitution matrix is proposed where the physicochemical similarities between any pair of given amino acids is calculated. These similarities were calculated based on the 25 amino acids indices selected, where each one represents a unique biological protein feature.

In order to show the applicability and robustness of the proposed method, a case study was presented by using 32 CD4 protein sequences extracted from the UniProt online database. For this study, two sets of results were presented, one for constructing a rooted dendrogram that is used as a guide tree for multiple sequence alignments. The results suggests that the proposed method for constructing guide trees can identify hidden similarities between the protein sequences which are not directly observable using only the homology information. By using this guide tree, two subgroups were selected, one representing all the primates presented in the protein dataset and one that include all the protein sequences extracted from the animals that belong to the Artiodactyla order. The second set was generated by applying MSA to the selected subgroups and comparing the results with other MSA methods in the literature like ClustalW2, MAFFT and T-Coffee MSA methods.

The results show that the proposed method yields a more reliable MSA for the protein sequences can be performed. For the Alignment 1, the proposed method as well as the CLUSTALW, MAFFT, and T-COFFEE produced identical MSAs. For Alignment 2, the MSAs produced from all the methods were highly similar, with the proposed method having the highest pairwise percent identity of CD4 aligned protein sequences in comparison to ClustalW2, MAFFT and T-Coffee MSA methods. Additionally, as the results show the proposed method has the highest average percent identity.

In conclusion, the proposed MSA method is not biased to specific groups of protein sequences [236] as the values for the substitution matrix are calculated from the amino acid indices, and not from the protein sequences. Additionally, for the proposed MSA, the same substitution matrix can be considered regarding the protein sequence's homology to be aligned or the mutation rate presented. A correlation to the physical characterisations of the

amino acids the substitution matrix derived from can be achieved, while different similarity matrices can be generated by considering different amino acids physical characterisations, that each amino acid indices represents.

# Chapter 6

# Complex Informational Spectrum for the Analysis of Protein Sequences

## 6.1 Introduction

If it is considered that a protein's biological function is controlled by the selective ability of the protein to interact with selected elements in the environment the following argument arises: how is this selective ability achieved? Several attempts have been made to decode the rules that help drive biological functions of the proteins directly from the primary structure of a protein sequence. One common method used for analysing protein sequences to determine biological functions is based on the search for similarities in the arrangements between the groups of sequences. One example is the Basic Local Alignment Search Tool (BLAST) [241]. Another method for analysing macro-module sequences is to extract structural and physicochemical features such as amino acid and dipeptide composition derived from primary structure of a protein sequence [62; 242]. These features can be used for prediction of protein structural classes [31; 93], functional classes [149; 151; 152; 243] and protein-protein interactions [73; 148; 244].

In recent years, signal processing techniques have been used in bioinformatics to extract information that is expected to reveal a protein's biological function [40; 41; 42]. One of the methods that uses Discrete Fourier Transform (DFT) is Informational Spectrum Analysis (ISA) [40; 91; 92]. In previous applications where ISA was used for each group of proteins analysed [40; 91; 92] there was a group of proteins that corresponds to specific

peaks in the frequency spectrum. Therefore, the method suggests [87; 92; 146; 245] that every biological function corresponds to one unique or a set of unique peaks. The importance of this general conclusion is that specific biological functions can be extracted from protein sequences using the signal processing techniques by identifying significant features of the frequencies, which are not found in unrelated frequencies.

Complementary information such as real and imaginary frequency spectra can be derived from DFT, which has successfully been used in various areas, including biomedicine [246]. However, complex SP concept was not previously explored in the analysis of protein sequences. A new method, the complex informational spectrum [245], was developed to explore results obtained from all three frequency spectra for the analysis of protein sequences.

In the traditional approach, due to the complex nature of proteins and their functional groups, the use of only the absolute spectrum in the analysis of protein sequences can be insufficient. Biologically related features of protein sequences can be more distinct either in the real or the imaginary spectrum. Various applications, such as, development of new drugs [247], identification of important protein sequence's domains [42; 248] and investigation of protein sequences interaction [249], where ISA and RRM are already applied in the literature, CISA will also be applicable and will be able to contribute additional information.

To be able to proceed with current signal processing techniques a set of numerical values must be assigned to nucleotides or amino acids [28] in order to transform protein sequences into signals. These values should generally represent natural biological characteristics of the macro-modules with which they are paired and be relevant to the biological activity of each module. These properties are taken as any of the amino acid indices deposited in the AAIndex database such as electron-ion interaction potential (EIIP) [87; 88], hydrophobicity [88; 89], solubility [88; 89] or molecular weight [88; 89]. Additional amino acid indices were included in the analysis that were not included in AAIndex. Such indices include long-range contacts [111] and relative connectivity [29].

In this chapter, Complex Informational Spectrum Analysis (CISA) was introduced. Application of the CISA in the influenza virus is also presented in order to show usefulness and robustness of the method developed. Using an expanded set of amino acid indices further supports this. Additionally, the Complex Informational Spectrum for Analysis of Protein Sequences (CISAPS) web server is presented, which can be freely accessed to

extract features of proteins from their amino acid sequences.

## 6.2 Methods And Materials

### 6.2.1 Signal Processing-Based for the Analysis of Protein Sequence

By using digital signal processing techniques the goal is to extract information that can be related to biological functions of proteins. Various signal processing methods have been used in bioinformatics for analysing protein sequences in recent years; one of the most common methods is the informational spectrum analysis (ISA) [40; 91; 92]. For the ISA method to be implemented for the analysis of protein sequences, Discrete Fourier Transform (DFT) is applied after each amino acid of the protein sequences is expressed as numerical sequences by using various amino acid indices. A special case of ISA is the Resonant Recognition Model [40; 87; 91; 92] where the EIIP amino acid index [87] is used to encode alphabetical protein sequences into numerical sequences. ISA reveals that functionally related protein sequences common peaks appear in the informational spectrum whereas they do not appear in functionally unrelated sequences. This is directly related to the biological property of the amino acid index used. In previous studies, ISA uses DFT to extract parameters using the absolute spectrum. However, DFT that generates complex output (imaginary and real frequency spectra) has shown to produce complementary information in various fields such as Doppler ultrasound in medicine [246], polar solvation dynamics in the femtosecond evolution [250], time-domain sum-frequency generation spectroscopy using mid-infrared pulse shaping [251], hydrophobic oil droplet-water interface for the orientation and charge of water [252] and noisy speech enhancement [253].

To the best of our knowledge, complex signal processing concept has not been explored for the analysis of protein sequences. Therefore, for the first time, this chapter is concerned with the development of the complex informational spectrum (CISA) for the analysis of groups of proteins using their sequence information. This study therefore aims at deriving absolute, real and imaginary spectra from DFT for a given set of proteins. They will then be used to extract characteristic frequency parameters for the group of proteins under study. This piece of information can be used to characterise and classify protein sequences. In order for researchers to apply the method in their own set of proteins without any knowledge of SP or complex SP concept, a freely accessible web server (CISAPS

web-server) is also developed and presented in this chapter.

### 6.2.2 Preprocessing of Protein Sequences

Previous studies [146] have shown that techniques such as zero-padding and windowing may affect the analysis and the features extracted from proteins sequences. Before applying the informational spectrum analysis to protein sequences, these techniques used in signal processing needs to be considered.

#### 6.2.2.1 Windowing

The first technique considered is windowing, which tries to reduce spectral leakage [254] by multiplying a pre-calculated window with the encoded numerical sequences. Spectral leakage is caused when processing finite-length signals using frequency analysis of infinite signals in which it appears that some energy has leaked out of the primary signal spectrum into neighboring frequencies. As the literature shows in other applications [255; 256] that windowing can reduce or even eliminate spectral leakage when frequency analysis and DFT are used. Various different types of windows exist in the literature like Hamming [147; 257], Hanning [257] and Rectangular [257] window. In this case, the Hamming window is used and can be defined in Equation 6.1 and presented in Fiq. 6.1.

$$w = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) \quad n = 0 \leq n \leq N-1 \tag{6.1}$$

where N is the total amino acids in a protein sequence.

#### 6.2.2.2 Zero-padding

The second technique used is zero-padding [258; 259] in which a specified number of zero elements is added to the end of each sequence in order to increase signal length before DFT is applied to the signals. This technique is essential for CISA as the given protein sequences may not be of the same length thus making CISA unrealisable. Seven different resolutions were used for the analysis of influenza A neuraminidase proteins. The first signal length used is 470, which is the maximum protein length of the influenza A protein subtypes. The remaining six signal lengths used are 512 ($2^9$), 1024 ($2^{10}$), 2048 ($2^{11}$), 4096 ($2^{12}$), 8192 ($2^{13}$) and 16384 ($2^{14}$).

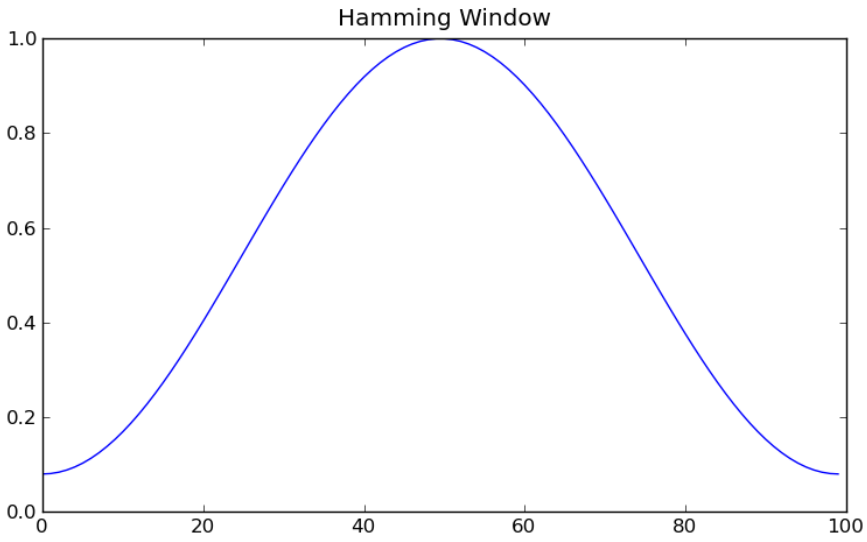



Figure 6.1: Hamming Window

### 6.2.3 Complex Informational Spectrum Analysis

The Discrete Fourier Transform (DFT) is defined as follows

$$X(n) = \sum_{m=0}^{N-1} x(m) e^{-j(2\pi/N)nm} \quad n = 0, 1, ..., N-1 \tag{6.2}$$

where $x(m)$ is the $m$th member of the numerical series, N is the total number of points in the series, and $X(n)$ are coefficients of the DFT. As the DFT coefficients consisted of two mirror parts, only the first half of the series $(N/2)$ points) will be hereafter considered. The following formula determines the maximal frequency ($F$) of all the signals (proteins) in the spectrum

$$F = \frac{1}{2d} \tag{6.3}$$

where $d$ is the distance between points of the sequence.

If it is assumed that all points of the sequence are equidistant with distance $d = 1$ then the maximum frequency in the spectrum can be found to be $F = 1/2(1) = 0.5$. This shows that the frequency range does not depend on the number of points in the sequence but only

the resolution of the spectrum.

As the definition of DFT is complex which is shown in Eq. 6.2, the output of DFT also yields a complex sequence and can be represented as follows

$$X(n) = (R(n) + jI(n)), \quad n = 0, 1, ..., (N-1)/2 \tag{6.4}$$

where $R(n)$ and $I(n)$ are the Real and Imaginary parts of the sequence, respectively.

They produce real and imaginary spectra results of which can be separately explored, and when combined, generate absolute spectrum. The absolute, real imaginary spectrum and complex informational spectrum can be formulated as follows

**Absolute Spectrum:**

$$S_a(n) = X(n)X^*(n) = |X(n)|^2, \quad n = 0, 1, ..., (N-1)/2 \tag{6.5}$$

where $S_a$ is the absolute spectrum for a specific protein, $X(n)$ are the DFT coefficients of the series $x(n)$ and $X*(n)$ are the complex conjugate.

**Real Spectrum:**

$$S_r(n) = |R(n)|^2, \quad n = 0, 1, ..., (N-1)/2 \tag{6.6}$$

where $S_r$ is the real spectrum for a specific protein, $R(n)$ are the real parts of DFT coefficients $X(n)$

**Imaginary Spectrum:**

$$S_i(n) = |I(n)|^2, \quad n = 0, 1, ..., (N-1)/2 \tag{6.7}$$

where $S_i$ is the imaginary spectrum for a specific protein, $I(n)$ are the Imaginary parts of DFT coefficients $X(n)$

**Complex Informational Spectrum:**

$$C_a = \prod_{m=1}^{M} S_{(a)}(m) \tag{6.8}$$

$$C_r = \prod_{m=1}^{M} S_{(r)}(m) \tag{6.9}$$

$$C_i = \prod_{m=1}^{M} S_{(i)}(m) \tag{6.10}$$

where $C_a$, $C_r$ and $C_i$ are the absolute, real and imaginary informational spectrum, respectively, and M is the number of protein sequences used for a specific class of proteins.

In order to scale Complex Informational Spectra in range 0 to 1, Equation 6.11 is used:

$$V_{a,r,i} = \frac{\sqrt{\sum_{n=0}^{L} C_{a,r,i}(n)}}{L} \tag{6.11}$$

where $L$ is the number of points in the Absolute ($C_a$), Real ($C_r$) and Imaginary Informational Spectrum ($C_i$).

The aim of this method is to determine a Characteristic Frequency Peak (CFP) using the informational spectrum for each spectrum (absolute, real and imaginary) that is expected to correlate with a biological function expressed by a group of protein sequences. To determine such a parameter, it is necessary to find common characteristics of the sequences with the same biological function.

CFP as a result of the CISA can be used to characterise and distinguish them from another group of proteins. However, the following conditions should be fulfilled for the CFP to be related to a biological function:

1. Only one CFP should exist for a group of protein sequences that share the same biological function.

2. For different biological functions the CFP is expected to be different.

## 6.3 Web Server Access

The CISAPS web server is available at http://cisaps.com/main/. As seen in Figure 6.2, the user can input the required information for the analysis using the input form.

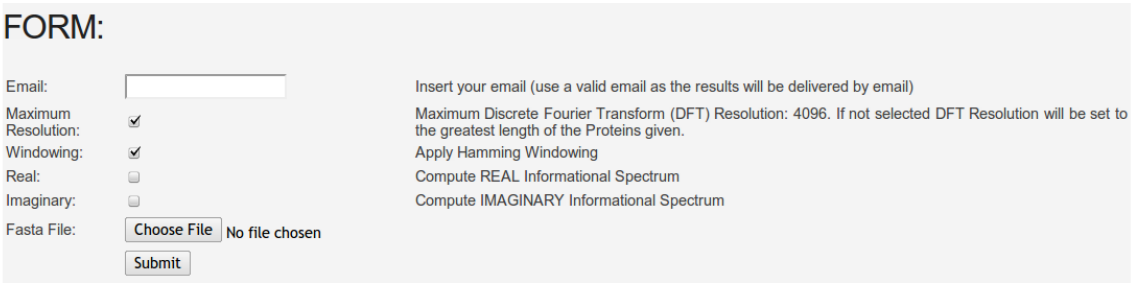



Figure 6.2: CISAPS Web Server Input Form

The mandatory information required is a valid email and protein sequences saved in FASTA format. The CISAPS web server can process up to 1000 protein sequences per analysis, where the length of any given protein should be between 8 and 4096. After a successful submission to the CISAPS web server, an email will be sent to the user with a description of the submitted data including number of proteins, unknown amino acids found in protein sequences and signal length used DFT. After the submission, protein sequences will be processed and an email will be sent to the user with the generated reports of the analysis. The email includes:

- A report of the CISA results grouped by CFP.
- A report of the CISA results listed by amino acid Index ID.
- Summary report of the occurrences of CFP.

## 6.4 Case Study: Analysing Influenza Neuraminidase Protein Sequences

During the twentieth century three major influenza A pandemics were recorded which were caused by H1N1, H2N2, and H3N2 viruses in this chronological order. In addition, H5N1 and H1N2 viruses are considered as current pandemic threats [260; 261]. Previous studies

[42] used influenza A subtypes to analyse the hemagglutinin (HA) gene with the RRM, aiming to identify new therapeutic targets for drug development by better understanding the interaction between the influenza virus and its receptors. For this analysis, the Neuraminidase (NA) gene of these five different subtypes of Influenza A virus was used, as it is the target for current antiviral drugs called neuraminidase inhibitors [262]. All data were collected from the Influenza Virus Resource database [263].

Influenza A H1N1 subtype virus is a subtype of influenza A virus and the most common cause of influenza in humans [264]. H1N1 first emerged in 1918 and was responsible for Spanish flu that killed 50 to 100 million people worldwide within a year (1918-1919) [265]. In 1947, a new H1N1 virus emerged through intrasubtype reassortment while the neuraminidase (NA) gene was preserved, which may have prevented the advancing of a new pandemic [266]. In 1957, H1N1 suddenly became extinct in humans and the reason is still not clear today. One probable explanation is that the development of high immunity to the H1N1 virus in conjunction with the development of immunity to the H2N2 Influenza virus led to the extinction of the virus. In 1977, the H1N1 virus reappeared in the former Soviet Union, Hong Kong, and north-eastern China [266]. Genetic analysis of the re-emerged H1N1 virus suggests that the strain had been conserved since 1950, and accidentally released from a laboratory facility. In April 2009, a new strain of H1N1 (S-OIV) was identified in the United States [264]. This new strain emerged from reassortment of NA and matrix genes from the Eurasian H1N1 influenza A swine virus and the remaining six gene segments from the H1N2 swine virus. For this subtype, four different groups of proteins were retrieved from the Influenza Virus Resource database, summarised in Table 6.4 and listed in Appendix D.

For the Influenza A H2N2 subtype, 76 proteins were sequenced before the period 1957 - 1968 from the Influenza Virus Resource database. H2N2 influenza viruses that could affect humans appeared in 1957; these were the result of antigenic shift from reassortment between already existing human H1N1 and avian H2N2 viruses [264]. H2N2 viruses possess the HA, NA, and PB1 gene fragments of an avian H2N2 virus whereas the remaining five gene fragments originated from the human H1N1 virus. H1N1 viruses were displaced by H2N2 viruses that were spreading quickly among humans, causing the Asian flu pandemic (1956-1958) which killed an estimated two million people worldwide [264].

For Influenza A H3N2 subtype, 200 proteins sequenced from the period 1968 - 2000 were retrieved from the Virus Resource database. H3N2 viruses emerged in 1968 by re-

| Influenza Subtype | Period | Number of Proteins |
|---|---|---|
| H1N1 | 27 | 1933 - 1946 |
| H1N1 | 12 | 1947 - 1957 |
| H1N1 | 48 | 1979 - 1989 |
| H1N1 | 200 | 2009 |
| H2N2 | 76 | 1957 - 1968 |
| H3N2 | 200 | 1968 - 2000 |
| H1N2 | 27 | 2001 - 2004 |
| H5N1 | 70 | 2005 - 2009 |

Table 6.1: Influenza Protein Sequences

assortment between circulating human H2N2 and avian H3 viruses [264]. These viruses adapted from H3 avian virus HA and PB1 genes and the six genes, including NA and, fragments of the already circulating human H2N2 viruses. H3N2 was responsible for the Hong Kong pandemic (1968-1969) which killed an estimated one million people worldwide.

For Influenza A H1N2 subtype, 27 proteins sequenced from the period 2001 - 2004 were retrieved from the Virus Resource database. The results of the genetically characterised H1N2 subtype [261] to determine the origin of all the eight gene segments showed that all H1N2 isolates were reassortants of classical swine H1N1 and triple reassortant H3N2 viruses. The neuraminidase (NA) and PB1 genes of the H1N2 isolates were of human origin, while the hemagglutinin (HA), nucleoprotein (NP), matrix (M), non-structural (NS), PA and PB2 polymerase genes were of avian or swine origin.

For Influenza A H5N1 subtype, 70 proteins sequenced from the period 2005 - 2009 in Asia were retrieved from the Virus Resource database. The H5N1 virus was created by combining various influenza A subtype virus [260]. Avian H3N8 contributes to H5N1 the PB2, PB1, NP and NS genes, while Avian H7N1 contributes to the M gene. H5N3 has the highest nucleotide similarity to H5N1 for the PA gene, which suggests that it has contributed to the PA and HA gene. Finally, avian H1N1 supplied the NA gene [260].

## 6.5 Results and Discussion

In this chapter, three sets of results were presented. The first set of results shows the effect of windowing and zero-padding on the results extracted from the complex informational

spectrum analysis. The second set of results refers to the analysis of the influenza neuraminidase protein sequences. For this case study, each H1N1, H5N1, H2N2, H3N2 and H1N2 NA protein file was submitted independently to the CISAPS server, and their results were retrieved by using the reports generated for absolute, real and imaginary informational spectrum CFPs. For the third analysis, computationally generated amino acid indices as presented in Chapter 3, were compared with original amino acid indices in relation to the CISA. The results for the third analysis were obtained using the acid and basic bovine growth factor protein sequences as presented in Chapter 2.

### 6.5.1 Effects of Windowing and Zero-Padding

For each of the Influenza A subtypes, two CFP are extracted for each signal length used; one where the windowing is applied ($w$) and one where windowing is suppressed ($\psi$). The results for the Influenza A neuraminidase subtypes are presented in Tables 6.3, 6.4 and 6.5 for CISA, AIS, RIS and IIS respectively and show that windowing and zero-padding have key impacts on CFP extracted.

Significant changes in the AIS, RIS and IIS have been observed for all the Influenza A NA subtypes when signal length is increased from 470 to 16384 through zero-padding. For example, in the case of subtype H1N1, in AIS was found to be 0.1681 when the signal length was 470, but it shifted to 0.0730 when the signal length was increased to 16384. Similar results for other subtypes are also observed in subtype H5N1 (0.1681 to 0.4839), H1N2 (0.3170 to 0.3970) and H3N2 (0.3170 to 0.3970). For RIS, in subtype H1N1 the CFP shifted from 0.4106 to 0.1687, H5N1 from 0.4830 to 0.3181, H1N2 from 0.3170 to 0.3965 and H2N2 from 0.3170 to 0.4091. For IIS the CFP shifted in subtype H5N1 from 0.3170 to 0.4844 and H3N2 from 0.3447 to 0.3972.

Additionally, significant changes in the AIS, RIS, and IIS can be observed for all Influenza A NA subtypes by applying the hamming window to the encoded protein sequences before CISA. For CFP in AIS where windowing is applied, significant changes are observed in subtype H1N1 (0.1687 to 0.0732), H5N1 (0.4839 to 0.0739), H1N2 (0.3970 to 0.4584), H2N2 (0.3972 to 0.4859), and H3N2 (0.3970 to 0.4586). For RIS the CFP shifted in subtype H5N1 from 0.3181 to 0.0731, H1N2 from 0.3965 to 0.4587, H2N2 from 0.4091 to 0.4863 and H3N2 from 0.3171 to 0.4586. For IIS the CFP shifted in subtype H5N1 (0.4844 to 0.0742), H1N2 (0.3973 to 0.4577), H2N2 (0.3973 to 0.4852) and H3N2

(0.3972 to 0.1889).

By considering the average percentage identity shown in Table 6.2 and the information retrieved from the literature regarding the Influenza A NA proteins the best match with CFP extracted from influenza A subtypes using CISA as shown in Tables 6.3-6.5 is obtained when the signal length of 4096 and windowing are both applied.

Table 6.2: Average Percent Identity

| | H1N1 | H1N2 | H2N2 | H3N2 | H5N1 |
|---|---|---|---|---|---|
| H1N1 | 93% | - | - | - | - |
| H1N2 | 40% | 98% | - | - | - |
| H2N2 | 42% | 86% | 96% | - | - |
| H3N2 | 40% | 88% | 86% | 94% | - |
| H5N1 | 83% | 41% | 43% | 41% | 96% |

### 6.5.2 Case study Results

By submitting each H1N1, H5N1, H2N2, H3N2 and H1N2 NA protein file independently to the CISAPS server, and using the reports generated for absolute, real and imaginary informational spectrum CFPs, results were retrieved. All the results obtained and reports generated can be found in Supplement 3. For the analysis, methods used in signal processing to extract better results were considered. As the Influenza A protein sequences have different lengths, maximum DFT resolution as well as windowing was also applied to the signals (protein sequences) in order to reduce spectral leakage as discussed in sub-section 6.2.2. A similar CFP between influenza A subtypes would suggest a close relationship between two protein classes for the particular feature that the amino acid index represents. By using minimum and maximum thresholds two sets of amino acid indices were retrieved. The first set represents amino acid indices with identical or closely related CFPs while the second set retrieved represent amino acids with more distributed CFPs. Two sets of tables are created to illustrate these results; Tables 6.10, 6.11 and 6.12 show amino acid indices that present highly similar cases, where Tables 6.13, 6.14 and 6.15 show amino acid indices that present the most distinct cases according to CFPs. Further information regarding amino acid indices shown in Tables 6.10 to 6.15 can be retrieved from the web server by using the assigned ID number.

```
Characteristic Frequency Peak (CFP): 0.1423 appears 6 times
Index ID: 2, Index Name: ARGP820101
Index ID: 14, Index Name: BULH740101
Index ID: 132, Index Name: JOND750101
Index ID: 355, Index Name: SIMZ760101
Index ID: 364, Index Name: TANS770102
Index ID: 395, Index Name: ZIMJ680101
```

Figure 6.3: H1N1 2009 Absolute Report sample obtained from the CISAPS web server

After extracting the results from the CISAPS web server, the next step in the analysis is to discover if any of the biological features represented in amino acid indices from Tables 6.10 - 6.15 can be related to previous work in the literature. The following associations were achieved:

- The results indicate that Hydrophobicity plays an important role for the neuraminidase gene, as it appears multiple times with different amino acid indices. Identification numbers of these amino acid indices that represent hydrophobicity are 56, 57, 58, 242 and 513. The literature supports [267; 268; 269] that the hydrophobic region of the influenza neuraminidase gene plays an important role informing the functionality of the gene [268; 269] and that it is a potential target for new antiviral drugs [268; 269].

- According to the literature, Protein Kinase C (PK-C) that is represented in amino acid 76, appears to play an important role in distinguishing various H5N1 subtypes [270].

- Another protein feature that is utilised from H1N1 subtype mutants [271] is linker propensity, which is represented in amino acid indices 434 and 496.

- Finally, as previous work shows, neuraminidase active sites present high polarity [272] that is represented in amino acid index 111.

As the importance of the amino acid indices that represent hydrophobicity, PK-C and linker propensity to the neuraminidase gene is established it can be concluded that the rest of the amino acid indices which appear in Tables 6.10 - 6.15 have a higher degree of

association than the rest of the amino acid indices in the database. Further investigation is required regarding the biological relationship of these indices to the influenza A NA gene. One of the promising results is indices 557 that represent short- and medium-range non-bonded energy [111], which only appears in the imaginary spectrum.

In the literature, when informational spectrum analysis is used [40; 91; 92] only the absolute spectrum is considered. As the results show, only the use of the absolute spectrum to determine how two or more protein classes are related according to CFP is not sufficient. Several amino acid indices do not appear in the absolute spectrum and have significant biological importance to the influenza A NA gene. One example are indices 111 and 513 that represent polarity and hydrophobicity respectively. Additionally, amino acid indices for the real informational spectrum are 154, 242 and 427 and for the imaginary informational spectrum are 403, 421, 463 and 557, which do not appear in the absolute spectrum and may also be biologically significant.

Table 6.3: Absolute Spectra Results

| N | H1N1 | | H5N1 | | H1N2 | | H2N2 | | H3N2 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\psi$ | w | $\psi$ | w | $\psi$ | w | $\psi$ | w | $\psi$ | w |
| 470 | 0.1681 | 0.0745 | 0.1681 | 0.0745 | 0.3170 | 0.4574 | 0.4085 | 0.1894 | 0.3170 | 0.1894 |
| 512 | 0.0742 | 0.0742 | 0.4844 | 0.0742 | 0.3965 | 0.4590 | 0.4082 | 0.1504 | 0.3965 | 0.4590 |
| 1024 | 0.1689 | 0.0732 | 0.4844 | 0.0742 | 0.3975 | 0.458 | 0.3975 | 0.1504 | 0.3975 | 0.4590 |
| 2048 | 0.0737 | 0.0737 | 0.4839 | 0.0737 | 0.3970 | 0.4585 | 0.3970 | 0.4858 | 0.3970 | 0.4585 |
| 4096 | 0.0737 | 0.0735 | 0.4839 | 0.0740 | 0.3970 | 0.4585 | 0.3972 | 0.4858 | 0.3970 | 0.4585 |
| 8192 | 0.0737 | 0.0735 | 0.4839 | 0.0739 | 0.3970 | 0.4584 | 0.3972 | 0.4860 | 0.3971 | 0.4585 |
| 16384 | 0.0737 | 0.0735 | 0.4839 | 0.0739 | 0.3970 | 0.4584 | 0.3972 | 0.4859 | 0.3970 | 0.4586 |

Table 6.4: Real Spectra Results

| N | H1N1 | | H5N1 | | H1N2 | | H2N2 | | H3N2 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\psi$ | w | $\psi$ | w | $\psi$ | w | $\psi$ | w | $\psi$ | w |
| 470 | 0.4106 | 0.4915 | 0.4830 | 0.4872 | 0.3170 | 0.5000 | 0.3170 | 0.5000 | 0.3170 | 0.3170 |
| 512 | 0.2188 | 0.2188 | 0.3184 | 0.3184 | 0.3965 | 0.4590 | 0.3926 | 0.4863 | 0.3965 | 0.4590 |
| 1024 | 0.0732 | 0.0732 | 0.4834 | 0.0732 | 0.3965 | 0.4590 | 0.4092 | 0.4863 | 0.3174 | 0.4590 |
| 2048 | 0.0732 | 0.0732 | 0.4834 | 0.0732 | 0.3965 | 0.4585 | 0.4092 | 0.4863 | 0.3169 | 0.4585 |
| 4096 | 0.1687 | 0.0732 | 0.3181 | 0.0732 | 0.3965 | 0.4587 | 0.4092 | 0.4863 | 0.3171 | 0.4585 |
| 8192 | 0.1687 | 0.0731 | 0.3181 | 0.0731 | 0.3965 | 0.4587 | 0.4091 | 0.4863 | 0.3171 | 0.4586 |
| 16384 | 0.1687 | 0.0732 | 0.3181 | 0.0731 | 0.3965 | 0.4587 | 0.4091 | 0.4863 | 0.3171 | 0.4586 |

Table 6.5: Imaginary Spectra Results

| N | H1N1 | | H5N1 | | H1N2 | | H2N2 | | H3N2 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\psi$ | w | $\psi$ | w | $\psi$ | w | $\psi$ | w | $\psi$ | w |
| 470 | 0.0745 | 0.0745 | 0.3170 | 0.0745 | 0.4085 | 0.4574 | 0.4085 | 0.4851 | 0.3447 | 0.4596 |
| 512 | 0.0742 | 0.0742 | 0.4844 | 0.0742 | 0.4082 | 0.2227 | 0.4082 | 0.1777 | 0.4082 | 0.2109 |
| 1024 | 0.0742 | 0.0742 | 0.4844 | 0.0742 | 0.3975 | 0.4580 | 0.3975 | 0.4854 | 0.3975 | 0.3975 |
| 2048 | 0.0742 | 0.0742 | 0.4844 | 0.0742 | 0.3975 | 0.4575 | 0.3975 | 0.4854 | 0.3975 | 0.1890 |
| 4096 | 0.0742 | 0.0742 | 0.4844 | 0.0742 | 0.3972 | 0.4578 | 0.3972 | 0.4854 | 0.3972 | 0.1890 |
| 8192 | 0.0741 | 0.0741 | 0.4844 | 0.0742 | 0.3973 | 0.4578 | 0.3972 | 0.4852 | 0.3972 | 0.1890 |
| 16384 | 0.0741 | 0.0741 | 0.4844 | 0.0742 | 0.3973 | 0.4577 | 0.3973 | 0.4852 | 0.3972 | 0.1889 |

### 6.5.3 Comparison of Generated and Original Amino Acid Indices in respect to Complex informational Spectrum Analysis

To compare the computationally generated Amino Acid indices (AAI) with the original AAI, CISA [146; 245] was applied to extract biologically related features. To demonstrate the application of CISA two protein sequences and four computationally generated AAI were used in the analysis. The protein sequences selected are (1) acid bovine fibroblast growth factor (AAA66188) [94] and (2) bovine fibroblast growth factor (AAA30517) [94].

The four computationally generated AAI that yielded a variance of 1.0 and appeared in all five clustering methods are selected, as shown in Table 6.6. All the AAI within the selected clusters represents the same or similar biological feature of the protein sequences. The original AAIs that were used to generate the four new computationally generated AAI are summarised as follows:

- Cluster 1
  - BROC820101: Retention coefficient in TFA [273]
  - BROC820102: Retention coefficient in HFBA [273]
- Cluster 2
  - PRAM820102: Slope in regression analysis x 1.0E1 [274]
  - PRAM820103: Correlation coefficient in regression analysis [274]
- Cluster 3
  - MIYS990101: Relative partition energies derived by the Bethe approximation [144]

Table 6.6: Generated Amino Acid Indices Used in Complex Informational Spectrum Analysis

| Original Index ID | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | | 249 | | 522 | | 524 | | | | | |
| 13 | | 250 | | 523 | | 525 | | | | | |
| Original Index Name | | | | | | | | | | | |
| BROC820101 | | PRAM820102 | | MIYS990101 | | MIYS990103 | | | | | |
| BROC820102 | | PRAM820103 | | MIYS990102 | | MIYS990104 | | | | | |
| Method (Threshold) | Generated ID | Method (Threshold) | Generated ID | Method (Threshold) | Generated ID | Method (Threshold) | Generated ID | | | | |
| Single Linkage (1) | 5 | Single Linkage (1) | a* | Single Linkage (1) | a* | Single Linkage (1) | a* | | | | |
| Single Linkage (0.65) | 8 | Single Linkage (0.65) | a* | Single Linkage (0.65) | 128 | Single Linkage (0.65) | 129 | | | | |
| Complete Linkage (1) | a* | Complete Linkage (1) | 44 | Complete Linkage (1) | a* | Complete Linkage (1) | a* | | | | |
| Complete Linkage (0.65) | 8 | Complete Linkage (0.65) | 126 | Complete Linkage (0.65) | 208 | Complete Linkage (0.65) | 209 | | | | |
| Average Linkage (1 and 0.4) | 1 | Average Linkage (1 and 0.4) | 38 | Average Linkage (1 and 0.4) | a* | Average Linkage (1 and 0.4) | a* | | | | |
| New Amino Generated Acid Index Values | | | | | | | | | | | |
| A | 0.23307 | 0.20028 | | 0.25918 | | -0.0265 | | | | | |
| R | -0.4846 | -0.3670 | | -0.1966 | | -0.1833 | | | | | |
| N | -0.1766 | 0.00588 | | -0.2860 | | -0.0198 | | | | | |
| D | 0.02147 | 0.03952 | | 0.19862 | | 0.14371 | | | | | |
| C | 0.45094 | 0.54348 | | -0.3526 | | 0.05804 | | | | | |
| Q | -0.1447 | -0.0816 | | -0.1573 | | 0.07268 | | | | | |
| E | 0.08204 | -0.1199 | | -0.0982 | | -0.0735 | | | | | |
| G | 0.10366 | 0.49572 | | -0.2163 | | 0.00164 | | | | | |
| H | -0.2872 | 0.01452 | | -0.0376 | | 0.07138 | | | | | |
| I | -0.3567 | -0.1286 | | -0.2027 | | 0.13444 | | | | | |
| L | 0.28669 | -0.0211 | | 0.27279 | | 0.34624 | | | | | |
| K | -0.0579 | -0.0446 | | 0.44546 | | -0.0292 | | | | | |
| M | 0.09762 | 0.15824 | | 0.12288 | | 0.21084 | | | | | |
| F | 0.25294 | -0.1306 | | 0.14407 | | -0.1377 | | | | | |
| P | -0.0519 | 0.09938 | | 0.26827 | | -0.1499 | | | | | |
| S | -0.0101 | -0.0024 | | -0.0089 | | 0.18126 | | | | | |
| T | -0.0136 | -0.0944 | | -0.1073 | | 0.22954 | | | | | |
| W | -0.1881 | -0.4104 | | -0.1542 | | -0.6538 | | | | | |
| Y | 0.14166 | -0.0967 | | 0.28038 | | -0.3816 | | | | | |
| V | 0.10127 | -0.0596 | | -0.1739 | | 0.20548 | | | | | |

a* The specified amino acid indices' cluster did not appear in the analysis.

- MIYS990102: Optimized relative partition energies - method A [144]

- Cluster 4

  - MIYS990103: Optimized relative partition energies - method B [144]
  - MIYS990104: Optimized relative partition energies - method C [144]

In order for the CISA method to be applied, the protein sequences used must have the same length. Zero-padding is used to increase the length of the proteins to 512. Additionally, the Hamming windowing was applied to the protein sequences used. Characteristic Frequency Peak (CFP) is obtained for each AAI, including the original and computationally generated AAI. Tables 6.7, 6.8 and 6.9 show the extracted CFP values from the protein sequences. Results show, that for each of the original AAIs, which belongs to the same cluster, the extracted CFP value is similar or identical to the other AAIs that belong to the same cluster. However, results in tables 6.7, 6.8 and 6.9 shows AAI that belong to another cluster and represent a different biological feature can result to the same CFP. For example, cluster 1, 3 and 4 resulted to an approximate identical CFP in AIS, RIS and IIS, only cluster 2 can be distinguished. By using computationally generated AAI for CIS analysis, a clear separation of the CFP extracted values for the clusters can be achieved. In addition, clusters that represent each class with variance 1.0 the CFP extracted in all three possible spectra are distinct (Table 6.6).

Table 6.7: Absolute Informational Spectrum Results

| Generated index ID | Index Name | Original Index CFP | Generated Index CFP |
|---|---|---|---|
| 12 | BROC820101 | 0.0955 | 0.1599 |
| 13 | BROC820102 | 0.095 | |
| 249 | PRAM820102 | 0.2131 | 0.2566 |
| 250 | PRAM820103 | 0.1899 | |
| 522 | MIYS990101 | 0.0957 | 0.4507 |
| 523 | MIYS990102 | 0.0957 | |
| 524 | MIYS990103 | 0.0962 | 0.3582 |
| 525 | MIYS990104 | 0.0962 | |

Table 6.8: Real Informational Spectrum Results

| Generated index ID | Index Name | Original Index CFP | Generated Index CFP |
|---|---|---|---|
| 12 | BROC820101 | 0.0952 | 0.1628 |
| 13 | BROC820102 | 0.0952 | |
| 249 | PRAM820102 | 0.2153 | 0.4744 |
| 250 | PRAM820103 | 0.1863 | |
| 522 | MIYS990101 | 0.0955 | 0.4475 |
| 523 | MIYS990102 | 0.0955 | |
| 524 | MIYS990103 | 0.0955 | 0.3843 |
| 525 | MIYS990104 | 0.0957 | |

Table 6.9: Imaginary Informational Spectrum Results

| Generated index ID | Index Name | Original Index CFP | Generated Index CFP |
|---|---|---|---|
| 12 | BROC820101 | 0.0986 | 0.1597 |
| 13 | BROC820102 | 0.0986 | |
| 249 | PRAM820102 | 0.0981 | 0.2214 |
| 250 | PRAM820103 | 0.1833 | |
| 522 | MIYS990101 | 0.0989 | 0.4512 |
| 523 | MIYS990102 | 0.0989 | |
| 524 | MIYS990103 | 0.0989 | 0.3601 |
| 525 | MIYS990104 | 0.0989 | |

## 6.6 Conclusions

In this chapter a web-based server is developed and presented , named CISAPS, which provides complex informational spectrum analysis for protein sequences. As the results show protein classes that present similarities or differences according to the CFP in specific amino acid indices, then it is probable that these classes are related with the protein feature that the specific amino acid represents. Furthermore, the use of only the absolute spectrum in the analysis of protein sequences using the informational spectrum analysis is proven to be insufficient, as biologically related features to the analysis of influenza A subtypes appear individually either in the real or the imaginary spectrum.

In the literature, various areas exist where ISA and RRM have been successfully applied. For these areas, CISA will also be applicable, and will be able to contribute additional information as discussed below.

**Development of New Drugs:**
Bioinformatics has become an important component in drug discovery in the recent years, by accelerating this complex, expensive and time-consuming process. ISA, in combination with the EIIP scale index can successfully be applied in the bioinformatics model for the discovery and development of new drugs. As the EIIP scale index represents the interaction potential of amino acids, the development time of a new drug can considerably decrease by applying ISA or CISA in the following ways:

1. By extracting key features such as the CFP of compounds that have shown activity against the target disease and comparing them against molecular databases.

2. By using ISA and CISA, the selected compounds can be modified to increase the desire biological activity.

3. Additionally potential target areas can be identified by selecting protein or nucleotide sequences domains.

An example of applying ISA in the area of drug discovery can be found in [247] where this technique was applied in development of HIV entry inhibitors.

**Identification of important protein sequence's domains:**
In biology, similar or identical nucleotide or protein sequences are called conserved sequences that can occur across different species or are presented in different molecules within the same organism. In influenza research area, the identification of such as a conserved domain is essential, especially any receptor binding related domain to the development of influenza inhibitors. By using ISA, the informational [42; 248] and structural [42; 248] features as well multiple conserved domain [275] of HA with receptor-virus interaction were investigated that related with receptor-virus interaction. These studies were intended to expand the collection of key regions by discovering multiple domains of H1N1 and H5N1 HA subtype 1 that can alter the receptor binding model. Using the same approach, mutations F71S, T128S, E302K and M314L, in the H1N1 HA gene are recognised

necessary for the human interaction. Additionally, positions 94D, 196D and 274D in the H1N1 HA were marked as important hot-spots for mutations. One of these mutations hot-spots, D274E, is already identified in H1N1 isolates and its contribution to the human host adaptation is identified. Furthermore, the results in these studies propose that the Influenza subtype H1N1 HA gene will persist in mutating, which could further promote the human interaction. These results were extracted using CFP at frequencies 0.055 and 0.295. Another study that uses ISA aims to predict amino acid residues in highly conserved domains of the hormone prolactin (PRL) [276]. In this study, ISA was implemented with the EIIP scale index to extract the CFP of the PRL hormone and to determine which amino acids contribute more to these frequencies, and therefore to the PRL biological function. By using ISA, the highly conserved regions by using ISA were determined in amino-terminal and C-terminus regions of PRL. As the paper [276] proposes, predictions correspond with experimentally tested residues using site-direct mutagenesis and photoaffinity labeling.

**Investigation of Protein Sequences Interaction:**
Another bioinformatics area in which ISA is applied is the analysis of protein sequence interaction. By using ISA with EIIP index scale's interactions between oncogene, IL-2, and p53 tumor suppressor proteins were analysed [249]. In order to investigate the common interactions of these protein sequences, CFP needs to be determined. As the results of this study had showed, ISA can be effectively used to extract features from protein sequences related to their common biological function. All three interactive protein sequences used share the CFP at frequency 0.0322. This identified feature is a distinguishing feature of oncogene proteins and can be used to characterise promotion of uncontrolled cell growth. Furthermore, anti-cancerous properties can be identified using CFP features and peptides can be designed to exhibit only these characteristics. As these results [249] show, ISA and CISA can provide a new method to understand information presented in a protein sequence's primary structure. Finally, these results can be used to contribute significantly in the development of new biomaterials by accelerating complex costly and time consuming procedures.

Additionally, an example is given where CIS analysis is applied to the generated amino acid indices in comparison to the original amino acid indices, as described in chapter 3. The results indicate that a clear separation of the CFP extracted for each of the clusters

can be achieved, in which distinctive protein features are represented. In contrast, in the individual amino acid indices, different protein features may result in the same or similar CFP.

This web-based server enables researchers with little knowledge of signal processing methods to apply and include complex informational spectrum analysis to their work. Furthermore, CISAPS uses a collection of 611 unique amino acid indices, each one representing a different property, to perform the analysis. Moreover, in this chapter, various technical issues such as signal length and windowing that may affect the analysis are also addressed.

Finally, a comparison is given between the original and generated amino acid indices in regard to the CISA. As the results show by using CISA and the generated amino acid indices, CFP extracted in all three possible spectra are more distinct in relation to the biological features the amino acid indices represent, in comparison to the original amino acid indices. More experimental study is required for the use of CISA with the generated amino acid indices as presented in Chapter 3, and their practical applications in classification and characterisation of protein sequences.

Table 6.10: Characteristic Frequency Peak Similarities in Absolute Informational Spectrum

| ID | H1N1 1933 | H1N1 1947 | H1N1 1979 | H1N1 2009 | H5N1 | H1N2 | H2N2 | H3N2 |
|---|---|---|---|---|---|---|---|---|
| 56 | 0.142 | 0.1418 | 0.1418 | 0.1276 | 0.1423 | 0.122 | 0.1218 | 0.122 |
| 57 | 0.1276 | 0.1276 | 0.1276 | 0.1276 | 0.1274 | 0.122 | 0.1218 | 0.122 |
| 58 | 0.142 | 0.142 | 0.142 | 0.1276 | 0.1425 | 0.122 | 0.122 | 0.122 |
| 84 | 0.082 | 0.0817 | 0.082 | 0.0815 | 0.0817 | 0.081 | 0.0813 | 0.081 |
| 113 | 0.1208 | 0.1208 | 0.1208 | 0.1205 | 0.1208 | 0.122 | 0.122 | 0.122 |
| 126 | 0.082 | 0.0817 | 0.0817 | 0.0817 | 0.0815 | 0.081 | 0.0813 | 0.081 |
| 173 | 0.082 | 0.0817 | 0.082 | 0.0817 | 0.0815 | 0.0808 | 0.0813 | 0.081 |
| 333 | 0.0817 | 0.0815 | 0.0817 | 0.0815 | 0.0813 | 0.0805 | 0.081 | 0.0808 |
| 369 | 0.082 | 0.0817 | 0.082 | 0.0817 | 0.0815 | 0.081 | 0.0813 | 0.0813 |
| 416 | 0.0815 | 0.0815 | 0.0815 | 0.0813 | 0.081 | 0.0803 | 0.0808 | 0.0805 |
| 436 | 0.1271 | 0.1271 | 0.1271 | 0.1271 | 0.1269 | 0.122 | 0.122 | 0.122 |
| 544 | 0.1274 | 0.1274 | 0.1274 | 0.1274 | 0.1271 | 0.1218 | 0.1218 | 0.1218 |
| 589 | 0.0822 | 0.082 | 0.082 | 0.0817 | 0.082 | 0.0815 | 0.0815 | 0.0815 |

Table 6.11: Characteristic Frequency Peak Similarities in Real Informational Spectrum

| ID | H1N1 1933 | H1N1 1947 | H1N1 1979 | H1N1 2009 | H5N1 | H1N2 | H2N2 | H3N2 |
|---|---|---|---|---|---|---|---|---|
| 56 | 0.1269 | 0.1271 | 0.1423 | 0.1271 | 0.1418 | 0.1218 | 0.1218 | 0.1218 |
| 57 | 0.1269 | 0.1271 | 0.1271 | 0.1271 | 0.1274 | 0.1218 | 0.1218 | 0.1218 |
| 58 | 0.1269 | 0.1425 | 0.1423 | 0.1425 | 0.1418 | 0.1218 | 0.1218 | 0.1218 |
| 84 | 0.0827 | 0.0813 | 0.0813 | 0.0813 | 0.0817 | 0.0813 | 0.0813 | 0.0813 |
| 111 | 0.1269 | 0.1271 | 0.1271 | 0.1271 | 0.1274 | 0.1218 | 0.1218 | 0.122 |
| 113 | 0.121 | 0.1213 | 0.1213 | 0.1213 | 0.121 | 0.1218 | 0.1218 | 0.1218 |
| 126 | 0.0827 | 0.0813 | 0.0813 | 0.0813 | 0.0817 | 0.0813 | 0.0815 | 0.0815 |
| 173 | 0.0827 | 0.0813 | 0.0813 | 0.0813 | 0.0817 | 0.0813 | 0.0815 | 0.0815 |
| 242 | 0.1269 | 0.1271 | 0.1271 | 0.1271 | 0.1271 | 0.1218 | 0.1218 | 0.1218 |
| 333 | 0.0827 | 0.081 | 0.0813 | 0.0813 | 0.0817 | 0.0813 | 0.0813 | 0.0813 |
| 369 | 0.0827 | 0.0813 | 0.0813 | 0.0813 | 0.0817 | 0.0813 | 0.0815 | 0.0815 |
| 416 | 0.0825 | 0.081 | 0.0813 | 0.0813 | 0.0817 | 0.0813 | 0.0813 | 0.0813 |
| 436 | 0.1269 | 0.1271 | 0.1271 | 0.1271 | 0.1274 | 0.1215 | 0.1215 | 0.1215 |
| 513 | 0.1269 | 0.1271 | 0.1271 | 0.1271 | 0.1274 | 0.1218 | 0.1218 | 0.1218 |
| 544 | 0.1269 | 0.1271 | 0.1271 | 0.1271 | 0.1274 | 0.1218 | 0.1218 | 0.1218 |
| 589 | 0.0827 | 0.0813 | 0.0813 | 0.0813 | 0.082 | 0.0815 | 0.0815 | 0.0815 |

Table 6.12: Characteristic Frequency Peak Similarities in Imaginary Informational Spectrum

| ID | H1N1 1933 | H1N1 1947 | H1N1 1979 | H1N1 2009 | H5N1 | H1N2 | H2N2 | H3N2 |
|---|---|---|---|---|---|---|---|---|
| 56 | 0.1279 | 0.1415 | 0.1415 | 0.1281 | 0.1428 | 0.1227 | 0.1225 | 0.1227 |
| 57 | 0.1281 | 0.1281 | 0.1281 | 0.1281 | 0.1428 | 0.1227 | 0.1227 | 0.1225 |
| 58 | 0.1281 | 0.1415 | 0.1415 | 0.1281 | 0.1428 | 0.1227 | 0.1225 | 0.1227 |
| 84 | 0.0817 | 0.082 | 0.0822 | 0.0822 | 0.0827 | 0.0803 | 0.0803 | 0.0803 |
| 113 | 0.1201 | 0.1203 | 0.1203 | 0.1203 | 0.1201 | 0.1227 | 0.1227 | 0.1225 |
| 126 | 0.0817 | 0.0822 | 0.0822 | 0.0822 | 0.0808 | 0.0805 | 0.0805 | 0.0805 |
| 173 | 0.0817 | 0.0822 | 0.0825 | 0.0822 | 0.0808 | 0.0803 | 0.0805 | 0.0805 |
| 333 | 0.0817 | 0.0822 | 0.0822 | 0.0822 | 0.0808 | 0.0803 | 0.0805 | 0.0803 |
| 369 | 0.0817 | 0.0822 | 0.0822 | 0.0822 | 0.0808 | 0.0803 | 0.0805 | 0.0803 |
| 403 | 0.0378 | 0.0698 | 0.0698 | 0.0698 | 0.0713 | 0.0698 | 0.0698 | 0.0698 |
| 416 | 0.0815 | 0.082 | 0.082 | 0.082 | 0.0805 | 0.0803 | 0.0803 | 0.0803 |
| 436 | 0.1281 | 0.1281 | 0.1281 | 0.1281 | 0.1262 | 0.1225 | 0.1225 | 0.1225 |
| 463 | 0.4846 | 0.4832 | 0.4832 | 0.4832 | 0.4841 | 0.4851 | 0.4849 | 0.4851 |
| 544 | 0.1281 | 0.1281 | 0.1281 | 0.1281 | 0.1428 | 0.1227 | 0.1227 | 0.1225 |
| 589 | 0.0817 | 0.0822 | 0.0825 | 0.0822 | 0.083 | 0.0805 | 0.0825 | 0.0825 |

Table 6.13: Characteristic Frequency Peak Differences in Absolute Informational Spectrum

| ID | H1N1 1933 | H1N1 1947 | H1N1 1979 | H1N1 2009 | H5N1 | H1N2 | H2N2 | H3N2 |
|---|---|---|---|---|---|---|---|---|
| 73 | 0.4885 | 0.4712 | 0.471 | 0.471 | 0.4707 | 0.02 | 0.02 | 0.0203 |
| 81 | 0.4837 | 0.4839 | 0.0251 | 0.0576 | 0.0588 | 0.4861 | 0.4861 | 0.4861 |
| 110 | 0.4605 | 0.4605 | 0.4607 | 0.0203 | 0.02 | 0.0205 | 0.0203 | 0.0205 |
| 285 | 0.0586 | 0.0586 | 0.4341 | 0.4344 | 0.4346 | 0.0207 | 0.4363 | 0.0205 |
| 359 | 0.4888 | 0.4888 | 0.4885 | 0.0188 | 0.4885 | 0.0215 | 0.458 | 0.0215 |
| 373 | 0.4893 | 0.4898 | 0.49 | 0.4893 | 0.489 | 0.0769 | 0.0761 | 0.0764 |
| 375 | 0.4297 | 0.4292 | 0.4283 | 0.4305 | 0.43 | 0.0381 | 0.0378 | 0.0378 |
| 496 | 0.4463 | 0.4463 | 0.4466 | 0.3902 | 0.0234 | 0.0683 | 0.0686 | 0.0686 |
| 536 | 0.4602 | 0.4602 | 0.4605 | 0.4602 | 0.4602 | 0.0203 | 0.0207 | 0.0205 |
| 574 | 0.0395 | 0.0395 | 0.0395 | 0.0393 | 0.3502 | 0.4863 | 0.4858 | 0.4861 |
| 588 | 0.0576 | 0.0573 | 0.0573 | 0.0576 | 0.0583 | 0.4863 | 0.4861 | 0.4863 |

Table 6.14: Characteristic Frequency Peak Differences in Real Informational Spectrum

| ID | H1N1 1933 | H1N1 1947 | H1N1 1979 | H1N1 2009 | H5N1 | H1N2 | H2N2 | H3N2 |
|---|---|---|---|---|---|---|---|---|
| 73 | 0.4641 | 0.4702 | 0.4705 | 0.4705 | 0.4824 | 0.0195 | 0.0195 | 0.0195 |
| 76 | 0.4702 | 0.4841 | 0.4841 | 0.0859 | 0.0727 | 0.409 | 0.021 | 0.021 |
| 81 | 0.0246 | 0.4841 | 0.0249 | 0.1796 | 0.0591 | 0.4858 | 0.4858 | 0.4858 |
| 110 | 0.4602 | 0.4605 | 0.4605 | 0.0193 | 0.021 | 0.4566 | 0.0195 | 0.4566 |
| 154 | 0.0581 | 0.4841 | 0.0561 | 0.0561 | 0.0591 | 0.3773 | 0.4858 | 0.4858 |
| 343 | 0.4902 | 0.1083 | 0.1083 | 0.0815 | 0.49 | 0.1088 | 0.4339 | 0.4339 |
| 359 | 0.4883 | 0.4888 | 0.4885 | 0.0193 | 0.0188 | 0.0215 | 0.4585 | 0.0215 |
| 373 | 0.491 | 0.4888 | 0.489 | 0.4888 | 0.4905 | 0.0781 | 0.0761 | 0.0761 |
| 375 | 0.43 | 0.428 | 0.4278 | 0.43 | 0.4305 | 0.0371 | 0.0371 | 0.0371 |
| 427 | 0.0224 | 0.0227 | 0.0227 | 0.3895 | 0.0224 | 0.4366 | 0.4366 | 0.3199 |
| 536 | 0.4605 | 0.4605 | 0.4607 | 0.0212 | 0.4595 | 0.0193 | 0.0212 | 0.0212 |
| 574 | 0.0403 | 0.04 | 0.04 | 0.0398 | 0.039 | 0.4863 | 0.4861 | 0.4861 |
| 588 | 0.0581 | 0.0581 | 0.0581 | 0.0581 | 0.0588 | 0.4861 | 0.4858 | 0.4858 |

Table 6.15: Characteristic Frequency Peak Differences in Imaginary Informational Spectrum

| ID | H1N1 1933 | H1N1 1947 | H1N1 1979 | H1N1 2009 | H5N1 | H1N2 | H2N2 | H3N2 |
|---|---|---|---|---|---|---|---|---|
| 73 | 0.4632 | 0.4714 | 0.4714 | 0.4714 | 0.4837 | 0.0205 | 0.0203 | 0.0205 |
| 76 | 0.0722 | 0.4832 | 0.4832 | 0.0717 | 0.4839 | 0.408 | 0.0754 | 0.0754 |
| 81 | 0.0259 | 0.4834 | 0.0261 | 0.0571 | 0.0578 | 0.4871 | 0.4868 | 0.4868 |
| 359 | 0.4893 | 0.4898 | 0.4898 | 0.0183 | 0.4888 | 0.0224 | 0.0205 | 0.0205 |
| 375 | 0.4307 | 0.429 | 0.429 | 0.4309 | 0.4295 | 0.0381 | 0.0381 | 0.0381 |
| 285 | 0.4334 | 0.1096 | 0.4336 | 0.4336 | 0.4346 | 0.0203 | 0.4356 | 0.0205 |
| 343 | 0.4893 | 0.1074 | 0.1074 | 0.0825 | 0.489 | 0.0637 | 0.4351 | 0.4348 |
| 421 | 0.0232 | 0.0237 | 0.0237 | 0.3902 | 0.0234 | 0.4358 | 0.4356 | 0.4356 |
| 434 | 0.4653 | 0.0237 | 0.0237 | 0.4653 | 0.0234 | 0.4356 | 0.4356 | 0.4356 |
| 536 | 0.4649 | 0.185 | 0.4597 | 0.4597 | 0.4605 | 0.0203 | 0.0203 | 0.0203 |
| 557 | 0.0569 | 0.02 | 0.02 | 0.0571 | 0.0195 | 0.4085 | 0.4361 | 0.4361 |
| 574 | 0.0393 | 0.039 | 0.039 | 0.0388 | 0.0403 | 0.4873 | 0.4851 | 0.4854 |
| 588 | 0.0569 | 0.4029 | 0.0571 | 0.0571 | 0.0578 | 0.4868 | 0.4868 | 0.4868 |

# Chapter 7

# Investigation into the Effects of an Individual Amino Acid on Protein Function by Means of Discrete Fourier Transform

## 7.1 Introduction

In recent years, a number of tools that are capable of directly identifying the mechanism by which proteins interact with their environment have been developed. Several of these tools are able to directly identify rules from primary protein structures. Two popular tools are the Basic Local Alignment Search Tool (BLAST) [241] and Protein Feature Server (PROFEAT) [56]. The former searches for similarities in the arrangements of amino acids in proteins while the latter extracts structural and physicochemical features from protein sequences, which can then be combined with statistical and/or machine learning methods for prediction of properties of proteins [79; 277; 278; 279]. Capturing sequence order is an important but challenging problem, for which there have been various methods developed [280; 281; 282].

Matching biological functions to features that are extracted by signal processing is another technique that is available in order to deal with capturing sequence order information. An example of this is the Resonant Recognition Model (RRM) [91; 92] that uses Discrete

Fourier Transform (DFT) and electron-ion interaction potential (EIIP) amino acid scale [87]. By using this method, characteristic features of proteins or families of proteins, can be identified which can be related to a specific protein function. With RRM each of the protein classes under study can be related to the biological function that they represent, by identifying a unique peak or set of peaks extracted from the frequency spectrum. Each unique peak extracted using RRM is called a Characteristic Frequency Peak (CFP). Further information regarding the RRM and Complex Resonant Recognition Model (CRRM) can be found in Chapters 2 and 6, respectively.

After identifying a unique peak or set of peaks which can be related to specific biological function by using signal processing techniques, a question that arises is which amino acid of the protein sequences contributes most to the specific feature extracted? One approach [42; 247] in the literature uses RRM to determine which section of protein is a dominant contributor to CFP. This section can be identified by measuring the magnitude of DFT at the CFP position, and selecting the window with the highest value. In this study, a new approach that measures the effect of individual amino acids of protein sequences upon CFP extracted from RRM is developed and presented. For this analysis, five different protein classes of influenza A virus Neuraminidase (NA) protein, which includes H1N1, H1N2, H2N2, H3N2 and H5N1 subtypes, are used to show applicability and robustness of the method developed.

## 7.2 Methodology

In this section, the influence of individual amino acid on the common frequency peak and on the absolute spectrum is described [283]. The difference between these methods is on the selection of the frequency. The former method selects the frequency based on the common frequency peak that is calculated by using the RRM. The latter method calculates the effect of individual amino acids from the entire frequency spectrum and selects the frequency that presents the highest change.

### 7.2.1 Influence of Individual Amino Acid on the Common Frequency Peak

The following algorithm must be followed to determine how individual amino acid can affect the magnitude of DFT at the CFP position.

**STEP 1:** Calculate the CFP position for given protein sequences using Informational Spectrum Analysis.

**STEP 2:** Calculate the magnitude of DFT at the CFP position for all original protein sequences.

**STEP 3:** Remove single amino acid at position X from original protein sequences with length N.

**STEP 4:** Recalculate the magnitude of DFT at the CFP position for all modified protein sequences.

**STEP 5:** Compare DFT at the CFP position between original and modified protein sequences.

**STEP 6:** Repeat STEP 3 to 5 for all N amino acids in protein sequences.

After all the six steps are completed the outcome is a measurement for all individual amino acids of all given protein sequences.

### 7.2.2 Influence of Individual Amino Acid on the Absolute Spectrum

In the previous section an algorithm is described to determine the effects of individual amino acids on the CFP in the spectrum determined by RRM. In this section an additional algorithm will be described on how individual amino acid can affect the entire absolute spectrum in different locations and at different magnitude.

**STEP 1:** Calculate the absolute spectrum for given protein sequences.

**STEP 2:** Remove single amino acid at position X from original protein sequences with length N.

**STEP 3:** Recalculate the absolute spectrum for all modified protein sequences.

**STEP 4:** Compare the absolute spectrum between original and modified protein sequences.

**STEP 5:** Repeat STEP 2 to 4 for all N amino acids in protein sequences.

**STEP 6:** Select the frequency with the maximum value which represent the maximum effect in the absolute spectrum

After all five steps are completed the outcome is a measurement for all individual amino acids of all given protein sequences.

## 7.3 Case Study: Analysing Influenza A NA protein Sequences

### 7.3.1 Protein Sequences

For the analysis five Influenza A Neuraminidase (NA) proteins were retrieved from the Influenza Virus Resource data set [263]. Further information regarding the protein sequences used can be found in Table 7.1. A pairwise percent identity for all the proteins that belong to the family of influenza NA proteins was calculated by using CLUSTALW [200] and is given in Table 7.2. A high pairwise percent identity is observed between H1N1 and H5N1 (87%), H1N2 and H2N2 (85%) and H1N2 and H3N2 (87%) NA proteins. Low percent identity is observed between H1N1 and H1N2 (40%), H1N1 and H2N2 (42%), H1N1 and H3N2 (39%), H5N1 and H1N2 (41%), H5N1 and H2N2 (42%) and H5N1 and H3N2 (41%).

Table 7.1: Influenza A NA protein Sequences Used for the Case Study

| ID | Name | Length | Reference |
|---|---|---|---|
| ADK33724 | A/Aarhus/INS242/2009(H1N1) | 469 | [284] |
| ADG59213 | A/Anhui/1/2005(H5N1) | 449 | [285] |
| CAD29972 | A/Egypt/84/2001(H1N2) | 469 | [286] |
| ABO52305 | A/Albany/3/1958(H2N2) | 469 | [284] |
| ACF22356 | A/Hong Kong/1-2-MA21-2/1968(H3N2) | 469 | [284] |

### 7.3.2 Results

In this section, the results produced the influence of individual amino acid on the common frequency peak and on the absolute spectrum will be presented and discussed.

Table 7.2: Pairwise Percent Identity

| | H1N1 | H1N2 | H2N2 | H3N2 |
|---|---|---|---|---|
| H1N2 | 40% | - | - | - |
| H2N2 | 42% | 85% | - | - |
| H3N2 | 39% | 87% | 94% | - |
| H5N1 | 87% | 41% | 42% | 41% |

#### 7.3.2.1 Influence of Individual Amino Acid on The Common Frequency Peak

By following the method described in Section 7.2.1 the following results are obtained for the five influenza A NA proteins. Step 1 requires that the CFP position for the protein sequences is calculated. For the five influenza A NA proteins, Figure 7.1 shows the CFP was found to be 0.4583. By following the remaining Steps 2 to 6, the influence of all amino acids of Influenza NA proteins is calculated. Figures 7.2, 7.3, 7.4, 7.5 and 7.6 show the results for H1N1, H5N1, H1N2, H2N2 and H3N2 NA proteins respectively. Each point in the figures gives the impact of a particular amino acid on DFT at the CFP position. All the results are presented in percentage to the original value of DFT at the CFP position. A threshold can be used to adjust the identified location according to the individual amino acid impact. For this analysis, a 60% threshold is applied to the results in order to obtain a sufficient area from the protein sequences. Four key areas are identified by applying the 60% threshold to the results.

These areas present the highest impact on DFT at CFP position. The key areas are A, B and C as displayed in Figures 7.2 and 7.3 between H1N1 and H5N1 NA proteins. Area D is displayed in figures 7.4, 7.5 and 7.6 is also identified between H1N2, H2N2 and H3N2 NA proteins. All amino acids that correspond to areas A, B, C and D can be found in Tables 7.3, 7.4, 7.5, 7.6 and 7.7 for H1N1, H5N1, H1N2, H2N2 and H3N2, respectively .

Table 7.3: High impact areas for H1N1 NA protein

| Area | Residues | Sequence |
|---|---|---|
| A | 159-185 | MSCPIGEVPSPYNSRFESVAWSASACH |
| B | 193-301 | IGISGPDNGAVAVLKYNGIITDTIKSWRN |
| | | NILRTQESECACVNGSCFTVMTDGPSD |
| | | GQASYKIFRIEKGKIVKSVEMNAPNYH |
| | | YEECSCYPDSSEITCVCRDNWHGSNR |
| C | 325-348 | NPRPNDKTGSCGPVSSNGANGVKG |

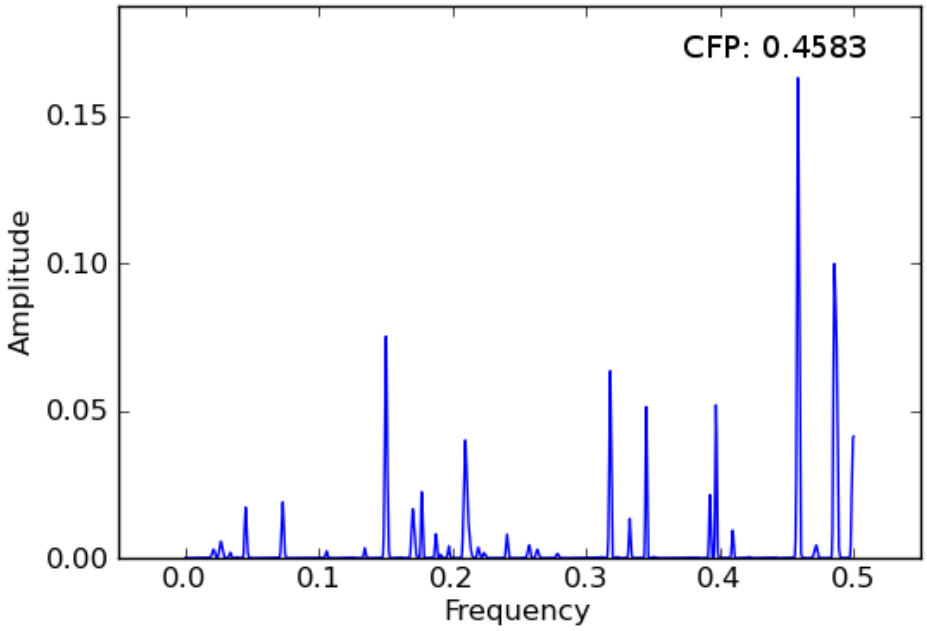



Figure 7.1: Informational Spectrum Analysis Results

By using the identified areas as shown in Tables 7.3-7.7 three segments that exist unchanged in influenza A proteins between H1N1 and H5N1 and two segments between H1N2, H2N2 and H3N2 NA proteins are identified. For H1N1 and H5N1 NA proteins the following segments are identified

- **PSPYNSRFESVAWS** from A area,
- **IGISGPDNGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTVMTS** from B1 area and
- **EITCVCRDNWHGSN** from B2 area.

For H1N2, H2N2 and H3N2 NA proteins the following segments are identified

- **GRLVDSIGSWS** and
- **ILRTQESECVCINGTC** from area D.

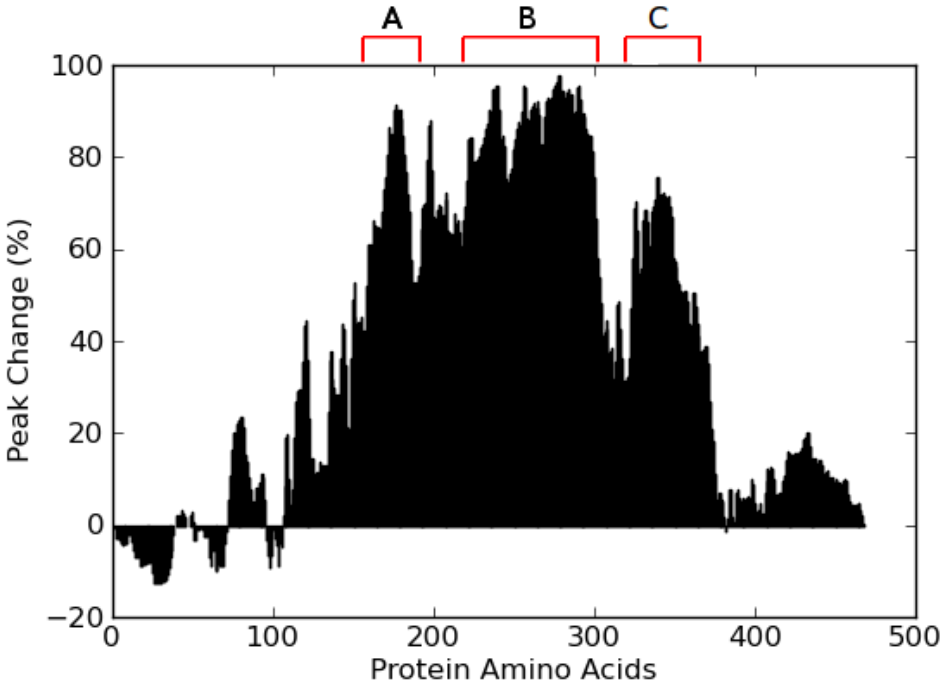



Figure 7.2: H1N1 Results

#### 7.3.2.2 Influence of Individual Amino Acid to Absolute Spectrum

For this analysis, one H1N1 NA protein is selected, A/Brevig Mission/1/1918 H1N1 (3BEQ) [287], for which extensive information exists in the literature regarding the primary, secondary and tertiary structures, as well as the binding site with NA inhibitors. By following the method described in Section 7.2.2 the following results are obtained for the five influenza A NA proteins. Step 1 requires that the absolute spectrum for the protein sequences is calculated. As steps 2 and 3 describe, by removing single amino acid from the H1N1 NA protein, the absolute spectrum is recalculated for the modified sequence. The subsequent step is to compare these spectra calculated, as Figure 7.7 shows where the $X$ axis represents frequencies and the $Y$ axis represents the amino acids of the protein sequence. The next step in the analysis is to select the frequency with the maximum value, which represents the maximum effect of individual amino acids in the absolute spectrum. This frequency is 0.3735 as it is highlighted in Figure 7.7 and can be observed in more detail in Figure 7.8. Furthermore, by using Figure 7.8, three consecutive areas, $S_1$, $S_2$ and $S_3$, are identified. Additionally, the identified areas, $S_1$, $S_2$ and $S_3$, correspond to specific amino

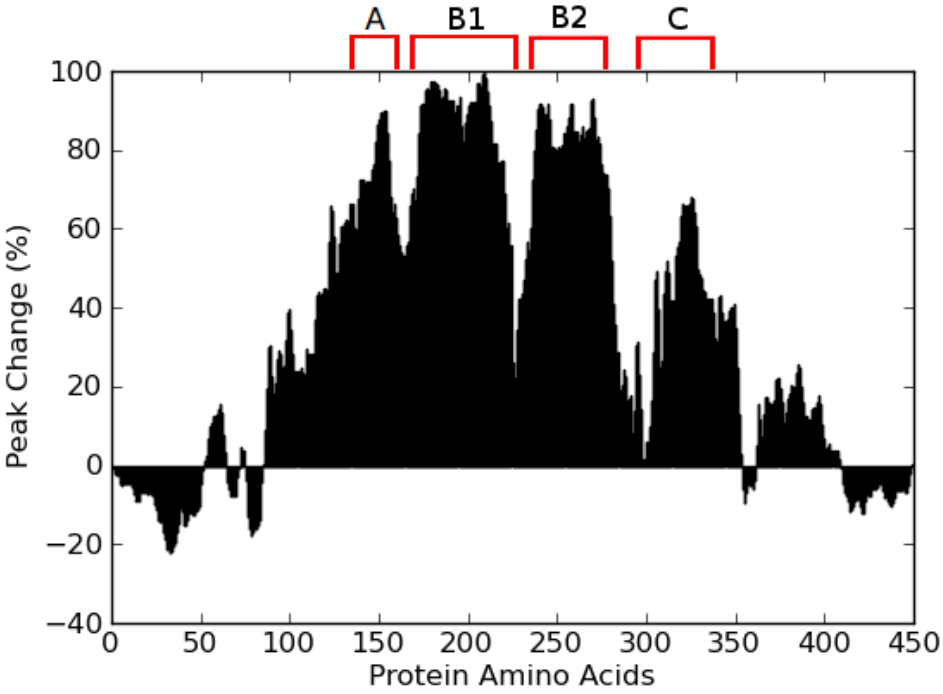



Figure 7.3: H5N1 Results

acids on the protein sequence under investigation. Figure 7.9 highlights these amino acids selected with colour coding on the tertiary structure of H1N1 NA as Figures 7.10 and 7.11 show. As the results indicate, the identified areas $S_1$, $S_2$ and $S_3$ converge in an active site within the structure that appears in the tertiary structure of H1N1 NA as Figure 7.10 shows. From the literature [287], zanamivir, which is an antiviral, binds to this location as Figure 7.11 shows.

## 7.4 Conclusions and Discussions

Upon identification of a new protein, it is important to single out these amino acids responsible for the structural classification of the protein, as well as the amino acids contributing to the protein's specific biological characterisation. In this chapter, a novel approach is presented to identify and quantify this cause and effect relationship between amino acid and protein. Two methods are presented in this chapter; the first method takes into consideration the frequency peak, CFP, which it calculated by using the RRM, and the sec-

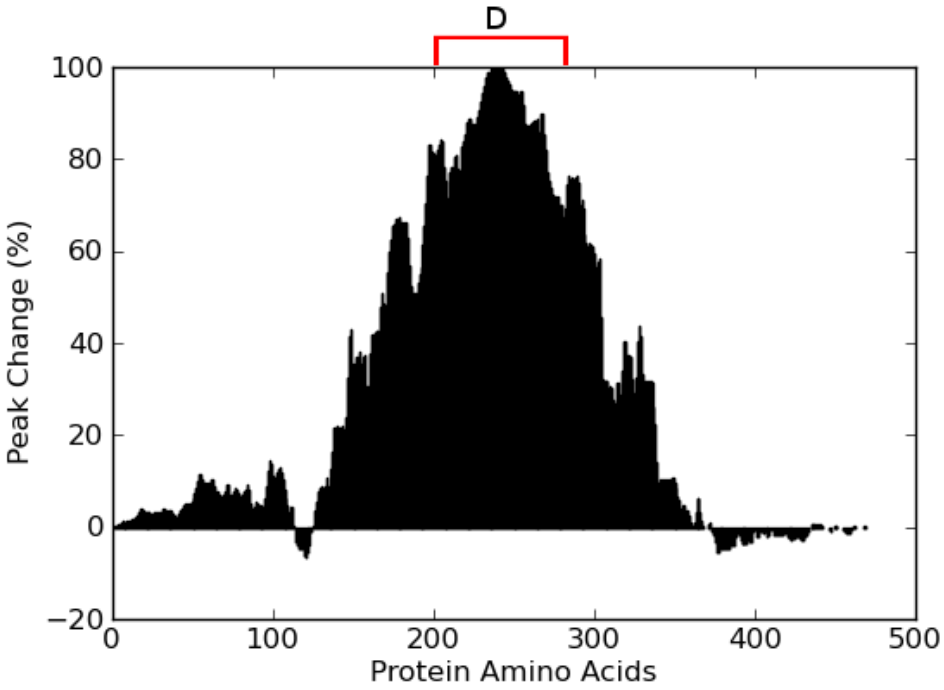



Figure 7.4: H1N2 Results

ond technique considers the entire absolute spectrum. Applicability and robustness of the methods are shown on a case study where five different protein families of the influenza A virus NA proteins, which includes H1N1, H1N2, H2N2, H3N2 and H5N1 NA proteins, are studied.

For the first method described, which studies the influence of individual amino acid to CFP for each of the Influenza A NA proteins studied, areas A, B, C and D that have a high impact on DFT at CFP position as shown in Figures 7.2 to 7.6, are identified. The analyses identified five segments, three between H1N1 and H5N1 and two between H1N2, H2N2 and H3N2 and suggested that they play a key role in Influenza A NA protein functionality and can potentially be considered as target areas for future antiviral drugs and vaccines such as neuraminidase inhibitors [262]. For the second method, which studies the influence of individual amino acid, the frequency 0.3735 is highlighted to present the highest impact to absolute spectrum. Furthermore, by using this distinguished frequency three consecutive areas, $S_1$, $S_2$ and $S_3$ are identified. As the results show the identified areas $S_1$, $S_2$ and $S_3$ converge in an active site within the structure that appears in the tertiary structure of

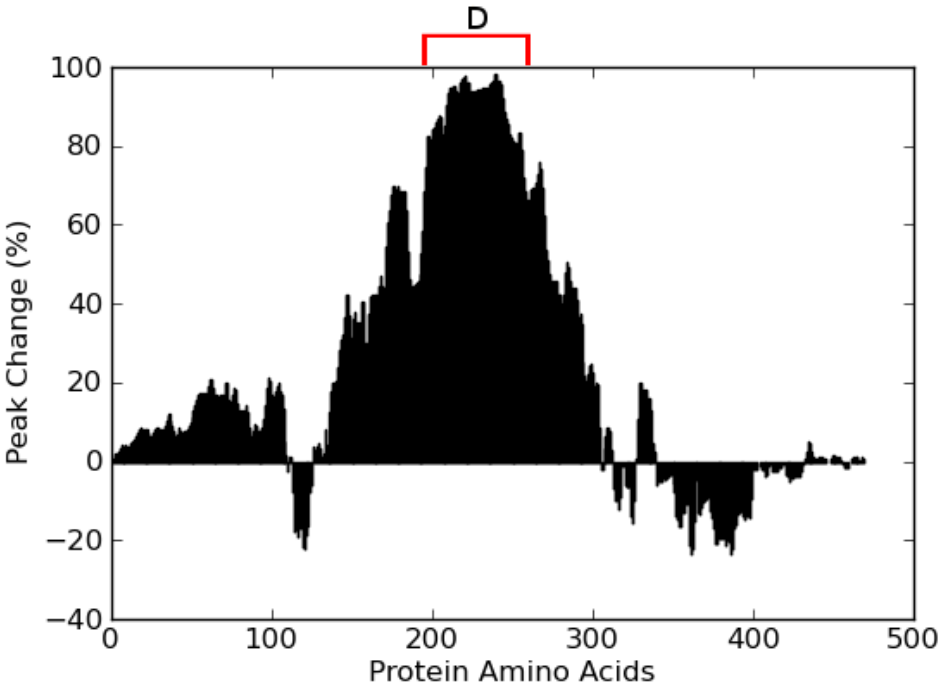



Figure 7.5: H2N2 Results

H1N1 NA as shown in Figures 7.10, where according to the literature [287], the antiviral zanamivir binds to this location as shown in Figure 7.11.

The biological functionality, represented by the amino acid index used to encode protein sequences to numerical sequences, can directly be linked to the identified segments of the protein sequences. In this study, EIIP was used, but more than 611 unique amino acid indices exist as discussed in in Chapter 3. These amino acid indices can be used to encode protein sequences and thus utilised for further analysis that may help reveal additional important segments of the protein sequences under study.

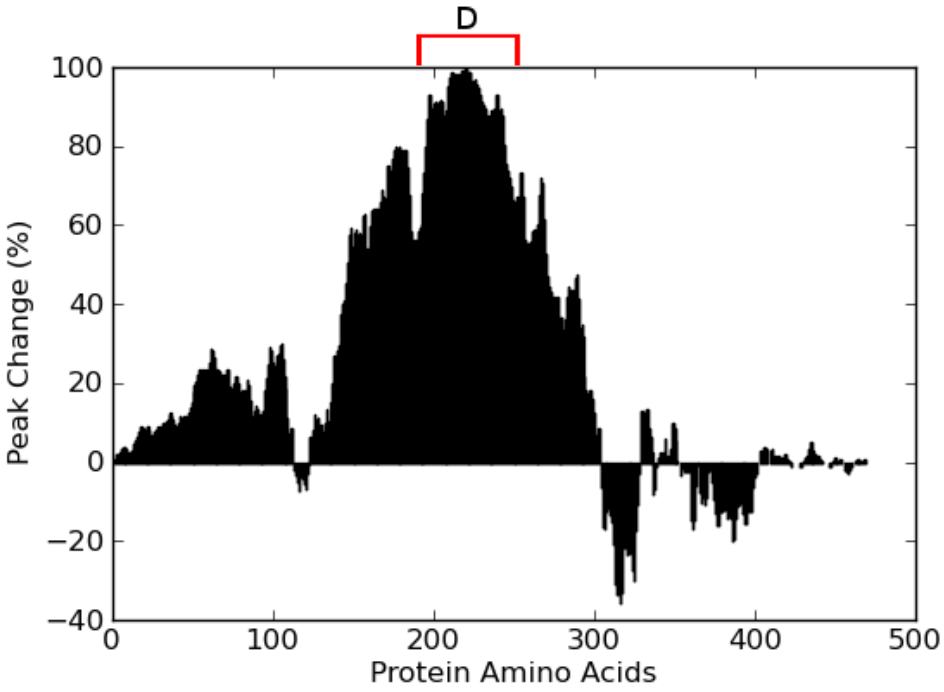



Figure 7.6: H3N2 Results

Table 7.4: High impact areas for H5N1 NA protein

| Area | Residues | Sequence |
|---|---|---|
| A | 138-160 | LMSCPVGEAPSPYNSRFESVAWS |
| B1 | 167-223 | GTSWLTIGISGPDNGAVAVLKYNGIITDT |
| | | IKSWRNNILRTQESECACVNGSCFTVMT |
| B2 | 235-280 | IFKMEKGKVVKSVELNAPNYHYEECSC |
| | | YPDAGEITCVCRDNWHGSN |
| C | 319-328 | SPNGAYGIKG |

Table 7.5: High impact areas for H1N2 NA protein

| Area | Residues | Sequence |
|---|---|---|
| D | 193-299 | CVTGDDKNATASFIYNGRLVDSIGSWS |
| | | KKILRTQESECVCINGTCAVVMTDGSA |
| | | SGKADTKILFIEEGKIGHTSLLSGSAQ |
| | | HVEECSCYPRYPGVRCVCRDNWKGSN |

Table 7.6: High impact areas for H2N2 NA protein

| Area | Residues | Sequence |
|---|---|---|
| D | 195-270 | TGDDRNATASFIYDGRLVDSIGSWSQN ILRTQESECVCINGTCTVVMTDGSASG RADTRILFIKEGKIVHISPLSG |

Table 7.7: High impact areas for H3N2 NA protein

| Area | Residues | Sequence |
|---|---|---|
| D | 192-256 | VCITGDDKNATASFIYDGRLVDSIGSW SQNILRTQESECVCINGTCTVVMTDGS ASGRADTRILF |

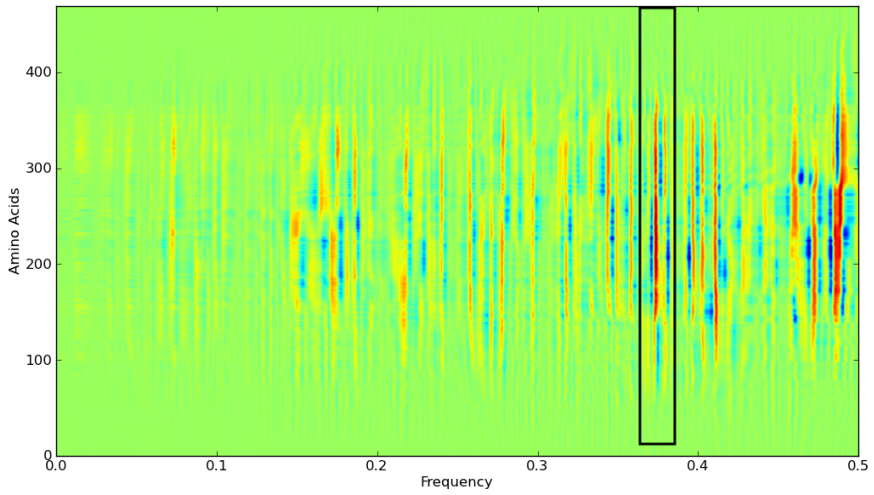



Figure 7.7: Effects of single amino acid on absolute spectrum for H1N1 NA protein

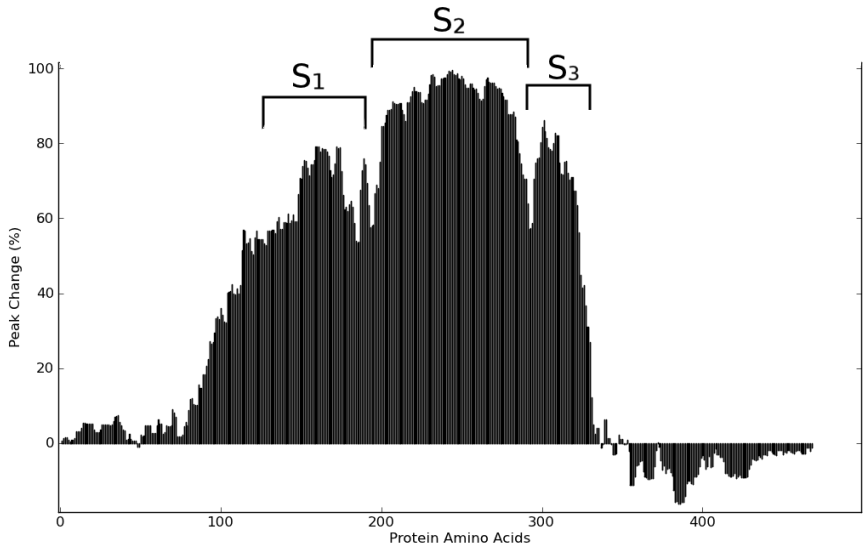



Figure 7.8: Results for frequency 0.3735 for H1N1 NA protein

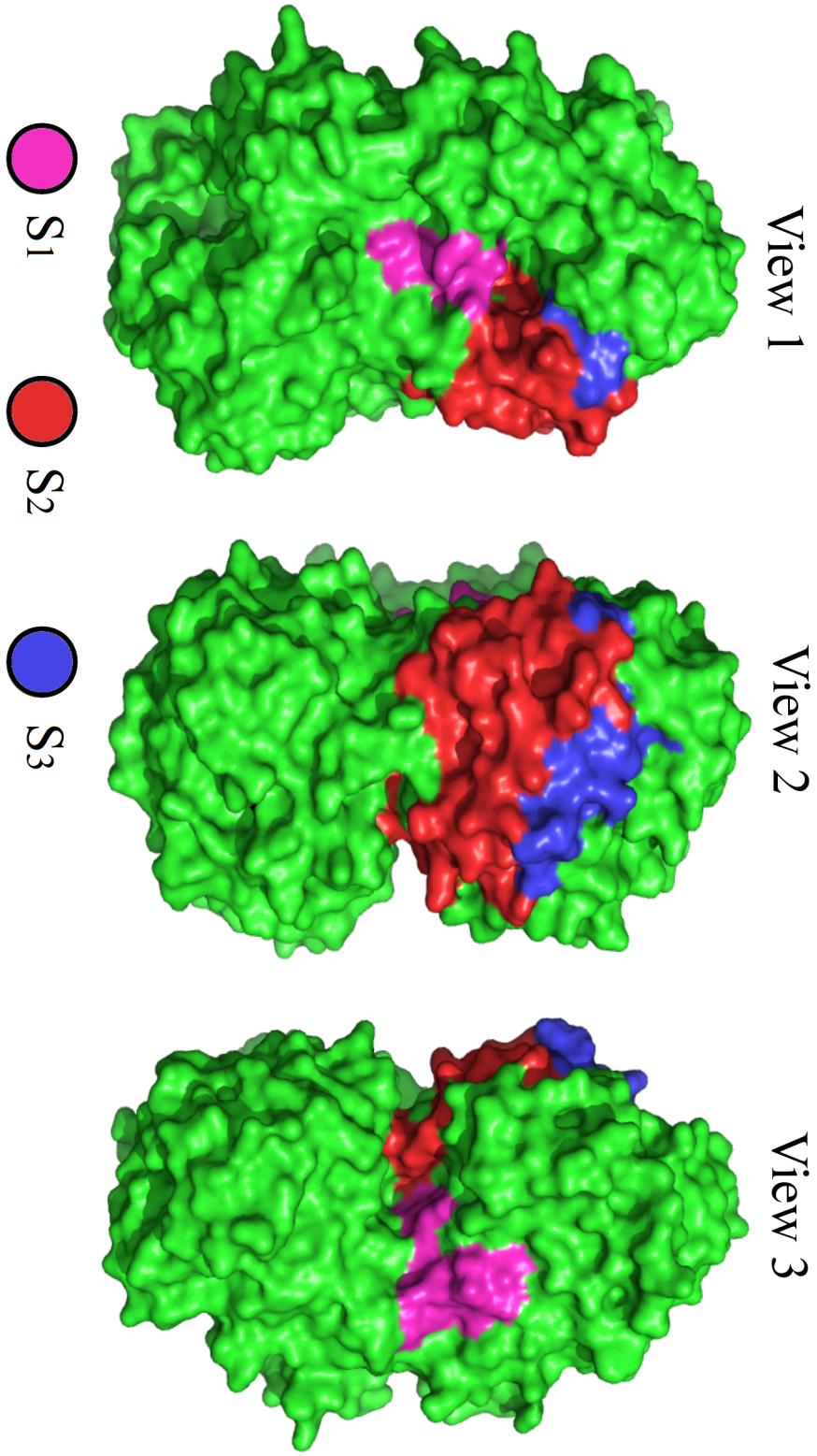



Figure 7.9: Views of H1N1 NA protein with the identified areas

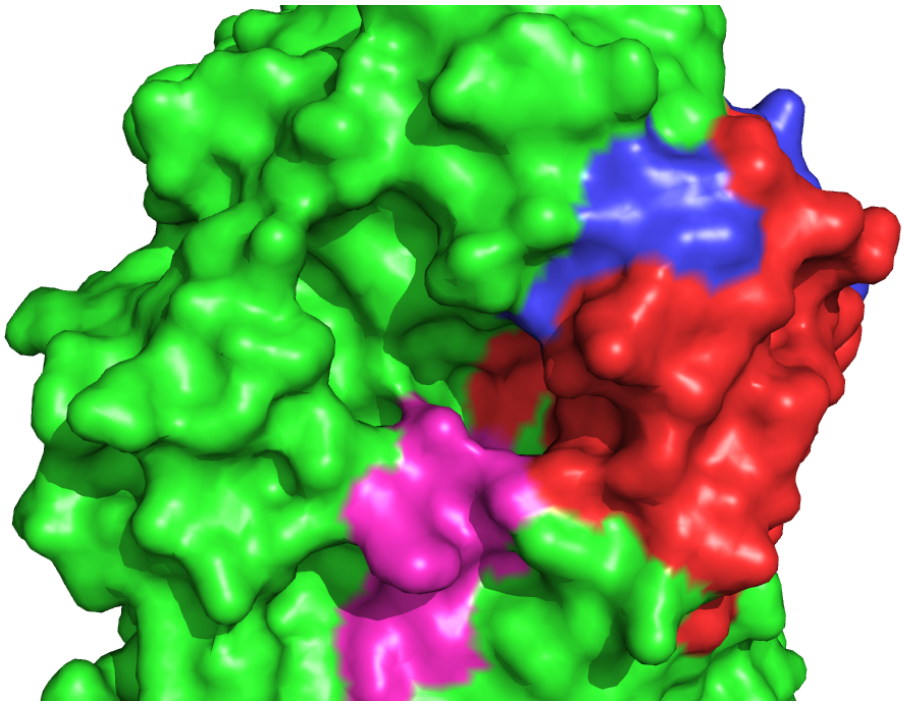


Figure 7.10: H1N1 NA protein active site within the structure

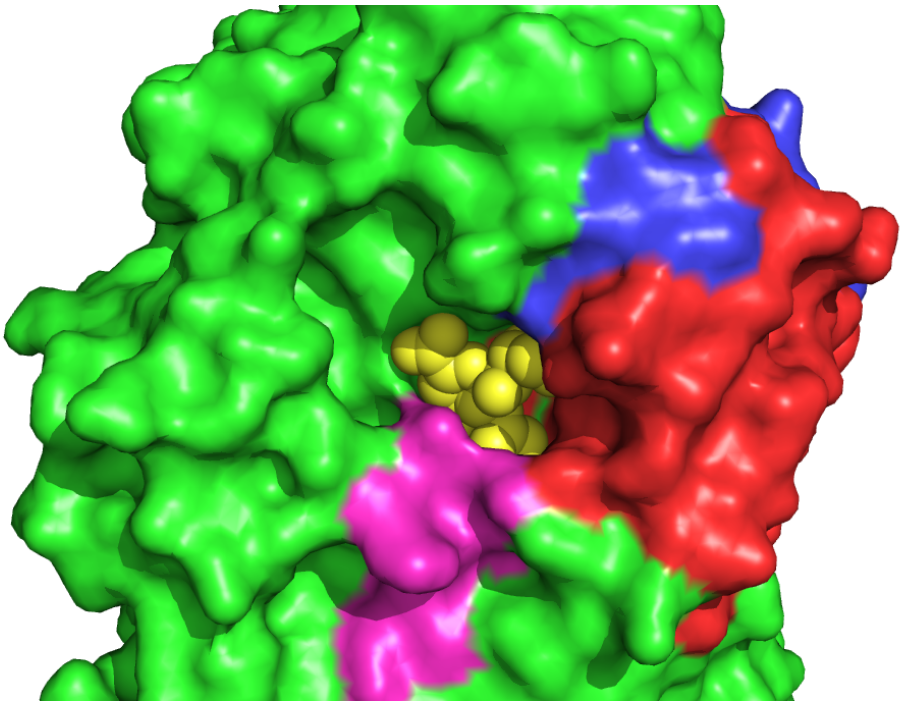


Figure 7.11: H1N1 NA protein active site within the structure with zanamivir

# Chapter 8

# Signal-processing based bioinformatics approach for Subgroup Discovery and Classification of Protein Sequences

## 8.1 Introduction

In recent years, decoding the rules that drive biological functions of proteins directly from their primary structures, has become a subject of intensive research with one example being the Basic Local Alignment Search Tool (BLAST) [241]. Signal processing-base techniques such as Resonant Recognition Model (RRM) [40; 91; 92] and Complex Resonant Recognition Model (CRRM) [245] have been introduced in bioinformatics to extract information that is expected to match protein biological functions. The study is performed using the algorithms that help derive meaningful knowledge from the proteins based on features extracted from the signal processing techniques. One such example is the examination of the relationship between a different protein sequences subgroups. These results can then help develop therapy and/or vaccines.

Signal processing techniques can generate a large amount of information, which can be related to a protein's biological function. The RRM and CRRM are only two of the techniques that try to identify which of the features extracted are related to the protein biological function. New methods are required to be able to identify all the important features that can be related to the bioinformatics problem, and to discard any ineffective or

noisy data. In this chapter, two algorithms will be used to be able to determine if a feature, or a set of features, extracted from protein sequences using signal-processing techniques can be used to characterise different protein classes. The algorithms used are the Subgroup Discovery (SD) [288; 289] and Support Vector Machine (SVM) [153; 154].

SD can have both predictive and descriptive orientation, and its objective is to find interpretable rules to describe relationships in the data. SD algorithms have been applied to bioinformatics problems such as cancer diagnosis [290; 291; 292; 293; 294]. These types of bioinformatics problems can be characterised as complex, in respect to the number of variables (between 7000 and 22,000), and with a low number of cases (not more than 200 cases). In various applications [290; 291; 292; 293; 294], good results of the SD algorithms in solving bioinformatics problems were presented, where novelty and interpretable models were obtained. The Non-dominated Multiobjective Evolutionary Algorithm for Extracting Fuzzy Rules in Subgroup Discovery (NMEEF-SD) [295] is employed as it is a novel approach which has been shown to yield more accurate results [295]. These properties have been considered by the experts in applying NMEEF-SD to the influenza A virus problem with several objectives, for obtaining a model, which can describe relationships in an interpretable way, and for predicting the type of the virus for new proteins introduced in the data set.

Additionally, SVM [153; 154] is implemented to create a classification model, which can be used to model relationships between protein sequences. SVM is a supervised statistical learning method that analyses data and recognises patterns for classification. The SVM takes a set of input data and predicts to which of two or more possible classes each given input belongs. SVM is selected as it can produce accurate and robust classification results on a established theoretical basis even when input data are noisy or non-linearly separable [296; 297].

Therefore, an exhaustive experimental study with the NMEEF-SD algorithm and SVM is presented for the problem, where different configurations of the algorithms are used to find the best model. By using the NMEEF-SD algorithm, the study is tackled from two different perspectives. On the one hand, a complete study from the point of view of SD task is performed to find the best parameters employed for the algorithm to solve the problem with respect to interpretability, novelty and precision. On the other hand, a predictive analysis is performed to show the ability of the NMEEF-SD algorithm to classify new proteins in the different virus studied. By using the SVM algorithm, the study tries to find

the best parameters to build the most accurate classification model. Furthermore, F-score is used in combination with SVM to select the top-related features for the classification model.

Finally, the predictive models obtained by the NMEEF-SD and SVM with the complete data set is presented to show the more representative subgroups and classification models created for the influenza A virus problem. An exhaustive study of different neuraminidase genes of influenza A virus subtypes is presented in this chapter: H1N1, H2N2, H3N2 and H5N1 NA subtypes are used in this analysis. These protein sequences were chosen for the high percentage identity they demonstrate within individual influenza subtype classes and the high variation they display in percent identity.

The chapter is organised as follows: Section 8.2 presents the methods and materials used in this chapter. Section 8.2.1 describes the signal processing methods used to extract protein related features. Sections 8.2.2 and 8.2.4 describes the methods used in this analysis, NMEEF-SD and SVM respectively. Section 8.3 presents a case study with influenza subtypes sequences and the results obtained by employing NMEEF-SD and SVM. Finally, conclusions are discussed in Section 8.4.

## 8.2 Methods and Materials

### 8.2.1 Signal Processing For Protein Sequence Analysis

By using digital signal processing techniques, the goal is to extract information that can be related to biological functions of proteins. Various methods have been used in bioinformatics for analysing protein sequences in recent years, and one of the most common methods is the RRM [40; 91; 92] and CRRM [245]. Previous studies [42] used influenza A subtypes to analyse the hemagglutinin (HA) gene, with RRM aiming to identify new therapeutic targets for drug development by better understanding the interaction of the influenza virus and its receptors.

In contrast to previous studies, the analysis was performed directly to absolute spectrum, which is derived by applying Discrete Fourier Transform (DFT) to each numerical encoded protein sequence. Electron-ion interaction potential (EIIP) [87; 88] amino acid index, as shown in Table 2.3, is used as express protein sequences to numerical sequences in order to be able to apply DFT. For the analysis of influenza A virus proteins, as the

sequences have different lengths, zero-padding was used to extend all protein sequences to $N = 512$ thus the output of the absolute spectrum is 256 features. Further information regarding RRM, CRRM can be found in Chapters 2 and 6, respectively.

### 8.2.2 Subgroup Discovery Technique [1]

The concept of Subgroup Discovery Technique (SD) was initially introduced by Kloesgen [288] and Wrobel [289]. SD [288; 289] is a data mining technique that aims at discovering relationships between properties of a given data set in respect to a specific property defined by the user. An induced subgroup can be represented by a rule ($R$) [298; 299]:

$$R : Cond \rightarrow Class$$

where $Class$ represents the target variable for SD, and $Cond$ is a combination of attribute-value features that describe a statistical distribution in respect to the $Class$.

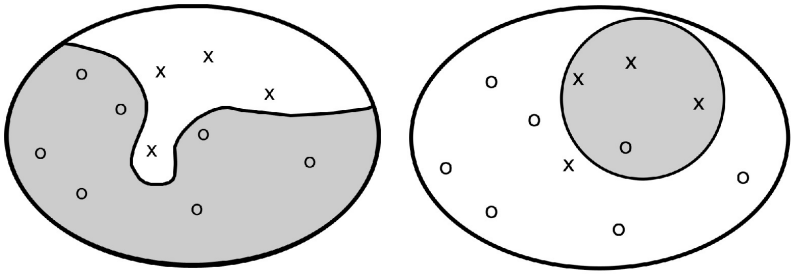
SD has a combination of predictive and descriptive orientation, and it is differentiated from classification techniques because SD attempts to describe knowledge for the data while a classifier attempts to predict it. Furthermore, the model obtained by a SD algorithm is usually simple and interpretable, while that obtained by a classifier is complex and precise.

As Figure 8.1 shows, the model obtained from the classifier (Fig. 8.1(a)) is more complex than the model obtained from the SD technique (Fig. 8.1(b)). In addition, the accuracy of the classification model is greater than the accuracy obtained by the SD model, but, with respect to the interpretability the best results are obtained from the SD model. In conclusion, the SD algorithm obtains simple models to describe behaviour of the data with a good level of accuracy.

Despite the lack of consistency of the quality measures utilised in SD methods, different studies [295; 300; 301; 302; 303] propose measures like sensitivity, unusualness, confidence and significance as the most important measurements of the quality of the subgroups.

- Sensitivity [288] is used to quantify the quality of individual rules according to the individual patterns of interest covered. It is a measure with precision and generality characteristics and it can be computed as:

[1] This section is mainly covered in a collaboration with Cristbal Jos Carmona del Jesus to investigate applications of Signal Processing and Subgroup Discovery Techniques in analysis of protein sequences.



(a) Classification Model (b) Subgroup Discovery Model

Figure 8.1: Illustration of The Difference Between Classification and Sub-group Discovery

$$Sens(R) = \frac{tp}{Pos} \tag{8.1}$$

where $tp$ are the examples correctly classified, and $Pos$ are the examples of the class.

- Unusualness [304] attempts to obtain a trade-off between generality, interest and precision in the results, and it can be computed as:

$$Unus(R) = \frac{tp + fp}{n_s} \cdot \left( \frac{tp}{tp + fp} - \frac{Pos}{n_s} \right) \tag{8.2}$$

where $fp$ are the examples incorrectly classified, and $n_s$ are the examples of the data set.

- Confidence [305] measures the relative frequency of examples satisfying the complete rule among those satisfying only the antecedent. It can be computed as:

$$Conf(R) = \frac{tp}{tp + fp} \tag{8.3}$$

- Significance [288] indicates the significance of a finding, if measured by the likelihood ratio of a rule, and it can be computed as:

$$Sign(R) = 2 \cdot \sum_{k=1}^{n_c} tp_k \cdot log \frac{tp_k}{Pos_k \cdot \frac{tp_k + fp_k}{n_s}} \tag{8.4}$$

where $n_c$ is the number of classes to study. It must be noted that, although each rule is for a specific $Class$, the significance measures the novelty in the distribution impartially for all the values of the class, hence the use of the summation.

With this set of quality measurements the subgroups obtained by the SD model are evaluated in a correct and quantifiable way.

### 8.2.3 NMEEF-SD: Non-dominated Multi-objective Evolutionary Algorithm For Extracting Fuzzy Rules in Subgroup Discovery [2]

In the literature different Evolutionary Fuzzy Systems (EFSs) can be found, such as SDIGA [306], MESDIF [307] and NMEEF-SD [295] in solving SD problems. Despite the good behaviour of this group of algorithms to solve the SD problem, in this chapter the NMEEF-SD algorithm is employed because it is a novelty approach which obtains significant and accurate results as can be observed in [295].

The NMEEF-SD [295] is an EFS whose objective is to extract descriptive fuzzy and/or crisp rules for the SD task, depending on the type of variables presented in the given problem. This technique is based on a multi-objective approach, the Nondominated Sorting Genetic Algorithm II (NSGA-II) [308] algorithm, which is a computationally fast multi-objective evolutionary algorithm based on a non-dominated sorting approach, and on the use of elitism.

The fuzzy logic is used to represent the continuous features, by using linguistic variables, which allow data mining processes to use numerical features without the need of incensement of the interpretability of the extracted knowledge by their discretisation. A linguistic variable such as temperature may have a value such as low, medium or high. The import aspect of linguistic variables is that they can be modified via linguistic boundaries associated with certain functions. The continuous variables are considered linguistic, and the fuzzy sets corresponding to the linguistic labels can be specified by the user or defined by means of a uniform partition, if knowledge from experts is not available.

With respect to the representation of the rules, NMEEF-SD employs the "*Chromosome = Rule*" approach, where only the antecedent is represented in the chromosome and the consequent is prefixed to one of the possible values of the target feature in the evolution.

[2] This section is mainly covered in a collaboration with Cristbal Jos Carmona del Jesus to investigate applications of Signal Processing and Subgroup Discovery Techniques in analysis of protein sequences.

Therefore, the algorithm must be executed as many times as the number of different values the target variable contains. It uses an integer representation model with as many genes as variables contained in the original data set without considering the target variable. Thus, the set of possible values for the categorical features is indicated by the problem, and for numerical variables it is the set of linguistic terms determined heuristically, or with expert information.

The quality measures considered as objectives in the evolutionary process are selected depending on the nature of the problem. In this chapter, NMEEF-SD uses $Sensitivity$ and $Unusualness$ as objectives for the multi-objective approach. The set of rules extracted for the algorithm are filtered with respect to a minimum confidence threshold, which is defined as a parameter of the algorithm. This confidence is a variation of the $Confidence$ presented in this chapter, which is the $Fuzzy\ Confidence$ [306]. A single operation scheme of NMEEF-SD algorithm can be observed in Algorithm 1.

**Algorithm 1** NMEEF-SD Algorithm

```
Generate the initial (parents) population
while (number of evaluations) is not reached do
  Generate offspring population
  Join the parent and offspring population in a combined one
  Generate all non-dominated fronts of the combined population
  if Pareto front (Front 1) evolves then
    Apply NSGA-II evolution
  else
    Apply Re-initialisation Based On Coverage
  end if
end while
Return the individuals of the Pareto front which reach a fuzzy confidence threshold
```

The main disadvantage of the classifiers in bioinformatics problems is the general lack of interpretability for the models obtained. These models that are obtained from classifiers have, as the main objective, accuracy. Additionally, in the majority of situations, these models are complex and they use a wide number of variables to describe different classes of the data set. In this way, it is very difficult for the experts to analyse and understand the behaviour of the different classes under investigation.

However, by the use of SD techniques simple and interpretable models can be extracted, where only some rules with a low number of variables representing each class are

obtained. The use of the NMEEF-SD algorithm in bioinformatics problems also facilitates the analysis by the experts because it uses linguistic labels in all the variables of the data set.

The search of interesting rules for the SD algorithms is another advantage provided by the NMEEF-SD algorithm. The use of $unusualness$ and $sensitivity$ as objective vectors in the multi-objective approach also provides a maximisation, not only for these measures but also for other measures in SD, as significance and confidence, because the unusualness and sensitivity have precision, novelty and coverage properties in their definitions.

Finally, the NMEEF-SD algorithm can be also studied like a classifier to see the behaviour of the algorithm to predict new data introduced. Thus, the experts can distinguish different groups of data with simple rules, which were obtained, and allow experts to have a clearer understanding of the given problem

### 8.2.4 Support Vector Machines

A support vector machine, (SVM) [153; 154] is a supervised statistical learning method that analyses data and recognises patterns for classification. The SVM takes a set of input data and predicts in which of two possible classes each given input belongs. A more detailed description of SVM can be found in Section 4.2.3. SVM is used in this analysis as it can produce accurate and robust classification results on a established theoretical basis even when input data are noisy or non-linearly separable [296; 297]. For this analysis the LIBSVM [156] tool was used to build a classification model. Furthermore, a 5-fold cross-validation was used in combination with F-score to find the optimum number of useful features that can be used to predict Influenza A neuraminidase subtypes without sacrificing any accuracy. In addition, grid search was used which is provided by LIBSVM to find the optimal parameters for the predictive model.

### 8.2.5 Feature Selection Using F-score

Feature selection [309; 310] is the technique of selecting relevant features for building robust classification models. Furthermore, feature selection is a particularly important step in analysing the data from many experimental techniques as they often include a large number of variables but low number of samples. By removing redundant features from the

data, feature selection can improve the performance of classification techniques like SVM in the following ways:

- Reduce data dimensionality.
- Improve the generalisation capability of the classification model.
- Speed up learning process.
- Improve model interpretability.

F-score [311] is one of the simplest but effective techniques, which measures the separation of two sets of real numbers. By using vectors $x_k$ , $k = 1, ..., m$, as input data, the number of positive and negative samples are $n_+$ and $n_-$ , respectively. F-score can be defined as follows for the $ith$ feature:

$$F(i) \equiv \frac{(\overline{x}_i^{(+)} - \overline{x}_i)^2 + (\overline{x}_i^{(-)} - \overline{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n+}(x_{k,i}^{(+)} - \overline{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n-}(x_{k,i}^{(-)} - \overline{x}_i^{(-)})^2} \tag{8.5}$$

where $\overline{x}_i$, $\overline{x}_i^{(+)}$, $\overline{x}_i^{(-)}$ are the averages of the $ith$ feature of the positive and negative data sets. The $ith$ feature of the $kth$ positive instance is $x_{k,i}^{(+)}$, and $x_{k,i}^{(-)}$ is the $ith$ feature of the $kth$ negative instance. In Equation 8.5 the numerator reveals the separation between the positive and negative sets, while the denominator points to the one within each of the two sets. The higher the F-score is, the higher the probability the feature under investigation is more separable using the given classes. Therefore, F-score is used as a feature selection criterion with SVM.

## 8.3 Case Study - Influenza A Neuraminidase Protein Sequence

### 8.3.1 Protein Sequences

Influenza A virus belongs to the orthomyxoviridae family of viruses and can affect mainly birds and some mammals. The Influenza A virus genome consists of eight single genes; the hemagglutinin (HA) gene, the neuraminidase (NA) gene, the nucleoprotein (NP) gene, the

matrix proteins (M) gene, the non-structural proteins (NS) gene and three RNA polymerase (PA, PB1, PB2) genes. Rarely do human pandemic outbreaks arise when the influenza A virus is transmitted from wild birds to domestic poultry. During the twentieth century, three major influenza pandemics were recorded, which were caused by H1N1, H2N2, and H3N2 viruses. In addition, the H5N1 virus is considered as a current pandemic thread. For this analysis, as Table 8.1 shows, four different subtypes of Influenza A virus Neuraminidase (NA) gene were used, as it is the target for current antiviral drugs, called neuraminidase inhibitors [262]. The complete list of protein sequences used in this study can be found in Appendix D.

Table 8.1: Influenza A Virus Neuraminidase Proteins

| Subtype | No of Sequences | Period |
|---|---|---|
| H1N1 | 200 | 2009 |
| H2N2 | 76 | 1957-1968 |
| H3N2 | 200 | 1968-2000 |
| H5N1 | 70 | 2005-2009 |

For influenza A subtypes 200 H1N1 NA proteins from 2009, 76 H2N2 NA proteins from the period 1957-1968, 200 H3N2 NA proteins from the period 1968-2000 and 70 H5N1 NA proteins from the period 2005-2009 were collected from the Influenza Virus Resource data set [263]. The relationship of influenza subtypes in respect of NA gene is shown in the following:

- H1N1 from 2009 is the result of reassortment between Eurasian H1N1 influenza A swine virus and H1N2 swine virus [264]. H1N1 retains the NA gene from Eurasian H1N1 influenza A swine virus.
- H2N2 from the period 1957-1968 is the result of reassortment between existing human H1N1 and avian H2N2 viruses [264]. H2N2 retains the NA gene from the avian H2N2 virus.
- H3N2 from the period 1968-2000 is the result of reassortment between circulating human H2N2 and avian H3 viruses [264]. H3N2 retains the NA gene from human H2N2 virus.
- H5N1 from the period 2005-2009 was created by combining various influenza A subtype viruses [260] where H5N1 retains the NA gene from avian H1N1 virus.

Percentage identity is a measurement used to determine the similarity between protein sequences. By using CLUSTALW, a freely available online tool [264], the pairwise percent identity of all the influenza A NA genes was calculated. Table 8.2 shows the average percent identity between all the classes.

Table 8.2: Average Pairwise Percent Identity

| | H1N1 | H2N2 | H3N2 | H5N1 |
|---|---|---|---|---|
| H1N1 | 93% | - | - | - |
| H2N2 | 42% | 96% | - | - |
| H3N2 | 40% | 86% | 94% | - |
| H5N1 | 83% | 43% | 41% | 96% |

As Table 8.2 shows, the percent identity within each individual influenza subtype class is very high with 93%, 96%, 94% and 96% for H1N1 NA, H2N2 NA, H3N2 NA and H5N1 NA influenza A subtypes In contrast to the individual class, percent identity from different classes may vary significantly, with high average percent identity of 83% between H1N1 and H5N1 and 86% between H2N2. Very low average percent identity was determined between H1N1 and H2N2 with 42%, H1N1 and H3N2 with 40%, H5N1 and H2N2 with 43%, and finally H5N1 and H3N2 with 41% average percent identity.

The main objective of this chapter is to find relationships between different proteins in the influenza A virus problem, with respect to different classes of these viruses. To follow this objective the problem is analysed and studied with a SD approach, the NMEEF-SD algorithm. In addition, a classification model is built based on SVM with features selections based on F-score. The problem has a high dimensionality, and it is composed for 256 features and 546 proteins sequences, where the proteins are distributed in the classes with 200 for class H1N1, 76 for H2N2, 200 for H3N2 and 70 for class H5N1.

All features used have a real domain, and therefore, they are continuous features, i.e. 256 continuous variables. The NMEEF-SD algorithm considers the continuous variables as linguistic fuzzy variables with fuzzy logic. More specifically, in this chapter, uniform partitions with triangular membership functions are used, as shown in Figure 8.2 for a variable with three linguistic labels $\{Low,\ Medium,\ High\}$. The parameters employed by NMEEF-SD are presented in Table 8.3.

Due to the non-deterministic nature of the NMEEF-SD, 5-fold cross validation is used to verify the results. In this way, the results shown are the averages of the results ob-

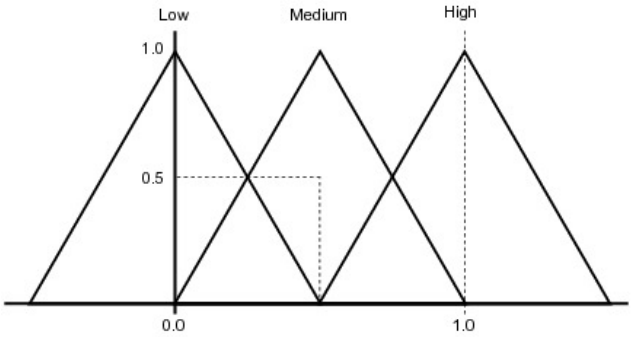



Figure 8.2: Example of Fuzzy Partition For A Continuous Variable With Three Labels

Table 8.3: Parameters For The NMEEF-SD Algorithm

| Parameters employed by NMEEF-SD algorithm | |
|---|---|
| Population size | 50 |
| Evaluations | 10000 |
| Crossover probability | 0.60 |
| Linguistic Labels | 3, 5, 7 and 9 |
| Mutation probability | 0.1 |
| Re-initialisation based on coverage | (50% of biased) |
| Minimum confidence | 0.2, 0.4 and 0.6 |
| Representation of the rule | Canonical |

tained for each fold. Therefore, the following average results in the experimental study in the tables can be observed: the number of linguistic labels employed ($LLs$), the minimum confidence threshold used ($Min_{Cnf}$), number of rules ($\sharp Rules$), number of variables ($\sharp Var$), significance ($SIGN$), unusualness ($UNUS$), sensitivity ($SENS$) and confidence ($CONF$).

### 8.3.2 Analysis of The Results Obtained For The NMEEF-SD Algorithm

Due to the complexity of the problem, it is necessary to use a diverse number of linguistic labels to find the best results. The algorithm can use different minimum confidence threshold to return rules that are fairly accurate. Therefore, for this study, 3, 5, 7 and 9 linguistic labels were used, with a minimum confidence threshold of 0.2, 0.4 and 0.6 for each one, as mentioned in Table 8.3. In this way, the NMEEF-SD algorithm is executed 25 times for

Table 8.4: Results Obtained For The NMEEF-SD Algorithm in The Experimental Study For The Influenza A Virus Problem

| $LLs$ | $Min_{Cnf}$ | $\sharp Rules$ | $\sharp Var$ | $SIGN$ | $UNUS$ | $SENS$ | $CONF$ |
|---|---|---|---|---|---|---|---|
| 3 | 0.2 | 4.60 | 2.79 | 57.945 | 0.153 | **1.000** | 0.747 |
| | 0.4 | 3.80 | 2.65 | 61.653 | 0.174 | **1.000** | 0.811 |
| | 0.6 | 2.60 | 2.73 | **66.967** | **0.190** | **1.000** | 0.849 |
| 5 | 0.2 | 3.40 | 2.13 | 47.628 | 0.125 | 0.990 | 0.708 |
| | 0.4 | 3.00 | 2.17 | 50.925 | 0.134 | 0.992 | 0.767 |
| | 0.6 | 2.20 | 2.10 | 54.155 | 0.148 | **1.000** | 0.807 |
| 7 | 0.2 | 3.00 | 2.28 | 47.832 | 0.110 | 0.963 | 0.760 |
| | 0.4 | 2.40 | 2.42 | 47.094 | 0.113 | 0.939 | 0.854 |
| | 0.6 | 1.60 | 2.37 | 52.038 | 0.127 | 0.938 | **0.911** |
| 9 | 0.2 | 1.60 | 2.00 | 40.257 | 0.092 | 0.952 | 0.585 |
| | 0.4 | 1.40 | 2.00 | 39.211 | 0.099 | 0.944 | 0.631 |
| | 0.6 | 0.60 | 0.80 | 17.191 | 0.048 | 0.378 | 0.394 |

each combination of parameters, and the average is shown for each row in Table 8.4. The best result for each quality measure is highlighted.

As Table 8.4 shows, the best results are obtained by using 3 linguistic labels. Table 8.5 shows the results obtained for each class in the experimental study for the influenza A virus problem with 3 linguistic labels. The results presented in this table are obtained using 5-fold cross validation for each class. For the subgroups obtained for a minimum confidence threshold of 0.6 the results indicate that there are not subgroups to describe all the classes. This is because the confidence threshold is too high in order to obtain good results in all the classes. Therefore, the results obtained in this configuration must be discarded. The best results obtained for the NMEEF-SD algorithm are obtained with 3 linguistic labels and minimum confidence of 0.2 and 0.4. Additionally, analyses related to the SD task for each class 3 linguistic labels and minimum confidence of 0.4 are presented below:

- The subgroups obtained for the H1N1 subtype have a high interpretability due to the low number of variables, where, in general, the subgroups obtained have less than 3 variables. The values for the significance and unusualness are the highest in respect to the values obtained from the remaining subtypes. The results obtained for this subtype are 69.9%, 19.9%, 100% and 84.5% for significance, unusualness, sensitivity and confidence respectively. Furthermore, the relation between sensitivity

Table 8.5: Results Obtained For The NMEEF-SD Algorithm For Each Class in The Experimental Study For The Influenza A Virus Problem With 3 Linguistic Labels

| $Min_{Cnf}$ | $Class$ | $\sharp Rules$ | $\sharp Var$ | $SIGN$ | $UNUS$ | $SENS$ | $CONF$ |
|---|---|---|---|---|---|---|---|
| 0.2 | H1N1 | 8.00 | 2.88 | 69.868 | 0.199 | 1.000 | 0.849 |
| | H2N2 | 5.00 | 3.20 | 44.562 | 0.101 | 1.000 | 0.543 |
| | H3N2 | 6.00 | 2.50 | 64.036 | 0.178 | 1.000 | 0.812 |
| | H5N1 | 5.00 | 2.60 | 44.907 | 0.102 | 1.000 | 0.717 |
| 0.4 | H1N1 | 8.00 | 2.88 | 69.868 | 0.199 | 1.000 | 0.849 |
| | H2N2 | 3.00 | 2.33 | 41.860 | 0.107 | 1.000 | 0.601 |
| | H3N2 | 5.00 | 2.40 | 67.831 | 0.193 | 1.000 | 0.835 |
| | H5N1 | 3.00 | 3.00 | 45.190 | 0.104 | 1.000 | 0.768 |
| 0.6 | H1N1 | 7.00 | 3.00 | 70.349 | 0.202 | 1.000 | 0.867 |
| | H2N2 | 0.00 | 0.00 | 0.000 | 0.000 | 0.000 | 0.000 |
| | H3N2 | 5.00 | 2.40 | 67.831 | 0.193 | 1.000 | 0.835 |
| | H5N1 | 1.00 | 3.00 | 44.923 | 0.101 | 1.000 | 0.867 |

and confidence is very good as the algorithm obtains subgroups where all the protein sequences for the class are covered.

- For the H2N2 subtype, the subgroups obtained are with the lower number of variables, and the interpretability is excellent. The values of significance and unusualness are also high considering that this class has a low number of protein sequences. The results obtained for this subtype are 41.9%, 10.7%, 100% and 60.1% for significance, unusualness, sensitivity and confidence, respectively.

- For the H3N2 subgroup the best subgroups are obtained along with the H1N1 subtype. The results obtained for this subtype are 67.8%, 19.3%, 100% and 83.5% for significance, unusualness, sensitivity and confidence, respectively.

- H5N1 is the class with the lowest number of protein sequences. Despite this fact, the results of sensitivity and confidence are very interesting as the subgroups cover protein sequences of the subtype with high confidence. The results obtained for this subtype are 45.2%, 10.4%, 100% and 76.8% for significance, unusualness, sensitivity and confidence, respectively.

Despite the objective of the NMEEF-SD algorithm being to obtain general and unusual rules to describe relations between the properties of the proteins with respect to different

Table 8.6: Predictive Results Obtained By NMEEF-SD Algorithm With 3 Linguistic Labels And A Minimum Confidence of 0.2 For The Influenza A Virus Problem

| Class | H1N1 | H2N2 | H3N2 | H5N1 |
|---|---|---|---|---|
| H1N1 | 0.965±0.055 | 0.000±0.000 | 0.000±0.000 | 0.025±0.055 |
| H2N2 | 0.000±0.000 | 0.799±0.413 | 0.201±0.413 | 0.000±0.000 |
| H3N2 | 0.000±0.000 | 0.132±0.086 | 0.868±0.086 | 0.000±0.000 |
| H5N1 | 0.217±0.424 | 0.000±0.000 | 0.000±0.000 | 0.729±0.424 |

Table 8.7: Predictive Results Obtained By NMEEF-SD Algorithm With 3 Linguistic Labels And A Minimum Confidence of 0.4 For The Influenza A Virus Problem

| Class | H1N1 | H2N2 | H3N2 | H5N1 |
|---|---|---|---|---|
| H1N1 | 0.975±0.056 | 0.000±0.000 | 0.000±0.000 | 0.025±0.056 |
| H2N2 | 0.107±0.174 | 0.413±0.537 | 0.481±0.457 | 0.000±0.000 |
| H3N2 | 0.000±0.025 | 0.043±0.112 | 0.925±0.106 | 0.000±0.000 |
| H5N1 | 0.629±0.458 | 0.000±0.000 | 0.029±0.064 | 0.343±0.480 |

type of virus, the algorithm has also a good behaviour as a classifier, as can be observed in the following analysis. Table 8.6 shows the confusion matrix for the accuracy of the model extracted by NMEEF-SD with three linguistic labels and a minimum confidence threshold of 0.2, and Table 8.7 shows the confusion matrix of the model with the same linguistic labels and 0.4 of minimum confidence. The results presented in both tables are the average of the 5-fold cross validation and the standard deviation for each one.

The total accuracy for the complete data set is $0.872 \pm 0.051$ for Table 8.6, and $0.797 \pm 0.069$ for Table 8.7. As can be observed in this study, for the model extracted by NMEEF-SD algorithm good precision in classifying new examples can be obtained. The best results are obtained with the minimum confidence threshold set to 0.2. In conclusion, NMEEF-SD shows good behaviour of the SD algorithms in discovering relationships in the analysis and classification of influenza A NA subtypes. An analysis for each subtype of virus can be observed below:

- For the H1N1 subtype class the average accuracy is 0.965±0.055, where the misclassified proteins were as H5N1.

- For the H2N2 subtype class the average accuracy is 0.799±0.413, where the misclassified proteins were as H3N2.

Table 8.8: Results of Subgroups Obtained For Each Class Using The NMEEF-SD Algorithm

| *Subgroup* | *SIGN* | *UNUS* | *SENS* | *CONF* |
|---|---|---|---|---|
| $IF\ (f44 = Low\ AND\ f97 = Low)\ THEN\ Cl = H1N1$ | 363.485 | 0.224 | 1.000 | 0.966 |
| $IF\ (f9 = Low\ AND\ f54 = Low\ f153 = Low\ AND\ f217 = Low)\ THEN\ Cl = H2N2$ | 227.960 | 0.105 | 1.000 | 0.600 |
| $IF\ (f8 = Low)\ THEN\ Cl = H3N2$ | 373.894 | 0.182 | 1.000 | 0.730 |
| $IF\ (f141 = Low\ AND\ f207 = Low\ AND\ f219 = Low)\ THEN\ Cl = H3N2$ | 309.357 | 0.196 | 0.995 | 0.966 |
| $IF\ (f115 = Low)\ THEN\ Cl = H5N1$ | 188.813 | 0.097 | 1.000 | 0.677 |

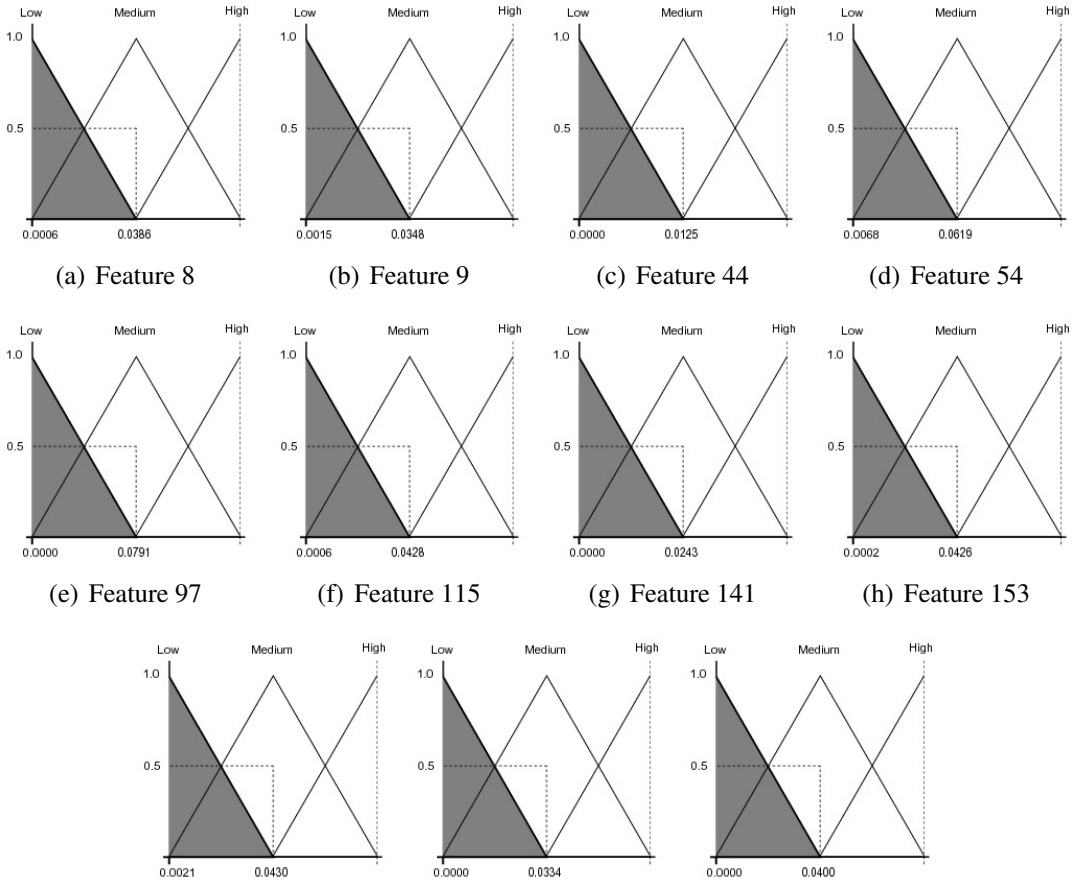
- For the H3N2 subtype class the average accuracy is 0.868±0.086, where the misclassified proteins were as H2N2.
- For the H5N1 subtype class the average accuracy is 0.729±0.424, where the misclassified proteins were as H1N1.

This analysis shows a strong correlation between the features extracted from the protein sequences using absolute spectrum and proteins percentage identity between classes as shown in Table 8.2. Only subtype classes that present high percentage identity between them as H1N1 with H5N1 subtype (83%) and H2N2 with H3N2 subtype (86%) were partially misclassification.

### 8.3.3 Fuzzy Subgroups Extracted By NMEEF-SD

Once determined that NMEEF-SD with 3 linguistic labels and a minimum confidence threshold of 0.2 obtains the best results for the influenza A virus problem, a new experiment was performed using the complete data set in order to analyse the subgroups obtained by NMEEF-SD. A detailed description of the NMEEF-SD algorithm can be found in [295]. Table 8.8 shows the subgroups obtained for the NMEEF-SD algorithm for each class with 3 linguistic labels and a minimum confidence of 0.2, where the variable $fx$ corresponds with the feature number $x$. The number of features that are included in the rules produced by the NMEEF-SD algorithm are limited to 11. In Figure 8.3, the linguistic representations for each variable found in the subgroups extracted by the algorithm can be observed. In addition, the table presents the results associated for each subgroup.

As it can be observed in Table 8.8, from the results of unusualness and significance, innovative information regarding the classification of NA subtypes introduced can be in-




(a) Feature 8 (b) Feature 9 (c) Feature 44 (d) Feature 54

(e) Feature 97 (f) Feature 115 (g) Feature 141 (h) Feature 153

(i) Feature 207 (j) Feature 217 (k) Feature 219

Figure 8.3: Linguistic Representations of The Continuous Feature of The Model Extracted By The NMEEF-SD Algorithm

troduced. Furthermore, the sensitivity obtained for the majority of the subgroups is the maximum level, and the confidence is very high with values higher than 0.6 and some very close to the maximum level. These good relations between the values of sensitivity and confidence present subgroups with high quality. In addition, the interpretability of these rules is excellent with subgroups, which in any case do not exceed of the four features.

Other methods that use signal processing techniques to extract biologically related features in order to characterise protein sequences, like Resonant Recognition Model in Hemagglutinin gene [42], and Complex Resonant Recognition for Neuraminidase gene [245] use informational spectrum analysis to retrieve these features. The extracted features are then used to characterise a specific class or compare it with another protein class based

on the common frequency peak [245]. By using NMEEF-SD algorithm simple rules, as Table 8.8 shows, can be extracted based on the features retrieve from the absolute spectrum. By using these features life-saving knowledge can be extracted and associated with the Influenza protein's sequences. These rules, created based on the features extracted, can then help develop therapy and/or vaccines. For the influenza problem, the rules created are based on 11 features extracted using DFT for all subtype classes. For example, for an unknown protein sequence, to be able to determine in which subgroup it belongs only 11 features of the absolute spectrum need to be considered and not the whole spectrum, where for the influenza A problem it consists of 256 variables. The importance of this outcome is that by using the NMEEF-SD algorithm biologically related positions are selected in the absolute spectrum for a problem, and a model is constructed with simple rules to characterise all protein classes. A detailed analysis for the subgroups extracted for each subtype of virus is shown below:

- For the H1N1 NA gene one rule with two features is obtained to describe this subtype, features 44 and 97. For feature 44, as Figure 8.3(c) shows, the interval for $Low$ is from 0.0 to 0.0125, and for feature 97, as Figure 8.3(e) shows, the interval for $Low$ is from 0.0006 to 0.0791. The SD results obtained for this subtype are very good with all the examples covered with 96.6% accuracy. In addition, the unusualness and significance values are very high, which shows an unusual behaviour of these properties to characterise this subtype of virus.

- For the H2N2 NA gene, one rule with four features is obtained to describe this subtype, features 9, 54, 153 and 217. For feature 9, as Figure 8.3(b) shows, the interval for $Low$ is from 0.0015 to 0.0348, for feature 54, as Figure 8.3(d) shows, the interval for $Low$ is from 0.0068 to 0.0619, for feature 153, as Figure 8.3(h) shows, the interval for $Low$ is from 0.0002 to 0.0426, and finally, for feature 217, as Figure 8.3(j) shows, the interval for $Low$ is from 0.0 to 0.0334. All the protein sequences for this subtype of virus are covered as the sensitivity is equal to 100.0%, with a 60.0% of success. The value of significance indicates a relative significance of this subtype of virus with respect to the others. Despite this subtype having a low number of instances the unusualness value is important. The interpretability of this subgroup is excellent with one subgroup represented with only one variable.

- For the H3N2 NA gene two rules are obtained to describe this subtype. For the

first rule, feature 8 is used and for the second rule features 141, 207 and 219 are used. For feature 8, as Figure 8.3(a) shows, the interval for $Low$ is from 0.006 to 0.0386, for feature 141, as Figure 8.3(g) shows, the interval for $Low$ is from 0.0 to 0.0243, for feature 207, as Figure 8.3(i) shows, the interval for $Low$ is from 0.0021 to 0.0430, and finally, for feature 219, as Figure 8.3(k) shows, the interval for $Low$ is from 0.0 to 0.04. To represent this subtype two different subgroups can be observed. For the first rule, a general subgroup is obtained with only one feature where all the examples for the subtypes are covered with 73.0% of proteins predicted correctly. For the second rule, a more specific subgroup is obtained with three features where 99.5% of proteins are covered with a 96.6% success. These relations between the sensitivity and confidence show a good rule to describe and classify new instances for this type of subtype.

- For the H5N1 NA gene, one feature, feature 115 of the absolute spectra was created to classify this subtype. For feature 115, as Figure 8.3(f) shows, the interval for $Low$ is from 0.0006 to 0.0428. This subgroup covers all the proteins of this subtype with a success of 67.7%, and the results for significance and unusualness are interesting, considering that this subtype has the lower number of instances of the data set. The interpretability of this subgroup is excellent, with one subgroup extracted with only one variable.

### 8.3.4 Classification Models Based on Support Vector Machines

By using SVM and F-score, a classification model was constructed for the Influenza A neuraminidase gene subtypes. By using F-score the goal is to select the most separable features extracted from influenza subtypes and create a predictive model without sacrificing any of the accuracy obtained by using all the features extracted. Figure 8.4 shows F-score value for all the features extracted from the protein sequences.

In order to build an accurate and generalised predictive model, 5-fold cross-validation was used. In combination with F-score, the minimum number of useful features that can be used to predict Influenza A neuraminidase subtypes without sacrificing any accuracy was set to 20. This number of features was discovered by manually eliminating features and repeating the analysis. Table 8.9 shows the best 20 features used with their calculated F-score. Figures 8.5 and 8.6 shows a plot of the best 20 features. In each subplot 8.5(a) -

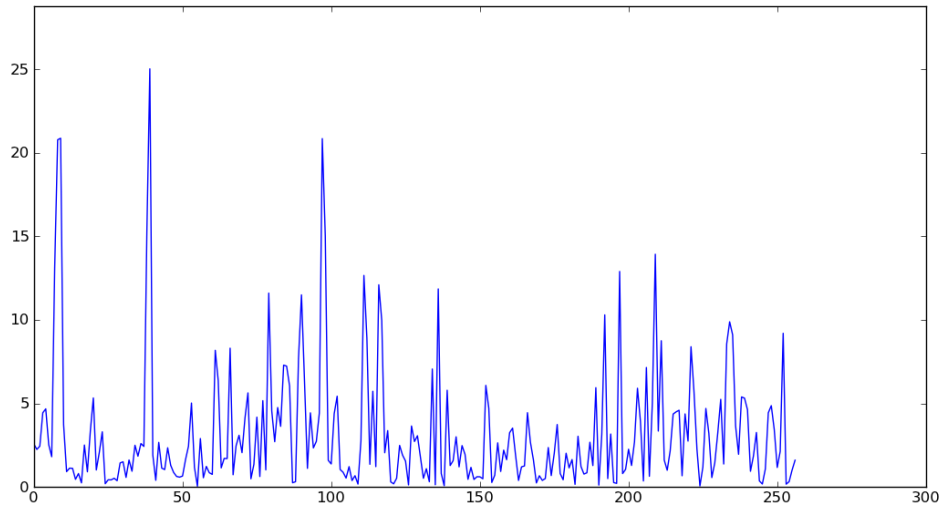



Figure 8.4: Feature Scores Based on F-score

8.5(h) a clear separation for most of the influenza subtypes can be observed.

The total accuracy for the complete data set is $0.983 \pm 0.006$. As the results show, good precision in classifying new protein sequences can be obtained. In addition, it can be observed in this study, the results obtained for the SVM based classification model as expected, have higher accuracy in comparison to the results obtained from the NMEEF-SD algorithm. An analysis for each influenza subtype can be observed below:

- For the H1N1 subtype class the average accuracy is $1.0 \pm 0.0$.
- For the H2N2 subtype class the average accuracy is $0.92 \pm 0.05$, where the misclassified proteins were as H3N2.
- For the H3N2 subtype class the average accuracy is $0.99 \pm 0.01$, where the misclassified proteins were as H2N2.
- For the H5N1 subtype class the average accuracy is $0.96 \pm 0.04$, where the misclassified proteins were as H1N1.

This analysis shows a strong correlation between the features extracted from the protein sequences and refined using F-score, with protein percentage identity between classes as

Table 8.9: Top 20 Features In Order of Importance Based on F-score

| | Feature | Score | | Feature | Score |
|---|---|---|---|---|---|
| 1 | 39 | 25.0333 | 11 | 116 | 12.1021 |
| 2 | 9 | 20.8762 | 12 | 136 | 11.8541 |
| 3 | 97 | 20.8564 | 13 | 79 | 11.6046 |
| 4 | 8 | 20.7845 | 14 | 90 | 11.4991 |
| 5 | 38 | 15.2451 | 15 | 192 | 10.3009 |
| 6 | 98 | 15.0310 | 16 | 117 | 9.9620 |
| 7 | 209 | 13.9315 | 17 | 234 | 9.8864 |
| 8 | 7 | 13.1166 | 18 | 252 | 9.1989 |
| 9 | 197 | 12.9016 | 19 | 236 | 9.1070 |
| 10 | 111 | 12.6627 | 20 | 113 | 8.9069 |

Table 8.10: Classification Results For SVM Predictive Model

| Class | H1N1 | H2N2 | H3N2 | H5N1 |
|---|---|---|---|---|
| H1N1 | **1.00 ± 0.00** | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 |
| H2N2 | 0.00 ± 0.00 | **0.92 ± 0.05** | 0.08 ± 0.05 | 0.00 ± 0.00 |
| H3N2 | 0.00 ± 0.00 | 0.01 ± 0.01 | **0.99 ± 0.01** | 0.00 ± 0.00 |
| H5N1 | 0.04 ± 0.04 | 0.00 ± 0.00 | 0.00 ± 0.00 | **0.96 ± 0.04** |

shown in Table 8.2. As in the case with the analysis with the NMEEF-SD algorithm, the most challenging part is to separate subtypes that present high percentage identity between them. As the bibliography indicates [260; 264], there is a clear biological connection between these influenza subtypes.

## 8.4 Conclusions

In this chapter, the influenza A virus characterisation problem is tackled through a SD algorithm, which can provide ancillary knowledge to the experts. The main objective of the case study was to find interpretable knowledge within the influenza A virus problem to describe the relationships between subtypes of this virus. For this purpose, one of the most representative SD algorithms was applied, the NMEEF-SD algorithm. NMEEF-SD is based on an EFS which is suitable for extracting rules with few features, i.e. interpretable, in order to facilitate the comprehensibility of the subgroups by the experts.

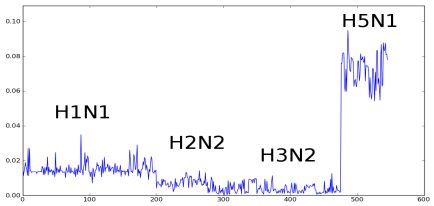
NMEEF-SD obtains representative subgroups for each subtype of virus of the problem.

These subgroups show an unusual and significant behaviour, and they also represent the total examples for each class with good confidence. The NMEEF-SD algorithm obtains a good level of precision when classifying new proteins to include in the data set. Furthermore, the rules extracted in order to classify new examples are interpretable because the algorithm employs linguistic labels to represent the continuous features, and because the number of features for each subgroup is very low.

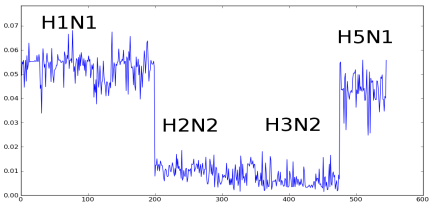
This case study offers the community a new point of view in the analysis of the influenza A virus with a novelty technique characterised by its interpretability, which obtains simple rules to represent different subtypes of virus. In this way, the model can classify an unknown protein sequence in a subtype of virus with only 11 features by using the NMEEF-SD and 20 features using SVM of the absolute spectrum instead of the whole spectrum, which consists of 256 features.

Additionally, in this chapter, an SVM classification model was built and tested with very high predictive accuracy. Furthermore, it has been demonstrated that by using signal processing techniques, in this case DFT, useful features can be extracted from protein sequences. By using F-score to refine these features and select the most appropriate, with the combination of SVM, an excellent classifier can be obtained for protein sequences.

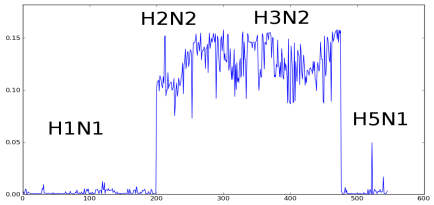
Finally in this chapter, the influenza A problem EIIP amino acid index (Table 2.3) is used to determine the absolute spectrum for each protein sequence. However, in the literature, 611 amino acid indices [312] exist to represent different biological features, and they can be used to construct different models in future works. Further information regarding the amino acid indices can be found in Chapter 3.



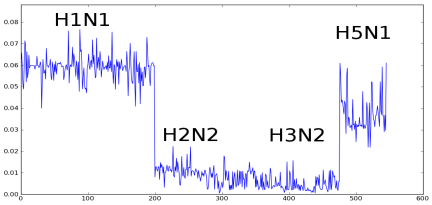

(a) Feature 39



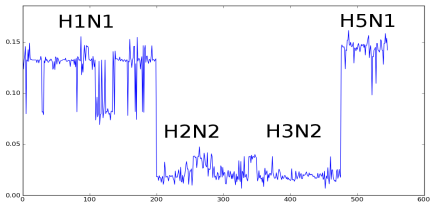

(b) Feature 9



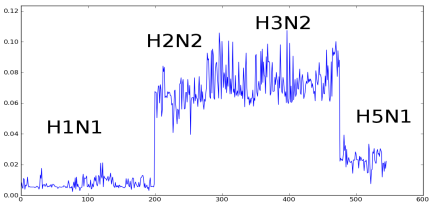

(c) Feature 97



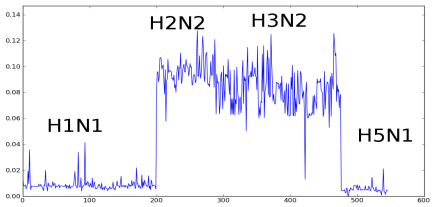

(d) Feature 8



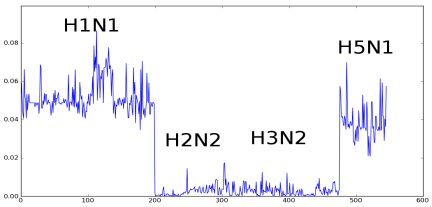

(e) Feature 38



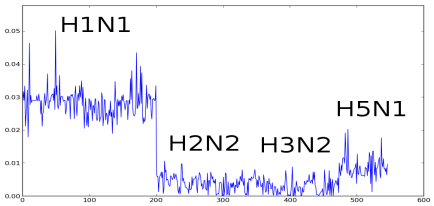

(f) Feature 98



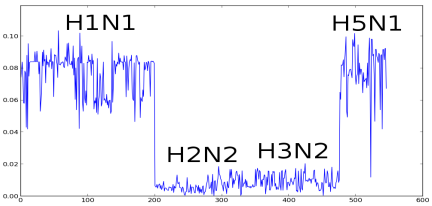

(g) Feature 209




(h) Feature 7

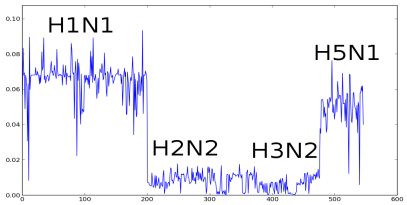



(i) Feature 197

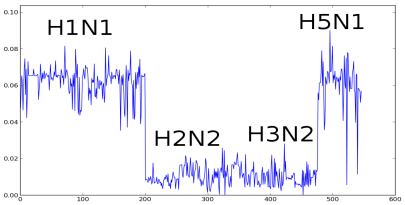



(j) Feature 111

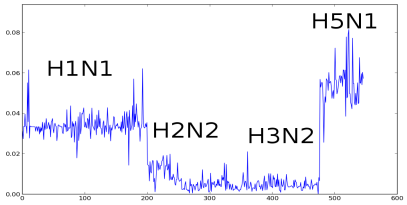
Figure 8.5: Top 20 Features In Order of Importance Based on F-score (1-10)



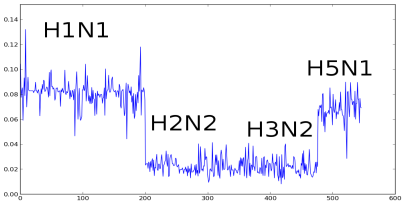

(a) Feature 116



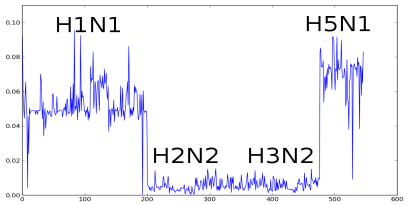

(b) Feature 136



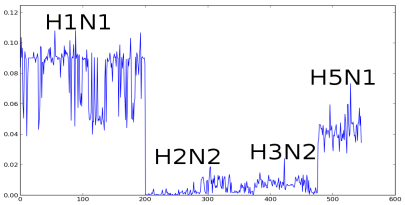

(c) Feature 79



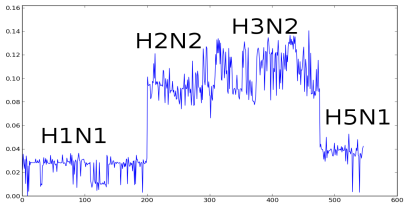

(d) Feature 90



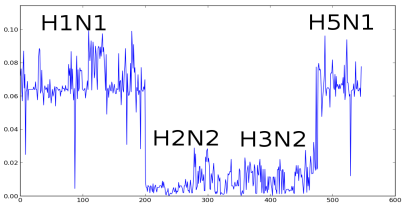

(e) Feature 192



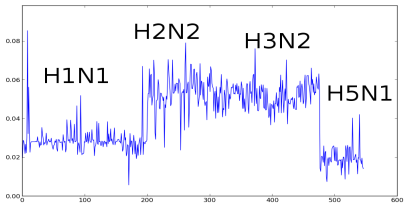

(f) Feature 117



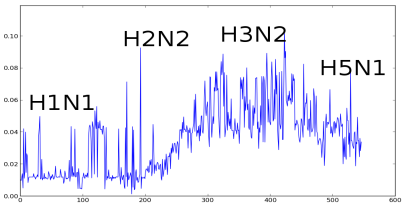

(g) Feature 234




(h) Feature 252




(i) Feature 236




(j) Feature 113

Figure 8.6: Top 20 Features In Order of Importance Based on F-score (10-20)

# Chapter 9

# Conclusions and Future Work

## 9.1 Research Summary

Protein sequencing has produced overwhelming amounts of protein sequences especially in the last decade [52; 53; 54; 55]. Nevertheless, the majority of the proteins' functional and structural classes sequenced are still unknown, and experimental methods currently used to determine these properties are very expensive, laborious and time consuming [15; 16]. Consequently, automated computational methods are urgently required to accurately and reliably predict functional and structural classes of the proteins. Several bioinformatics methods have been developed to determine such properties of proteins directly from their sequence information [56]. These involve signal processing methods, which are discussed and analysed in the literature review and have recently become popular in the bioinformatics area. They have been investigated for the analysis of DNA and protein sequences, and are shown to be useful and generally help better characterise the sequences [41; 57].

The outcome of this study is a series of bioinformatics systems that considers signal processing techniques for the analysis of protein sequences as a signal, and hence are expected to be capable of better characterising the proteins. This is also further improved by fusing multiple protein characteristics and pattern recognition methods. This research study, brought a novel concept to characterising the proteins and helped to predict structural and functional classes of the proteins from primary sequence information with greater reliability and accuracy. In addition, this is the most comprehensive and consistent study that considers the use of amino acid indices. It is expected that this work will provide guidelines to other areas of study in proteomics the main aim of which is automated prediction

of a protein's properties by using its primary structure.

## 9.2 Contribution to Knowledge

This thesis makes the following main original contributions:

- Amino acid indices have various applications and can represent diverse features of the protein sequences and amino acids. As the majority of indices included in the database have similar features, Chapter 3 proposes a set of computationally derived indices that summarise and better represent the original group of the indices. For this analysis, the hierarchical clustering and principal component analysis (PCA) were used. By using the hierarchical clustering methods the amino acid indices with similar features are clustered together. The next step is to use PCA on these clusters to computationally derive an amino acid index that should be able to represent the original amino acid indices included in the cluster. In the AAID database, 611 amino acid indices exist. By using the computational generated amino acid indices the search space can be reduced without losing any valuable information. The search space can be reduced to 542, 478, 521, 395 and 544 amino acid indices, by using single (1.0), single (0.65), complete (1.0), complete (0.65) and average (1.0 and 0.4) linkage, respectively. Additionally, in Chapter 3, the largest database of amino acid indices is developed, which contains all the latest published amino acid indices and presented in the Amino Acid Index Database (AAID) web-server (http://cisaps.com/indices).

- In Chapter 4, a study is carried out in order to find a unique and universal set of best discriminating amino acid indices for the characterisation of allergenic proteins. This analysis extracts features directly from protein sequences by using Discrete Fourier Transform (DFT) to build a classification model based on Support Vector Machines (SVM), for allergenic proteins. In summary, all the amino acid indices identified from the study characterises different aspects of hydrophobicity of allergenic proteins (Table 4.2).

  An extensive analysis was performed by using the Optimised relative partition energies - method C [144] amino acid index which is found to be the best amino acid index for characterising the allergen protein sequences. The classification model developed performs better in comparison to the AllerHunter classification model, which is

considered to be the latest and most precise available online tool for classification of Allergenic protein sequences [134]. As the results show, AllerHunter is more biased to non-allergen protein sequences as sensitivity, specificity and Matthews correlation coefficient (MCC) values show. This method of classification of allergenic and non-allergenic protein sequence is more reliable, and it is better for creating a generalised classification model.

- In chapter 5, a new method was proposed for performing protein multiple sequence alignments. For this method, Discrete Fourier Transform was used to construct a new distance matrix in combination with the multiple amino acid indices that were used to encode protein sequences into numerical sequences. The distance matrix is important as it can be used to construct a dendrogram that will act as a guide for Multiple Sequence Alignment (MSA) in which the global alignment is estimated by a series of pairwise alignments. The amino acid indices selected for this analysis are based on general and widely accepted features of the amino acids. Additionally, these indices are used to construct a similarity matrix. In this chapter, a new type of substitution matrix is proposed where the physicochemical similarities between any given amino acids is calculated. These similarities were calculated based on the 25 amino acids indices selected, where each one represents a unique biological protein feature.

  Additionally, in this Chapter, a case study was presented by using 32 CD4 protein sequences extracted from the UniProt database. The results show that the proposed method yields a more reliable alignment in comparison to the state-of-art MSA programs, like CLUSTALW, MAFFT, and T-COFFEE. Furthermore, the proposed MSA method is not biased to specific groups of protein sequences [236] as the values for the similarity matrix are calculated from the amino acid indices, and not from the protein sequences. Finally, for the proposed MSA, the same similarity matrix can be considered regarding the protein sequence's homology to be aligned or the mutation rate presented. A correlation to the physical characterisations of the amino acids the similarity matrix derived from can be achieved, while different similarity matrices can be generated by considering distinctive amino acids physical characterisations, that each amino acid index represents.

- In chapter 6, Complex Informational Spectrum Analysis (CISA) is developed and

presented. As the results show, when protein classes present similarities or differences according to the Common Frequency Peak (CFP) in specific amino acid indices, then it is probable that these classes are related to the protein feature that the specific amino acid represents. Furthermore, the use of only the absolute spectrum in the analysis of protein sequences using the informational spectrum analysis is proven to be insufficient, as biologically related features within the analysis of influenza A subtypes appear individually either in the real or the imaginary spectrum. Some of the applications of Informational Spectrum Analysis (ISA) and Resonant Recognition Model (RRM) that are already applied in the literature and CISA will also be applicable, and will be able to contribute additional information to the development of new drugs [247], identification of important protein sequences' domains and investigation of protein sequence interactions.

In this Chapter, a web-based server is also developed and presented, named CISAPS, which provides CISA for protein sequences. This web-based server enables researchers with little knowledge of signal processing methods to apply CISA to their work. Furthermore, CISAPS uses a collection of 611 unique amino acid indices, each one representing a different property, to perform the analysis. Moreover, in this chapter, various technical issues such as signal length and windowing that may affect the analysis are also addressed.

- Upon identification of a new protein, it is important to single out amino acid responsible for the structural and functional classification of the protein, as well as the amino acids contributing to the protein's specific biological characterisation. In Chapter 7, a novel approach is presented to identify and quantify the relationship between individual amino acids and the protein. Two methods are presented in this Chapter; the first takes into consideration the frequency peak, CFP, which it calculated from the RRM, and the second considers the entire absolute spectrum. Applicability and robustness of the methods are shown on a case study where five different protein families of the influenza A virus NA genes, which includes H1N1, H1N2, H2N2, H3N2 and H5N1 NA genes, are studied.

- In chapter 8, the influenza A virus problem is tackled through a Subgroup Discovery (SD) algorithm, which can provide ancillary knowledge to the experts. The main objective of the case study was to derive interpretable knowledge for the influenza

A virus problem and to consequently better describe the relationships between subtypes of this virus. For this purpose, one of the most representative SD algorithms was applied, namely the Non-dominated Multiobjective Evolutionary Algorithm for Extracting Fuzzy Rules in Subgroup Discovery (NMEEF-SD) algorithm. NMEEF-SD is based on Evolutionary Fuzzy System (EFS), which is suitable for extracting rules with small number of features, in order to facilitate the comprehensibility of the subgroups by the experts. The NMEEF-SD algorithm obtains a good level of precision when classifying new proteins to include in the data set. Furthermore, the rules extracted in order to classify new examples are interpretable because the algorithm employs linguistic labels to represent the continuous features, and the number of features for each subgroup is very low.

Additionally, in this chapter, an SVM classification model was built and tested with very high predictive accuracy. Furthermore, it has been demonstrated that by using signal processing techniques, in this case DFT, useful features can be extracted from protein sequences. By using F-score to refine these features and select the most appropriate, with the combination of SVM, an excellent classifier can be obtained for protein sequences. This case study offers the community a new point of view in the analysis of the influenza A virus with a novelty technique characterised by its interpretability, which obtains simple rules to represent different subtypes of virus. In this way, the model can classify an unknown protein sequence in a subtype of virus with only 11 features by using the NMEEF-SD and 20 features using SVM of the absolute spectrum instead of the whole spectrum, which consists of 256 features.

## 9.3 Future Work

So far, a series of bioinformatics systems that considers signal-processing techniques for analysis of the protein sequence have been discussed. In this section, the possible directions for future research are going to be discussed.

- As presented in Chapter 3, 611 amino acid indices exist to represent distinctive biological features, and they can be used in the development of different models in future works. In various methods developed and presented in this thesis, such as Chapter 5, for performing multiple protein sequence alignments, Chapter 7 for investigat-

ing the effects of an individual amino acid on protein function based on Discrete Fourier Transform or Chapter 8 for subgroup discovery and classification of protein sequences, a limited set of the available amino acid indices was used. Further investigation regarding the effects of the complete set of amino acid indices should be carried out.

- In Chapter 3 novel amino acid indices were produced and presented by using the hierarchical cluster analysis and principal component analysis. Further evaluation needs to be contacted regarding the quality of these generated amino acid indices. Finally the effect of these generated amino acid indices on existing methods needs to be investigated.

- Further research regarding the formulation of the gap penalties in multiple sequence alignments is required. In the analysis presented in chapter 5 gap penalties are assigned to the value $4*(min(similarityscore))$. In future works, as the gap penalties are the only variables in the proposed approach that can be taken into consideration for the homological similarity to the protein sequences to be aligned, a new adaptive model needs to be constructed that considers the protein sequences' homology.

  Additionally, for the alignment of protein sequences, 25 amino acid indices were used to construct the similarity matrix. In Chapter 3, 611 amino acid indices are described that represent a specific protein feature. In future works, the effect of individual amino acid index to the alignment of protein sequences needs to be studied.

- The Complex Resonant Recognition Model [245] can generate the real and imaginary spectrum in addition to the absolute spectrum was developed. As indicated in Chapter 6, these additional spectrum's can also be used to contribute supplemental information to the analysis of protein sequences. Additional work needs to be contacted regarding the contribution of these frequency spectrums to other methods where signal-processing methods are used to classify or characterise protein sequences.

- Finally, other signal processing techniques as presented in Chapter 2, can be utilise for characterisation and classification of protein sequences.

# References


[1] T. Attwood, A. Gisel, N. Eriksson, and E. Bongcam-Rudloff, "Concepts, historical milestones and the central place of bioinformatics in modern biology: A european perspective," *Bioinformatics-Trends and Methodologies*, 2011. 1

[2] P. Hogeweg, "The roots of bioinformatics in theoretical biology," *PLoS computational biology*, vol. 7, no. 3, p. e1002021, 2011. 1

[3] D. Benson, M. Boguski, D. Lipman, and J. Ostell, "Genbank," *Nucleic acids research*, vol. 25, no. 1, pp. 1–6, 1997. 1

[4] A. Bairoch, R. Apweiler, C. Wu, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, *et al.*, "The universal protein resource (uniprot)," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D154–D159, 2005. 1, 2, 44, 45, 50, 79

[5] C. O'Donovan, M. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, and R. Apweiler, "High-quality protein knowledge resource: Swiss-prot and trembl," *Briefings in Bioinformatics*, vol. 3, no. 3, pp. 275–284, 2002. 2

[6] D. Baker and A. Sali, "Protein structure prediction and structural genomics," *Science's STKE*, vol. 294, no. 5540, p. 93, 2001. 2

[7] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton, "Cath-a hierarchic classification of protein domain structures," *Structure*, vol. 5, no. 8, pp. 1093–1109, 1997. 2

[8] T. Blundell, B. Sibanda, R. Montalvão, S. Brewerton, V. Chelliah, C. Worth, N. Harmer, O. Davies, and D. Burke, "Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery,"

*Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 361, no. 1467, pp. 413–423, 2006. 2

[9] L. Ravichandran, A. Papandreou-Suppappola, A. Spanias, Z. Lacroix, and C. Legendre, "DNA sequence alignment using the matching pursuit decomposition," in *Genomic Signal Processing and Statistics, 2008. GENSiPS 2008. IEEE International Workshop on*, p. 14, 2008. 2

[10] J. Pevsner, *Pairwise Sequence Alignment*, pp. 46–98. John Wiley & Sons, Inc., 2009. 2

[11] J. Cuff and G. Barton, "Application of multiple sequence alignment profiles to improve protein secondary structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 40, no. 3, pp. 502–511, 2000. 2, 61, 62

[12] A. Delcher, D. Harmon, S. Kasif, O. White, and S. Salzberg, "Improved microbial gene identification with glimmer," *Nucleic acids research*, vol. 27, no. 23, pp. 4636–4641, 1999. 2

[13] A. Cambon-Thomsen, "The social and ethical issues of post-genomic human biobanks," *Nature Reviews Genetics*, vol. 5, no. 11, pp. 866–873, 2004. 2

[14] W. Wolfensberger, "Ethical issues in research with human subjects," *Science*, vol. 155, no. 3758, p. 47, 1967. 2

[15] K. D. Rao and M. Swamy, "Analysis of genomics and proteomics using DSP techniques," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 55, no. 1, p. 370378, 2008. 2, 7, 163

[16] D. W. Mount, *Bioinformatics: sequence and genome analysis*. CSHL press, 2004. 2, 61, 62, 64, 66, 163

[17] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine theme article*, vol. vol. 18, No 4, p. 820, 2001. 3, 15

[18] M. J. Zaki and C. Bystroff, *Protein Structure Prediction 2nd Edition*. Humana press, 2007. 3, 15

[19] J. Pevsner, *Bioinformatics and functional genomics*. Wiley Online Library, 2003. 3

[20] A. Streitwieser, C. Heathcock, and E. Kosower, *Introduction to organic chemistry*. Macmillan New York, 1992. 4, 5

[21] H. Stoker, *General, Organic, and Biological Chemistry*. Brooks/Cole Pub Co, 2012. 5

[22] A. Hughes, *Amino Acids, Peptides and Proteins in Organic Chemistry: Building Blocks, Catalysis and Coupling Chemistry*, vol. 3. Wiley-VCH, 2011. 5, 73, 80

[23] P. Ertl, "Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups," *Journal of chemical information and computer sciences*, vol. 43, no. 2, pp. 374–380, 2003. 5

[24] G. Fasman, *Practical handbook of biochemistry and molecular biology*. CRC, 1989. 5, 29, 73

[25] R. Grantham, "Amino acid difference formula to help explain protein evolution," *Science*, vol. 185, no. 4154, p. 862, 1974. 5, 29, 73

[26] O. Mayo and D. Brock, *The biochemical genetics of man*. Cambridge Univ Press, 1972. 5, 29, 73

[27] P. Sneath, "Relations between chemical structure and biological activity in peptides," *Journal of Theoretical Biology*, vol. 12, no. 2, pp. 157 – 195, 1966. 5

[28] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "Aaindex: amino acid index database, progress report 2008," *Nucleic acids research*, vol. 36, no. suppl 1, p. D202, 2008. 5, 29, 30, 31, 100

[29] J. Huang, S. Kawashima, and M. Kanehisa, "New amino acid indices based on residue network topology," *Genome Informatics*, vol. 18, pp. 152–161, 2007. 6, 29, 31, 52, 53, 56, 73, 100, 216

[30] S.S., Nanuwa, A. Dziurla, and H. Seker, "Weighted amino acid composition based on amino acid indices for prediction of protein structural classes," in *IEEE ITAB 2009*, November 2009. 6, 29

[31] H. Seker, "Novel weighted amino acid composition for prediction of protein structural classes within the context of multi-sensor data fusion approach," in *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on*, p. 16, 2008. 6, 16, 29, 31, 99, 216

[32] D. Sarda, G. Chua, K. Li, and A. Krishnan, "pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties," *BMC bioinformatics*, vol. 6, no. 1, p. 152, 2005. 6, 29

[33] M. B. N. H. L. C. Kazemian, M., "A new expertness index for assessment of secondary structure prediction engines," *Computational Biology and Chemistry*, vol. 31, no. 1, pp. 44–47, 2007. cited By (since 1996) 4. 6, 29

[34] L. Kurgan, W. Stach, and J. Ruan, "Novel scales based on hydrophobicity indices for secondary protein structure," *Journal of theoretical biology*, vol. 248, no. 2, pp. 354–366, 2007. 6, 29, 31, 216

[35] L. E. Zhao, G., "An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity," *Protein Science*, vol. 15, no. 8, pp. 1987–2001, 2006. cited By (since 1996) 26. 6, 29

[36] P. Sneath, "Relations between chemical structure and biological activity in peptides," *Journal of theoretical biology*, vol. 12, no. 2, pp. 157–195, 1966. 6, 30

[37] O. T. Nishikawa, K., "Prediction of the surface-interior diagram of globular proteins by an empirical method.," *International Journal of Peptide and Protein Research*, vol. 16, no. 1, pp. 19–32, 1980. cited By (since 1996) 33. 6, 30

[38] O. T. Nishikawa, K., "Radial locations of amino acid residues in a globular protein: Correlation with the sequence," *Journal of Biochemistry*, vol. 100, no. 4, pp. 1043–1047, 1986. cited By (since 1996) 46. 6, 30

[39] Z. X. F. Z.-P. Han, P., "Predicting disordered regions in proteins using the profiles of amino acid indices," *BMC Bioinformatics*, vol. 10, no. SUPPL. 1, 2009. cited By (since 1996) 4. 6, 30

[40] E. Pirogova and I. Cosic, "Bioactive peptide design using the resonant recognition model," *Nonlinear Biomed Phys.*, p. 17, 2007. 7, 16, 17, 99, 101, 112, 139, 141

[41] J. V. Lorenzo-Ginori, "Digital signal processing in the analysis of genomic sequences.," *Current Bioinformatics*, vol. 4(1), p. 2840, 2009. 7, 8, 99, 163

[42] V. Veljkovic, N. Veljkovic, C. Muller, S. Muller, S. Glisic, V. Perovic, and H. Kohler, "Characterization of conserved properties of hemagglutinin of h5n1 and human influenza viruses: possible consequences for therapy and infection control," *BMC Structural Biology*, vol. 9, no. 1, p. 21, 2009. 7, 99, 100, 107, 117, 124, 141, 155

[43] S. Qian and D. Chen, "Joint time-frequency analysis," *Signal Processing Magazine, IEEE*, vol. 16, no. 2, pp. 52–67, 1999. 7

[44] U. Zölzer, *Digital audio signal processing*. Wiley Online Library, 2008. 7

[45] J. Parker, *Algorithms for image processing and computer vision*. Wiley Publishing, 2010. 7

[46] J. Benesty, *Springer handbook of speech processing*. Springer Verlag, 2008. 7

[47] L. Rabiner and B. Juang, *Fundamentals of speech recognition*, vol. 103. Prentice hall, 1993. 7

[48] I. Glover and P. Grant, *Digital communications*. Pearson Prentice Hall, 2009. 7

[49] M. Akay, V. Marmarelis, R. Merletti, P. Parker, D. Westwick, and R. Kearney, "Nonlinear biomedical signal processing, volume 1, fuzzy logic, neural networks, and new algorithms," 1996. 7

[50] Q. Fang, "Prediction of the active sites of the fibroblast growth factors using continuous wavelet transform and the resonant recognition model," *Proc. Inaugural Conf. Victorian Chapter IEEE EMBS*, pp. pp. 211–214, 1999. 7, 25, 26

[51] P. Goupillaud, A. Grossmann, and J. Morlet, "Cycle-octave and related transforms in seismic signal analysis," *Geoexploration*, vol. 23, no. 1, pp. 85–102, 1984. 7, 25

[52] L. Martens, H. Hermjakob, P. Jones, M. Adamski, C. Taylor, K. Gevaert, J. Vandekerckhove, R. Apweiler, *et al.*, "Pride: the proteomics identifications database," *Proteomics*, vol. 5, no. 13, pp. 3537–3545, 2005. 7, 163

[53] J. A. Vizcaíno, R. Côté, F. Reisinger, H. Barsnes, J. M. Foster, J. Rameseder, H. Hermjakob, and L. Martens, "The proteomics identifications database: 2010 update," *Nucleic acids research*, vol. 38, no. suppl 1, pp. D736–D742, 2010. 7, 163

[54] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, *et al.*, "The universal protein resource (uniprot)," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D154–D159, 2005. 7, 163

[55] R. Apweiler, M. J. Martin, C. O'Donovan, M. Magrane, Y. Alam-Faruque, E. Alpi, R. Antunes, J. Arganiska, E. B. Casanova, B. Bely, *et al.*, "Update on activities at the universal protein resource (uniprot) in 2013," *NUCLEIC ACIDS RESEARCH*, vol. 41, no. D 1, pp. D43–D47, 2013. 7, 163

[56] L. Z.R., "Profeat: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence.," *Nucleic Acids Res.*, vol. 34, pp. 32–37, 2006. 7, 14, 123, 163

[57] V. P., "The role of signal-processing concepts in genomics and proteomics.," *Journal of the Franklin Institute*, vol. 341(1), p. 2004, 2004. 8, 163

[58] H. Rao, F. Zhu, G. Yang, Z. Li, and Y. Chen, "Update of profeat: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence," *Nucleic acids research*, vol. 39, no. suppl 2, p. W385, 2011. 11, 12, 14

[59] L. Han, C. Cai, Z. Ji, Z. Cao, J. Cui, and Y. Chen, "Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach," *Nucleic acids research*, vol. 32, no. 21, pp. 6437–6444, 2004. 11

[60] A. Castillo, H. Gutierrez, J. Monzon, and A. Urrutia, "Protein amino acid composition: A genomic signature of encephalization in mammals," *PLoS ONE*, vol. 6, no. 11, 2011. 12, 29

[61] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function, and Genetics*, vol. 43(3), pp. 246–255, Proteins: Structure, Function, and Genetics. 12, 13, 14, 29

[62] P. Petrilli, "Classification of protein sequences by their dipeptide composition," *Computer applications in the biosciences: CABIOS*, vol. 9, no. 2, p. 205, 1993. 12, 99

[63] R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, vol. 79. Wiley-Vch, 2008. 12

[64] K. Chou, "Prediction of protein subcellular locations by incorporating quasi-sequence-order effect," *Biochemical and Biophysical Research Communications*, vol. 278, no. 2, pp. 477–483, 2000. 12

[65] L. Han, J. Cui, H. Lin, Z. Ji, Z. Cao, Y. Li, and Y. Chen, "Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity," *Proteomics*, vol. 6, no. 14, pp. 4023–4037, 2006. 12

[66] K. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, p. 10, 2005. 12

[67] C. Zhang, K. Chou, and G. Maggiora, "Predicting protein structural classes from amino acid composition: application of fuzzy clustering," *Protein engineering*, vol. 8, no. 5, pp. 425–435, 1995. 12

[68] M. Bhasin and G. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *Journal of Biological Chemistry*, vol. 279, no. 22, pp. 23262–23266, 2004. 12

[69] S. Sahu and G. Panda, "A novel feature representation method based on chou's pseudo amino acid composition for protein structural class prediction," *Computational Biology and Chemistry*, vol. 34, no. 5, pp. 320–327, 2010. 14

[70] Z. Li, X. Zhou, Z. Dai, and X. Zou, "Prediction of protein structural classes by chous pseudo amino acid composition: approached using continuous wavelet transform

and principal component analysis," *Amino acids*, vol. 37, no. 2, pp. 415–425, 2009. 14, 21

[71] J. Qiu, J. Huang, S. Shi, and R. Liang, "Using the concept of chous pseudo amino acid composition to predict enzyme family classes: An approach with support vector machine based on discrete wavelet transform," *Protein and Peptide Letters*, vol. 17, no. 6, pp. 715–722, 2010. 14

[72] X. Jiang, R. Wei, Y. Zhao, and T. Zhang, "Using chous pseudo amino acid composition based on approximate entropy and an ensemble of adaboost classifiers to predict protein subnuclear location," *Amino Acids*, vol. 34, no. 4, pp. 669–675, 2008. 14

[73] J. Bock and D. Gough, "Predicting protein–protein interactions from primary structure," *Bioinformatics*, vol. 17, no. 5, p. 455, 2001. 14, 99

[74] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. Krogan, S. Chung, A. Emili, M. Snyder, J. Greenblatt, and M. Gerstein, "A bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, p. 449, 2003. 14

[75] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting subcellular localization of proteins based on their n-terminal amino acid sequence," *Journal of molecular biology*, vol. 300, no. 4, pp. 1005–1016, 2000. 14

[76] J. Shi, S. Zhang, Q. Pan, and G. Zhou, "Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution," *Amino Acids*, vol. 35, no. 2, pp. 321–327, 2008. 14

[77] H. Lin, "The modified mahalanobis discriminant for predicting outer membrane proteins by using chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 252, no. 2, pp. 350–356, 2008. 14

[78] Y. Diao, D. Ma, Z. Wen, J. Yin, J. Xiang, and M. Li, "Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and lempel-ziv complexity," *Amino Acids*, vol. 34, no. 1, pp. 111–117, 2008. 14

[79] G. Zhang, H. Li, J. Gao, and B. Fang, "Predicting lipase types by improved chous pseudo-amino acid composition," *Protein and Peptide Letters*, vol. 15, no. 10, pp. 1132–1137, 2008. 14, 123

[80] K. Chou and C. Zhang, "Predicting protein folding types by distance functions that make allowances for amino acid interactions.," *Journal of Biological Chemistry*, vol. 269, no. 35, pp. 22014–22020, 1994. 14

[81] J. Guo, N. Rao, G. Liu, Y. Yang, and G. Wang, "Predicting protein folding rates using the concept of chou's pseudo amino acid composition," *Journal of computational chemistry*, 2011. 14

[82] G. Zhou and Y. Cai, "Predicting protease types by hybridizing gene ontology and pseudo amino acid composition," *PROTEINS: Structure, Function, and Bioinformatics*, vol. 63, no. 3, pp. 681–684, 2006. 14

[83] Y. Fang, Y. Guo, Y. Feng, and M. Li, "Predicting dna-binding proteins: approached from chous pseudo amino acid composition and other specific sequence features," *Amino Acids*, vol. 34, no. 1, pp. 103–109, 2008. 14

[84] W. Landschulz, P. Johnson, and S. McKnight, "The leucine zipper: a hypothetical structure common to a new class of dna binding proteins," *Science*, vol. 240, no. 4860, p. 1759, 1988. 14

[85] C. Chen, L. Chen, X. Zou, and P. Cai, "Prediction of protein secondary structure content by using the concept of chous pseudo amino acid composition and support vector machine," *Protein and peptide letters*, vol. 16, no. 1, pp. 27–31, 2009. 14, 46

[86] U. Seiffert, B. Hammer, S. Kaski, and T. Villmann, "Neural networks and machine learning in bioinformatics - theory and applications," *ESANN 2006, 26-28 April 2006, Brugge, Belgium*, 2006. 15

[87] V. Veljkovic, I. Cosic, B. Dimitrijevic, and D. LalovicC, "Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?," *IEEE Transaction on Biomedical Engineering*, vol. 32, no. 5, pp. 337–341, 1985. 15, 17, 29, 73, 100, 101, 124, 141

[88] K. Gopalakrishnan, R. Zadeh, K. Najarian, and A. Darvish, "Computational analysis and classification of p53 mutants according to primary structure," in *2004 IEEE Computational Systems Bioinformatics Conference, Proceedings*, pp. 694–695, 2004. 15, 100, 141

[89] K. Gopalakrishnan and K. Najarian, "Prediction of protein function using signal processing of biochemical properties," in *Proceedings of the IEEE Computer Society Conference on Bioinformatics*, p. 536, IEEE Computer Society, 2003. 15, 100

[90] X. Liu, "A modified resonant recognition model to predict protein-protein interaction," *Frontiers of Biology in China*, vol. 2, p. 268 , 2007. 16

[91] E. Pirogova, "Analysis of amino acid parameters in the resonant recognition model," *Proceedings of the International Conference on Bioelectromagnetism*, p. 71, 1998. 16, 17, 99, 101, 112, 123, 139, 141

[92] I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules?Theory and applications.," *IEEE transactions on bio-medical engineering.*, vol. 41, p. 1101, 1994. 16, 17, 99, 100, 101, 112, 123, 139, 141

[93] J. Grassmann, M. Reczko, S. Suhai, and L. Edler, "Protein fold class prediction: new methods of statistical classification," in *Proceedings of the ISMB*, pp. 6–10, 1999. 16, 99

[94] I. S. Wei Zhang, *Computational and Statistical Approaches to Genomics*. Springer, 2005. 17, 113

[95] K. Gröchenig, *Foundations of time-frequency analysis*. Birkhauser, 2001. 21, 22

[96] H. H and K. S, "Prediction of hydrophobic cores of proteins using wavelet analysis.," *Genome Inform Ser*, no. 8, p. 6170, 1997. 21

[97] K. Li, P. Issac, and A. Krishnan, "Predicting allergenic proteins using wavelet transform," *Bioinformatics*, vol. 20, no. 16, p. 2572, 2004. 21, 44

[98] K. Chen, J. T. Huzil, H. Freedman, P. Ramachandran, A. Antoniou, J. A. Tuszynski, and L. Kurgan, "Identification of tubulin drug binding sites and prediction of relative

differences in binding affinities to tubulin isotypes using digital signal processing," *Journal of Molecular Graphics and Modelling*, vol. 27, pp. 497–505, Nov. 2008. 21

[99] S. Guo and Z. Yi-Sheng, "An integrative algorithm for predicting protein coding regions," in *IEEE Asia-Pacific Conference on Circuits and Systems, Proceedings, APCCAS*, pp. 438–441, 2008. 21

[100] Z. ning Wen, K. long Wang, M. long Li, F. sheng Nie, and Y. Yang, "Analyzing functional similarity of protein sequences with discrete wavelet transform," *Computational Biology and Chemistry*, vol. 29, no. 3, pp. 220–228, 2005. 24

[101] C. Torrence and G. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61–78, 1998. 26

[102] F. C. A. P. Eisenhaber, F., "Prediction of secondary structural content of proteins from their amino acid composition alone. ii. the paradox with secondary structural class," *Proteins: Structure, Function and Genetics*, vol. 25, no. 2, pp. 169–179, 1996. 29

[103] I. F. A. P.-F. C. Eisenhaber, F., "Prediction of secondary structural content of proteins from their amino acid composition alone. i. new analytic vector decomposition methods," *Proteins: Structure, Function and Genetics*, vol. 25, no. 2, pp. 157–168, 1996. 29

[104] S. Roy, D. Martinez, H. Platero, T. Lane, and M. Werner-Washburne, "Exploiting amino acid composition for predicting protein-protein interactions," *PLoS ONE*, vol. 4, p. e7813, 11 2009. 29

[105] L. X. R. E.-W. R. Hansen, J.C., "Intrinsic protein disorder, amino acid composition, and histone terminal domains," *Journal of Biological Chemistry*, vol. 281, no. 4, pp. 1853–1856, 2006. 29

[106] H. Nakashima and K. Nishikawa, "The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins," *FEBS letters*, vol. 303, no. 2-3, pp. 141–146, 1992. 29

[107] I. Saha, U. Maulik, S. Bandyopadhyay, and D. Plewczynski, "Fuzzy clustering of physicochemical and biochemical properties of amino acids," *Amino Acids*, pp. 1–12. 10.1007/s00726-011-1106-9. 30

[108] A. G., "Interpretable numerical descriptors of amino acid space," *Journal of Comput. Biology*, vol. 16(5), pp. 703–723, 2009. 30

[109] W. Atchley, J. Zhao, A. Fernandes, and T. Drüke, "Solving the protein sequence metric problem," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 18, p. 6395, 2005. 31, 75, 216

[110] M. Zviling, H. Leonov, and I. T. Arkin, "Genetic algorithm-based optimization of hydrophobicity tables," *BIOINFORMATICS -OXFORD-*, vol. 21, no. 11, p. 26512656, 2005. 31, 216

[111] L. Fernández, J. Caballero, J. Abreu, and M. Fernández, "Amino acid sequence autocorrelation vectors and bayesian-regularized genetic neural networks for modeling protein conformational stability: Gene v protein mutants," *Proteins: Structure, Function, and Bioinformatics*, vol. 67, no. 4, pp. 834–852, 2007. 31, 52, 53, 56, 73, 100, 112, 216, 217

[112] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. Wilkins, R. Appel, and A. Bairoch, "Protein identification and analysis tools on the expasy server," *The proteomics protocols handbook*, pp. 571–607, 2005. 31, 44, 52, 56, 73, 216, 217

[113] N. Asakawa, N. Sakiyama, R. Teshima, and S. Mitaku, "Characteristic amino acid distribution around segments unique to allergens," *Journal of biochemistry*, vol. 147, no. 1, p. 127, 2010. 31, 217

[114] M. Marx and R. Larsen, *Introduction to mathematical statistics and its applications*. Pearson/Prentice Hall, 2006. 31

[115] I. T. Jolliffe, "Principal component analysis," *New York: Springer-Verlag*, 1986. 33

[116] J. Psychol, "1. pearson k: On lines and planes of closest fit to systems of points in space.," *Phil Mag*, vol. 6, no. 2, pp. 559–572, 1901. 33

[117] A. B. Kay, "Overview of allergy and allergic diseases: with a view to the future," *British medical bulletin*, vol. 56, no. 4, pp. 843–864, 2000. 43

[118] O. Ivanciuc, T. Garcia, M. Torres, C. Schein, and W. Braun, "Characteristic motifs for families of allergenic proteins," *Molecular immunology*, vol. 46, no. 4, pp. 559–568, 2009. 43

[119] F. Shakib, A. Ghaemmaghami, and H. Sewell, "The molecular basis of allergenicity," *Trends in immunology*, vol. 29, no. 12, pp. 633–642, 2008. 43

[120] M. Wills-Karp, "Allergen-specific pattern recognition receptor pathways," *Current opinion in immunology*, 2010. 43

[121] D. Marsh, "Allergens and the genetics of allergy," *The antigens*, vol. 3, pp. 271–359, 1975. 43

[122] H. Breiteneder and C. Mills, "Structural bioinformatic approaches to understand cross-reactivity," *Molecular nutrition & food research*, vol. 50, no. 7, pp. 628–632, 2006. 43

[123] R. Aalberse *et al.*, "Assessment of allergen cross-reactivity," *Clin Mol Allergy*, vol. 5, no. 2, 2007. 43

[124] C. Schein, O. Ivanciuc, and W. Braun, "Bioinformatics approaches to classifying allergens and predicting cross-reactivity," *Immunology and allergy clinics of North America*, vol. 27, no. 1, pp. 1–27, 2007. 43

[125] C. Traidl-Hoffmann, T. Jakob, and H. Behrendt, "Determinants of allergenicity," *Journal of Allergy and Clinical Immunology*, vol. 123, no. 3, pp. 558–566, 2009. 43

[126] S. Eisenbarth, D. Piggott, J. Huleatt, I. Visintin, C. Herrick, and K. Bottomly, "Lipopolysaccharide-enhanced, toll-like receptor 4–dependent t helper cell type 2 responses to inhaled antigen," *The Journal of experimental medicine*, vol. 196, no. 12, p. 1645, 2002. 43

[127] V. Redecke, H. H
"acker, S. Datta, A. Fermin, P. Pitha, D. Broide, and E. Raz, "Cutting edge: activa-

tion of toll-like receptor 2 induces a th2 immune response and promotes experimental asthma," *The Journal of Immunology*, vol. 172, no. 5, p. 2739, 2004. 43

[128] H. Hammad, M. Chieppa, F. Perros, M. Willart, R. Germain, and B. Lambrecht, "House dust mite allergen induces asthma via tlr4 triggering of airway structural cells," *Nature medicine*, vol. 15, no. 4, p. 410, 2009. 43

[129] G. Deslée, A. Charbonnier, H. Hammad, G. Angyalosi, I. Tillie-Leblond, A. Mantovani, A. Tonnel, and J. Pestel, "Involvement of the mannose receptor in the uptake of der p 1, a major mite allergen, by human dendritic cells," *The Journal of allergy and clinical immunology*, vol. 110, no. 5, pp. 763–770, 2002. 43

[130] P. Royer, M. Emara, C. Yang, A. Al-Ghouleh, P. Tighe, N. Jones, H. Sewell, F. Shakib, L. Martinez-Pomares, and A. Ghaemmaghami, "The mannose receptor mediates the uptake of diverse native allergens by dendritic cells and determines allergen-induced t cell polarization through modulation of ido activity," *The Journal of Immunology*, vol. 185, no. 3, p. 1522, 2010. 43

[131] M. Li, A. Gustchina, J. Glesner, S. W
"unschmann, L. Vailes, M. Chapman, A. Pomés, and A. Wlodawer, "Carbohydrates contribute to the interactions between cockroach allergen bla g 2 and a monoclonal antibody," *The Journal of Immunology*, vol. 186, no. 1, p. 333, 2011. 43

[132] R. Furmonaviciene, B. Sutton, C. Laughton, H. Sewell, and F. Shakib, "The definition of allergen-specific molecular surface features: new insights into allergenicity," *Bioinformatics*, vol. 21, pp. 4201–4204, 2005. 43, 44, 52, 54

[133] F. Glaser, T. Pupko, I. Paz, R. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal, "Consurf: identification of functional regions in proteins by surface-mapping of phylogenetic information," *Bioinformatics*, vol. 19, no. 1, p. 163, 2003. 44

[134] H. Muh, J. Tong, and M. Tammi, "Allerhunter: a svm-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins," *PloS one*, vol. 4, no. 6, p. e5861, 2009. 44, 45, 50, 51, 52, 165

[135] Z. Zhang, J. Koh, G. Zhang, K. Choo, M. Tammi, and J. Tong, "Allertool: a web server for predicting allergenicity and allergic cross-reactivity in proteins," *Bioinformatics*, vol. 23, no. 4, p. 504, 2007. 44

[136] Å. Björklund, D. Soeria-Atmadja, A. Zorzet, U. Hammerling, and M. Gustafsson, "Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins," *Bioinformatics*, vol. 21, no. 1, p. 39, 2005. 44

[137] J. Cui, L. Han, H. Li, C. Ung, Z. Tang, C. Zheng, Z. Cao, and Y. Chen, "Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties," *Molecular immunology*, vol. 44, no. 4, pp. 514–520, 2007. 44

[138] A. Zorzet, M. Gustafsson, and U. Hammerling, "Prediction of food protein allergenicity: A bio-informatic learning systems approach," *In Silico Biology*, vol. 2, no. 4, pp. 525–534, 2002. 44

[139] D. Soeria-Atmadja, A. Zorzet, M. Gustafsson, and U. Hammerling, "Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms," *International archives of allergy and immunology*, vol. 133, no. 2, pp. 101–112, 2004. 44

[140] R. Aalberse, "Structural biology of allergens," *Journal of allergy and clinical immunology*, vol. 106, no. 2, pp. 228–238, 2000. 44

[141] W. Thomas, B. Hales, and W. Smith, "Structural biology of allergens," *Current Allergy and Asthma Reports*, vol. 5, no. 5, pp. 388–393, 2005. 44

[142] M. Chapman, A. Pomés, H. Breiteneder, and F. Ferreira, "Nomenclature and structural biology of allergens," *Journal of allergy and clinical immunology*, vol. 119, no. 2, pp. 414–420, 2007. 44

[143] R. Hileman, A. Silvanovich, R. Goodman, E. Rice, G. Holleschak, J. Astwood, and S. Hefle, "Bioinformatic methods for allergenicity assessment using a comprehensive allergen database," *International archives of allergy and immunology*, vol. 128, no. 4, pp. 280–291, 2000. 44, 45, 50

[144] S. Miyazawa and R. Jernigan, "Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues," *Proteins: Structure, Function, and Bioinformatics*, vol. 34, no. 1, pp. 49–68, 1999. 44, 52, 56, 113, 115, 164

[145] G. Rose, A. Geselowitz, G. Lesser, R. Lee, and M. Zehfus, "Hydrophobicity of amino acid residues in globular proteins," *Science*, vol. 229, no. 4716, p. 834, 1985. 44, 52, 53, 56

[146] C. Chrysostomou, H. Seker, and N. Aydin, "Effects of windowing and zero-padding on complex resonant recognition model for protein sequence analysis," in *Proceedings of EMBC 2011*, (Boston, USA), pp. 4955–8, August 2011. 46, 100, 102, 113

[147] R. Blackman and J. Tukey, "The Measurement of Power Spectra, 190 pp," *New York*, 1958. 46, 102

[148] S. Lo, C. Cai, Y. Chen, and M. Chung, "Effect of training datasets on support vector machine prediction of protein-protein interactions," *Proteomics*, vol. 5, no. 4, pp. 876–884, 2005. 46, 99

[149] L. Han, C. CaiI, S. Lo, M. Chung, and Y. Chen, "Prediction of RNA-binding proteins from primary sequence by a support vector machine approach," *Rna*, vol. 10, no. 3, p. 355, 2004. 46, 99

[150] S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, no. 8, pp. 721–728, 2001. 46

[151] C. Cai, L. Han, Z. Ji, and Y. Chen, "Enzyme family classification by support vector machines," *PROTEINS: Structure, Function, and Bioinformatics*, vol. 55, no. 1, pp. 66–76, 2004. 47, 99

[152] R. Karchin, K. Karplus, and D. Haussler, "Classifying G-protein coupled receptors with support vector machines," *Bioinformatics*, vol. 18, no. 1, p. 147, 2002. 47, 99

[153] B. E. Boser and et al., "A training algorithm for optimal margin classifiers," in *PROCEEDINGS OF THE 5TH ANNUAL ACM WORKSHOP ON COMPUTATIONAL LEARNING THEORY*, pp. 144–152, ACM Press, 1992. 47, 140, 146

[154] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. 47, 140, 146

[155] A. Aizerman, E. Braverman, and L. Rozoner, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and remote control*, vol. 25, pp. 821–837, 1964. 47

[156] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 47, 146

[157] R. Christensen, M. Enuameh, M. Noyes, M. Brodsky, S. Wolfe, and G. Stormo, "Recognition models to predict dna-binding specificities of homeodomain proteins," *Bioinformatics*, vol. 28, no. 12, pp. i84–i89, 2012. 47

[158] Y. Ou, S. Chen, and M. Gromiha, "Classification of transporters using efficient radial basis function networks with position-specific scoring matrices and biochemical properties," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 7, pp. 1789–1797, 2010. 47

[159] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011. 48

[160] D. Olson and D. Delen, *Advanced data mining techniques*. Springer Verlag, 2008. 48

[161] M. Abramowitz and I. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, vol. 55. Dover publications, 1964. 48, 49

[162] C. J. van Rijsbergen, *Information Retrieval*. Butterworth, 1979. 48, 49

[163] B. Matthews *et al.*, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme.," *Biochimica et biophysica acta*, vol. 405, no. 2, p. 442, 1975. 48, 49

[164] A. Orriols-Puig and E. Bernadó-Mansilla, "Evolutionary rule-based systems for imbalanced data sets," *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, vol. 13, no. 3, pp. 213–225, 2009. 49

[165] J. Wal, "Structure and function of milk allergens," *Allergy*, vol. 56, pp. 35–38, 2001. 52, 54, 55

[166] K. Mengumpun, C. Tayapiwatana, R. Hamilton, P. Sangsupawanich, and R. Wititsuwannakul, "Hydrophobic allergens from the bottom fraction membrane of hevea brasiliensis," *Asian Pacific Journal of Allergy and Immunology*, vol. 26, no. 2-3, pp. 129–136, 2010. 52, 54

[167] M. Gijzen, S. Miller, K. Kuflu, R. Buzzell, and B. Miki, "Hydrophobic protein synthesized in the pod endocarp adheres to the seed surface," *Plant physiology*, vol. 120, no. 4, p. 951, 1999. 52, 54, 55

[168] H. Naderi-Manesh, M. Sadeghi, S. Arab, and A. Moosavi Movahedi, "Prediction of protein surface accessibility with information theory," *Proteins: Structure, Function, and Bioinformatics*, vol. 42, no. 4, pp. 452–459, 2001. 52, 53, 56

[169] G. Fasman, *Prediction of protein structure and the principles of protein conformation*. Springer Us, 1989. 52, 56

[170] S. Katsura and M. Takizawa, "Bethe lattice and the bethe approximation," *Prog. Theor. Phys*, vol. 51, pp. 82–98, 1974. 52

[171] M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, and A. Sarai, "Importance of surrounding residues for protein stability of partially buried mutations.," *Journal of biomolecular structure & dynamics*, vol. 18, no. 2, p. 281, 2000. 53

[172] S. Maleki, R. Kopper, D. Shin, C. Park, C. Compadre, H. Sampson, A. Burks, and G. Bannon, "Structure of the major peanut allergen ara h 1 may protect ige-binding epitopes from degradation," *The journal of Immunology*, vol. 164, no. 11, p. 5844, 2000. 55

[173] A. Trompette, S. Divanovic, A. Visintin, C. Blanchard, R. Hegde, R. Madan, P. Thorne, M. Wills-Karp, T. Gioannini, J. Weiss, *et al.*, "Allergenicity resulting

from functional mimicry of a toll-like receptor complex protein," *Nature*, vol. 457, no. 7229, pp. 585–588, 2008. 55

[174] S. Kumar, K. Tamura, and M. Nei, "Mega3: integrated software for molecular evolutionary genetics analysis and sequence alignment," *Briefings in bioinformatics*, vol. 5, no. 2, pp. 150–163, 2004. 61

[175] S. Henikoff, J. Henikoff, and S. Pietrokovski, "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations.," *Bioinformatics*, vol. 15, no. 6, pp. 471–479, 1999. 61

[176] A. Phillips, D. Janies, and W. Wheeler, "Multiple sequence alignment in phylogenetic analysis," *Molecular Phylogenetics and Evolution*, vol. 16, no. 3, pp. 317–330, 2000. 61

[177] D. Feng and R. Doolittle, "Progressive sequence alignment as a prerequisitetto correct phylogenetic trees," *Journal of molecular evolution*, vol. 25, no. 4, pp. 351–360, 1987. 61, 62, 64

[178] S. Needleman and C. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970. 61, 66, 71

[179] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. . Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman, "The pfam protein families database," *Nucleic acids research*, vol. 36, no. SUPPL. 1, pp. D281–D288, 2008. 61

[180] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton, "Jalview version 2-a multiple sequence alignment editor and analysis workbench," *Bioinformatics*, vol. 25, no. 9, pp. 1189–1191, 2009. 61

[181] A. Marchler-Bauer, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. Deweese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, A. Tasneem, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, and S. H. Bryant, "Cdd: Specific functional annotation with

the conserved domain database," *Nucleic acids research*, vol. 37, no. SUPPL. 1, pp. D205–D210, 2009. 61

[182] T. J. P. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek, "Ensembl 2009," *Nucleic acids research*, vol. 37, no. SUPPL. 1, pp. D690–D697, 2009. 61

[183] D. Lipman and W. Pearson, "Rapid and sensitive protein similarity searches," *Science*, vol. 227, no. 4693, p. 1435, 1985. 61, 63, 64

[184] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, *et al.*, "Clustal w and clustal x version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007. 61, 63, 65

[185] C. Notredame, D. Higgins, and J. Heringa, "T-coffee: a novel method for fast and accurate multiple sequence alignment1," *Journal of molecular biology*, vol. 302, no. 1, pp. 205–217, 2000. 61, 63, 65, 66

[186] K. Katoh and H. Toh, "Recent developments in the mafft multiple sequence alignment program," *Briefings in bioinformatics*, vol. 9, no. 4, pp. 286–298, 2008. 61, 63, 66

[187] D. Lipman, S. Altschul, and J. Kececioglu, "A tool for multiple sequence alignment," *Proceedings of the National Academy of Sciences*, vol. 86, no. 12, p. 4412, 1989. 61

[188] S. Manavski and G. Valle, "Cuda compatible gpu cards as efficient hardware accelerators for smith-waterman sequence alignment," *BMC bioinformatics*, vol. 9, no. Suppl 2, p. S10, 2008. 61

[189] C. Kemena and C. Notredame, "Upcoming challenges for multiple sequence alignment methods in the high-throughput era," *Bioinformatics*, vol. 25, no. 19, pp. 2455–2465, 2009. 62

[190] J. Thompson, F. Plewniak, and O. Poch, "A comprehensive comparison of multiple sequence alignment programs," *Nucleic Acids Research*, vol. 27, no. 13, pp. 2682–2690, 1999. 62, 64, 66

[191] M. Hirosawa, Y. Totoki, M. Hoshida, and M. Ishikawa, "Comprehensive study on iterative algorithms of multiple sequence alignment," *Computer applications in the biosciences: CABIOS*, vol. 11, no. 1, pp. 13–18, 1995. 62, 64

[192] R. Hughey and A. Krogh, "Hidden markov models for sequence analysis: extension and analysis of the basic method," *Computer applications in the biosciences: CABIOS*, vol. 12, no. 2, pp. 95–107, 1996. 62, 64

[193] O. Gotoh, "Optimal alignment between groups of sequences and its application to multiple sequence alignment.," *Comput Appl Biosci*, vol. 9, pp. 361–370, Jun 1993. 62

[194] E. Althaus, A. Caprara, H.-P. Lenhof, and K. Reinert, "Multiple sequence alignment with arbitrary gap costs: computing an optimal solution using polyhedral combinatorics.," *Bioinformatics*, vol. 18 Suppl 2, pp. S4–S16, 2002. 62

[195] M. Vingron and M. S. Waterman, "Sequence alignment and penalty choice. review of concepts, case studies and implications.," *J Mol Biol*, vol. 235, pp. 1–12, Jan 1994. 62

[196] D. Higgins and P. Sharp, "Clustal: a package for performing multiple sequence alignment on a microcomputer," *Gene*, vol. 73, no. 1, pp. 237–244, 1988. 63, 65

[197] M. Dayhoff, R. Schwartz, and B. Orcutt, "A model of evolutionary change in proteins," *Atlas of protein sequence and structure*, vol. 5, pp. 345–352, 1972. 64, 67, 75

[198] T. Smith, M. Waterman, *et al.*, "Identification of common molecular subsequences," *J. mol. Biol*, vol. 147, no. 1, pp. 195–197, 1981. 64

[199] W. Wilbur and D. Lipman, "Rapid similarity searches of nucleic acid and protein data banks," *Proceedings of the National Academy of Sciences*, vol. 80, no. 3, p. 726, 1983. 65

[200] J. Thompson, D. Higgins, and T. Gibson, "Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic acids research*, vol. 22, no. 22, p. 4673, 1994. 65, 66, 126

[201] F. Jeanmougin, J. Thompson, M. Gouy, D. Higgins, T. Gibson, *et al.*, "Multiple sequence alignment with clustal x.," *Trends in biochemical sciences*, vol. 23, no. 10, p. 403, 1998. 65

[202] F. Sievers, A. Wilm, D. Dineen, T. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, *et al.*, "Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega," *Molecular Systems Biology*, vol. 7, no. 1, 2011. 65

[203] G. Vogt, T. Etzold, and P. Argos, "An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited," *Journal of molecular biology*, vol. 249, no. 4, pp. 816–831, 1995. 66

[204] P. Lio and N. Goldman, "Models of molecular evolution and phylogeny," *Genome research*, vol. 8, no. 12, pp. 1233–1244, 1998. 67

[205] C. Kosiol and N. Goldman, "Different versions of the dayhoff rate matrix," *Molecular Biology and Evolution*, vol. 22, no. 2, pp. 193–199, 2005. 67

[206] S. Meyn, R. Tweedie, and P. Glynn, *Markov chains and stochastic stability*, vol. 2. Cambridge University Press Cambridge, 2009. 67

[207] S. Henikoff and J. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, p. 10915, 1992. 67, 75

[208] A. Zomaya, *Handbook of nature-inspired and innovative computing: integrating classical models with emerging technologies*. Springer, 2005. 69

[209] J. Henikoff, S. Henikoff, and S. Pietrokovski, "New features of the blocks database servers," *Nucleic acids research*, vol. 27, no. 1, pp. 226–228, 1999. 69

[210] S. Eddy, "Where did the blosum62 alignment score matrix come from?," *Nature Biotechnology*, vol. 22, no. 8, pp. 1035–1036, 2004. 69

[211] M. Styczynski, K. Jensen, I. Rigoutsos, and G. Stephanopoulos, "Blosum62 miscalculations improve search performance," *Nature biotechnology*, vol. 26, no. 3, pp. 274–275, 2008. 69

[212] G. H. Gonnet, M. A. Cohen, and S. A. Benner, "Exhaustive matching of the entire protein sequence database.," *Science*, vol. 256, pp. 1443–1445, Jun 1992. 69, 71

[213] R. Sokal and C. Michener, "A statistical method for evaluating systematic relationships," *Univ. Kans. Sci. Bull.*, vol. 38, pp. 1409–1438, 1958. 71, 77

[214] Z. Ying-Ding and Q. Xian-Xia, "Research on optimal multiple sequence alignment," in *E-Business and E-Government (ICEE), 2010 International Conference on*, pp. 5500–5505, IEEE, 2010. 72

[215] X. Xia and W. H. Li, "What amino acid properties affect protein evolution?," *J Mol Evol*, vol. 47, pp. 557–564, Nov 1998. 73, 80

[216] S. Woolley, J. Johnson, M. J. Smith, K. A. Crandall, and D. A. McClellan, "Treesaap: selection on amino acid properties using phylogenetic trees.," *Bioinformatics*, vol. 19, pp. 671–672, Mar 2003. 73, 80

[217] G. Singh, *Chemistry of amino-acids and proteins*. Discovery Publishing House, 2007. 73, 80

[218] P. Manavalan and P. Ponnuswamy, "Hydrophobic character of amino acid residues in globular proteins," 1978. 73

[219] R. Wolfenden, P. Cullis, and C. Southgate, "Water, protein folding, and the genetic code," *Science*, vol. 206, no. 4418, p. 575, 1979. 73

[220] P. ARGOS, J. Rao, and P. HARGRAVE, "Structural prediction of membrane-bound proteins," *European Journal of Biochemistry*, vol. 128, no. 2-3, pp. 565–575, 1982. 73

[221] J. ZimmermanNaomi and R. Simha, "The characterization of amino acid sequences in proteins by statistical methods," *Journal of theoretical biology*, vol. 21, no. 2, pp. 170–201, 1968. 73

[222] L. Acid, D. Citrulline, and D. HCI, "Heat capacities, absolute entropies, and entropies of formation of amino acids and related compounds," *Handbook of biochemistry and molecular biology*, vol. 1, no. 154.33, p. 109, 1984. 73

[223] H. Zhou and Y. Zhou, "Quantifying the effect of burial of amino acid residues on protein stability," *PROTEINS: Structure, Function, and Bioinformatics*, vol. 54, no. 2, pp. 315–322, 2004. 73

[224] M. Oobatake and T. Ooi, "An analysis of non-bonded energy of proteins," *Journal of Theoretical Biology*, vol. 67, no. 3, pp. 567–584, 1977. 73

[225] R. Wolfenden, L. Andersson, P. Cullis, and C. Southgate, "Affinities of amino acid side chains for solvent water," *Biochemistry*, vol. 20, no. 4, pp. 849–855, 1981. 73

[226] J. FAUCHÈRE, M. Charton, L. Kier, A. Verloop, and V. Pliska, "Amino acid side chain parameters for correlation studies in biology and pharmacology," *International journal of peptide and protein research*, vol. 32, no. 4, pp. 269–278, 1988. 73

[227] J. Kyte and R. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of molecular biology*, vol. 157, no. 1, pp. 105–132, 1982. 73

[228] R. Bhaskaran and P. Ponnuswamy, "Positional flexibilities of amino acid residues in globular proteins," *International Journal of Peptide and Protein Research*, vol. 32, no. 4, pp. 241–255, 1988. 73

[229] Y. Yu and S. Altschul, "The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions," *Bioinformatics*, vol. 21, no. 7, pp. 902–911, 2005. 75

[230] T. Wu and D. Brutlag, "Discovering empirically conserved amino acid substitution groups in databases of protein families," in *Proceedings of the 4th International Conference on Intelligent Systems for Molecular Biology*, pp. 230–240, 1996. 75

[231] D. Horner, W. Pirovano, and G. Pesole, "Correlated substitution analysis and the prediction of amino acid structural contacts," *Briefings in bioinformatics*, vol. 9, no. 1, pp. 46–56, 2008. 75

[232] Y. Huang and C. Bystroff, "Improved pairwise alignments of proteins in the twilight zone using local structure predictions," *Bioinformatics*, vol. 22, no. 4, pp. 413–422, 2006. 75

[233] D. Rice and D. Eisenberg, "A 3d-1d substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence1," *Journal of molecular biology*, vol. 267, no. 4, pp. 1026–1038, 1997. 75

[234] S. Gong and T. Blundell, "Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures.," *PLoS computational biology*, vol. 4, no. 10, p. e1000179, 2008. 75

[235] N. Goonesekere and B. Lee, "Context-specific amino acid substitution matrices and their use in the detection of protein homologs," *Proteins: Structure, Function, and Bioinformatics*, vol. 71, no. 2, pp. 910–919, 2008. 75

[236] S. Henikoff and J. Henikoff, "Performance evaluation of amino acid substitution matrices," *Proteins: Structure, Function, and Bioinformatics*, vol. 17, no. 1, pp. 49–61, 1993. 76, 97, 165

[237] A. Bernard, *Leucocyte typing: human leucocyte differentiation antigens detected by monoclonal antibodies: specification, classification, nomenclature*. Springer, 1984. 78

[238] P. Kwong, R. Wyatt, J. Robinson, R. Sweet, J. Sodroski, and W. Hendrickson, "Structure of an hiv gp 120 envelope glycoprotein in complex with the cd4 receptor and a neutralizing human antibody," *NATURE-LONDON-*, pp. 648–659, 1998. 79

[239] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez, "A new bioinformatics analysis tools framework at embl–ebi," *Nucleic Acids Research*, vol. 38, no. suppl 2, pp. W695–W699, 2010. 80

[240] D. Wilson and D. Reeder, *Mammal species of the world: a taxonomic and geographic reference*, vol. 1. Johns Hopkins Univ Pr, 2005. 83

[241] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990. 99, 123, 139

[242] K. Mohammed, "Amino acid composition," Nov. 20 1973. US Patent 3,773,930. 99

[243] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S. Kim, "Recognition of a protein fold in the context of the SCOP classification," *Proteins: Structure, Function, and Bioinformatics*, vol. 35, no. 4, pp. 401–407, 1999. 99

[244] J. Bonk and D. Gough, "Whole-proteome interaction mining," *Bioinformatics*, vol. 19, no. 1, pp. 125–135, 2003. 99

[245] C. Chrysostomou, H. Seker, N. Aydin, and P. Haris, "Complex resonant recognition model in analysing influenza a virus subtype protein sequences," in *10th IEEE International Conference on Information Technology and Applications in Biomedicine*, (Corfu, Greece), pp. 1–4, November 2010. 100, 113, 139, 141, 155, 156, 168

[246] N. Aydin, H.S. Markus, "Directional wavelet transform in the context of complex quadrature doppler signals.," *IEEE Signal Processing Letters.*, vol. 10(7), pp. 278–280, 2000. 100, 101

[247] V. Veljkovic, N. Veljkovic, J. Este, A. Huther, and U. Dietrich, "Application of the EIIP/ISM bioinformatics concept in development of new drugs," *Current medicinal chemistry*, vol. 14, no. 4, pp. 441–453, 2007. 100, 117, 124, 166

[248] V. Veljkovic, H. Niman, S. Glisic, N. Veljkovic, V. Perovic, and C. Muller, "Identification of hemagglutinin structural domain and polymorphisms which may modulate swine h1n1 interactions with human receptor," *BMC Structural Biology*, vol. 9, no. 1, p. 62, 2009. 100, 117

[249] E. Pirogova, M. Akay, and I. Cosic, "Investigating the interaction between oncogene and tumor suppressor protein," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 13, no. 1, pp. 10–15, 2009. 100, 118

[250] J. Hybl, A. Yu, D. Farrow, and D. Jonas, "Polar solvation dynamics in the femtosecond evolution of two-dimensional fourier transform spectra," *The Journal of Physical Chemistry A*, vol. 106, no. 34, pp. 7651–7654, 2002. 101

[251] J. Laaser, W. Xiong, and M. Zanni, "Time-domain sfg spectroscopy using mid-ir pulse shaping: Practical and intrinsic advantages," *The Journal of Physical Chemistry B*, 2011. 101

[252] R. Vacha, S. Rick, P. Jungwirth, A. de Beer, H. de Aguiar, J. Samson, and S. Roke, "The orientation and charge of water at the hydrophobic oil droplet-water interface," *Journal of the American Chemical Society*, 2011. 101

[253] Y. Zhang and Y. Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement," *Speech Communication*, vol. 55, no. 4, pp. 509–522, 2013. 101

[254] A. Girgis and F. Ham, "A quantitative study of pitfalls in the FFT," *Aerospace and Electronic Systems, IEEE Transactions on*, no. 4, pp. 434–439, 1980. 102

[255] F. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978. 102

[256] D. Agrez, "Improving phase estimation with leakage minimization," *Instrumentation and Measurement, IEEE Transactions on*, vol. 54, no. 4, pp. 1347–1353, 2005. 102

[257] A. Oppenheim, R. Schafer, J. Buck, *et al.*, *Discrete-time signal processing*, vol. 2. Prentice hall Upper Saddle Riverˆ eN. JNJ, 1989. 102

[258] R. Henry and P. GRAEFE, "Zero padding as a means of improving definition of computed spectra," *MANUSCRIPT REPORT SERIES NO 20, 1971. 10 P, 4 FIG, 1 REF.*, 1971. 102

[259] D. Sundararajan, *The discrete Fourier transform: theory, algorithms and applications*. World Scientific Pub Co Inc, 2001. 102

[260] M. M. Mukhtar, S. T. Rasool, D. Song, C. Zhu, Q. Hao, Y. Zhu, and J. Wu, "Origin of highly pathogenic H5N1 avian influenza virus in China and genetic characterization of donor and recipient viruses," *JOURNAL OF GENERAL VIROLOGY*, vol. 88, pp. 3094–3099, NOV 2007. 106, 108, 148, 159

[261] Y. Choi, S. Goyal, M. Farnham, and H. Joo, "Phylogenetic analysis of H1N2 isolates of influenza A virus from pigs in the United States," *Virus research*, vol. 87, no. 2, pp. 173–179, 2002. 106, 108

[262] A. Moscona, "Neuraminidase inhibitors for influenza," *New England Journal of Medicine*, vol. 353, no. 13, p. 1363, 2005. 107, 131, 148

[263] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, "The influenza virus resource at the National Center for Biotechnology Information," *Journal of virology*, vol. 82, no. 2, p. 596, 2008. 107, 126, 148

[264] D. Morens, J. Taubenberger, and A. Fauci, "The persistent legacy of the 1918 influenza virus," *The New England journal of medicine*, vol. 361, no. 3, p. 225, 2009. 107, 108, 148, 149, 159

[265] D. Morens and A. Fauci, "The 1918 influenza pandemic: insights for the 21st century," *Journal of Infectious Diseases*, vol. 195, no. 7, p. 1018, 2007. 107

[266] S. Zimmer and D. Burke, "Historical perspectiveemergence of influenza a (h1n1) viruses," *New England Journal of Medicine*, vol. 361, no. 3, pp. 279–285, 2009. 107

[267] C. Maring, V. Stoll, C. Zhao, M. Sun, A. Krueger, K. Stewart, D. Madigan, W. Kati, Y. Xu, R. Carrick, *et al.*, "Structure-based characterization and optimization of novel hydrophobic binding interactions in a series of pyrrolidine influenza neuraminidase inhibitors," *Journal of medicinal chemistry*, vol. 48, no. 12, pp. 3980–3990, 2005. 111

[268] J. Varghese, "Development of neuraminidase inhibitors as anti-influenza virus drugs," *Drug Development Research*, vol. 46, no. 3-4, pp. 176–196, 1999. 111

[269] R. Russell, L. Haire, D. Stevens, P. Collins, Y. Lin, G. Blackburn, A. Hay, S. Gamblin, and J. Skehel, "The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design," *Nature*, vol. 443, no. 7107, pp. 45–49, 2006. 111

[270] T. Anwar, S. Lal, and A. Khan, "In silico analysis of genes nucleoprotein, neuraminidase and hemagglutinin: a comparative study on different strains of influenza A (Bird flu) virus sub-type H5N1," *In silico biology*, vol. 6, no. 3, pp. 161–168, 2006. 111

[271] D. Sharma, A. Rawat, S. Srivastava, R. Srivastava, and A. Kumar, "Comparative Sequence Analysis on Different Strains of Swine Influenza Virus Sub-type H1N1 for Neuraminidase and Hemagglutinin," *Journal of Proteomics & Bioinformatics*, vol. 3, no. 2, pp. 55–60, 2010. 111

[272] V. Stoll, K. Stewart, C. Maring, S. Muchmore, V. Giranda, Y. Yu-gui, G. Wang, Y. Chen, M. Sun, C. Zhao, *et al.*, "Influenza neuraminidase inhibitors: structure-based design of a novel inhibitor series," *Biochemistry*, vol. 42, no. 3, pp. 718–727, 2003. 111

[273] C. Browne, H. Bennett, and S. Solomon, "The isolation of peptides by high-performance liquid chromatography using predicted elution positions," *Analytical biochemistry*, vol. 124, no. 1, pp. 201–208, 1982. 113

[274] M. Prabhakaran and P. Ponnuswamy, "Shape and surface features of globular proteins," *Macromolecules*, vol. 15, no. 2, pp. 314–320, 1982. 113

[275] W. Hu, "Identification of highly conserved domains in hemagglutinin associated with the receptor binding specificity of influenza viruses: 2009 h1n1, avian h5n1, and swine h1n2," *Journal of Biomedical Science and Engineering*, vol. 3, no. 2, pp. 114–123, 2010. 117

[276] C. Hejase de Trad, Q. Fang, and I. Cosic, "The resonant recognition model (rrm) predicts amino acid residues in highly conserved regions of the hormone prolactin (prl)," *Biophysical chemistry*, vol. 84, no. 2, pp. 149–157, 2000. 118

[277] G. Zhang and B. Fang, "Predicting the cofactors of oxidoreductases based on amino acid composition distribution and chou's amphiphilic pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 253, no. 2, pp. 310–315, 2008. 123

[278] C. Chen, L. Chen, X. Zou, and P. Cai, "Predicting protein structural class based on multi-features fusion," *Journal of theoretical biology*, vol. 253, no. 2, pp. 388–392, 2008. 123

[279] L. Han, C. Zheng, B. Xie, J. Jia, X. Ma, F. Zhu, H. Lin, X. Chen, and Y. Chen, "Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness," *Drug discovery today*, vol. 12, no. 7, pp. 304–313, 2007. 123

[280] J. Park, S. Dietmann, A. Heger, and L. Holm, "Estimating the significance of sequence order in protein secondary structure and prediction," *Bioinformatics*, vol. 16, no. 11, pp. 978–987, 2000. 123

[281] X. Du and J. Cheng, "Inferring protein-protein interactions from sequence using sequence order information," in *Computer Science and Education (ICCSE), 2010 5th International Conference on*, pp. 481–486, IEEE, 2010. 123

[282] H. Shen and K. Chou, "Predicting protein fold pattern with functional domain and sequential evolution information," *Journal of Theoretical Biology*, vol. 256, no. 3, pp. 441–446, 2009. 123

[283] C. Chrysostomou, H. Seker, and N. Aydin, "Investigation into the effects of an individual amino acid on protein function by means of a resonant recognition model," in *Proceedings of the 5th international conference on Convergence and hybrid information technology*, ICHIT'11, (Berlin, Heidelberg), pp. 229–236, Springer-Verlag, 2011. 124

[284] L. Simonsen, G. Bernabe, K. Lacourciere, R. Taylor, and M. Giovanni, "The niaid influenza genome sequencing project," *National Institute of Allergy and Infectious Diseases, NIH*, pp. 109–113, 2008. 126

[285] Y. Li, J. Shi, G. Zhong, G. Deng, G. Tian, J. Ge, X. Zeng, J. Song, D. Zhao, L. Liu, *et al.*, "Continued evolution of h5n1 influenza viruses in wild birds, domestic poultry, and humans in china from 2004 to 2009," *Journal of virology*, vol. 84, no. 17, pp. 8389–8397, 2010. 126

[286] V. Gregory, M. Bennett, M. Orkhan, S. Al Hajjar, N. Varsano, E. Mendelson, M. Zambon, J. Ellis, A. Hay, and Y. Lin, "Emergence of influenza a h1n2 reassortant viruses in the human population during 2001," *Virology*, vol. 300, no. 1, pp. 1–7, 2002. 126

[287] X. Xu, X. Zhu, R. Dwek, J. Stevens, and I. Wilson, "Structural characterization of the 1918 influenza virus h1n1 neuraminidase," *Journal of virology*, vol. 82, no. 21, pp. 10493–10501, 2008. 129, 130, 132

[288] W. Klösgen, "Explora: A multipattern and multistrategy discovery assistant," in *Advances in knowledge discovery and data mining*, pp. 249–271, American Association for Artificial Intelligence, 1996. 140, 142, 143

[289] S. Wrobel, "An algorithm for multi-relational discovery of subgroups," *Principles of Data Mining and Knowledge Discovery*, pp. 78–87, 1997. 140, 142

[290] D. Gamberger, N. Lavrac, F. Zelezny, and J. Tolar, "Induction of comprehensible models for gene expression datasets by subgroup discovery methodology," *Journal of biomedical informatics*, vol. 37, no. 4, pp. 269–284, 2004. 140

[291] N. Lavrač, "Subgroup discovery techniques and applications," *Advances in Knowledge Discovery and Data Mining*, pp. 13–20, 2005. 140

[292] I. Trajkovski, F. Zelezny, N. Lavrac, and J. Tolar, "Relational descriptive analysis of gene expression data," in *Proceeding of the 2006 conference on STAIRS 2006: Proceedings of the Third Starting AI Researchers' Symposium*, pp. 184–195, IOS Press, 2006. 140

[293] I. Trajkovski, F. Zelezny, N. Lavrac, and J. Tolar, "Learning relational descriptions of differentially expressed gene groups," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 38, no. 1, pp. 16–25, 2008. 140

[294] F. Železnỳ, J. Tolar, N. Lavrac, and O. Štepánková, "Relational subgroup discovery for gene expression data mining," 2005. 140

[295] C. Carmona, P. González, M. del Jesus, and F. Herrera, "Nmeef-sd: non-dominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery," *Fuzzy Systems, IEEE Transactions on*, vol. 18, no. 5, pp. 958–970, 2010. 140, 142, 144, 154

[296] M. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998. 140, 146

[297] I. Steinwart and A. Christmann, *Support vector machines*. Springer Verlag, 2008. 140, 146

[298] D. Gamberger and N. Lavrac, "Expert-guided subgroup discovery: Methodology and application," *Arxiv preprint arXiv:1106.4576*, 2011. 142

[299] N. Lavrač, B. Cestnik, D. Gamberger, and P. Flach, "Decision support through subgroup discovery: Three case studies and the lessons learned," *Machine Learning*, vol. 57, no. 1, pp. 115–143, 2004. 142

[300] C. Carmona, P. González, M. del Jesus, and F. Herrera, "Non-dominated multiobjective evolutionary algorithm based on fuzzy rules extraction for subgroup discovery," *Hybrid Artificial Intelligence Systems*, pp. 573–580, 2009. 142

[301] C. Carmona, P. González, M. Jesus, and F. Herrera, "An analysis of evolutionary algorithms with different types of fuzzy rules in subgroup discovery," in *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on*, pp. 1706–1711, IEEE, 2009. 142

[302] C. Carmona, P. González, M. del Jesus, C. Romero, and S. Ventura, "Evolutionary algorithms for subgroup discovery applied to e-learning data," in *Education Engineering (EDUCON), 2010 IEEE*, pp. 983–990, IEEE, 2010. 142

[303] C. Carmona, P. González, M. del Jesus, M. Navío-Acosta, and L. Jiménez-Trevino, "Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department," *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, pp. 1–14, 2010. 142

[304] N. Lavrač, P. Flach, and B. Zupan, "Rule evaluation measures: A unifying view," *Inductive Logic Programming*, pp. 174–185, 1999. 143

[305] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. Verkamo, *et al.*, "Fast discovery of association rules," *Advances in knowledge discovery and data mining*, vol. 12, pp. 307–328, 1996. 143

[306] M. del Jesus, P. González, F. Herrera, and M. Mesonero, "Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing," *Fuzzy Systems, IEEE Transactions on*, vol. 15, no. 4, pp. 578–592, 2007. 144, 145

[307] M. del Jesus, P. González, and F. Herrera, "Multiobjective genetic algorithm for extracting subgroup discovery fuzzy rules," in *Computational Intelligence in Multicriteria Decision Making, IEEE Symposium on*, pp. 50–57, IEEE, 2007. 144

[308] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *Evolutionary Computation, IEEE Transactions on*, vol. 6, no. 2, pp. 182–197, 2002. 144

[309] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 1-4, pp. 131–156, 1997. 146

[310] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*, vol. 454. Springer, 1998. 146

[311] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*, vol. 207. Springer, 2006. 147

[312] S. Kawashima, H. Ogata, and M. Kanehisa, "AAindex: amino acid index database," *Nucleic Acids Research*, vol. 27, no. 1, p. 368, 1999. 160

# Appendix A

# Literature Review

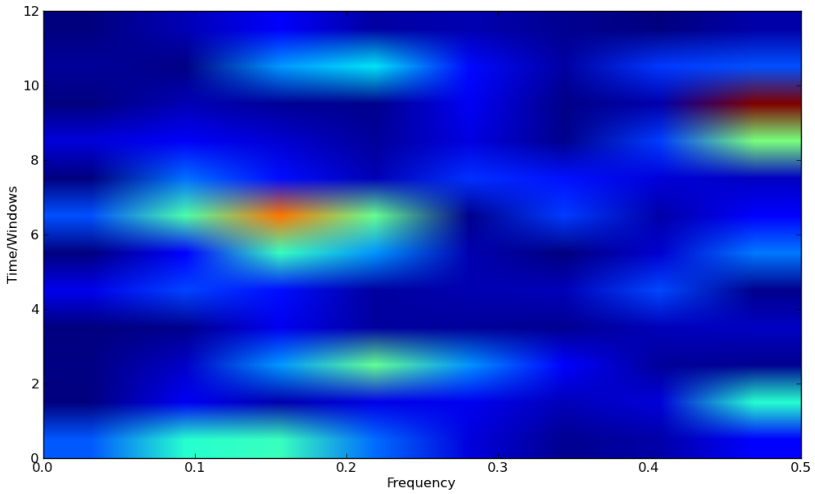



Figure A.1: Short-Space Fourier of Acid Bovine FGF Protein (Window: 10% - Overlap: 25%)

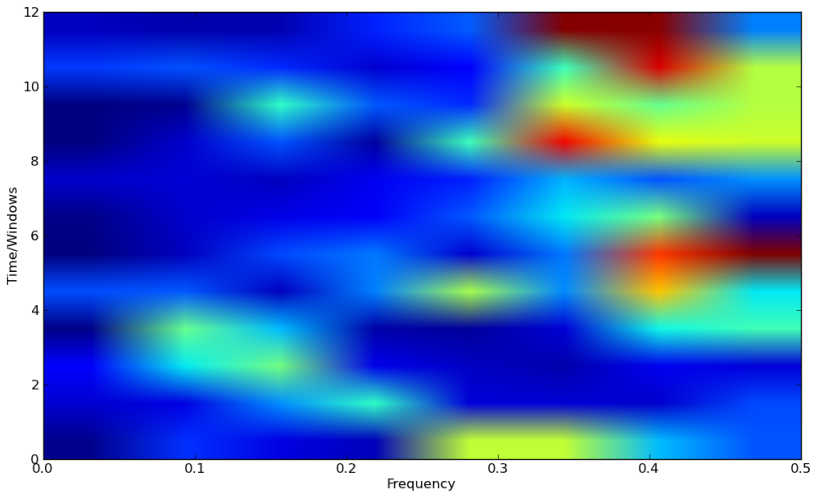



Figure A.2: Short-Space Fourier of Basic Bovine FGF Protein (Window:10% - Overlap: 25%)

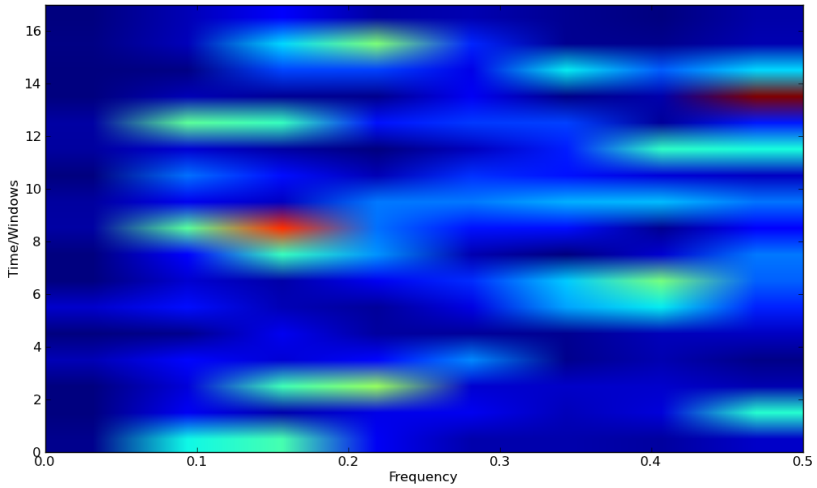



Figure A.3: Short-Space Fourier of Acid Bovine FGF Protein (Window: 10% - Overlap: 50%)

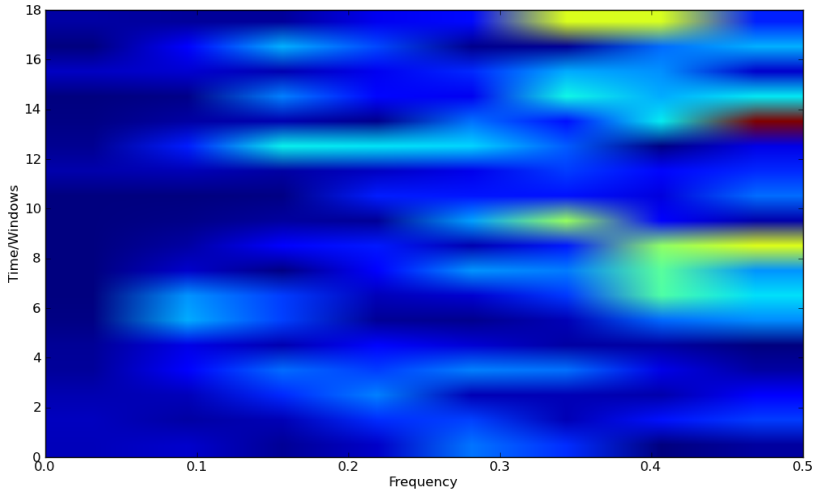



Figure A.4: Short-Space Fourier of Basic Bovine FGF Protein (Window: 10% - Overlap: 50%)

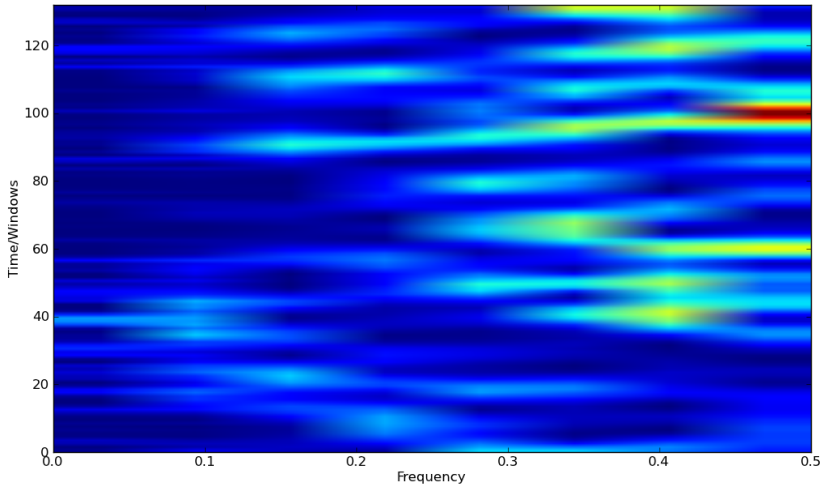



Figure A.5: Short-Space Fourier of Acid Bovine FGF Protein (Window: 10% - Overlap: 99%)

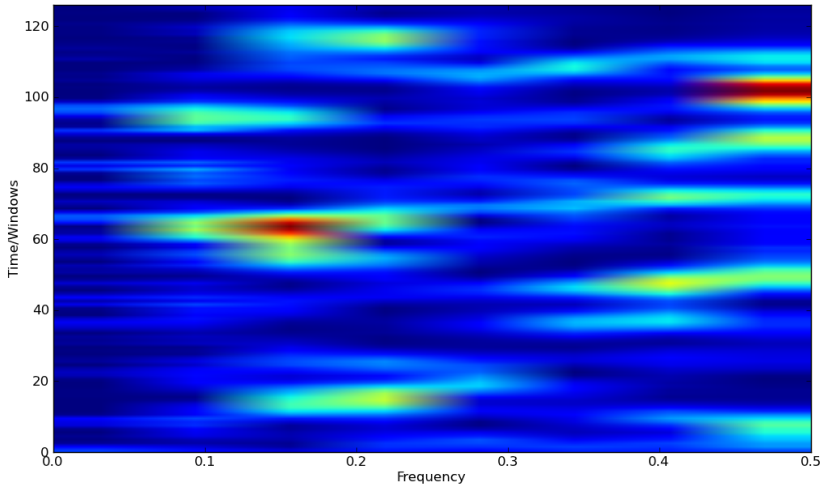



Figure A.6: Short-Space Fourier of Basic Bovine FGF Protein (Window: 10% - Overlap: 99%)

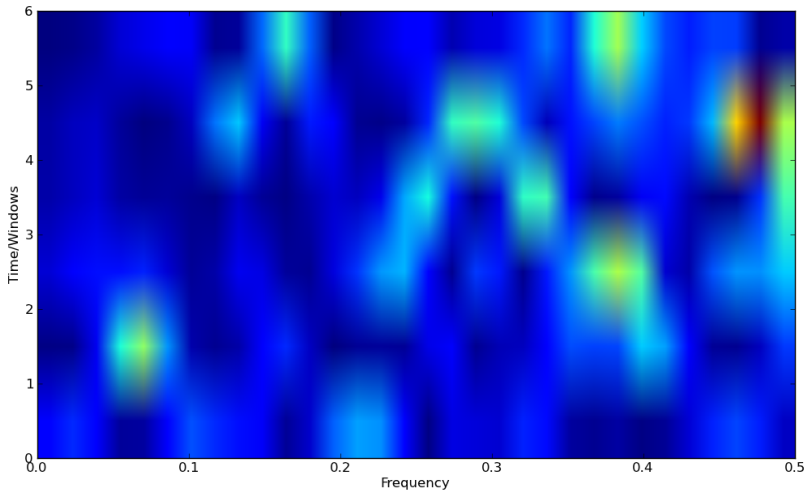



Figure A.7: Short-Space Fourier of Acid Bovine FGF Protein (Window: 28% - Overlap: 50%)

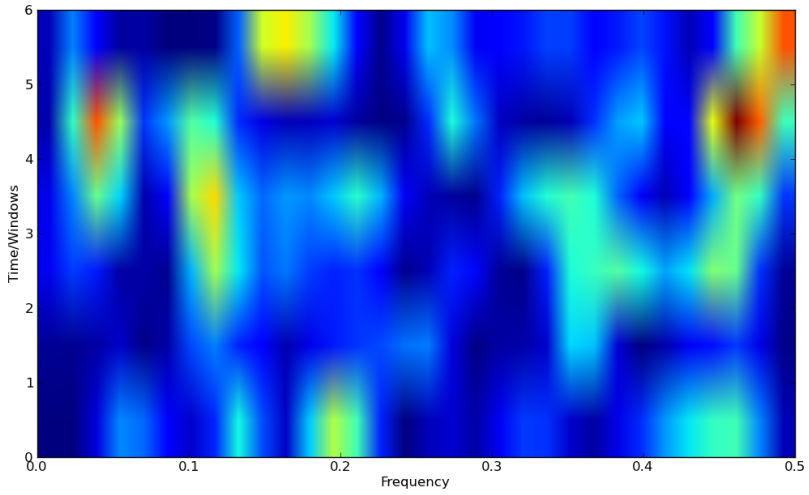



Figure A.8: Short-Space Fourier of Basic Bovine FGF Protein (Window: 28% - Overlap: 50%)

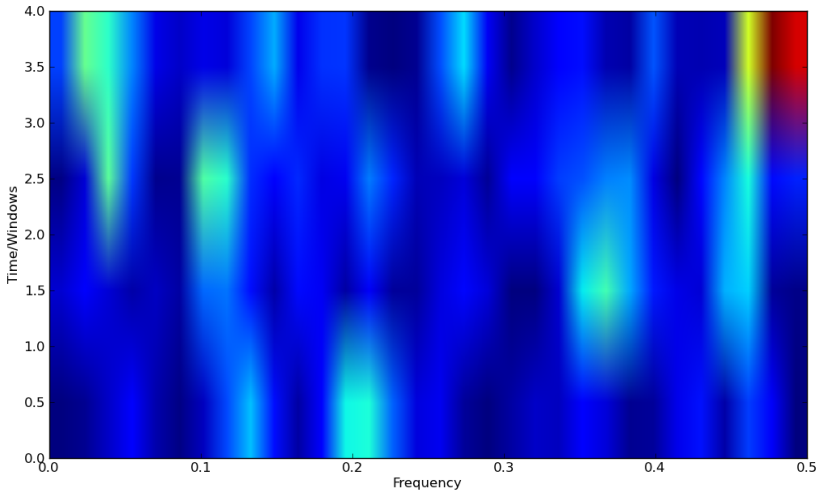



Figure A.9: Short-Space Fourier of Acid Bovine FGF Protein (Window: 40% - Overlap: 50%)

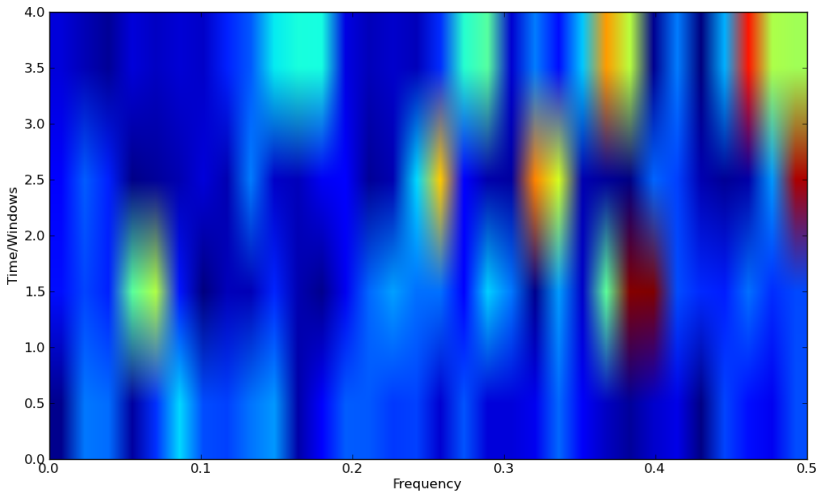



Figure A.10: Short-Space Fourier of Basic Bovine FGF Protein (Window: 40% - Overlap: 50%)

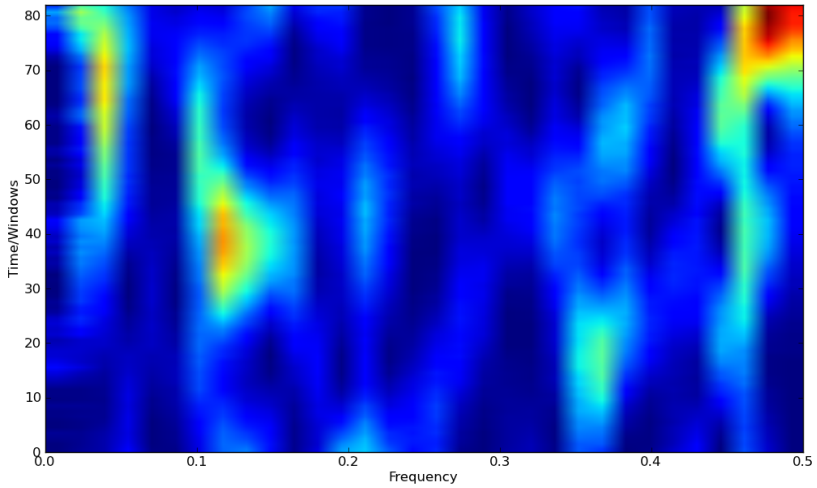



Figure A.11: Short-Space Fourier of Acid Bovine FGF Protein (Window: 40% - Overlap: 99%)

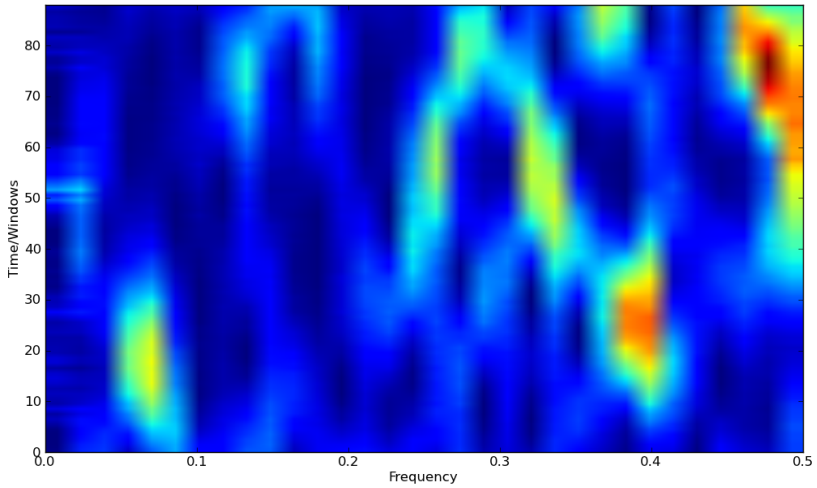



Figure A.12: Short-Space Fourier of Basic Bovine FGF Protein (Window: 40% - Overlap: 99%)

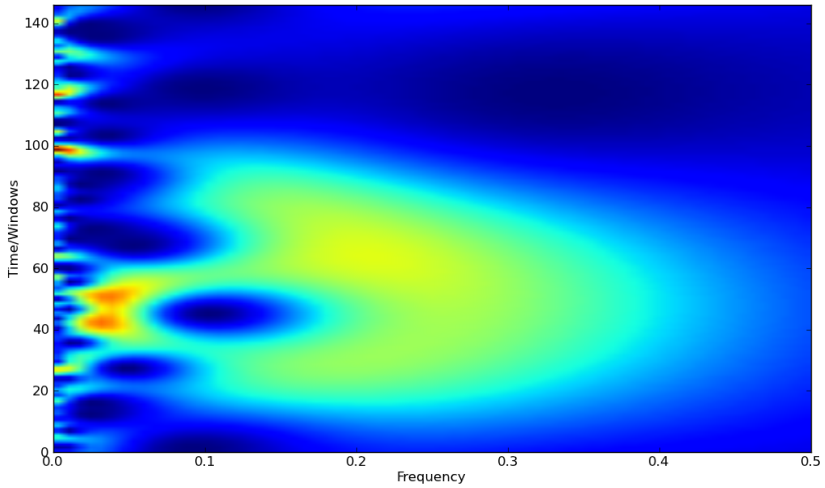



Figure A.13: Paul Wavelet Transform of Acid Bovine FGF Protein

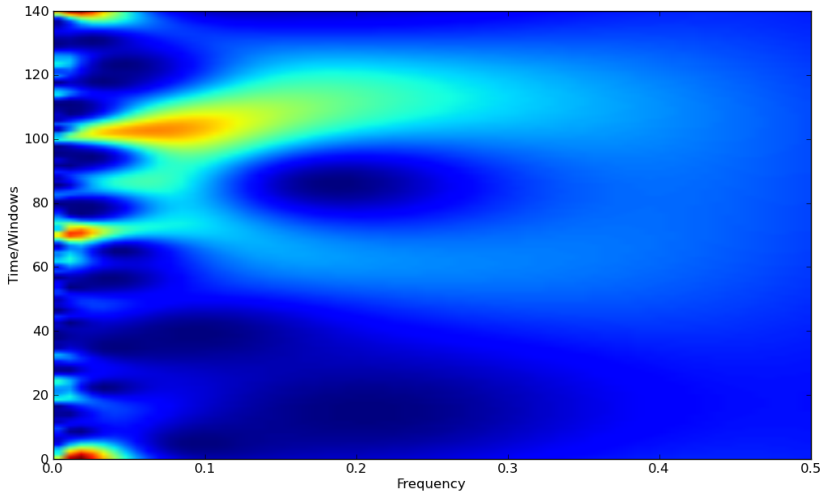



Figure A.14: Paul Wavelet Transform of Basic Bovine FGF Protein

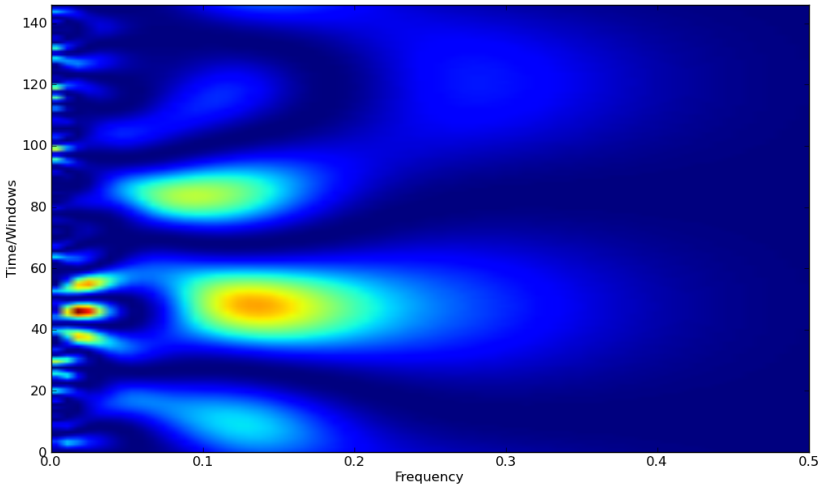



Figure A.15: Mexican Hat Wavelet Transform of Acid Bovine FGF Protein

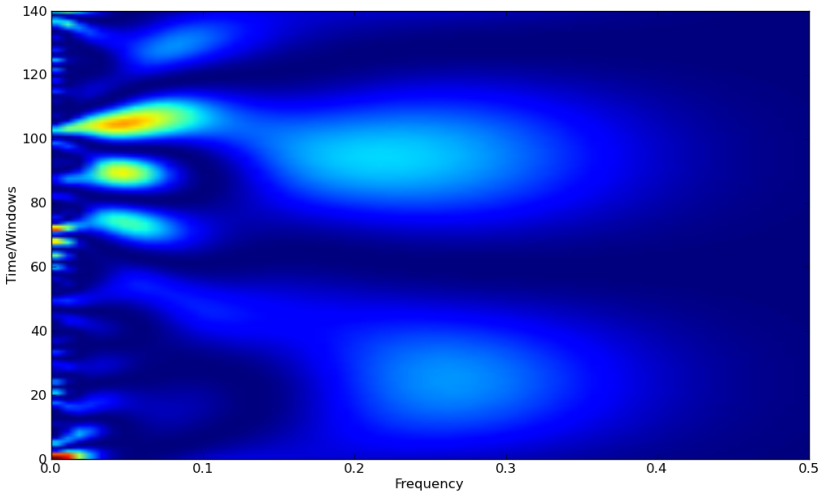



Figure A.16: Mexican Hat Wavelet Transform of Basic Bovine FGF Protein

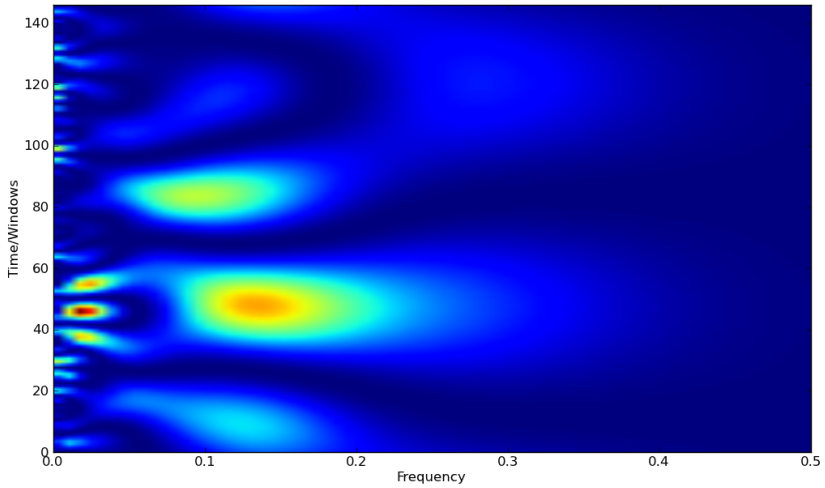



Figure A.17: Derivative of Gaussian Wavelet Transform of Acid Bovine FGF Protein

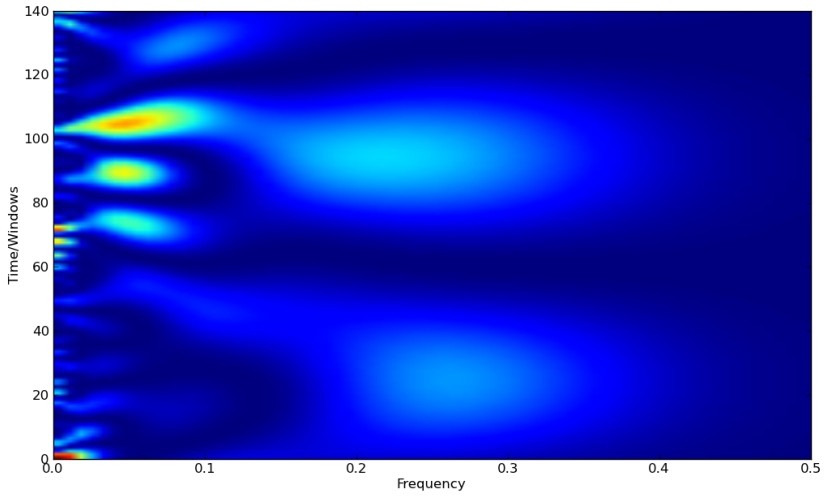



Figure A.18: Derivative of Gaussian Wavelet Transform of Basic Bovine FGF Protein

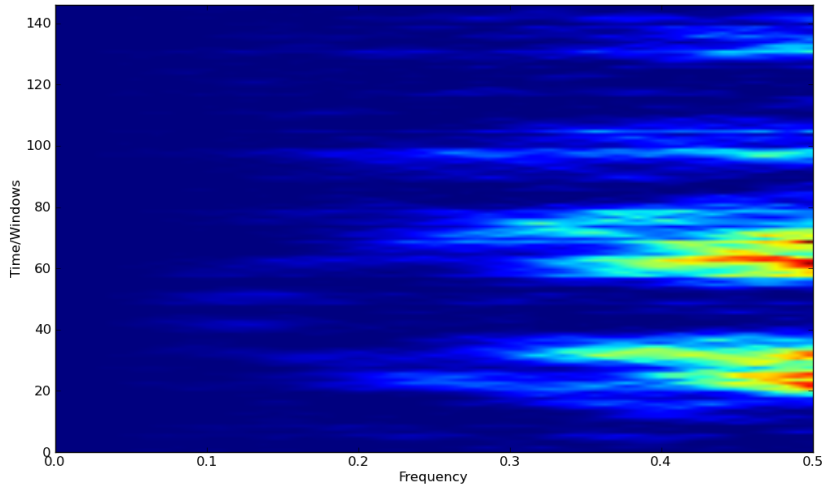



Figure A.19: Haar Wavelet Transform of Acid Bovine FGF Protein

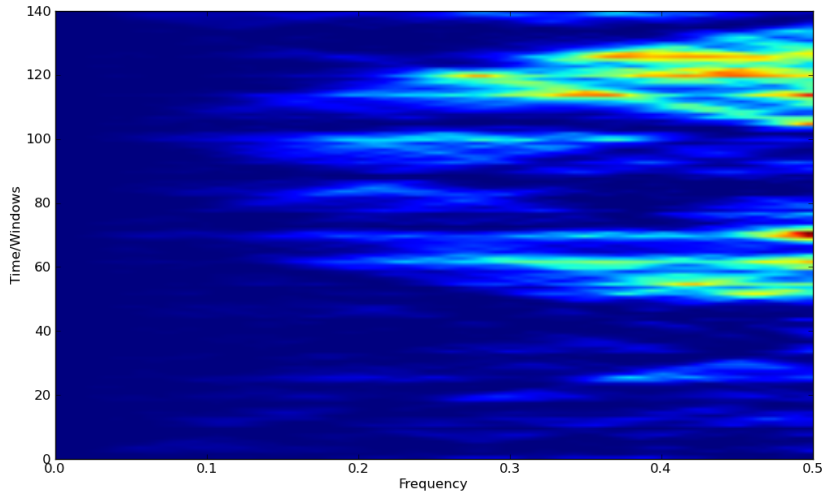



Figure A.20: Haar Wavelet Transform of Basic Bovine FGF Protein

# Appendix B

# Amino Acid Indices

## B. AMINO ACID INDICES

Table B.1: Amino Acid Indices from the Literature that were not included in the AAIndex database.

| ID | Description | Reference | ID | Description | Reference |
|---|---|---|---|---|---|
| 529 | Factor 1 | [109] | 571 | Solvent-accessible surface area for denatured | [111] |
| 530 | Factor 2 | [109] | 572 | Solvent-accessible surface area for native | [111] |
| 531 | Factor 3 | [109] | 573 | Solvent-accessible surface area for unfolding | [111] |
| 532 | Factor 4 | [109] | 574 | Gibbs free energy change of hydration for unfolding | [111] |
| 533 | Factor 5 | [109] | 575 | Gibbs free energy change of hydration for denatured | [111] |
| 534 | Hydrophobicity | [31] | 576 | Gibbs free energy change of hydration for native protein | [111] |
| 535 | Mass | [31] | 577 | Unfolding enthalpy change of hydration | [111] |
| 536 | pK1(a-CO2H | [31] | 578 | Unfolding entropy change of hydration | [111] |
| 537 | pK2(NH3 | [31] | 579 | Unfolding hydration heat capacity change | [111] |
| 538 | pI(NH3 | [31] | 580 | Unfolding Gibbs free energy | [111] |
| 539 | Goldman-Engelman-Steitz scale | [110] | 581 | Unfolding enthalpy | [111] |
| 540 | GA set 1 | [110] | 582 | Unfolding entropy changes of side-chain | [111] |
| 541 | GA set 2 | [110] | 583 | Unfolding Gibbs free energy change | [111] |
| 542 | GA set 3 | [110] | 584 | Unfolding enthalpy change | [111] |
| 543 | Relative connectivity | [29] | 585 | Unfolding entropy change of protein | [111] |
| 544 | Relative Clustering Coefficient | [29] | 586 | Volume | [111] |
| 545 | Relative Closeness | [29] | 587 | Shape | [111] |
| 546 | Relative Betweenness | [29] | 588 | Flexibility | [111] |
| 547 | Compressibility | [111] | 589 | Pf-s | [111] |
| 548 | Surrounding hydrophobicity | [111] | 590 | Hydrophobic Parameter | [34] |
| 549 | Polarity | [111] | 591 | Recognition factors | [112] |
| 550 | Isoelectric point | [111] | 592 | Hydrophobicity | [112] |
| 551 | Equilibrium constant | [111] | 593 | Hydrophobicity | [112] |
| 552 | Molecular weight | [111] | 594 | Hydrophobicity | [112] |
| 553 | Bulkiness | [111] | 595 | Hydrophobic | [112] |
| 554 | Chromatographic index | [111] | 596 | Mobilities chromatography paper (RF | [112] |
| 555 | Refractive index | [111] | 597 | Molar fraction 2001 buried residues | [112] |

Table B.1: (continued)

| | | | | | |
|---|---|---|---|---|---|
| 556 | Normalized consensus hydrophobicity | [111] | 598 | Average flexibility index | [112] |
| 557 | Short- and medium-range nonbonded energy | [111] | 599 | Conformational parameter alpha helix | [112] |
| 558 | Medium-range nonbonded energy | [111] | 600 | Conformational parameter beta turn | [112] |
| 559 | Total nonbonded energy | [111] | 601 | overall amino acid composition | [112] |
| 560 | Coil tendencies | [111] | 602 | Number of codon coding | [112] |
| 561 | Helical contact area | [111] | 603 | Membrane buried helix parameter | [112] |
| 562 | Mean RMS fluctuational displacement | [111] | 604 | Antigenicity value X 10. | [112] |
| 563 | Buriedness | [111] | 605 | Hydrophobicity indices ph 7.5 | [112] |
| 564 | N-terminal of a-helix | [111] | 606 | Transmembrane tendency | [112] |
| 565 | C-terminal of a-helix | [111] | 607 | Molar fraction accessible residues | [112] |
| 566 | Middle of a-helix | [111] | 608 | Conformational parameter beta sheet | [112] |
| 567 | Partial specific volume | [111] | 609 | Conformational parameter coil | [112] |
| 568 | Average medium contacts | [111] | 610 | Amino acid composition | [112] |
| 569 | Long-range contacts | [111] | 611 | Characterizing distribution of allergen-unique segments | [113] |
| 570 | Combined surrounding hydrophobicity | [111] | | | |

Table B.2: PCA Generated Amino Acid Indices With Single Linkage 1

| Variance | Indices | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 17 | -0.19 | -0.11 | 0.08 | 0.09 | 0.13 | -0.16 | -0.20 | -0.41 | -0.01 | 0.51 | 0.44 | -0.18 | -0.10 | 0.02 | -0.02 | -0.10 | -0.20 | 0.04 | -0.01 | 0.37 |
| 1 | 3 4 | -0.33 | -0.16 | 0.05 | 0.01 | 0.66 | 0.27 | -0.03 | -0.05 | 0.10 | -0.19 | 0.24 | 0.01 | -0.31 | -0.06 | 0.06 | 0.21 | -0.01 | -0.24 | -0.21 | -0.01 |
| 1 | 5 564 | -0.15 | -0.26 | -0.06 | -0.26 | 0.33 | 0.31 | 0.27 | -0.06 | -0.15 | -0.02 | 0.26 | 0.15 | -0.58 | 0.16 | -0.09 | 0.16 | 0.02 | 0.12 | -0.19 | 0.03 |
| 1 | 8 598 | 0.10 | 0.11 | -0.23 | -0.05 | 0.16 | -0.20 | 0.22 | -0.23 | -0.35 | -0.16 | 0.25 | 0.27 | 0.19 | -0.43 | 0.09 | 0.23 | -0.32 | 0.20 | -0.06 | 0.20 |
| 1 | 12 13 | 0.23 | -0.48 | -0.18 | 0.02 | 0.45 | -0.14 | 0.08 | 0.10 | -0.29 | -0.36 | 0.29 | -0.06 | 0.10 | 0.25 | -0.05 | -0.01 | -0.01 | -0.19 | 0.14 | 0.10 |
| 1 | 15 59 | 0.17 | -0.42 | -0.04 | 0.17 | -0.24 | -0.04 | 0.02 | 0.18 | -0.02 | 0.27 | 0.03 | 0.06 | 0.04 | -0.18 | -0.05 | 0.43 | 0.11 | -0.28 | -0.47 | 0.25 |
| 1 | 21 79 | -0.08 | -0.03 | 0.35 | 0.33 | -0.11 | -0.07 | -0.09 | -0.23 | 0.10 | 0.21 | 0.19 | -0.20 | 0.04 | -0.03 | 0.56 | -0.10 | -0.31 | -0.27 | -0.03 | -0.23 |
| 1 | 22 80 | 0.18 | -0.08 | -0.02 | -0.41 | 0.37 | -0.10 | -0.49 | 0.26 | 0.17 | -0.03 | -0.03 | -0.14 | 0.22 | 0.14 | 0.15 | -0.09 | -0.18 | 0.30 | -0.27 | 0.05 |
| 1 | 39 225 | 0.22 | -0.43 | -0.14 | 0.56 | -0.12 | -0.33 | 0.18 | 0.24 | 0.01 | 0.07 | -0.08 | 0.06 | -0.23 | 0.16 | -0.22 | -0.10 | 0.09 | -0.16 | 0.04 | 0.18 |
| 1 | 42 566 | 0.06 | -0.25 | -0.38 | 0.02 | -0.36 | -0.33 | 0.05 | -0.05 | -0.27 | 0.25 | 0.15 | 0.38 | 0.25 | 0.03 | -0.15 | 0.12 | 0.28 | -0.13 | 0.20 | 0.14 |
| 1 | 52 346 | 0.09 | -0.13 | 0.12 | 0.01 | 0.02 | -0.26 | 0.23 | 0.08 | -0.28 | -0.30 | 0.13 | -0.05 | -0.18 | -0.43 | 0.47 | 0.06 | -0.22 | 0.19 | 0.30 | 0.14 |
| 1 | 65 135 | -0.04 | -0.23 | 0.38 | 0.32 | -0.15 | 0.16 | 0.43 | 0.16 | -0.37 | -0.16 | -0.05 | -0.15 | 0.00 | 0.03 | 0.11 | -0.07 | -0.22 | 0.18 | 0.05 | -0.39 |
| 1 | 85 110 | -0.07 | -0.01 | -0.21 | 0.45 | -0.47 | -0.06 | 0.08 | -0.37 | 0.33 | -0.07 | -0.07 | -0.16 | 0.32 | 0.24 | -0.19 | 0.13 | -0.05 | -0.06 | 0.14 | 0.09 |
| 1 | 93 421 | -0.47 | -0.27 | 0.31 | 0.00 | 0.13 | 0.02 | -0.14 | 0.29 | 0.32 | 0.13 | -0.30 | -0.24 | -0.17 | -0.07 | -0.14 | 0.00 | 0.26 | -0.11 | 0.20 | 0.23 |
| 1 | 98 228 | 0.14 | -0.48 | 0.26 | -0.24 | 0.10 | 0.26 | -0.09 | 0.15 | -0.01 | -0.06 | -0.07 | -0.03 | -0.11 | 0.00 | 0.49 | -0.06 | 0.34 | -0.16 | -0.31 | -0.12 |
| 1 | 100 230 | 0.00 | 0.21 | -0.18 | -0.13 | 0.24 | 0.47 | -0.47 | 0.00 | -0.17 | -0.28 | 0.22 | -0.05 | 0.11 | -0.23 | 0.03 | 0.05 | 0.32 | -0.09 | -0.19 | 0.16 |
| 1 | 105 234 | 0.02 | -0.33 | 0.17 | 0.24 | 0.18 | -0.13 | 0.21 | 0.11 | -0.25 | -0.11 | -0.25 | 0.13 | 0.31 | 0.37 | -0.44 | 0.03 | -0.03 | 0.18 | -0.21 | -0.19 |
| 1 | 106 428 | -0.14 | -0.15 | 0.21 | -0.13 | 0.49 | -0.44 | 0.22 | 0.06 | -0.11 | 0.31 | -0.28 | -0.23 | -0.12 | 0.02 | 0.03 | 0.08 | 0.17 | -0.09 | -0.17 | 0.28 |
| 1 | 115 153 | -0.08 | -0.16 | -0.09 | 0.19 | -0.07 | -0.09 | 0.19 | -0.09 | -0.08 | -0.05 | -0.05 | -0.16 | -0.06 | -0.03 | 0.90 | -0.10 | -0.08 | -0.01 | -0.03 | -0.05 |
| 1 | 118 588 | -0.04 | -0.12 | 0.13 | 0.26 | -0.33 | 0.21 | 0.33 | 0.21 | -0.21 | -0.12 | -0.11 | -0.07 | 0.53 | 0.16 | -0.29 | 0.11 | -0.20 | -0.15 | -0.08 | -0.21 |
| 1 | 124 175 | -0.11 | 0.12 | 0.03 | 0.12 | 0.06 | 0.04 | -0.13 | 0.40 | -0.40 | -0.30 | -0.38 | -0.20 | -0.08 | -0.10 | 0.01 | 0.37 | 0.28 | 0.09 | 0.29 | -0.14 |
| 1 | 139 608 | 0.08 | -0.26 | -0.28 | 0.13 | 0.15 | 0.32 | -0.48 | 0.41 | 0.08 | -0.21 | 0.21 | -0.07 | -0.26 | -0.19 | 0.22 | -0.06 | 0.12 | -0.07 | 0.17 | -0.01 |
| 1 | 154 157 | 0.38 | -0.12 | 0.12 | 0.19 | -0.08 | 0.10 | 0.11 | -0.18 | 0.27 | -0.44 | -0.01 | -0.37 | -0.41 | 0.22 | -0.11 | 0.20 | -0.06 | 0.19 | 0.10 | -0.12 |
| 1 | 166 275 | 0.01 | -0.31 | 0.07 | 0.17 | -0.02 | -0.11 | -0.13 | 0.02 | -0.10 | -0.02 | 0.14 | 0.03 | -0.24 | -0.34 | 0.33 | -0.12 | -0.17 | 0.35 | -0.15 | 0.59 |
| 1 | 176 555 | 0.95 | -0.02 | -0.08 | -0.09 | 0.03 | -0.06 | -0.06 | -0.15 | -0.04 | -0.05 | -0.05 | -0.04 | -0.04 | 0.00 | -0.09 | -0.12 | -0.09 | 0.06 | 0.01 | -0.08 |
| 1 | 196 455 | 0.32 | -0.22 | 0.04 | 0.06 | -0.25 | 0.08 | 0.08 | 0.48 | 0.04 | -0.14 | -0.20 | -0.31 | 0.18 | 0.25 | -0.12 | -0.44 | 0.00 | 0.24 | 0.04 | -0.14 |
| 1 | 198 208 | -0.26 | -0.13 | 0.26 | 0.02 | 0.29 | -0.03 | 0.03 | -0.63 | -0.05 | 0.48 | -0.11 | -0.02 | -0.02 | 0.21 | -0.01 | 0.03 | -0.20 | -0.06 | 0.13 | 0.09 |
| 1 | 205 465 | -0.07 | -0.19 | 0.00 | -0.15 | 0.21 | -0.09 | -0.16 | 0.31 | 0.03 | 0.55 | -0.25 | 0.05 | 0.10 | -0.35 | 0.03 | -0.37 | -0.10 | 0.32 | 0.02 | 0.10 |
| 1 | 212 529 | -0.13 | -0.13 | 0.03 | 0.14 | -0.08 | 0.25 | -0.26 | -0.36 | -0.35 | 0.25 | -0.06 | 0.09 | -0.41 | 0.46 | 0.09 | 0.24 | -0.04 | 0.01 | 0.08 | 0.19 |
| 1 | 233 252 | -0.16 | -0.06 | 0.18 | -0.07 | -0.58 | 0.33 | 0.30 | 0.04 | -0.15 | 0.24 | -0.04 | -0.14 | 0.32 | 0.26 | -0.23 | -0.05 | 0.08 | -0.20 | 0.10 | -0.16 |
| 1 | 236 347 | -0.26 | 0.11 | 0.28 | 0.13 | -0.07 | -0.15 | -0.06 | -0.04 | 0.14 | 0.31 | -0.01 | -0.46 | -0.31 | 0.18 | 0.08 | -0.06 | 0.42 | 0.05 | -0.39 | 0.08 |
| 1 | 251 527 | 0.04 | -0.15 | -0.08 | 0.06 | 0.25 | -0.32 | 0.18 | 0.13 | -0.35 | 0.22 | -0.15 | 0.27 | -0.24 | 0.12 | 0.32 | -0.08 | -0.18 | -0.42 | 0.26 | 0.16 |
| 1 | 269 270 | -0.21 | -0.18 | -0.11 | 0.29 | 0.06 | 0.40 | -0.17 | 0.24 | 0.12 | -0.20 | 0.08 | -0.01 | 0.26 | -0.04 | -0.55 | 0.06 | 0.07 | 0.14 | -0.33 | 0.09 |
| 1 | 271 281 | -0.16 | 0.18 | 0.06 | -0.27 | -0.20 | 0.24 | 0.00 | 0.32 | -0.42 | 0.24 | -0.13 | 0.51 | 0.03 | -0.09 | -0.21 | 0.12 | -0.03 | 0.16 | -0.19 | -0.16 |
| 1 | 282 361 | -0.02 | -0.28 | -0.18 | 0.34 | -0.38 | 0.31 | -0.09 | 0.04 | -0.07 | 0.00 | 0.11 | 0.50 | 0.01 | -0.13 | -0.22 | -0.01 | -0.09 | 0.36 | 0.01 | -0.25 |
| 1 | 292 293 | 0.04 | -0.18 | 0.13 | -0.25 | 0.09 | 0.14 | 0.01 | 0.22 | -0.07 | 0.46 | -0.24 | -0.31 | 0.26 | -0.23 | -0.17 | -0.28 | -0.07 | -0.16 | 0.30 | 0.31 |
| 1 | 315 556 | 0.17 | -0.10 | -0.19 | -0.28 | 0.19 | 0.08 | -0.11 | 0.22 | -0.29 | -0.12 | 0.17 | 0.55 | 0.11 | -0.34 | 0.32 | -0.07 | -0.06 | -0.24 | -0.11 | 0.10 |
| 1 | 335 364 | 0.10 | 0.00 | -0.47 | -0.06 | 0.03 | 0.21 | -0.01 | -0.11 | 0.49 | 0.24 | 0.29 | -0.06 | 0.00 | -0.12 | 0.24 | -0.17 | -0.09 | -0.36 | 0.12 | -0.26 |
| 1 | 340 341 | -0.05 | -0.03 | -0.16 | -0.18 | 0.22 | 0.06 | 0.31 | -0.22 | -0.04 | 0.01 | -0.19 | -0.21 | -0.19 | -0.02 | 0.69 | 0.02 | -0.06 | -0.29 | 0.20 | 0.13 |
| 1 | 362 515 | 0.24 | -0.17 | -0.16 | -0.11 | 0.16 | -0.18 | -0.14 | 0.22 | -0.26 | 0.34 | 0.18 | -0.06 | 0.06 | -0.24 | 0.21 | 0.21 | 0.15 | -0.26 | -0.47 | 0.27 |
| 1 | 376 473 | -0.19 | -0.37 | 0.01 | 0.00 | -0.26 | 0.41 | -0.10 | 0.02 | 0.49 | -0.10 | -0.10 | -0.07 | 0.42 | -0.03 | 0.03 | -0.26 | 0.20 | -0.17 | 0.05 | 0.02 |
| 1 | 390 391 | 0.01 | -0.15 | -0.06 | 0.04 | -0.03 | 0.53 | -0.34 | 0.10 | 0.22 | -0.40 | 0.16 | 0.19 | -0.27 | -0.05 | 0.06 | 0.33 | 0.01 | -0.33 | -0.03 | 0.04 |
| 1 | 392 393 | -0.09 | -0.04 | -0.31 | -0.17 | 0.49 | -0.16 | 0.39 | 0.02 | 0.15 | -0.16 | -0.03 | -0.03 | -0.02 | -0.22 | -0.18 | -0.10 | 0.13 | 0.47 | 0.08 | -0.23 |
| 1 | 394 431 | 0.05 | -0.26 | 0.07 | 0.11 | -0.22 | -0.10 | -0.17 | 0.27 | -0.46 | 0.08 | 0.31 | -0.44 | 0.13 | 0.29 | -0.02 | 0.24 | 0.12 | -0.20 | 0.19 | 0.01 |
| 1 | 397 549 | 0.07 | -0.01 | 0.07 | 0.00 | 0.07 | 0.07 | 0.00 | 0.07 | -0.01 | -0.96 | 0.07 | 0.00 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 |
| 1 | 402 417 | 0.20 | -0.17 | 0.18 | 0.34 | -0.02 | -0.17 | 0.08 | 0.26 | 0.02 | -0.24 | -0.26 | -0.41 | -0.15 | -0.11 | 0.08 | 0.41 | 0.31 | 0.02 | -0.27 | -0.10 |
| 1 | 407 411 | -0.06 | -0.08 | -0.11 | -0.03 | 0.21 | -0.09 | 0.19 | 0.35 | -0.33 | 0.18 | 0.15 | -0.39 | -0.18 | 0.30 | -0.02 | -0.06 | -0.02 | 0.20 | -0.46 | 0.26 |
| 1 | 409 414 | 0.03 | 0.16 | 0.00 | -0.02 | 0.16 | -0.26 | -0.06 | -0.09 | -0.48 | 0.20 | -0.12 | -0.35 | 0.36 | 0.28 | 0.02 | -0.08 | 0.16 | -0.35 | 0.17 | 0.27 |

Table B.2: (continued)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 429 430 | -0.38 | -0.15 | 0.28 | 0.52 | 0.30 | -0.17 | -0.28 | -0.24 | 0.08 | 0.26 | 0.13 | -0.08 | -0.22 | 0.11 | 0.08 | -0.22 | -0.01 | -0.09 | -0.01 | 0.11 |
| 1 | 439 440 | -0.07 | -0.42 | 0.11 | -0.28 | 0.12 | 0.27 | -0.22 | 0.12 | 0.30 | -0.13 | 0.45 | -0.02 | -0.10 | -0.13 | 0.10 | -0.04 | 0.04 | -0.10 | -0.32 | 0.33 |
| 1 | 447 448 | 0.19 | -0.20 | 0.00 | -0.39 | -0.21 | -0.19 | -0.44 | 0.18 | 0.18 | 0.21 | 0.23 | 0.27 | 0.10 | 0.17 | 0.16 | 0.05 | -0.39 | -0.08 | 0.17 | -0.01 |
| 1 | 460 607 | -0.07 | -0.03 | -0.36 | -0.04 | -0.13 | -0.03 | 0.01 | -0.41 | 0.05 | 0.07 | 0.37 | 0.41 | -0.14 | 0.09 | -0.16 | 0.20 | -0.26 | -0.16 | 0.35 | 0.23 |
| 1 | 462 463 | 0.48 | -0.18 | 0.13 | -0.07 | 0.05 | 0.16 | -0.20 | 0.25 | 0.05 | -0.02 | 0.01 | -0.33 | 0.18 | -0.10 | -0.16 | 0.27 | 0.20 | -0.01 | -0.20 | -0.50 |
| 1 | 482 484 | -0.06 | -0.45 | 0.14 | 0.38 | 0.21 | -0.25 | -0.34 | 0.13 | 0.23 | 0.09 | -0.12 | -0.26 | 0.11 | -0.29 | 0.22 | 0.00 | 0.28 | 0.09 | -0.11 | 0.00 |
| 1 | 483 499 | 0.32 | -0.31 | 0.01 | 0.21 | 0.19 | -0.15 | -0.10 | 0.03 | -0.12 | -0.34 | -0.26 | -0.44 | 0.18 | 0.03 | 0.46 | 0.05 | 0.09 | 0.18 | -0.06 | 0.03 |
| 1 | 501 502 | -0.12 | -0.03 | 0.24 | 0.51 | -0.12 | 0.10 | 0.35 | -0.03 | -0.02 | -0.25 | -0.20 | -0.44 | -0.23 | 0.04 | -0.15 | 0.12 | 0.06 | 0.31 | 0.05 | -0.19 |
| 1 | 513 570 | -0.13 | -0.30 | -0.08 | 0.15 | 0.21 | -0.28 | 0.38 | 0.03 | 0.08 | -0.12 | -0.05 | 0.00 | 0.13 | 0.47 | -0.03 | -0.49 | 0.20 | -0.10 | 0.11 | -0.18 |
| 1 | 531 533 | -0.10 | -0.45 | 0.00 | 0.14 | -0.09 | -0.09 | 0.48 | -0.27 | -0.29 | 0.19 | -0.05 | -0.31 | 0.10 | 0.24 | 0.05 | -0.21 | 0.08 | 0.17 | 0.19 | 0.22 |
| 1 | 582 585 | 0.23 | -0.51 | 0.03 | 0.24 | 0.41 | 0.01 | 0.17 | 0.34 | -0.11 | -0.24 | -0.17 | -0.07 | -0.02 | 0.02 | 0.10 | 0.13 | 0.09 | -0.30 | -0.28 | -0.07 |
| 1 | 583 584 | 0.14 | -0.43 | -0.22 | 0.11 | 0.32 | -0.21 | 0.12 | -0.12 | -0.28 | 0.12 | 0.11 | -0.07 | 0.24 | -0.21 | 0.06 | 0.30 | 0.21 | -0.44 | 0.11 | 0.12 |
| 0.9999 | 78 396 553 | -0.09 | 0.15 | -0.07 | 0.05 | -0.05 | -0.26 | -0.17 | 0.12 | 0.53 | 0.32 | -0.47 | -0.23 | -0.05 | -0.13 | -0.02 | 0.14 | 0.33 | -0.19 | 0.05 | 0.04 |
| 0.9989 | 169 474 548 | -0.15 | -0.40 | -0.11 | -0.03 | -0.02 | 0.03 | 0.36 | 0.14 | -0.17 | -0.13 | 0.21 | 0.33 | 0.30 | -0.11 | 0.30 | -0.20 | -0.23 | -0.19 | -0.22 | 0.30 |
| 0.9959 | 72 535 552 586 | -0.19 | 0.14 | -0.01 | 0.03 | 0.55 | 0.03 | 0.07 | -0.24 | -0.10 | -0.05 | -0.05 | 0.05 | 0.65 | -0.22 | -0.16 | -0.07 | -0.02 | -0.20 | -0.10 | -0.10 |
| 0.9953 | 191 193 195 | 0.10 | -0.03 | -0.47 | -0.09 | 0.06 | 0.06 | 0.09 | -0.09 | 0.01 | -0.29 | 0.23 | -0.14 | 0.39 | -0.08 | 0.18 | 0.03 | 0.32 | 0.28 | -0.42 | -0.15 |

Table B.3: PCA Generated Amino Acid Indices With Single Linkage 0.65

| Variance | Indices | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 17 | -0.19 | -0.11 | 0.08 | 0.09 | 0.13 | -0.16 | -0.20 | -0.41 | -0.01 | 0.51 | 0.44 | -0.18 | -0.10 | 0.02 | -0.02 | -0.10 | -0.20 | 0.04 | -0.01 | 0.37 |
| 1 | 2 132 | 0.17 | -0.19 | 0.23 | 0.36 | -0.03 | -0.07 | 0.09 | -0.03 | 0.17 | -0.21 | -0.23 | 0.37 | -0.42 | 0.04 | 0.00 | -0.13 | -0.13 | 0.38 | -0.03 | -0.34 |
| 1 | 3 4 | -0.33 | -0.16 | 0.05 | 0.01 | 0.66 | 0.27 | -0.03 | -0.05 | 0.10 | -0.19 | 0.24 | 0.01 | -0.31 | -0.06 | 0.06 | 0.21 | -0.01 | -0.24 | -0.21 | -0.01 |
| 1 | 5 564 | -0.15 | -0.26 | -0.06 | -0.26 | 0.33 | 0.31 | 0.27 | -0.06 | -0.15 | -0.02 | 0.26 | 0.15 | -0.58 | 0.16 | -0.09 | 0.16 | 0.02 | 0.12 | -0.19 | 0.03 |
| 1 | 8 598 | 0.10 | 0.11 | -0.23 | -0.05 | 0.16 | -0.20 | 0.22 | -0.23 | -0.35 | -0.16 | 0.25 | 0.27 | 0.19 | -0.43 | 0.09 | 0.23 | -0.32 | 0.20 | -0.06 | 0.20 |
| 1 | 9 109 | -0.46 | 0.04 | 0.03 | 0.51 | 0.09 | 0.01 | -0.04 | -0.25 | -0.03 | 0.22 | 0.22 | -0.23 | 0.00 | 0.11 | -0.12 | 0.37 | -0.13 | -0.33 | 0.07 | -0.09 |
| 1 | 10 319 | -0.21 | -0.30 | 0.24 | -0.15 | 0.01 | -0.11 | 0.20 | -0.35 | 0.16 | 0.13 | 0.07 | 0.04 | 0.38 | -0.35 | 0.28 | 0.16 | 0.12 | -0.38 | -0.09 | 0.15 |
| 1 | 12 13 | 0.23 | -0.48 | -0.18 | 0.02 | 0.45 | -0.14 | 0.08 | 0.10 | -0.29 | -0.36 | 0.29 | -0.06 | 0.10 | 0.25 | -0.05 | -0.01 | -0.01 | -0.19 | 0.14 | 0.10 |
| 1 | 15 59 | 0.17 | -0.42 | -0.04 | 0.17 | -0.24 | -0.04 | 0.02 | 0.18 | -0.02 | 0.27 | 0.03 | 0.06 | 0.04 | -0.18 | -0.05 | 0.43 | 0.11 | -0.28 | -0.47 | 0.25 |
| 1 | 21 79 | -0.08 | -0.03 | 0.35 | 0.33 | -0.11 | -0.07 | -0.09 | -0.23 | 0.10 | 0.21 | 0.19 | -0.20 | 0.04 | -0.03 | 0.56 | -0.10 | -0.31 | -0.27 | -0.03 | -0.23 |
| 1 | 22 80 | 0.18 | -0.08 | -0.02 | -0.41 | 0.37 | -0.10 | -0.49 | 0.26 | 0.17 | -0.03 | -0.03 | -0.14 | 0.22 | 0.14 | 0.15 | -0.09 | -0.18 | 0.30 | -0.27 | 0.05 |
| 1 | 24 560 | 0.04 | -0.97 | 0.07 | 0.06 | 0.06 | 0.05 | 0.05 | 0.07 | 0.05 | 0.04 | 0.04 | 0.05 | 0.04 | 0.04 | 0.07 | 0.06 | 0.05 | 0.04 | 0.05 | 0.04 |
| 1 | 35 505 | -0.03 | -0.04 | -0.11 | -0.04 | -0.41 | -0.04 | 0.11 | 0.04 | 0.08 | 0.20 | -0.34 | 0.10 | 0.57 | -0.22 | 0.36 | 0.11 | 0.10 | -0.30 | -0.13 | 0.00 |
| 1 | 37 227 | 0.12 | 0.11 | 0.30 | -0.05 | -0.72 | 0.17 | 0.05 | 0.06 | 0.02 | -0.19 | 0.19 | 0.23 | -0.24 | -0.27 | -6e-05 | -0.11 | 0.05 | 0.23 | 0.08 | -0.04 |
| 1 | 38 223 | -0.37 | 0.25 | 0.22 | -0.19 | -0.08 | 0.63 | -0.04 | -0.02 | 0.25 | -0.31 | 0.07 | -0.11 | 0.17 | -0.12 | -0.11 | -0.06 | 0.08 | -0.26 | 0.05 | -0.05 |
| 1 | 39 225 | 0.22 | -0.43 | -0.14 | 0.56 | -0.12 | -0.33 | 0.18 | 0.24 | 0.01 | 0.07 | -0.08 | 0.06 | -0.23 | 0.16 | -0.22 | -0.10 | 0.09 | -0.16 | 0.04 | 0.18 |
| 1 | 40 53 | -0.03 | -0.14 | 0.07 | 0.14 | 0.00 | 0.18 | 0.24 | 0.21 | 0.31 | -0.06 | -0.12 | 0.22 | -0.16 | -0.23 | -0.72 | 0.07 | -0.07 | -0.09 | 0.19 | -0.01 |
| 1 | 41 565 | 0.19 | -0.13 | 0.13 | 0.17 | 0.10 | -0.05 | -0.58 | -0.24 | -0.28 | -0.03 | -0.05 | -0.14 | -0.02 | 0.52 | 0.26 | 0.06 | 0.21 | -0.11 | 0.01 | -0.02 |
| 1 | 42 566 | 0.06 | -0.25 | -0.38 | 0.02 | -0.36 | -0.33 | 0.05 | -0.05 | -0.27 | 0.25 | 0.15 | 0.38 | 0.25 | 0.03 | -0.15 | 0.12 | 0.28 | -0.13 | 0.20 | 0.14 |
| 1 | 50 366 | 0.39 | -0.49 | -0.38 | -0.20 | 0.02 | 0.24 | 0.25 | 0.05 | 0.01 | 0.16 | 0.26 | -0.26 | -0.15 | -0.22 | 0.12 | -0.07 | -0.03 | 0.22 | 0.05 | 0.03 |
| 1 | 51 123 | -0.02 | -0.33 | -0.06 | 0.10 | -0.74 | 0.30 | -0.19 | 0.20 | 0.13 | 0.00 | 0.18 | 0.02 | 0.08 | 0.15 | 0.06 | -0.16 | 0.11 | -0.08 | 0.13 | 0.12 |
| 1 | 52 346 | 0.09 | -0.13 | 0.12 | 0.01 | 0.02 | -0.26 | 0.23 | 0.08 | -0.28 | -0.30 | 0.13 | -0.05 | -0.18 | -0.43 | 0.47 | 0.06 | -0.22 | 0.19 | 0.30 | 0.14 |
| 1 | 56 58 | 0.33 | -0.09 | -0.12 | -0.04 | -0.07 | 0.06 | 0.31 | -0.01 | -0.09 | 0.34 | 0.04 | -0.14 | 0.20 | -0.33 | 0.03 | 0.37 | -0.41 | -0.39 | -0.06 | 0.08 |
| 1 | 64 137 | -0.13 | -0.11 | 0.39 | -0.36 | 0.07 | -0.01 | -0.47 | 0.17 | -0.02 | 0.24 | -0.17 | 0.30 | -0.01 | 0.11 | -0.43 | 0.18 | -0.04 | -0.01 | 0.15 | 0.13 |
| 1 | 65 135 | -0.04 | -0.23 | 0.38 | 0.32 | -0.15 | 0.16 | 0.43 | 0.16 | -0.37 | -0.16 | -0.05 | -0.15 | 0.00 | 0.03 | 0.11 | -0.07 | -0.22 | 0.18 | 0.05 | -0.39 |
| 1 | 67 603 | -0.21 | 0.15 | 0.25 | 0.01 | 0.58 | -0.13 | -0.30 | -0.10 | 0.43 | -0.01 | -0.21 | 0.16 | 0.18 | -0.19 | -0.20 | -0.09 | -0.22 | 0.01 | -0.02 | -0.10 |
| 1 | 68 534 | 0.22 | -0.29 | 0.45 | -0.04 | 0.24 | -0.29 | -0.30 | 0.11 | 0.17 | -0.33 | 0.19 | 0.25 | -0.15 | -0.19 | 0.01 | 0.15 | -0.22 | 0.08 | 0.12 | -0.18 |
| 1 | 72 552 | 0.01 | -0.02 | 0.25 | 0.46 | -0.19 | 0.13 | 0.34 | 0.03 | 0.14 | -0.26 | -0.26 | -0.27 | -0.43 | -0.03 | -0.06 | 0.21 | 0.09 | -0.05 | 0.16 | -0.24 |
| 1 | 76 133 | -0.08 | -0.55 | 0.16 | 0.43 | -0.13 | 0.05 | 0.22 | -0.10 | 0.31 | -0.09 | -0.09 | 0.08 | -0.03 | -0.38 | 0.26 | 0.06 | 0.12 | -0.21 | 0.03 | -0.06 |
| 1 | 77 590 | -0.04 | 0.00 | 0.00 | 0.00 | 0.96 | -0.02 | -0.01 | -0.03 | -0.04 | -0.11 | -0.11 | 0.01 | -0.09 | -0.11 | -0.06 | -0.04 | -0.04 | -0.13 | -0.07 | -0.08 |
| 1 | 85 110 | -0.07 | -0.01 | -0.21 | 0.45 | -0.47 | -0.06 | 0.08 | -0.37 | 0.33 | -0.07 | -0.07 | -0.16 | 0.32 | 0.24 | -0.19 | 0.13 | -0.05 | -0.06 | 0.14 | 0.09 |
| 1 | 89 324 | 0.14 | -0.22 | -0.04 | -0.51 | -0.10 | 0.43 | 0.14 | -0.10 | -0.04 | -0.22 | -0.10 | 0.02 | 0.14 | -0.10 | -0.22 | -0.22 | 0.14 | 0.38 | 0.26 | 0.20 |
| 1 | 92 450 | 0.03 | -0.29 | 0.33 | -0.30 | 0.00 | -0.09 | 0.37 | -0.05 | -0.09 | -0.23 | 0.04 | -0.28 | 0.02 | 0.04 | -0.11 | 0.42 | 0.44 | -0.03 | -0.03 | -0.22 |
| 1 | 93 421 | -0.47 | -0.27 | 0.31 | 0.00 | 0.13 | 0.02 | -0.14 | 0.29 | 0.32 | 0.13 | -0.30 | -0.24 | -0.17 | -0.07 | -0.14 | 0.00 | 0.26 | -0.11 | 0.20 | 0.23 |
| 1 | 98 228 | 0.14 | -0.48 | 0.26 | -0.24 | 0.10 | 0.26 | -0.09 | 0.15 | -0.01 | -0.06 | -0.07 | -0.03 | -0.11 | 0.00 | 0.49 | -0.06 | 0.34 | -0.16 | -0.31 | -0.12 |
| 1 | 100 230 | 0.00 | 0.21 | -0.18 | -0.13 | 0.24 | 0.47 | -0.47 | 0.00 | -0.17 | -0.28 | 0.22 | -0.05 | 0.11 | -0.23 | 0.03 | 0.05 | 0.32 | -0.09 | -0.19 | 0.16 |
| 1 | 104 107 | -0.08 | -0.08 | 0.58 | 0.36 | 0.00 | 0.29 | -0.27 | -0.02 | -0.19 | 0.26 | 0.08 | -0.26 | 0.02 | 0.07 | -0.27 | -0.17 | -0.17 | 0.02 | -0.20 | 0.02 |
| 1 | 105 234 | 0.02 | -0.33 | 0.17 | 0.24 | 0.18 | -0.13 | 0.21 | 0.11 | -0.25 | -0.11 | -0.25 | 0.13 | 0.31 | 0.37 | -0.44 | 0.03 | -0.03 | 0.18 | -0.21 | -0.19 |
| 1 | 106 428 | -0.14 | -0.15 | 0.21 | -0.13 | 0.49 | -0.44 | 0.22 | 0.06 | -0.11 | 0.31 | -0.28 | -0.23 | -0.12 | 0.02 | 0.03 | 0.08 | 0.17 | -0.09 | -0.17 | 0.28 |
| 1 | 111 385 | -0.09 | -0.12 | -0.16 | 0.41 | 0.20 | -0.28 | 0.46 | -0.06 | -0.32 | 0.09 | 0.18 | -0.03 | 0.07 | 0.13 | -0.19 | -0.26 | -0.37 | 0.17 | 0.06 | 0.11 |
| 1 | 115 153 | -0.08 | -0.16 | -0.09 | 0.19 | -0.07 | -0.09 | 0.19 | -0.09 | -0.08 | -0.05 | -0.05 | -0.16 | -0.06 | -0.03 | 0.90 | -0.10 | -0.08 | -0.01 | -0.03 | -0.05 |
| 1 | 118 588 | -0.04 | -0.12 | 0.13 | 0.26 | -0.33 | 0.21 | 0.33 | 0.21 | -0.21 | -0.12 | -0.11 | -0.07 | 0.53 | 0.16 | -0.29 | 0.11 | -0.20 | -0.15 | -0.08 | -0.21 |
| 1 | 119 170 | 0.15 | -0.10 | 0.02 | 0.29 | -0.14 | 0.08 | -0.39 | 0.15 | 0.17 | 0.08 | -0.27 | -0.14 | 0.21 | 0.02 | -0.57 | -0.04 | 0.35 | 0.13 | 0.16 | -0.15 |
| 1 | 120 171 | -0.04 | -0.10 | -0.10 | -0.03 | -0.45 | -0.02 | -0.27 | 0.72 | 0.06 | -0.03 | -0.06 | -0.02 | 0.13 | 0.12 | 0.30 | -0.08 | -0.03 | -0.19 | 0.08 | 0.02 |
| 1 | 124 175 | -0.11 | 0.12 | 0.03 | 0.12 | 0.06 | 0.04 | -0.13 | 0.40 | -0.40 | -0.30 | -0.38 | -0.20 | -0.08 | -0.10 | 0.01 | 0.37 | 0.28 | 0.09 | 0.29 | -0.14 |
| 1 | 127 508 | 0.26 | -0.41 | -0.01 | 0.15 | -0.20 | -0.09 | -0.11 | 0.43 | -0.22 | -0.20 | -0.22 | 0.15 | -0.26 | 0.10 | 0.20 | 0.27 | 0.32 | -0.09 | 0.10 | -0.17 |
| 1 | 128 563 | -0.20 | 0.24 | -0.16 | -0.06 | -0.17 | -0.10 | -0.24 | -0.25 | 0.28 | 0.22 | 0.14 | 0.15 | 0.03 | 0.32 | -0.12 | -0.34 | -0.05 | 0.06 | 0.50 | -0.23 |
| 1 | 134 454 | 0.17 | 0.11 | -0.13 | 0.00 | -0.29 | -0.02 | 0.03 | 0.61 | 0.16 | 0.23 | -0.24 | 0.13 | 0.28 | -0.02 | -0.33 | -0.29 | -0.20 | -0.08 | -0.04 | -0.08 |

Table B.3: (continued)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 139 608 | 0.08 | -0.26 | -0.28 | 0.13 | 0.15 | 0.32 | -0.48 | 0.41 | 0.08 | -0.21 | 0.21 | -0.07 | -0.26 | -0.19 | 0.22 | -0.06 | 0.12 | -0.07 | 0.17 | -0.01 |
| 1 | 140 408 | 0.04 | -0.26 | -0.08 | 0.38 | 0.04 | -0.33 | 0.05 | 0.05 | 0.38 | -0.23 | 0.49 | 0.11 | -0.09 | -0.18 | -0.05 | -0.13 | 0.08 | 0.16 | -0.35 | -0.07 |
| 1 | 151 481 | 0.57 | 0.41 | -0.09 | 0.00 | -0.02 | 0.07 | -0.35 | 0.15 | -0.08 | 0.00 | -0.01 | 0.09 | -0.03 | -0.02 | 0.19 | -0.32 | -0.06 | -0.44 | -0.06 | 0.00 |
| 1 | 154 157 | 0.38 | -0.12 | 0.12 | 0.19 | -0.08 | 0.10 | 0.11 | -0.18 | 0.27 | -0.44 | -0.01 | -0.37 | -0.41 | 0.22 | -0.11 | 0.20 | -0.06 | 0.19 | 0.10 | -0.12 |
| 1 | 156 432 | 0.21 | -0.15 | -0.34 | -0.10 | 0.00 | -0.31 | -0.13 | 0.08 | 0.15 | 0.44 | -0.04 | -0.17 | -0.13 | 0.16 | -0.29 | 0.14 | 0.34 | -0.07 | -0.15 | 0.39 |
| 1 | 161 253 | -0.32 | -0.04 | 0.01 | 0.03 | 0.23 | 0.25 | -0.30 | 0.05 | -0.33 | -0.08 | -0.35 | -0.03 | 0.46 | 0.20 | 0.09 | -0.26 | 0.26 | 0.22 | -0.02 | -0.08 |
| 1 | 162 222 | -0.01 | -0.28 | -0.29 | -0.02 | 0.01 | -0.31 | 0.09 | 0.05 | 0.23 | -0.35 | 0.05 | -0.02 | -0.29 | 0.46 | 0.22 | 0.09 | 0.03 | 0.39 | -0.17 | 0.13 |
| 1 | 163 224 | 0.29 | -0.24 | -0.31 | -0.02 | -0.34 | -0.10 | 0.41 | -0.18 | 0.06 | -0.02 | 0.08 | 0.07 | -0.09 | 0.39 | -0.25 | 0.04 | 0.38 | 0.10 | -0.19 | -0.06 |
| 1 | 166 275 | 0.01 | -0.31 | 0.07 | 0.17 | -0.02 | -0.11 | -0.13 | 0.02 | -0.10 | -0.02 | 0.14 | 0.03 | -0.24 | -0.34 | 0.33 | -0.12 | -0.17 | 0.35 | -0.15 | 0.59 |
| 1 | 174 371 | -0.16 | -0.43 | 0.14 | 0.22 | 0.26 | -0.01 | -0.02 | -0.15 | -0.15 | -0.04 | -0.09 | -0.16 | 0.59 | -0.12 | 0.09 | 0.24 | -0.01 | -0.23 | 0.23 | -0.21 |
| 1 | 176 555 | 0.95 | -0.02 | -0.08 | -0.09 | 0.03 | -0.06 | -0.06 | -0.15 | -0.04 | -0.05 | -0.05 | -0.04 | -0.04 | 0.00 | -0.09 | -0.12 | -0.09 | 0.06 | 0.01 | -0.08 |
| 1 | 179 180 | 0.06 | -0.38 | 0.25 | 0.30 | -0.39 | 0.21 | 0.18 | 0.16 | -0.36 | -0.08 | 0.02 | -0.29 | -0.15 | 0.25 | -0.15 | 0.04 | 0.30 | 0.05 | 0.10 | -0.11 |
| 1 | 181 295 | 0.03 | -0.24 | 0.00 | -0.25 | 0.24 | 0.18 | -0.44 | -0.03 | -0.05 | -0.12 | -0.19 | -0.23 | -0.29 | 0.06 | 0.41 | 0.27 | 0.12 | 0.27 | 0.00 | 0.26 |
| 1 | 182 296 | 0.03 | -0.03 | 0.15 | 0.25 | -0.29 | -0.19 | 0.39 | -0.32 | 0.11 | 0.06 | 0.15 | 0.28 | 0.41 | -0.07 | -0.37 | -0.27 | -0.10 | 0.01 | -0.07 | -0.14 |
| 1 | 184 593 | 0.21 | -0.01 | -0.02 | -0.07 | 0.68 | 0.03 | -0.15 | 0.00 | -0.09 | -0.02 | 0.10 | 0.17 | -0.01 | -0.21 | -0.23 | 0.12 | 0.05 | -0.42 | -0.31 | 0.17 |
| 1 | 188 601 | 0.38 | -0.25 | 0.10 | 0.21 | -0.11 | 0.04 | 0.11 | 0.37 | 0.08 | -0.41 | -0.31 | -0.47 | -0.11 | 0.11 | 0.20 | 0.04 | 0.02 | 0.00 | 0.09 | -0.11 |
| 1 | 190 192 | 0.04 | -0.06 | -0.44 | -0.08 | 0.21 | 0.11 | 0.09 | -0.21 | 0.13 | -0.34 | 0.13 | -0.23 | 0.33 | -0.03 | 0.26 | -0.04 | 0.32 | 0.29 | -0.24 | -0.23 |
| 1 | 191 193 | -0.14 | -0.07 | 0.49 | -0.01 | -0.11 | -0.12 | -0.18 | 0.06 | -0.02 | 0.38 | -0.11 | 0.07 | -0.28 | 0.17 | -0.20 | -0.02 | -0.29 | -0.23 | 0.47 | 0.12 |
| 1 | 196 455 | 0.32 | -0.22 | 0.04 | 0.06 | -0.25 | 0.08 | 0.08 | 0.48 | 0.04 | -0.14 | -0.20 | -0.31 | 0.18 | 0.25 | -0.12 | -0.44 | 0.00 | 0.24 | 0.04 | -0.14 |
| 1 | 198 208 | -0.26 | -0.13 | 0.26 | 0.02 | 0.29 | -0.03 | 0.03 | -0.63 | -0.05 | 0.48 | -0.11 | -0.02 | -0.02 | 0.21 | -0.01 | 0.03 | -0.20 | -0.06 | 0.13 | 0.09 |
| 1 | 201 457 | 0.10 | -0.51 | 0.12 | 0.41 | -0.17 | 0.01 | 0.36 | 0.14 | 0.15 | -0.07 | 0.29 | -0.43 | -0.06 | 0.00 | -0.16 | -0.06 | -0.18 | 0.01 | 0.02 | 0.04 |
| 1 | 205 465 | -0.07 | -0.19 | 0.00 | -0.15 | 0.21 | -0.09 | -0.16 | 0.31 | 0.03 | 0.55 | -0.25 | 0.05 | 0.10 | -0.35 | 0.03 | -0.37 | -0.10 | 0.32 | 0.02 | 0.10 |
| 1 | 209 247 | -0.30 | -0.10 | -0.02 | 0.11 | -0.15 | -0.01 | -0.18 | 0.19 | -0.44 | 0.17 | 0.21 | -0.06 | -0.44 | 0.03 | 0.45 | 0.21 | 0.14 | -0.07 | 0.02 | 0.24 |
| 1 | 212 529 | -0.13 | -0.13 | 0.03 | 0.14 | -0.08 | 0.25 | -0.26 | -0.36 | -0.35 | 0.25 | -0.06 | 0.09 | -0.41 | 0.46 | 0.09 | 0.24 | -0.04 | 0.01 | 0.08 | 0.19 |
| 1 | 215 216 | -0.12 | -0.17 | 0.22 | 0.24 | -0.17 | 0.24 | 0.37 | -0.36 | 0.33 | -0.29 | 0.06 | 0.33 | 0.03 | -0.06 | 0.05 | -0.03 | 0.01 | -0.35 | -0.06 | -0.25 |
| 1 | 233 252 | -0.16 | -0.06 | 0.18 | -0.07 | -0.58 | 0.33 | 0.30 | 0.04 | -0.15 | 0.24 | -0.04 | -0.14 | 0.32 | 0.26 | -0.23 | -0.05 | 0.08 | -0.20 | 0.10 | -0.16 |
| 1 | 236 347 | -0.26 | 0.11 | 0.28 | 0.13 | -0.07 | -0.15 | -0.06 | -0.04 | 0.14 | 0.31 | -0.01 | -0.46 | -0.31 | 0.18 | 0.08 | -0.06 | 0.42 | 0.05 | -0.39 | 0.08 |
| 1 | 238 387 | 0.13 | -0.38 | -0.10 | 0.18 | -0.14 | 0.06 | 0.04 | 0.00 | -0.53 | 0.05 | -0.18 | -0.17 | 0.01 | 0.03 | 0.50 | 0.19 | 0.31 | -0.13 | 0.01 | 0.14 |
| 1 | 240 241 | -0.01 | -0.04 | 0.08 | 0.04 | -0.52 | -0.13 | 0.07 | 0.01 | 0.37 | 0.28 | 0.18 | 0.33 | -0.04 | -0.16 | -0.20 | -0.42 | -0.14 | -0.09 | 0.16 | 0.21 |
| 1 | 251 527 | 0.04 | -0.15 | -0.08 | 0.06 | 0.25 | -0.32 | 0.18 | 0.13 | -0.35 | 0.22 | -0.15 | 0.27 | -0.24 | 0.12 | 0.32 | -0.08 | -0.18 | -0.42 | 0.26 | 0.16 |
| 1 | 264 265 | -0.27 | -0.16 | 0.02 | 0.02 | 0.23 | -0.44 | 0.30 | -0.03 | -0.25 | 0.22 | -0.03 | -0.11 | 0.16 | 0.09 | -0.35 | 0.26 | 0.30 | -0.20 | -0.06 | 0.29 |
| 1 | 269 270 | -0.21 | -0.18 | -0.11 | 0.29 | 0.06 | 0.40 | -0.17 | 0.24 | 0.12 | -0.20 | 0.08 | -0.01 | 0.26 | -0.04 | -0.55 | 0.06 | 0.07 | 0.14 | -0.33 | 0.09 |
| 1 | 271 281 | -0.16 | 0.18 | 0.06 | -0.27 | -0.20 | 0.24 | 0.00 | 0.32 | -0.42 | 0.24 | -0.13 | 0.51 | 0.03 | -0.09 | -0.21 | 0.12 | -0.03 | 0.16 | -0.19 | -0.16 |
| 1 | 279 280 | -0.33 | -0.14 | 0.10 | 0.07 | -0.06 | -0.32 | -0.16 | 0.10 | 0.12 | 0.47 | -0.14 | 0.07 | -0.10 | 0.06 | -0.27 | -0.10 | -0.12 | -0.03 | 0.28 | 0.50 |
| 1 | 282 361 | -0.02 | -0.28 | -0.18 | 0.34 | -0.38 | 0.31 | -0.09 | 0.04 | -0.07 | 0.00 | 0.11 | 0.50 | 0.01 | -0.13 | -0.22 | -0.01 | -0.09 | 0.36 | 0.01 | -0.25 |
| 1 | 289 290 | -0.03 | -0.34 | -0.01 | -0.26 | -0.26 | -0.13 | 0.01 | 0.42 | 0.26 | 0.08 | 0.35 | 0.43 | 0.09 | -0.27 | 0.13 | -0.10 | -0.15 | -0.21 | 0.01 | -0.01 |
| 1 | 292 293 | 0.04 | -0.18 | 0.13 | -0.25 | 0.09 | 0.14 | 0.01 | 0.22 | -0.07 | 0.46 | -0.24 | -0.31 | 0.26 | -0.23 | -0.17 | -0.28 | -0.07 | -0.16 | 0.30 | 0.31 |
| 1 | 315 556 | 0.17 | -0.10 | -0.19 | -0.28 | 0.19 | 0.08 | -0.11 | 0.22 | -0.29 | -0.12 | 0.17 | 0.55 | 0.11 | -0.34 | 0.32 | -0.07 | -0.06 | -0.24 | -0.11 | 0.10 |
| 1 | 335 364 | 0.10 | 0.00 | -0.47 | -0.06 | 0.03 | 0.21 | -0.01 | -0.11 | 0.49 | 0.24 | 0.29 | -0.06 | 0.00 | -0.12 | 0.24 | -0.17 | -0.09 | -0.36 | 0.12 | -0.26 |
| 1 | 340 341 | -0.05 | -0.03 | -0.16 | -0.18 | 0.22 | 0.06 | 0.31 | -0.22 | -0.04 | 0.01 | -0.19 | -0.21 | -0.19 | -0.02 | 0.69 | 0.02 | -0.06 | -0.29 | 0.20 | 0.13 |
| 1 | 343 348 | -0.09 | -0.25 | -0.21 | 0.03 | 0.38 | 0.08 | 0.13 | 0.14 | 0.28 | -0.18 | -0.26 | -0.01 | 0.13 | 0.25 | -0.46 | -0.30 | 0.13 | 0.31 | -0.13 | 0.00 |
| 1 | 350 573 | 0.18 | -0.32 | -0.12 | -0.11 | 0.32 | 0.04 | -0.18 | -0.02 | -0.39 | 0.51 | 0.02 | -0.03 | 0.31 | -0.16 | 0.10 | -0.02 | -0.08 | -0.33 | 0.14 | 0.15 |
| 1 | 352 479 | 0.21 | -0.11 | -0.17 | -0.07 | -0.41 | -0.05 | 0.41 | -0.16 | -0.08 | -0.13 | 0.09 | -0.22 | 0.03 | -0.09 | 0.18 | -0.15 | 0.06 | 0.40 | 0.45 | -0.18 |
| 1 | 362 515 | 0.24 | -0.17 | -0.16 | -0.11 | 0.16 | -0.18 | -0.14 | 0.22 | -0.26 | 0.34 | 0.18 | -0.06 | 0.06 | -0.24 | 0.21 | 0.21 | 0.15 | -0.26 | -0.47 | 0.27 |
| 1 | 376 473 | -0.19 | -0.37 | 0.01 | 0.00 | -0.26 | 0.41 | -0.10 | 0.02 | 0.49 | -0.10 | -0.10 | -0.07 | 0.42 | -0.03 | 0.03 | -0.26 | 0.20 | -0.17 | 0.05 | 0.02 |
| 1 | 386 504 | -0.07 | -0.09 | -0.08 | 0.45 | -0.06 | -0.08 | -0.09 | -0.06 | -0.08 | -0.06 | -0.06 | -0.07 | -0.06 | -0.07 | 0.84 | -0.07 | -0.07 | -0.07 | -0.08 | -0.07 |
| 1 | 390 391 | 0.01 | -0.15 | -0.06 | 0.04 | -0.03 | 0.53 | -0.34 | 0.10 | 0.22 | -0.40 | 0.16 | 0.19 | -0.27 | -0.05 | 0.06 | 0.33 | 0.01 | -0.33 | -0.03 | 0.04 |
| 1 | 392 393 | -0.09 | -0.04 | -0.31 | -0.17 | 0.49 | -0.16 | 0.39 | 0.02 | 0.15 | -0.16 | -0.03 | -0.03 | -0.02 | -0.22 | -0.18 | -0.10 | 0.13 | 0.47 | 0.08 | -0.23 |
| 1 | 394 431 | 0.05 | -0.26 | 0.07 | 0.11 | -0.22 | -0.10 | -0.17 | 0.27 | -0.46 | 0.08 | 0.31 | -0.44 | 0.13 | 0.29 | -0.02 | 0.24 | 0.12 | -0.20 | 0.19 | 0.01 |

Table B.3: (continued)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 396 553 | 0.00 | -0.03 | 0.24 | 0.00 | -0.02 | -0.65 | -0.02 | 0.09 | 0.23 | -0.10 | -0.10 | 0.08 | -0.05 | -0.08 | -0.06 | 0.02 | -0.04 | 0.64 | -0.06 | -0.10 |
| 1 | 397 549 | 0.07 | -0.01 | 0.07 | 0.00 | 0.07 | 0.07 | 0.00 | 0.07 | -0.01 | -0.96 | 0.07 | 0.00 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 |
| 1 | 398 550 | 0.06 | -0.11 | 0.08 | 0.17 | 0.09 | 0.07 | -0.94 | 0.06 | 0.00 | 0.06 | 0.06 | -0.08 | 0.06 | 0.07 | 0.05 | 0.07 | 0.07 | 0.06 | 0.07 | 0.06 |
| 1 | 399 554 | -0.07 | 0.00 | -0.01 | 0.02 | 0.02 | -0.06 | 0.02 | -0.01 | -0.05 | -0.17 | -0.18 | 0.01 | 0.85 | -0.19 | -0.14 | -0.03 | -0.07 | 0.32 | -0.14 | -0.13 |
| 1 | 402 417 | 0.20 | -0.17 | 0.18 | 0.34 | -0.02 | -0.17 | 0.08 | 0.26 | 0.02 | -0.24 | -0.26 | -0.41 | -0.15 | -0.11 | 0.08 | 0.41 | 0.31 | 0.02 | -0.27 | -0.10 |
| 1 | 403 404 | -0.07 | -0.16 | 0.10 | -0.17 | -0.37 | 0.00 | -0.19 | -0.19 | -0.14 | 0.02 | -0.27 | 0.31 | -0.01 | 0.16 | -0.26 | 0.11 | 0.45 | 0.43 | 0.08 | 0.18 |
| 1 | 407 411 | -0.06 | -0.08 | -0.11 | -0.03 | 0.21 | -0.09 | 0.19 | 0.35 | -0.33 | 0.18 | 0.15 | -0.39 | -0.18 | 0.30 | -0.02 | -0.06 | -0.02 | 0.20 | -0.46 | 0.26 |
| 1 | 409 414 | 0.03 | 0.16 | 0.00 | -0.02 | 0.16 | -0.26 | -0.06 | -0.09 | -0.48 | 0.20 | -0.12 | -0.35 | 0.36 | 0.28 | 0.02 | -0.08 | 0.16 | -0.35 | 0.17 | 0.27 |
| 1 | 420 434 | 0.28 | 0.44 | 0.10 | -0.12 | -0.18 | 0.10 | 0.06 | 0.14 | -0.35 | -0.43 | 0.08 | 0.31 | 0.03 | -0.07 | 0.17 | 0.06 | -0.26 | 0.12 | -0.23 | -0.24 |
| 1 | 426 427 | -0.31 | -0.07 | 0.07 | -0.36 | 0.20 | -0.23 | 0.48 | 0.35 | 0.14 | 0.19 | -0.09 | -0.23 | 0.02 | 0.15 | -0.14 | -0.37 | 0.01 | 0.00 | 0.12 | 0.06 |
| 1 | 429 430 | -0.38 | -0.15 | 0.28 | 0.52 | 0.30 | -0.17 | -0.28 | -0.24 | 0.08 | 0.26 | 0.13 | -0.08 | -0.22 | 0.11 | 0.08 | -0.22 | -0.01 | -0.09 | -0.01 | 0.11 |
| 1 | 437 438 | 0.14 | -0.38 | 0.27 | -0.10 | 0.14 | 0.04 | -0.33 | 0.02 | -0.04 | -0.37 | 0.13 | 0.16 | 0.03 | 0.04 | 0.00 | 0.58 | 0.01 | -0.05 | -0.30 | 0.02 |
| 1 | 439 440 | -0.07 | -0.42 | 0.11 | -0.28 | 0.12 | 0.27 | -0.22 | 0.12 | 0.30 | -0.13 | 0.45 | -0.02 | -0.10 | -0.13 | 0.10 | -0.04 | 0.04 | -0.10 | -0.32 | 0.33 |
| 1 | 445 446 | 0.41 | -0.13 | 0.21 | -0.28 | -0.29 | 0.01 | -0.20 | 0.31 | -0.07 | 0.06 | 0.09 | -0.38 | 0.00 | -0.01 | 0.27 | 0.29 | -0.10 | -0.33 | -0.06 | 0.19 |
| 1 | 447 448 | 0.19 | -0.20 | 0.00 | -0.39 | -0.21 | -0.19 | -0.44 | 0.18 | 0.18 | 0.21 | 0.23 | 0.27 | 0.10 | 0.17 | 0.16 | 0.05 | -0.39 | -0.08 | 0.17 | -0.01 |
| 1 | 459 461 | 0.31 | -0.10 | -0.12 | 0.12 | -0.13 | -0.16 | 0.43 | 0.43 | 0.12 | 0.08 | -0.02 | -0.53 | 0.02 | 0.06 | -0.19 | -0.33 | 0.04 | 0.01 | 0.03 | -0.10 |
| 1 | 460 607 | -0.07 | -0.03 | -0.36 | -0.04 | -0.13 | -0.03 | 0.01 | -0.41 | 0.05 | 0.07 | 0.37 | 0.41 | -0.14 | 0.09 | -0.16 | 0.20 | -0.26 | -0.16 | 0.35 | 0.23 |
| 1 | 462 463 | 0.48 | -0.18 | 0.13 | -0.07 | 0.05 | 0.16 | -0.20 | 0.25 | 0.05 | -0.02 | 0.01 | -0.33 | 0.18 | -0.10 | -0.16 | 0.27 | 0.20 | -0.01 | -0.20 | -0.50 |
| 1 | 471 472 | 0.19 | -0.42 | -0.17 | -0.35 | 0.18 | -0.18 | -0.43 | 0.04 | -0.16 | 0.27 | 0.27 | -0.07 | 0.19 | 0.26 | 0.09 | 0.06 | -0.07 | 0.14 | -0.07 | 0.25 |
| 1 | 474 548 | 0.83 | -0.03 | -0.37 | -0.01 | -0.06 | 0.33 | -0.01 | -0.02 | -0.03 | -0.08 | -0.06 | -0.18 | -0.05 | -0.05 | 0.00 | -0.02 | -0.02 | -0.06 | -0.05 | -0.07 |
| 1 | 482 484 | -0.06 | -0.45 | 0.14 | 0.38 | 0.21 | -0.25 | -0.34 | 0.13 | 0.23 | 0.09 | -0.12 | -0.26 | 0.11 | -0.29 | 0.22 | 0.00 | 0.28 | 0.09 | -0.11 | 0.00 |
| 1 | 483 499 | 0.32 | -0.31 | 0.01 | 0.21 | 0.19 | -0.15 | -0.10 | 0.03 | -0.12 | -0.34 | -0.26 | -0.44 | 0.18 | 0.03 | 0.46 | 0.05 | 0.09 | 0.18 | -0.06 | 0.03 |
| 1 | 488 493 | 0.15 | -0.21 | -0.28 | 0.08 | -0.39 | -0.28 | -0.05 | -0.07 | 0.16 | 0.20 | 0.00 | 0.32 | 0.24 | -0.12 | -0.25 | -0.12 | -0.16 | 0.31 | 0.41 | 0.05 |
| 1 | 501 502 | -0.12 | -0.03 | 0.24 | 0.51 | -0.12 | 0.10 | 0.35 | -0.03 | -0.02 | -0.25 | -0.20 | -0.44 | -0.23 | 0.04 | -0.15 | 0.12 | 0.06 | 0.31 | 0.05 | -0.19 |
| 1 | 510 539 | 0.01 | -0.66 | 0.23 | 0.36 | -0.01 | 0.16 | 0.34 | 0.05 | -0.12 | -0.12 | -0.06 | 0.22 | -0.13 | -0.15 | 0.09 | -0.13 | -0.19 | -0.03 | 0.22 | -0.07 |
| 1 | 512 605 | -0.02 | 0.30 | 0.05 | -0.18 | 0.09 | 0.05 | -0.09 | 0.04 | 0.44 | 0.13 | 0.11 | -0.42 | 0.16 | -0.10 | 0.06 | -0.02 | -0.01 | -0.02 | -0.64 | 0.09 |
| 1 | 513 570 | -0.13 | -0.30 | -0.08 | 0.15 | 0.21 | -0.28 | 0.38 | 0.03 | 0.08 | -0.12 | -0.05 | 0.00 | 0.13 | 0.47 | -0.03 | -0.49 | 0.20 | -0.10 | 0.11 | -0.18 |
| 1 | 514 543 | 0.49 | -0.46 | 0.12 | 0.11 | 0.10 | 0.05 | -0.37 | -0.20 | 0.04 | -0.35 | -0.03 | 0.17 | -0.15 | 0.08 | 0.09 | 0.16 | 0.06 | -0.22 | 0.23 | 0.08 |
| 1 | 516 517 | 0.13 | 0.14 | 0.13 | 0.27 | 0.18 | -0.41 | -0.07 | 0.26 | 0.11 | 0.06 | -0.22 | -0.21 | -0.20 | -0.09 | -0.48 | 0.30 | -0.15 | 0.32 | 0.00 | -0.08 |
| 1 | 518 519 | -0.02 | -0.42 | 0.02 | 0.20 | 0.57 | 0.09 | 0.29 | -0.08 | -0.06 | -0.04 | -0.20 | 0.13 | 0.00 | -0.02 | -0.40 | -0.02 | -0.22 | 0.08 | 0.24 | -0.14 |
| 1 | 522 523 | 0.26 | -0.20 | -0.29 | 0.20 | -0.35 | -0.16 | -0.10 | -0.22 | -0.04 | -0.20 | 0.27 | 0.45 | 0.12 | 0.14 | 0.27 | -0.01 | -0.11 | -0.15 | 0.28 | -0.17 |
| 1 | 524 525 | -0.03 | -0.18 | -0.02 | 0.14 | 0.06 | 0.07 | -0.07 | 0.00 | 0.07 | 0.13 | 0.35 | -0.03 | 0.21 | -0.14 | -0.15 | 0.18 | 0.23 | -0.65 | -0.38 | 0.21 |
| 1 | 531 533 | -0.10 | -0.45 | 0.00 | 0.14 | -0.09 | -0.09 | 0.48 | -0.27 | -0.29 | 0.19 | -0.05 | -0.31 | 0.10 | 0.24 | 0.05 | -0.21 | 0.08 | 0.17 | 0.19 | 0.22 |
| 1 | 574 575 | 0.04 | -0.41 | -0.04 | -0.17 | 0.21 | 0.00 | -0.14 | 0.01 | 0.12 | -0.10 | -0.09 | -0.47 | 0.04 | 0.09 | -0.06 | 0.17 | -0.04 | 0.49 | 0.43 | -0.07 |
| 1 | 582 585 | 0.23 | -0.51 | 0.03 | 0.24 | 0.41 | 0.01 | 0.17 | 0.34 | -0.11 | -0.24 | -0.17 | -0.07 | -0.02 | 0.02 | 0.10 | 0.13 | 0.09 | -0.30 | -0.28 | -0.07 |
| 1 | 583 584 | 0.14 | -0.43 | -0.22 | 0.11 | 0.32 | -0.21 | 0.12 | -0.12 | -0.28 | 0.12 | 0.11 | -0.07 | 0.24 | -0.21 | 0.06 | 0.30 | 0.21 | -0.44 | 0.11 | 0.12 |

Table B.4: PCA Generated Amino Acid Indices With Complete Linkage 1

| Variance | Indices | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 19 363 | -0.04 | -0.40 | -0.05 | -0.18 | -0.06 | 0.41 | -0.20 | 0.01 | -0.02 | -0.01 | -0.10 | -0.09 | 0.04 | -0.33 | 0.14 | -0.02 | 0.19 | 0.56 | -0.12 | 0.28 |
| 1 | 20 283 | 0.10 | -0.07 | 0.14 | -0.24 | 0.15 | -0.03 | 0.04 | -0.33 | -0.14 | 0.31 | -0.15 | 0.05 | -0.33 | -0.09 | 0.00 | 0.33 | 0.20 | -0.43 | 0.06 | 0.42 |
| 1 | 22 80 | 0.18 | -0.08 | -0.02 | -0.41 | 0.37 | -0.10 | -0.49 | 0.26 | 0.17 | -0.03 | -0.03 | -0.14 | 0.22 | 0.14 | 0.15 | -0.09 | -0.18 | 0.30 | -0.27 | 0.05 |
| 1 | 31 284 | 0.21 | 0.19 | 0.25 | 0.08 | 0.04 | 0.10 | 0.19 | -0.03 | 0.02 | -0.57 | -0.15 | -0.11 | 0.26 | -0.02 | -0.56 | -0.01 | 0.02 | 0.09 | 0.15 | -0.15 |
| 1 | 34 509 | 0.10 | 0.48 | -0.14 | 0.14 | 0.32 | 0.02 | -0.01 | 0.21 | 0.21 | -0.24 | -0.02 | -0.58 | -0.13 | -0.16 | -0.01 | -0.02 | -0.10 | 0.21 | -0.18 | -0.08 |
| 1 | 35 505 | -0.03 | -0.04 | -0.11 | -0.04 | -0.41 | -0.04 | 0.11 | 0.04 | 0.08 | 0.20 | -0.34 | 0.10 | 0.57 | -0.22 | 0.36 | 0.11 | 0.10 | -0.30 | -0.13 | 0.00 |
| 1 | 39 225 | 0.22 | -0.43 | -0.14 | 0.56 | -0.12 | -0.33 | 0.18 | 0.24 | 0.01 | 0.07 | -0.08 | 0.06 | -0.23 | 0.16 | -0.22 | -0.10 | 0.09 | -0.16 | 0.04 | 0.18 |
| 1 | 44 344 | -0.05 | -0.16 | 0.10 | 0.30 | -0.02 | 0.11 | 0.04 | -0.51 | -0.25 | 0.00 | -0.15 | -0.42 | -0.15 | 0.02 | 0.24 | 0.18 | -0.01 | 0.38 | 0.28 | 0.04 |
| 1 | 49 321 | -0.10 | 0.03 | -0.29 | 0.10 | 0.66 | 0.07 | -0.10 | -0.19 | 0.40 | -0.14 | 0.02 | -0.21 | -0.13 | 0.00 | 0.17 | -0.22 | -0.26 | 0.13 | -0.01 | 0.07 |
| 1 | 60 185 | -0.15 | 0.02 | 0.04 | -0.08 | 0.00 | 0.47 | -0.18 | -0.09 | 0.13 | 0.10 | -0.07 | -0.53 | 0.12 | -0.32 | -0.20 | 0.21 | 0.18 | 0.41 | -0.03 | -0.02 |
| 1 | 63 117 | 0.02 | -0.16 | 0.17 | -0.27 | -0.57 | 0.21 | 0.07 | -0.17 | -0.37 | 0.02 | 0.00 | -0.10 | -0.05 | 0.17 | 0.33 | -0.03 | 0.32 | 0.00 | 0.26 | 0.14 |
| 1 | 70 86 | -0.01 | -0.21 | 0.54 | 0.05 | -0.04 | 0.44 | -0.15 | -0.01 | 0.21 | -0.22 | -0.19 | -0.23 | -0.35 | -0.21 | -0.04 | 0.25 | 0.12 | 0.10 | 0.06 | -0.10 |
| 1 | 74 90 | 0.01 | -0.29 | -0.23 | 0.25 | -0.06 | -0.13 | 0.08 | -0.27 | 0.02 | -0.13 | 0.21 | -0.33 | 0.03 | 0.41 | -0.15 | -0.14 | 0.18 | 0.53 | 0.04 | -0.02 |
| 1 | 87 425 | -0.01 | 0.18 | 0.29 | 0.13 | -0.04 | 0.03 | 0.07 | -0.02 | 0.26 | 0.10 | -0.01 | -0.58 | -0.35 | 0.12 | -0.39 | -0.06 | 0.17 | -0.20 | 0.28 | 0.02 |
| 1 | 100 230 | 0.00 | 0.21 | -0.18 | -0.13 | 0.24 | 0.47 | -0.47 | 0.00 | -0.17 | -0.28 | 0.22 | -0.05 | 0.11 | -0.23 | 0.03 | 0.05 | 0.32 | -0.09 | -0.19 | 0.16 |
| 1 | 106 428 | -0.14 | -0.15 | 0.21 | -0.13 | 0.49 | -0.44 | 0.22 | 0.06 | -0.11 | 0.31 | -0.28 | -0.23 | -0.12 | 0.02 | 0.03 | 0.08 | 0.17 | -0.09 | -0.17 | 0.28 |
| 1 | 111 385 | -0.09 | -0.12 | -0.16 | 0.41 | 0.20 | -0.28 | 0.46 | -0.06 | -0.32 | 0.09 | 0.18 | -0.03 | 0.07 | 0.13 | -0.19 | -0.26 | -0.37 | 0.17 | 0.06 | 0.11 |
| 1 | 114 397 549 | 0.00 | -0.62 | 0.19 | 0.29 | -0.21 | 0.17 | 0.22 | 0.02 | -0.51 | -0.04 | -0.04 | 0.04 | -0.09 | 0.08 | -0.05 | 0.12 | 0.08 | 0.15 | 0.20 | -0.02 |
| 1 | 115 153 | -0.08 | -0.16 | -0.09 | 0.19 | -0.07 | -0.09 | 0.19 | -0.09 | -0.08 | -0.05 | -0.05 | -0.16 | -0.06 | -0.03 | 0.90 | -0.10 | -0.08 | -0.01 | -0.03 | -0.05 |
| 1 | 124 175 | -0.11 | 0.12 | 0.03 | 0.12 | 0.06 | 0.04 | -0.13 | 0.40 | -0.40 | -0.30 | -0.38 | -0.20 | -0.08 | -0.10 | 0.01 | 0.37 | 0.28 | 0.09 | 0.29 | -0.14 |
| 1 | 130 545 | 0.15 | -0.28 | 0.09 | 0.21 | -0.20 | -0.08 | 0.17 | 0.31 | -0.35 | 0.26 | 0.20 | -0.06 | -0.01 | 0.00 | 0.17 | 0.01 | -0.15 | -0.01 | -0.60 | 0.17 |
| 1 | 134&454 | 0.17 | 0.11 | -0.13 | 0.00 | -0.29 | -0.02 | 0.03 | 0.61 | 0.16 | 0.23 | -0.24 | 0.13 | 0.28 | -0.02 | -0.33 | -0.29 | -0.20 | -0.08 | -0.04 | -0.08 |
| 1 | 139 608 | 0.08 | -0.26 | -0.28 | 0.13 | 0.15 | 0.32 | -0.48 | 0.41 | 0.08 | -0.21 | 0.21 | -0.07 | -0.26 | -0.19 | 0.22 | -0.06 | 0.12 | -0.07 | 0.17 | -0.01 |
| 1 | 143 148 | -0.15 | -0.21 | -0.28 | 0.42 | 0.13 | -0.10 | -0.03 | 0.04 | -0.51 | 0.06 | 0.09 | -0.22 | -0.21 | 0.13 | 0.03 | 0.31 | 0.40 | 0.11 | -0.04 | 0.02 |
| 1 | 147 221 | 0.00 | 0.13 | 0.09 | 0.40 | -0.23 | 0.30 | 0.09 | -0.10 | 0.00 | -0.01 | -0.05 | -0.25 | -0.22 | -0.20 | 0.32 | -0.18 | -0.49 | 0.24 | 0.24 | -0.10 |
| 1 | 149 297 | -0.23 | -0.20 | 0.03 | 0.27 | 0.11 | 0.60 | 0.00 | -0.01 | -0.18 | 0.10 | -0.08 | -0.14 | 0.17 | 0.11 | -0.50 | 0.01 | -0.01 | 0.07 | 0.14 | -0.27 |
| 1 | 151 481 | 0.57 | 0.41 | -0.09 | 0.00 | -0.02 | 0.07 | -0.35 | 0.15 | -0.08 | 0.00 | -0.01 | 0.09 | -0.03 | -0.02 | 0.19 | -0.32 | -0.06 | -0.44 | -0.06 | 0.00 |
| 1 | 152 395 | 0.14 | -0.04 | 0.07 | 0.19 | 0.28 | -0.19 | 0.18 | -0.29 | 0.24 | 0.30 | 0.04 | 0.22 | -0.22 | 0.11 | -0.32 | -0.30 | -0.01 | -0.50 | 0.09 | 0.01 |
| 1 | 156 432 | 0.21 | -0.15 | -0.34 | -0.10 | 0.00 | -0.31 | -0.13 | 0.08 | 0.15 | 0.44 | -0.04 | -0.17 | -0.13 | 0.16 | -0.29 | 0.14 | 0.34 | -0.07 | -0.15 | 0.39 |
| 1 | 167 246 | 0.06 | 0.12 | 0.12 | -0.01 | 0.15 | -0.17 | -0.11 | 0.06 | -0.30 | 0.18 | -0.16 | 0.23 | 0.03 | -0.40 | 0.00 | 0.16 | 0.05 | -0.58 | 0.26 | 0.32 |
| 1 | 174 371 | -0.16 | -0.43 | 0.14 | 0.22 | 0.26 | -0.01 | -0.02 | -0.15 | -0.15 | -0.04 | -0.09 | -0.16 | 0.59 | -0.12 | 0.09 | 0.24 | -0.01 | -0.23 | 0.23 | -0.21 |
| 1 | 187 530 | 0.19 | 0.51 | 0.24 | 0.00 | -0.10 | -0.11 | 0.29 | -0.11 | -0.09 | 0.01 | -0.59 | 0.07 | 0.15 | -0.25 | -0.17 | -0.03 | -0.02 | -0.11 | 0.20 | -0.05 |
| 1 | 189 532 | 0.06 | -0.22 | 0.09 | 0.11 | -0.26 | -0.07 | -0.17 | 0.04 | -0.03 | 0.23 | 0.11 | -0.55 | 0.28 | 0.22 | -0.13 | 0.18 | 0.40 | -0.09 | 0.13 | -0.32 |
| 1 | 196 455 | 0.32 | -0.22 | 0.04 | 0.06 | -0.25 | 0.08 | 0.08 | 0.48 | 0.04 | -0.14 | -0.20 | -0.31 | 0.18 | 0.25 | -0.12 | -0.44 | 0.00 | 0.24 | 0.04 | -0.14 |
| 1 | 198 208 | -0.26 | -0.13 | 0.26 | 0.02 | 0.29 | -0.03 | 0.03 | -0.63 | -0.05 | 0.48 | -0.11 | -0.02 | -0.02 | 0.21 | -0.01 | 0.03 | -0.20 | -0.06 | 0.13 | 0.09 |
| 1 | 205 465 | -0.07 | -0.19 | 0.00 | -0.15 | 0.21 | -0.09 | -0.16 | 0.31 | 0.03 | 0.55 | -0.25 | 0.05 | 0.10 | -0.35 | 0.03 | -0.37 | -0.10 | 0.32 | 0.02 | 0.10 |
| 1 | 209 247 | -0.30 | -0.10 | -0.02 | 0.11 | -0.15 | -0.01 | -0.18 | 0.19 | -0.44 | 0.17 | 0.21 | -0.06 | -0.44 | 0.03 | 0.45 | 0.21 | 0.14 | -0.07 | 0.02 | 0.24 |
| 1 | 212 529 | -0.13 | -0.13 | 0.03 | 0.14 | -0.08 | 0.25 | -0.26 | -0.36 | -0.35 | 0.25 | -0.06 | 0.09 | -0.41 | 0.46 | 0.09 | 0.24 | -0.04 | 0.01 | 0.08 | 0.19 |
| 1 | 218 380 | 0.12 | 0.19 | -0.14 | -0.37 | 0.11 | -0.05 | -0.17 | 0.01 | 0.02 | 0.03 | 0.00 | 0.44 | -0.02 | -0.46 | -0.19 | 0.35 | 0.23 | 0.17 | -0.30 | 0.05 |
| 1 | 219 478 | 0.49 | -0.35 | 0.11 | 0.12 | -0.36 | -0.11 | 0.10 | 0.39 | 0.19 | 0.04 | 0.11 | -0.38 | 0.10 | -0.03 | 0.03 | -0.24 | -0.20 | 0.01 | -0.04 | 0.03 |
| 1 | 220 298 | -0.22 | 0.29 | -0.38 | -0.25 | -0.17 | 0.13 | 0.10 | 0.12 | -0.03 | -0.02 | 0.06 | 0.07 | 0.06 | -0.07 | -0.14 | 0.13 | 0.52 | 0.33 | -0.36 | -0.15 |
| 1 | 233 252 | -0.16 | -0.06 | 0.18 | -0.07 | -0.58 | 0.33 | 0.30 | 0.04 | -0.15 | 0.24 | -0.04 | -0.14 | 0.32 | 0.26 | -0.23 | -0.05 | 0.08 | -0.20 | 0.10 | -0.16 |
| 1 | 239 480 | -0.17 | -0.10 | -0.21 | -0.32 | 0.14 | -0.05 | -0.05 | -0.13 | 0.67 | 0.02 | 0.00 | 0.18 | 0.25 | -0.02 | 0.24 | -0.18 | 0.09 | -0.29 | -0.16 | 0.10 |
| 1 | 249 250 | -0.20 | 0.37 | -0.01 | -0.04 | -0.54 | 0.08 | 0.12 | -0.50 | -0.01 | 0.13 | 0.02 | 0.04 | -0.16 | 0.13 | -0.10 | 0.00 | 0.09 | 0.41 | 0.10 | 0.06 |
| 1 | 251 378 527 | -0.07 | 0.19 | -0.20 | -0.03 | 0.54 | -0.20 | -0.14 | 0.01 | 0.07 | -0.12 | -0.14 | -0.45 | 0.24 | -0.18 | 0.36 | 0.19 | 0.16 | -0.18 | 0.05 | -0.10 |
| 1 | 257 299 | -0.03 | -0.20 | 0.32 | 0.39 | -0.12 | -0.28 | -0.02 | -0.08 | 0.30 | 0.01 | -0.09 | 0.06 | -0.25 | 0.32 | -0.42 | 0.11 | -0.23 | -0.09 | 0.02 | 0.29 |
| 1 | 259 327 | -0.29 | -0.23 | 0.08 | 0.16 | 0.23 | 0.08 | 0.50 | -0.26 | -0.23 | -0.12 | 0.38 | -0.24 | -0.07 | 0.06 | 0.25 | 0.19 | -0.17 | -0.09 | -0.11 | -0.13 |
| 1 | 261 262 | -0.14 | 0.07 | -0.04 | 0.64 | -0.07 | -0.14 | 0.01 | -0.44 | -0.05 | -0.02 | -0.24 | -0.32 | 0.17 | 0.20 | 0.14 | 0.06 | 0.11 | 0.24 | -0.02 | -0.15 |

Table B.4: (continued)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 271 281 | -0.16 | 0.18 | 0.06 | -0.27 | -0.20 | 0.24 | 0.00 | 0.32 | -0.42 | 0.24 | -0.13 | 0.51 | 0.03 | -0.09 | -0.21 | 0.12 | -0.03 | 0.16 | -0.19 | -0.16 |
| 1 | 272 401 | -0.08 | -0.03 | -0.29 | -0.13 | 0.22 | 0.06 | -0.39 | 0.13 | -0.02 | 0.34 | -0.24 | -0.01 | 0.10 | -0.06 | -0.37 | -0.11 | 0.30 | 0.07 | 0.04 | 0.48 |
| 1 | 279 280 | -0.33 | -0.14 | 0.10 | 0.07 | -0.06 | -0.32 | -0.16 | 0.10 | 0.12 | 0.47 | -0.14 | 0.07 | -0.10 | 0.06 | -0.27 | -0.10 | -0.12 | -0.03 | 0.28 | 0.50 |
| 1 | 282 361 | -0.02 | -0.28 | -0.18 | 0.34 | -0.38 | 0.31 | -0.09 | 0.04 | -0.07 | 0.00 | 0.11 | 0.50 | 0.01 | -0.13 | -0.22 | -0.01 | -0.09 | 0.36 | 0.01 | -0.25 |
| 1 | 294 537 | 0.15 | -0.12 | -0.07 | 0.07 | 0.29 | -0.56 | -0.20 | 0.12 | 0.23 | 0.09 | -0.12 | 0.04 | -0.22 | -0.38 | 0.41 | 0.22 | -0.08 | 0.06 | 0.07 | 0.00 |
| 1 | 310 451 | 0.12 | -0.31 | -0.49 | 0.28 | 0.07 | 0.21 | -0.28 | 0.20 | -0.23 | -0.05 | 0.12 | 0.03 | -0.10 | 0.18 | -0.18 | 0.14 | 0.40 | 0.03 | -0.26 | 0.10 |
| 1 | 335 364 | 0.10 | 0.00 | -0.47 | -0.06 | 0.03 | 0.21 | -0.01 | -0.11 | 0.49 | 0.24 | 0.29 | -0.06 | 0.00 | -0.12 | 0.24 | -0.17 | -0.09 | -0.36 | 0.12 | -0.26 |
| 1 | 349 558 | -0.38 | -0.15 | 0.11 | 0.04 | 0.04 | 0.24 | 0.14 | -0.08 | 0.15 | 0.36 | 0.07 | 0.32 | 0.00 | 0.10 | 0.20 | -0.31 | -0.32 | -0.22 | 0.11 | -0.41 |
| 1 | 369 589 | 0.03 | -0.09 | 0.38 | 0.36 | -0.35 | -0.41 | 0.07 | -0.42 | 0.37 | -0.12 | 0.04 | 0.00 | 0.09 | -0.03 | -0.05 | 0.19 | -0.11 | 0.12 | 0.00 | -0.08 |
| 1 | 402 417 | 0.20 | -0.17 | 0.18 | 0.34 | -0.02 | -0.17 | 0.08 | 0.26 | 0.02 | -0.24 | -0.26 | -0.41 | -0.15 | -0.11 | 0.08 | 0.41 | 0.31 | 0.02 | -0.27 | -0.10 |
| 1 | 407 411 | -0.06 | -0.08 | -0.11 | -0.03 | 0.21 | -0.09 | 0.19 | 0.35 | -0.33 | 0.18 | 0.15 | -0.39 | -0.18 | 0.30 | -0.02 | -0.06 | -0.02 | 0.20 | -0.46 | 0.26 |
| 1 | 423 435 | -0.17 | -0.04 | -0.12 | 0.13 | 0.59 | 0.07 | -0.12 | -0.31 | -0.20 | 0.01 | -0.01 | 0.17 | -0.24 | -0.07 | 0.40 | -0.06 | -0.08 | 0.21 | -0.32 | 0.16 |
| 1 | 424 507 | 0.16 | -0.20 | 0.03 | 0.15 | -0.01 | -0.05 | -0.28 | 0.10 | 0.05 | -0.37 | -0.36 | 0.22 | 0.14 | 0.45 | -0.17 | -0.18 | 0.44 | -0.14 | 0.03 | -0.01 |
| 1 | 441 579 | 0.19 | -0.13 | 0.11 | 0.28 | -0.18 | 0.02 | 0.22 | -0.24 | 0.11 | -0.40 | -0.25 | 0.10 | -0.40 | 0.23 | 0.22 | -0.07 | -0.14 | 0.39 | 0.10 | -0.15 |
| 1 | 462 463 | 0.48 | -0.18 | 0.13 | -0.07 | 0.05 | 0.16 | -0.20 | 0.25 | 0.05 | -0.02 | 0.01 | -0.33 | 0.18 | -0.10 | -0.16 | 0.27 | 0.20 | -0.01 | -0.20 | -0.50 |
| 1 | 464 597 | 0.13 | 0.22 | 0.25 | 0.18 | -0.60 | 0.00 | -0.04 | -0.16 | -0.18 | 0.05 | 0.01 | 0.02 | -0.08 | 0.07 | -0.04 | -0.20 | 0.13 | -0.09 | 0.55 | -0.19 |
| 1 | 475 511 | 0.27 | -0.17 | -0.27 | 0.02 | 0.31 | -0.26 | 0.04 | 0.09 | -0.07 | -0.11 | -0.11 | -0.14 | 0.49 | -0.46 | 0.26 | 0.08 | -0.10 | 0.07 | -0.15 | 0.20 |
| 1 | 482 484 | -0.06 | -0.45 | 0.14 | 0.38 | 0.21 | -0.25 | -0.34 | 0.13 | 0.23 | 0.09 | -0.12 | -0.26 | 0.11 | -0.29 | 0.22 | 0.00 | 0.28 | 0.09 | -0.11 | 0.00 |
| 1 | 486 487 | -0.04 | 0.12 | -0.09 | -0.08 | 0.20 | 0.51 | -0.33 | -0.15 | -0.18 | 0.02 | 0.18 | 0.18 | -0.16 | -0.42 | -0.04 | -0.24 | 0.22 | 0.29 | -0.13 | 0.15 |
| 1 | 489 494 | 0.22 | 0.00 | -0.42 | -0.13 | -0.18 | -0.25 | -0.13 | -0.26 | -0.13 | 0.27 | 0.37 | -0.01 | 0.35 | 0.23 | 0.31 | -0.17 | -0.22 | 0.06 | 0.06 | 0.03 |
| 1 | 501 502 | -0.12 | -0.03 | 0.24 | 0.51 | -0.12 | 0.10 | 0.35 | -0.03 | -0.02 | -0.25 | -0.20 | -0.44 | -0.23 | 0.04 | -0.15 | 0.12 | 0.06 | 0.31 | 0.05 | -0.19 |
| 1 | 503 602 | -0.07 | -0.07 | 0.20 | 0.44 | 0.04 | -0.13 | 0.27 | 0.21 | 0.13 | -0.08 | -0.38 | 0.25 | 0.27 | 0.01 | -0.33 | -0.09 | -0.28 | -0.05 | -0.34 | -0.01 |
| 1 | 514 543 | 0.49 | -0.46 | 0.12 | 0.11 | 0.10 | 0.05 | -0.37 | -0.20 | 0.04 | -0.35 | -0.03 | 0.17 | -0.15 | 0.08 | 0.09 | 0.16 | 0.06 | -0.22 | 0.23 | 0.08 |
| 1 | 518 519 | -0.02 | -0.42 | 0.02 | 0.20 | 0.57 | 0.09 | 0.29 | -0.08 | -0.06 | -0.04 | -0.20 | 0.13 | 0.00 | -0.02 | -0.40 | -0.02 | -0.22 | 0.08 | 0.24 | -0.14 |
| 1 | 557 577 | 0.17 | -0.06 | 0.02 | -0.22 | 0.26 | -0.08 | -0.05 | -0.07 | 0.31 | -0.23 | 0.12 | -0.42 | 0.15 | -0.19 | -0.36 | 0.42 | 0.03 | 0.06 | 0.31 | -0.15 |
| 1 | 561 571 | -0.19 | -0.13 | -0.02 | 0.05 | 0.13 | 0.13 | 0.26 | -0.42 | 0.08 | 0.07 | 0.11 | 0.14 | 0.22 | -0.16 | 0.40 | -0.02 | -0.12 | -0.57 | 0.17 | -0.15 |
| 1 | 582 585 | 0.23 | -0.51 | 0.03 | 0.24 | 0.41 | 0.01 | 0.17 | 0.34 | -0.11 | -0.24 | -0.17 | -0.07 | -0.02 | 0.02 | 0.10 | 0.13 | 0.09 | -0.30 | -0.28 | -0.07 |
| 0.9996 | 399 554 596 | -0.09 | -0.14 | 0.20 | -0.11 | 0.01 | 0.46 | -0.26 | -0.39 | 0.33 | -0.03 | -0.09 | -0.14 | -0.18 | 0.11 | 0.49 | -0.07 | 0.14 | -0.01 | -0.03 | -0.20 |
| 0.999 | 521 522 523 | 0.05 | 0.30 | -0.11 | -0.33 | 0.12 | 0.10 | 0.19 | 0.05 | 0.37 | -0.04 | -0.18 | -0.03 | 0.18 | -0.21 | -0.19 | 0.08 | -0.09 | -0.60 | 0.20 | 0.17 |
| 0.9989 | 169 474 548 | -0.15 | -0.40 | -0.11 | -0.03 | -0.02 | 0.03 | 0.36 | 0.14 | -0.17 | -0.13 | 0.21 | 0.33 | 0.30 | -0.11 | 0.30 | -0.20 | -0.23 | -0.19 | -0.22 | 0.30 |
| 0.9972 | 8 142 598 | -0.33 | 0.42 | -0.23 | 0.36 | 0.00 | -0.32 | 0.03 | 0.00 | -0.22 | 0.29 | 0.05 | -0.09 | -0.16 | -0.02 | 0.26 | -0.27 | -0.12 | -0.02 | 0.30 | 0.07 |
| 0.9959 | 72 535 552 586 | -0.19 | 0.14 | -0.01 | 0.03 | 0.55 | 0.03 | 0.07 | -0.24 | -0.10 | -0.05 | -0.05 | 0.05 | 0.65 | -0.22 | -0.16 | -0.07 | -0.02 | -0.20 | -0.10 | -0.10 |
| 0.9953 | 191 193 195 | 0.10 | -0.03 | -0.47 | -0.09 | 0.06 | 0.06 | 0.09 | -0.09 | 0.01 | -0.29 | 0.23 | -0.14 | 0.39 | -0.08 | 0.18 | 0.03 | 0.32 | 0.28 | -0.42 | -0.15 |

Table B.5: PCA Generated Amino Acid Indices With Complete Linkage 0.65

| Variance | Indices | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 17 | -0.19 | -0.11 | 0.08 | 0.09 | 0.13 | -0.16 | -0.20 | -0.41 | -0.01 | 0.51 | 0.44 | -0.18 | -0.10 | 0.02 | -0.02 | -0.10 | -0.20 | 0.04 | -0.01 | 0.37 |
| 1 | 2 132 | 0.17 | -0.19 | 0.23 | 0.36 | -0.03 | -0.07 | 0.09 | -0.03 | 0.17 | -0.21 | -0.23 | 0.37 | -0.42 | 0.04 | 0.00 | -0.13 | -0.13 | 0.38 | -0.03 | -0.34 |
| 1 | 3 4 | -0.33 | -0.16 | 0.05 | 0.01 | 0.66 | 0.27 | -0.03 | -0.05 | 0.10 | -0.19 | 0.24 | 0.01 | -0.31 | -0.06 | 0.06 | 0.21 | -0.01 | -0.24 | -0.21 | -0.01 |
| 1 | 5 564 | -0.15 | -0.26 | -0.06 | -0.26 | 0.33 | 0.31 | 0.27 | -0.06 | -0.15 | -0.02 | 0.26 | 0.15 | -0.58 | 0.16 | -0.09 | 0.16 | 0.02 | 0.12 | -0.19 | 0.03 |
| 1 | 8 598 | 0.10 | 0.11 | -0.23 | -0.05 | 0.16 | -0.20 | 0.22 | -0.23 | -0.35 | -0.16 | 0.25 | 0.27 | 0.19 | -0.43 | 0.09 | 0.23 | -0.32 | 0.20 | -0.06 | 0.20 |
| 1 | 9 109 | -0.46 | 0.04 | 0.03 | 0.51 | 0.09 | 0.01 | -0.04 | -0.25 | -0.03 | 0.22 | 0.22 | -0.23 | 0.00 | 0.11 | -0.12 | 0.37 | -0.13 | -0.33 | 0.07 | -0.09 |
| 1 | 10 319 | -0.21 | -0.30 | 0.24 | -0.15 | 0.01 | -0.11 | 0.20 | -0.35 | 0.16 | 0.13 | 0.07 | 0.04 | 0.38 | -0.35 | 0.28 | 0.16 | 0.12 | -0.38 | -0.09 | 0.15 |
| 1 | 12 13 | 0.23 | -0.48 | -0.18 | 0.02 | 0.45 | -0.14 | 0.08 | 0.10 | -0.29 | -0.36 | 0.29 | -0.06 | 0.10 | 0.25 | -0.05 | -0.01 | -0.01 | -0.19 | 0.14 | 0.10 |
| 1 | 15 59 | 0.17 | -0.42 | -0.04 | 0.17 | -0.24 | -0.04 | 0.02 | 0.18 | -0.02 | 0.27 | 0.03 | 0.06 | 0.04 | -0.18 | -0.05 | 0.43 | 0.11 | -0.28 | -0.47 | 0.25 |
| 1 | 18 604 | -0.15 | -0.05 | 0.18 | -0.05 | 0.25 | -0.03 | 0.13 | 0.17 | -0.29 | 0.42 | -0.10 | -0.27 | -7e-05 | 0.41 | -0.43 | 0.03 | 0.09 | -0.35 | 0.00 | 0.05 |
| 1 | 19 363 | -0.04 | -0.40 | -0.05 | -0.18 | -0.06 | 0.41 | -0.20 | 0.01 | -0.02 | -0.01 | -0.10 | -0.09 | 0.04 | -0.33 | 0.14 | -0.02 | 0.19 | 0.56 | -0.12 | 0.28 |
| 1 | 20 283 | 0.10 | -0.07 | 0.14 | -0.24 | 0.15 | -0.03 | 0.04 | -0.33 | -0.14 | 0.31 | -0.15 | 0.05 | -0.33 | -0.09 | 0.00 | 0.33 | 0.20 | -0.43 | 0.06 | 0.42 |
| 1 | 21 79 | -0.08 | -0.03 | 0.35 | 0.33 | -0.11 | -0.07 | -0.09 | -0.23 | 0.10 | 0.21 | 0.19 | -0.20 | 0.04 | -0.03 | 0.56 | -0.10 | -0.31 | -0.27 | -0.03 | -0.23 |
| 1 | 22 80 | 0.18 | -0.08 | -0.02 | -0.41 | 0.37 | -0.10 | -0.49 | 0.26 | 0.17 | -0.03 | -0.03 | -0.14 | 0.22 | 0.14 | 0.15 | -0.09 | -0.18 | 0.30 | -0.27 | 0.05 |
| 1 | 23 449 | -0.06 | -0.08 | 0.20 | -0.22 | -0.01 | 0.18 | 0.02 | 0.05 | 0.14 | 0.06 | 0.08 | -0.25 | -0.04 | -0.10 | -0.07 | 0.02 | -0.37 | -0.14 | 0.75 | -0.18 |
| 1 | 24 560 | 0.04 | -0.97 | 0.07 | 0.06 | 0.06 | 0.05 | 0.05 | 0.07 | 0.05 | 0.04 | 0.04 | 0.05 | 0.04 | 0.04 | 0.07 | 0.06 | 0.05 | 0.04 | 0.05 | 0.04 |
| 1 | 25 278 | -0.06 | -0.10 | -0.15 | -0.06 | 0.22 | -0.22 | 0.27 | 0.14 | 0.10 | -0.16 | 0.16 | -0.16 | 0.36 | 0.06 | -0.27 | 0.16 | 0.44 | -0.23 | -0.03 | -0.46 |
| 1 | 27 330 | 0.21 | -0.22 | -0.22 | -0.31 | -0.02 | 0.00 | 0.13 | -0.06 | -0.09 | 0.04 | -0.20 | 0.40 | -0.13 | 0.22 | -0.10 | 0.16 | 0.21 | -0.40 | -0.09 | 0.47 |
| 1 | 28 470 | 0.10 | -0.09 | 0.10 | 0.10 | 0.10 | -0.28 | -0.27 | 0.10 | -0.24 | 0.10 | 0.10 | 0.11 | -0.47 | -0.47 | 0.10 | 0.10 | 0.10 | 0.34 | 0.33 | 0.10 |
| 1 | 31 284 | 0.21 | 0.19 | 0.25 | 0.08 | 0.04 | 0.10 | 0.19 | -0.03 | 0.02 | -0.57 | -0.15 | -0.11 | 0.26 | -0.02 | -0.56 | -0.01 | 0.02 | 0.09 | 0.15 | -0.15 |
| 1 | 34 509 | 0.10 | 0.48 | -0.14 | 0.14 | 0.32 | 0.02 | -0.01 | 0.21 | 0.21 | -0.24 | -0.02 | -0.58 | -0.13 | -0.16 | -0.01 | -0.02 | -0.10 | 0.21 | -0.18 | -0.08 |
| 1 | 35 505 | -0.03 | -0.04 | -0.11 | -0.04 | -0.41 | -0.04 | 0.11 | 0.04 | 0.08 | 0.20 | -0.34 | 0.10 | 0.57 | -0.22 | 0.36 | 0.11 | 0.10 | -0.30 | -0.13 | 0.00 |
| 1 | 36 243 | 0.31 | -0.02 | -0.16 | 0.29 | 0.16 | 0.03 | -0.26 | -0.33 | -0.01 | -0.05 | -0.02 | -0.23 | -0.09 | -0.15 | 0.12 | 0.42 | -0.13 | -0.05 | -0.29 | 0.45 |
| 1 | 37 227 | 0.12 | 0.11 | 0.30 | -0.05 | -0.72 | 0.17 | 0.05 | 0.06 | 0.02 | -0.19 | 0.19 | 0.23 | -0.24 | -0.27 | -6e-05 | -0.11 | 0.05 | 0.23 | 0.08 | -0.04 |
| 1 | 38 223 | -0.37 | 0.25 | 0.22 | -0.19 | -0.08 | 0.63 | -0.04 | -0.02 | 0.25 | -0.31 | 0.07 | -0.11 | 0.17 | -0.12 | -0.11 | -0.06 | 0.08 | -0.26 | 0.05 | -0.05 |
| 1 | 39 225 | 0.22 | -0.43 | -0.14 | 0.56 | -0.12 | -0.33 | 0.18 | 0.24 | 0.01 | 0.07 | -0.08 | 0.06 | -0.23 | 0.16 | -0.22 | -0.10 | 0.09 | -0.16 | 0.04 | 0.18 |
| 1 | 40 53 | -0.03 | -0.14 | 0.07 | 0.14 | 0.00 | 0.18 | 0.24 | 0.21 | 0.31 | -0.06 | -0.12 | 0.22 | -0.16 | -0.23 | -0.72 | 0.07 | -0.07 | -0.09 | 0.19 | -0.01 |
| 1 | 41 565 | 0.19 | -0.13 | 0.13 | 0.17 | 0.10 | -0.05 | -0.58 | -0.24 | -0.28 | -0.03 | -0.05 | -0.14 | -0.02 | 0.52 | 0.26 | 0.06 | 0.21 | -0.11 | 0.01 | -0.02 |
| 1 | 42 566 | 0.06 | -0.25 | -0.38 | 0.02 | -0.36 | -0.33 | 0.05 | -0.05 | -0.27 | 0.25 | 0.15 | 0.38 | 0.25 | 0.03 | -0.15 | 0.12 | 0.28 | -0.13 | 0.20 | 0.14 |
| 1 | 44 344 | -0.05 | -0.16 | 0.10 | 0.30 | -0.02 | 0.11 | 0.04 | -0.51 | -0.25 | 0.00 | -0.15 | -0.42 | -0.15 | 0.02 | 0.24 | 0.18 | -0.01 | 0.38 | 0.28 | 0.04 |
| 1 | 45 186 | 0.08 | -0.37 | -0.08 | 0.44 | 0.32 | -0.33 | -0.36 | 0.21 | 0.25 | -0.08 | 0.11 | -0.26 | 0.03 | 0.08 | -0.03 | 0.02 | 0.19 | -0.25 | 0.05 | -0.03 |
| 1 | 46 377 | 0.01 | -0.13 | -0.42 | -0.09 | -0.33 | 0.10 | 0.20 | 0.02 | -0.13 | 0.25 | 0.33 | 0.02 | -0.18 | 0.11 | 0.34 | -0.03 | 0.01 | 0.40 | -0.32 | -0.16 |
| 1 | 49 321 | -0.10 | 0.03 | -0.29 | 0.10 | 0.66 | 0.07 | -0.10 | -0.19 | 0.40 | -0.14 | 0.02 | -0.21 | -0.13 | 0.00 | 0.17 | -0.22 | -0.26 | 0.13 | -0.01 | 0.07 |
| 1 | 50 366 | 0.39 | -0.49 | -0.38 | -0.20 | 0.02 | 0.24 | 0.25 | 0.05 | 0.01 | 0.16 | 0.26 | -0.26 | -0.15 | -0.22 | 0.12 | -0.07 | -0.03 | 0.22 | 0.05 | 0.03 |
| 1 | 51 123 | -0.02 | -0.33 | -0.06 | 0.10 | -0.74 | 0.30 | -0.19 | 0.20 | 0.13 | 0.00 | 0.18 | 0.02 | 0.08 | 0.15 | 0.06 | -0.16 | 0.11 | -0.08 | 0.13 | 0.12 |
| 1 | 52 346 | 0.09 | -0.13 | 0.12 | 0.01 | 0.02 | -0.26 | 0.23 | 0.08 | -0.28 | -0.30 | 0.13 | -0.05 | -0.18 | -0.43 | 0.47 | 0.06 | -0.22 | 0.19 | 0.30 | 0.14 |
| 1 | 55 497 | -0.16 | 0.07 | 0.09 | 0.06 | 0.03 | 0.03 | 0.50 | -0.11 | -0.35 | -0.11 | 0.08 | -0.18 | -0.40 | 0.40 | -0.25 | 0.16 | 0.16 | 0.18 | 0.02 | -0.23 |
| 1 | 56 58 | 0.33 | -0.09 | -0.12 | -0.04 | -0.07 | 0.06 | 0.31 | -0.01 | -0.09 | 0.34 | 0.04 | -0.14 | 0.20 | -0.33 | 0.03 | 0.37 | -0.41 | -0.39 | -0.06 | 0.08 |
| 1 | 60 185 | -0.15 | 0.02 | 0.04 | -0.08 | 0.00 | 0.47 | -0.18 | -0.09 | 0.13 | 0.10 | -0.07 | -0.53 | 0.12 | -0.32 | -0.20 | 0.21 | 0.18 | 0.41 | -0.03 | -0.02 |
| 1 | 63 117 | 0.02 | -0.16 | 0.17 | -0.27 | -0.57 | 0.21 | 0.07 | -0.17 | -0.37 | 0.02 | 0.00 | -0.10 | -0.05 | 0.17 | 0.33 | -0.03 | 0.32 | 0.00 | 0.26 | 0.14 |
| 1 | 64 137 | -0.13 | -0.11 | 0.39 | -0.36 | 0.07 | -0.01 | -0.47 | 0.17 | -0.02 | 0.24 | -0.17 | 0.30 | -0.01 | 0.11 | -0.43 | 0.18 | -0.04 | -0.01 | 0.15 | 0.13 |
| 1 | 65 135 | -0.04 | -0.23 | 0.38 | 0.32 | -0.15 | 0.16 | 0.43 | 0.16 | -0.37 | -0.16 | -0.05 | -0.15 | 0.00 | 0.03 | 0.11 | -0.07 | -0.22 | 0.18 | 0.05 | -0.39 |
| 1 | 67 603 | -0.21 | 0.15 | 0.25 | 0.01 | 0.58 | -0.13 | -0.30 | -0.10 | 0.43 | -0.01 | -0.21 | 0.16 | 0.18 | -0.19 | -0.20 | -0.09 | -0.22 | 0.01 | -0.02 | -0.10 |
| 1 | 68 534 | 0.22 | -0.29 | 0.45 | -0.04 | 0.24 | -0.29 | -0.30 | 0.11 | 0.17 | -0.33 | 0.19 | 0.25 | -0.15 | -0.19 | 0.01 | 0.15 | -0.22 | 0.08 | 0.12 | -0.18 |
| 1 | 70 86 | -0.01 | -0.21 | 0.54 | 0.05 | -0.04 | 0.44 | -0.15 | -0.01 | 0.21 | -0.22 | -0.19 | -0.23 | -0.35 | -0.21 | -0.04 | 0.25 | 0.12 | 0.10 | 0.06 | -0.10 |
| 1 | 72 552 | 0.01 | -0.02 | 0.25 | 0.46 | -0.19 | 0.13 | 0.34 | 0.03 | 0.14 | -0.26 | -0.26 | -0.27 | -0.43 | -0.03 | -0.06 | 0.21 | 0.09 | -0.05 | 0.16 | -0.24 |
| 1 | 74 90 | 0.01 | -0.29 | -0.23 | 0.25 | -0.06 | -0.13 | 0.08 | -0.27 | 0.02 | -0.13 | 0.21 | -0.33 | 0.03 | 0.41 | -0.15 | -0.14 | 0.18 | 0.53 | 0.04 | -0.02 |
| 1 | 75 418 | -0.13 | 0.52 | 0.41 | 0.01 | 0.00 | 0.24 | 0.09 | -0.30 | 0.03 | -0.23 | -0.27 | 0.25 | 0.07 | -0.02 | 0.09 | -0.11 | -0.01 | -0.12 | -0.15 | -0.37 |

Table B.5: (continued)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 76 133 | -0.08 | -0.55 | 0.16 | 0.43 | -0.13 | 0.05 | 0.22 | -0.10 | 0.31 | -0.09 | -0.09 | 0.08 | -0.03 | -0.38 | 0.26 | 0.06 | 0.12 | -0.21 | 0.03 | -0.06 |
| 1 | 77 590 | -0.04 | 0.00 | 0.00 | 0.00 | 0.96 | -0.02 | -0.01 | -0.03 | -0.04 | -0.11 | -0.11 | 0.01 | -0.09 | -0.11 | -0.06 | -0.04 | -0.04 | -0.13 | -0.07 | -0.08 |
| 1 | 82 84 | -0.34 | -0.01 | -0.28 | -0.18 | 0.27 | -0.06 | 0.01 | 0.07 | -0.09 | -0.34 | -0.10 | -0.03 | -0.07 | 0.23 | 0.57 | 0.16 | 0.05 | 0.17 | 0.23 | -0.25 |
| 1 | 83 388 | 0.27 | -0.10 | 0.23 | 0.34 | 0.02 | 0.29 | 0.33 | -0.17 | -0.09 | -0.15 | -0.30 | 0.08 | -0.25 | -0.29 | 0.13 | 0.09 | -0.14 | 0.23 | -0.29 | -0.23 |
| 1 | 85 110 | -0.07 | -0.01 | -0.21 | 0.45 | -0.47 | -0.06 | 0.08 | -0.37 | 0.33 | -0.07 | -0.07 | -0.16 | 0.32 | 0.24 | -0.19 | 0.13 | -0.05 | -0.06 | 0.14 | 0.09 |
| 1 | 87 425 | -0.01 | 0.18 | 0.29 | 0.13 | -0.04 | 0.03 | 0.07 | -0.02 | 0.26 | 0.10 | -0.01 | -0.58 | -0.35 | 0.12 | -0.39 | -0.06 | 0.17 | -0.20 | 0.28 | 0.02 |
| 1 | 89 324 | 0.14 | -0.22 | -0.04 | -0.51 | -0.10 | 0.43 | 0.14 | -0.10 | -0.04 | -0.22 | -0.10 | 0.02 | 0.14 | -0.10 | -0.22 | -0.22 | 0.14 | 0.38 | 0.26 | 0.20 |
| 1 | 92 450 | 0.03 | -0.29 | 0.33 | -0.30 | 0.00 | -0.09 | 0.37 | -0.05 | -0.09 | -0.23 | 0.04 | -0.28 | 0.02 | 0.04 | -0.11 | 0.42 | 0.44 | -0.03 | -0.03 | -0.22 |
| 1 | 93 421 | -0.47 | -0.27 | 0.31 | 0.00 | 0.13 | 0.02 | -0.14 | 0.29 | 0.32 | 0.13 | -0.30 | -0.24 | -0.17 | -0.07 | -0.14 | 0.00 | 0.26 | -0.11 | 0.20 | 0.23 |
| 1 | 96 159 | 0.30 | -0.32 | 0.05 | 0.03 | 0.06 | 0.08 | 0.08 | 0.26 | 0.03 | -0.04 | 0.02 | -0.33 | -0.07 | -0.22 | -0.13 | 0.24 | 0.03 | 0.44 | -0.53 | 0.01 |
| 1 | 98 228 | 0.14 | -0.48 | 0.26 | -0.24 | 0.10 | 0.26 | -0.09 | 0.15 | -0.01 | -0.06 | -0.07 | -0.03 | -0.11 | 0.00 | 0.49 | -0.06 | 0.34 | -0.16 | -0.31 | -0.12 |
| 1 | 100 230 | 0.00 | 0.21 | -0.18 | -0.13 | 0.24 | 0.47 | -0.47 | 0.00 | -0.17 | -0.28 | 0.22 | -0.05 | 0.11 | -0.23 | 0.03 | 0.05 | 0.32 | -0.09 | -0.19 | 0.16 |
| 1 | 102 231 | -0.02 | -0.19 | 0.02 | -0.01 | 0.14 | -0.13 | 0.45 | -0.05 | -0.10 | -0.09 | -0.45 | -0.07 | 0.38 | -0.06 | 0.04 | -0.15 | -0.03 | -0.16 | -0.06 | 0.54 |
| 1 | 104 107 | -0.08 | -0.08 | 0.58 | 0.36 | 0.00 | 0.29 | -0.27 | -0.02 | -0.19 | 0.26 | 0.08 | -0.26 | 0.02 | 0.07 | -0.27 | -0.17 | -0.17 | 0.02 | -0.20 | 0.02 |
| 1 | 105 234 | 0.02 | -0.33 | 0.17 | 0.24 | 0.18 | -0.13 | 0.21 | 0.11 | -0.25 | -0.11 | -0.25 | 0.13 | 0.31 | 0.37 | -0.44 | 0.03 | -0.03 | 0.18 | -0.21 | -0.19 |
| 1 | 106 428 | -0.14 | -0.15 | 0.21 | -0.13 | 0.49 | -0.44 | 0.22 | 0.06 | -0.11 | 0.31 | -0.28 | -0.23 | -0.12 | 0.02 | 0.03 | 0.08 | 0.17 | -0.09 | -0.17 | 0.28 |
| 1 | 111 385 | -0.09 | -0.12 | -0.16 | 0.41 | 0.20 | -0.28 | 0.46 | -0.06 | -0.32 | 0.09 | 0.18 | -0.03 | 0.07 | 0.13 | -0.19 | -0.26 | -0.37 | 0.17 | 0.06 | 0.11 |
| 1 | 115 153 | -0.08 | -0.16 | -0.09 | 0.19 | -0.07 | -0.09 | 0.19 | -0.09 | -0.08 | -0.05 | -0.05 | -0.16 | -0.06 | -0.03 | 0.90 | -0.10 | -0.08 | -0.01 | -0.03 | -0.05 |
| 1 | 118 588 | -0.04 | -0.12 | 0.13 | 0.26 | -0.33 | 0.21 | 0.33 | 0.21 | -0.21 | -0.12 | -0.11 | -0.07 | 0.53 | 0.16 | -0.29 | 0.11 | -0.20 | -0.15 | -0.08 | -0.21 |
| 1 | 119 170 | 0.15 | -0.10 | 0.02 | 0.29 | -0.14 | 0.08 | -0.39 | 0.15 | 0.17 | 0.08 | -0.27 | -0.14 | 0.21 | 0.02 | -0.57 | -0.04 | 0.35 | 0.13 | 0.16 | -0.15 |
| 1 | 120 171 | -0.04 | -0.10 | -0.10 | -0.03 | -0.45 | -0.02 | -0.27 | 0.72 | 0.06 | -0.03 | -0.06 | -0.02 | 0.13 | 0.12 | 0.30 | -0.08 | -0.03 | -0.19 | 0.08 | 0.02 |
| 1 | 124 175 | -0.11 | 0.12 | 0.03 | 0.12 | 0.06 | 0.04 | -0.13 | 0.40 | -0.40 | -0.30 | -0.38 | -0.20 | -0.08 | -0.10 | 0.01 | 0.37 | 0.28 | 0.09 | 0.29 | -0.14 |
| 1 | 126 173 | -0.08 | -0.37 | 0.30 | 0.12 | -0.60 | 0.01 | 0.04 | -0.18 | 0.14 | 0.03 | -0.03 | -0.12 | 0.25 | 0.01 | 0.13 | 0.30 | -0.20 | 0.20 | -0.16 | 0.20 |
| 1 | 127 508 | 0.26 | -0.41 | -0.01 | 0.15 | -0.20 | -0.09 | -0.11 | 0.43 | -0.22 | -0.20 | -0.22 | 0.15 | -0.26 | 0.10 | 0.20 | 0.27 | 0.32 | -0.09 | 0.10 | -0.17 |
| 1 | 128 563 | -0.20 | 0.24 | -0.16 | -0.06 | -0.17 | -0.10 | -0.24 | -0.25 | 0.28 | 0.22 | 0.14 | 0.15 | 0.03 | 0.32 | -0.12 | -0.34 | -0.05 | 0.06 | 0.50 | -0.23 |
| 1 | 130 545 | 0.15 | -0.28 | 0.09 | 0.21 | -0.20 | -0.08 | 0.17 | 0.31 | -0.35 | 0.26 | 0.20 | -0.06 | -0.01 | 0.00 | 0.17 | 0.01 | -0.15 | -0.01 | -0.60 | 0.17 |
| 1 | 131 559 | -0.12 | -0.37 | 0.36 | 0.41 | 0.00 | 0.29 | 0.21 | -0.12 | -0.01 | 0.15 | -0.18 | -0.39 | -0.26 | 0.29 | 0.01 | 0.04 | -0.17 | 0.01 | -0.07 | -0.10 |
| 1 | 134 454 | 0.17 | 0.11 | -0.13 | 0.00 | -0.29 | -0.02 | 0.03 | 0.61 | 0.16 | 0.23 | -0.24 | 0.13 | 0.28 | -0.02 | -0.33 | -0.29 | -0.20 | -0.08 | -0.04 | -0.08 |
| 1 | 139 608 | 0.08 | -0.26 | -0.28 | 0.13 | 0.15 | 0.32 | -0.48 | 0.41 | 0.08 | -0.21 | 0.21 | -0.07 | -0.26 | -0.19 | 0.22 | -0.06 | 0.12 | -0.07 | 0.17 | -0.01 |
| 1 | 140 408 | 0.04 | -0.26 | -0.08 | 0.38 | 0.04 | -0.33 | 0.05 | 0.05 | 0.38 | -0.23 | 0.49 | 0.11 | -0.09 | -0.18 | -0.05 | -0.13 | 0.08 | 0.16 | -0.35 | -0.07 |
| 1 | 143 148 | -0.15 | -0.21 | -0.28 | 0.42 | 0.13 | -0.10 | -0.03 | 0.04 | -0.51 | 0.06 | 0.09 | -0.22 | -0.21 | 0.13 | 0.03 | 0.31 | 0.40 | 0.11 | -0.04 | 0.02 |
| 1 | 147 221 | 0.00 | 0.13 | 0.09 | 0.40 | -0.23 | 0.30 | 0.09 | -0.10 | 0.00 | -0.01 | -0.05 | -0.25 | -0.22 | -0.20 | 0.32 | -0.18 | -0.49 | 0.24 | 0.24 | -0.10 |
| 1 | 149 297 | -0.23 | -0.20 | 0.03 | 0.27 | 0.11 | 0.60 | 0.00 | -0.01 | -0.18 | 0.10 | -0.08 | -0.14 | 0.17 | 0.11 | -0.50 | 0.01 | -0.01 | 0.07 | 0.14 | -0.27 |
| 1 | 151 481 | 0.57 | 0.41 | -0.09 | 0.00 | -0.02 | 0.07 | -0.35 | 0.15 | -0.08 | 0.00 | -0.01 | 0.09 | -0.03 | -0.02 | 0.19 | -0.32 | -0.06 | -0.44 | -0.06 | 0.00 |
| 1 | 152 395 | 0.14 | -0.04 | 0.07 | 0.19 | 0.28 | -0.19 | 0.18 | -0.29 | 0.24 | 0.30 | 0.04 | 0.22 | -0.22 | 0.11 | -0.32 | -0.30 | -0.01 | -0.50 | 0.09 | 0.01 |
| 1 | 154 157 | 0.38 | -0.12 | 0.12 | 0.19 | -0.08 | 0.10 | 0.11 | -0.18 | 0.27 | -0.44 | -0.01 | -0.37 | -0.41 | 0.22 | -0.11 | 0.20 | -0.06 | 0.19 | 0.10 | -0.12 |
| 1 | 156 432 | 0.21 | -0.15 | -0.34 | -0.10 | 0.00 | -0.31 | -0.13 | 0.08 | 0.15 | 0.44 | -0.04 | -0.17 | -0.13 | 0.16 | -0.29 | 0.14 | 0.34 | -0.07 | -0.15 | 0.39 |
| 1 | 160 568 | -0.05 | -0.40 | -0.28 | 0.16 | 0.12 | 0.11 | 0.36 | -0.29 | 0.12 | 0.07 | -0.24 | 0.19 | -0.04 | -0.24 | 0.16 | 0.23 | 0.23 | -0.22 | -0.27 | 0.29 |
| 1 | 161 253 | -0.32 | -0.04 | 0.01 | 0.03 | 0.23 | 0.25 | -0.30 | 0.05 | -0.33 | -0.08 | -0.35 | -0.03 | 0.46 | 0.20 | 0.09 | -0.26 | 0.26 | 0.22 | -0.02 | -0.08 |
| 1 | 162 222 | -0.01 | -0.28 | -0.29 | -0.02 | 0.01 | -0.31 | 0.09 | 0.05 | 0.23 | -0.35 | 0.05 | -0.02 | -0.29 | 0.46 | 0.22 | 0.09 | 0.03 | 0.39 | -0.17 | 0.13 |
| 1 | 163 224 | 0.29 | -0.24 | -0.31 | -0.02 | -0.34 | -0.10 | 0.41 | -0.18 | 0.06 | -0.02 | 0.08 | 0.07 | -0.09 | 0.39 | -0.25 | 0.04 | 0.38 | 0.10 | -0.19 | -0.06 |
| 1 | 166 275 | 0.01 | -0.31 | 0.07 | 0.17 | -0.02 | -0.11 | -0.13 | 0.02 | -0.10 | -0.02 | 0.14 | 0.03 | -0.24 | -0.34 | 0.33 | -0.12 | -0.17 | 0.35 | -0.15 | 0.59 |
| 1 | 167 246 | 0.06 | 0.12 | 0.12 | -0.01 | 0.15 | -0.17 | -0.11 | 0.06 | -0.30 | 0.18 | -0.16 | 0.23 | 0.03 | -0.40 | 0.00 | 0.16 | 0.05 | -0.58 | 0.26 | 0.32 |
| 1 | 172 382 | -0.41 | -0.17 | -0.04 | 0.36 | -0.34 | 0.14 | 0.15 | -0.20 | 0.37 | -0.06 | -0.10 | -0.12 | 0.40 | -0.05 | 0.05 | -0.15 | -0.23 | 0.06 | 0.24 | 0.10 |
| 1 | 174 371 | -0.16 | -0.43 | 0.14 | 0.22 | 0.26 | -0.01 | -0.02 | -0.15 | -0.15 | -0.04 | -0.09 | -0.16 | 0.59 | -0.12 | 0.09 | 0.24 | -0.01 | -0.23 | 0.23 | -0.21 |
| 1 | 176 555 | 0.95 | -0.02 | -0.08 | -0.09 | 0.03 | -0.06 | -0.06 | -0.15 | -0.04 | -0.05 | -0.05 | -0.04 | -0.04 | 0.00 | -0.09 | -0.12 | -0.09 | 0.06 | 0.01 | -0.08 |
| 1 | 179 180 | 0.06 | -0.38 | 0.25 | 0.30 | -0.39 | 0.21 | 0.18 | 0.16 | -0.36 | -0.08 | 0.02 | -0.29 | -0.15 | 0.25 | -0.15 | 0.04 | 0.30 | 0.05 | 0.10 | -0.11 |
| 1 | 181 295 | 0.03 | -0.24 | 0.00 | -0.25 | 0.24 | 0.18 | -0.44 | -0.03 | -0.05 | -0.12 | -0.19 | -0.23 | -0.29 | 0.06 | 0.41 | 0.27 | 0.12 | 0.27 | 0.00 | 0.26 |
| 1 | 182 296 | 0.03 | -0.03 | 0.15 | 0.25 | -0.29 | -0.19 | 0.39 | -0.32 | 0.11 | 0.06 | 0.15 | 0.28 | 0.41 | -0.07 | -0.37 | -0.27 | -0.10 | 0.01 | -0.07 | -0.14 |

Table B.5: (continued)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 184 593 | 0.21 | -0.01 | -0.02 | -0.07 | 0.68 | 0.03 | -0.15 | 0.00 | -0.09 | -0.02 | 0.10 | 0.17 | -0.01 | -0.21 | -0.23 | 0.12 | 0.05 | -0.42 | -0.31 | 0.17 |
| 1 | 187 530 | 0.19 | 0.51 | 0.24 | 0.00 | -0.10 | -0.11 | 0.29 | -0.11 | -0.09 | 0.01 | -0.59 | 0.07 | 0.15 | -0.25 | -0.17 | -0.03 | -0.02 | -0.11 | 0.20 | -0.05 |
| 1 | 188 601 | 0.38 | -0.25 | 0.10 | 0.21 | -0.11 | 0.04 | 0.11 | 0.37 | 0.08 | -0.41 | -0.31 | -0.47 | -0.11 | 0.11 | 0.20 | 0.04 | 0.02 | 0.00 | 0.09 | -0.11 |
| 1 | 189 532 | 0.06 | -0.22 | 0.09 | 0.11 | -0.26 | -0.07 | -0.17 | 0.04 | -0.03 | 0.23 | 0.11 | -0.55 | 0.28 | 0.22 | -0.13 | 0.18 | 0.40 | -0.09 | 0.13 | -0.32 |
| 1 | 190 192 | 0.04 | -0.06 | -0.44 | -0.08 | 0.21 | 0.11 | 0.09 | -0.21 | 0.13 | -0.34 | 0.13 | -0.23 | 0.33 | -0.03 | 0.26 | -0.04 | 0.32 | 0.29 | -0.24 | -0.23 |
| 1 | 191 193 | -0.14 | -0.07 | 0.49 | -0.01 | -0.11 | -0.12 | -0.18 | 0.06 | -0.02 | 0.38 | -0.11 | 0.07 | -0.28 | 0.17 | -0.20 | -0.02 | -0.29 | -0.23 | 0.47 | 0.12 |
| 1 | 196 455 | 0.32 | -0.22 | 0.04 | 0.06 | -0.25 | 0.08 | 0.08 | 0.48 | 0.04 | -0.14 | -0.20 | -0.31 | 0.18 | 0.25 | -0.12 | -0.44 | 0.00 | 0.24 | 0.04 | -0.14 |
| 1 | 198 208 | -0.26 | -0.13 | 0.26 | 0.02 | 0.29 | -0.03 | 0.03 | -0.63 | -0.05 | 0.48 | -0.11 | -0.02 | -0.02 | 0.21 | -0.01 | 0.03 | -0.20 | -0.06 | 0.13 | 0.09 |
| 1 | 201 457 | 0.10 | -0.51 | 0.12 | 0.41 | -0.17 | 0.01 | 0.36 | 0.14 | 0.15 | -0.07 | 0.29 | -0.43 | -0.06 | 0.00 | -0.16 | -0.06 | -0.18 | 0.01 | 0.02 | 0.04 |
| 1 | 203 204 | 0.03 | 0.16 | -0.29 | 0.13 | 0.24 | -0.02 | 0.28 | -0.16 | -0.11 | 0.01 | -0.40 | -0.26 | -0.11 | -0.07 | -0.13 | 0.04 | 0.59 | -0.15 | 0.00 | 0.23 |
| 1 | 205 465 | -0.07 | -0.19 | 0.00 | -0.15 | 0.21 | -0.09 | -0.16 | 0.31 | 0.03 | 0.55 | -0.25 | 0.05 | 0.10 | -0.35 | 0.03 | -0.37 | -0.10 | 0.32 | 0.02 | 0.10 |
| 1 | 209 247 | -0.30 | -0.10 | -0.02 | 0.11 | -0.15 | -0.01 | -0.18 | 0.19 | -0.44 | 0.17 | 0.21 | -0.06 | -0.44 | 0.03 | 0.45 | 0.21 | 0.14 | -0.07 | 0.02 | 0.24 |
| 1 | 210 381 | -0.05 | -0.01 | 0.10 | 0.23 | 0.14 | -0.19 | 0.16 | -0.20 | 0.38 | -0.26 | -0.21 | -0.02 | 0.11 | 0.17 | -0.12 | 0.45 | -0.43 | 0.14 | -0.09 | -0.29 |
| 1 | 211 313 | 0.11 | -0.64 | 0.08 | 0.09 | 0.09 | 0.05 | -0.35 | 0.09 | 0.35 | 0.26 | 0.10 | 0.13 | 0.13 | -0.09 | 0.09 | 0.11 | 0.03 | -0.33 | -0.24 | -0.06 |
| 1 | 212 529 | -0.13 | -0.13 | 0.03 | 0.14 | -0.08 | 0.25 | -0.26 | -0.36 | -0.35 | 0.25 | -0.06 | 0.09 | -0.41 | 0.46 | 0.09 | 0.24 | -0.04 | 0.01 | 0.08 | 0.19 |
| 1 | 215 216 | -0.12 | -0.17 | 0.22 | 0.24 | -0.17 | 0.24 | 0.37 | -0.36 | 0.33 | -0.29 | 0.06 | 0.33 | 0.03 | -0.06 | 0.05 | -0.03 | 0.01 | -0.35 | -0.06 | -0.25 |
| 1 | 218 380 | 0.12 | 0.19 | -0.14 | -0.37 | 0.11 | -0.05 | -0.17 | 0.01 | 0.02 | 0.03 | 0.00 | 0.44 | -0.02 | -0.46 | -0.19 | 0.35 | 0.23 | 0.17 | -0.30 | 0.05 |
| 1 | 219 478 | 0.49 | -0.35 | 0.11 | 0.12 | -0.36 | -0.11 | 0.10 | 0.39 | 0.19 | 0.04 | 0.11 | -0.38 | 0.10 | -0.03 | 0.03 | -0.24 | -0.20 | 0.01 | -0.04 | 0.03 |
| 1 | 220 298 | -0.22 | 0.29 | -0.38 | -0.25 | -0.17 | 0.13 | 0.10 | 0.12 | -0.03 | -0.02 | 0.06 | 0.07 | 0.06 | -0.07 | -0.14 | 0.13 | 0.52 | 0.33 | -0.36 | -0.15 |
| 1 | 226 237 | 0.12 | 0.00 | 0.04 | -0.11 | -0.04 | -0.41 | 0.19 | -0.29 | -0.11 | 0.35 | 0.08 | -0.10 | 0.04 | 0.11 | -0.27 | 0.09 | 0.06 | -0.36 | 0.56 | 0.05 |
| 1 | 229 258 | 0.31 | -0.03 | 0.24 | 0.26 | -0.23 | -0.36 | -0.14 | 0.39 | -0.34 | -0.08 | 0.04 | -0.25 | 0.21 | -0.06 | 0.04 | -0.15 | 0.16 | 0.30 | -0.14 | -0.18 |
| 1 | 232 375 | 0.02 | -0.09 | -0.07 | 0.05 | 0.16 | -0.05 | 0.34 | -0.25 | -0.28 | -0.16 | -0.20 | 0.10 | -0.36 | 0.09 | 0.43 | 0.16 | -0.25 | 0.06 | 0.44 | -0.14 |
| 1 | 233 252 | -0.16 | -0.06 | 0.18 | -0.07 | -0.58 | 0.33 | 0.30 | 0.04 | -0.15 | 0.24 | -0.04 | -0.14 | 0.32 | 0.26 | -0.23 | -0.05 | 0.08 | -0.20 | 0.10 | -0.16 |
| 1 | 236 347 | -0.26 | 0.11 | 0.28 | 0.13 | -0.07 | -0.15 | -0.06 | -0.04 | 0.14 | 0.31 | -0.01 | -0.46 | -0.31 | 0.18 | 0.08 | -0.06 | 0.42 | 0.05 | -0.39 | 0.08 |
| 1 | 238 387 | 0.13 | -0.38 | -0.10 | 0.18 | -0.14 | 0.06 | 0.04 | 0.00 | -0.53 | 0.05 | -0.18 | -0.17 | 0.01 | 0.03 | 0.50 | 0.19 | 0.31 | -0.13 | 0.01 | 0.14 |
| 1 | 239 480 | -0.17 | -0.10 | -0.21 | -0.32 | 0.14 | -0.05 | -0.05 | -0.13 | 0.67 | 0.02 | 0.00 | 0.18 | 0.25 | -0.02 | 0.24 | -0.18 | 0.09 | -0.29 | -0.16 | 0.10 |
| 1 | 240 241 | -0.01 | -0.04 | 0.08 | 0.04 | -0.52 | -0.13 | 0.07 | 0.01 | 0.37 | 0.28 | 0.18 | 0.33 | -0.04 | -0.16 | -0.20 | -0.42 | -0.14 | -0.09 | 0.16 | 0.21 |
| 1 | 242 546 | -0.02 | -0.06 | -0.20 | 0.09 | -0.07 | 0.19 | 0.19 | 0.31 | -0.35 | -0.25 | -0.13 | -0.25 | 0.49 | -0.07 | -0.06 | 0.34 | -0.04 | -0.16 | -0.19 | 0.27 |
| 1 | 249 250 | -0.20 | 0.37 | -0.01 | -0.04 | -0.54 | 0.08 | 0.12 | -0.50 | -0.01 | 0.13 | 0.02 | 0.04 | -0.16 | 0.13 | -0.10 | 0.00 | 0.09 | 0.41 | 0.10 | 0.06 |
| 1 | 251 527 | 0.04 | -0.15 | -0.08 | 0.06 | 0.25 | -0.32 | 0.18 | 0.13 | -0.35 | 0.22 | -0.15 | 0.27 | -0.24 | 0.12 | 0.32 | -0.08 | -0.18 | -0.42 | 0.26 | 0.16 |
| 1 | 257 299 | -0.03 | -0.20 | 0.32 | 0.39 | -0.12 | -0.28 | -0.02 | -0.08 | 0.30 | 0.01 | -0.09 | 0.06 | -0.25 | 0.32 | -0.42 | 0.11 | -0.23 | -0.09 | 0.02 | 0.29 |
| 1 | 259 327 | -0.29 | -0.23 | 0.08 | 0.16 | 0.23 | 0.08 | 0.50 | -0.26 | -0.23 | -0.12 | 0.38 | -0.24 | -0.07 | 0.06 | 0.25 | 0.19 | -0.17 | -0.09 | -0.11 | -0.13 |
| 1 | 261 262 | -0.14 | 0.07 | -0.04 | 0.64 | -0.07 | -0.14 | 0.01 | -0.44 | -0.05 | -0.02 | -0.24 | -0.32 | 0.17 | 0.20 | 0.14 | 0.06 | 0.11 | 0.24 | -0.02 | -0.15 |
| 1 | 263 338 | -0.29 | 0.01 | 0.26 | -0.24 | 0.32 | 0.08 | -0.03 | 0.04 | 0.29 | -0.66 | 0.05 | 0.13 | 0.13 | 0.19 | -0.02 | -0.04 | -0.06 | -0.25 | 0.13 | -0.03 |
| 1 | 264 265 | -0.27 | -0.16 | 0.02 | 0.02 | 0.23 | -0.44 | 0.30 | -0.03 | -0.25 | 0.22 | -0.03 | -0.11 | 0.16 | 0.09 | -0.35 | 0.26 | 0.30 | -0.20 | -0.06 | 0.29 |
| 1 | 267 268 | 0.35 | 0.15 | 0.07 | -0.01 | 0.05 | 0.13 | -0.01 | -0.37 | -0.20 | -0.05 | 0.26 | -0.32 | 0.04 | 0.03 | -0.36 | -0.23 | 0.36 | -0.27 | 0.11 | 0.28 |
| 1 | 269 270 | -0.21 | -0.18 | -0.11 | 0.29 | 0.06 | 0.40 | -0.17 | 0.24 | 0.12 | -0.20 | 0.08 | -0.01 | 0.26 | -0.04 | -0.55 | 0.06 | 0.07 | 0.14 | -0.33 | 0.09 |
| 1 | 271 281 | -0.16 | 0.18 | 0.06 | -0.27 | -0.20 | 0.24 | 0.00 | 0.32 | -0.42 | 0.24 | -0.13 | 0.51 | 0.03 | -0.09 | -0.21 | 0.12 | -0.03 | 0.16 | -0.19 | -0.16 |
| 1 | 272 401 | -0.08 | -0.03 | -0.29 | -0.13 | 0.22 | 0.06 | -0.39 | 0.13 | -0.02 | 0.34 | -0.24 | -0.01 | 0.10 | -0.06 | -0.37 | -0.11 | 0.30 | 0.07 | 0.04 | 0.48 |
| 1 | 279 280 | -0.33 | -0.14 | 0.10 | 0.07 | -0.06 | -0.32 | -0.16 | 0.10 | 0.12 | 0.47 | -0.14 | 0.07 | -0.10 | 0.06 | -0.27 | -0.10 | -0.12 | -0.03 | 0.28 | 0.50 |
| 1 | 282 361 | -0.02 | -0.28 | -0.18 | 0.34 | -0.38 | 0.31 | -0.09 | 0.04 | -0.07 | 0.00 | 0.11 | 0.50 | 0.01 | -0.13 | -0.22 | -0.01 | -0.09 | 0.36 | 0.01 | -0.25 |
| 1 | 285 301 | 0.11 | -0.25 | 0.24 | -0.23 | -0.40 | -0.09 | -0.01 | 0.34 | 0.39 | -0.31 | 0.00 | 0.29 | -0.10 | -0.18 | 0.00 | 0.31 | 0.03 | 0.05 | 0.03 | -0.22 |
| 1 | 287 288 | -0.06 | 0.29 | -0.08 | 0.28 | -0.16 | 0.10 | 0.08 | -0.19 | -0.16 | -0.21 | 0.01 | -0.46 | 0.12 | 0.22 | 0.01 | 0.13 | 0.43 | 0.22 | -0.33 | -0.21 |
| 1 | 289 290 | -0.03 | -0.34 | -0.01 | -0.26 | -0.26 | -0.13 | 0.01 | 0.42 | 0.26 | 0.08 | 0.35 | 0.43 | 0.09 | -0.27 | 0.13 | -0.10 | -0.15 | -0.21 | 0.01 | -0.01 |
| 1 | 292 293 | 0.04 | -0.18 | 0.13 | -0.25 | 0.09 | 0.14 | 0.01 | 0.22 | -0.07 | 0.46 | -0.24 | -0.31 | 0.26 | -0.23 | -0.17 | -0.28 | -0.07 | -0.16 | 0.30 | 0.31 |
| 1 | 294 537 | 0.15 | -0.12 | -0.07 | 0.07 | 0.29 | -0.56 | -0.20 | 0.12 | 0.23 | 0.09 | -0.12 | 0.04 | -0.22 | -0.38 | 0.41 | 0.22 | -0.08 | 0.06 | 0.07 | 0.00 |
| 1 | 302 442 | 0.01 | 0.11 | 0.47 | -0.08 | 0.15 | -0.01 | 0.10 | 0.11 | -0.25 | -0.12 | 0.39 | 0.20 | -0.48 | -0.02 | -0.41 | -0.09 | -0.02 | -0.02 | 0.11 | -0.15 |
| 1 | 310 451 | 0.12 | -0.31 | -0.49 | 0.28 | 0.07 | 0.21 | -0.28 | 0.20 | -0.23 | -0.05 | 0.12 | 0.03 | -0.10 | 0.18 | -0.18 | 0.14 | 0.40 | 0.03 | -0.26 | 0.10 |
| 1 | 315 556 | 0.17 | -0.10 | -0.19 | -0.28 | 0.19 | 0.08 | -0.11 | 0.22 | -0.29 | -0.12 | 0.17 | 0.55 | 0.11 | -0.34 | 0.32 | -0.07 | -0.06 | -0.24 | -0.11 | 0.10 |

Table B.5: (continued)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 322 354 | -0.23 | -0.11 | 0.35 | 0.23 | 0.00 | 0.19 | 0.07 | -0.19 | 0.43 | -0.11 | 0.04 | 0.00 | 0.04 | 0.27 | -0.40 | -0.04 | -0.08 | -0.43 | 0.15 | -0.19 |
| 1 | 328 329 | -0.26 | -0.09 | 0.02 | -0.07 | 0.34 | -0.33 | -0.18 | 0.02 | -0.29 | 0.10 | 0.06 | -0.19 | -0.07 | 0.35 | -0.09 | 0.02 | 0.06 | 0.39 | -0.23 | 0.43 |
| 1 | 333 416 | 0.16 | -0.15 | 0.03 | 0.06 | 0.24 | 0.02 | -0.37 | 0.16 | -0.47 | 0.23 | 0.21 | 0.04 | 0.19 | -0.36 | 0.21 | 0.30 | -0.26 | -0.20 | -0.02 | 0.00 |
| 1 | 335 364 | 0.10 | 0.00 | -0.47 | -0.06 | 0.03 | 0.21 | -0.01 | -0.11 | 0.49 | 0.24 | 0.29 | -0.06 | 0.00 | -0.12 | 0.24 | -0.17 | -0.09 | -0.36 | 0.12 | -0.26 |
| 1 | 340 341 | -0.05 | -0.03 | -0.16 | -0.18 | 0.22 | 0.06 | 0.31 | -0.22 | -0.04 | 0.01 | -0.19 | -0.21 | -0.19 | -0.02 | 0.69 | 0.02 | -0.06 | -0.29 | 0.20 | 0.13 |
| 1 | 343 348 | -0.09 | -0.25 | -0.21 | 0.03 | 0.38 | 0.08 | 0.13 | 0.14 | 0.28 | -0.18 | -0.26 | -0.01 | 0.13 | 0.25 | -0.46 | -0.30 | 0.13 | 0.31 | -0.13 | 0.00 |
| 1 | 349 558 | -0.38 | -0.15 | 0.11 | 0.04 | 0.04 | 0.24 | 0.14 | -0.08 | 0.15 | 0.36 | 0.07 | 0.32 | 0.00 | 0.10 | 0.20 | -0.31 | -0.32 | -0.22 | 0.11 | -0.41 |
| 1 | 350 573 | 0.18 | -0.32 | -0.12 | -0.11 | 0.32 | 0.04 | -0.18 | -0.02 | -0.39 | 0.51 | 0.02 | -0.03 | 0.31 | -0.16 | 0.10 | -0.02 | -0.08 | -0.33 | 0.14 | 0.15 |
| 1 | 352 479 | 0.21 | -0.11 | -0.17 | -0.07 | -0.41 | -0.05 | 0.41 | -0.16 | -0.08 | -0.13 | 0.09 | -0.22 | 0.03 | -0.09 | 0.18 | -0.15 | 0.06 | 0.40 | 0.45 | -0.18 |
| 1 | 357 547 | -0.07 | -0.58 | -0.28 | -0.26 | 0.25 | -0.16 | -0.04 | 0.11 | -0.02 | 0.25 | 0.23 | 0.04 | -0.09 | 0.30 | -0.21 | -0.06 | 0.19 | -0.07 | 0.19 | 0.27 |
| 1 | 360 412 | 0.26 | -0.08 | 0.12 | 0.39 | -0.33 | -0.25 | 0.02 | 0.31 | -0.01 | -0.46 | 0.13 | 0.16 | 0.34 | -0.13 | -0.11 | 0.14 | -0.19 | 0.01 | -0.17 | -0.12 |
| 1 | 362 515 | 0.24 | -0.17 | -0.16 | -0.11 | 0.16 | -0.18 | -0.14 | 0.22 | -0.26 | 0.34 | 0.18 | -0.06 | 0.06 | -0.24 | 0.21 | 0.21 | 0.15 | -0.26 | -0.47 | 0.27 |
| 1 | 369 589 | 0.03 | -0.09 | 0.38 | 0.36 | -0.35 | -0.41 | 0.07 | -0.42 | 0.37 | -0.12 | 0.04 | 0.00 | 0.09 | -0.03 | -0.05 | 0.19 | -0.11 | 0.12 | 0.00 | -0.08 |
| 1 | 376 473 | -0.19 | -0.37 | 0.01 | 0.00 | -0.26 | 0.41 | -0.10 | 0.02 | 0.49 | -0.10 | -0.10 | -0.07 | 0.42 | -0.03 | 0.03 | -0.26 | 0.20 | -0.17 | 0.05 | 0.02 |
| 1 | 383 536 | -0.15 | -0.31 | 0.41 | 0.13 | 0.35 | -0.09 | 0.03 | 0.00 | 0.23 | -0.21 | -0.23 | -0.19 | -0.10 | 0.29 | 0.23 | 0.00 | 0.17 | -0.37 | 0.05 | -0.24 |
| 1 | 386 504 | -0.07 | -0.09 | -0.08 | 0.45 | -0.06 | -0.08 | -0.09 | -0.06 | -0.08 | -0.06 | -0.06 | -0.07 | -0.06 | -0.07 | 0.84 | -0.07 | -0.07 | -0.07 | -0.08 | -0.07 |
| 1 | 389 477 | 0.29 | -0.11 | 0.18 | 0.43 | -0.06 | -0.04 | 0.01 | 0.34 | -0.33 | -0.29 | -0.09 | -0.14 | -0.31 | 0.32 | 0.19 | 0.08 | -0.29 | 0.03 | -0.08 | -0.11 |
| 1 | 390 391 | 0.01 | -0.15 | -0.06 | 0.04 | -0.03 | 0.53 | -0.34 | 0.10 | 0.22 | -0.40 | 0.16 | 0.19 | -0.27 | -0.05 | 0.06 | 0.33 | 0.01 | -0.33 | -0.03 | 0.04 |
| 1 | 392 393 | -0.09 | -0.04 | -0.31 | -0.17 | 0.49 | -0.16 | 0.39 | 0.02 | 0.15 | -0.16 | -0.03 | -0.03 | -0.02 | -0.22 | -0.18 | -0.10 | 0.13 | 0.47 | 0.08 | -0.23 |
| 1 | 394 431 | 0.05 | -0.26 | 0.07 | 0.11 | -0.22 | -0.10 | -0.17 | 0.27 | -0.46 | 0.08 | 0.31 | -0.44 | 0.13 | 0.29 | -0.02 | 0.24 | 0.12 | -0.20 | 0.19 | 0.01 |
| 1 | 396 553 | 0.00 | -0.03 | 0.24 | 0.00 | -0.02 | -0.65 | -0.02 | 0.09 | 0.23 | -0.10 | -0.10 | 0.08 | -0.05 | -0.08 | -0.06 | 0.02 | -0.04 | 0.64 | -0.06 | -0.10 |
| 1 | 397 549 | 0.07 | -0.01 | 0.07 | 0.00 | 0.07 | 0.07 | 0.00 | 0.07 | -0.01 | -0.96 | 0.07 | 0.00 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 |
| 1 | 398 550 | 0.06 | -0.11 | 0.08 | 0.17 | 0.09 | 0.07 | -0.94 | 0.06 | 0.00 | 0.06 | 0.06 | -0.08 | 0.06 | 0.07 | 0.05 | 0.07 | 0.07 | 0.06 | 0.07 | 0.06 |
| 1 | 399 554 | -0.07 | 0.00 | -0.01 | 0.02 | 0.02 | -0.06 | 0.02 | -0.01 | -0.05 | -0.17 | -0.18 | 0.01 | 0.85 | -0.19 | -0.14 | -0.03 | -0.07 | 0.32 | -0.14 | -0.13 |
| 1 | 400 419 | -0.13 | -0.12 | 0.26 | 0.09 | 0.29 | 0.00 | -0.40 | -0.18 | -0.09 | -0.04 | -0.04 | 0.05 | 0.02 | -0.11 | 0.32 | 0.17 | -0.52 | 0.33 | -0.16 | 0.25 |
| 1 | 402 417 | 0.20 | -0.17 | 0.18 | 0.34 | -0.02 | -0.17 | 0.08 | 0.26 | 0.02 | -0.24 | -0.26 | -0.41 | -0.15 | -0.11 | 0.08 | 0.41 | 0.31 | 0.02 | -0.27 | -0.10 |
| 1 | 403 404 | -0.07 | -0.16 | 0.10 | -0.17 | -0.37 | 0.00 | -0.19 | -0.19 | -0.14 | 0.02 | -0.27 | 0.31 | -0.01 | 0.16 | -0.26 | 0.11 | 0.45 | 0.43 | 0.08 | 0.18 |
| 1 | 405 406 | 0.22 | -0.08 | -0.05 | 0.38 | 0.17 | 0.23 | -0.14 | 0.06 | 0.39 | -0.32 | 0.04 | -0.21 | 0.15 | 0.07 | -0.54 | -0.10 | 0.10 | -0.12 | -0.20 | -0.06 |
| 1 | 407 411 | -0.06 | -0.08 | -0.11 | -0.03 | 0.21 | -0.09 | 0.19 | 0.35 | -0.33 | 0.18 | 0.15 | -0.39 | -0.18 | 0.30 | -0.02 | -0.06 | -0.02 | 0.20 | -0.46 | 0.26 |
| 1 | 409 414 | 0.03 | 0.16 | 0.00 | -0.02 | 0.16 | -0.26 | -0.06 | -0.09 | -0.48 | 0.20 | -0.12 | -0.35 | 0.36 | 0.28 | 0.02 | -0.08 | 0.16 | -0.35 | 0.17 | 0.27 |
| 1 | 410 495 | 0.29 | -0.10 | -0.25 | 0.23 | 0.08 | 0.10 | 0.27 | 0.03 | -0.20 | 0.46 | -0.42 | 0.07 | -0.28 | -0.08 | -0.08 | -0.07 | 0.13 | -0.28 | -0.15 | 0.25 |
| 1 | 420 434 | 0.28 | 0.44 | 0.10 | -0.12 | -0.18 | 0.10 | 0.06 | 0.14 | -0.35 | -0.43 | 0.08 | 0.31 | 0.03 | -0.07 | 0.17 | 0.06 | -0.26 | 0.12 | -0.23 | -0.24 |
| 1 | 423 435 | -0.17 | -0.04 | -0.12 | 0.13 | 0.59 | 0.07 | -0.12 | -0.31 | -0.20 | 0.01 | -0.01 | 0.17 | -0.24 | -0.07 | 0.40 | -0.06 | -0.08 | 0.21 | -0.32 | 0.16 |
| 1 | 424 507 | 0.16 | -0.20 | 0.03 | 0.15 | -0.01 | -0.05 | -0.28 | 0.10 | 0.05 | -0.37 | -0.36 | 0.22 | 0.14 | 0.45 | -0.17 | -0.18 | 0.44 | -0.14 | 0.03 | -0.01 |
| 1 | 426 427 | -0.31 | -0.07 | 0.07 | -0.36 | 0.20 | -0.23 | 0.48 | 0.35 | 0.14 | 0.19 | -0.09 | -0.23 | 0.02 | 0.15 | -0.14 | -0.37 | 0.01 | 0.00 | 0.12 | 0.06 |
| 1 | 429 430 | -0.38 | -0.15 | 0.28 | 0.52 | 0.30 | -0.17 | -0.28 | -0.24 | 0.08 | 0.26 | 0.13 | -0.08 | -0.22 | 0.11 | 0.08 | -0.22 | -0.01 | -0.09 | -0.01 | 0.11 |
| 1 | 437 438 | 0.14 | -0.38 | 0.27 | -0.10 | 0.14 | 0.04 | -0.33 | 0.02 | -0.04 | -0.37 | 0.13 | 0.16 | 0.03 | 0.04 | 0.00 | 0.58 | 0.01 | -0.05 | -0.30 | 0.02 |
| 1 | 439 440 | -0.07 | -0.42 | 0.11 | -0.28 | 0.12 | 0.27 | -0.22 | 0.12 | 0.30 | -0.13 | 0.45 | -0.02 | -0.10 | -0.13 | 0.10 | -0.04 | 0.04 | -0.10 | -0.32 | 0.33 |
| 1 | 441 579 | 0.19 | -0.13 | 0.11 | 0.28 | -0.18 | 0.02 | 0.22 | -0.24 | 0.11 | -0.40 | -0.25 | 0.10 | -0.40 | 0.23 | 0.22 | -0.07 | -0.14 | 0.39 | 0.10 | -0.15 |
| 1 | 445 446 | 0.41 | -0.13 | 0.21 | -0.28 | -0.29 | 0.01 | -0.20 | 0.31 | -0.07 | 0.06 | 0.09 | -0.38 | 0.00 | -0.01 | 0.27 | 0.29 | -0.10 | -0.33 | -0.06 | 0.19 |
| 1 | 447 448 | 0.19 | -0.20 | 0.00 | -0.39 | -0.21 | -0.19 | -0.44 | 0.18 | 0.18 | 0.21 | 0.23 | 0.27 | 0.10 | 0.17 | 0.16 | 0.05 | -0.39 | -0.08 | 0.17 | -0.01 |
| 1 | 459 461 | 0.31 | -0.10 | -0.12 | 0.12 | -0.13 | -0.16 | 0.43 | 0.43 | 0.12 | 0.08 | -0.02 | -0.53 | 0.02 | 0.06 | -0.19 | -0.33 | 0.04 | 0.01 | 0.03 | -0.10 |
| 1 | 460 607 | -0.07 | -0.03 | -0.36 | -0.04 | -0.13 | -0.03 | 0.01 | -0.41 | 0.05 | 0.07 | 0.37 | 0.41 | -0.14 | 0.09 | -0.16 | 0.20 | -0.26 | -0.16 | 0.35 | 0.23 |
| 1 | 462 463 | 0.48 | -0.18 | 0.13 | -0.07 | 0.05 | 0.16 | -0.20 | 0.25 | 0.05 | -0.02 | 0.01 | -0.33 | 0.18 | -0.10 | -0.16 | 0.27 | 0.20 | -0.01 | -0.20 | -0.50 |
| 1 | 464 597 | 0.13 | 0.22 | 0.25 | 0.18 | -0.60 | 0.00 | -0.04 | -0.16 | -0.18 | 0.05 | 0.01 | 0.02 | -0.08 | 0.07 | -0.04 | -0.20 | 0.13 | -0.09 | 0.55 | -0.19 |
| 1 | 471 472 | 0.19 | -0.42 | -0.17 | -0.35 | 0.18 | -0.18 | -0.43 | 0.04 | -0.16 | 0.27 | 0.27 | -0.07 | 0.19 | 0.26 | 0.09 | 0.06 | -0.07 | 0.14 | -0.07 | 0.25 |
| 1 | 474 548 | 0.83 | -0.03 | -0.37 | -0.01 | -0.06 | 0.33 | -0.01 | -0.02 | -0.03 | -0.08 | -0.06 | -0.18 | -0.05 | -0.05 | 0.00 | -0.02 | -0.02 | -0.06 | -0.05 | -0.07 |
| 1 | 475 511 | 0.27 | -0.17 | -0.27 | 0.02 | 0.31 | -0.26 | 0.04 | 0.09 | -0.07 | -0.11 | -0.11 | -0.14 | 0.49 | -0.46 | 0.26 | 0.08 | -0.10 | 0.07 | -0.15 | 0.20 |
| 1 | 482 484 | -0.06 | -0.45 | 0.14 | 0.38 | 0.21 | -0.25 | -0.34 | 0.13 | 0.23 | 0.09 | -0.12 | -0.26 | 0.11 | -0.29 | 0.22 | 0.00 | 0.28 | 0.09 | -0.11 | 0.00 |

Table B.5: (continued)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 483 | 499 | 0.32 | -0.31 | 0.01 | 0.21 | 0.19 | -0.15 | -0.10 | 0.03 | -0.12 | -0.34 | -0.26 | -0.44 | 0.18 | 0.03 | 0.46 | 0.05 | 0.09 | 0.18 | -0.06 | 0.03 |
| 1 | 486 | 487 | -0.04 | 0.12 | -0.09 | -0.08 | 0.20 | 0.51 | -0.33 | -0.15 | -0.18 | 0.02 | 0.18 | 0.18 | -0.16 | -0.42 | -0.04 | -0.24 | 0.22 | 0.29 | -0.13 | 0.15 |
| 1 | 488 | 493 | 0.15 | -0.21 | -0.28 | 0.08 | -0.39 | -0.28 | -0.05 | -0.07 | 0.16 | 0.20 | 0.00 | 0.32 | 0.24 | -0.12 | -0.25 | -0.12 | -0.16 | 0.31 | 0.41 | 0.05 |
| 1 | 489 | 494 | 0.22 | 0.00 | -0.42 | -0.13 | -0.18 | -0.25 | -0.13 | -0.26 | -0.13 | 0.27 | 0.37 | -0.01 | 0.35 | 0.23 | 0.31 | -0.17 | -0.22 | 0.06 | 0.06 | 0.03 |
| 1 | 501 | 502 | -0.12 | -0.03 | 0.24 | 0.51 | -0.12 | 0.10 | 0.35 | -0.03 | -0.02 | -0.25 | -0.20 | -0.44 | -0.23 | 0.04 | -0.15 | 0.12 | 0.06 | 0.31 | 0.05 | -0.19 |
| 1 | 503 | 602 | -0.07 | -0.07 | 0.20 | 0.44 | 0.04 | -0.13 | 0.27 | 0.21 | 0.13 | -0.08 | -0.38 | 0.25 | 0.27 | 0.01 | -0.33 | -0.09 | -0.28 | -0.05 | -0.34 | -0.01 |
| 1 | 510 | 539 | 0.01 | -0.66 | 0.23 | 0.36 | -0.01 | 0.16 | 0.34 | 0.05 | -0.12 | -0.12 | -0.06 | 0.22 | -0.13 | -0.15 | 0.09 | -0.13 | -0.19 | -0.03 | 0.22 | -0.07 |
| 1 | 512 | 605 | -0.02 | 0.30 | 0.05 | -0.18 | 0.09 | 0.05 | -0.09 | 0.04 | 0.44 | 0.13 | 0.11 | -0.42 | 0.16 | -0.10 | 0.06 | -0.02 | -0.01 | -0.02 | -0.64 | 0.09 |
| 1 | 513 | 570 | -0.13 | -0.30 | -0.08 | 0.15 | 0.21 | -0.28 | 0.38 | 0.03 | 0.08 | -0.12 | -0.05 | 0.00 | 0.13 | 0.47 | -0.03 | -0.49 | 0.20 | -0.10 | 0.11 | -0.18 |
| 1 | 514 | 543 | 0.49 | -0.46 | 0.12 | 0.11 | 0.10 | 0.05 | -0.37 | -0.20 | 0.04 | -0.35 | -0.03 | 0.17 | -0.15 | 0.08 | 0.09 | 0.16 | 0.06 | -0.22 | 0.23 | 0.08 |
| 1 | 516 | 517 | 0.13 | 0.14 | 0.13 | 0.27 | 0.18 | -0.41 | -0.07 | 0.26 | 0.11 | 0.06 | -0.22 | -0.21 | -0.20 | -0.09 | -0.48 | 0.30 | -0.15 | 0.32 | 0.00 | -0.08 |
| 1 | 518 | 519 | -0.02 | -0.42 | 0.02 | 0.20 | 0.57 | 0.09 | 0.29 | -0.08 | -0.06 | -0.04 | -0.20 | 0.13 | 0.00 | -0.02 | -0.40 | -0.02 | -0.22 | 0.08 | 0.24 | -0.14 |
| 1 | 522 | 523 | 0.26 | -0.20 | -0.29 | 0.20 | -0.35 | -0.16 | -0.10 | -0.22 | -0.04 | -0.20 | 0.27 | 0.45 | 0.12 | 0.14 | 0.27 | -0.01 | -0.11 | -0.15 | 0.28 | -0.17 |
| 1 | 524 | 525 | -0.03 | -0.18 | -0.02 | 0.14 | 0.06 | 0.07 | -0.07 | 0.00 | 0.07 | 0.13 | 0.35 | -0.03 | 0.21 | -0.14 | -0.15 | 0.18 | 0.23 | -0.65 | -0.38 | 0.21 |
| 1 | 531 | 533 | -0.10 | -0.45 | 0.00 | 0.14 | -0.09 | -0.09 | 0.48 | -0.27 | -0.29 | 0.19 | -0.05 | -0.31 | 0.10 | 0.24 | 0.05 | -0.21 | 0.08 | 0.17 | 0.19 | 0.22 |
| 1 | 541 | 542 | 0.08 | -0.21 | 0.59 | -0.32 | -0.17 | 0.01 | 0.37 | -0.02 | -0.07 | 0.01 | 0.08 | 0.11 | -0.05 | -0.05 | -0.26 | -0.06 | 0.23 | 0.20 | -0.35 | -0.10 |
| 1 | 557 | 577 | 0.17 | -0.06 | 0.02 | -0.22 | 0.26 | -0.08 | -0.05 | -0.07 | 0.31 | -0.23 | 0.12 | -0.42 | 0.15 | -0.19 | -0.36 | 0.42 | 0.03 | 0.06 | 0.31 | -0.15 |
| 1 | 561 | 571 | -0.19 | -0.13 | -0.02 | 0.05 | 0.13 | 0.13 | 0.26 | -0.42 | 0.08 | 0.07 | 0.11 | 0.14 | 0.22 | -0.16 | 0.40 | -0.02 | -0.12 | -0.57 | 0.17 | -0.15 |
| 1 | 574 | 575 | 0.04 | -0.41 | -0.04 | -0.17 | 0.21 | 0.00 | -0.14 | 0.01 | 0.12 | -0.10 | -0.09 | -0.47 | 0.04 | 0.09 | -0.06 | 0.17 | -0.04 | 0.49 | 0.43 | -0.07 |
| 1 | 582 | 585 | 0.23 | -0.51 | 0.03 | 0.24 | 0.41 | 0.01 | 0.17 | 0.34 | -0.11 | -0.24 | -0.17 | -0.07 | -0.02 | 0.02 | 0.10 | 0.13 | 0.09 | -0.30 | -0.28 | -0.07 |
| 1 | 583 | 584 | 0.14 | -0.43 | -0.22 | 0.11 | 0.32 | -0.21 | 0.12 | -0.12 | -0.28 | 0.12 | 0.11 | -0.07 | 0.24 | -0.21 | 0.06 | 0.30 | 0.21 | -0.44 | 0.11 | 0.12 |

Table B.6: PCA Generated Amino Acid Indices With Average Linkage 1 and 0.45

| Variance | Indices | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 12 13 | -0.23 | 0.48 | 0.18 | -0.02 | -0.45 | 0.14 | -0.08 | -0.10 | 0.29 | 0.36 | -0.29 | 0.06 | -0.10 | -0.25 | 0.05 | 0.01 | 0.01 | 0.19 | -0.14 | -0.10 |
| 1 | 18 604 | 0.15 | 0.05 | -0.18 | 0.05 | -0.25 | 0.03 | -0.13 | -0.17 | 0.29 | -0.42 | 0.10 | 0.27 | 7e-05 | -0.41 | 0.43 | -0.03 | -0.09 | 0.35 | 0.00 | -0.05 |
| 1 | 25 591 | -0.05 | 0.15 | -0.11 | -0.19 | 0.02 | 0.21 | 0.12 | -0.17 | -0.21 | -0.21 | 0.08 | 0.02 | 0.20 | -0.14 | 0.19 | 0.39 | 0.03 | 0.39 | -0.15 | -0.57 |
| 1 | 27 330 | 0.21 | -0.22 | -0.22 | -0.31 | -0.02 | 0.00 | 0.13 | -0.06 | -0.09 | 0.04 | -0.20 | 0.40 | -0.13 | 0.22 | -0.10 | 0.16 | 0.21 | -0.40 | -0.09 | 0.47 |
| 1 | 31 284 | -0.21 | -0.19 | -0.25 | -0.08 | -0.04 | -0.10 | -0.19 | 0.03 | -0.02 | 0.57 | 0.15 | 0.11 | -0.26 | 0.02 | 0.56 | 0.01 | -0.02 | -0.09 | -0.15 | 0.15 |
| 1 | 35 505 | 0.03 | 0.04 | 0.11 | 0.04 | 0.41 | 0.04 | -0.11 | -0.04 | -0.08 | -0.20 | 0.34 | -0.10 | -0.57 | 0.22 | -0.36 | -0.11 | -0.10 | 0.30 | 0.13 | 0.00 |
| 1 | 39 225 | 0.22 | -0.43 | -0.14 | 0.56 | -0.12 | -0.33 | 0.18 | 0.24 | 0.01 | 0.07 | -0.08 | 0.06 | -0.23 | 0.16 | -0.22 | -0.10 | 0.09 | -0.16 | 0.04 | 0.18 |
| 1 | 40 53 | 0.03 | 0.14 | -0.07 | -0.14 | 0.00 | -0.18 | -0.24 | -0.21 | -0.31 | 0.06 | 0.12 | -0.22 | 0.16 | 0.23 | 0.72 | -0.07 | 0.07 | 0.09 | -0.19 | 0.01 |
| 1 | 44 334 | 0.25 | 0.12 | -0.37 | 0.21 | -0.08 | 0.39 | 0.10 | -0.19 | -0.30 | 0.13 | 0.03 | -0.12 | 0.13 | -0.54 | 0.25 | -0.13 | -0.04 | 0.08 | -0.03 | 0.11 |
| 1 | 52 346 | 0.09 | -0.13 | 0.12 | 0.01 | 0.02 | -0.26 | 0.23 | 0.08 | -0.28 | -0.30 | 0.13 | -0.05 | -0.18 | -0.43 | 0.47 | 0.06 | -0.22 | 0.19 | 0.30 | 0.14 |
| 1 | 65 135 | 0.04 | 0.23 | -0.38 | -0.32 | 0.15 | -0.16 | -0.43 | -0.16 | 0.37 | 0.16 | 0.05 | 0.15 | 0.00 | -0.03 | -0.11 | 0.07 | 0.22 | -0.18 | -0.05 | 0.39 |
| 1 | 74 90 | 0.01 | -0.29 | -0.23 | 0.25 | -0.06 | -0.13 | 0.08 | -0.27 | 0.02 | -0.13 | 0.21 | -0.33 | 0.03 | 0.41 | -0.15 | -0.14 | 0.18 | 0.53 | 0.04 | -0.02 |
| 1 | 75 418 | 0.13 | -0.52 | -0.41 | -0.01 | 0.00 | -0.24 | -0.09 | 0.30 | -0.03 | 0.23 | 0.27 | -0.25 | -0.07 | 0.02 | -0.09 | 0.11 | 0.01 | 0.12 | 0.15 | 0.37 |
| 1 | 86 509 | 0.09 | -0.22 | -0.29 | 0.19 | 0.12 | 0.05 | 0.04 | 0.13 | 0.07 | -0.38 | -0.06 | 0.20 | 0.05 | -0.16 | 0.64 | -0.10 | -0.09 | -0.21 | 0.17 | -0.25 |
| 1 | 92 450 | 0.03 | -0.29 | 0.33 | -0.30 | 0.00 | -0.09 | 0.37 | -0.05 | -0.09 | -0.23 | 0.04 | -0.28 | 0.02 | 0.04 | -0.11 | 0.42 | 0.44 | -0.03 | -0.03 | -0.22 |
| 1 | 100 230 | 0.00 | -0.21 | 0.18 | 0.13 | -0.24 | -0.47 | 0.47 | 0.00 | 0.17 | 0.28 | -0.22 | 0.05 | -0.11 | 0.23 | -0.03 | -0.05 | -0.32 | 0.09 | 0.19 | -0.16 |
| 1 | 104 107 | 0.08 | 0.08 | -0.58 | -0.36 | 0.00 | -0.29 | 0.27 | 0.02 | 0.19 | -0.26 | -0.08 | 0.26 | -0.02 | -0.07 | 0.27 | 0.17 | 0.17 | -0.02 | 0.20 | -0.02 |
| 1 | 106 428 | -0.14 | -0.15 | 0.21 | -0.13 | 0.49 | -0.44 | 0.22 | 0.06 | -0.11 | 0.31 | -0.28 | -0.23 | -0.12 | 0.02 | 0.03 | 0.08 | 0.17 | -0.09 | -0.17 | 0.28 |
| 1 | 114 397 549 | 0.00 | -0.62 | 0.19 | 0.29 | -0.21 | 0.17 | 0.22 | 0.02 | -0.51 | -0.04 | -0.04 | 0.04 | -0.09 | 0.08 | -0.05 | 0.12 | 0.08 | 0.15 | 0.20 | -0.02 |
| 1 | 121 226 | 0.05 | -0.25 | -0.05 | -0.29 | 0.07 | 0.31 | -0.33 | 0.16 | -0.53 | 0.14 | -0.21 | -0.11 | 0.23 | 0.28 | -0.05 | 0.19 | -0.06 | 0.18 | 0.24 | 0.01 |
| 1 | 124 175 | 0.11 | -0.12 | -0.03 | -0.12 | -0.06 | -0.04 | 0.13 | -0.40 | 0.40 | 0.30 | 0.38 | 0.20 | 0.08 | 0.10 | -0.01 | -0.37 | -0.28 | -0.09 | -0.29 | 0.14 |
| 1 | 139 608 | 0.08 | -0.26 | -0.28 | 0.13 | 0.15 | 0.32 | -0.48 | 0.41 | 0.08 | -0.21 | 0.21 | -0.07 | -0.26 | -0.19 | 0.22 | -0.06 | 0.12 | -0.07 | 0.17 | -0.01 |
| 1 | 145 547 | -0.30 | 0.37 | -0.25 | -0.06 | -0.08 | 0.04 | 0.17 | -0.08 | 0.37 | -0.03 | -0.05 | -0.12 | -0.16 | 0.43 | -0.50 | 0.00 | 0.01 | 0.21 | 0.08 | -0.05 |
| 1 | 149 297 | -0.23 | -0.20 | 0.03 | 0.27 | 0.11 | 0.60 | 0.00 | -0.01 | -0.18 | 0.10 | -0.08 | -0.14 | 0.17 | 0.11 | -0.50 | 0.01 | -0.01 | 0.07 | 0.14 | -0.27 |
| 1 | 151 481 | -0.57 | -0.41 | 0.09 | 0.00 | 0.02 | -0.07 | 0.35 | -0.15 | 0.08 | 0.00 | 0.01 | -0.09 | 0.03 | 0.02 | -0.19 | 0.32 | 0.06 | 0.44 | 0.06 | 0.00 |
| 1 | 152 395 | 0.14 | -0.04 | 0.07 | 0.19 | 0.28 | -0.19 | 0.18 | -0.29 | 0.24 | 0.30 | 0.04 | 0.22 | -0.22 | 0.11 | -0.32 | -0.30 | -0.01 | -0.50 | 0.09 | 0.01 |
| 1 | 156 432 | 0.21 | -0.15 | -0.34 | -0.10 | 0.00 | -0.31 | -0.13 | 0.08 | 0.15 | 0.44 | -0.04 | -0.17 | -0.13 | 0.16 | -0.29 | 0.14 | 0.34 | -0.07 | -0.15 | 0.39 |
| 1 | 167 246 | 0.06 | 0.12 | 0.12 | -0.01 | 0.15 | -0.17 | -0.11 | 0.06 | -0.30 | 0.18 | -0.16 | 0.23 | 0.03 | -0.40 | 0.00 | 0.16 | 0.05 | -0.58 | 0.26 | 0.32 |
| 1 | 187 237 | -0.27 | 0.47 | -0.03 | -0.14 | 0.32 | 0.05 | -0.24 | -0.10 | 0.22 | -0.07 | -0.03 | -0.08 | -0.04 | -0.48 | 0.26 | 0.06 | -0.06 | 0.20 | -0.24 | 0.21 |
| 1 | 196 455 | 0.32 | -0.22 | 0.04 | 0.06 | -0.25 | 0.08 | 0.08 | 0.48 | 0.04 | -0.14 | -0.20 | -0.31 | 0.18 | 0.25 | -0.12 | -0.44 | 0.00 | 0.24 | 0.04 | -0.14 |
| 1 | 198 208 | -0.26 | -0.13 | 0.26 | 0.02 | 0.29 | -0.03 | 0.03 | -0.63 | -0.05 | 0.48 | -0.11 | -0.02 | -0.02 | 0.21 | -0.01 | 0.03 | -0.20 | -0.06 | 0.13 | 0.09 |
| 1 | 205 465 | 0.07 | 0.19 | 0.00 | 0.15 | -0.21 | 0.09 | 0.16 | -0.31 | -0.03 | -0.55 | 0.25 | -0.05 | -0.10 | 0.35 | -0.03 | 0.37 | 0.10 | -0.32 | -0.02 | -0.10 |
| 1 | 209 247 | 0.30 | 0.10 | 0.02 | -0.11 | 0.15 | 0.01 | 0.18 | -0.19 | 0.44 | -0.17 | -0.21 | 0.06 | 0.44 | -0.03 | -0.45 | -0.21 | -0.14 | 0.07 | -0.02 | -0.24 |
| 1 | 213 388 | -0.28 | 0.36 | -0.17 | -0.08 | -0.29 | 0.14 | 0.04 | 0.42 | -0.32 | 0.30 | -0.15 | 0.12 | -0.24 | -0.03 | -0.24 | 0.00 | 0.20 | -0.12 | 0.09 | 0.26 |
| 1 | 219 478 | -0.49 | 0.35 | -0.11 | -0.12 | 0.36 | 0.11 | -0.10 | -0.39 | -0.19 | -0.04 | -0.11 | 0.38 | -0.10 | 0.03 | -0.03 | 0.24 | 0.20 | -0.01 | 0.04 | -0.03 |
| 1 | 220 298 | 0.22 | -0.29 | 0.38 | 0.25 | 0.17 | -0.13 | -0.10 | -0.12 | 0.03 | 0.02 | -0.06 | -0.07 | -0.06 | 0.07 | 0.14 | -0.13 | -0.52 | -0.33 | 0.36 | 0.15 |
| 1 | 233 252 | 0.16 | 0.06 | -0.18 | 0.07 | 0.58 | -0.33 | -0.30 | -0.04 | 0.15 | -0.24 | 0.04 | 0.14 | -0.32 | -0.26 | 0.23 | 0.05 | -0.08 | 0.20 | -0.10 | 0.16 |
| 1 | 249 250 | 0.20 | -0.37 | 0.01 | 0.04 | 0.54 | -0.08 | -0.12 | 0.50 | 0.01 | -0.13 | -0.02 | -0.04 | 0.16 | -0.13 | 0.10 | 0.00 | -0.09 | -0.41 | -0.10 | -0.06 |
| 1 | 254 338 | 0.23 | 0.03 | -0.17 | 0.37 | -0.38 | -0.05 | 0.08 | 0.18 | 0.13 | -0.04 | -0.17 | -0.06 | 0.17 | -0.08 | -0.07 | 0.17 | 0.09 | 0.02 | -0.66 | 0.20 |
| 1 | 267 268 | -0.35 | -0.15 | -0.07 | 0.01 | -0.05 | -0.13 | 0.01 | 0.37 | 0.20 | 0.05 | -0.26 | 0.32 | -0.04 | -0.03 | 0.36 | 0.23 | -0.36 | 0.27 | -0.11 | -0.28 |
| 1 | 269 270 | 0.21 | 0.18 | 0.11 | -0.29 | -0.06 | -0.40 | 0.17 | -0.24 | -0.12 | 0.20 | -0.08 | 0.01 | -0.26 | 0.04 | 0.55 | -0.06 | -0.07 | -0.14 | 0.33 | -0.09 |
| 1 | 271 281 | 0.16 | -0.18 | -0.06 | 0.27 | 0.20 | -0.24 | 0.00 | -0.32 | 0.42 | -0.24 | 0.13 | -0.51 | -0.03 | 0.09 | 0.21 | -0.12 | 0.03 | -0.16 | 0.19 | 0.16 |
| 1 | 289 290 | 0.03 | 0.34 | 0.01 | 0.26 | 0.26 | 0.13 | -0.01 | -0.42 | -0.26 | -0.08 | -0.35 | -0.43 | -0.09 | 0.27 | -0.13 | 0.10 | 0.15 | 0.21 | -0.01 | 0.01 |
| 1 | 305 568 | 0.07 | -0.28 | -0.12 | 0.27 | -0.40 | 0.14 | 0.03 | -0.29 | 0.02 | 0.22 | -0.32 | 0.20 | -0.37 | 0.06 | -0.18 | 0.28 | 0.16 | 0.18 | 0.11 | 0.22 |
| 1 | 310 451 | 0.12 | -0.31 | -0.49 | 0.28 | 0.07 | 0.21 | -0.28 | 0.20 | -0.23 | -0.05 | 0.12 | 0.03 | -0.10 | 0.18 | -0.18 | 0.14 | 0.40 | 0.03 | -0.26 | 0.10 |
| 1 | 322 354 | 0.23 | 0.11 | -0.35 | -0.23 | 0.00 | -0.19 | -0.07 | 0.19 | -0.43 | 0.11 | -0.04 | 0.00 | -0.04 | -0.27 | 0.40 | 0.04 | 0.08 | 0.43 | -0.15 | 0.19 |
| 1 | 335 364 | 0.10 | 0.00 | -0.47 | -0.06 | 0.03 | 0.21 | -0.01 | -0.11 | 0.49 | 0.24 | 0.29 | -0.06 | 0.00 | -0.12 | 0.24 | -0.17 | -0.09 | -0.36 | 0.12 | -0.26 |
| 1 | 403 404 | 0.07 | 0.16 | -0.10 | 0.17 | 0.37 | 0.00 | 0.19 | 0.19 | 0.14 | -0.02 | 0.27 | -0.31 | 0.01 | -0.16 | 0.26 | -0.11 | -0.45 | -0.43 | -0.08 | -0.18 |

Table B.6: (continued)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 407 411 | 0.06 | 0.08 | 0.11 | 0.03 | -0.21 | 0.09 | -0.19 | -0.35 | 0.33 | -0.18 | -0.15 | 0.39 | 0.18 | -0.30 | 0.02 | 0.06 | 0.02 | -0.20 | 0.46 | -0.26 |
| 1 | 429 430 | -0.38 | -0.15 | 0.28 | 0.52 | 0.30 | -0.17 | -0.28 | -0.24 | 0.08 | 0.26 | 0.13 | -0.08 | -0.22 | 0.11 | 0.08 | -0.22 | -0.01 | -0.09 | -0.01 | 0.11 |
| 1 | 441 579 | 0.19 | -0.13 | 0.11 | 0.28 | -0.18 | 0.02 | 0.22 | -0.24 | 0.11 | -0.40 | -0.25 | 0.10 | -0.40 | 0.23 | 0.22 | -0.07 | -0.14 | 0.39 | 0.10 | -0.15 |
| 1 | 462 463 | 0.48 | -0.18 | 0.13 | -0.07 | 0.05 | 0.16 | -0.20 | 0.25 | 0.05 | -0.02 | 0.01 | -0.33 | 0.18 | -0.10 | -0.16 | 0.27 | 0.20 | -0.01 | -0.20 | -0.50 |
| 1 | 464 597 | 0.13 | 0.22 | 0.25 | 0.18 | -0.60 | 0.00 | -0.04 | -0.16 | -0.18 | 0.05 | 0.01 | 0.02 | -0.08 | 0.07 | -0.04 | -0.20 | 0.13 | -0.09 | 0.55 | -0.19 |
| 1 | 470 580 | 0.10 | -0.19 | -0.27 | -0.22 | -0.20 | -0.15 | -0.13 | 0.08 | -0.16 | 0.23 | 0.22 | 0.45 | 0.02 | -0.27 | 0.23 | -0.29 | -0.06 | 0.41 | 0.01 | 0.20 |
| 1 | 489 494 | 0.22 | 0.00 | -0.42 | -0.13 | -0.18 | -0.25 | -0.13 | -0.26 | -0.13 | 0.27 | 0.37 | -0.01 | 0.35 | 0.23 | 0.31 | -0.17 | -0.22 | 0.06 | 0.06 | 0.03 |
| 1 | 501 502 | -0.12 | -0.03 | 0.24 | 0.51 | -0.12 | 0.10 | 0.35 | -0.03 | -0.02 | -0.25 | -0.20 | -0.44 | -0.23 | 0.04 | -0.15 | 0.12 | 0.06 | 0.31 | 0.05 | -0.19 |
| 1 | 513 570 | 0.13 | 0.30 | 0.08 | -0.15 | -0.21 | 0.28 | -0.38 | -0.03 | -0.08 | 0.12 | 0.05 | 0.00 | -0.13 | -0.47 | 0.03 | 0.49 | -0.20 | 0.10 | -0.11 | 0.18 |
| 1 | 514 543 | -0.49 | 0.46 | -0.12 | -0.11 | -0.10 | -0.05 | 0.37 | 0.20 | -0.04 | 0.35 | 0.03 | -0.17 | 0.15 | -0.08 | -0.09 | -0.16 | -0.06 | 0.22 | -0.23 | -0.08 |
| 1 | 518 519 | 0.02 | 0.42 | -0.02 | -0.20 | -0.57 | -0.09 | -0.29 | 0.08 | 0.06 | 0.04 | 0.20 | -0.13 | 0.00 | 0.02 | 0.40 | 0.02 | 0.22 | -0.08 | -0.24 | 0.14 |
| 1 | 582 585 | 0.23 | -0.51 | 0.03 | 0.24 | 0.41 | 0.01 | 0.17 | 0.34 | -0.11 | -0.24 | -0.17 | -0.07 | -0.02 | 0.02 | 0.10 | 0.13 | 0.09 | -0.30 | -0.28 | -0.07 |
| 0.9989 | 169 474 548 | 0.15 | 0.40 | 0.11 | 0.03 | 0.02 | -0.03 | -0.36 | -0.14 | 0.17 | 0.13 | -0.21 | -0.33 | -0.30 | 0.11 | -0.30 | 0.20 | 0.23 | 0.19 | 0.22 | -0.30 |
| 0.9957 | 8 424 598 | 0.35 | -0.20 | 0.12 | -0.15 | -0.26 | 0.19 | 0.28 | -0.37 | 0.23 | -0.42 | 0.16 | 0.17 | 0.31 | -0.12 | 0.20 | -0.10 | -0.06 | -0.05 | -0.15 | -0.13 |
| 0.9953 | 191 193 195 | 0.10 | -0.03 | -0.47 | -0.09 | 0.06 | 0.06 | 0.09 | -0.09 | 0.01 | -0.29 | 0.23 | -0.14 | 0.39 | -0.08 | 0.18 | 0.03 | 0.32 | 0.28 | -0.42 | -0.15 |

# Appendix C

# Signal-processing based bioinformatics approach for Multiple Protein Sequence Alignment

## C. SIGNAL-PROCESSING BASED BIOINFORMATICS APPROACH FOR MULTIPLE PROTEIN SEQUENCE ALIGNMENT

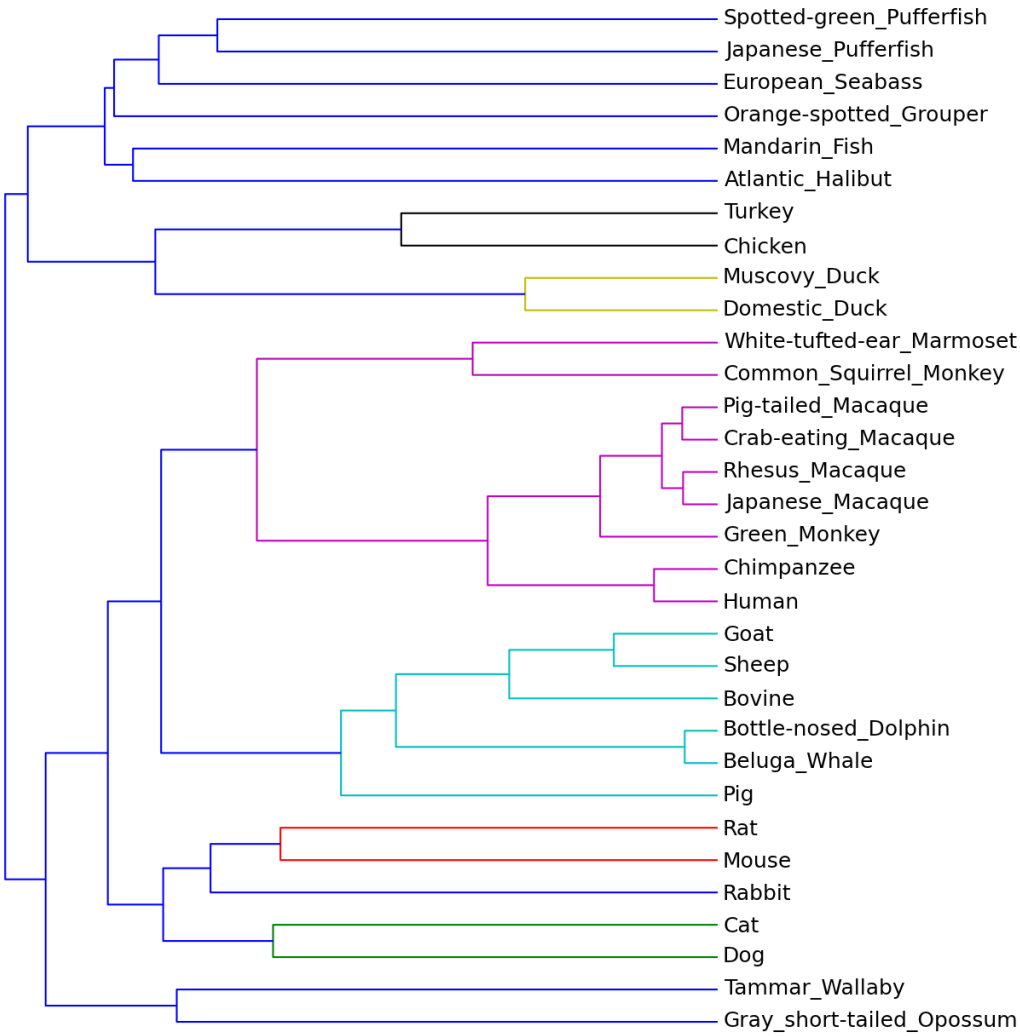



Figure C.1: Dendrogram using DFT and Bulkiness

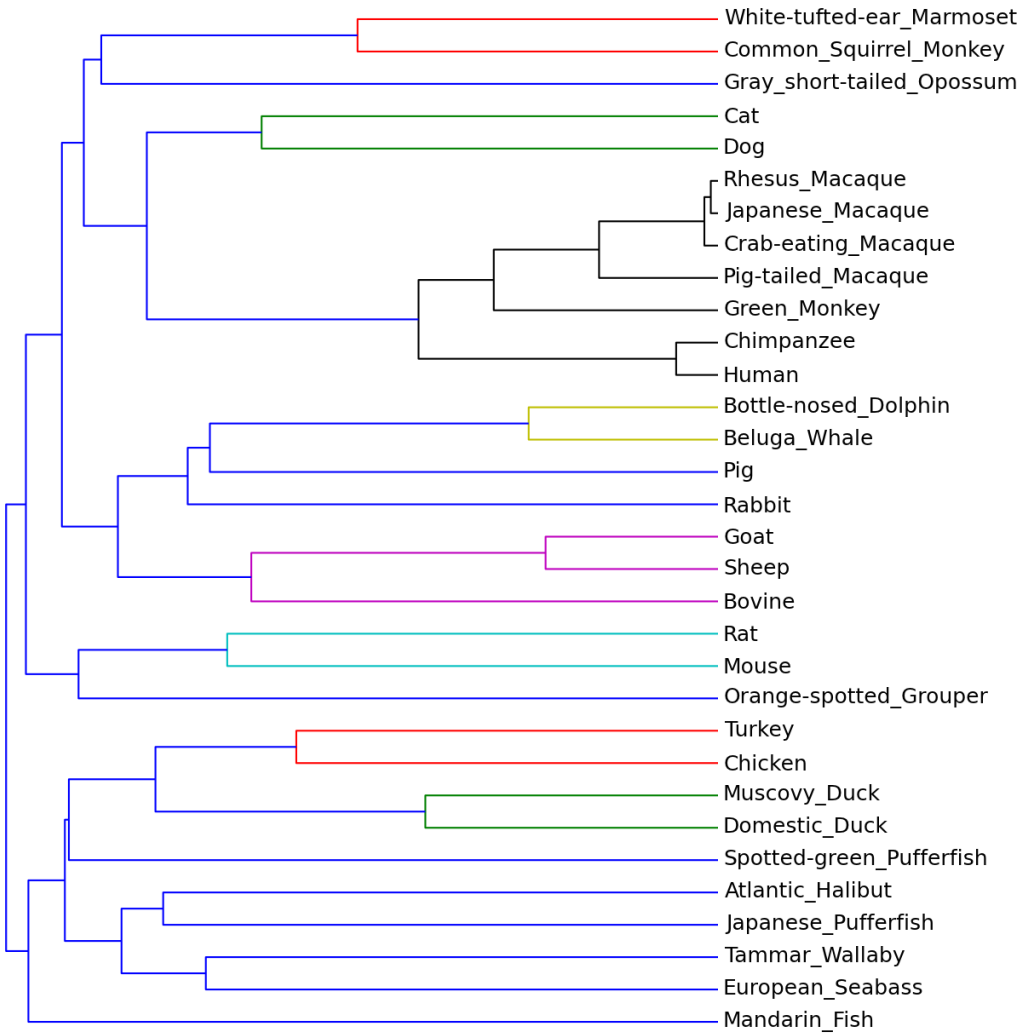



Figure C.2: Dendrogram using DFT and Isoelectric Point

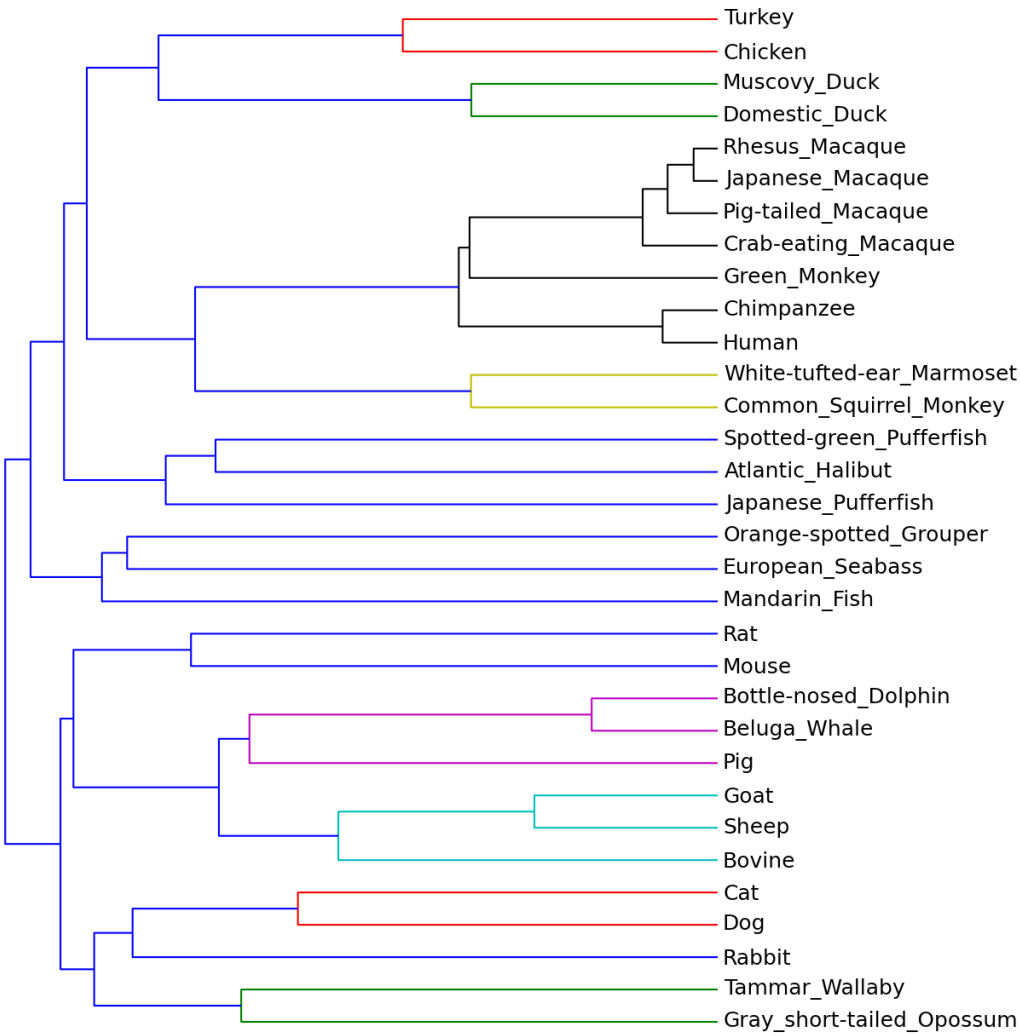



Figure C.3: Dendrogram using DFT and Absolute Entropy


Figure C.4: Dendrogram using DFT and Size

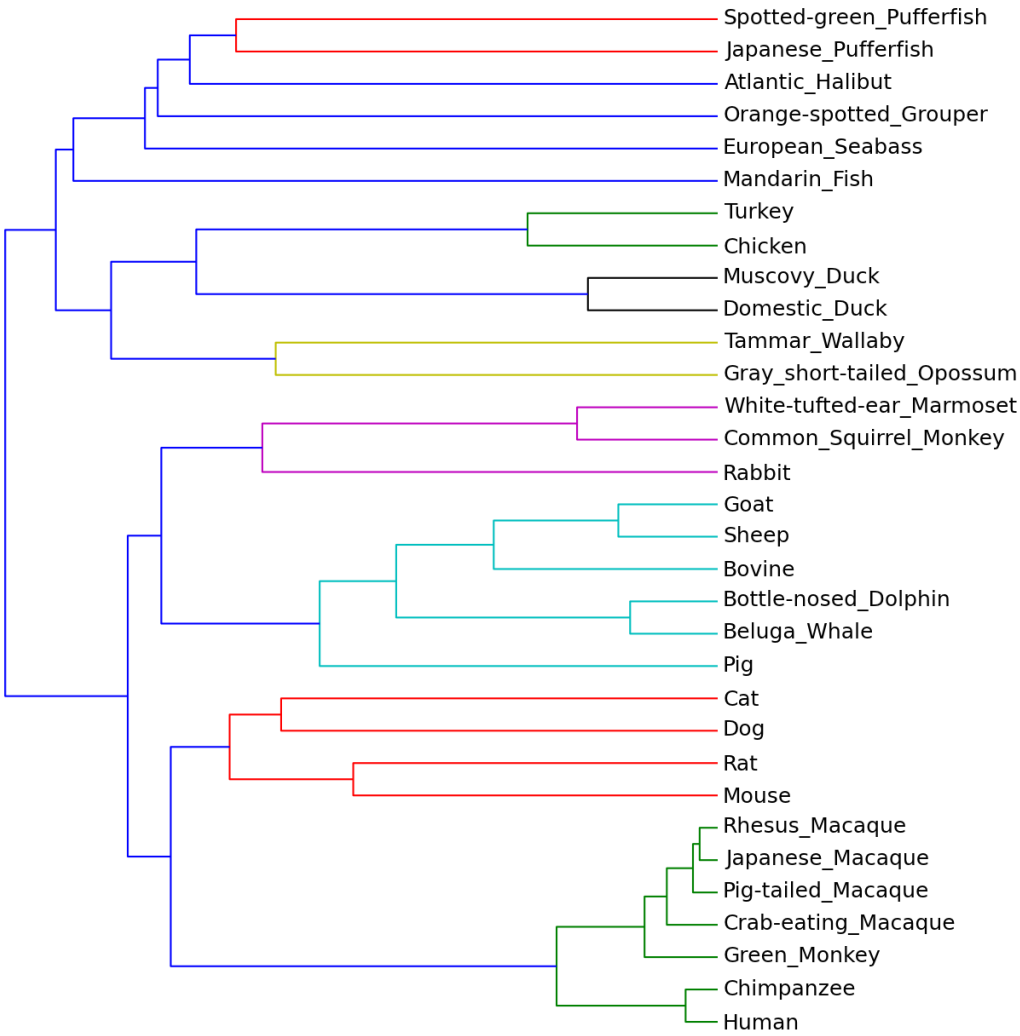



Figure C.5: Dendrogram using DFT and Polarity


Figure C.6: Dendrogram using DFT and Volume

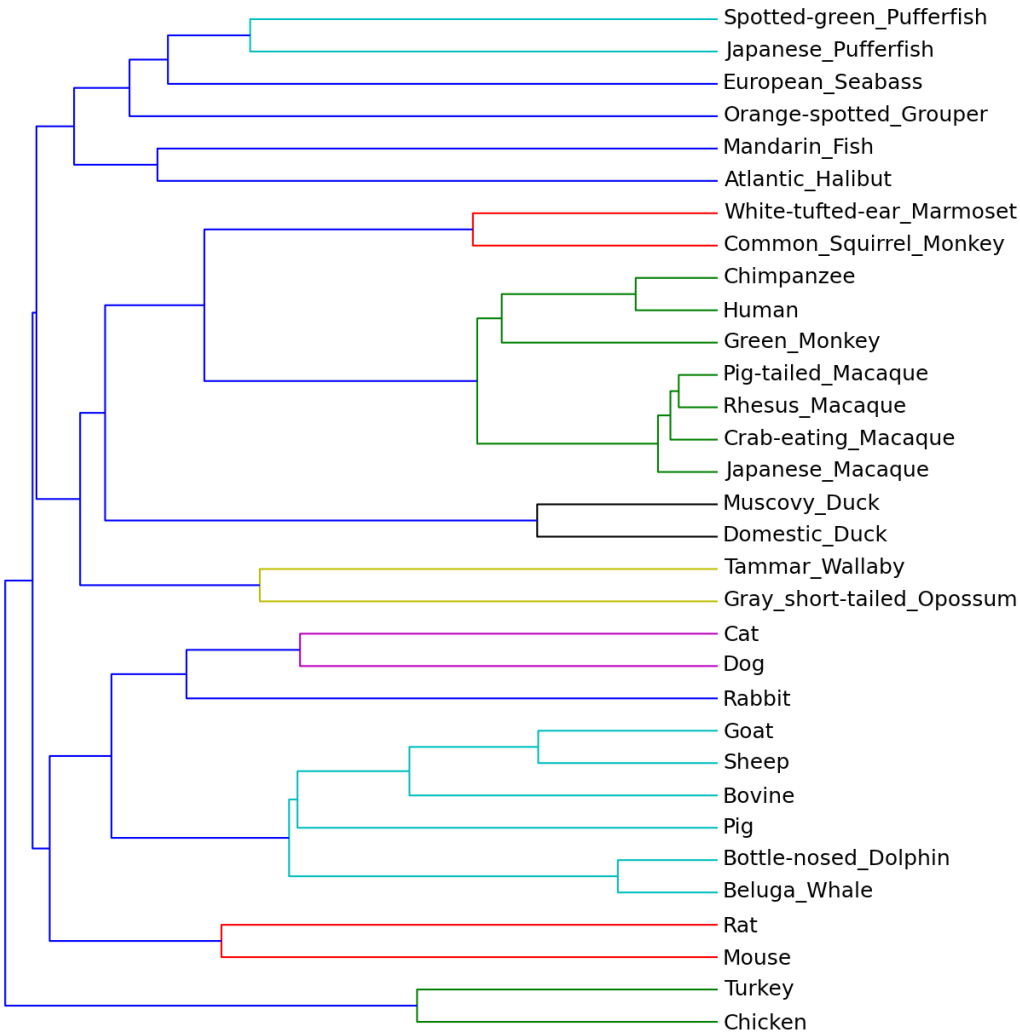



Figure C.7: Dendrogram using DFT and Molecular Weight


Figure C.8: Dendrogram using DFT and Melting Point

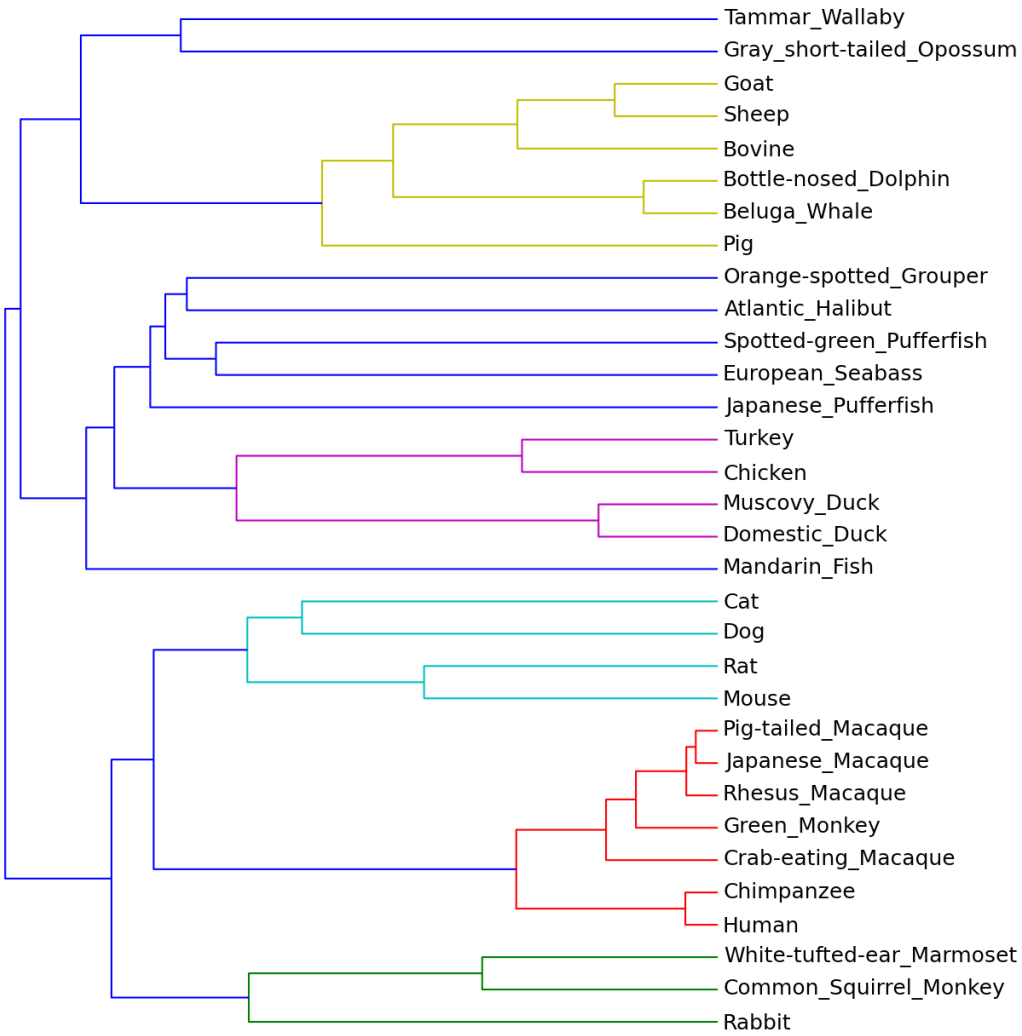



Figure C.9: Dendrogram using DFT and Hydrophobicity Index


Figure C.10: Dendrogram using DFT and the stability Scale from the Knowledge-Based atom-atom Potential

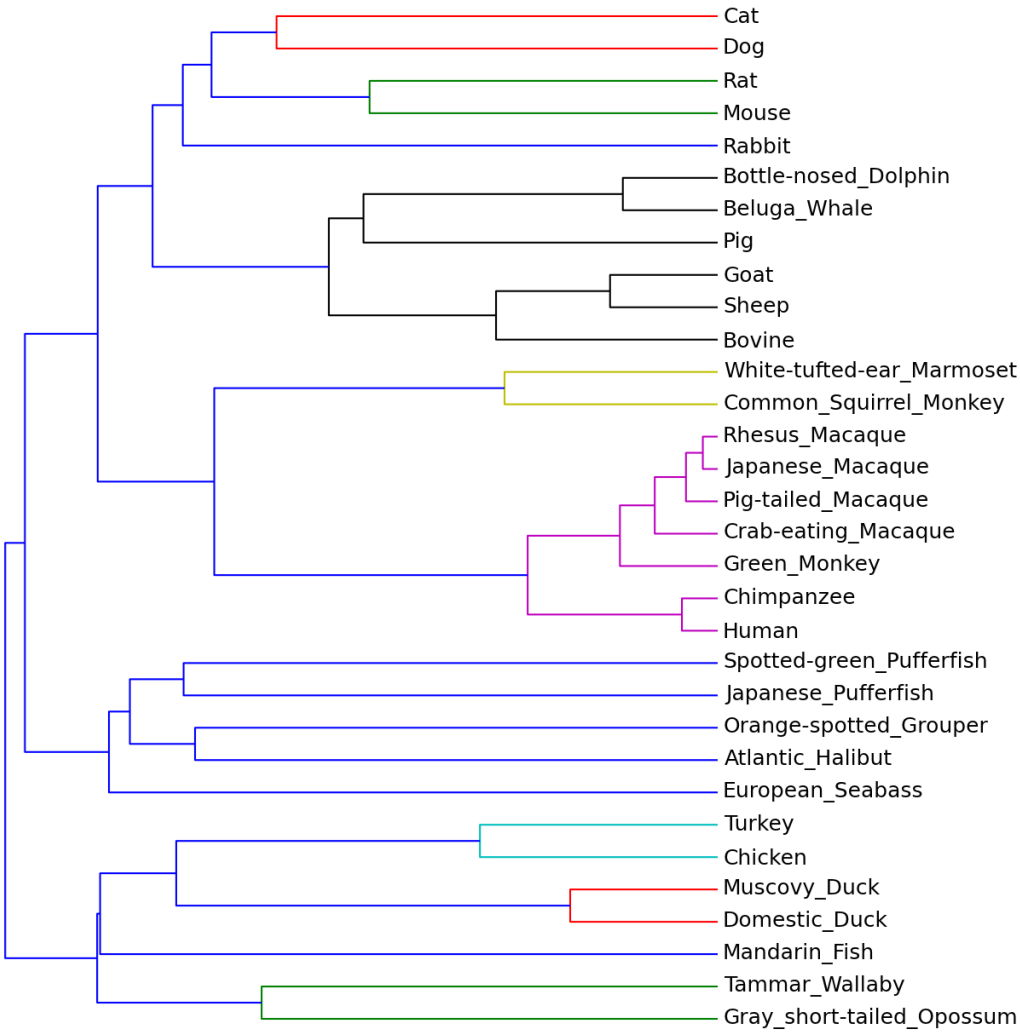



Figure C.11: Dendrogram using DFT and Long Range non-Bonded Energy per Atom

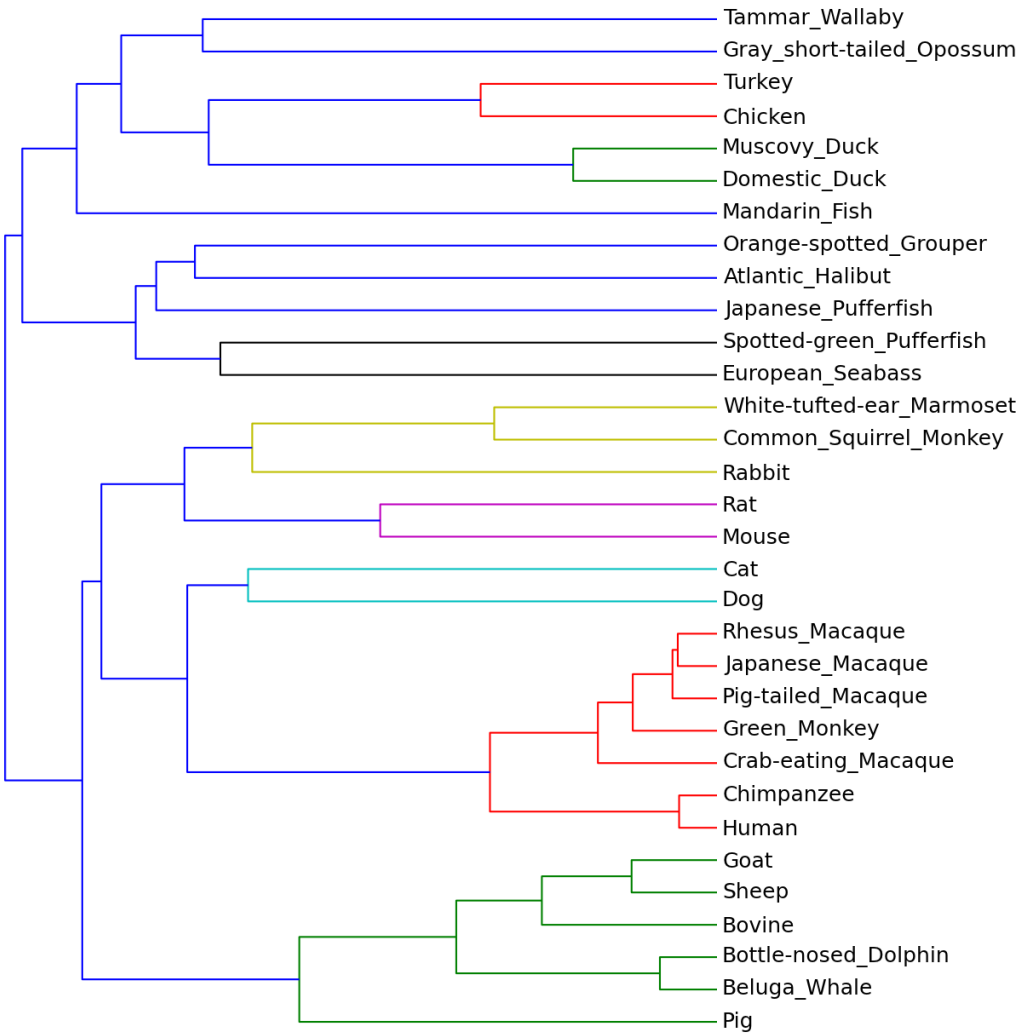



Figure C.12: Dendrogram using DFT and Average Surrounding Hydrophobicity

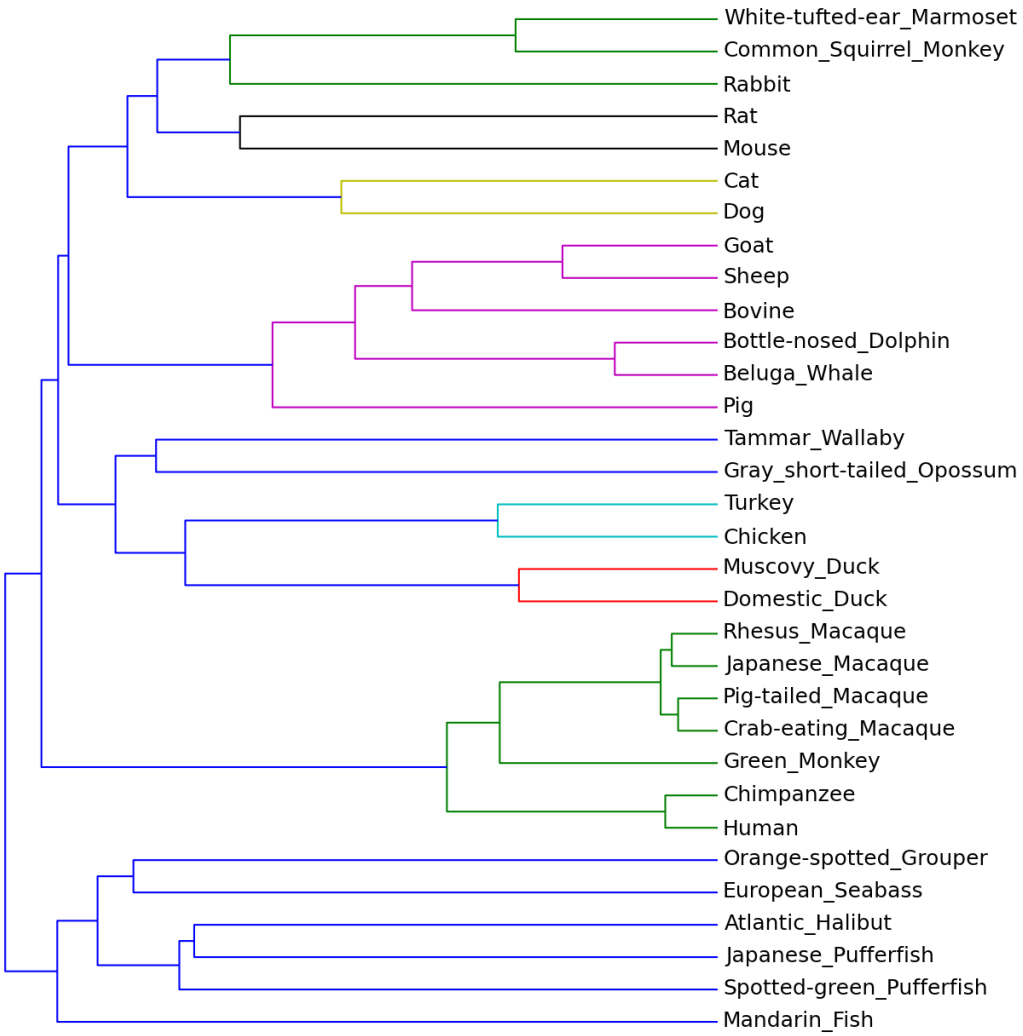



Figure C.13: Dendrogram using DFT and Hydrophobicity Index

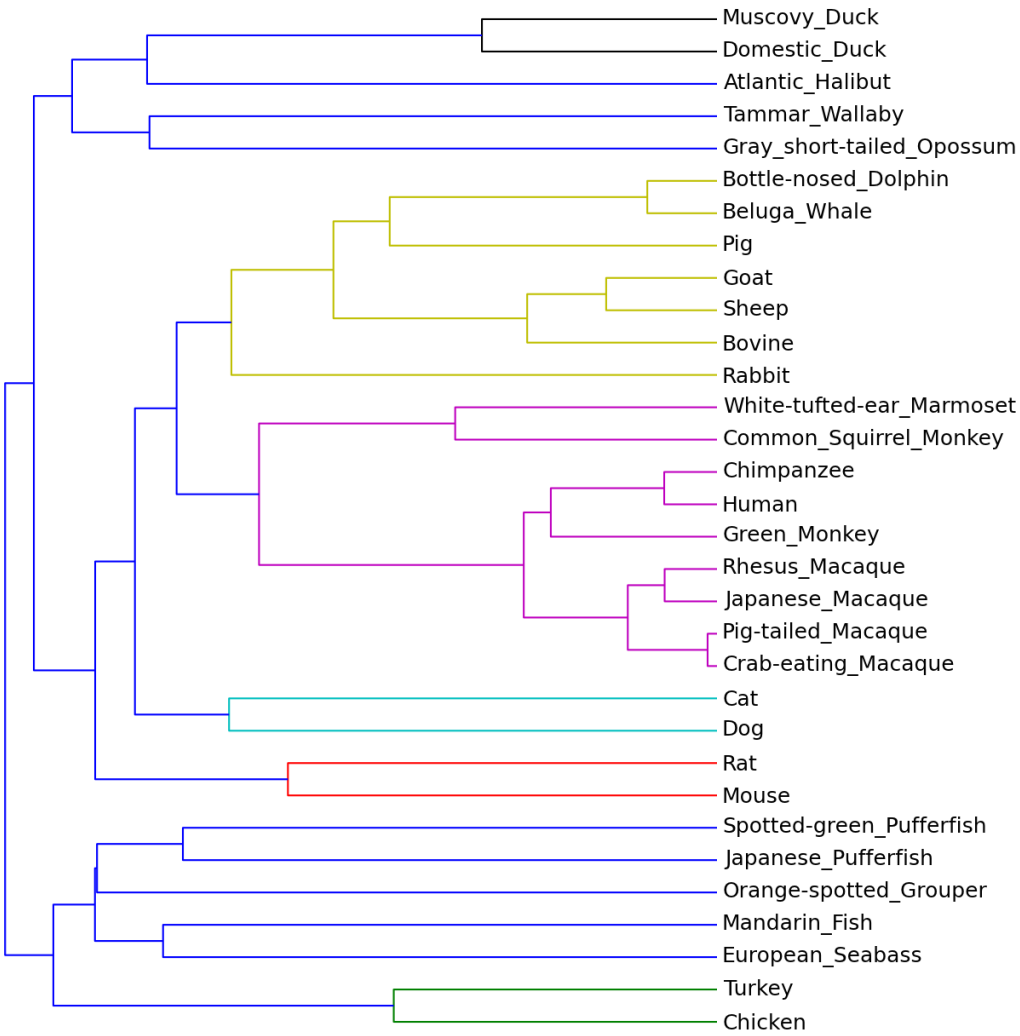



Figure C.14: Dendrogram using DFT and Hydration Potential

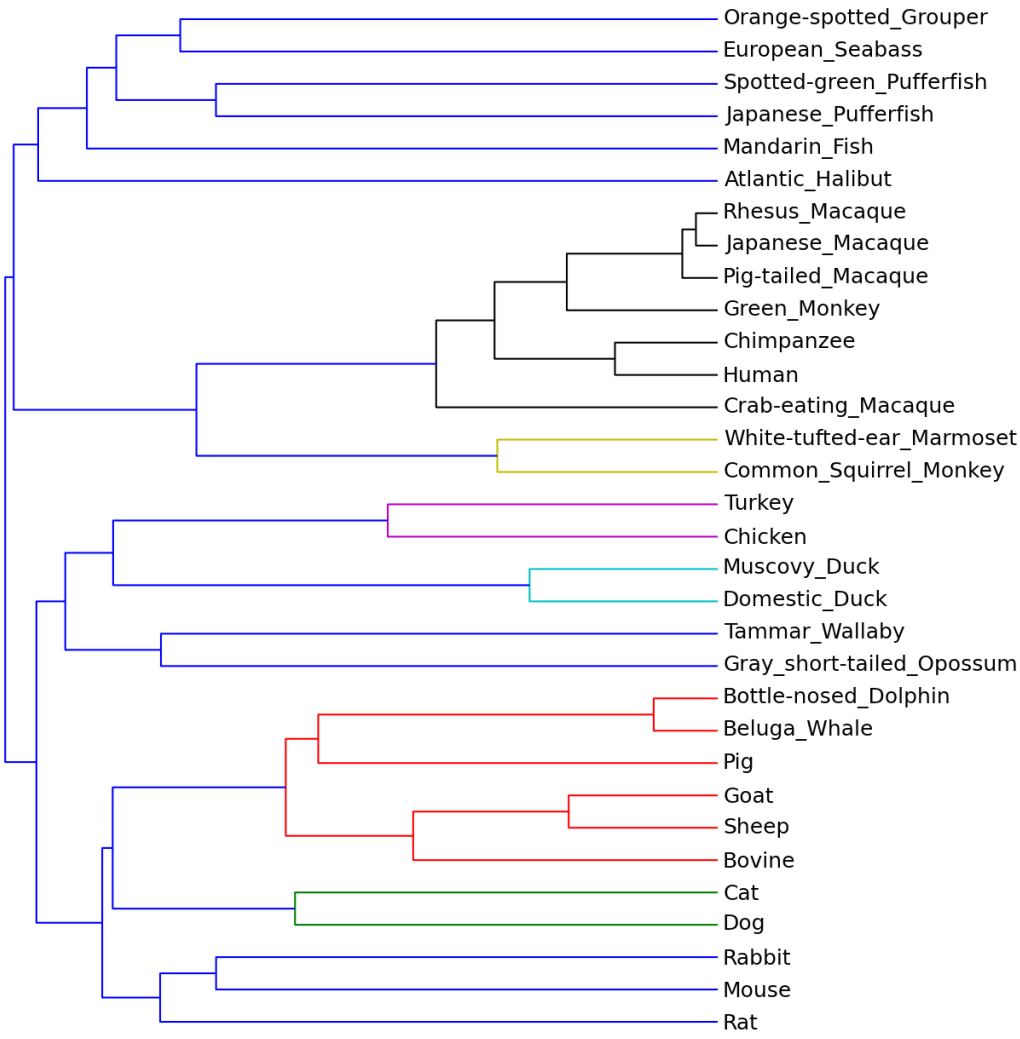



Figure C.15: Dendrogram using DFT and Smoothed Upsilon Steric Parameter

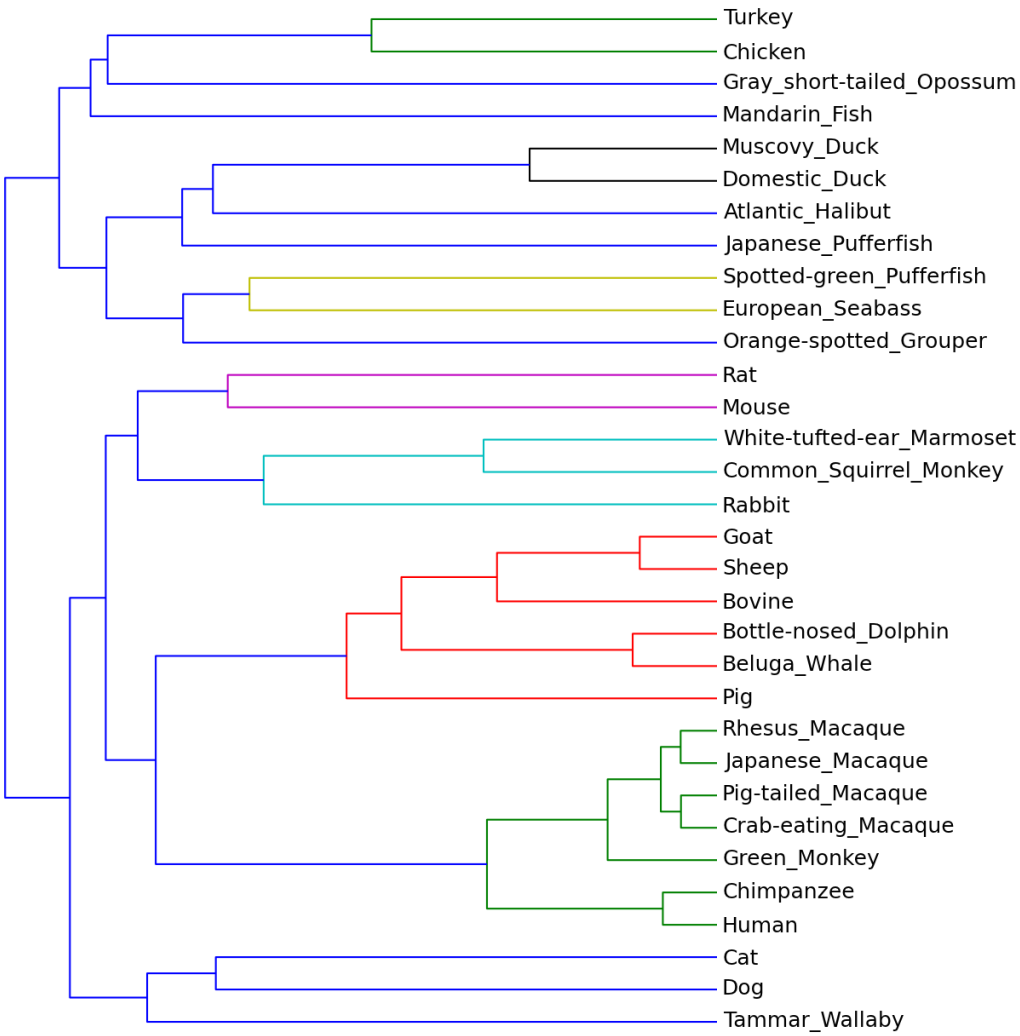



Figure C.16: Dendrogram using DFT and Hydrophobicity Index

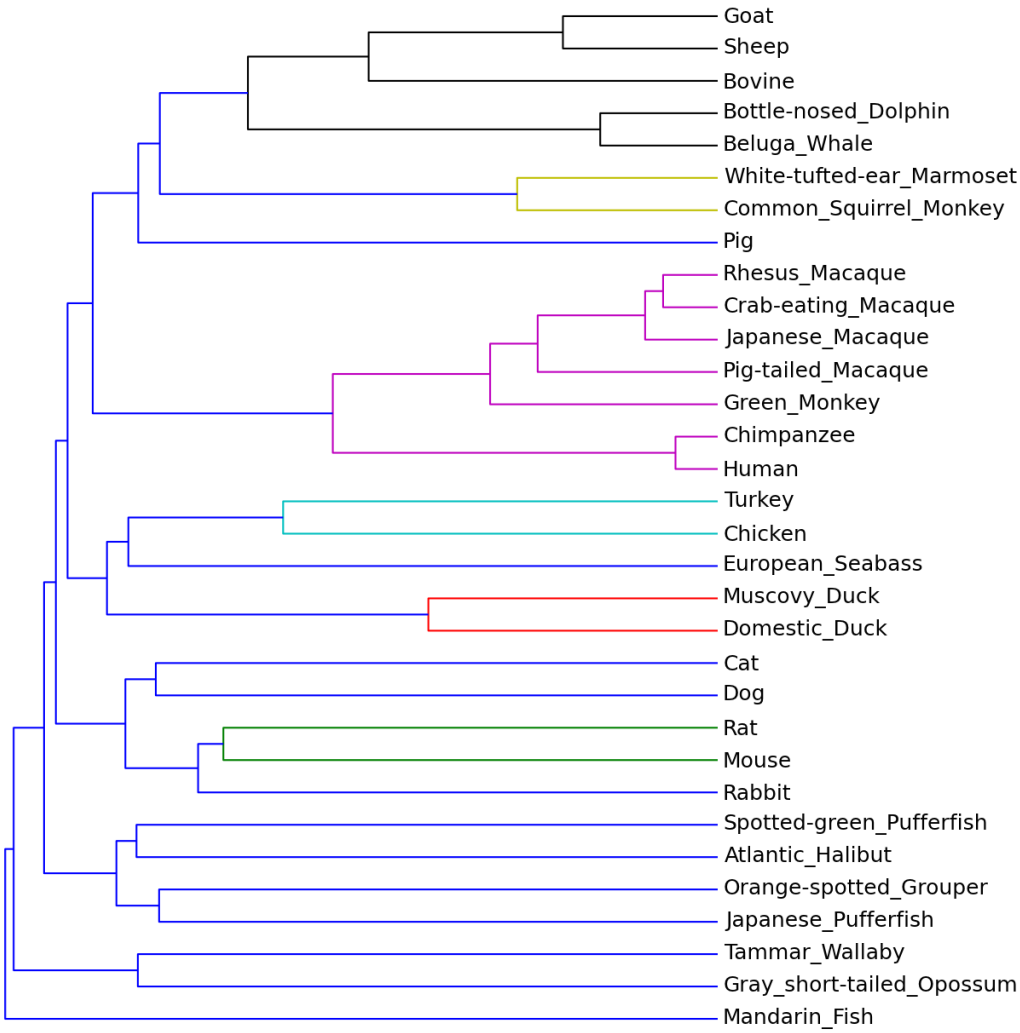



Figure C.17: Dendrogram using DFT and Electron-Ion Interaction Potential

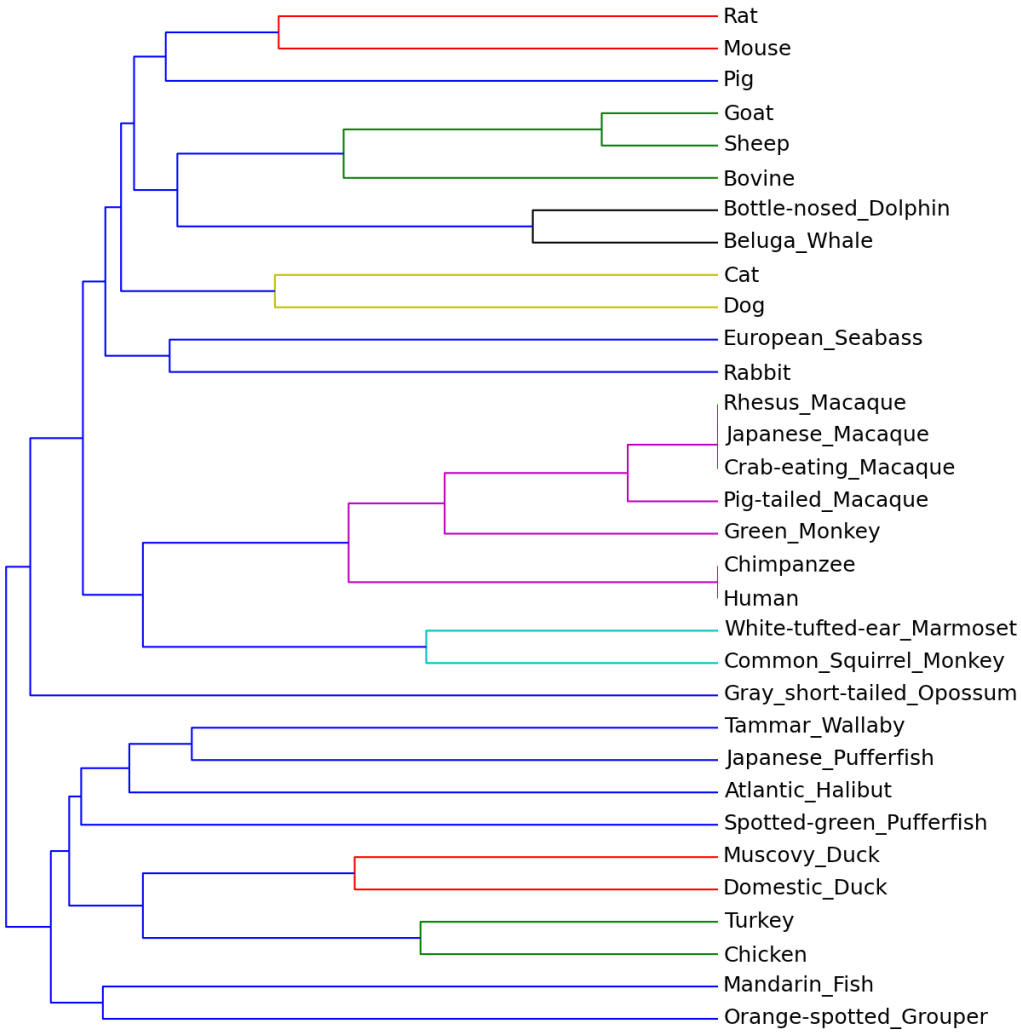



Figure C.18: Dendrogram using DFT and Positive Charge

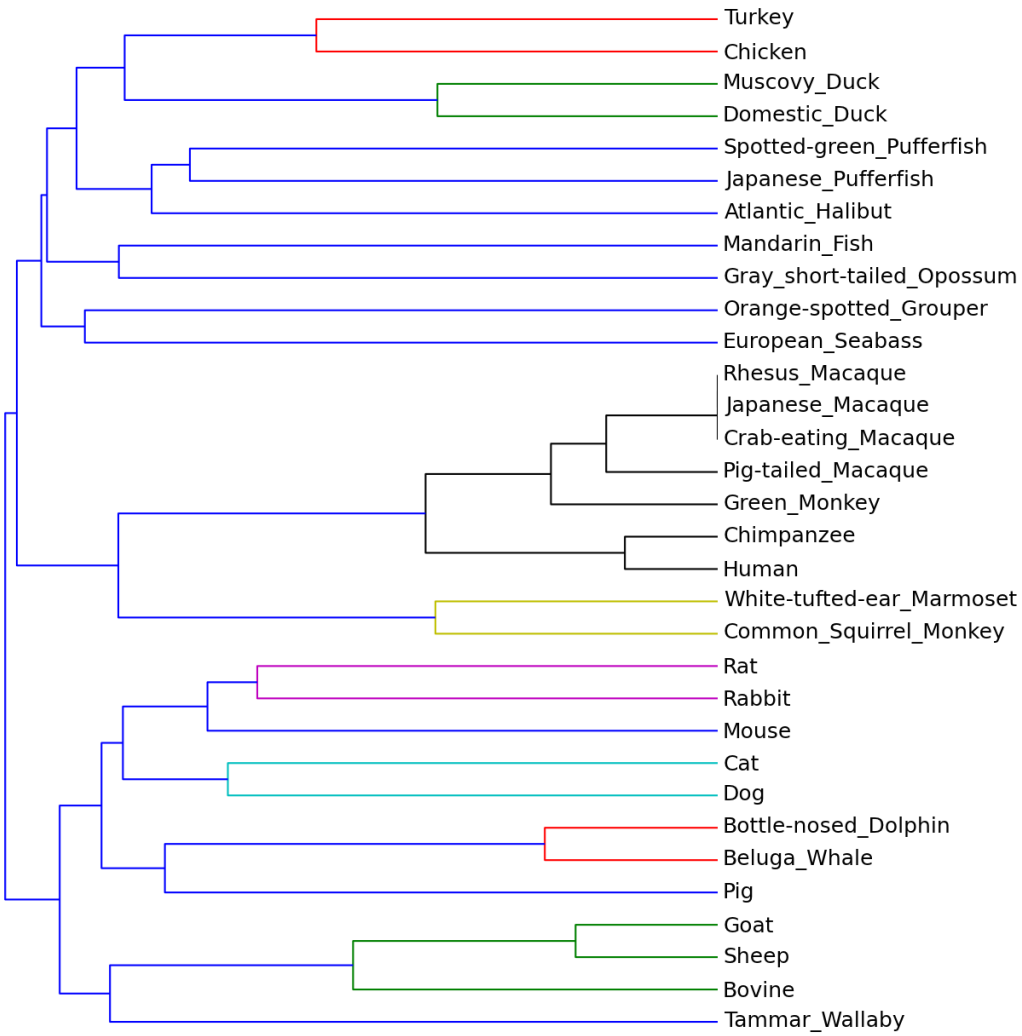



Figure C.19: Dendrogram using DFT and Negative Charge

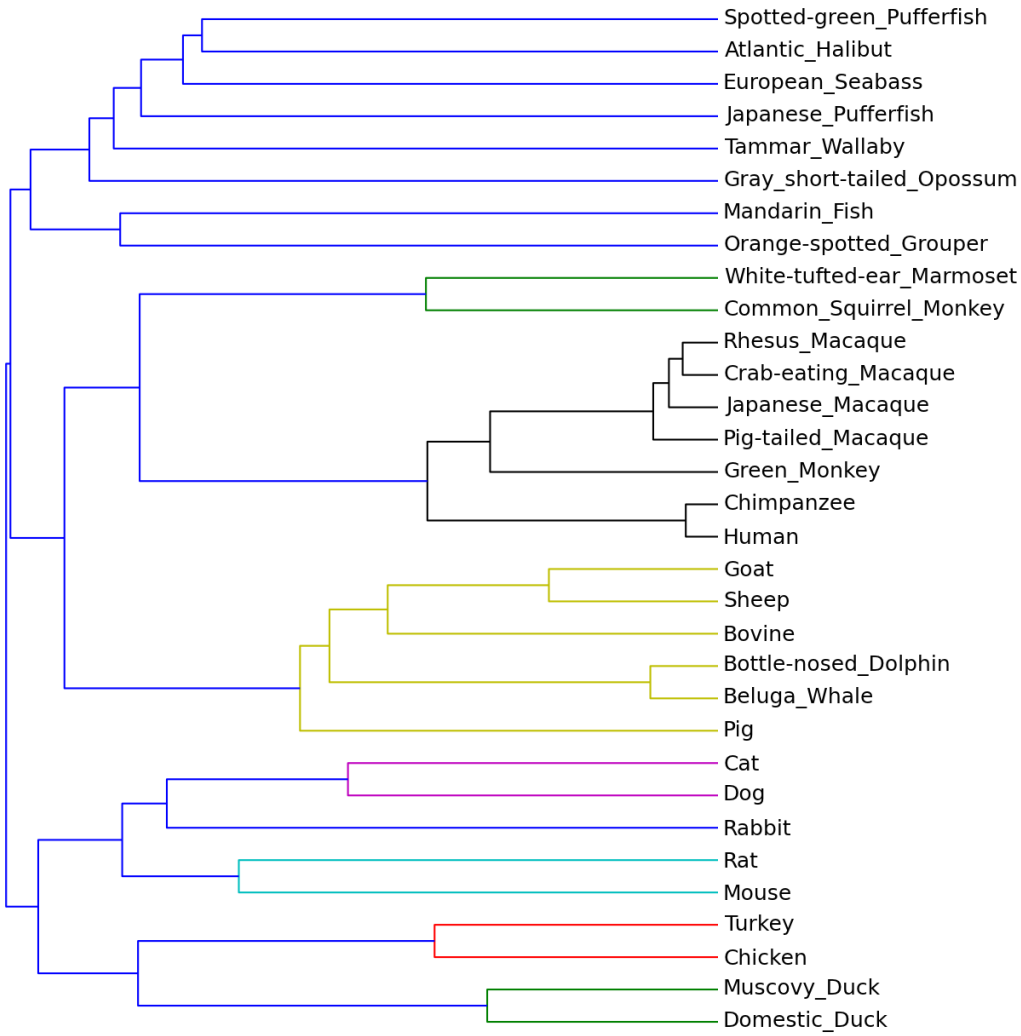



Figure C.20: Dendrogram using DFT and Number of Hydrogen Bond Donors


Figure C.21: Dendrogram using DFT and Hydropathy Index

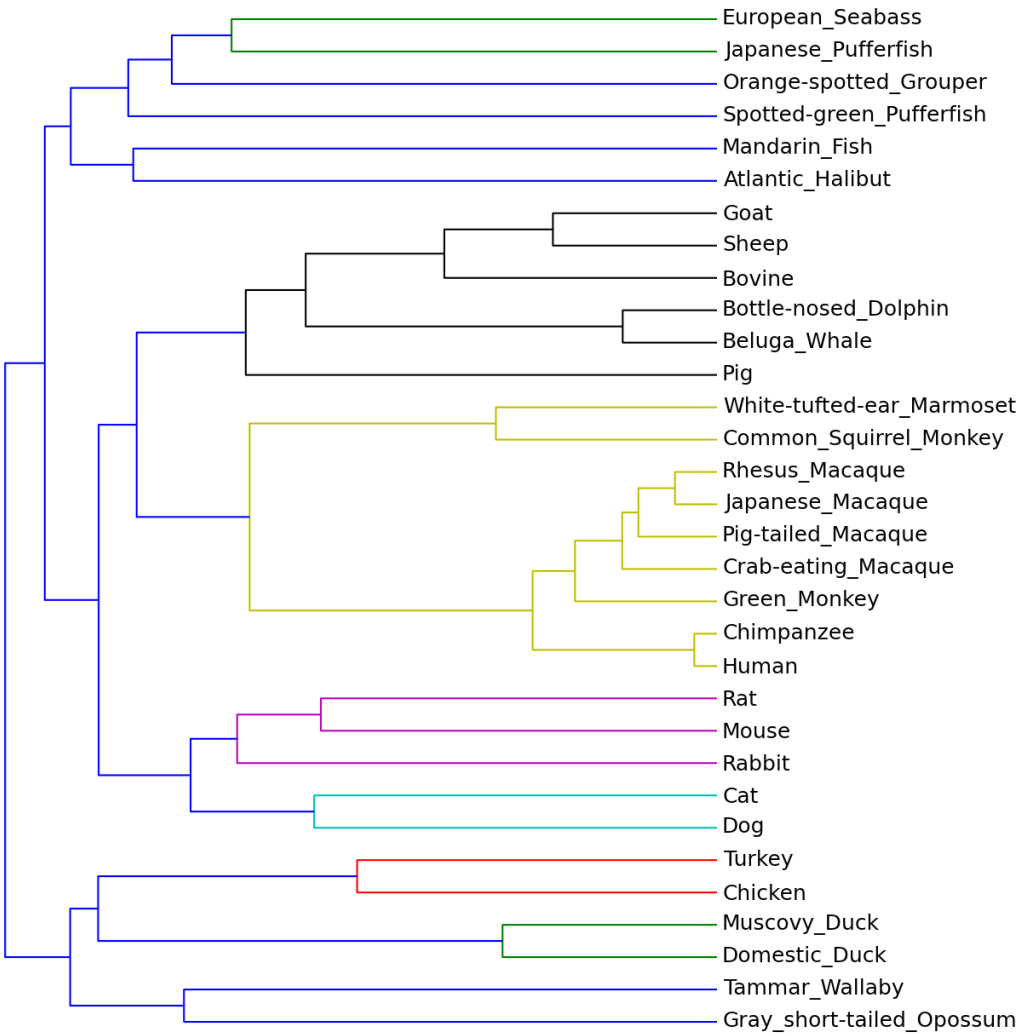



Figure C.22: Dendrogram using DFT and Average Flexibility Indices

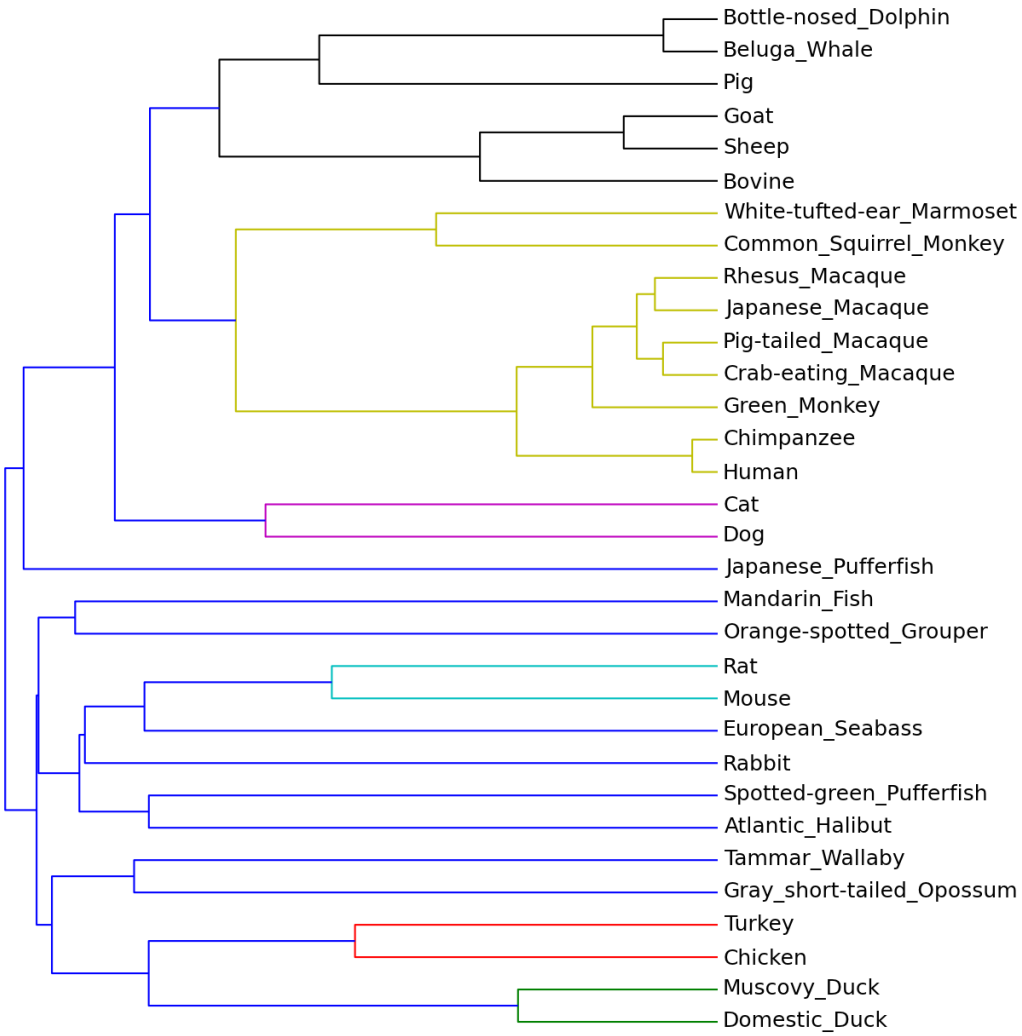



Figure C.23: Dendrogram using DFT and Recognition Factors

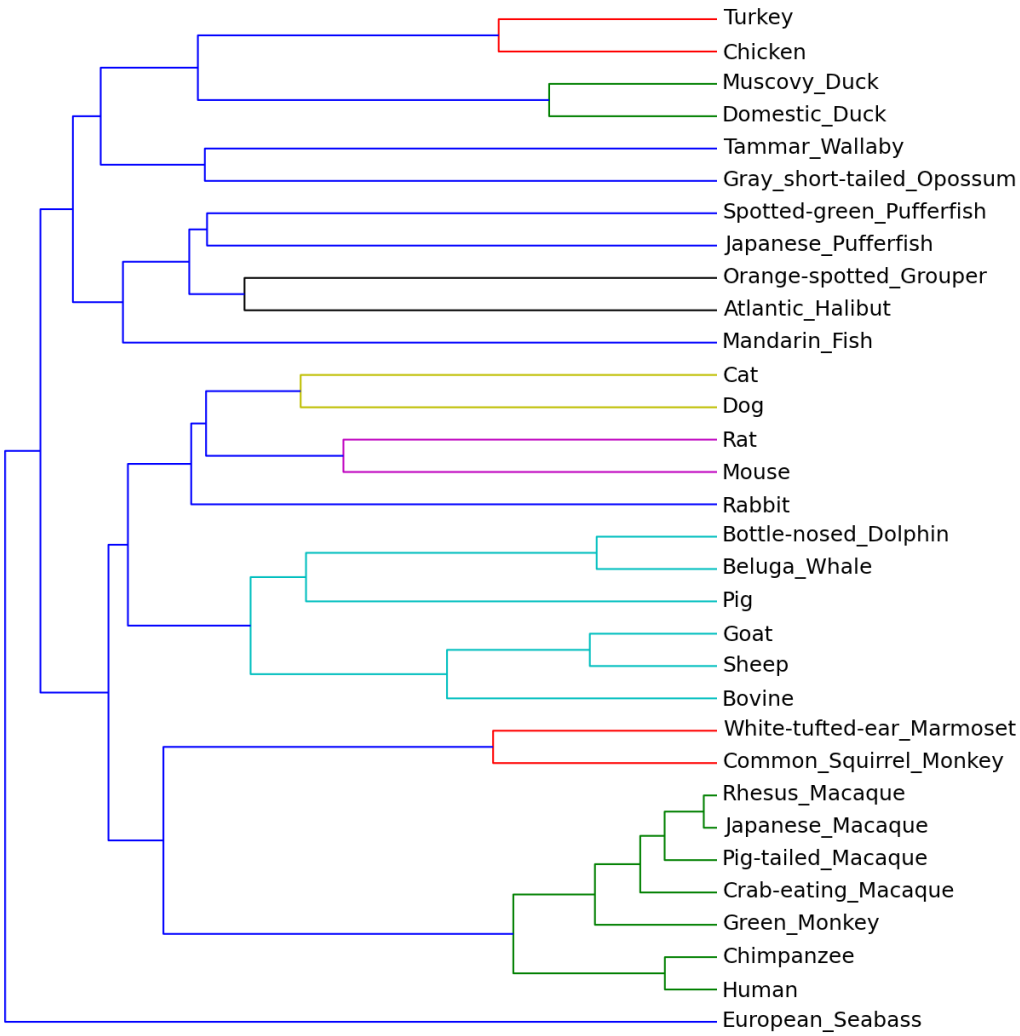



Figure C.24: Dendrogram using DFT and Long-Range Contacts

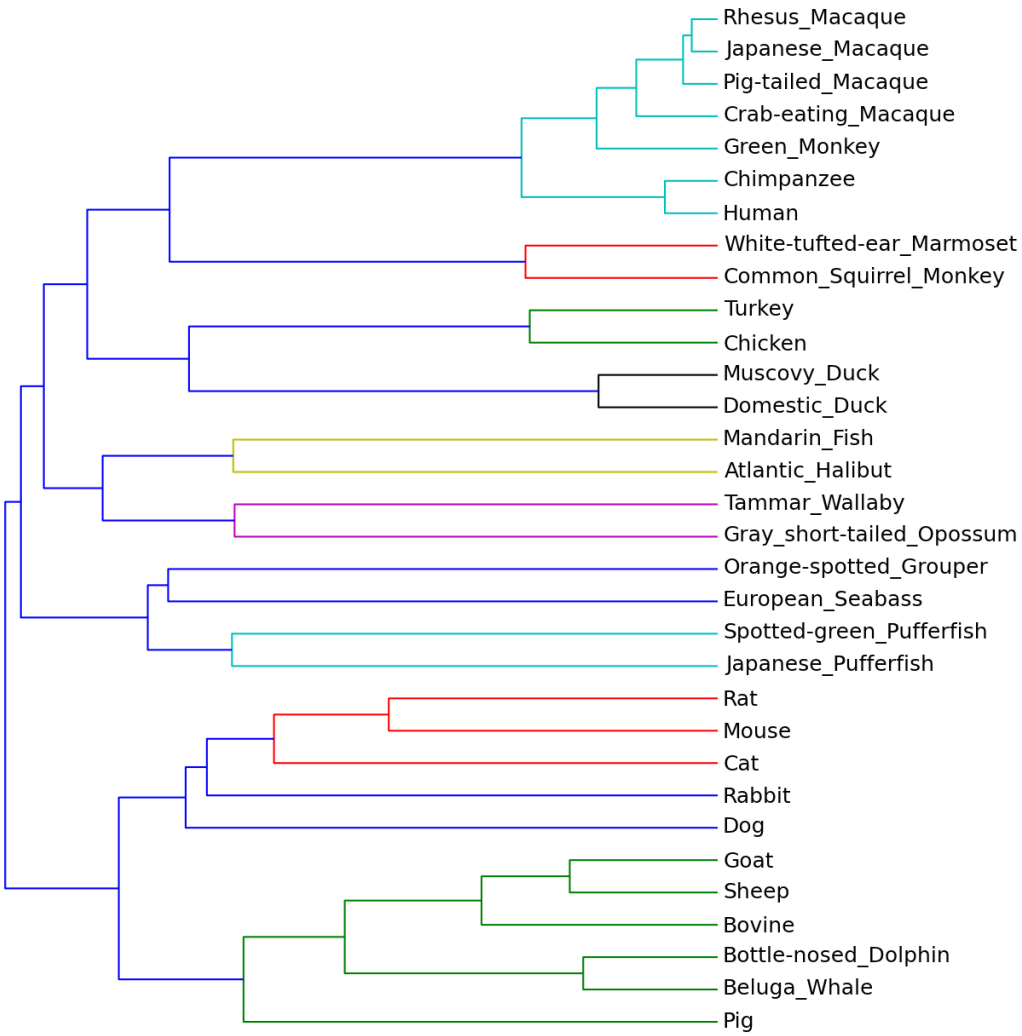



Figure C.25: Dendrogram using DFT and Relative Connectivity

# Appendix D

# List of Influenza Neuraminidase A Protein Sequences

## D. LIST OF INFLUENZA NEURAMINIDASE A PROTEIN SEQUENCES

Table D.1: H1N1 1933-1946 Protein Sequences

| Uniprot ID | Description | Uniprot ID | Description |
|---|---|---|---|
| ABP49330 | A/AA/Huston/1945 | ABD77678 | A/Puerto Rico/8/1934 |
| ABO38057 | A/AA/Marton/1943 | AAM75160 | A/Puerto Rico/8/34/Mount Sinai |
| ABD62845 | A/Bellamy/1942 | ACV49537 | A/United Kingdom/1/1933 |
| ABD79115 | A/Cameron/1946 | AAA91327 | A/WS/1933 |
| ABO38354 | A/Henry/1936 | ABF47958 | A/WSN/1933 |
| ABI20829 | A/Hickox/1940 | AAA43397 | A/WSN/1933 |
| AAM75501 | A/Marton/43 | AAA91328 | A/WSN/1933 |
| ACV49559 | A/Melbourne/1/1946 | ACF54601 | A/WSN/1933 |
| ABD62784 | A/Melbourne/1935 | AAF77045 | A/Weiss/1943 |
| AAA91326 | A/NWS/1933 | AAM76692 | A/Weiss/1943 |
| ABO38387 | A/Phila/1935 | ABD79104 | A/Weiss/1943 |
| ACV49548 | A/Puerto Rico/8-1/1934 | ABD77799 | A/Wilson-Smith/1933 |
| ACR15351 | A/Puerto Rico/8-SV14/1934 | ABF21334 | A/Wilson-Smith/1933 |
| AAA43412 | A/Puerto Rico/8/1934(Cambridge) | | |

Table D.2: H1N1 1947-1957 Protein Sequences

| Uniprot ID | Description | Uniprot ID | Description |
|---|---|---|---|
| ABP49484 | A/Albany/12/1951 | AAF77037 | A/Fort Monmouth/1/1947 |
| ABQ44474 | A/Albany/14/1951 | ABD61738 | A/Fort Worth/1950 |
| ABN59404 | A/Albany/4835/1948 | AAA43797 | A/Leningrad/1954 |
| ABD15262 | A/Denver/1957 | ABD60969 | A/Malaysia/1954 |
| AAM76693 | A/Fort Monmouth/1/1947 | AAM75502 | A/Rhodes/1947 |
| ABD77810 | A/Fort Monmouth/1/1947 | ABN59437 | A/Roma/1949 |

Table D.3: H1N1 1979-1989 Protein Sequences

| Uniprot ID | Description | Uniprot ID | Description |
|---|---|---|---|
| ABW36314 | A/Albany/8/1979 | ABF47828 | A/Memphis/16/1983 |
| ABD95342 | A/Baylor/11515/1982 | ABG88347 | A/Memphis/2/1983 |
| ABD77821 | A/Baylor/11735/1982 | ABK40549 | A/Memphis/24/1983 |
| ABO52261 | A/Baylor/4052/1981 | ABK40593 | A/Memphis/30/1983 |
| ABP49341 | A/California/10/1978 | ABM66889 | A/Memphis/31/1983 |
| ABY81352 | A/California/45/1978 | ABN50903 | A/Memphis/49/1983 |
| CAA33354 | A/Chile/1/1983 | ABI92305 | A/Memphis/6/1983 |
| AAA96671 | A/Chile/1/1983 | ABQ44397 | A/New Jersey/1976 |
| ABO38343 | A/Chile/1/1983 | AAA43210 | A/New Jersey/8/1976 |
| ABO52800 | A/Christ's Hospital/157/1982 | ACQ99824 | A/New Jersey/8/1976 |
| BAA06714 | A/Hokkaido/11/1988 | ABU80403 | A/Ohio/3559/1988 |
| ABD60947 | A/Hong Kong/117/1977 | ACK99446 | A/Siena/10/1989 |

Table D.3: (continued)

| | | | |
|---|---|---|---|
| ABO38365 | A/India/6263/1980 | ACV49669 | A/Siena/4/1987 |
| AAM76694 | A/India/80 | ACL12264 | A/Siena/9/1989 |
| AAA43435 | A/Kiev/59/1979 | ABO38398 | A/Singapore/6/1986 |
| ABO33009 | A/Maryland/2/1980 | ABF21330 | A/Taiwan/01/1986 |
| ABN50759 | A/Memphis/1/1979 | ABO44126 | A/Texas/2922-3/1986 |
| ABG88336 | A/Memphis/1/1983 | ABO44137 | A/Tientsin/78/1977 |
| ABP49352 | A/Memphis/1/1984 | ABO38409 | A/Tonga/14/1984 |
| ABQ44419 | A/Memphis/1/1987 | AAA43449 | A/USSR/90/1977 |
| ABI30381 | A/Memphis/10/1983 | ABD60936 | A/USSR/92/1977 |
| ABF47707 | A/Memphis/11/1978 | ABV45841 | A/Wisconsin/301/1976 |
| ABM22249 | A/Memphis/12/1986 | BAA06718 | A/Yamagata/120/1986 |
| ABI20862 | A/Memphis/14/1983 | BAA06720 | A/Yamagata/32/1989 |

Table D.4: H1N1 2009 Protein Sequences

| Uniprot ID | Description | Uniprot ID | Description |
|---|---|---|---|
| ADK33724 | A/Aarhus/INS242/2009 | ADJ80697 | A/California/VRDL107/2009 |
| ADK33814 | A/Aarhus/INS253/2009 | ADJ80707 | A/California/VRDL108/2009 |
| ADK33824 | A/Aarhus/INS254/2009 | ADD21777 | A/California/VRDL11/2009 |
| ADA83034 | A/Abakan/02/2009 | ADJ80727 | A/California/VRDL110/2009 |
| ADM14968 | A/Addis Ababa/WR2848N/2009 | ADJ80747 | A/California/VRDL112/2009 |
| ACU44306 | A/Alaska/01/2009 | ADJ80861 | A/California/VRDL123/2009 |
| ADM14469 | A/Amman/WR0060N/2009 | ADJ80896 | A/California/VRDL126/2009 |
| ADM14702 | A/Amman/WR1335T/2009 | ADJ80913 | A/California/VRDL127/2009 |
| ADC32390 | A/Ancona/15/2009 | ADJ80940 | A/California/VRDL129/2009 |
| ADC32407 | A/Ancona/254/2009 | ADM07463 | A/California/VRDL149/2009 |
| ADC32403 | A/Ancona/310/2009 | ADN06281 | A/California/VRDL152/2009 |
| ADC32405 | A/Ancona/508/2009 | ADN06314 | A/California/VRDL155/2009 |
| ADM14711 | A/Ankara/WR1429T/2009 | ADN06336 | A/California/VRDL167/2009 |
| ADA83582 | A/Argentina/HNRG13/2009 | ADN06347 | A/California/VRDL175/2009 |
| ADA83588 | A/Argentina/HNRG14/2009 | ADD21807 | A/California/VRDL18/2009 |
| ADA83592 | A/Argentina/HNRG15/2009 | ADN89053 | A/California/VRDL180/2009 |
| ADA83596 | A/Argentina/HNRG16/2009 | ADN89064 | A/California/VRDL186/2009 |
| ADA83600 | A/Argentina/HNRG17/2009 | ADN89108 | A/California/VRDL194/2009 |
| ADA83604 | A/Argentina/HNRG18/2009 | ADJ81006 | A/California/VRDL2/2010 |
| ADA83608 | A/Argentina/HNRG20/2009 | ADN89119 | A/California/VRDL200/2009 |
| ADA83612 | A/Argentina/HNRG21/2009 | ADN89130 | A/California/VRDL202/2009 |
| ADA83616 | A/Argentina/HNRG23/2009 | ADN89218 | A/California/VRDL220/2009 |
| ADA83634 | A/Argentina/HNRG31/2009 | ADN06358 | A/California/VRDL229/2009 |
| ADA83642 | A/Argentina/HNRG33/2009 | ADB89354 | A/California/VRDL23/2009 |
| ADB24796 | A/Argentina/HNRG36/2009 | ADN89317 | A/California/VRDL232/2009 |
| ADA83649 | A/Argentina/HNRG37/2009 | ADN89328 | A/California/VRDL235/2009 |
| ADA83657 | A/Argentina/HNRG40/2009 | ADN89361 | A/California/VRDL238/2009 |
| ADA83665 | A/Argentina/HNRG42/2009 | ADN89405 | A/California/VRDL249/2009 |
| ADA83681 | A/Argentina/HNRG5/2009 | ADN89427 | A/California/VRDL252/2009 |

Table D.4: (continued)

| | | | |
|---|---|---|---|
| ACU44276 | A/Arkansas/01/2009 | ADN89449 | A/California/VRDL256/2009 |
| ACU44287 | A/Arkansas/02/2009 | ADN89471 | A/California/VRDL264/2009 |
| ADL39469 | A/Arkansas/WRAIR1249P/2009 | ADN89493 | A/California/VRDL270/2009 |
| ADG59596 | A/Astrakhan/CRIE-CHRM/2009 | ADN89515 | A/California/VRDL275/2009 |
| ADG42176 | A/Athens/INS123/2009 | ADN89559 | A/California/VRDL283/2009 |
| ADH02001 | A/Athens/INS153/2009 | ADN89581 | A/California/VRDL289/2009 |
| ADG42336 | A/Athens/INS155/2009 | ADB89404 | A/California/VRDL29/2009 |
| ADG42366 | A/Athens/INS159/2009 | ADB89474 | A/California/VRDL36/2009 |
| ADG42386 | A/Athens/INS161/2009 | ADN89801 | A/California/VRDL371/2009 |
| ADG42406 | A/Athens/INS163/2009 | ADB89254 | A/California/VRDL4/2009 |
| ADK33844 | A/Athens/INS259/2009 | ADB89534 | A/California/VRDL45/2009 |
| ADK33874 | A/Athens/INS262/2009 | ADB89544 | A/California/VRDL48/2009 |
| ADK21826 | A/Athens/INS274/2009 | ADB89554 | A/California/VRDL50/2009 |
| ADM31671 | A/Athens/INS329/2009 | ADB89584 | A/California/VRDL53/2009 |
| ADM32884 | A/Athens/INS330/2009 | ADB89274 | A/California/VRDL6/2009 |
| ADM31691 | A/Athens/INS332/2009 | ADB89614 | A/California/VRDL61/2009 |
| ADM31731 | A/Athens/INS336/2009 | ADB89654 | A/California/VRDL68/2009 |
| ADM31761 | A/Athens/INS339/2009 | ADJ81046 | A/California/VRDL7/2010 |
| ADM31771 | A/Athens/INS340/2009 | ADJ81056 | A/California/VRDL8/2010 |
| ADM31781 | A/Athens/INS341/2009 | ADF87354 | A/California/VRDL83/2009 |
| ADM31791 | A/Athens/INS342/2009 | ADF87384 | A/California/VRDL86/2009 |
| ADM31801 | A/Athens/INS344/2009 | ADG42636 | A/California/VRDL88/2009 |
| ADM31811 | A/Athens/INS345/2009 | ADG42646 | A/California/VRDL89/2009 |
| ADM13074 | A/Athens/INS390/2010 | ADG42676 | A/California/VRDL92/2009 |
| ADO00743 | A/Athens/INS393/2010 | ADM14601 | A/California/WR1316P/2009 |
| ADM13094 | A/Athens/INS396/2010 | ADM14611 | A/California/WR1317P/2009 |
| ADM13224 | A/Athens/INS412/2010 | ADM14639 | A/California/WR1320P/2009 |
| ACQ42240 | A/Auckland/1/2009 | ADM14648 | A/California/WR1321P/2009 |
| ACR08499 | A/Auckland/1/2009 | ADL39447 | A/California/WRAIR1243P/2009 |
| ACR40631 | A/Auckland/1/2009 | ACT68161 | A/Canada-MB/RV2023/2009 |
| ACR01020 | A/Auckland/4/2009 | ACT68169 | A/Canada-QC/RV1759/2009 |
| ADD22607 | A/Australia/27/2009 | ACT68170 | A/Canada-SK/RV1767/2009 |
| ADB89694 | A/Australia/3/2009 | ACU31177 | A/Canada-SK/RV2486/2009 |
| ADD22678 | A/Australia/39/2009 | ADH29478 | A/Che/NIV658/2009 |
| ADD22688 | A/Australia/40/2009 | ADO12236 | A/Chile/158/2009 |
| ADD22755 | A/Australia/47/2009 | ADO12276 | A/Chile/1599/2009 |
| ADD22765 | A/Australia/48/2009 | ADG59615 | A/Chita/CRIE-8/2009 |
| ADD22964 | A/Australia/69/2009 | ACQ83302 | A/Christchurch/2/2009 |
| ADD23085 | A/Australia/73/2009 | ACR01016 | A/Christchurch/2/2009 |
| ADD21887 | A/Australia/9/2009 | ACU44294 | A/Colorado/07/2009 |
| ADM13294 | A/Bangkok/INS424/2010 | ADD75091 | A/Copenhagen/INS96/2009 |
| ADM13334 | A/Bangkok/INS428/2010 | ACU44310 | A/Delaware/01/2009 |
| ADI49395 | A/Barcelona/INS190/2009 | ACV67183 | A/District of Columbia/03/2009 |
| ADM86443 | A/Barcelona/INS378/2009 | ADD23253 | A/District of Columbia/INS17/2009 |
| ACZ98548 | A/Beijing/718/2009 | ADD23263 | A/District of Columbia/INS18/2009 |
| ACZ98554 | A/Beijing/719/2009 | ADD23273 | A/District of Columbia/INS19/2009 |
| ACZ98562 | A/Beijing/720/2009 | ADK32680 | A/District of Columbia/INS226/2009 |

Table D.4: (continued)

| | | | |
|---|---|---|---|
| ADC39000 | A/Bishkek/03/2009 | ADD23323 | A/District of Columbia/INS27/2009 |
| ACZ97474 | A/Blore/NIV236/2009 | ADD74961 | A/District of Columbia/INS44/2009 |
| ADG42596 | A/Bochum/INS187/2009 | ACU56927 | A/Ekaterinburg/01/2009 |
| ADK33774 | A/Bochum/INS249/2009 | ACU44283 | A/Florida/03/2009 |
| ACY77870 | A/Bogota/WR0090N/2009 | ACR08565 | A/Florida/04/2009 |
| ADL39150 | A/Bogota/WRAIR0088N/2009 | ACU44302 | A/Florida/07/2009 |
| ADK21846 | A/Bonn/INS277/2009 | ACS72707 | A/Florida/09/2009 |
| ADI48645 | A/Boston/106/2009 | ACS72708 | A/Florida/10/2009 |
| ADI48675 | A/Boston/115/2009 | ACV67175 | A/Florida/16/2009 |
| ADI48685 | A/Boston/116/2009 | ADL32457 | A/Frankfurt/INS301/2009 |
| ADI48755 | A/Boston/124/2009 | ADM31481 | A/Frankfurt/INS302/2009 |
| ADI49253 | A/Boston/132/2009 | ADM31501 | A/Frankfurt/INS305/2009 |
| ADI49345 | A/Boston/141/2009 | ADM13134 | A/Frankfurt/INS401/2009 |
| ADO25095 | A/Boston/584/2009 | ADM13144 | A/Frankfurt/INS402/2010 |
| ADO25105 | A/Boston/591/2009 | ADM14720 | A/Ft Carson/WR1446P/2009 |
| ACT33116 | A/Brandenburg/20/2009 | ADM14730 | A/Ft Carson/WR1448P/2009 |
| ACT79155 | A/Brawley/40082/2009 | ADF80463 | A/Gangwon/1805/2009 |
| ACZ97356 | A/Brownsville/39H/2009 | ACS72680 | A/Georgia/03/2009 |
| ADK33734 | A/Brussels/INS243/2009 | ACX56269 | A/Ghom/1550/2009 |
| ACP41107 | A/California/04/2009 | ADJ40647 | A/Guam/NHRC0002/2009 |
| ACP41931 | A/California/05/2009 | ADL29753 | A/Guam/NHRC0010/2009 |
| ACT36688 | A/California/07/2009 | ADL32020 | A/Guam/NHRC0015/2009 |
| ACT36692 | A/California/12/2009 | ADL32357 | A/Guam/NHRC0023/2009 |
| ADG42736 | A/California/VRDL100/2009 | ADM12894 | A/Guam/NHRC0030/2009 |

Table D.5: H2N2 Protein Sequences

| Uniprot ID | Description | Uniprot ID | Description |
|---|---|---|---|
| AAA43413 | A/RI/5-/1957 | AAO46221 | A/Panama/1/61 |
| AAA96710 | A/Leningrad/134/17/1957 | AAO46222 | A/Taiwan/1/1962 |
| AAA96711 | A/Leningrad/134/47/1957 | AAO46225 | A/Netherlands/60/62 |
| BAC77660 | A/Japan/305/1957 | AAO46226 | A/Albany/1/63 |
| AAO46211 | A/Singapore/1/1957 | AAO46227 | A/Netherlands/65/63 |
| AAO46212 | A/Chile/13/57 | BAC77661 | A/Murakami/4/64 |
| AAO46213 | A/Davis/1/57 | AAO46228 | A/Taiwan/1964 |
| AAO46214 | A/El Salvador/2/1957 | ABF21331 | A/Taiwan/1964 |
| BAD16637 | A/Adachi/2/1957 | ACD85190 | A/Berlin/3/1964 |
| BAD16638 | A/Singapore/1/1957 | ACD85223 | A/Cottbus/1/1964 |
| AAT66420 | A/Japan/305/1957 | BAA04722 | A/Kumamoto/5/1965 |
| ABF21326 | A/Japan/305/1957 | AAO46230 | A/NewJersey/3/65 |
| ABF82433 | A/RI/5+/1957 | AAO46231 | A/Albany/1/65 |
| ABF82434 | A/RI/5+/1957 | AAO46232 | A/Pittsburgh/2/65 |
| ABF82435 | A/RI/5+/1957 | AAO46233 | A/Kumamoto/1/1965 |
| BAF48642 | A/Singapore/1/1957 | ACD56327 | A/Moscow/1019/1965 |
| ABP49462 | A/Albany/22/1957 | ACD85201 | A/Potsdam/2/1965 |

## D. LIST OF INFLUENZA NEURAMINIDASE A PROTEIN SEQUENCES

Table D.5: (continued)

| | | | |
|---|---|---|---|
| ACD56294 | A/Ann Arbor/23/1957 | AAO46234 | A/Berkeley/1/1966 |
| ACD85234 | A/Guiyang/1/1957 | AAO46236 | A/Canada/1/1966 |
| ACF54491 | A/Singapore/1-MA12/1957 | AAO46237 | A/Panama/1/66 |
| ACF54524 | A/Singapore/1-MA12D/1957 | ACI25727 | A/Czech Republic/1/1966 |
| ACU79962 | A/Japan/305/1957 | AAB05621 | A/Tokyo/3/1967 |
| ABI84962 | A/Japan/305/1957 | AAO46238 | A/Panama/1/67 |
| ADG59707 | A/El Salvador/2/1957 | AAO46239 | A/AnnArbor/7/1967 |
| ADG59718 | A/El Salvador/2-Q226L/1957 | AAO46240 | A/England/10/67 |
| ABO38310 | A/Albany/26/1957 | AAO46242 | A/Poland/5/67 |
| AAO46215 | A/Albany/6/58 | AAO46243 | A/Montevideo/2208/67 |
| AAO46216 | A/Malaya/16/58 | AAO46244 | A/Georgia/1/1967 |
| ABO38299 | A/Albany/4/1958 | AAO46245 | A/Tokyo/3/1967 |
| ABO52305 | A/Albany/3/1958 | AAO46246 | A/Johannesburg/567/1967 |
| ABP49440 | A/Albany/24/1958 | ABO38704 | A/Albany/6/1967 |
| ABO52239 | A/Albany/5/1958 | ABO44060 | A/Albany/9/1967 |
| AAO46217 | A/Sao Paolo/3/1959 | ABO44104 | A/Albany/8/1967 |
| AAO46218 | A/Victoria/15681/59 | ACD85245 | A/Tashkent/1046/1967 |
| ABO44093 | A/Albany/1/1959 | ACD85256 | A/Johannesburg/617/1967 |
| AAO46219 | A/Ann Arbor/6/1960 | AAO46248 | A/Korea/426/1968 |
| ABQ01358 | A/Albany/1/1960 | AAO46249 | A/Berkeley/1/1968 |
| AAO46220 | A/England/1/61 | ACF41694 | A/Berkeley/1/1968 |

Table D.6: H3N2 Protein Sequences

| Uniprot ID | Description | Uniprot ID | Description |
|---|---|---|---|
| BAD16642 | A/Aichi/2/1968 | ACF36462 | A/Hong Kong/CUHK10132/1999 |
| ABQ53724 | A/Alabama/01/1998 | ACF36463 | A/Hong Kong/CUHK10133/1999 |
| ABQ01369 | A/Albany/1/1976 | ACF36464 | A/Hong Kong/CUHK10151/1999 |
| ABO33072 | A/Albany/14/1978 | ACF36469 | A/Hong Kong/CUHK10537/1998 |
| ABP49495 | A/Albany/15/1976 | ACF36470 | A/Hong Kong/CUHK10554/1998 |
| ABO52360 | A/Albany/17/1968 | ACF36472 | A/Hong Kong/CUHK10632/1998 |
| ABO44071 | A/Albany/18/1968 | ACF36473 | A/Hong Kong/CUHK10660/1998 |
| ABN51124 | A/Albany/19/1968 | ACF36474 | A/Hong Kong/CUHK10756/1998 |
| ABO52338 | A/Albany/20/1974 | ACF36475 | A/Hong Kong/CUHK10954/1998 |
| ABS49913 | A/Albany/3/1970 | ACF36477 | A/Hong Kong/CUHK10960/1998 |
| ABO52349 | A/Albany/4/1977 | ACF36478 | A/Hong Kong/CUHK11163/1999 |
| ABO52316 | A/Albany/42/1975 | ACF36485 | A/Hong Kong/CUHK12160/1997 |
| ABO52371 | A/Albany/6/1970 | ACF36491 | A/Hong Kong/CUHK12563/1997 |
| AAG49328 | A/Athens/135/1999 | ACF36493 | A/Hong Kong/CUHK12572/1999 |
| AAG49333 | A/Athens/16/1998 | ACF36494 | A/Hong Kong/CUHK12580/1999 |
| AAG49330 | A/Athens/2/1997 | ACF36496 | A/Hong Kong/CUHK12794/1997 |
| ABQ53717 | A/Athens/23/1997 | ACF36497 | A/Hong Kong/CUHK12897/1999 |
| AAG49323 | A/Athens/76/1998 | ACF36512 | A/Hong Kong/CUHK13763/1999 |
| ABU80180 | A/Auckland/583/2000 | ACF36513 | A/Hong Kong/CUHK14627/1999 |
| ABS50023 | A/Auckland/600/2000 | ACF36514 | A/Hong Kong/CUHK14672/1999 |

Table D.6: (continued)

| | | | |
|---|---|---|---|
| AAA43386 | A/Bangkok/1/1979 | ACF36517 | A/Hong Kong/CUHK16614/1998 |
| ABF21324 | A/Bangkok/1/1979 | ACF36518 | A/Hong Kong/CUHK17464/1998 |
| ACF41859 | A/Beijing/32/1992 | ACF36519 | A/Hong Kong/CUHK17603/1998 |
| ABF21325 | A/Beijing/353/1989 | ACF36520 | A/Hong Kong/CUHK17643/1998 |
| ABB46395 | A/Beijing/39/1975 | ACF36521 | A/Hong Kong/CUHK17697/1998 |
| ABQ53708 | A/Brazil/207/1996 | ACF36523 | A/Hong Kong/CUHK17707/1998 |
| ABQ53726 | A/Brazil/97/1997 | ACF36525 | A/Hong Kong/CUHK18036/1998 |
| AAQ10332 | A/Buenos Aires/4057/95 | ACF36526 | A/Hong Kong/CUHK18194/1998 |
| AAQ10333 | A/Buenos Aires/4064/95 | ACF36527 | A/Hong Kong/CUHK18218/1998 |
| AAQ10339 | A/Buenos Aires/4559/96 | ACF36529 | A/Hong Kong/CUHK18351/1998 |
| ABQ53694 | A/Bur/23/1996 | ACF36530 | A/Hong Kong/CUHK18358/1998 |
| AAO46480 | A/Canada/2/70 | ACF36531 | A/Hong Kong/CUHK18610/1998 |
| ABQ53695 | A/Canada/2974/1997 | ACF36533 | A/Hong Kong/CUHK19579/1998 |
| ABQ53701 | A/Canada/318/1996 | ACF36536 | A/Hong Kong/CUHK20173/2000 |
| AAY88205 | A/Canada/33312/99 | ACF36538 | A/Hong Kong/CUHK20200/2000 |
| ABC85790 | A/Canterbury/17/2000 | ACF36539 | A/Hong Kong/CUHK20213/1997 |
| ABD60837 | A/Canterbury/3/2000 | ACF36540 | A/Hong Kong/CUHK20217/1997 |
| ABD16295 | A/Canterbury/56/2000 | ACF36541 | A/Hong Kong/CUHK20236/1997 |
| ABF82654 | A/Canterbury/62/2000 | ACF36542 | A/Hong Kong/CUHK20292/2000 |
| ABJ09110 | A/Canterbury/75/2000 | ACF36545 | A/Hong Kong/CUHK20320/1997 |
| AAY88201 | A/Charlottesville/10/99 | ACF36546 | A/Hong Kong/CUHK20523/1997 |
| AAY88204 | A/Charlottesville/49/99 | ACF36548 | A/Hong Kong/CUHK20992/1997 |
| ABQ53702 | A/Dakar/5/1997 | ACF36554 | A/Hong Kong/CUHK21250/2000 |
| ABQ53703 | A/Delaware/1/1997 | ACF36566 | A/Hong Kong/CUHK22013/1997 |
| ABU92840 | A/Denmark/201/2000 | ACF36567 | A/Hong Kong/CUHK22048/1997 |
| ABU92841 | A/Denmark/202/2000 | ACF36568 | A/Hong Kong/CUHK22072/2000 |
| ABU92842 | A/Denmark/203/2000 | ACF36569 | A/Hong Kong/CUHK22078/2000 |
| ABQ53698 | A/Denmark/22511/1998 | ACF36570 | A/Hong Kong/CUHK22087/2000 |
| ABU92771 | A/Denmark/38/2000 | ACF36572 | A/Hong Kong/CUHK22163/2000 |
| CAD29985 | A/Denmark/41/2000 | ACF36598 | A/Hong Kong/CUHK26846/2000 |
| CAD29960 | A/Denmark/511/1998 | ACF36600 | A/Hong Kong/CUHK26969/2000 |
| ABQ53727 | A/Eng/23/1996 | ACF36601 | A/Hong Kong/CUHK26980/2000 |
| AAO46488 | A/England/42/1972 | ACF36602 | A/Hong Kong/CUHK27157/2000 |
| AAO46476 | A/England/878/1969 | ACF36603 | A/Hong Kong/CUHK27183/2000 |
| CAD29993 | A/Finland/620/99 | ACF36604 | A/Hong Kong/CUHK27348/2000 |
| AAO46470 | A/Georgia/122/68 | ACF36605 | A/Hong Kong/CUHK27374/2000 |
| AAG49325 | A/Greece/106/98 | ACF36606 | A/Hong Kong/CUHK28038/2000 |
| AAG49326 | A/Greece/109/99 | ACF36608 | A/Hong Kong/CUHK28044/2000 |
| AAG49327 | A/Greece/132/99 | ACF36616 | A/Hong Kong/CUHK31448/1999 |
| AAG49334 | A/Greece/18/98 | ACF36617 | A/Hong Kong/CUHK31490/1999 |
| ABC67568 | A/Guandong/243/1972 | ACF36618 | A/Hong Kong/CUHK31510/1999 |
| ABQ53709 | A/Guangdong/57/1998 | ACF36619 | A/Hong Kong/CUHK32796/1999 |
| ACF41738 | A/Hong Kong/1-1-MA-12/1968 | ACF36629 | A/Hong Kong/CUHK33829/1999 |
| ACF22213 | A/Hong Kong/1-1/1968 | ACF36633 | A/Hong Kong/CUHK33915/1999 |
| ACF41782 | A/Hong Kong/1-11-MA21-1/1968 | ACF36647 | A/Hong Kong/CUHK41114/1997 |
| ACF41815 | A/Hong Kong/1-12-MA21-1/1968 | ACF36648 | A/Hong Kong/CUHK41222/1997 |
| ACU79907 | A/Hong Kong/1-6-MA21-1/1968 | ACF36649 | A/Hong Kong/CUHK41459/1997 |

## D. LIST OF INFLUENZA NEURAMINIDASE A PROTEIN SEQUENCES

Table D.6: (continued)

| | | | |
|---|---|---|---|
| ACF54447 | A/Hong Kong/1-8-MA21-1/1968 | ACF36651 | A/Hong Kong/CUHK41507/1997 |
| ACF54458 | A/Hong Kong/1-8-MA21-3/1968 | ACF36652 | A/Hong Kong/CUHK41757/1997 |
| ACF41749 | A/Hong Kong/1-9-MA21-1/1968 | ACF36654 | A/Hong Kong/CUHK4245/1997 |
| ACF41760 | A/Hong Kong/1-9-MA21-2/1968 | ACF36658 | A/Hong Kong/CUHK4391/1997 |
| ACF41771 | A/Hong Kong/1-9-MA21-3/1968 | ACF36661 | A/Hong Kong/CUHK4542/1997 |
| AAK51726 | A/Hong Kong/1/1968 | ACF36663 | A/Hong Kong/CUHK4803/1997 |
| ACU79874 | A/Hong Kong/1/1968 | ACF36668 | A/Hong Kong/CUHK50552/1998 |
| ABB46406 | A/Hong Kong/1/1982 | ACF36671 | A/Hong Kong/CUHK50722/1998 |
| ABB04341 | A/Hong Kong/11/1973 | AAO46486 | A/Hungary/2/1971 |
| AAK63827 | A/Hong Kong/1143/99 | ABQ53718 | A/Inverness/580868097/1997 |
| AAK63828 | A/Hong Kong/1143/99 | CAD29992 | A/Ireland/10586/99 |
| AAK63829 | A/Hong Kong/1144/99 | ABQ53710 | A/Japan/1268/1998 |
| AAK70429 | A/Hong Kong/1179/99 | CAD29974 | A/Johannesburg/33/1994 |
| AAK70431 | A/Hong Kong/1180/99 | BAD16645 | A/Kumamoto/55/76 |
| AAK70432 | A/Hong Kong/1180/99 | ABF21327 | A/Leningrad/360/1986 |
| ABB04297 | A/Hong Kong/14/1974 | CAD29984 | A/Lyon/1242/2000 |
| ABB04909 | A/Hong Kong/14/1992 | CAC87416 | A/Lyon/2573/1998 |
| CAC40040 | A/Hong Kong/1774/99 | AAO46468 | A/Malaysia/221/68 |
| CAD29961 | A/Hong Kong/1789/2000 | ABR20817 | A/Mecklenburg-Vorpommern/9/99 |
| ABB04319 | A/Hong Kong/2/1988 | BAC77664 | A/Memphis/1/1971 |
| ABB04330 | A/Hong Kong/24/1985 | ABB96377 | A/Memphis/1/1977 |
| ABB04920 | A/Hong Kong/26/1983 | ABA43339 | A/Memphis/1/1990 |
| ABB80037 | A/Hong Kong/3/1969 | ABB96355 | A/Memphis/101/1974 |
| ABD60793 | A/Hong Kong/33/1973 | BAD16643 | A/Memphis/102/1972 |
| ABB79802 | A/Hong Kong/4/1984 | ABC84545 | A/Memphis/103/1972 |
| ABB04931 | A/Hong Kong/43/1975 | ABC97377 | A/Memphis/105/1972 |
| ABB46550 | A/Hong Kong/45/1980 | ABD61760 | A/Memphis/105/1976 |
| ABB04286 | A/Hong Kong/46/1980 | ABD16743 | A/Memphis/108/1976 |
| ABC40622 | A/Hong Kong/49/1974 | ABB96333 | A/Memphis/12/1978 |
| ABB80026 | A/Hong Kong/50/1972 | ABG37211 | A/Memphis/12/1985 |
| ABQ97206 | A/Hong Kong/68 | ABB96512 | A/Memphis/13/1988 |
| ABB04953 | A/Hong Kong/7/1984 | ABB96344 | A/Memphis/18/1978 |
| ACF36461 | A/Hong Kong/CUHK10100/1999 | ABB96322 | A/Memphis/2/1978 |

Table D.7: H1N2 Protein Sequences

| Uniprot ID | Description | Uniprot ID | Description |
|---|---|---|---|
| CAD29972 | A/Egypt/84/2001 | ABU92910 | A/Denmark/86/2003 |
| AAN64882 | A/Wisconsin/12/2001 | ABU92911 | A/Denmark/56/2003 |
| AAN64883 | A/Texas/7/2001 | ABU92912 | A/Denmark/12/2003 |
| AAN64888 | A/Egypt/21181/2001 | ABB83018 | A/New York/481/2003 |
| AAN64890 | A/India/77251/2001 | AAZ79552 | A/New York/300/2003 |
| CAD35672 | A/England/627/01 | ABB53732 | A/New York/491/2003 |
| CAD35673 | A/England/691/01 | ABA42283 | A/New York/226/2003 |
| ABU50446 | A/Pennsylvania/1/2001 | AAZ74377 | A/New York/296/2003 |
| BAD02347 | A/Yokohama/22/2002 | ABB83141 | A/New York/492/2003 |
| ABU50429 | A/Pennsylvania/1/2002 | ABD94946 | A/New York/229/2003 |
| ABC40634 | A/New York/417/2002 | ABB53606 | A/New York/219/2003 |
| AAY78942 | A/New York/78/2002 | ABU50434 | A/Virginia/20/2003 |
| ABU50432 | A/New York/26/2002 | ABK57099 | A/Philippines/344/2004 |
| ABB02784 | A/New York/217/2002 | | |

Table D.8: H5N1 ASIA Protein Sequences

| Uniprot ID | Description | Uniprot ID | Description |
|---|---|---|---|
| ABU94738 | A/Anhui/1/2005 | ABI49417 | A/Indonesia/CDC759/2006 |
| ADG59213 | A/Anhui/1/2005 | ABL31746 | A/Indonesia/CDC835/2006 |
| ABV23945 | A/Azerbaijan/002-115/2006 | ABL31782 | A/Indonesia/CDC887/2006 |
| ABV23981 | A/Azerbaijan/006-207/2006 | ABL07010 | A/Indonesia/CDC938/2006 |
| ABV23961 | A/Azerbaijan/011-162/2006 | ABL07032 | A/Indonesia/CDC940/2006 |
| ABO10176 | A/Cambodia/JP52a/2005 | ACB87566 | A/Jiangsu/2/2007 |
| ADM95394 | A/Cambodia/Q0405047/2006 | BAH24021 | A/Shanghai/1/2006 |
| ACI06179 | A/Cambodia/R0405050/2007 | ABC72646 | A/Thailand/676/2005 |
| ADM95372 | A/Cambodia/S1211394/2008 | ABJ98533 | A/Thailand/NA60/2005 |
| ABI16506 | A/China/GD01/2006 | ACU46645 | A/Thailand/NBL1/2006 |
| ABX57872 | A/China/GD02/2006 | ABD16286 | A/Thailand/NK165/2005 |
| ADG59235 | A/Guangxi/1/2005 | ABJ98528 | A/Thailand/RPFE/2005 |
| ABW06346 | A/Indonesia/160H/2005 | ABQ58916 | A/Turkey/12/2006 |
| ABW06357 | A/Indonesia/175H/2005 | ABQ58918 | A/Turkey/651242/2006 |
| ABW06314 | A/Indonesia/245H/2005 | AAZ72720 | A/Vietnam/BL-014/2005 |
| ABW06303 | A/Indonesia/283H/2006 | AAZ72721 | A/Vietnam/DT-036/2005 |
| ABW06294 | A/Indonesia/286H/2006 | AAZ72722 | A/Vietnam/HG-207/2005 |
| ABW06254 | A/Indonesia/321H/2006 | ABO10180 | A/Vietnam/HN30408/2005 |
| ABW06243 | A/Indonesia/341H/2006 | ABY19420 | A/Vietnam/HN31242/2007 |
| ABW06107 | A/Indonesia/5/2005 | ABY19421 | A/Vietnam/HN31242/2007 |
| ABW06199 | A/Indonesia/535H/2006 | ABO10179 | A/Vietnam/JP14/2005 |
| ABW06210 | A/Indonesia/538H/2006 | ABO10178 | A/Vietnam/JP4207/2005 |
| ABW06159 | A/Indonesia/560H/2006 | ABE97719 | A/Vietnam/CL105/2005 |
| ABW06376 | A/Indonesia/6/2005 | ABB76120 | A/Vietnam/CL115/2005 |
| ABM90480 | A/Indonesia/CDC1032/2007 | ADF83610 | A/Vietnam/HN31388M1/2007 |
| ABM90513 | A/Indonesia/CDC1046/2007 | ADF83606 | A/Vietnam/HN31432M/2008 |

Table D.8: (continued)

| | | | |
|---|---|---|---|
| ABI36200 | A/Indonesia/CDC523/2006 | ABF56651 | A/Vietnam/PEV16T/2005 |
| ABI36146 | A/Indonesia/CDC594/2006 | ADF83617 | A/Vietnam/UT30408III/2005 |
| ABI36314 | A/Indonesia/CDC610/2006 | ADF83616 | A/Vietnam/UT30850/2005 |
| ABI36325 | A/Indonesia/CDC623/2006 | ADF83615 | A/Vietnam/UT31203A/2007 |
| ABI36347 | A/Indonesia/CDC624/2006 | ADF83611 | A/Vietnam/UT31312II/2007 |
| ABI36380 | A/Indonesia/CDC634/2006 | ADF83609 | A/Vietnam/UT31394II/2008 |
| ABI36465 | A/Indonesia/CDC644T/2006 | ADF83608 | A/Vietnam/UT31412II/2008 |
| ABI36435 | A/Indonesia/CDC669/2006 | ADF83607 | A/Vietnam/UT31413II/2008 |
| ABI49409 | A/Indonesia/CDC742/2006 | ABG23658 | A/Zhejiang/16/2006 |

# Appendix E

# List of Allergen and Non-Allergen Protein Sequences

# E. LIST OF ALLERGEN AND NON-ALLERGEN PROTEIN SEQUENCES

Table E.1: Uniprot IDs for Allergens Training Set - Allerhunter vs Allergenonline Database

| ID | ID | ID | ID | ID | ID |
|---|---|---|---|---|---|
| P83885.1 | ABH06347.1 | ACK76297.1 | Q9M7M9.1 | AAK58515.1 | P02761.1 |
| CAB58171.1 | P82971.1 | ACK76299.1 | Q9M7M8.1 | AAD05375.1 | Q63213 |
| BAC77154.1 | ACI01048.1 | BAD74060.2 | Q9LEI8.1 | AAO33897.1 | XP_001067036.1 |
| CAK50389.1 | AAA30478.1 | P16311.2 | CAA11041.1 | AAL92578.1 | Q870B9.1 |
| BAF43534.1 | AAA30479.1 | BAC53948.1 | CAA11042.1 | ACZ57582.1 | AAT37679.1 |
| Q7M1Y0 | AAA30480.1 | BAA04558.1 | AAR98518.1 | AAY88919.1 | P01089.2 |
| Q7M1X6 | AAB29137.1 | AAM64112.1 | CAE85467.1 | I53806 | Q91483.3 |
| G37396 | CAA76847.1 | BAA01241.1 | 1WKX | E53806 | Q91482.1 |
| A59055 | NP_851341.1 | AAB30829.1 | Q39967.3 | F53806 | ACI68103.1 |
| Q9BMK4.1 | P02754.3 | AAL47677.1 | ABN03965.1 | G53806 | P83181_1 |
| Q7M4I5.1 | P00711.2 | AAA99805.1 | CAD24068.1 | D53806 | P83181_2 |
| P01502.1 | Q28133.1 | P49275.2 | CAA93121.1 | H53806 | P83181_3 |
| MEHB2 | ABW98943.1 | Q26456.1 | AAG42255.1 | BAE54432.1 | P83181_4 |
| P01501.1 | ABW98945.1 | AAM19082.1 | CAB10765.1 | P86432.1 | ACS34771.1 |
| NP_001035360.1 | ABW98953.1 | P39673.1 | CAB10766.1 | P86431.1 | AAX11262.1 |
| ABF21077.1 | ACG59280.1 | P49276.2 | CAA10140.1 | BAF95206.1 | AAT99258.1 |
| ABF21078.1 | P02769.4 | AAB27594.1 | P43216.1 | AAA86533.1 | ACO34813.1 |
| ABD51779.1 | NP_776945.1 | CAI05850.1 | O44119.1 | Q948T6.2 | AAX11261.1 |
| P00630.3 | Q28050.1 | CAI05849.1 | AAC48288.1 | AAF72991.1 | CAX32966.1 |
| Q08169.1 | CAA29664.1 | P16312.1 | CAA42832.1 | BAA07774.1 | CAX32967.1 |
| P81943_1 | CAA32835.1 | AAY84563.1 | CAA35188.1 | BAA07710.1 | AAO15607.1 |
| P81943_2 | AAA62707.1 | ABA39437.1 | P16968.2 | BAA07711.1 | AAX37321.1 |
| P81943_3 | AAA30430.1 | ABB52642.1 | AAA32970.1 | BAA07712.1 | CAQ68366.1 |
| P81943_4 | AAA30429.1 | 2AS8 | P32936.2 | BAA07713.1 | BAH10151.1 |
| P92918.1 | AAA30428.1 | ABC73706.1 | CAA41956.1 | AAB99797.1 | D37396 |
| AAD29409.1 | AAA30413.1 | CAK22338.1 | CAA49555.1 | BAB71741.1 | Q7M1Y1 |
| P49372.1 | AAA30431.1 | CAA46317.1 | CAA08836.1 | Q01882.2 | C37396 |
| AAT00595.1 | AAA30433.1 | P46419.1 | CAA46705.1 | Q01883.2 | AAP06493.1 |
| AAT00594.1 | P80207.1 | P49273.1 | AAP94213.1 | BAC19997.1 | AAC67308.1 |
| AAT00596.1 | P80208.1 | ABG76196.1 | AAP15199.1 | BAC20650.1 | BAC66618.1 |
| AAM93157.1 | S65144 | ABV66255.1 | AAP15200.1 | Q40638.2 | CAX32965.1 |
| ABI17154.1 | S65145 | 1A9V | AAM54366.1 | BAC20657.1 | Q9S8H2 |
| ABP97433.1 | AAN86249.1 | ABY53034.1 | AAM54365.1 | BAA01998.1 | CAH92627.1 |
| 1W2Q | P69198.1 | 3F5V | AAF18269.1 | BAA01996.1 | CAH92630.1 |
| ACA79908.1 | P69196.1 | ACG58378.1 | AAB41308.1 | BAA07772.1 | Q7M263 |
| AAK96887.1 | S65143 | ACI32128.1 | ACI47547.1 | BAA07773.1 | BAE54429.1 |
| ACN62248.1 | P81729.1 | CAQ68249.1 | AAW29810.1 | ACA96507.1 | BAE54430.1 |
| ABW17159.1 | P69197.1 | CAQ68250.1 | P81294.1 | BAF47264.1 | AAK15089.1 |
| ACH91862.1 | P69199.1 | AAA28296.1 | Q9FY19.1 | O61379.1 | ACI41244.1 |
| 3C3V | AAN11300.1 | P08176.2 | P81295.1 | BAF47266.1 | ACH85188.1 |
| ABX56711.1 | P30575.1 | CAD38361.1 | CAC48400.1 | BAF47265.1 | Q9XHP2 |
| ABX75045.1 | O18873.1 | CAD38362.1 | O64943.2 | P55958.1 | AAK15088.1 |
| Q647G9.1 | O18874.1 | CAD38363.1 | AAR21072.1 | CAA65122.1 | AAD42943.1 |
| ADB96066.1 | CAD82911.1 | CAD38364.1 | AAR21071.1 | Q9T0M8.1 | CAA62910.1 |
| P43237.1 | CAD82912.1 | CAD38365.1 | AAF80166.1 | Q9XG85.1 | P15322.2 |
| P43238.1 | AAB30434.1 | CAD38366.1 | Q9LD79.2 | Q40905.1 | AAX77383.1 |
| AAM78596.1 | CAA76841.1 | CAD38367.1 | AAF80164.1 | P43217.3 | AAX77384.1 |
| AAN77576.1 | CAB64867.1 | CAD38368.1 | CAD87731.1 | 2008179A | CAA62909.1 |
| AAC63045.1 | CAD10376.1 | CAD38369.1 | CAD87730.1 | CAA65123.1 | CAA62908.1 |
| AAD47382.1 | P00784.1 | CAD38370.1 | AAQ73492.1 | AAB46819.1 | CAA62911.1 |
| AAD55587.1 | ABZ81043.1 | CAD38371.1 | AAQ73493.1 | Q7M1E8 | CAA62912.1 |
| AAD56337.1 | ABZ81042.1 | AAB69424.1 | AAQ73494.1 | AAB36008.1 | P16348.1 |
| AAD56719.1 | ABZ81041.1 | CAA75141.1 | P80384.2 | AAB36009.1 | P20347.3 |
| AAQ91847.1 | CAA47357.1 | AAM21322.1 | Q9NFZ4.1 | AAB36010.1 | P30941.2 |
| AAB22817.1 | P38949.2 | P49278.1 | Q9U5P1.1 | AAB36011.1 | P15476.2 |
| AAL37561.1 | P38950.2 | 1KTJ | S66499 | AAB36012.1 | ABB16985.1 |
| AAM46958.1 | CAB02214.1 | CAD38372.1 | CAD32313.1 | AAB46820.1 | ABA81885.1 |
| CAG26895.1 | CAB02206.1 | CAD38373.1 | CAD32314.1 | ACA23876.1 | AAA33819.1 |
| CAK50834.1 | CAB02207.1 | CAD38374.1 | 2118249B | AAO15713.1 | CAA27588.1 |
| ACE07186.1 | CAB02208.1 | CAD38375.1 | 2118249A | BAF47262.1 | CAA27571.1 |
| ACE07187.1 | CAB02212.1 | CAD38376.1 | Q9U5P2.1 | AAX11194.1 | CAA31575.1 |
| ACE07188.1 | CAB02215.1 | CAD38377.1 | Q9U1G2.1 | AAB34785.1 | CAA45723.1 |
| ACE07189.1 | CAB02216.1 | CAD38378.1 | AAQ73484.1 | AAF71379.1 | BAH10156.1 |
| AAO24900.1 | CAB02217.1 | CAD38379.1 | AAQ73486.1 | AAM33821.1 | AAF65312.1 |
| AAX85388.1 | AAB20453.1 | CAD38381.1 | AAQ73487.1 | AAF23726.1 | AAF65313.1 |
| P0C088.1 | ABZ81044.1 | CAD38382.1 | AAQ73488.1 | AAG44693.2 | AAC97370.1 |
| Q8H2C9.3 | ABZ81040.1 | CAD38383.1 | AAQ73489.1 | AAD25926.1 | P35777.2 |
| Q8H2C8.3 | CAA64868.1 | AAA19973.1 | AAQ73490.1 | AAD25995.1 | AAB36117.1 |
| CAD23613.1 | CAD10374.1 | AAD38942.1 | AAQ73491.1 | AAG44480.1 | AAB36119.1 |
| CAD23614.1 | ACJ23863.1 | P14004.2 | CAP17694.1 | Q96X46.3 | AAB36120.1 |
| CAD23611.1 | ACJ23861.1 | AAB32842.1 | CAC84590.2 | Q92260.1 | AAB36121.1 |
| BAH09387.1 | ACJ23862.1 | CAD69036.1 | CAC84593.2 | AAD42074.1 | AAT95008.1 |

Table E.1: (continued)

| | | | | | |
|---|---|---|---|---|---|
| AAD13651.1 | P83507.1 | P49277.1 | O82015.2 | AAR17475.1 | P35775.1 |
| AAD13652.1 | P83508.1 | CAC09234.1 | CAA54819.1 | AAG44478.1 | P35778.2 |
| AAB93837.1 | Q7M1E7.1 | AAO73464.1 | AAZ91659.1 | ADB92492.1 | P35779.2 |
| AAB93839.1 | BAF32143.1 | AAN02510.1 | AAL07320.1 | ADB92493.1 | P35776.2 |
| AAD13646.1 | Q96385.1 | P49274.1 | ABC02750.1 | ADD17628.1 | P34071.1 |
| ACN32322.1 | Q9N2R3.1 | AAX47076.1 | ACM89179.1 | AAC34736.1 | P20723.1 |
| AAD13648.1 | Q8LGR0.1 | AAX37326.1 | ACC76803.1 | AAB82404.1 | 1ESF |
| AAD13650.1 | ACR77509.1 | ABC96702.1 | ABI98020.1 | AAD13533.1 | CAJ43561.1 |
| AAD13644.1 | AAL92870.1 | Q05108.1 | CAA10350.1 | AAC34312.1 | P06886.1 |
| AAD13645.1 | AAL92871.1 | P10737.3 | Q9SC98 | AAC34737.1 | AAS75831.1 |
| AAD13647.1 | BAF47267.1 | P10736.1 | Q7M1X5.1 | AAB09632.1 | P00791.3 |
| AAD13649.1 | O96764.2 | P53357.1 | P14947.1 | AAB62731.1 | S43242 |
| AAB26195.1 | P02221.2 | Q06478.1 | CAA51775.1 | AAB63595.1 | S43243 |
| Q06811.2 | P02222.2 | P49371.1 | P14948.1 | AAD19606.1 | S43244 |
| Q9UVU3 | P02231.1 | P83340.1 | Q40240.1 | Q25641.1 | ACT33296.1 |
| Q8J0P4.2 | P02224.2 | P81217.1 | Q40237.1 | AAP13554.1 | Q7M4K8 |
| Q8NKF4.2 | P84296.1 | P82615 | AAD20386.1 | AAX33728.1 | AAK63089.1 |
| Q96X30.3 | P12548.1 | P35747.1 | CAB64344.1 | Q9UB83.1 | AAK63088.1 |
| Q9UUZ6.2 | P84298.1 | P82615.3 | CAH92637.1 | 3EBW | BAE54431.1 |
| CAA06305.1 | P12549.1 | Q95182.1 | AAA63279.1 | ACJ37391.1 | BAE46763.1 |
| AAF86369.1 | P12550.1 | P81216.1 | P14946.2 | ACS14052.1 | BAH10155.1 |
| CAA59419.1 | P02227.1 | BAF47269.1 | AAA63278.1 | AAL86701.1 | AAF07903.2 |
| CAB44442.1 | P02228.1 | BAF47268.1 | ACB05815.1 | AAG08988.1 | AAD52013.1 |
| P40292.3 | P02229.2 | BAF76431.1 | ABR21772.1 | CAB01591.1 | AAD52012.1 |
| AAC61261.1 | P02230.1 | BAF76430.1 | ABR21771.1 | P56167.1 | CAD23374.1 |
| AAB07620.1 | P84160.1 | AAC82350.1 | CAD10377.1 | Q41260.1 | CAA31396.1 |
| P79017.2 | P84177.2 | Q9TZZ2.2 | AAL29690.1 | P56166.1 | CAA25593.1 |
| AAK49451.1 | P84161.1 | BAB79444.1 | AAL75449.1 | P56165.1 | CAA26383.1 |
| AAB95638.1 | P84159.1 | AAX57578.1 | AAL75450.1 | P56164.1 | CAA26384.1 |
| CAA04959.1 | CAH03799.1 | ABC18306.1 | CAJ19706.1 | AAB27445.1 | CAA26385.1 |
| AAB60779.1 | ABQ59329.1 | O23878.1 | CAJ19705.1 | ADC80502.1 | CAA31685.1 |
| O42799.2 | AAK67491.1 | O23880.1 | ADC55380.1 | ADC80503.1 | CAA26847.1 |
| CAB64688.1 | CAB39376.1 | Q9XFM4.1 | Q01940.1 | AAC25998.1 | CAA30570.1 |
| CAA11266.1 | CAA50326.1 | ABQ10638.1 | P56577.1 | O82040.1 | CAA24934.1 |
| CAA73782.1 | CAA96548.1 | AAF34634.1 | P56578.1 | ABB78007.1 | CAA43361.1 |
| AAM43909.1 | CAA96549.1 | ABO93594.1 | AAD25927.1 | CAD38384.1 | AAA34272.1 |
| P67875.1 | AAD48405.1 | ABI32184.1 | CAI43283.4 | CAD38385.1 | AAA34274.1 |
| Q92450.3 | AAG40329.1 | AAZ76743.1 | 2J23 | CAD38386.1 | AAA34275.1 |
| O60024.2 | AAG40330.1 | Q8WNR9.1 | CAD68071.1 | CAD38387.1 | AAA34276.1 |
| EAL90121.1 | AAG40331.1 | P30440.1 | CAA09883.1 | CAD38388.1 | AAA34279.1 |
| EAL87477.1 | CAC14168.1 | CAA44343.1 | CAA09884.1 | CAD38389.1 | AAA34280.1 |
| EAL85811.1 | AAK01235.1 | CAA44344.1 | CAA09885.1 | CAD38390.1 | AAA34281.1 |
| EAL89830.1 | AAK01236.1 | P30438.2 | CAA09886.2 | CAD38391.1 | AAA34282.1 |
| AAD13106.1 | AAK28533.1 | AAC37318.1 | CAA09887.4 | CAD38392.1 | AAA34283.1 |
| CAB06417.1 | AAL73404.1 | CAA44345.1 | AAT80662.1 | CAD38393.1 | AAA34284.1 |
| AAA32702.1 | Q08407.3 | Q5VFH6.1 | AAT80659.1 | CAD38394.1 | AAA34285.1 |
| P12547.2 | AAO65960.1 | NP_001041618.1 | AAT80649.1 | CAD38395.1 | AAA34286.1 |
| P0C1B3.1 | ACO56333.1 | P49064.1 | AAR22488.1 | CAD38396.1 | AAA34287.1 |
| P29600.1 | AAL86739.1 | ACD65080.1 | AAX18307.1 | CAD38397.1 | AAA34288.1 |
| P00780.1 | CAA50328.1 | ACD65081.1 | AAX18318.1 | CAD87529.1 | AAA34289.1 |
| AAG31026.1 | CAA50325.1 | CAJ85641.1 | AAX19848.1 | 1NLX | AAB02788.1 |
| BAA05540.1 | BAH10152.1 | ABD39049.1 | AAX19851.1 | CAD54670.2 | AAA17741.1 |
| BAF46896.1 | AAK96889.1 | CAJ85642.1 | AAX19854.1 | CAG24374.1 | O22116 |
| Q7M3Y8_1 | AAC61869.1 | CAJ85646.1 | AAX19856.1 | CAF32567.2 | CAA27052.1 |
| Q7M3Y8_2 | AAW81034.1 | Q5ULZ4_3 | AAX19858.1 | CAF32566.2 | CAA59338.1 |
| BAH10149.1 | BAF32110.1 | Q5ULZ4_2 | AAX19860.1 | P43214.1 | CAA59339.1 |
| CAA38363.1 | BAF32116.1 | CAJ85644.1 | Q9FSG7.1 | P43213.1 | CAA59340.1 |
| P04403.2 | BAF32119.1 | AAQ83588.1 | Q40280.3 | Q8H6L7.1 | CAA24933.1 |
| AAO38859.1 | BAF32122.1 | AAV74343.1 | Q9M5X7.1 | 2118271A | CAA61944.1 |
| CAA07328.1 | BAF32128.1 | AAQ08947.1 | CAK93713.1 | CAQ55938.1 | CAA61945.1 |
| CAA07320.1 | BAF32130.1 | BAH10153.1 | CAK93753.1 | CAQ55939.1 | CAA61943.1 |
| 1CQA | BAF32133.1 | AAN73248.1 | CAK93757.1 | CAQ55940.1 | BAA11251.1 |
| A45786 | BAF32134.1 | Q8TFM8.1 | CAT99618.1 | CAQ55941.1 | AAB35353.1 |
| CAA54696.1 | BAF32137.1 | Q8TFM9.1 | CAT99619.1 | 1N10 | BAA12318.1 |
| CAA54695.1 | BAF51970.1 | P02622.1 | CAT99617.1 | CAA81613.1 | CAA72273.1 |
| CAA54694.1 | BAF45320.1 | AAK63086.1 | CAT99612.1 | P35079.1 | O22108 |
| P43177.2 | P18632.2 | AAK63087.1 | CAT99611.1 | CAA70609.1 | CAC14917.1 |
| CAA05186.1 | BAB86286.1 | CAM56785.1 | Q941P6 | CAA70610.1 | CAI79052.1 |
| CAA05187.1 | P43212.1 | CAM56786.1 | Q941P5 | CAB42886.1 | CAI78902.1 |
| CAA05188.1 | BAC23082.1 | CAA23681.1 | CAA88833.1 | S32101 | P81496.1 |
| CAA05190.1 | BAC23083.1 | P01012.2 | CAA96534.1 | S38584 | CAI64396.1 |
| CAA07318.1 | BAC23084.1 | CAA23682.1 | AAD26546.1 | Q7M1L8 | CAI64397.1 |
| CAA07319.1 | AAK27264.1 | 1JTI | AAD26547.1 | CAA50281.1 | CAI64398.1 |

Table E.1: (continued)

| | | | | | |
|---|---|---|---|---|---|
| CAA07323.1 | BAD77932.1 | CAA26040.1 | AAD26548.1 | Q40963.2 | BAE20328.1 |
| CAA07324.1 | BAA05543.1 | P02789.2 | AAD26552.1 | AAK25823.1 | CAR82265.1 |
| CAA07325.1 | BAF32105.1 | P00698.1 | AAD26553.1 | 1L3P | CAR82266.1 |
| CAA07326.1 | CAD92666.1 | AAA48944.1 | AAD26554.1 | CAA52753.1 | CAR82267.1 |
| CAA07327.1 | AAW69549.1 | 1UHG | AAD26555.1 | AAC25994.1 | ACG59281.1 |
| CAA07329.1 | P83834_1 | P19121.2 | AAD26558.1 | CAB05372.1 | P08819.2 |
| CAA07330.1 | P83834_2 | NP_001106132.1 | CAD10375.1 | AAC16525.1 | ABS58503.1 |
| CAA04823.1 | Q39547.1 | ACJ04729.1 | AAO25113.1 | AAC16526.1 | ACE82288.1 |
| CAA04826.1 | P83834_3 | CAX32963.1 | P43211.2 | AAC16527.1 | ACE82289.1 |
| CAA04827.1 | ACB45874.1 | P01005.1 | Q9XF40.1 | AAC16528.1 | ACE82290.1 |
| CAA04828.1 | AAP13533.2 | XP_385781.1 | Q9XF41.1 | CAB05371.1 | ACE82291.1 |
| CAA04829.1 | CAC05258.1 | AAF82096.1 | Q9XF42.1 | CAA76556.1 | ACE82292.1 |
| 1QMR | ABK78766.1 | ADD19989.1 | CAD46559.1 | CAA76557.1 | CAZ76052.1 |
| 1LLT | Q9SCG9.1 | ADD19985.1 | CAD46561.1 | 2023228A | CBA13559.1 |
| P15494.2 | ACY01951.1 | ADD18879.1 | CAD46560.1 | CAA76558.1 | P24296.2 |
| 1B6F | CAC37790.2 | CAA33217.1 | AAB35897.1 | AAC25995.1 | P82977.2 |
| AAD26560.1 | AAR21075.1 | CAA37044.1 | AAT80665.1 | AAC25997.1 | CAA35238.1 |
| AAD26561.1 | AAR21074.1 | CAA45777.1 | AAT80664.1 | CAD10390.1 | CAA35598.1 |
| AAD26562.1 | AAF72625.1 | CAA45778.1 | BAF47263.1 | ABO36677.1 | CAA42453.1 |
| P45431.2 | AAF72626.1 | P22895.1 | CAA73720.1 | ABG73110.1 | CAA35597.1 |
| P43176.2 | AAF72627.1 | AAA33947.1 | Q25456.1 | ABG73109.1 | CAA43331.1 |
| P43178.2 | AAF72628.1 | AAB01374.1 | AAG08989.1 | ABG73108.1 | AAB21323.1 |
| P43180.2 | AAF72629.1 | AAA33964.1 | P85894.1 | ABU42022.1 | S16031 |
| P43183.2 | AAL14078.1 | AAA33965.1 | P02762.2 | CAF25233.1 | CAA34709.1 |
| P43184.2 | AAK96255.1 | AAB23463.1 | AAK54834.1 | CAF25232.1 | CAA39099.1 |
| P43185.2 | AAL14077.1 | AAB23464.1 | AAB50883.1 | CAD80019.1 | CAA36063.1 |
| CAA96541.1 | AAL14079.1 | AAB23482.1 | Q07932.1 | CAC41633.1 | CAA44473.1 |
| CAA96546.1 | AAF80379.2 | AAB23483.1 | 2206305A | CAC41634.1 | AAA34290.1 |
| CAA96539.1 | CAA69670.1 | CAA56343.1 | Q26464.1 | CAC41635.1 | ACL36923.1 |
| CAA96540.1 | CAA62634.1 | CAA26478.1 | BAH10150.1 | ABY21305.1 | AAU11502.1 |
| CAA96542.1 | CAA01909.1 | AAB09252.1 | CAE17317.1 | Q8GT41.1 | ABZ81991.1 |
| CAA96543.1 | CAA01910.1 | BAA25899.1 | CAE17316.1 | CAE52833.1 | ABU97480.1 |
| CAA96544.1 | AAB28566.1 | BAB21619.1 | BAE54433.1 | CAC85911.1 | ABU97479.1 |
| CAA96547.1 | AAB28567.1 | ABA54897.1 | AAO22133.1 | A60373 | ABQ96644.1 |
| CAA96545.1 | AAB32317.1 | P82947.1 | AAO22132.1 | P22285.1 | ABM53751.1 |
| CAB02155.1 | O04701.1 | ABA54898.1 | AAN18044.1 | CAA10348.1 | O02380.1 |
| CAB02156.1 | CAC83659.1 | P26987.1 | AAQ10281.1 | A60372 | AAT40866.1 |
| CAB02157.1 | CAC83658.1 | ABU97472.1 | AAQ10280.1 | F37396 | P35781.1 |
| CAB02158.1 | CAD20406.1 | ACD36976.1 | AAQ10279.1 | CAA10520.1 | P35782.1 |
| CAB02159.1 | 2103117A | ACD36975.1 | AAQ10278.1 | AAG42254.1 | P81657.1 |
| CAB02160.1 | CAA10345.1 | ACD36974.1 | AAQ10277.1 | P22284.1 | P35783.1 |
| CAB02161.1 | AAK62278.1 | ACD36978.1 | AAQ10276.1 | P22286.1 | CAL59818.1 |
| P43179.2 | CAD20405.1 | AAB34755.1 | AAQ10274.1 | AAA29793.1 | CAL59819.1 |
| P43186.2 | P93124.1 | A57106 | AAQ10271.1 | Q9U6V9.1 | P35784.1 |
| AAP37482.1 | P82946_4 | CAA11755.1 | AAQ10268.1 | Q9U6W0.1 | CAJ28930.1 |
| P25816.1 | P82946_3 | O65809.1 | AAQ08190.1 | AAT95010.1 | CAJ28931.1 |
| P43187.1 | P82946_2 | CAA35691.1 | ABP58627.1 | AAS67044.1 | P35760.1 |
| Q39419.1 | P82946_1 | CAA26575.1 | ABP58632.1 | AAP37412.1 | P51528.1 |
| AAG22740.1 | AAP96759.1 | CAA26723.1 | ABP58633.1 | AAS67043.1 | ABC73068.1 |
| CAC84116.1 | BAB88129.1 | CAA33215.1 | ABP58635.1 | AAS67042.1 | P35785.1 |
| B45786 | O04298.1 | CAA33216.1 | ABP58636.1 | AAS67041.1 | P35786.1 |
| AAB20452.1 | Q8SAE6.1 | CAA60533.1 | ABP58637.1 | AAT95009.1 | P35787.1 |
| BAB21489.1 | BAA13604.1 | CAA55977.1 | ABX26131.1 | P35759.1 | 1QNX |
| BAB21490.1 | CAB03715.1 | Q9U5P7.1 | ABX26132.1 | P35780.1 | CAI77218.1 |
| BAB21491.1 | CAB03716.1 | Q9NFQ4.1 | ABX26134.1 | P83542.1 | 2ATM |
| AAB25850.1 | AAL76932.1 | AAQ54603.1 | ABX26138.1 | P83377.1 | AAB48072.1 |
| AAB25851.1 | AAB01092.1 | BAH10148.1 | ABX26139.1 | BAH59276.1 | P49370.1 |
| ABP04043.1 | P42039.3 | AAG08987.1 | ABX26140.1 | P81651.2 | AAA30333.1 |
| ABP04044.1 | CAA55070.1 | P23110.1 | ABX26141.1 | AAD29411.1 | CAB42887.1 |
| ABP35603.1 | P42040.2 | CAA75506.1 | ABX26143.1 | AAF26449.1 | AAX19889.1 |
| ACJ37389.1 | P42059.1 | O97192.1 | ABX26145.1 | AAB38064.1 | P80274.1 |
| ABX57814.1 | P40918.1 | ABN03966.1 | ABX26147.1 | AAS47035.1 | P80273.2 |
| ACF53836.1 | CAD38166.1 | ABN09653.1 | ABX54842.1 | AAS47036.1 | P33556.1 |
| ACF53837.1 | CAD42710.1 | ABN09654.1 | ABX54844.1 | AAS47037.1 | CAR48256.1 |
| AAD13530.2 | P0C0Y5.1 | ABN09655.1 | ABX54849.1 | AAC02632.1 | CAI64400.1 |
| AAD13531.1 | P40108.2 | Q9LEI9.1 | ABX54859.1 | 1H2O | ABD79094.1 |
| P54958.1 | AAX14379.1 | Q9LEJ0.1 | ABX54862.1 | P82534.1 | ABD79095.1 |
| AAA87851.1 | P42038.1 | Q7Y1X1.1 | ABX54864.1 | P82952.1 | ABD79096.1 |
| O18598.3 | CAI05848.1 | ABW34946.1 | ABX54866.1 | 3EHK | ABD79097.1 |
| 1YG9 | ABA39436.1 | 3F55 | ABX54869.1 | AAL91662.1 | ABD79098.1 |
| AAB29344.1 | ABA39438.1 | ACY91851.1 | ABX54876.1 | ACE80972.1 | ABF81661.1 |
| AAB29345.1 | 2A0A | ACZ74626.1 | 2JON | ACE80970.1 | ABF81662.1 |
| AAF72534.1 | AAP35077.1 | P15252.2 | P19963.2 | AAV40850.1 | ABG81312.1 |

Table E.1: (continued)

| | | | | | |
|---|---|---|---|---|---|
| ABH06346.1 | Q967Z0.1 | AAA16792.1 | A38968 | ABB78006.1 | ABG81313.1 |
| ABH06348.1 | Q23939.2 | CAB53458.1 | AAB32652.2 | P81402.1 | ABG81314.1 |
| ABH06344.1 | AAX34048.1 | CAC13961.1 | A53806 | 2B5S | ABG81315.1 |
| Q17284.1 | ABO84963.1 | CAC42881.1 | B53806 | CAD37201.1 | ABG81316.1 |
| ABH06352.1 | ABO84964.1 | AAL25839.1 | C53806 | CAD37202.1 | ABG81317.1 |
| ABH06359.1 | ABO84966.1 | AAA87456.1 | CAA73038.1 | BAH10154.1 | ABG81318.1 |
| ABU97466.1 | ABO84967.1 | AAP87281.1 | CAA73037.1 | AAC24001.1 | 2HCZ |
| 2JMH | ABO84968.1 | O82803.1 | CAA73036.1 | AAD29410.1 | Q1ZYQ8.2 |
| AAK58415.1 | ABO84969.1 | P02877.2 | O24169.1 | AAC13315.1 | P0C1Y5.1 |
| AAM83103.1 | ABU49605.1 | CAA05978.1 | O24170.1 | D53288 | P19656.1 |
| AAD10850.1 | ABU68318.1 | AAF25553.1 | O24171.1 | ABZ81046.1 | 2209273A |
| AAM10779.1 | ABY28115.1 | AAC27724.1 | P80740.2 | ABZ81045.1 | AAB86960.1 |
| AAQ24541.1 | ACK76300.1 | CAA75312.1 | O24172.1 | CAC83046.1 | AAO45607.1 |
| AAQ24542.1 | ACK76291.1 | 1G5U | P81430.2 | CAC95152.1 | AAO45608.1 |
| AAP35071.1 | ACK76292.1 | Q9STB6.1 | AAF31152.1 | CAC83047.1 | AAK56124.1 |
| ABH06350.1 | ACK76296.1 | Q9M7N0.1 | Q9M7R0.1 | CAC95153.1 | AAX40948.1 |

Table E.2: Uniprot IDs for Allergens Training Set - Allerhunter vs Uniprot Database

| ID | ID | ID | ID | ID | ID | ID | ID |
|---|---|---|---|---|---|---|---|
| P81217 | Q10G40 | Q9LEH8 | O04403 | Q0CBV0 | P86755 | Q8J1L4 | P60623 |
| P18153 | Q946J4 | P24337 | P81651 | O59939 | P86758 | Q24702 | Q94424 |
| P92918 | Q6H677 | P80514 | P82534 | Q00645 | P86763 | Q9M0C2 | Q39547 |
| Q7M1G9 | Q7XT40 | P78983 | P81402 | P04959 | P86766 | Q9ZV52 | P81531 |
| P82258 | Q0DZ85 | P40918 | O04004 | P11073 | P86770 | Q9ZP41 | P86472 |
| P79017 | Q7X6J9 | Q92260 | P0C088 | P18209 | P86772 | Q9LEJ0 | Q6TPK4 |
| Q28133 | Q5W6Z9 | P40292 | P86838 | P04960 | P86776 | Q9LEI9 | Q8WNR9 |
| O18874 | Q9LZT4 | P49370 | P84160 | P0C1A4 | P86777 | Q9HDT3 | O04298 |
| Q00855 | Q10S70 | Q5D7H4 | P84161 | P0C1A5 | P86746 | Q96X30 | P49275 |
| P49278 | Q9SVE5 | Q08169 | P19656 | P16530 | P86754 | P42040 | P49276 |
| Q9TZZ2 | Q7XCL0 | P49371 | Q9M5X7 | P39116 | P86760 | Q96X46 | P39675 |
| P80384 | Q9LZT5 | Q9U6V9 | Q9M5X8 | P40973 | P86767 | Q870B9 | P49277 |
| P82952 | Q8H274 | P86687 | Q9M5X6 | Q59671 | P02622 | O14638 | Q8J1M3 |
| Q965E2 | Q5Z980 | P0CH89 | Q8NJ52 | Q51915 | Q90YK9 | P83340 | Q9UW98 |
| O02380 | O23547 | P86875 | O49813 | P72242 | P02620 | B2D0J5 | P12319 |
| O01949 | Q850K7 | Q8NIN9 | P01012 | Q60140 | P59747 | Q7Y1X1 | P20489 |
| P82259 | Q9LDR9 | P16968 | Q06478 | P40972 | P33050 | O97370 | P12371 |
| O60024 | Q7XUD0 | P01085 | Q6Q252 | Q56806 | Q65DC2 | Q1ZYQ8 | Q01362 |
| Q5VFH6 | Q9LNU3 | P13691 | P53357 | O24554 | Q8GCB2 | Q8H7T4 | P20490 |
| P80741 | Q4PNY1 | P01083 | Q6Q251 | O43099 | B1B6T1 | P0C1Y5 | P13386 |
| P82947 | Q9LDJ3 | P10846 | Q6Q250 | P14292 | Q9WYR4 | P04721 | P30438 |
| P81943 | Q7G6Z2 | P81367 | Q6Q249 | P14293 | B1L969 | P04722 | P30440 |
| O96870 | Q9M9P0 | P81368 | Q9U6W0 | P83181 | Q10464 | P04723 | Q92949 |
| P14004 | Q4PR52 | P01084 | P83542 | P69196 | O65200 | P04724 | Q5ULZ4 |
| Q39967 | Q9FMA0 | P28041 | A2VBC4 | P69197 | P85126 | P04725 | Q8J1X7 |
| Q9U5P2 | Q4PR51 | P32936 | Q68KK0 | Q8VWY6 | Q01881 | P04726 | P81010 |
| P83563 | O80622 | P16850 | P0CH87 | P69198 | Q01883 | P04727 | Q17040 |
| O24172 | Q4PR50 | P16851 | Q3ZU95 | P69199 | Q01882 | P04728 | P02863 |
| O42799 | Q9M2S9 | P17314 | P51528 | O64943 | O49065 | P18573 | P38950 |
| Q26456 | Q69XV9 | P34951 | P0CH47 | Q8VWY7 | P15252 | P48060 | P38948 |
| P49273 | Q9ZSI1 | P11643 | P0CH86 | Q84V36 | Q8NKF4 | Q9CWG1 | Q41183 |
| Q9U1G2 | Q4PR49 | P01086 | P49369 | O81092 | P49148 | P84159 | P93124 |
| P81430 | Q9LQ07 | P01087 | Q9BMK4 | P58171 | P50344 | Q32LB5 | P82946 |
| Q9M7R0 | Q4PR48 | P16969 | Q7M4I5 | O81701 | P42037 | Q6UWM5 | P43216 |
| P43187 | Q7G6Z5 | P16159 | P00630 | Q39419 | Q9UUZ6 | Q9CQ35 | P81294 |
| P85063 | Q9SZM1 | P16852 | P82971 | P94092 | P42039 | O18598 | Q9LLT1 |
| P79085 | Q10RK1 | Q5EF78 | Q7M4I6 | O82040 | Q8TFM9 | P46419 | P81826 |
| Q5EZC5 | Q9FL81 | Q5BLY4 | P00784 | P85444 | P67875 | Q6R4B4 | P14946 |
| Q5EZ82 | Q10KN4 | P83207 | Q00002 | P35792 | Q28050 | P43317 | P14947 |
| O94095 | Q9FL80 | P35225 | O34819 | P35793 | P26987 | P02877 | P14948 |
| P42058 | Q4PR44 | P01005 | P94449 | Q08697 | P18632 | P43213 | P83466 |
| P49274 | Q9FL79 | P93105 | A1CFS2 | P08299 | O43290 | P43214 | Q41260 |
| Q7Z1K3 | Q4PR43 | P01088 | B0XT36 | P09042 | P80740 | P43215 | P24396 |
| P83885 | Q9FL76 | P02538 | B8N6W5 | P83834 | P86254 | P10414 | Q9LFP5 |
| P16348 | Q4PR42 | P43393 | Q4WV10 | Q05968 | Q92450 | P0C0Y4 | Q940Q1 |
| P59704 | Q9FL77 | Q9D2H2 | A2R3I1 | Q40374 | Q8MQS8 | P0C0Y5 | P0C1C0 |
| P50635 | Q4PR41 | C7E3T4 | Q01172 | P11670 | P0CH88 | P02762 | Q6CZT4 |
| Q5GQ85 | Q9FL78 | Q95PM9 | Q2TXM4 | Q9U6R6 | Q7M4I3 | P83507 | Q93WF1 |
| P54958 | Q2QP13 | P85261 | Q0CYL8 | Q00008 | Q7Z269 | P02761 | Q9FM66 |
| Q16937 | Q7XE35 | P84527 | Q00374 | Q64LH1 | P30941 | P86703 | Q93Z25 |

# E. LIST OF ALLERGEN AND NON-ALLERGEN PROTEIN SEQUENCES

Table E.2: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P85446 | Q4PR40 | P02754 | Q5BAU9 | Q42449 | O82803 | Q07932 | O65388 |
| P85445 | Q4PR39 | P00711 | Q12639 | Q8H2C9 | P0C8D4 | Q26464 | P0C1C1 |
| Q09095 | Q8W2X8 | P13447 | A1CYC2 | P84177 | P0C8D6 | Q9NJA9 | Q6CZT3 |
| Q09114 | Q75I75 | P82615 | B0YCL3 | O65812 | P0C8D5 | Q8MUF6 | Q9M9S2 |
| Q09097 | Q6YYW5 | Q948T6 | Q4WIS6 | Q9XF40 | P0C8D7 | Q967Z0 | P0C1C2 |
| P83508 | P0C1Y4 | O82015 | A2QFN7 | O24169 | P0C8D8 | Q13765 | Q6CZT2 |
| Q9BQE9 | Q9C554 | Q7M1X5 | Q00205 | Q9XG85 | Q9UW97 | Q67A25 | P86739 |
| P85983 | Q7XWU8 | P85894 | Q2TWM1 | P35079 | Q8J077 | P43217 | P86744 |
| P15494 | Q38866 | P00698 | Q5BA61 | O65809 | Q8WQ47 | O04404 | P86745 |
| P45431 | Q40636 | P39674 | A2QW65 | P49232 | Q9UW02 | Q40905 | P86749 |
| P43176 | O80932 | P39673 | Q5MBK3 | Q64LH2 | Q8TFM8 | P0C1C3 | P86756 |
| P43177 | Q40637 | P43211 | Q2UCT7 | Q8H2C8 | P83958 | Q9C8G4 | P86761 |
| P43178 | O48818 | Q40280 | Q5BEB9 | Q9STB6 | P81370 | P15721 | P86764 |
| P43179 | A2Y5R6 | Q01940 | B8NVB7 | Q9XF41 | P50694 | P15722 | P86768 |
| P43180 | Q0DHB7 | P56577 | A2RBL2 | O24170 | Q9FSG7 | Q9FXD8 | P86773 |
| P43183 | Q38864 | P56578 | P22864 | Q9T0M8 | P85814 | O64510 | P86774 |
| P43184 | Q6ZGU9 | P68407 | Q2TXS4 | O24650 | Q9NAS5 | Q9SRH4 | P86778 |
| P43185 | Q38865 | Q8LW54 | Q5AQJ1 | Q93YG7 | Q9N2R3 | Q9M8Z8 | P86431 |
| P43186 | Q9M4X7 | P01502 | P24112 | O65810 | O96764 | Q9LRM5 | Q91482 |
| P85412 | Q9LN94 | P01504 | P27027 | P49233 | A2V735 | B0XT32 | Q90YK8 |
| P54962 | Q852A1 | P01501 | B3GQR3 | Q64LH0 | Q23939 | B8NE46 | Q05109 |
| O23791 | O22874 | P59261 | Q0CFF7 | Q9M7N0 | O18416 | Q4WIT0 | P81656 |
| P02662 | Q9XHX0 | P59262 | Q0CZD4 | Q9XF42 | Q9GZ71 | A2QV36 | P35759 |
| P92919 | Q9LZ99 | P68408 | B0Y0L8 | O24171 | O97192 | Q9C2Z0 | P35780 |
| P08819 | Q4PR53 | P68409 | B8N316 | O24282 | O44119 | Q2U8R6 | P83377 |
| P80325 | Q9SKU2 | Q9BPX6 | A2R6A1 | P49234 | Q9NFZ4 | P0C1A3 | P86686 |
| Q9JKC0 | P58738 | P86158 | Q8NJK6 | Q9M7M9 | Q25456 | Q5AVN4 | Q7Z156 |
| P81729 | Q40638 | P56924 | Q2UJA7 | Q9M7M8 | Q9GZ69 | P0C1A2 | B2MVK7 |
| Q8LGR0 | Q9SHY6 | P27759 | Q5B3J8 | Q9LEI8 | O61379 | A1CYB8 | P85840 |
| P86473 | O24230 | P27760 | A1DEH0 | Q94JN2 | A1KYZ2 | B8NBC2 | P35783 |
| P42059 | Q9M0I2 | P27761 | P16311 | Q9XF37 | Q9UB83 | P86741 | P35784 |
| Q6PSU2 | Q336T5 | P28744 | P16312 | Q9SQI9 | Q9GZ70 | P86743 | P81657 |
| Q647G9 | Q9SHD1 | P56164 | P08176 | P25816 | Q7M3Y8 | P86747 | P35760 |
| Q9UW00 | Q94LR4 | P56165 | P25780 | P85984 | P02789 | P86750 | P86870 |
| Q8J0P4 | Q9M203 | P56166 | Q1EIQ3 | Q93YI9 | O64432 | P86752 | Q05110 |
| P54107 | Q7XT39 | P56167 | Q03211 | Q84V37 | P45669 | P86757 | Q2L6Z1 |
| Q9XSD3 | Q7XCA7 | P43174 | Q7M1E7 | Q5EF31 | P85413 | P86759 | P85860 |
| Q25641 | Q9LD07 | Q40240 | P43212 | Q5FX67 | P35775 | P86762 | C9WMM5 |
| P83908 | Q10T32 | Q40962 | Q9FY19 | O04725 | P35776 | P86765 | B1A4F7 |
| Q2XXR2 | Q07154 | P43175 | Q6H9K0 | Q8SAE6 | A5X2H7 | P86769 | B2D0J4 |
| Q2XXR1 | Q7XCG7 | Q40237 | Q8H6L7 | P0C0Y3 | P35778 | P86771 | Q9LDH0 |
| Q2XXQ8 | P85293 | Q40963 | Q56S59 | O81982 | P35779 | P86775 | Q6H676 |
| Q2XXP2 | Q55G31 | P22284 | Q8GT41 | Q941H7 | Q9NH75 | P86779 | Q9FSI9 |
| Q2XXP1 | Q54PA4 | P22285 | P82242 | O49894 | P35777 | P86432 | P55958 |
| Q8JI40 | Q7KWS2 | P22286 | Q9LJ42 | Q94JN3 | P35781 | Q91483 | Q2TZY0 |
| Q7ZTA0 | Q86AV4 | P27762 | Q9LTZ0 | Q9XF39 | P10736 | Q90YK7 | P86753 |
| Q7ZT99 | Q55G32 | P00304 | Q9SCP2 | Q8GSL5 | D4P2Y4 | P86740 | |
| Q2XXQ0 | Q54J35 | P02878 | Q93Z04 | Q8GT39 | P35782 | P86742 | |
| Q09GJ9 | Q54C80 | P38949 | Q9SVQ6 | Q9XF38 | P10737 | P86748 | |
| B0VXV6 | Q54C78 | Q96385 | Q944R1 | P81295 | Q05108 | P86751 | |
| Q8JI39 | O76821 | Q08407 | O65456 | Q9LD79 | A9QQ26 | P35785 | |
| Q7ZZN9 | Q17284 | Q9SCG9 | O65457 | O50001 | A9YME1 | P35786 | |
| P79845 | Q9U5P1 | O04701 | Q9C5M8 | O24248 | C0ITL3 | P35787 | |

Table E.3: Uniprot IDs for Non-Allergens Training Set - Allerhunter vs Allergenonline Database

| ID | ID | ID | ID | ID | ID | | |
|---|---|---|---|---|---|---|---|
| Q03965 | P07369 | P42780 | P36400 | P58724 | P58724 | O76242 | P0A212 |
| P06026 | Q32AF9 | P08712 | P02030 | Q8NZJ2 | Q8NZJ2 | Q9CIT0 | Q9X5A6 |
| Q9SSK7 | P02185 | P04462 | Q91507 | Q5LRY6 | Q5LRY6 | Q9JZ53 | O88339 |
| P31110 | P02688 | Q08713 | Q5KVE7 | Q8TI34 | Q8TI34 | Q8K4D8 | P16866 |
| P02561 | O76014 | P84787 | P39825 | Q725C4 | Q725C4 | P56682 | Q9ZC70 |
| O62654 | P29528 | P48492 | Q52701 | P29720 | P29720 | P91919 | P13390 |
| P59120 | Q91055 | P68388 | Q97BX7 | P60876 | P60876 | Q9JL96 | Q819Y8 |
| P55949 | Q9XW16 | P09205 | P43080 | P00617 | P00617 | Q9K4V3 | P00217 |
| Q9UBL6 | P17576 | P83597 | Q5G270 | Q28300 | Q28300 | Q8XEJ1 | O25475 |
| P11376 | P93508 | P15241 | P50700 | Q9PRA4 | Q9PRA4 | Q8BIK6 | Q5X105 |
| O77691 | P02199 | O48897 | P06753 | Q8ZTY0 | Q8ZTY0 | O75094 | Q38802 |
| Q99P61 | O33665 | P12532 | P33185 | P19923 | P19923 | O47669 | Q98C27 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P24534 | P09989 | P08481 | P29111 | Q7U8F2 | Q7U8F2 | P02976 | P05411 |
| O73727 | Q9XY07 | Q9FE63 | O96507 | P53128 | P53128 | P52726 | Q01449 |
| Q5XD48 | Q8NBS9 | P38647 | P84755 | Q8YVH1 | Q8YVH1 | Q29490 | Q8VMY3 |
| Q8LFM0 | Q9R013 | P16293 | Q12891 | P15804 | P15804 | P15425 | Q8IUC8 |
| Q5PFK7 | P16347 | O43940 | Q9NR61 | Q6MEM7 | Q6MEM7 | P35037 | Q5R941 |
| P08759 | P41041 | P48047 | P11032 | Q8FSA1 | Q8FSA1 | O15990 | P60310 |
| P25296 | Q09566 | P21760 | P54223 | Q65LX4 | Q65LX4 | P37714 | Q6FST5 |
| P27495 | Q7VZT5 | Q17192 | Q9DGJ2 | Q52480 | Q52480 | P17475 | Q92IL9 |
| O02654 | O43290 | P02639 | Q03197 | Q5K915 | Q5K915 | O54760 | Q04875 |
| P60662 | P10872 | Q5RFM9 | P33718 | O26024 | O26024 | P07207 | P30191 |
| Q9UXS6 | Q5GTE5 | P27447 | P58795 | Q04971 | Q04971 | P50477 | Q9XCA6 |
| Q4WMB6 | O58389 | P69756 | P04688 | P0C113 | P0C113 | P58407 | O95954 |
| P48820 | Q9W1V6 | O70559 | P14084 | Q94986 | Q94986 | Q8HXQ4 | Q50264 |
| Q7MF13 | P47768 | P70564 | Q9XJ54 | Q8Y2D9 | Q8Y2D9 | O43187 | O19905 |
| Q9AVB0 | Q8NJR0 | Q9UXA8 | P24548 | Q34806 | Q34806 | P20613 | Q752I1 |
| P61917 | P11153 | P48491 | O17473 | P14826 | P14826 | Q8WWM9 | Q58867 |
| Q8IUG1 | Q9ZRF1 | Q9Z6J6 | P13277 | Q7SIB3 | Q7SIB3 | Q8XKU4 | P22342 |
| Q8NFD2 | Q94504 | Q5B4E7 | Q6CAB5 | P0AG45 | P0AG45 | P05123 | Q07430 |
| P14126 | O70220 | P98020 | Q28153 | O67637 | O67637 | Q27384 | Q9JLI8 |
| Q9SKQ0 | P23412 | P27523 | P20305 | P0C0K0 | P0C0K0 | P41981 | Q9WVM1 |
| Q8FXX2 | P27693 | O23320 | Q11101 | P27344 | P27344 | O54758 | O84799 |
| Q9C7I7 | P02627 | Q3YSL9 | P02194 | P44044 | P44044 | P46728 | Q10232 |
| Q8HZ58 | P56533 | P09737 | P17992 | P43819 | P43819 | P41044 | Q71WF2 |
| Q9R0R1 | Q27877 | Q9HGY9 | P39872 | Q9KQX6 | Q9KQX6 | P35418 | Q08012 |
| P24541 | P50346 | P97435 | P07197 | P0AFK3 | P0AFK3 | P29545 | P15556 |
| P27489 | P05595 | P35478 | Q28668 | P17632 | P17632 | P00710 | P0A9I9 |
| Q7N5I5 | Q6FSD5 | Q8VHK8 | P05088 | Q9HNE2 | Q9HNE2 | O44010 | P30204 |
| P01321 | Q29135 | P05607 | P83052 | Q9Z0S8 | Q9Z0S8 | P05693 | O30333 |
| Q94518 | Q40392 | P35616 | P02629 | Q902F9 | Q902F9 | Q8AVA3 | Q5XI53 |
| Q12667 | P69750 | P10592 | Q46WD4 | Q7VLR8 | Q7VLR8 | P50684 | P13397 |
| Q10992 | P00741 | P05318 | P23240 | Q667K0 | Q667K0 | P23604 | P65625 |
| P09477 | O18750 | Q9JLZ1 | P17879 | Q9NPP4 | Q9NPP4 | P47647 | Q9KVE0 |
| P31001 | O84332 | P11140 | P47037 | Q6GJX7 | Q6GJX7 | Q9IB50 | Q7UTS2 |
| O97763 | P38566 | P29602 | P80894 | P07866 | P07866 | Q7RTY8 | P11471 |
| O81918 | Q93WU7 | P04163 | P64409 | O30391 | O30391 | P00925 | P21448 |
| P44758 | P49101 | Q6YNR6 | P93046 | P0AF69 | P0AF69 | P02631 | P54438 |
| P07632 | P68373 | Q7N835 | P19639 | Q8CTR0 | Q8CTR0 | P50338 | P01919 |
| Q83Q04 | Q9WUW3 | Q5RF01 | Q01644 | P44440 | P44440 | Q9Y778 | Q7V1N1 |
| P67888 | Q05876 | P02113 | P00523 | Q8FF19 | Q8FF19 | P84339 | Q759T1 |
| P35591 | O54266 | Q23758 | Q800A0 | O83934 | O83934 | Q67SV9 | Q63313 |
| Q7LZE4 | Q7MTV8 | Q9DB20 | Q9K715 | P0A556 | P0A556 | O16305 | P43600 |
| P38113 | Q9U615 | P01343 | P56116 | P0A6R7 | P0A6R7 | P15242 | P66096 |
| P11117 | Q8CVI9 | Q4QP81 | Q01177 | P53877 | P53877 | O97788 | P46666 |
| Q4IPH4 | Q4L6T0 | Q5XLE4 | Q8CG14 | P09380 | P09380 | Q09607 | Q8FL40 |
| O22347 | P24396 | P29599 | Q8JIU7 | Q6TC33 | Q6TC33 | P39035 | P32785 |
| Q9BYR5 | P45797 | P33791 | P81695 | P57112 | P57112 | P91791 | O15239 |
| Q42972 | Q8YFY2 | Q42842 | Q9WVG5 | Q99XL0 | Q99XL0 | P34107 | P26978 |
| P05599 | P42820 | P23695 | Q95M30 | Q9MAS5 | Q9MAS5 | P98159 | P63876 |
| P68987 | P04724 | P05413 | P79180 | Q9D174 | Q9D174 | P81368 | Q14656 |
| Q5HJK3 | P29871 | P50919 | Q6P7Q4 | Q51429 | Q51429 | Q8ZN80 | P70186 |
| Q9ZA11 | Q8NUR2 | Q9H2H8 | P17750 | Q9P9E9 | Q9P9E9 | P07887 | Q6NC50 |
| P00774 | Q9GKK3 | P04745 | P36204 | Q9LK94 | Q9LK94 | P14831 | Q9CNX9 |
| Q04948 | Q65G89 | Q61316 | Q09023 | P37979 | P37979 | Q06806 | P98000 |
| Q17334 | P83336 | Q8X187 | Q40114 | Q02879 | Q02879 | P04464 | Q9Y5G6 |
| Q7VGK6 | P21195 | P27524 | P48451 | O67270 | O67270 | Q14814 | P54981 |
| P33008 | P05946 | Q91XA9 | O82530 | Q6EW16 | Q6EW16 | Q39591 | Q924C3 |
| Q99MG9 | P40973 | Q92968 | P12549 | Q9D8V7 | Q9D8V7 | P81188 | Q8K9G7 |
| Q6AXR4 | P02539 | Q42611 | P35661 | Q5U9D2 | Q5U9D2 | O70557 | Q87A28 |
| Q9ZD15 | Q9D7P9 | Q9L7P1 | Q9SCX3 | Q2G8X1 | Q2G8X1 | P12794 | P43661 |
| P32957 | Q10994 | Q8Z3E0 | Q6T3B0 | P0AAZ2 | P0AAZ2 | Q9GLP2 | P60330 |
| P61443 | Q37649 | Q29144 | P08709 | Q9SS44 | Q9SS44 | P30883 | P73910 |
| Q4UJV5 | Q9FID0 | P04106 | P05410 | Q8P5S6 | Q8P5S6 | P92998 | O04104 |
| Q8CQ84 | Q2UGK2 | P00771 | O13309 | Q03655 | Q03655 | P46409 | Q72KF7 |
| Q9ASS6 | P02144 | Q47VL0 | Q9RUV2 | Q72IV9 | Q72IV9 | P97820 | P65377 |
| P83368 | P59888 | Q4FNT6 | P82460 | Q8BH66 | Q8BH66 | Q64425 | Q4ZPN9 |
| Q9ULU8 | P69377 | P10975 | Q9TTN0 | Q817W3 | Q817W3 | P68379 | Q864U6 |
| P36375 | Q91W90 | P08480 | P68190 | P43982 | P43982 | Q7W519 | P83366 |
| Q9D6L8 | Q29290 | Q05816 | P15231 | Q867B5 | Q867B5 | Q5I2M5 | Q70Y16 |
| O08460 | Q9XZP2 | Q95VF7 | O49073 | Q6BMK7 | Q6BMK7 | Q87WD5 | O66132 |
| P68243 | P15331 | Q9LDA4 | P37235 | P0ABL0 | P0ABL0 | Q3MSM4 | P10805 |
| P09794 | P81399 | P50262 | P13919 | Q88L76 | Q88L76 | Q8UH55 | P0AFK4 |
| P43377 | P04365 | Q14767 | P04971 | Q49408 | Q49408 | P02201 | P73950 |
| P56221 | Q74J64 | P61632 | P01087 | Q9VAZ3 | Q9VAZ3 | Q45551 | Q8KAY6 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Q8P855 | P20753 | Q75C78 | Q96A23 | P32567 | P32567 | P36186 | Q8LVH2 |
| P28814 | Q96L50 | P71016 | Q96WL3 | Q89B17 | Q89B17 | Q20222 | Q8D341 |
| O14497 | Q925H1 | P37392 | P09860 | Q9C8N7 | Q9C8N7 | P28517 | Q5PGK0 |
| P37109 | Q759R7 | Q9Y493 | Q6AY56 | Q8RFN0 | Q8RFN0 | Q41001 | P26170 |
| Q9BEB0 | P02640 | P23346 | P56271 | P02846 | P02846 | Q7T6X8 | Q98Q77 |
| Q64702 | P05544 | P68393 | P35211 | Q9KUP7 | Q9KUP7 | P02223 | P0A0X7 |
| Q9ZDC6 | P50698 | P14376 | P53650 | O84460 | O84460 | Q9LSY7 | P33808 |
| Q08279 | Q8HXP7 | P62345 | P05580 | Q9N0T2 | Q9N0T2 | Q9F3P9 | Q9NRE2 |
| P22738 | Q9U6D3 | Q12798 | Q75E44 | P64516 | P64516 | P26740 | P96022 |
| P19269 | P0C1B4 | O64393 | P04691 | P29953 | P29953 | P80463 | Q9TU45 |
| Q9V1T5 | Q5R5A3 | Q8H7U1 | Q28810 | P16798 | P16798 | P81105 | P0AAT3 |
| Q03682 | Q38798 | Q6BWC0 | Q96506 | P75060 | P75060 | Q24214 | O10350 |
| P13851 | P80168 | Q29137 | P11841 | Q7UYK5 | Q7UYK5 | Q6CW24 | Q7URP3 |
| O24356 | P31161 | P24894 | Q9XI01 | P33198 | P33198 | P32419 | P60444 |
| P49863 | Q9YH85 | P32261 | O03893 | O24312 | O24312 | P22431 | Q7JEW0 |
| P15797 | Q9SN86 | P02554 | P80229 | Q8FJ51 | Q8FJ51 | Q43645 | Q9SV84 |
| Q05524 | Q32J53 | Q9NRR2 | P27677 | Q62EU9 | Q62EU9 | P21619 | O45291 |
| P82727 | P06873 | Q92EU4 | P41387 | Q9KNK0 | Q9KNK0 | Q68Y22 | Q3K9J0 |
| P02829 | Q8K6Y8 | Q9FEG8 | Q8T0W0 | Q9K096 | Q9K096 | P16349 | Q8FAJ3 |
| P22774 | P12844 | P20136 | P17715 | Q75W64 | Q75W64 | Q5HVP5 | P13411 |
| Q9D7Q0 | P12020 | Q38910 | P24452 | P68405 | P68405 | O65351 | P95242 |
| P51649 | P10791 | Q03045 | P43376 | Q8XZH5 | Q8XZH5 | P04780 | O83209 |
| Q9NPB3 | Q8Y0Q3 | P12460 | Q42976 | Q50HP3 | Q50HP3 | Q8JI39 | P38841 |
| Q9Y337 | P20801 | Q96251 | P51650 | Q7VDV0 | Q7VDV0 | Q8MSS1 | P76214 |
| P55059 | O67149 | P12068 | Q39649 | Q8H185 | Q8H185 | Q9H3S3 | P46058 |
| Q9FE20 | Q07078 | Q92429 | P20152 | P0AG00 | P0AG00 | Q9QXZ0 | P60219 |
| P60163 | Q6FWL5 | Q8UW11 | Q9PGJ7 | Q22516 | Q22516 | P80429 | Q15012 |
| P13566 | P09644 | P07518 | P02029 | P80109 | P80109 | P33625 | Q979F2 |
| Q10265 | P48285 | P05386 | P12062 | P57627 | P57627 | P00701 | O13620 |
| P29844 | Q17040 | P09493 | P07481 | Q31XF7 | Q31XF7 | Q9XF89 | Q50028 |
| P82735 | Q9X6W9 | Q94F62 | Q9U639 | P82372 | P82372 | P15737 | P10905 |
| P06702 | P30407 | Q92079 | P52576 | Q8TYL6 | Q8TYL6 | O93826 | Q5WTI8 |
| O48737 | Q8I1F6 | Q3UP87 | Q9YGQ1 | P53143 | P53143 | P08486 | Q8TGX1 |
| O35648 | P12658 | Q9MIY7 | P22686 | P19119 | P19119 | P93171 | P35972 |
| Q5G267 | P29875 | P04109 | P05587 | P54570 | P54570 | Q9BXB1 | P36557 |
| Q9GL24 | P42666 | Q08697 | Q6MT06 | P64998 | P64998 | P07895 | P50914 |
| P02165 | Q5Z3N4 | P22325 | Q93VR4 | P65948 | P65948 | P49614 | P63418 |
| P04372 | P04813 | P24707 | P52242 | P16043 | P16043 | Q9QXV8 | P14156 |
| P42329 | P35417 | Q14563 | Q5R537 | P71575 | P71575 | P68384 | P40357 |
| P00746 | Q3Z6P1 | Q7VRL0 | Q8BWF0 | P26281 | P26281 | P02548 | Q5R9J5 |
| P81286 | Q6BIB1 | Q87EN0 | Q741U7 | Q8PNS8 | Q8PNS8 | Q8Y0B5 | Q9YGY0 |
| Q86RQ7 | P16474 | P15312 | Q10657 | P60027 | P60027 | Q29485 | P30704 |
| Q8RDX7 | P0C0Q6 | P98024 | Q9T076 | O84120 | O84120 | O65493 | Q5F5P7 |
| P00403 | Q7LZP9 | Q9FZ27 | O93429 | Q5K3U4 | Q5K3U4 | P43232 | P0AG96 |
| P40121 | Q9RJH9 | P50679 | Q9TWF8 | Q7N6I5 | Q7N6I5 | Q9BMQ6 | P65428 |
| Q9NFL4 | Q9QZH3 | Q6NWF6 | P12839 | Q967G1 | Q967G1 | Q5XFN2 | P55179 |
| P18655 | P50261 | Q9DEP0 | P06671 | Q41373 | Q41373 | P25860 | Q83A78 |
| Q42641 | Q53NL5 | P12306 | Q9M5J9 | Q5X5C5 | Q5X5C5 | P14715 | Q8KLU8 |
| P77972 | P32623 | Q6M075 | P09643 | P56927 | P56927 | Q9XHP0 | Q894G7 |
| P25155 | P68945 | O73763 | Q06076 | P70091 | P70091 | P26857 | Q5HFB6 |
| P42926 | O81919 | Q41783 | P20365 | P35108 | P35108 | P58521 | P0A793 |
| P48610 | P10669 | P52409 | Q59623 | P82838 | P82838 | P79845 | P55982 |
| Q8KCH7 | Q74ZJ0 | P08631 | P17538 | Q9KC86 | Q9KC86 | O47671 | Q86119 |
| O49293 | Q95227 | P46427 | Q9W5U2 | Q5FTZ0 | Q5FTZ0 | Q6IE32 | O74878 |
| P10733 | Q98PL5 | P31416 | P50023 | Q5XDL8 | Q5XDL8 | O88520 | Q88WK9 |
| Q8MKG2 | P02544 | Q3SB13 | Q7VIE3 | P80815 | P80815 | Q9CWU6 | Q45219 |
| Q589G4 | P36580 | O81755 | Q39752 | P69383 | P69383 | P15007 | P00020 |
| P36184 | Q95V58 | P29763 | P28590 | Q5PA83 | Q5PA83 | O81772 | Q4A5X2 |
| Q6AW21 | Q7M4G0 | O73860 | P29445 | Q9Y224 | Q9Y224 | P30270 | O31554 |
| Q5X866 | P41973 | P26011 | O00548 | Q9ZB72 | Q9ZB72 | Q9BY71 | Q8Y634 |
| Q39709 | Q87WP0 | P48890 | P43236 | Q8DGS4 | Q8DGS4 | P15290 | P36501 |
| P98048 | P39450 | Q3MSU9 | Q39817 | Q9SX55 | Q9SX55 | Q9TTC6 | O43761 |
| P58502 | Q9JW31 | Q8HXQ0 | P49060 | O83437 | O83437 | P41114 | Q9T1S6 |
| P04105 | P00332 | P11833 | P55063 | P56440 | P56440 | P42849 | P10826 |
| P50453 | P21807 | Q8SPE8 | Q9CN86 | P31881 | P31881 | P05579 | Q83EL6 |
| O73888 | Q6AC76 | P20472 | Q08331 | Q6PBT9 | Q6PBT9 | P47670 | P40423 |
| Q9ULH4 | Q8TD31 | Q9H4F8 | Q5P1H5 | P0ADA9 | P0ADA9 | P21820 | O50630 |
| P17661 | Q9FE01 | P96985 | P61944 | Q92733 | Q92733 | P13190 | Q4L5P4 |
| P05154 | P07216 | P22803 | Q8N1G4 | Q8XVW4 | Q8XVW4 | P06706 | Q00239 |
| Q59679 | P19805 | P0C0I6 | Q4WLG9 | Q9X278 | Q9X278 | P08853 | Q9USK0 |
| P26727 | P52263 | Q52828 | P00775 | P0A3M2 | P0A3M2 | P0A390 | Q5JDM6 |
| P35636 | Q92452 | P10597 | P80284 | P53339 | P53339 | Q3K5Y1 | P27550 |
| Q68F79 | P01335 | P41294 | Q63088 | Q00390 | Q00390 | Q9PVK2 | Q97ES6 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P53375 | P82205 | P49258 | P11034 | Q8TZJ9 | Q8TZJ9 | Q96UQ7 | P34246 |
| P02563 | Q9LKI8 | P02636 | Q7W0M8 | Q9RN02 | Q9RN02 | Q10989 | Q9P3W4 |
| Q4P7H2 | Q10100 | P29525 | Q91ZJ9 | O03522 | O03522 | Q9ESD1 | Q9UKN8 |
| P30893 | P09229 | Q08114 | Q659U1 | P83468 | P83468 | Q815K8 | Q8WXA9 |
| P00748 | P27656 | Q8PJK3 | P04656 | Q5X1A3 | Q5X1A3 | Q41484 | P38348 |
| O14463 | Q8KG25 | Q8PC41 | P07406 | Q51366 | Q51366 | Q05870 | Q9ZM66 |
| Q6KIB0 | Q8TQR2 | P53715 | P58801 | O83066 | O83066 | P54627 | P44894 |
| P68278 | Q5R440 | Q9HFN9 | P20279 | Q2QLD1 | Q2QLD1 | P02696 | Q89933 |
| Q95PP1 | Q975I1 | P04630 | Q8T6A5 | Q8P272 | Q8P272 | P09232 | Q74G62 |
| Q4WPF8 | P61186 | Q00645 | Q05746 | Q5PKT6 | Q5PKT6 | O65457 | Q9UK28 |
| Q3SNS6 | P82353 | Q95225 | P84297 | P53230 | P53230 | Q93WH6 | Q69014 |
| P14140 | P35032 | Q8T938 | Q9GL10 | Q8RC26 | Q8RC26 | P06396 | Q02767 |
| P83365 | P52401 | P54222 | Q9P602 | P61368 | P61368 | P49174 | Q9XDB4 |
| P68409 | P26199 | P28755 | P25777 | P12616 | P12616 | P30811 | Q3C256 |
| O04011 | P93258 | P23547 | O59808 | Q97HD1 | Q97HD1 | Q7WYP6 | P37064 |
| Q5NPS6 | O01725 | Q8P5D7 | O97943 | Q48254 | Q48254 | Q73Q16 | P75597 |
| Q24388 | Q7LZM5 | O04153 | Q5YQ30 | P56175 | P56175 | Q43387 | O59804 |
| Q6D182 | Q62073 | P34113 | Q42443 | Q5BL44 | Q5BL44 | Q06283 | Q8YGT8 |
| Q9Y8H8 | P22618 | P50419 | P07322 | P47508 | P47508 | P35175 | Q7L3T8 |
| P43318 | P30415 | Q9YHV4 | P27161 | P40851 | P40851 | Q7XA39 | Q44145 |
| P02183 | Q865F7 | P17182 | Q6FW50 | Q98IG7 | Q98IG7 | Q9ZVR7 | Q5M7K0 |
| Q98TA8 | P60175 | Q43008 | P04654 | Q9PHT7 | Q9PHT7 | O02640 | P32111 |
| Q03467 | Q03686 | P15194 | Q86WK6 | Q6FUY5 | Q6FUY5 | P09800 | Q97W59 |
| Q00593 | Q05944 | P93535 | P43231 | Q5RE43 | Q5RE43 | Q07449 | Q58898 |
| P13029 | P20231 | P01010 | O66105 | Q8ZBN1 | Q8ZBN1 | O61998 | Q8R216 |
| P07246 | Q6MEY2 | P07758 | Q8X191 | Q3V8B3 | Q3V8B3 | O77011 | P0ACD0 |
| P19943 | P48593 | Q02977 | P83506 | P46748 | P46748 | P33719 | Q9CLM1 |
| P29877 | P37154 | P51460 | P21326 | Q83I17 | Q83I17 | Q8L1Z5 | P39200 |
| Q9UIV8 | P17540 | P50121 | Q8QLK1 | P65897 | P65897 | P06797 | P36520 |
| O23609 | Q8UH56 | Q25632 | P02598 | Q9Y6N7 | Q9Y6N7 | O47496 | P22412 |
| P38234 | P00717 | Q43723 | P02168 | Q96Q80 | Q96Q80 | P34504 | P55853 |
| P42683 | Q8Z1A8 | P16404 | P50452 | Q5UQY1 | Q5UQY1 | O69298 | Q5N2P7 |
| P15590 | P84708 | Q3YUZ7 | Q43731 | P48626 | P48626 | P52016 | Q83HT7 |
| Q5RD69 | P84342 | P10976 | P15309 | P46484 | P46484 | P06198 | Q01349 |
| P25783 | Q9SW93 | P69045 | P09421 | Q6A5S3 | Q6A5S3 | P09220 | P16892 |
| O53021 | P34108 | P08552 | P17514 | P33440 | P33440 | Q27640 | Q57715 |
| P33763 | Q96519 | P93407 | P46487 | P34221 | P34221 | P24704 | Q40194 |
| P84795 | Q8GUQ5 | Q9PHR3 | P98119 | Q8FD47 | Q8FD47 | P49230 | Q57905 |
| P0AAB1 | P84516 | P09093 | Q5YNI0 | Q96DT0 | Q96DT0 | P00783 | P67454 |
| P14812 | P12277 | Q28399 | Q8YRB0 | P22081 | P22081 | P58514 | P08494 |
| Q29146 | P46271 | P32821 | P24664 | P33804 | P33804 | P00773 | P0A8Z4 |
| P50454 | Q9SP02 | P34466 | P97677 | Q3KKY8 | Q3KKY8 | P12845 | Q87EH6 |
| P00714 | P13929 | P22324 | Q64424 | Q97QJ8 | Q97QJ8 | P50688 | P23552 |
| P18322 | O32482 | P57999 | P39462 | Q58726 | Q58726 | Q5RC27 | P41799 |
| P52227 | Q39994 | O08716 | Q5UR27 | Q9NPC8 | Q9NPC8 | Q28988 | P18278 |
| Q9WWW2 | O75093 | P00707 | P83344 | Q88A48 | Q88A48 | Q81JK8 | Q8TV06 |
| P33828 | P26738 | Q39794 | P10562 | Q8W425 | Q8W425 | O00187 | O54983 |
| P28756 | Q9ZPN9 | P87047 | P27797 | P04581 | P04581 | P48779 | Q9RT91 |
| P36221 | P84299 | P98092 | P19753 | P56300 | P56300 | P09642 | O97935 |
| P80174 | P61277 | O60911 | Q8PNT0 | Q8TQ11 | Q8TQ11 | Q6BTB1 | Q8DS88 |
| Q9PTD7 | Q8L5C6 | Q01958 | Q66FQ2 | P57487 | P57487 | Q92563 | Q8ZU97 |
| P37832 | P33522 | Q70JN8 | P61893 | P52320 | P52320 | P27894 | Q8HS39 |
| Q00897 | Q9S744 | Q41160 | P56252 | O20162 | O20162 | Q8ZKP7 | Q42946 |
| O80912 | Q43284 | P36187 | Q61362 | P00654 | P00654 | P35047 | Q87RQ4 |
| Q7RJG2 | O43396 | Q5X5J7 | P09435 | P64841 | P64841 | P12710 | P84281 |
| P11152 | Q9HHB9 | P42328 | P35684 | Q16994 | Q16994 | P06605 | P40366 |
| Q9IA97 | P07683 | Q80TR4 | Q5R2J5 | Q58123 | Q58123 | P31695 | P11134 |
| Q6VAF7 | P30275 | P00765 | P29881 | Q9ZXI0 | Q9ZXI0 | P51818 | Q81JW5 |
| Q9JX17 | Q9P4T6 | Q29614 | Q9G2X2 | Q95LF4 | Q95LF4 | P58773 | P0C0W8 |
| P27322 | Q6FLU4 | P52578 | Q8ZGW0 | Q8YUR5 | Q8YUR5 | P16044 | Q03004 |
| P07951 | P81906 | Q9FMA6 | Q9SPV5 | Q8XID7 | Q8XID7 | P26986 | Q6D025 |
| P56634 | P94391 | P50675 | P78385 | Q7N798 | Q7N798 | P50689 | Q7VHK3 |
| P54985 | O74238 | P20698 | P41099 | P68175 | P68175 | Q02906 | Q9Z2A8 |
| Q6PL31 | Q4L3K3 | Q29371 | P19799 | Q74GT9 | Q74GT9 | P13745 | Q89BU5 |
| P11591 | P09189 | Q07943 | Q6L8G9 | Q5N5I8 | Q5N5I8 | Q63W07 | Q5RJQ0 |
| P21568 | P52588 | P00413 | Q8YW74 | Q59321 | Q59321 | P10790 | Q9P7L5 |
| Q5HNW6 | P30946 | P64074 | Q42684 | Q61609 | Q61609 | P27163 | P27324 |
| Q6L711 | P52249 | P98184 | Q03684 | Q6HDT6 | Q6HDT6 | P62759 | Q87SD2 |
| Q8Z7N9 | P24774 | Q3MSM3 | P10049 | Q9JIL5 | Q9JIL5 | P70059 | Q8NT25 |
| P73728 | Q5PG91 | Q6S4N8 | P58518 | Q47377 | Q47377 | P93000 | Q8EY59 |
| Q02226 | Q9FT36 | Q5E9A1 | Q8MI17 | P37144 | P37144 | P02621 | Q9HKM6 |
| O65252 | P30922 | Q8SPH5 | Q00762 | Q987C8 | Q987C8 | Q37411 | Q9KK72 |
| P50661 | P50446 | P34690 | P50667 | P53677 | P53677 | Q11083 | P77845 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| O15991 | P60331 | P42529 | Q28042 | Q8DC44 | Q8DC44 | P07092 | Q8PWZ4 |
| P42897 | P30187 | P14584 | Q9M8X1 | Q05862 | Q05862 | Q38PR9 | O14358 |
| Q9LDB4 | P05105 | P07897 | P54197 | P44031 | P44031 | Q9XFI1 | Q8A5W4 |
| O80822 | P13538 | Q39697 | P12308 | P0A440 | P0A440 | O97944 | O75841 |
| Q9RJZ6 | P05566 | Q37595 | P12882 | P52357 | P52357 | Q5WLS0 | P45512 |
| P0A6Y8 | Q47XI6 | P18242 | Q8MUC1 | P21482 | P21482 | P02690 | P81488 |
| P02170 | P54318 | Q9VG58 | Q8CN24 | Q8U9K5 | Q8U9K5 | P12615 | Q07093 |
| Q9WUA1 | P12847 | P23565 | Q94738 | P21653 | P21653 | O68901 | P37458 |
| O02705 | P15107 | Q8R0Y6 | P22778 | Q85AW7 | Q85AW7 | P07498 | Q75G34 |
| P54640 | P02855 | Q28986 | P49189 | Q63US8 | Q63US8 | P57870 | Q9UG56 |
| P03949 | P09327 | Q5KKX7 | Q02971 | O42300 | O42300 | P30838 | O26934 |
| P16350 | O86103 | P25809 | Q9YWK4 | Q8S9F1 | Q8S9F1 | P06660 | Q7VDU8 |
| P24044 | P50699 | P0A4L3 | Q9X519 | Q9UHL4 | Q9UHL4 | P98033 | Q62764 |
| Q03013 | P19330 | P48494 | Q37440 | P03285 | P03285 | P80356 | Q8N653 |
| P22531 | P15838 | Q6FDF8 | P24744 | Q74LN3 | Q74LN3 | O88279 | Q46J22 |
| Q9DGI9 | Q9CYL5 | P41131 | Q9ZPN8 | Q7MPD2 | Q7MPD2 | P05582 | P19389 |
| Q28661 | Q9NRD9 | Q27774 | Q55595 | P55505 | P55505 | P08052 | Q9WVR4 |
| P24478 | Q53197 | P26200 | Q59671 | P0A6E5 | P0A6E5 | Q9U4X5 | P06595 |
| P06768 | P21664 | P94040 | Q43560 | Q9APM5 | Q9APM5 | Q13231 | Q73FP5 |
| Q7Z1F8 | P06306 | P92132 | P24526 | P32893 | P32893 | P29117 | P29765 |
| Q61526 | P10936 | P24102 | P53640 | P32237 | P32237 | P02612 | Q6LVA9 |
| Q9CPN6 | P04905 | P19172 | P61628 | P08375 | P08375 | Q9FMA0 | O05434 |
| O74479 | P05585 | O44001 | P19329 | Q8ZH40 | Q8ZH40 | Q9P3X9 | P50853 |
| P08843 | O46894 | P07916 | Q56UD0 | P56989 | P56989 | P98038 | P31994 |
| P93285 | P50058 | P47738 | P33186 | Q57B47 | Q57B47 | P20907 | Q54710 |
| P33224 | P29215 | Q7A1Y7 | P59034 | Q49755 | Q49755 | Q06544 | P63382 |
| Q7Z794 | Q41495 | P25840 | Q7NIR1 | Q5X2Y1 | Q5X2Y1 | P09931 | Q95M71 |
| P84407 | P50253 | Q50639 | P37369 | P41796 | P41796 | Q27965 | Q60AK7 |
| Q8RI55 | P59222 | P14273 | Q4KFX8 | Q9Y8D3 | Q9Y8D3 | P27490 | Q6HDM4 |
| Q65T37 | Q8KEP3 | P41691 | Q01294 | P0A008 | P0A008 | O77013 | O64231 |
| O22373 | Q5U405 | Q7N8Y4 | Q8ENP4 | P61610 | P61610 | P11121 | O13929 |
| O77791 | Q95184 | Q9BQ16 | Q7M4G1 | Q962B2 | Q962B2 | P18966 | Q9TM35 |
| O88280 | P06871 | P02550 | Q56UD4 | O34669 | O34669 | P25815 | Q8TUU3 |
| P29501 | P52231 | Q9UTS0 | P81262 | Q5WYR3 | Q5WYR3 | Q99715 | O00767 |
| Q9AYP9 | P22327 | P09668 | P06868 | Q9PLN5 | Q9PLN5 | P02200 | Q81X04 |
| Q9BZP6 | Q9JI71 | Q8E8J1 | Q4K526 | Q8TCT8 | Q8TCT8 | Q05967 | Q8X9G9 |
| P07290 | P02664 | P12388 | P41208 | Q96N87 | Q96N87 | Q09433 | P19263 |
| Q9W5U3 | P12398 | Q8CD91 | Q7IKU8 | Q821T4 | Q821T4 | P11706 | P61968 |
| Q9P9L3 | P13116 | Q5GWS6 | P20387 | P39933 | P39933 | P09941 | P50411 |
| P18209 | Q43729 | P81714 | P32872 | Q88PK1 | Q88PK1 | P35041 | O14207 |
| P07688 | P22970 | P07591 | Q8SQI3 | P04337 | P04337 | P02588 | P33068 |
| P02567 | O15989 | P18426 | Q9GPU2 | P09729 | P09729 | Q59641 | P45855 |
| P01009 | P08108 | P81166 | P05563 | P35131 | P35131 | Q7LZQ3 | P65734 |
| Q90593 | P05564 | Q9FL92 | Q8ZTE6 | Q02577 | Q02577 | P50425 | Q7VL22 |
| P17989 | Q39239 | Q40401 | P36604 | Q93T12 | Q93T12 | Q6C4W6 | Q8TKT0 |
| P42757 | P04828 | P31863 | Q59519 | Q9KD70 | Q9KD70 | Q8AVA4 | P60755 |
| O28050 | P21674 | P17207 | Q05739 | O12934 | O12934 | P25782 | Q5YPZ5 |
| O76819 | P50683 | P52010 | Q37685 | Q99075 | Q99075 | Q10454 | P95337 |
| P50672 | Q9FL16 | P13213 | Q9NJU9 | Q3J9L7 | Q3J9L7 | Q930L2 | P34957 |
| P12078 | P47955 | Q12560 | P67941 | O62479 | O62479 | P35627 | Q66G80 |
| P02694 | P81702 | P02617 | P30821 | P14116 | P14116 | P35587 | Q29077 |
| Q9S795 | Q9Y2L9 | Q9STD3 | Q95M18 | P18802 | P18802 | P47949 | Q6YR70 |
| P81371 | P16563 | Q9I8V0 | O62650 | Q92100 | Q92100 | Q8MPZ7 | Q9RAE4 |
| P77929 | P06649 | P16895 | Q661T0 | Q5M869 | Q5M869 | Q43766 | P02532 |
| P15721 | Q90023 | P00750 | P54650 | Q3KHL6 | Q3KHL6 | P41311 | Q47745 |
| O93344 | P00991 | O28354 | Q9XSJ4 | P65352 | P65352 | Q9SS98 | Q9Z172 |
| P34697 | P57286 | Q6P9F7 | Q58405 | Q92PW3 | Q92PW3 | P52395 | P0A5N3 |
| P48006 | P45877 | P30157 | Q5G271 | P67546 | P67546 | Q05866 | Q5HKD6 |
| P50543 | P32930 | Q29116 | Q9THX6 | Q12587 | Q12587 | Q65H54 | Q55986 |
| Q62803 | Q07235 | Q37596 | P83180 | P30082 | P30082 | Q01807 | Q04235 |
| Q8HY82 | P33587 | Q8D2Q5 | Q7NZI3 | Q5PLZ1 | Q5PLZ1 | P92693 | Q9KD79 |
| P36368 | Q43129 | P15004 | P33675 | P36337 | P36337 | Q9N4X8 | O08686 |
| Q660W4 | P62936 | P19588 | P02186 | Q6NYB7 | Q6NYB7 | P23729 | P36189 |
| Q5R7B5 | P25349 | Q19948 | Q8CG48 | O88350 | O88350 | Q4R5Q0 | Q9CSB4 |
| Q9LLR7 | P60052 | P50420 | P41715 | P90521 | P90521 | P36214 | Q71YK6 |
| P27492 | P41797 | P24367 | P13667 | Q65MR4 | Q65MR4 | Q94CD8 | Q9HP20 |
| P48794 | P01329 | Q42403 | O96064 | Q89UZ9 | Q89UZ9 | Q9I8F9 | Q87N18 |
| O85348 | Q96R05 | P31657 | O67861 | Q8E5R4 | Q8E5R4 | P52248 | P58995 |
| Q6L8G5 | Q7YQC6 | Q9J8B9 | P84788 | Q8HS40 | Q8HS40 | P48670 | Q5F8V2 |
| O13811 | Q7MA35 | P27357 | Q07437 | Q7V0F0 | Q7V0F0 | P22953 | P47038 |
| P62963 | P08937 | P19092 | Q4P0V4 | Q5HWW3 | Q5HWW3 | Q9SSK5 | Q6YQC6 |
| P01344 | Q5HID3 | P31417 | P42693 | Q9N298 | Q9N298 | P29856 | P45768 |
| P44427 | P24660 | O24313 | P04778 | Q8K903 | Q8K903 | P42278 | Q05437 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Q99170 | Q5CZK3 | P13582 | P13406 | P45777 | P45777 | P12725 | P31273 |
| P41040 | Q8HZK3 | Q9FM66 | P30292 | Q8BWY3 | Q8BWY3 | Q7IC90 | O87391 |
| Q9UWF0 | P61859 | P62504 | P47721 | P56517 | P56517 | Q9P940 | Q08761 |
| Q8LD49 | Q62635 | Q8R4U0 | P52246 | Q60996 | Q60996 | P20147 | P47995 |
| P13020 | O83548 | P24549 | P11120 | Q817U4 | Q817U4 | P83516 | Q8TXZ9 |
| P82900 | Q8CK51 | Q66JT1 | Q9K596 | Q6NRE7 | Q6NRE7 | Q42642 | Q9MUC0 |
| O22263 | P43077 | O67124 | P25842 | P53382 | P53382 | Q27451 | Q5QQ50 |
| P29612 | Q6L8H4 | Q6TMG6 | P81440 | P69641 | P69641 | P11499 | Q47VY2 |
| P14540 | Q875P5 | P75512 | Q8N4C6 | O07025 | O07025 | P17928 | Q9PDV8 |
| P02468 | Q9PEQ0 | Q07068 | P58778 | Q22909 | Q22909 | Q87KQ4 | O19002 |
| Q02862 | P25372 | P11141 | Q8K3Q3 | Q6W8W3 | Q6W8W3 | O76260 | P0C075 |
| P08515 | P11484 | Q96183 | P82787 | P05957 | P05957 | P00742 | P42509 |
| Q7X222 | P26371 | P41209 | P43375 | Q91XB0 | Q91XB0 | Q26516 | P58742 |
| Q8N9N7 | P12243 | P54736 | Q9LVT4 | Q5UZR8 | Q5UZR8 | P00708 | P26907 |
| P17990 | P10058 | P39023 | P25381 | P52596 | P52596 | Q06015 | Q8PCM6 |
| P87066 | P07219 | Q9FKA4 | Q81J48 | O77404 | O77404 | Q9FCC9 | Q9UY16 |
| Q07439 | P26262 | Q923P0 | P82747 | P98086 | P98086 | P00940 | P51737 |
| P55051 | Q9N2D1 | P15437 | Q23092 | P50863 | P50863 | P62941 | Q7V9J7 |
| P14545 | P22623 | P02628 | P83748 | P36249 | P36249 | P98049 | Q9QZN1 |
| Q99P72 | P52581 | Q9SZB9 | P25096 | P47679 | P47679 | P02866 | Q9V0L2 |
| Q9XSD3 | Q05968 | Q41159 | P80049 | Q3M4N7 | Q3M4N7 | P52589 | Q92993 |
| P78695 | P14872 | P17946 | P32646 | Q8VS63 | Q8VS63 | Q4WE62 | Q14152 |
| P0C1A3 | Q9FG34 | Q05820 | P09036 | Q4UKB1 | Q4UKB1 | Q5R536 | P32397 |
| P15461 | P31725 | P50912 | P50703 | O57803 | O57803 | Q9K813 | Q99N08 |
| P25804 | P21148 | Q9CPU0 | P02202 | P64280 | P64280 | P50449 | P40354 |
| Q9LJ42 | Q96C19 | P40313 | P13104 | Q9I8S4 | Q9I8S4 | Q21735 | Q92147 |
| P02623 | P68425 | P29872 | Q55287 | Q7VA23 | Q7VA23 | Q10758 | Q73VI4 |
| Q48E62 | O97680 | P81365 | P46102 | Q8TW08 | Q8TW08 | Q9S772 | Q6FNV1 |
| Q02245 | Q8LGC6 | P93087 | Q8HZJ4 | Q5PCE2 | Q5PCE2 | Q9LHA7 | Q9PNX4 |
| P92662 | P02620 | P31683 | Q23762 | P38156 | P38156 | P15457 | Q9RUG6 |
| O77018 | P53444 | P41751 | Q49V52 | Q39208 | Q39208 | Q9I8X1 | P44231 |
| Q6F0Z7 | Q7NG49 | Q7RTV2 | P05124 | P47813 | P47813 | Q9ZSM8 | Q49V09 |
| P28491 | Q02816 | P08594 | P15223 | Q57AM9 | Q57AM9 | Q8DTS9 | Q72RB6 |
| P17718 | Q9LMP6 | P48384 | P61852 | P09807 | P09807 | P10639 | P16539 |
| Q9BE71 | P14632 | P41721 | P28514 | O93794 | O93794 | Q96YR5 | Q6GBV6 |
| P57492 | P00527 | Q9M0D2 | P05434 | Q9KLX6 | Q9KLX6 | Q86GF7 | Q92PB8 |
| P29508 | Q7VN27 | Q7WC32 | Q26534 | O02228 | O02228 | P20142 | P10544 |
| Q9D869 | P07440 | P26509 | O65198 | P16918 | P16918 | Q85FW4 | P96786 |
| P50702 | Q8BLY1 | P00713 | P51613 | Q4JAP5 | Q4JAP5 | P35030 | Q6YPM2 |
| P02180 | O00845 | Q24439 | P0C0V3 | P46245 | P46245 | Q25563 | P0AB08 |
| P05805 | Q00465 | P11602 | P34887 | O74501 | O74501 | P31987 | P43944 |
| P19446 | P40954 | P83048 | Q9H2Y9 | Q81MW4 | Q81MW4 | Q4PD66 | Q8YQ64 |
| Q43089 | Q659U0 | P26759 | P40953 | P07061 | P07061 | Q94714 | Q5ZRJ2 |
| Q2U6U0 | Q74AR6 | P56552 | Q27289 | Q59782 | Q59782 | P00502 | Q9HQ86 |
| P43372 | P13447 | O10364 | P22202 | Q493T4 | Q493T4 | O42941 | O95183 |
| P14398 | Q99MI1 | P02197 | P23499 | P65350 | P65350 | Q9RYM8 | Q8RC52 |
| P01334 | P14773 | Q9X2M2 | P05117 | P59379 | P59379 | O80932 | Q9PJV6 |
| P07505 | Q54179 | Q9NI62 | Q8X6C6 | Q6F8J3 | Q6F8J3 | P25526 | P50426 |
| P07154 | Q6VN46 | Q9KW14 | Q99954 | Q29542 | Q29542 | P10079 | P19100 |
| P02126 | Q5LWJ6 | Q5E6Q9 | P22334 | P17952 | P17952 | P05616 | P0AAL2 |
| Q8WX39 | P13588 | Q96X43 | Q313S2 | Q99952 | Q99952 | P23345 | Q9I3H7 |
| Q9ZRJ4 | P21543 | Q38900 | P00703 | O76828 | O76828 | P20721 | P56331 |
| P09645 | P14658 | Q9UAS2 | P28470 | Q865X1 | Q865X1 | P83447 | P09709 |
| Q9U0E6 | P00786 | P02187 | Q9I8X0 | Q7LFX5 | Q7LFX5 | Q9UBR2 | Q9W727 |
| P83629 | P20368 | O13479 | P20842 | Q9HET9 | Q9HET9 | Q95ND7 | Q9US08 |
| Q9UAE6 | Q6BKY9 | P11139 | P43019 | P02716 | P02716 | Q9M0C2 | Q9MFN3 |
| Q6LUA7 | P13916 | Q5F8Z2 | P0ACA9 | P32717 | P32717 | P52273 | Q6YQU7 |
| P76149 | Q41651 | Q8Z4S7 | P08089 | Q8HYJ3 | Q8HYJ3 | P52184 | P30668 |
| P42848 | O86165 | O03892 | P24368 | Q8PLY2 | Q8PLY2 | Q9WGE0 | Q9JF93 |
| Q9CR16 | P33126 | P23785 | P29339 | P55638 | P55638 | P30044 | P39369 |
| P62682 | Q865V1 | P07090 | Q9UWG2 | P68211 | P68211 | Q6GKD8 | P19360 |
| P45850 | Q7VP41 | P81641 | O51917 | Q8DGX5 | Q8DGX5 | Q9HFQ5 | P31720 |
| Q87T31 | P43736 | Q5JDJ0 | P33543 | Q96EQ0 | Q96EQ0 | Q9JK88 | P65861 |
| Q22866 | Q9Z1I7 | P18061 | P07443 | P80015 | P80015 | P54048 | Q747K3 |
| Q3J5T0 | P81708 | Q41448 | P61895 | P18033 | P18033 | P49663 | Q7MNM8 |
| Q8N1E2 | Q40831 | Q6IRU2 | P80273 | Q920D2 | Q920D2 | Q02088 | P24546 |
| Q8S8H3 | Q932F8 | P01320 | Q6BQZ1 | Q8VEC4 | Q8VEC4 | Q9UXZ0 | Q71WN4 |
| P29957 | P24702 | P18137 | P35793 | P28249 | P28249 | Q5ZYQ0 | Q06602 |
| P53651 | Q95146 | O54757 | P60370 | P57899 | P57899 | Q8RB68 | Q932M0 |
| O09210 | P46429 | Q08519 | P52259 | P69675 | P69675 | P48534 | P77717 |
| P91254 | Q9LDN9 | Q9LZT4 | Q9KUT3 | Q68RX7 | Q68RX7 | Q5R899 | Q8NE62 |
| Q17770 | P52254 | P04275 | P80034 | P62230 | P62230 | Q9SFF9 | O18229 |
| P19470 | P84789 | P02400 | P70623 | Q65LI4 | Q65LI4 | Q9NH48 | Q99N02 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Q03734 | Q9WXB9 | P25660 | Q43321 | P35905 | P35905 | O35023 | P59011 |
| P52399 | Q65RV7 | P07754 | P52256 | Q8TE76 | Q8TE76 | O18420 | P47876 |
| P55091 | P80050 | P35044 | Q9XLW8 | Q5PIU1 | Q5PIU1 | Q9FMR0 | Q7U4X5 |
| Q03156 | P34933 | Q21743 | Q8PLS0 | Q92PI3 | Q92PI3 | Q29524 | Q35873 |
| P02440 | Q8DKA0 | P21776 | P16061 | Q8M899 | Q8M899 | Q5E9Z2 | Q6G5T7 |
| O68191 | P11102 | Q9TWG0 | O15865 | P45710 | P45710 | P34952 | Q23628 |
| Q9K1I0 | Q59182 | O03889 | O46647 | Q91309 | Q91309 | Q26551 | Q21824 |
| P11517 | P02179 | P11582 | P11428 | P59334 | P59334 | Q5P339 | Q6TND4 |
| Q15399 | P10648 | O26110 | Q92QA1 | O28277 | O28277 | P36182 | Q81FP3 |
| Q6C662 | O08665 | Q886M3 | O49886 | O30118 | O30118 | P58062 | Q8G5W9 |
| Q8X4B4 | Q9CHS7 | P98042 | Q6S4N2 | P52448 | P52448 | P96763 | Q940G6 |
| Q04672 | P42638 | Q91CL9 | P08001 | Q928C2 | Q928C2 | Q96509 | P11961 |
| P46876 | P17897 | P25940 | Q9ZDF3 | Q8YP78 | Q8YP78 | P31025 | P0A6S2 |
| Q9FX85 | O33783 | P28583 | Q8BG58 | P08762 | P08762 | Q9M612 | P67304 |
| P27082 | P08537 | P25251 | Q607A5 | P42469 | P42469 | Q15782 | Q9WU49 |
| Q9ZP20 | Q4HXF6 | Q9BMK4 | Q72QL9 | Q62GV3 | Q62GV3 | P83410 | P08717 |
| Q9UIF9 | Q41114 | Q9EQC7 | Q3Z601 | Q4I8B6 | Q4I8B6 | P12795 | Q642A5 |
| Q80X76 | O44249 | Q5AXT6 | Q9Y2D1 | Q827R1 | Q827R1 | O35090 | P82812 |
| Q6HBF3 | P62735 | Q93Z04 | P04353 | Q5H390 | Q5H390 | Q9EPC5 | Q8CUT9 |
| Q96127 | P50545 | Q88PD5 | Q64118 | Q9H2X9 | Q9H2X9 | Q7YFV7 | Q5JHP1 |
| Q9PU28 | P07091 | Q6GJC8 | Q91Y47 | P26642 | P26642 | O52064 | P51486 |
| Q7MMR7 | O77012 | P50701 | P27051 | P0C070 | P0C070 | P14009 | P00154 |
| P37153 | P82683 | P10832 | Q9ZFC6 | P46725 | P46725 | Q8WXG8 | P54715 |
| Q3SIN4 | Q9XZ43 | Q9FHI7 | Q9FHJ2 | Q9D173 | Q9D173 | Q7ZZP9 | P25386 |
| P63083 | Q8CFG8 | P81428 | P60576 | P04453 | P04453 | P28842 | P31554 |
| P04690 | Q95239 | P52239 | Q03401 | O15259 | O15259 | Q97BG8 | Q8KBY5 |
| P19123 | Q39528 | Q40682 | P12471 | P0A7W4 | P0A7W4 | P64076 | Q91690 |
| Q9D8Y0 | Q01116 | Q37548 | P33049 | P62411 | P62411 | P02884 | Q7W0D0 |
| Q68HB4 | O35543 | P82734 | P29335 | P39618 | P39618 | Q3IC08 | P28710 |
| P33125 | P21227 | P02402 | P31713 | O07899 | O07899 | P41976 | O54055 |
| Q00248 | Q09055 | P10622 | Q5AEN1 | Q49ZF4 | Q49ZF4 | O75366 | Q45XI9 |
| P05597 | P36377 | Q03736 | P17445 | Q9CDW0 | Q9CDW0 | Q61696 | Q8S8U0 |
| Q5KIK5 | P42157 | P80425 | O47870 | Q7MIV3 | Q7MIV3 | Q8BMY7 | P80784 |
| P60226 | P02750 | Q29562 | Q05036 | Q6ENV2 | Q6ENV2 | P48774 | Q57209 |
| P29171 | P29522 | P52397 | P03952 | Q99034 | Q99034 | P15059 | P63723 |
| P02626 | P27350 | Q8CFG9 | P63185 | Q5E455 | Q5E455 | P10965 | P83415 |
| P08070 | P48869 | P08744 | P54014 | P53714 | P53714 | P67782 | P34917 |
| P00768 | P46088 | P11910 | P52905 | Q7Z5P4 | Q7Z5P4 | P06732 | P62999 |
| O08404 | Q7T3T2 | P02624 | P67937 | Q8ZM40 | Q8ZM40 | P29750 | Q00092 |
| Q63722 | Q75EZ3 | Q39693 | P33830 | O13034 | O13034 | P04266 | P20982 |
| P49182 | Q7Z1K4 | P42280 | P48978 | Q54714 | Q54714 | P50060 | Q889R0 |
| P11944 | Q9HXZ5 | Q922F4 | P08418 | P45883 | P45883 | P27168 | Q9YHA1 |
| Q3SJW8 | P29514 | P48666 | P44638 | Q9H869 | Q9H869 | Q83EL0 | Q6FEZ6 |
| P07371 | Q94B08 | P52290 | Q6RXL1 | Q62868 | Q62868 | Q43629 | P33986 |
| P40782 | P00443 | Q7SBX8 | P15846 | P45657 | P45657 | Q5XHX6 | Q9NQF3 |
| Q9ZCB9 | Q8WUA2 | Q9LQQ8 | Q6MPQ2 | O43676 | O43676 | P05491 | O61570 |
| Q17133 | P52405 | Q9CXG3 | Q02815 | P59780 | P59780 | P16079 | Q822U8 |
| Q9K849 | P83039 | P14710 | Q03683 | Q7N3Z5 | Q7N3Z5 | P99028 | Q7S055 |
| Q9TSN6 | Q9HFQ4 | P08551 | P16397 | Q4AAS1 | Q4AAS1 | Q6BI20 | P47660 |
| P00939 | P03954 | P19221 | Q91883 | P65206 | P65206 | Q68LC0 | P00014 |
| P02587 | P27741 | Q6INW9 | P19837 | Q5R0L7 | Q5R0L7 | Q9Y623 | Q81WL2 |
| Q8RGH4 | Q96L12 | P05546 | P27323 | Q45537 | Q45537 | Q8G1Z8 | Q00144 |
| Q898R2 | O83384 | P60180 | P25437 | Q9ZLW6 | Q9ZLW6 | P15459 | P22125 |
| P67979 | O74173 | Q88DU2 | P08419 | O74391 | O74391 | P13944 | Q6F7H0 |
| P35335 | Q99PW7 | P10605 | P26213 | Q6P026 | Q6P026 | Q90628 | Q92546 |
| P07761 | P80888 | Q7MYB3 | P53371 | Q9NST1 | Q9NST1 | O00097 | Q9ABF8 |
| P00763 | P30370 | P12708 | P19324 | Q62EU1 | Q62EU1 | P07237 | O88703 |
| Q8K0D2 | P61634 | P29025 | P58481 | Q864J2 | Q864J2 | Q9LVL2 | P29980 |
| Q6ZJJ1 | P50120 | Q9SR72 | Q95029 | Q9M9Y8 | Q9M9Y8 | P49924 | Q05933 |
| Q9SEU8 | P02828 | P28513 | P47547 | Q9NWD9 | Q9NWD9 | Q9Y219 | Q96PQ6 |
| Q5F5V4 | P93257 | P05605 | P34840 | Q71ZI2 | Q71ZI2 | P35394 | Q662D8 |
| Q43743 | P17713 | Q6G1F9 | Q7M1B9 | P31005 | P31005 | O93510 | P0A9A3 |
| Q9ES87 | P81647 | P00767 | P29835 | P15767 | P15767 | P50532 | P65571 |
| P98046 | Q8IFW4 | O66602 | Q9ZU91 | Q976J8 | Q976J8 | Q39165 | O14772 |
| P54626 | Q8HXP2 | Q6L8H1 | Q69AB2 | Q9PJN2 | Q9PJN2 | Q8NJR2 | P0AEH2 |
| P16094 | P93022 | Q8R4U2 | Q63969 | Q5M5U4 | Q5M5U4 | Q9FLB4 | P29625 |
| P16296 | P29597 | O46412 | Q61955 | Q8Z153 | Q8Z153 | P16049 | P93890 |
| P52396 | O14967 | P19246 | P01339 | P64017 | P64017 | P24633 | Q96563 |
| Q86XW9 | P52243 | P32954 | P62965 | P57690 | P57690 | Q9XGS0 | Q04561 |
| P29876 | O03890 | P11828 | Q47GZ8 | P03597 | P03597 | O57521 | Q96RU7 |
| Q5ZLV2 | Q91240 | P15108 | P02632 | P02930 | P02930 | Q8G6D5 | Q3B6F6 |
| P08628 | P14399 | Q5E326 | P27525 | P66937 | P66937 | P33505 | Q9NYV7 |
| P27798 | P22196 | Q91291 | P48721 | Q8IG42 | Q8IG42 | Q5XIW8 | P55536 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P08519 | Q43697 | Q37430 | Q39241 | Q8ZLE4 | Q8ZLE4 | P35144 | P80526 |
| P00608 | P04542 | P10462 | O65174 | Q9PE71 | Q9PE71 | Q9Y6U3 | Q965Q4 |
| P08734 | P08094 | Q94JX9 | P17945 | Q8KAY7 | Q8KAY7 | Q61483 | Q8WUQ7 |
| Q7T1K6 | P52261 | P15156 | O33756 | P33097 | P33097 | P19944 | P52803 |
| O97397 | P23202 | Q06343 | Q7LZB9 | Q8UFS3 | Q8UFS3 | P84775 | Q07806 |
| P67796 | P43689 | P18253 | P62339 | Q3M244 | Q3M244 | P80028 | Q6FN96 |
| Q96511 | Q27319 | P80353 | P82176 | P0A8C2 | P0A8C2 | Q9K717 | Q00502 |
| Q5KHA8 | Q29461 | Q9I8W8 | P47931 | Q80UW0 | Q80UW0 | P0AA30 | P00639 |
| Q6NCY4 | P16147 | P22954 | P80009 | Q5XEM3 | Q5XEM3 | P19664 | P50204 |
| P54651 | P36220 | P62757 | P10567 | Q9QZM3 | Q9QZM3 | Q91WP6 | Q9X0Y7 |
| P27084 | P53441 | Q5GYK4 | Q42616 | Q5GH68 | Q5GH68 | O88398 | Q07814 |
| P14211 | P82718 | O24386 | Q02498 | Q8KGG6 | Q8KGG6 | P39586 | Q7Z4Y8 |
| Q5K4E3 | Q8ENP5 | Q43744 | Q9SSK9 | Q893M2 | Q893M2 | O22959 | P19435 |
| P83053 | O04204 | P05933 | P87497 | Q31VX9 | Q31VX9 | O47672 | P48949 |
| P52241 | P98025 | P46226 | Q9LQ54 | P44426 | P44426 | Q838I4 | P21426 |
| P02827 | P32006 | P41753 | P47948 | Q8HZR3 | Q8HZR3 | P25250 | Q8D3A7 |
| Q9Z0E3 | Q7LZS8 | Q96520 | Q80TJ1 | P68077 | P68077 | P04264 | P75534 |
| P26214 | P01324 | P18288 | P42896 | Q57H64 | Q57H64 | P05940 | P42361 |
| O87864 | Q9ZV40 | P47103 | O00394 | Q75F02 | Q75F02 | P62150 | P25186 |
| Q91VE3 | Q83FF7 | P27541 | P29512 | O93479 | O93479 | P17801 | Q8FUA5 |
| P46208 | Q01546 | O84296 | Q37741 | Q72T26 | Q72T26 | P47201 | P16528 |
| Q25009 | Q8SPE9 | Q64268 | Q28295 | P25094 | P25094 | Q02917 | O21327 |
| P00719 | P03946 | Q9C3Z5 | P81366 | Q3JBV4 | Q3JBV4 | P42754 | Q6MBF8 |
| P03951 | P64264 | P01091 | O59838 | P07701 | P07701 | Q5K2P9 | O84032 |
| Q92BR6 | P59566 | P00999 | Q28542 | Q3Z5H3 | Q3Z5H3 | P80026 | P08740 |
| P16300 | P25274 | P27773 | P60623 | Q6G043 | Q6G043 | Q04189 | O13157 |
| P81648 | Q60854 | P30414 | P46225 | P99112 | P99112 | Q7YT16 | P63943 |
| Q95108 | P29416 | Q8VVC1 | Q8BFR2 | P47761 | P47761 | Q5RD23 | O60147 |
| Q12708 | P48171 | Q9BDJ3 | P01336 | Q98HB7 | Q98HB7 | Q6URW6 | P69432 |
| P48672 | P19540 | P42276 | Q9ZWC8 | Q32041 | Q32041 | Q9MA63 | Q8D235 |
| P14524 | P02118 | P02460 | P0C0L1 | Q2RQX7 | Q2RQX7 | Q96512 | Q81WX1 |
| Q5I2J3 | P22233 | P00564 | P54629 | Q9PQ00 | Q9PQ00 | P51588 | Q8TE85 |
| P54649 | O77458 | Q6D7Z6 | P80052 | Q8TRK7 | Q8TRK7 | P16228 | Q83ET0 |
| P10763 | P54357 | Q756H2 | P52268 | Q9PKR5 | Q9PKR5 | Q6YQT9 | Q5UPE8 |
| Q81SZ9 | P84577 | P02614 | O88780 | Q8ZVL3 | Q8ZVL3 | P80311 | Q47908 |
| P07483 | Q9NZU6 | P40047 | P00704 | P10903 | P10903 | Q9QYZ9 | Q9ZCZ4 |
| P14260 | Q9Y5Y6 | P28767 | Q818I1 | Q7MNN1 | Q7MNN1 | P12861 | P23251 |
| P05932 | Q7LZT1 | Q09011 | P45851 | Q81K75 | Q81K75 | P23400 | Q2VEE5 |
| P33048 | P14234 | P05976 | P12931 | Q5RBR1 | Q5RBR1 | O47667 | Q8ZKB1 |
| Q9JLK4 | Q93GF1 | P46263 | Q7LLZ8 | P19636 | P19636 | P00755 | Q8IA44 |
| Q8TKL5 | Q7VC41 | O78310 | P09762 | Q7XD65 | Q7XD65 | P10599 | O94964 |
| P24790 | Q9NFL6 | Q62471 | P01038 | Q27963 | Q27963 | P19059 | Q9WIK2 |
| Q84WM9 | Q6IFZ6 | Q9D695 | P04119 | P03385 | P03385 | Q4UVY7 | Q980I7 |
| P15460 | P08670 | Q6FF90 | Q9QYN3 | P58618 | P58618 | P11090 | Q9LQQ2 |
| P00756 | P02565 | Q9CQ07 | Q920J4 | P17162 | P17162 | P13396 | Q6F243 |
| P02772 | P64079 | Q01591 | Q03196 | O51470 | O51470 | Q43130 | Q09338 |
| P50117 | P02671 | P08003 | P11021 | P64267 | P64267 | P25779 | Q8DSX9 |
| Q41162 | Q6NI61 | Q9JVQ9 | Q10092 | Q06795 | Q06795 | P35564 | P32656 |
| P02463 | Q95199 | Q8BT20 | Q7XA42 | Q9DDA9 | Q9DDA9 | Q95224 | Q5LQP4 |
| Q5FX67 | P59996 | P72242 | Q8VBX1 | Q9UK17 | Q9UK17 | P29874 | Q11183 |
| Q9NJN7 | Q7SF72 | P11796 | P19140 | Q64143 | Q64143 | P82536 | P22499 |
| Q5K2P8 | Q06285 | P12311 | O64973 | Q6G125 | Q6G125 | P02147 | P0A4S4 |
| P51913 | P29027 | Q7LZT2 | O22349 | P0ADJ0 | P0ADJ0 | P29269 | Q9TTS3 |
| Q9FJB5 | P04111 | O31186 | P50686 | Q8P9X8 | Q8P9X8 | Q5LQL4 | Q7VKR8 |
| Q6VAF4 | P08175 | Q9DG68 | P42155 | P21789 | P21789 | Q28380 | Q10008 |
| Q5FSL5 | P97776 | P79105 | P07686 | P48301 | P48301 | P38596 | P10520 |
| P13796 | P27520 | Q7J6G2 | Q5JJV0 | Q68XS9 | Q68XS9 | P0A861 | P02723 |
| P41043 | P11018 | P78386 | P11232 | Q9VMD6 | Q9VMD6 | P21668 | P37737 |
| P05437 | P34688 | Q14554 | P12360 | Q7W5H7 | Q7W5H7 | P82281 | Q12948 |
| P37160 | P11857 | P04254 | Q9ZPP0 | O70252 | O70252 | Q9XD74 | O43159 |
| P10707 | O43353 | Q9VJ26 | Q9CD42 | Q8GPG1 | Q8GPG1 | P00275 | Q87L78 |
| P37842 | O81645 | Q9GSU4 | P08590 | P24833 | P24833 | Q04760 | P83382 |
| P15195 | P68176 | P26729 | Q9NZR2 | Q8K9A2 | Q8K9A2 | P14412 | Q9JHX6 |
| P40714 | P62712 | Q9UKX2 | Q7RAV5 | Q9KPT1 | Q9KPT1 | P24062 | Q8VDU5 |
| Q40646 | Q3SW76 | P00689 | P24705 | Q92R94 | Q92R94 | P02184 | P02919 |
| P16588 | P19442 | Q5R4E2 | Q8K981 | Q8U0U6 | Q8U0U6 | P05786 | Q5F5Y7 |
| Q8CPL5 | Q9DA01 | Q9S660 | Q9V2X0 | P49970 | P49970 | Q7ZT98 | O07103 |
| P81646 | P07052 | Q29411 | P06787 | P10696 | P10696 | P14085 | Q3Z637 |
| Q89KV6 | P29030 | P52252 | Q8ZW52 | P43544 | P43544 | Q07182 | Q8C4Y3 |
| Q00915 | Q39313 | O14313 | Q10991 | O13077 | O13077 | P06638 | Q03434 |
| Q6N5U6 | Q92335 | P83972 | Q9P4C2 | Q8R844 | Q8R844 | O01305 | O14910 |
| Q641X3 | P04667 | Q88CW6 | O65199 | P26847 | P26847 | O49813 | Q89L59 |
| Q41785 | P32255 | Q42372 | P14417 | Q13492 | Q13492 | Q04447 | Q9CNJ5 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| O70251 | Q6LTE2 | P07455 | P04573 | Q921C5 | Q921C5 | P42684 | Q5QD17 |
| P29242 | P25188 | Q26548 | Q4UYN4 | Q6FCI6 | Q6FCI6 | Q9YMP9 | Q9ZDG8 |
| P14804 | P54639 | P27056 | Q8TD86 | Q5UR02 | Q5UR02 | Q8BLY3 | P21973 |
| Q02126 | Q8DJ55 | Q91060 | Q56696 | Q6LYY1 | Q6LYY1 | Q90629 | O14753 |
| Q9YD10 | Q90239 | P41747 | Q75A33 | Q65ED8 | Q65ED8 | Q7X3X5 | Q9A2G0 |
| Q9SLH7 | P25334 | Q23894 | Q14533 | O15155 | O15155 | P19886 | P40585 |
| Q9DEN8 | P43370 | P46439 | P11047 | Q58884 | Q58884 | Q8SPQ0 | P92555 |
| O97646 | P02670 | P11839 | P31000 | Q6AF88 | Q6AF88 | P24626 | Q8UBM5 |
| P56503 | Q9L7Z1 | Q8W9N3 | Q21313 | P08031 | P08031 | P41362 | Q9Z223 |
| P38879 | Q7M1R5 | P42042 | P00699 | P11892 | P11892 | Q985M6 | Q5L589 |
| P29216 | P22217 | Q6PXP0 | P23032 | Q02847 | Q02847 | Q8VHJ4 | Q9RV58 |
| O13340 | Q94HW3 | P41962 | P26215 | P00312 | P00312 | P48675 | P24068 |
| P31696 | Q8EHL9 | Q9SK27 | Q8BLR2 | Q6L3Y2 | Q6L3Y2 | P83957 | Q62LD1 |
| Q9S8M0 | P19137 | P81074 | Q95085 | Q8KWS7 | Q8KWS7 | P36907 | Q02819 |
| P09893 | P11009 | O14412 | P33154 | P38830 | P38830 | Q74K78 | P49831 |
| O77698 | P06470 | O43447 | P04903 | P04239 | P04239 | O24385 | Q5JH51 |
| P11955 | P02618 | P16292 | P28546 | Q8CZY7 | Q8CZY7 | O77014 | Q6ZMR3 |
| O24554 | P00979 | P80038 | Q8T8M2 | P34409 | P34409 | P58480 | Q8K8H2 |
| Q8J0N3 | O22022 | Q9ZV04 | P55994 | Q7Z092 | Q7Z092 | P01314 | P81784 |
| O49044 | Q9FMA8 | Q7XVA8 | O02192 | Q9RCA2 | Q9RCA2 | P83579 | Q7MK65 |
| O55000 | Q47XA7 | P18988 | P14716 | Q9F7L6 | Q9F7L6 | P60279 | Q6GEV5 |
| Q14520 | P80247 | P07146 | P20846 | O84462 | O84462 | P02231 | P14991 |
| P50381 | P09446 | P09102 | P0C196 | O57474 | O57474 | P42412 | Q9ANP1 |
| Q61699 | Q59094 | P04814 | P24525 | Q08079 | Q08079 | P80164 | Q7MJ36 |
| P02048 | P08935 | P00764 | Q05482 | P55121 | P55121 | Q8G5I9 | Q9ZL36 |
| Q9BN01 | Q04709 | P08263 | Q9XF61 | P80809 | P80809 | O77788 | P14353 |
| P27047 | P00760 | P36185 | O22668 | Q15388 | Q15388 | Q8CN04 | P34878 |
| P00761 | P55053 | P27450 | P02116 | Q9JW45 | Q9JW45 | P11884 | Q5RC37 |
| P27538 | Q06382 | P30156 | P00431 | P60245 | P60245 | Q6XE38 | Q63085 |
| P80424 | O94178 | Q63811 | Q5HT16 | Q9ZN56 | Q9ZN56 | Q3SYF9 | P65436 |
| P48038 | P20462 | P01346 | Q27287 | P96112 | P96112 | Q985M5 | Q9MS61 |
| P51541 | P00563 | P29613 | Q9C8T9 | P25910 | P25910 | O19023 | Q07205 |
| P0C0F8 | P12307 | Q9N1R0 | Q6QNF8 | Q71Y90 | Q71Y90 | P29449 | Q9AXQ9 |
| P27094 | O65888 | P39037 | P59409 | Q4VZN1 | Q4VZN1 | Q9FIC9 | P29409 |
| P35617 | P11482 | P10040 | O02611 | P37781 | P37781 | P29450 | Q7WDS1 |
| O47674 | Q06012 | P04989 | Q39950 | Q5QUC5 | Q5QUC5 | Q9FJZ9 | O95336 |
| P64411 | P12065 | Q9UKX3 | P33712 | Q88VY1 | Q88VY1 | P37713 | Q8PJK5 |
| P07196 | P02160 | O77021 | Q9R6P9 | Q72C17 | Q72C17 | Q8YE76 | P34439 |
| P50268 | P61896 | P22497 | P19883 | O27606 | O27606 | P15054 | P33800 |
| O03851 | Q9XS41 | P28287 | Q07276 | P65530 | P65530 | P20369 | O83295 |
| Q05094 | Q38858 | P00743 | Q8TRU7 | P62933 | P62933 | P48616 | Q5F7H5 |
| P19135 | Q00451 | Q47UV9 | O04795 | Q3SJS2 | Q3SJS2 | O42824 | P29993 |
| Q96VN5 | Q41350 | Q61096 | P53645 | P98055 | P98055 | P81635 | Q9UIK4 |
| P80065 | Q60604 | O54761 | P91913 | Q5RA29 | Q5RA29 | Q6IE47 | O93122 |
| P12104 | P16295 | Q00827 | Q27352 | P53540 | P53540 | P52194 | O05052 |
| O08523 | P23546 | P04783 | P52267 | Q6AF52 | Q6AF52 | P43738 | Q8YQX9 |
| Q9ZMS6 | P08729 | P82103 | Q9W7L0 | Q694B6 | Q694B6 | Q9KV30 | Q10443 |
| Q59838 | P51458 | O84544 | P38057 | P75540 | P75540 | P50254 | Q7WTJ2 |
| P17742 | P05598 | P41694 | P09179 | Q9S763 | Q9S763 | P45960 | P35078 |
| P61189 | Q26734 | Q8UEY3 | Q8YP17 | Q49434 | Q49434 | Q28046 | Q605B9 |
| Q9GMY3 | Q689Z7 | Q00488 | Q61129 | Q9D2M8 | Q9D2M8 | P30201 | Q9Y2G3 |
| Q96NI6 | Q7MGR8 | Q9XFS7 | P02189 | Q58489 | Q58489 | P29615 | P69702 |
| P20030 | Q9PSM2 | P14659 | P20379 | Q04770 | Q04770 | P32038 | Q720K7 |
| P14392 | P24634 | P02157 | Q40143 | Q05974 | Q05974 | P19157 | Q8CSG0 |
| Q76HN1 | Q9FXD8 | P54653 | Q08806 | Q9H400 | Q9H400 | P50879 | Q7N6T6 |
| P16579 | Q7KW39 | P09210 | Q4QN33 | Q9A8T1 | Q9A8T1 | P57653 | Q8BSZ2 |
| Q9FI23 | Q8U259 | P12469 | P09652 | P44628 | P44628 | Q6TY83 | P46162 |
| P10973 | Q8BMB0 | Q9U757 | P14641 | P11449 | P11449 | P10764 | O64769 |
| Q61072 | Q8W4Y8 | Q13427 | P43606 | Q9HB40 | Q9HB40 | Q8KNX9 | P14746 |
| P14832 | O33686 | O33833 | P23578 | P04948 | P04948 | P03995 | Q9I8U0 |
| P43234 | Q9GZM7 | P02174 | P61188 | Q98D10 | Q98D10 | P13867 | P51184 |
| P30801 | Q5NVM9 | P23613 | O02772 | Q96Q45 | Q96Q45 | P49384 | Q4URG1 |
| Q9D787 | Q9GLW9 | P22676 | Q13443 | Q88SV5 | Q88SV5 | P05562 | Q9RBJ3 |
| P09203 | P16975 | P13868 | P27054 | Q9ESN9 | Q9ESN9 | Q49WA1 | P40190 |
| O75891 | P52778 | Q9D868 | Q5KWZ7 | Q81TQ3 | Q81TQ3 | P05942 | P0A7H3 |
| P21812 | Q9C7F7 | P43785 | P27824 | P15925 | P15925 | P56730 | P95879 |
| Q03681 | O22874 | Q5I2M7 | P50913 | Q5R655 | Q5R655 | P37495 | Q5NL81 |
| P53653 | Q9D7Q1 | Q8YCV3 | Q60675 | Q09677 | Q09677 | P00942 | Q83HA5 |
| Q9XDS7 | Q9JHW9 | P81706 | P07759 | Q47SA2 | Q47SA2 | P57078 | Q8NB50 |
| P80310 | Q03461 | Q6RHW4 | P33720 | Q00471 | Q00471 | P55054 | Q92AH0 |
| P35043 | Q9NFL5 | P22577 | P58517 | P43964 | P43964 | P40150 | P46536 |
| P32718 | P41770 | O35186 | P82159 | P29745 | P29745 | P26912 | O79413 |
| O97125 | Q96XE0 | P05600 | P41361 | P35259 | P35259 | P11501 | P45438 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P53634 | P00238 | P60496 | Q47JB0 | P66987 | P66987 | P08851 | P79728 |
| P47796 | P12707 | Q9GRJ1 | Q95149 | Q6ELW3 | Q6ELW3 | O01395 | P0C0V0 |
| P13239 | P29413 | Q9VN93 | Q91159 | P75138 | P75138 | Q9JIA9 | P0A7U7 |
| P35466 | P13367 | Q8L3R2 | P84343 | Q9KUI0 | Q9KUI0 | Q8HYE4 | P11801 |
| P02605 | P49149 | O96081 | P00449 | P42006 | P42006 | P02875 | Q92GX3 |
| P30841 | Q7M8Q0 | Q7SDV9 | Q8ZKC3 | P65647 | P65647 | Q6FF82 | Q6FZG1 |
| Q02455 | P06603 | Q4L966 | O57391 | P27220 | P27220 | P29031 | Q9UI09 |
| Q9NPH6 | P01318 | P08218 | Q9HPD4 | Q8UVV7 | Q8UVV7 | P50116 | Q6C3J3 |
| O94788 | Q9UQX0 | P52258 | P07858 | P22997 | P22997 | P47773 | Q63008 |
| P19134 | Q62356 | P19136 | Q8HEC3 | O34835 | O34835 | P42324 | O83626 |
| P12704 | Q07488 | Q8S8H9 | P79335 | Q5T442 | Q5T442 | P36361 | P73809 |
| P0A4J7 | Q9Y616 | P0C1A2 | P67971 | Q835H1 | Q835H1 | P58482 | Q9ZMZ0 |
| Q14203 | P72186 | P22549 | O81644 | P26008 | P26008 | P12035 | P17391 |
| P31984 | Q9EPC6 | Q39785 | Q96VP4 | O25909 | O25909 | Q12335 | Q8U0Q5 |
| P68471 | Q8SEM9 | P37159 | P42115 | Q85X34 | Q85X34 | Q9WTV1 | Q58149 |
| Q38905 | P02586 | Q25008 | Q9GZN4 | Q75TH5 | Q75TH5 | Q60817 | P04991 |
| Q02243 | Q7RV85 | Q9SHY6 | P33523 | Q10173 | Q10173 | Q4IE79 | Q96S97 |
| Q6C9P3 | P14278 | P47032 | P33679 | P28186 | P28186 | P08678 | Q7VDT5 |
| P41363 | P83610 | Q91X97 | Q84V83 | Q3IRL7 | Q3IRL7 | Q99P60 | Q07211 |
| Q9HIQ9 | Q9D0W5 | P68147 | Q06383 | Q9XD23 | Q9XD23 | P02564 | Q864F6 |
| Q9GLW7 | Q04721 | Q75AA5 | P54652 | Q5E176 | Q5E176 | Q9ULV8 | Q5WEI8 |
| P83790 | P38567 | Q5WZL9 | P75189 | Q646D9 | Q646D9 | P09005 | Q7VKF1 |
| Q87RH5 | Q94EG3 | Q10993 | Q90404 | P66978 | P66978 | P50114 | Q8YW05 |
| Q9CQV3 | P09457 | P49385 | O78750 | Q01753 | Q01753 | P17139 | Q9X109 |
| Q9UKR3 | Q59081 | P24629 | Q9HDE1 | P77260 | P77260 | Q4ZM62 | Q8ZJ87 |
| P20151 | O23317 | Q58939 | O24387 | P25685 | P25685 | P17512 | P79208 |
| Q09596 | P31515 | P02152 | Q9CWT6 | P0A4S7 | P0A4S7 | P17513 | P05076 |
| P00706 | Q8FQS7 | P10878 | Q96RM1 | Q5SHZ1 | Q5SHZ1 | P33188 | Q9M3L9 |
| P06029 | Q9BMK3 | P01011 | Q8G417 | O95750 | O95750 | Q7UNC6 | Q9A8T9 |
| Q15661 | P42821 | P02607 | P41978 | P16320 | P16320 | Q6AM97 | P35809 |
| Q93Z25 | P37227 | Q9PK66 | P30116 | Q8G340 | Q8G340 | Q90VZ3 | O23970 |
| Q9GKX8 | Q9ZRB7 | Q82EL5 | P46265 | P39954 | P39954 | Q9SLF7 | Q34952 |
| Q03402 | Q6BMB8 | P96190 | P43298 | P0A8H5 | P0A8H5 | P01330 | Q06168 |
| Q9UL52 | P10255 | Q38904 | P68014 | P35191 | P35191 | P50718 | Q95ED9 |
| Q7TMF5 | P58796 | Q8JFB2 | Q39529 | O59426 | O59426 | Q9DG83 | Q9D1J3 |
| Q06394 | P00718 | Q811B3 | P77526 | Q52095 | Q52095 | P27164 | Q3ZYV4 |
| P02599 | Q5B4R3 | Q97E05 | Q9D7D2 | Q5M5R4 | Q5M5R4 | P20918 | P81721 |
| P02870 | P00565 | Q9U5M4 | Q91053 | Q589A6 | Q589A6 | Q6XPR3 | Q12514 |
| Q92BN8 | Q28895 | P84076 | Q21697 | Q9KPH9 | Q9KPH9 | P15545 | Q8KGG2 |
| Q6CC82 | Q02171 | P16026 | Q5WEG2 | Q5FSP9 | Q5FSP9 | Q805F2 | P0A0F7 |
| Q75GB1 | O80803 | Q25519 | P19665 | Q10435 | Q10435 | Q9JLJ2 | P27956 |
| P21569 | P00688 | Q8WSW3 | Q6DAN0 | Q5QVL5 | Q5QVL5 | Q04756 | Q9C5J9 |
| P41937 | P29141 | P14649 | P80416 | Q55797 | Q55797 | Q5NZ69 | Q7U659 |
| Q5R544 | Q8R4K2 | P27166 | P81591 | Q8CXV3 | Q8CXV3 | P30562 | Q88IU4 |
| P01008 | P92478 | P00772 | P0A6Q2 | P96955 | P96955 | O17486 | Q97Q59 |
| P33685 | P82966 | P25071 | P66928 | O73557 | O73557 | Q01645 | P83571 |
| P10280 | P67974 | P87102 | Q6FJI3 | Q92R54 | Q92R54 | P00983 | P46152 |
| O70370 | Q9NRS4 | Q03420 | P20029 | Q39644 | Q39644 | Q9TTE1 | P21680 |
| P02686 | P04260 | P14830 | Q81FU5 | Q83I82 | Q83I82 | P02153 | Q5XDM4 |
| Q88YH3 | P10831 | Q5BKQ4 | P68171 | Q8FBH7 | Q8FBH7 | Q88QP2 | P12585 |
| P46259 | P13858 | Q5XBF8 | Q8IU80 | P87126 | P87126 | Q9ZZY8 | P60689 |
| Q7UP89 | P57393 | P93538 | P50568 | Q83HA1 | Q83HA1 | P93761 | Q8TTA9 |
| P83370 | P87800 | Q9RYG9 | P11863 | Q9NWC5 | Q9NWC5 | P81460 | P33669 |
| Q9M0I2 | Q63548 | P48594 | P19889 | P37711 | P37711 | P00126 | Q56195 |
| P52427 | Q43194 | P0A1Q2 | P04781 | O54713 | O54713 | Q5R9I9 | P25308 |
| P09104 | Q43119 | P79239 | P98027 | Q59059 | Q59059 | Q49Y22 | P01634 |
| P09865 | P02401 | Q6FQY4 | Q9C7F5 | Q6GG18 | Q6GG18 | Q5E8E5 | Q82DN7 |
| P30224 | Q4QL89 | Q71WX1 | P50422 | Q62167 | Q62167 | Q9D4J1 | Q9PQQ2 |
| P52476 | O18345 | P87222 | P48870 | Q32J27 | Q32J27 | P20855 | Q9RVY1 |
| O26824 | Q42971 | Q9XCB1 | P18160 | P42843 | P42843 | P02120 | Q4P9E5 |
| Q3IKQ2 | Q9SRH4 | Q5R482 | P19084 | Q9V1G4 | Q9V1G4 | P12701 | Q8PD71 |
| P32262 | Q9R0H5 | P15159 | P26791 | P55219 | P55219 | Q9YFM2 | Q57I25 |
| Q88MF9 | P80322 | Q05855 | P09462 | Q8TWK0 | Q8TWK0 | P52897 | Q7WMU9 |
| P00411 | Q7M3C1 | Q66ED8 | P14080 | P44150 | P44150 | P30721 | P83146 |
| P15947 | Q8JFZ2 | Q867B0 | Q5R1W8 | Q6UXV4 | Q6UXV4 | P50693 | Q8UFY5 |
| Q9SCP2 | Q9R1K9 | P29243 | Q9AA17 | Q9PC10 | Q9PC10 | P07867 | P22335 |
| P00541 | Q93GB7 | Q9C554 | Q05431 | Q4QNS2 | Q4QNS2 | O65091 | Q5HHK8 |
| Q27178 | O74300 | Q8GT95 | P57936 | Q8FF59 | Q8FF59 | P08299 | Q9DBG9 |
| P87178 | P00412 | Q4QRB4 | Q37370 | Q9CQQ8 | Q9CQQ8 | P51433 | Q5SHR6 |
| Q91508 | Q5AUG9 | P09930 | P52018 | P70408 | P70408 | O79404 | P44643 |
| Q9N1Q8 | P59527 | P30740 | Q13087 | P0AG89 | P0AG89 | P48205 | Q6Q972 |
| P14642 | P41351 | P00445 | Q9X5C7 | Q9HDV4 | Q9HDV4 | Q5ZVS1 | Q38898 |
| P50918 | P02552 | Q03603 | Q02958 | Q8NZN7 | Q8NZN7 | P02242 | P72505 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Q55611 | P09805 | Q5S248 | P08976 | P62252 | P62252 | Q7V483 | P03933 |
| Q5RBM7 | P53421 | Q99M74 | P53014 | Q60892 | Q60892 | P36362 | Q8EVF5 |
| Q6AMQ3 | Q6IP73 | P55031 | O27002 | Q04633 | Q04633 | P12331 | P10073 |
| Q9W6Y1 | P00434 | Q62406 | Q8HXP1 | O10355 | O10355 | P48976 | P11510 |
| P09755 | Q9RTU4 | P00941 | Q9TSR4 | Q6HD54 | Q6HD54 | Q751L8 | Q6F8H3 |
| Q59689 | P13046 | P23473 | Q6RI85 | P32899 | P32899 | P13647 | P97798 |
| O83326 | Q72F92 | Q42686 | P53648 | Q5JFD3 | Q5JFD3 | Q92876 | P38724 |
| Q9NZT1 | P0AC83 | Q5NGW8 | Q9Y8D9 | Q9CZR3 | Q9CZR3 | Q9U4X4 | P08315 |
| Q9SB81 | Q8PC56 | Q8JI38 | Q8FCA2 | P49748 | P49748 | Q9FL78 | Q4WHG1 |
| P51964 | P08582 | O81235 | Q95VA8 | Q9UFN0 | Q9UFN0 | P50021 | P72394 |
| Q2TZ33 | Q92901 | Q8R9D0 | P08107 | Q9HIN8 | Q9HIN8 | P52577 | P30822 |
| P80351 | O69174 | Q8FVH2 | P54680 | Q24180 | Q24180 | Q17020 | Q8EIG9 |
| Q7U3T1 | P29598 | P00740 | P49862 | P58106 | P58106 | P29787 | Q9PVW8 |
| Q7S7Z6 | Q94572 | P14397 | Q944R1 | Q5WDZ8 | Q5WDZ8 | Q5DU41 | Q5LRU0 |
| P32765 | P23284 | P59462 | O21400 | P95173 | P95173 | P38659 | P28495 |
| Q91GE3 | Q64119 | P36908 | Q8DL40 | Q90257 | Q90257 | P47945 | P29777 |
| P28582 | Q4ZNP7 | P05977 | Q8PV50 | Q646F4 | Q646F4 | P07093 | Q8G0U3 |
| P32882 | P08802 | P52403 | O54233 | P53864 | P53864 | Q4W1I8 | P51691 |
| P82279 | Q8XUZ8 | P33183 | Q5E3A7 | Q80T62 | Q80T62 | Q962P8 | P56199 |
| Q9BLG0 | P04253 | O47676 | P35467 | P41117 | P41117 | P68127 | Q6D001 |
| P04655 | Q01202 | P30115 | P16080 | Q720J7 | Q720J7 | P31226 | Q65G93 |
| P34329 | P29448 | Q6L0S7 | Q6GUL6 | Q8NXF1 | Q8NXF1 | Q9R1B9 | Q6GIC6 |
| P35049 | Q9HV51 | P29244 | O55035 | Q5DZB8 | Q5DZB8 | P98073 | P52685 |
| Q8YPH9 | Q8U477 | P24146 | P23431 | Q65KJ5 | Q65KJ5 | Q606T2 | Q58064 |
| P41977 | Q493S7 | Q7X5C9 | Q8TQ79 | P23006 | P23006 | P09671 | Q9PQX7 |
| P49066 | P17536 | Q5BIR5 | P00352 | Q6LHF5 | Q6LHF5 | P12067 | P07095 |
| Q9M088 | Q09928 | P60458 | P30188 | Q811W0 | Q811W0 | Q8LCT3 | Q7WC23 |
| P50342 | Q02437 | Q55585 | Q9ERU9 | P0A186 | P0A186 | P10323 | O65352 |
| P02241 | P15253 | Q5RBP6 | Q9W7L1 | P0A827 | P0A827 | P83037 | Q92E89 |
| P79268 | P04722 | Q9TU73 | Q923D2 | O32954 | O32954 | P04265 | Q82CI4 |
| Q11004 | P42222 | Q8NRS1 | Q814I6 | Q8A9S7 | Q8A9S7 | P25805 | Q9Y5S1 |
| P07411 | P02089 | P17642 | P22357 | Q5JDK7 | Q5JDK7 | Q6DHL5 | P18496 |
| O30563 | P18026 | Q9SRP5 | Q68GV9 | Q8N9R8 | Q8N9R8 | Q91WP0 | Q9M571 |
| Q9W092 | P11598 | P35842 | Q08420 | O94953 | O94953 | O22779 | P47410 |
| Q8P7T0 | P29289 | Q6GDJ1 | Q9QW30 | O33926 | O33926 | P91895 | Q5R6T1 |
| P05936 | P04809 | P69754 | P59261 | P01851 | P01851 | Q5PXS1 | Q31E63 |
| Q9LQ07 | P29870 | P16352 | P24987 | Q9V730 | Q9V730 | P02112 | Q99J56 |
| Q5JEV6 | P04779 | Q892R0 | P06858 | Q8FUA6 | Q8FUA6 | P43235 | Q58737 |
| P83616 | P17607 | O97341 | P35176 | Q83G56 | Q83G56 | Q82TU1 | Q9YGL6 |
| O87579 | Q6FNU6 | P52266 | P20866 | O67030 | O67030 | Q00008 | P77569 |
| Q8NJR4 | Q5ZTX1 | Q7LZM2 | P95576 | Q9HNZ1 | Q9HNZ1 | Q03302 | P31176 |
| P35237 | P29257 | P50697 | P14202 | Q9P7H8 | Q9P7H8 | P02853 | Q89HW8 |
| Q43646 | O60635 | P09769 | P32956 | Q98LU3 | Q98LU3 | Q9C413 | Q8XX54 |
| Q8S4P4 | P11237 | O89020 | P01040 | Q87TT6 | Q87TT6 | Q92Y27 | P16213 |
| P22085 | Q9WUD9 | P79811 | P0A618 | P84231 | P84231 | P15502 | Q90ZS6 |
| O18552 | Q68Y08 | Q9TA26 | Q8Z320 | P13340 | P13340 | O65719 | Q7VQN1 |
| Q28773 | P35173 | P28757 | O81332 | Q99935 | Q99935 | Q8VCA5 | P13788 |
| P37156 | Q90935 | P04421 | P29061 | Q9ES14 | Q9ES14 | P04879 | Q5VV43 |
| P34790 | Q9A7J9 | Q8G0P0 | P49063 | Q9Y496 | Q9Y496 | P04785 | Q4JQI1 |
| P98034 | P02226 | P68062 | Q9BGI1 | Q5X9Q9 | Q5X9Q9 | O93233 | Q8TH25 |
| Q60396 | P04723 | P27275 | O82709 | Q8UE43 | Q8UE43 | Q5QXL1 | P91374 |
| P10876 | O01673 | P00410 | Q80TG9 | P42521 | P42521 | P24733 | P41894 |
| Q42978 | Q07568 | P13535 | Q6SA95 | Q8PCE2 | Q8PCE2 | P58479 | P25118 |
| Q631M2 | P97493 | Q9ZJ80 | P33629 | Q64428 | Q64428 | P50695 | Q9RLB6 |
| O45035 | Q99574 | O15992 | Q5KVE6 | P06504 | P06504 | Q5TBE3 | P07297 |
| P09738 | P07730 | Q26565 | Q61UX1 | Q9XC89 | Q9XC89 | P05121 | P42218 |
| Q9JJV2 | Q5K2N3 | P17967 | P09223 | Q9MB33 | Q9MB33 | Q8LG89 | P17778 |
| P26823 | P34058 | Q9PB21 | Q5I2M4 | Q8RQD2 | Q8RQD2 | Q9UAL5 | P57303 |
| Q62426 | P83833 | Q8IFJ8 | P68298 | Q07851 | Q07851 | P02867 | Q12857 |
| P02046 | P14396 | P11418 | O42182 | Q31610 | Q31610 | Q60751 | Q89AL6 |
| P46437 | P81446 | Q9HEY7 | Q835R7 | Q96PD4 | Q96PD4 | P02547 | Q7VKB6 |
| P02638 | P14088 | P63018 | P00782 | P18251 | P18251 | O09164 | Q8UDN3 |
| O83889 | O33528 | P36718 | P82901 | Q9Y5F6 | Q9Y5F6 | P54630 | Q7MK12 |
| Q8KE61 | Q9UNK4 | P41219 | O13069 | Q44586 | Q44586 | Q42615 | P46066 |
| O01504 | P02869 | P42327 | Q63202 | P25213 | P25213 | P42156 | P24336 |
| P77674 | P11012 | Q941R6 | P99134 | Q98PK2 | Q98PK2 | Q7LZQ2 | Q8K941 |
| Q96190 | P29859 | P55275 | P19637 | P07457 | P07457 | P04939 | Q7N119 |
| Q6DGG0 | P80426 | Q95KR7 | P36910 | Q97BP1 | Q97BP1 | O00806 | P69406 |
| Q66D53 | Q9I3C5 | P19179 | P18258 | Q06691 | Q06691 | P06761 | Q5WZ43 |
| P93194 | Q8DPS0 | P96132 | P47944 | P73623 | P73623 | P57727 | P81433 |
| P57975 | P05095 | P11752 | Q9SEU7 | Q5R4X1 | Q5R4X1 | P21226 | P63255 |
| P36914 | P52251 | Q866X0 | P35416 | Q02909 | Q02909 | Q4KIH1 | Q9CXV1 |
| P43510 | Q6GCV9 | Q8NUM0 | Q8PJ38 | O84212 | O84212 | P45820 | Q6EV70 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P56208 | P16884 | Q4QXT9 | Q95081 | Q9X1F7 | Q9X1F7 | P30410 | P84078 |
| Q09139 | Q701N2 | P00992 | P22358 | P18750 | P18750 | P07900 | Q9YEL4 |
| Q8VQ15 | Q98JB6 | Q97L52 | P51407 | P12663 | P12663 | O74225 | Q4QMG7 |
| Q7DMN9 | Q02610 | Q8BJ66 | Q92155 | Q9JYQ9 | Q9JYQ9 | O83949 | P00827 |
| P19993 | Q6CF69 | P29752 | O19045 | P25124 | P25124 | Q5KFV5 | Q3J8S1 |
| Q601S2 | Q92HJ3 | P34189 | Q92Q98 | Q8D7K6 | Q8D7K6 | Q8DC62 | O07145 |
| O95925 | P18573 | P15165 | P00747 | Q9ZK27 | Q9ZK27 | P22777 | Q04236 |
| P20145 | Q7VNM6 | P21845 | Q12574 | Q8WWP5 | Q8WWP5 | Q604M4 | P03022 |
| P49864 | Q894P6 | P48500 | P41244 | P58412 | P58412 | Q14031 | Q8YPK5 |
| P29845 | Q9SSD1 | P22011 | P58477 | P68618 | P68618 | O14154 | P09961 |
| P10299 | P81167 | Q68WZ8 | P06239 | Q5N3D4 | Q5N3D4 | Q25464 | Q884I1 |
| P18320 | P60411 | P19685 | P80736 | P66976 | P66976 | Q8PMB0 | Q9Z3Z6 |
| P73263 | P00448 | Q7VIH4 | Q6GFJ3 | P67335 | P67335 | Q60557 | P40723 |
| Q95095 | P49118 | P00687 | P91778 | P59245 | P59245 | P36110 | P22765 |
| Q8TY90 | P35339 | P46264 | P51547 | P25798 | P25798 | Q9BPL7 | P60510 |
| P93176 | P07036 | Q8TZZ8 | Q9Y8E3 | Q9UUF2 | Q9UUF2 | P67892 | P34730 |
| P54114 | Q65W41 | Q5M561 | P80450 | Q8DBF7 | Q8DBF7 | Q90WX8 | Q9KCB0 |
| Q6PR54 | Q47HK2 | P23869 | P02191 | P13902 | P13902 | Q9SRP7 | P69004 |
| Q889X8 | Q16956 | Q96P63 | Q82HH5 | Q7SIB9 | Q7SIB9 | Q46845 | Q5LRI7 |
| P16227 | P04636 | P00995 | Q43871 | Q5HMA2 | Q5HMA2 | P38067 | P23190 |
| P83596 | O46202 | P49175 | P09244 | Q8EX97 | Q8EX97 | P55050 | P14332 |
| Q91166 | Q3APD2 | Q9LZ99 | P83959 | P0AFR9 | P0AFR9 | P07148 | Q8U9L4 |
| Q06000 | P02566 | P20533 | Q07288 | P26908 | P26908 | Q86UW7 | P55304 |
| P00712 | P15839 | Q9V3Q6 | Q6ZSA7 | Q6CC99 | Q6CC99 | P67942 | Q5LXR8 |
| Q68F90 | P36584 | P82187 | P83684 | P64832 | P64832 | Q07412 | Q8KTR0 |
| P08123 | Q9C2U0 | Q9M8X6 | Q6T499 | P28386 | P28386 | P55052 | Q7VKJ6 |
| Q9Y5K2 | Q9NZU7 | Q9SMU8 | Q73P50 | P0A0K8 | P0A0K8 | Q4ZMN7 | P0A9P9 |
| P30405 | Q9D9Q6 | Q8DI58 | Q7RA60 | P58074 | P58074 | Q701N9 | Q7WJ91 |
| Q9QZL0 | Q9ES18 | P59121 | Q62148 | Q16666 | Q16666 | P29062 | P32529 |
| Q7VC80 | P80190 | O03895 | P83721 | Q68WW1 | Q68WW1 | P02595 | Q9RUS2 |
| Q03610 | P60994 | Q99583 | Q9FRV1 | Q8X8Y9 | Q8X8Y9 | O77015 | P82284 |
| P50903 | O60041 | P25077 | P11590 | O70436 | O70436 | Q57580 | Q8ES61 |
| P81481 | Q70CP7 | P38662 | P98019 | Q7C3R3 | Q7C3R3 | Q92NH8 | P11667 |
| P43293 | P94317 | Q37605 | Q8GE63 | Q08014 | Q08014 | Q9NZ20 | P01350 |
| O14618 | Q9UU93 | Q8D2K1 | Q4R5F2 | Q9LEK8 | Q9LEK8 | Q7VQL4 | Q9EQG3 |
| P83667 | P62201 | Q5XD01 | P26301 | P21754 | P21754 | Q04108 | Q5RJK8 |
| Q8L4H4 | P68085 | O08461 | O24493 | Q887Q0 | Q887Q0 | Q9NJN8 | Q8FU05 |
| O81148 | P37900 | P22910 | Q9YHC3 | Q16658 | Q16658 | P67894 | P29535 |
| P68368 | P23035 | P08975 | P02178 | P04831 | P04831 | Q9M8X4 | Q46560 |
| P15232 | Q02438 | P20375 | P00990 | Q92QQ2 | Q92QQ2 | Q9ZP21 | P18101 |
| Q5R8S7 | P50447 | Q5X3M7 | P24021 | Q65JP5 | Q65JP5 | Q9MAG6 | O08400 |
| P23105 | Q9PKA0 | P47918 | O18344 | P57260 | P57260 | P53623 | Q8G768 |
| Q7LZI3 | P83051 | Q8ZBN2 | P35048 | O79672 | O79672 | P83443 | P58374 |
| P07711 | Q8BKV0 | P98140 | Q9PTU8 | P54176 | P54176 | P23535 | P0ABV9 |
| P21266 | Q93WD2 | P07453 | Q06964 | Q6FVF9 | Q6FVF9 | Q8S4P6 | Q9V3A4 |
| Q95P07 | P21610 | O81223 | P41826 | Q8RPZ9 | Q8RPZ9 | P33044 | P26583 |
| P83036 | P33525 | O24509 | P19961 | P75958 | P75958 | O79549 | Q97HE5 |
| P29873 | Q16651 | P82968 | Q8BM88 | Q75D20 | Q75D20 | Q9DGI8 | Q8PNR4 |
| Q8ZGV9 | P48035 | P22576 | Q57AD7 | P93819 | P93819 | P00405 | Q493D9 |
| Q3ZC49 | Q37472 | P41095 | P24091 | P14076 | P14076 | Q824B2 | Q5F5K6 |
| Q9PJK3 | O18783 | P20857 | Q5RC84 | Q8I6S2 | Q8I6S2 | Q5N1J4 | P16020 |
| P09206 | Q9DEN9 | P61187 | P16019 | P23747 | P23747 | P46426 | Q96CJ1 |
| Q94715 | Q9M069 | Q823U7 | Q96LZ3 | Q63DX6 | Q63DX6 | Q29545 | P40006 |
| Q9CK91 | Q9AG20 | P05993 | Q43473 | Q8S9G8 | Q8S9G8 | P54781 | Q53462 |
| P36215 | P10587 | P02121 | P16501 | P01569 | P01569 | Q8DFM0 | P11440 |
| Q87BR7 | Q37684 | P35046 | Q28789 | Q13233 | Q13233 | Q91510 | Q6LUJ4 |
| P02181 | Q42580 | P29377 | P93543 | Q98D89 | Q98D89 | P02177 | Q57QH7 |
| P54625 | Q12567 | P46843 | P52262 | Q9GZV4 | Q9GZV4 | P54624 | P10205 |
| P45959 | Q9XF98 | P04728 | P80367 | Q29627 | Q29627 | P19864 | P40689 |
| Q6NU09 | O65857 | P01331 | Q97SG9 | P34136 | P34136 | Q68Y23 | P0C069 |
| Q8NLY6 | Q7PT10 | P34652 | P66832 | Q9VCZ8 | Q9VCZ8 | P80889 | O29534 |
| P81726 | P23314 | Q5ACI8 | P21563 | P49811 | P49811 | P52015 | P97297 |
| P02228 | P17476 | Q06647 | P31816 | Q8Y065 | Q8Y065 | P00994 | O67623 |
| Q62CH7 | P08963 | Q8W3J8 | Q7LZR3 | P43722 | P43722 | Q9M8X3 | Q9KVD6 |
| Q9RH76 | P00787 | P31949 | P23645 | Q9H0U4 | Q9H0U4 | Q8X166 | P52663 |
| P22281 | Q8VWY7 | P41509 | Q9SUT2 | Q9HLJ2 | Q9HLJ2 | O76263 | P41207 |
| Q9QXX0 | P34953 | Q5WHG1 | Q9TWF9 | P49113 | P49113 | Q967Y8 | Q72NU6 |
| Q8HXV2 | Q91775 | P01002 | Q8C2S7 | Q9V280 | Q9V280 | Q08694 | P0A2R6 |
| P06867 | P0A5B9 | Q6MNF8 | Q8HXP9 | Q8UJ85 | Q8UJ85 | Q8K1H9 | Q57IW5 |
| O76459 | P10623 | Q7MYW9 | Q91511 | Q5PH81 | Q5PH81 | O76262 | P42321 |
| Q92J36 | Q8N475 | P52009 | P45665 | Q9PL33 | Q9PL33 | P02148 | P12235 |
| Q4R520 | P56598 | P27055 | P54315 | Q7A078 | Q7A078 | P09670 | Q93EU6 |
| P09213 | P05190 | Q59179 | P81482 | O83908 | O83908 | O54210 | Q5HM84 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P43312 | Q24895 | Q02905 | Q7ZZN8 | Q7NIM7 | Q7NIM7 | Q66BP3 | P14184 |
| P12332 | P51614 | O59605 | P00757 | P37793 | P37793 | Q28198 | P54286 |
| Q41073 | Q96543 | P07669 | P57005 | P29068 | P29068 | Q9HRY2 | O51933 |
| Q91WD9 | P28763 | P30233 | Q38879 | P52948 | P52948 | P47728 | P43853 |
| P36222 | Q87WQ1 | P02243 | Q03773 | Q6L3Z0 | Q6L3Z0 | P0A0T3 | Q6DE96 |
| O24418 | P00993 | P70124 | O69777 | Q99SN6 | Q99SN6 | O52191 | P84826 |
| P11670 | P35003 | Q8YHF0 | P84287 | P70453 | P70453 | P48364 | Q7UWN8 |
| P17094 | Q9UKQ9 | P84794 | P73789 | P57791 | P57791 | O74190 | Q5HW83 |
| Q43116 | Q27380 | P91252 | P39036 | P0A1P1 | P0A1P1 | P48979 | P35449 |
| P80293 | P82980 | O70560 | Q9M7T0 | P02741 | P02741 | P11778 | P10865 |
| P26913 | P51617 | P35792 | O45687 | Q8KA79 | Q8KA79 | P34723 | P04407 |
| P11481 | P68990 | Q96I82 | P07304 | P73071 | P73071 | P04362 | O70274 |
| P25773 | P25696 | Q06915 | P43233 | P35498 | P35498 | P48677 | Q9XPP0 |
| Q9GL21 | Q01548 | P15120 | P17979 | Q65I58 | Q65I58 | Q7NXI3 | P69095 |
| Q27249 | P63317 | P05490 | Q39362 | P62973 | P62973 | P33676 | Q4FR05 |
| P38693 | Q9Y6M0 | P42325 | P33688 | Q83485 | Q83485 | O59418 | P41070 |
| P07458 | Q7VKL7 | P14323 | Q8KE49 | O30167 | O30167 | P59584 | P69770 |
| Q6LMT1 | P12412 | P00195 | Q8BXA0 | Q7VYB9 | Q7VYB9 | Q23761 | P20982 |
| Q6JVE9 | Q8EW34 | P82771 | Q6AYZ1 | P38456 | P38456 | P25691 | Q889R0 |
| P12693 | P17085 | P77395 | P20473 | Q6YPX7 | Q6YPX7 | Q01556 | Q9YHA1 |
| P07979 | Q06930 | P17477 | P46418 | P69040 | P69040 | Q45KW7 | Q6FEZ6 |
| P18109 | Q38867 | P56677 | P50685 | P36254 | P36254 | Q9UNP9 | P33986 |
| Q26496 | O75340 | P68363 | P01338 | P17210 | P17210 | P92997 | Q9NQF3 |
| P29402 | P09678 | Q01100 | Q04004 | P57925 | P57925 | Q6Q487 | O61570 |
| Q9S7W1 | Q4FNP9 | Q9UHF7 | O94739 | O69629 | O69629 | Q01469 | Q822U8 |
| P12262 | Q00898 | O03891 | Q8P9Z3 | P32726 | P32726 | P02692 | Q7S055 |
| Q02916 | Q71WW9 | O08762 | P15944 | P02439 | P02439 | P08110 | P47660 |
| Q9LYN8 | Q8V4T7 | P0AFL3 | P06865 | O35001 | O35001 | Q42578 | P00014 |
| P83335 | P41383 | Q6IN38 | Q9M203 | Q5M3J8 | Q5M3J8 | P49366 | Q81WL2 |
| P15094 | P29616 | P81025 | P28761 | Q82J66 | Q82J66 | Q9ER58 | Q00144 |
| Q5FS51 | Q9ZPN7 | P50421 | P36378 | P62205 | P62205 | Q9XFH9 | P22125 |
| Q02253 | P11118 | P06399 | O75635 | Q7KV19 | Q7KV19 | P80511 | Q6F7H0 |
| P28764 | P54316 | O76265 | P26010 | O30851 | O30851 | P68375 | Q92546 |
| Q3A8C2 | Q9FMB0 | Q9ZPS9 | O48276 | Q6D446 | Q6D446 | Q9SCC8 | Q9ABF8 |
| Q9BUP0 | Q08048 | Q5B1Z0 | Q06014 | Q48EC9 | Q48EC9 | P08861 | O88703 |
| P25973 | P80146 | P40372 | P0C0I5 | Q16613 | Q16613 | Q39034 | P29980 |
| Q63SQ0 | P70079 | P08252 | Q05091 | P46175 | P46175 | P29878 | Q05933 |
| O95757 | Q02252 | P32848 | P75344 | P35543 | P35543 | Q00652 | Q96PQ6 |
| P87051 | P16270 | Q921I1 | P04117 | Q9A231 | Q9A231 | Q9FLV5 | Q662D8 |
| Q9M5J8 | O59827 | P22359 | Q24956 | Q89AC3 | Q89AC3 | Q8PPG8 | P0A9A3 |
| Q98DD1 | P08562 | P29530 | Q8NHR9 | Q8G2R6 | Q8G2R6 | P12796 | P65571 |
| O15905 | O00060 | P23744 | Q6RY07 | Q8FMG0 | Q8FMG0 | P46660 | O14772 |
| P25553 | P20163 | P12965 | Q9UKR0 | O87197 | O87197 | P55776 | P0AEH2 |
| Q02566 | Q7MP97 | P20060 | Q9LNL0 | Q8ETX2 | Q8ETX2 | P40924 | P29625 |
| P21550 | P53652 | P30232 | P06869 | Q90069 | Q90069 | P83210 | P93890 |
| P30741 | Q2U256 | Q9JK72 | P21251 | P34365 | P34365 | P07980 | Q96563 |
| Q9LCQ5 | Q9NY15 | P62283 | Q2I3H1 | Q6F211 | Q6F211 | Q833J0 | Q04561 |
| O35205 | Q9U4X2 | P42894 | Q6BXZ7 | Q5WLN0 | Q5WLN0 | P62749 | Q96RU7 |
| P13731 | Q62468 | Q8ZQ40 | Q9PHZ3 | Q8WY91 | Q8WY91 | P24924 | Q3B6F6 |
| P22323 | Q6G173 | P19418 | O61367 | Q2YDK4 | Q2YDK4 | P05545 | Q9NYV7 |
| Q9BZJ3 | P61897 | P98164 | P00766 | Q5UNU9 | Q5UNU9 | P60673 | P55536 |
| Q9Z7P5 | P11941 | P29029 | P12788 | Q9KWX0 | Q9KWX0 | P81156 | P80526 |
| Q40190 | P29513 | Q2SSB0 | P55653 | Q8EDX4 | Q8EDX4 | Q9GZ71 | Q965Q4 |
| Q9FMA9 | Q5KAB3 | P01039 | P08228 | Q8FGL1 | Q8FGL1 | Q8F4T8 | Q8WUQ7 |
| Q5R1W4 | Q04619 | P41210 | Q8VBT0 | P18351 | P18351 | P84293 | P52803 |
| P51819 | P61185 | Q27727 | P32045 | Q5ARK3 | Q5ARK3 | O35632 | Q07806 |
| P02169 | Q9Z2T6 | Q8NKX2 | Q93Z08 | P11804 | P11804 | P49741 | Q6FN96 |
| P11996 | Q9ZRR5 | P48668 | Q4J781 | P14749 | P14749 | P08207 | Q00502 |
| Q9LFP5 | P80216 | Q9ZPP1 | O47677 | Q02336 | Q02336 | P27464 | P00639 |
| P04467 | P48660 | P00769 | P16354 | P53144 | P53144 | P07338 | P50204 |
| P40414 | Q7NYF6 | P36181 | Q06684 | Q8DFG4 | Q8DFG4 | P12066 | Q9X0Y7 |
| P22271 | Q9WU84 | P77949 | P19122 | P82927 | P82927 | Q40832 | Q07814 |
| Q25417 | P18321 | Q6CUI0 | P24101 | Q5PAN5 | Q5PAN5 | P04414 | Q7Z4Y8 |
| P29163 | P67896 | Q86YQ2 | Q90627 | Q9Y5I0 | Q9Y5I0 | P11838 | P19435 |
| Q9LSP0 | P10821 | Q9NYJ7 | Q6D6E0 | Q5JGJ7 | Q5JGJ7 | P08249 | P48949 |
| P28037 | Q9ZSI1 | P13723 | O27743 | P26546 | P26546 | Q61112 | P21426 |
| O54263 | P40220 | Q42762 | P42639 | P13922 | P13922 | Q9ER04 | Q8D3A7 |
| P82177 | Q6AAB8 | Q17299 | Q28315 | Q59NP8 | Q59NP8 | O49066 | P75534 |
| Q9FIC8 | Q07167 | P35042 | Q8R4Z1 | Q9RZ05 | Q9RZ05 | P42211 | P42361 |
| P10056 | Q9M2S9 | P27165 | P09485 | Q99VJ0 | Q99VJ0 | P80248 | P25186 |
| Q43636 | P11995 | P92995 | P25249 | P0A0N5 | P0A0N5 | P17205 | Q8FUA5 |
| P27425 | Q7MHQ1 | O19010 | P25854 | Q9YER2 | Q9YER2 | P81548 | P16528 |
| P26457 | Q41359 | P02615 | Q9ZMW4 | Q9Z6J1 | Q9Z6J1 | P24986 | O21327 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P05609 | Q29463 | Q4WVU5 | P81497 | Q9TLR0 | Q9TLR0 | Q6NBB9 | Q6MBF8 |
| P35479 | Q25145 | Q7V9G2 | O74435 | P57696 | P57696 | P53642 | O84032 |
| P05571 | Q5NB25 | P11964 | P21609 | Q59628 | Q59628 | P19104 | P08740 |
| P52250 | Q9TXA7 | Q4P6X6 | Q07563 | Q8FVC0 | Q8FVC0 | Q61982 | O13157 |
| P02825 | Q86SG7 | P02151 | P22972 | Q8PB52 | Q8PB52 | P02630 | P63943 |
| P04104 | Q9GTX8 | Q01642 | P04764 | Q88VP8 | Q88VP8 | O68988 | O60147 |
| P23096 | O04895 | P52230 | O33605 | P03690 | P03690 | O14293 | P69432 |
| P09799 | O93390 | P29519 | Q9KPC5 | P31819 | P31819 | Q8K9E0 | Q8D235 |
| P98036 | P37835 | Q9D1G5 | P79294 | P50888 | P50888 | P29060 | Q81WX1 |
| Q5HFK7 | P27482 | O02367 | Q6IG00 | P0A8H8 | P0A8H8 | P84535 | Q8TE85 |
| P09233 | Q60477 | P05980 | Q661A3 | P14203 | P14203 | P68294 | Q83ET0 |
| Q09334 | P27493 | P40370 | P24724 | Q93542 | Q93542 | P14170 | Q5UPE8 |
| P43727 | P07385 | P49742 | P01328 | P41955 | P41955 | Q9SMH2 | Q47908 |
| Q27535 | P25007 | P15626 | Q8HZ60 | Q2MI82 | Q2MI82 | P47776 | Q9ZCZ4 |
| P05656 | P22599 | P25807 | O17271 | Q8TT89 | Q8TT89 | O35508 | P23251 |
| P02145 | P38568 | Q8R966 | Q7V377 | Q5VYS8 | Q5VYS8 | Q28990 | Q2VEE5 |
| P34227 | Q7LZM1 | O47681 | P98012 | P03358 | P03358 | Q29147 | Q8ZKB1 |
| P36183 | Q41258 | Q711T9 | P93479 | Q92379 | Q92379 | O91466 | Q8IA44 |
| Q00216 | P29240 | Q96419 | Q9SZH2 | Q9BYZ2 | Q9BYZ2 | P46530 | O94964 |
| P80370 | Q26636 | P05561 | Q17127 | O25989 | O25989 | O42242 | Q9WIK2 |
| P42779 | P29786 | P02871 | Q66ET0 | Q6P7N4 | Q6P7N4 | Q8B9D5 | Q980I7 |
| P29457 | Q8VXF2 | O30826 | P51673 | P39512 | P39512 | P80476 | Q9LQQ2 |
| Q29512 | P07323 | P37226 | Q9UM47 | Q83PD9 | Q83PD9 | P26737 | Q6F243 |
| Q69SV0 | P11404 | Q9DD65 | P02193 | P66115 | P66115 | Q29548 | Q09338 |
| P00197 | Q41480 | Q00387 | Q9BYP8 | Q43644 | Q43644 | Q8SQ30 | Q8DSX9 |
| P20015 | Q43681 | P62154 | P04777 | Q49WZ2 | Q49WZ2 | P32824 | P32656 |
| P58775 | P83332 | O74729 | Q6VAG1 | Q63DJ6 | Q63DJ6 | Q92598 | Q5LQP4 |
| Q05589 | P51880 | Q5XQN5 | P04960 | Q5HCX7 | Q5HCX7 | Q8CIZ8 | Q11183 |
| Q02779 | Q71U36 | P01035 | P44516 | Q6AN63 | Q6AN63 | O15231 | P22499 |
| P05017 | Q9DGJ0 | P61442 | P56567 | P55271 | P55271 | Q9BQE3 | P0A4S4 |
| P47710 | P08238 | Q5R9M3 | P12762 | P32132 | P32132 | P43082 | Q9TTS3 |
| P15845 | P26732 | P09204 | Q61554 | Q05800 | Q05800 | Q04869 | Q7VKR8 |
| Q752Y2 | Q08277 | P14643 | P29024 | Q9RA51 | Q9RA51 | Q9NP86 | Q10008 |
| Q9M9Q9 | Q41420 | P00987 | O19093 | Q9JJ61 | Q9JJ61 | P02594 | P10520 |
| P82733 | P07463 | Q9ZMN8 | P44429 | Q91695 | Q91695 | Q61810 | P02723 |
| Q6XZB0 | P87072 | Q9FL79 | Q06209 | Q8JFG0 | Q8JFG0 | P50680 | P37737 |
| P31151 | O23547 | Q6AW23 | Q8W4J9 | Q5YYX9 | Q5YYX9 | Q7UIR2 | Q12948 |
| Q09840 | Q05319 | P07456 | Q00978 | O83451 | O83451 | P41963 | O43159 |
| Q02509 | P20856 | P61133 | P74934 | P14370 | P14370 | Q5R4V1 | Q87L78 |
| Q63639 | Q12007 | P54213 | P83298 | Q5FFE7 | Q5FFE7 | P06703 | P83382 |
| P02542 | Q89Z05 | Q93V93 | O99819 | Q9HRT7 | Q9HRT7 | P18087 | Q9JHX6 |
| Q9HGY8 | Q9HXY6 | P52398 | Q92JR5 | P27655 | P27655 | P05319 | Q8VDU5 |
| P50061 | Q42999 | P47739 | Q87EQ7 | O67520 | O67520 | Q800D3 | P02919 |
| P08010 | P01504 | P29820 | Q5SX39 | Q97EI3 | Q97EI3 | Q9JLK3 | Q5F5Y7 |
| Q9R063 | Q9HV43 | P16233 | Q96510 | Q7MGS0 | Q7MGS0 | P82113 | O07103 |
| P34576 | Q8EBR0 | Q9M8X5 | P24296 | Q9PC00 | Q9PC00 | P38658 | Q3Z637 |
| P22971 | P31671 | Q01233 | P52779 | Q5QT56 | Q5QT56 | Q7RAH3 | Q8C4Y3 |
| Q8BSS9 | P41825 | P50696 | P0C1A5 | P83187 | P83187 | Q43691 | Q03434 |
| P46598 | Q59994 | Q6FTW6 | Q9R014 | P38142 | P38142 | O53553 | O14910 |
| P35908 | Q8WSF3 | Q805F3 | P08438 | Q68XN7 | Q68XN7 | P13602 | Q89L59 |
| Q93X49 | O53510 | Q27031 | Q60HC2 | O57486 | O57486 | P36374 | Q9CNJ5 |
| P02173 | P15276 | Q63Q03 | Q9D1Q6 | Q6BSA9 | Q6BSA9 | P02868 | Q5QD17 |
| Q68FR8 | O43240 | P81061 | P35036 | Q9Z7H2 | Q9Z7H2 | Q9HFY6 | Q9ZDG8 |
| Q5E983 | P29529 | P25719 | P50692 | P52861 | P52861 | P52017 | P21973 |
| Q43857 | Q7NQE6 | Q02817 | P53372 | P67463 | P67463 | P49923 | O14753 |
| Q9ZSW1 | Q8NJR6 | P02399 | Q877B6 | Q5HRS3 | Q5HRS3 | Q804W2 | Q9A2G0 |
| Q29577 | O16127 | P02879 | Q6W3C0 | Q97NL4 | Q97NL4 | Q8D3K2 | P40585 |
| O43318 | P05591 | P31514 | O77019 | Q9Y5Z0 | Q9Y5Z0 | P23802 | P92555 |
| O75478 | P19913 | Q6C0Z6 | P19583 | Q56B59 | Q56B59 | Q9JHF7 | Q8UBM5 |
| O25982 | P00982 | P48499 | P70663 | Q10733 | Q10733 | Q81IP0 | Q9Z223 |
| P28325 | P00526 | O12933 | P05565 | P22174 | P22174 | Q8Z303 | Q5L589 |
| P02150 | P21788 | P57796 | P15465 | Q7U313 | Q7U313 | Q5ZL43 | Q9RV58 |
| P34461 | Q3BWZ1 | Q9KT93 | O48677 | Q8PCY1 | Q8PCY1 | P0C0Y3 | P24068 |
| Q8FM78 | Q09665 | P14949 | Q9JLK7 | Q5Z1V9 | Q5Z1V9 | Q7WQ31 | Q62LD1 |
| P80420 | Q9JI18 | P05315 | P79687 | Q8HY12 | Q8HY12 | P52193 | Q02819 |
| P25781 | P26461 | P29183 | Q09751 | P32782 | P32782 | P20799 | P49831 |
| P02787 | P52402 | Q5QWR2 | Q42614 | P42253 | P42253 | Q5E238 | Q5JH51 |
| P06173 | P34826 | P69048 | P00981 | Q8YQC1 | Q8YQC1 | P35491 | Q6ZMR3 |
| Q6MTZ2 | Q8JI40 | P22684 | Q4L9D7 | P0C0D9 | P0C0D9 | Q5FB29 | Q8K8H2 |
| P81245 | P09042 | P70275 | Q04962 | Q9CEI4 | Q9CEI4 | P18968 | P81784 |
| Q9LJF3 | P14279 | Q9M9S2 | Q00016 | O53563 | O53563 | O94038 | Q7MK65 |
| P14296 | P01322 | Q7V1N4 | Q875L2 | Q36518 | Q36518 | O06430 | Q6GEV5 |
| P07335 | P05596 | Q9NFH9 | P25312 | Q9HNE7 | Q9HNE7 | Q9ZKE6 | P14991 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P11150 | O02739 | Q8XNC2 | Q6HNR9 | Q7U3B9 | Q7U3B9 | P23433 | Q9ANP1 |
| P29760 | Q06118 | P81406 | P34935 | Q43876 | Q43876 | P28766 | Q7MJ36 |
| O02323 | P13533 | P24787 | P28608 | P95113 | P95113 | Q9ZRA4 | Q9ZL36 |
| Q8RWQ9 | Q43872 | Q9NSD5 | Q6V115 | P81661 | P81661 | Q03509 | P14353 |
| P29515 | Q5KNY5 | P14834 | P02619 | Q7N7A5 | Q7N7A5 | O47670 | P34878 |
| O73798 | O97508 | Q9ZRA8 | Q11174 | Q46448 | Q46448 | P15426 | Q5RC37 |
| O32507 | Q62632 | Q25293 | P58519 | Q2IJ76 | Q2IJ76 | P24817 | Q63085 |
| P08222 | O60259 | P47033 | Q42418 | Q601M2 | Q601M2 | Q7NXH5 | P65436 |
| P09648 | P35040 | P52275 | Q6CGK4 | P63752 | P63752 | P79179 | Q9MS61 |
| Q9JJS8 | Q6P6T1 | P81926 | P46235 | P46508 | P46508 | P06704 | Q07205 |
| P09486 | Q9UIK5 | O61270 | P29828 | Q04429 | Q04429 | P0A4J5 | Q9AXQ9 |
| Q9SD46 | P11219 | P08124 | Q25481 | O05411 | O05411 | O88751 | P29409 |
| P49114 | P58520 | Q69DK8 | P00998 | Q5P7N1 | Q5P7N1 | P84843 | Q7WDS1 |
| P46436 | P21641 | P29290 | Q42891 | P53267 | P53267 | P45854 | O95336 |
| Q40312 | Q8K3K4 | P66940 | P35794 | Q9VVD9 | Q9VVD9 | Q88FB9 | Q8PJK5 |
| P83326 | O97490 | Q6VAF5 | Q8Z2V9 | Q69384 | Q69384 | Q9WUU7 | P34439 |
| P23632 | Q5F6W5 | Q96P15 | Q6AF43 | P21458 | P21458 | Q9KWN0 | P33800 |
| P43374 | Q13356 | P80566 | Q90474 | Q87LM3 | Q87LM3 | P36373 | O83295 |
| Q5RCW5 | P04054 | P48816 | P25002 | Q6Q7K0 | Q6Q7K0 | Q58806 | Q5F7H5 |
| Q9BQR3 | P81719 | Q9ER10 | P04746 | P51694 | P51694 | Q9JIQ8 | P29993 |
| Q5R1M5 | P62155 | Q9SZM1 | P97816 | Q4QMV2 | Q4QMV2 | Q76FS3 | Q9UIK4 |
| Q87H52 | P04725 | P37395 | P05307 | Q8Z9T1 | Q8Z9T1 | Q4K4K8 | O93122 |
| Q6L8G4 | Q875V0 | Q8K4B2 | P98039 | O88275 | O88275 | P28237 | O05052 |
| Q03835 | P21798 | Q9L4K1 | Q28933 | O08816 | O08816 | O82484 | Q8YQX9 |
| Q829W1 | Q9TVD0 | O47680 | Q9SQL5 | Q7MDF5 | Q7MDF5 | P33331 | Q10443 |
| Q9XFH8 | O96827 | P80010 | P27497 | Q01320 | Q01320 | P02122 | Q7WTJ2 |
| P08479 | P26882 | Q02439 | Q6IE49 | Q9PL34 | Q9PL34 | Q7YRZ7 | P35078 |
| P29034 | Q9AGW2 | P02155 | P48501 | P38546 | P38546 | O64483 | Q605B9 |
| Q22492 | P07509 | P35016 | Q17424 | P43755 | P43755 | P99029 | Q9Y2G3 |
| Q9QUL7 | P04721 | Q5G265 | P17892 | O45319 | O45319 | Q9LMC9 | P69702 |
| P05317 | P22683 | Q7MN85 | Q61555 | P00225 | P00225 | Q91ZU6 | Q720K7 |
| P80532 | Q28522 | O94128 | P53683 | Q9XEK7 | Q9XEK7 | P02689 | Q8CSG0 |
| Q5NHT8 | P13006 | P00331 | P02171 | P37218 | P37218 | Q9M8U1 | Q7N6T6 |
| Q922R8 | Q8GR70 | Q87SU7 | P16409 | Q7VTU0 | Q7VTU0 | P00793 | Q8BSZ2 |
| Q4I5R9 | O33603 | P22105 | P00780 | Q5WFG8 | Q5WFG8 | P26372 | P46162 |
| P02562 | P50717 | P23297 | P28268 | Q9Y3D5 | Q9Y3D5 | P32642 | O64769 |
| P80192 | Q9DGJ3 | P02584 | Q6CJG5 | P64487 | P64487 | O77797 | P14746 |
| P68468 | P34955 | Q7T6Y2 | Q6XHI1 | P50571 | P50571 | O47673 | Q9I8U0 |
| Q3A9Q7 | P51463 | P06796 | Q00896 | Q92478 | Q92478 | P14754 | P51184 |
| Q6BEA2 | Q91509 | P52012 | P22935 | Q46172 | Q46172 | Q29477 | Q4URG1 |
| P48496 | O75007 | P04122 | Q6L8G8 | Q03411 | Q03411 | P20682 | Q9RBJ3 |
| Q8L202 | Q8HZ57 | Q8KB35 | O23095 | Q67Q48 | Q67Q48 | P43503 | P40190 |
| P49347 | P08261 | Q750Y8 | P47895 | Q97F65 | Q97F65 | P62966 | P0A7H3 |
| P05689 | P08071 | Q6P6Q2 | Q757B7 | Q7A461 | Q7A461 | P62028 | P95879 |
| Q8TNN7 | O79673 | Q9LUV1 | P34791 | O59292 | O59292 | Q9DBI0 | Q5NL81 |
| P12330 | O24388 | Q8CIY2 | P44557 | Q10244 | Q10244 | O00206 | Q83HA5 |
| P29524 | Q9LDR9 | P50107 | O17449 | P32016 | P32016 | P52724 | Q8NB50 |
| P61503 | Q7RXA6 | Q9Z0K9 | P00716 | P69936 | P69936 | P78590 | Q92AH0 |
| Q96JJ7 | Q81FQ3 | P25792 | Q9NDS0 | Q6P1G2 | Q6P1G2 | P21588 | P46536 |
| Q9HFQ6 | Q4R4X8 | P51544 | P41235 | P66028 | P66028 | Q29100 | O79413 |
| P50687 | P58371 | O95071 | Q6G7Q1 | Q8TZV1 | Q8TZV1 | P84347 | P45438 |
| Q8R9T6 | P20724 | Q95PU1 | Q7SYC7 | P82049 | P82049 | Q8LPB4 | P79728 |
| P81830 | Q21193 | P31394 | O13786 | Q6AG99 | Q6AG99 | Q5WJE6 | P0C0V0 |
| Q9U1G6 | O24415 | Q6JLX1 | O08739 | P23168 | P23168 | P92119 | P0A7U7 |
| P01326 | P22742 | Q8Y4I3 | Q91318 | Q9XT86 | Q9XT86 | Q8XXP9 | P11801 |
| P23951 | P04904 | P37685 | Q83A95 | O93385 | O93385 | Q97QS2 | Q92GX3 |
| P06604 | Q9DG84 | P29032 | P64604 | Q89AP7 | Q89AP7 | P29879 | Q6FZG1 |
| Q98Q50 | P01003 | Q4WP12 | Q65N49 | P14529 | P14529 | O80994 | Q9UI09 |
| P29023 | P06471 | P24882 | P05877 | Q9V1R0 | Q9V1R0 | Q5ZM98 | Q6C3J3 |
| P69098 | P02603 | Q8Z331 | P37615 | Q9JWU6 | Q9JWU6 | P17478 | Q63008 |
| Q9SRG3 | Q9JLJ3 | P29375 | Q724L7 | Q8RFJ7 | Q8RFJ7 | Q6DTV9 | O83626 |
| P09654 | Q10716 | Q00746 | P24751 | P18386 | P18386 | P29118 | P73809 |
| P06870 | P02874 | Q5I0H9 | Q8MIT8 | P19099 | P19099 | P02054 | Q9ZMZ0 |
| O67470 | Q09637 | P51462 | P11283 | Q5RB31 | Q5RB31 | O65175 | P17391 |
| P60204 | P25070 | P97821 | Q9Y2M5 | P61365 | P61365 | O65388 | Q8U0Q5 |
| Q4I1Y1 | P32955 | P49383 | Q6PQZ2 | Q9S7A0 | Q9S7A0 | Q9JJG7 | Q58149 |
| Q9KNR1 | P43305 | Q07090 | Q6N5Q9 | Q9AB09 | Q9AB09 | Q98967 | P04991 |
| P80359 | Q6GLR7 | P79158 | P20236 | O08770 | O08770 | Q8XW40 | Q96S97 |
| P26733 | P54712 | P02856 | Q7A0Q6 | Q976G3 | Q976G3 | Q9NSB4 | Q7VDT5 |
| P07291 | Q8X1S6 | Q9KD10 | Q9TM06 | O50316 | O50316 | P05091 | Q07211 |
| Q06850 | P39634 | P46488 | Q65JH8 | Q9KRU5 | Q9KRU5 | P81367 | Q864F6 |
| Q41161 | Q6MWS8 | O97388 | Q5HP68 | P45142 | P45142 | P08294 | Q5WEI8 |
| P29140 | P23432 | P81178 | P55489 | P16449 | P16449 | P36013 | Q7VKF1 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| O77016 | Q10284 | P52404 | Q8E2H3 | Q8BZA7 | Q8BZA7 | O82018 | Q8YW05 |
| Q5TCX8 | P04250 | P98121 | Q8P9B9 | P65318 | P65318 | Q9H3U7 | Q9X109 |
| Q00042 | P34937 | P0C0I0 | Q92A79 | Q968A5 | Q968A5 | Q26503 | Q8ZJ87 |
| Q06814 | Q02765 | Q9PDT8 | O02754 | P14984 | P14984 | P53301 | P79208 |
| Q6D9D1 | Q9TTK8 | Q8CDN6 | O84072 | Q8IWJ2 | Q8IWJ2 | P41980 | P05076 |
| P22852 | P70375 | Q4IPB8 | P54660 | P56980 | P56980 | Q4JB40 | Q9M3L9 |
| Q874T7 | Q01643 | Q9H013 | P0A5A8 | Q97I46 | Q97I46 | Q9FRA7 | Q9A8T9 |
| P21902 | P28318 | Q8K1T0 | Q85FZ0 | Q7W2G9 | Q7W2G9 | P11214 | P35809 |
| P51782 | P34934 | Q9Z6B9 | Q5HFJ4 | O94810 | O94810 | Q66D46 | O23970 |
| O04012 | O32757 | Q71ZJ7 | Q00659 | P34882 | P34882 | Q8DEC2 | Q34952 |
| Q5NFG7 | P61894 | Q37106 | Q493T5 | Q09602 | Q09602 | P14640 | Q06168 |
| P16226 | Q9CWG1 | P04689 | Q92834 | Q94VK5 | Q94VK5 | Q43533 | Q95ED9 |
| Q8EHT7 | Q88YH4 | P30230 | Q88XT4 | Q5LW41 | Q5LW41 | P09230 | Q9D1J3 |
| P80679 | Q94571 | Q28035 | P04661 | P53616 | P53616 | P07310 | Q3ZYV4 |
| Q61001 | Q9GQV3 | P59565 | Q8F1M2 | P14583 | P14583 | P50019 | P81721 |
| Q42434 | Q5KEB7 | P15399 | P69506 | Q9VLP3 | Q9VLP3 | P33431 | Q12514 |
| Q76KF9 | P22457 | Q75A41 | P01963 | Q9BWW8 | Q9BWW8 | P35039 | Q8KGG2 |
| Q06077 | Q00874 | O09043 | Q5FM66 | Q50367 | Q50367 | Q7GCM7 | P0A0F7 |
| P91928 | P02094 | P05389 | P49626 | Q99335 | Q99335 | Q95228 | P27956 |
| Q8TMP0 | P20865 | Q8G6W1 | Q9KNS1 | O94225 | O94225 | P51225 | Q9C5J9 |
| P05575 | P51779 | Q9HNP3 | P60851 | P06746 | P06746 | P48497 | Q7U659 |
| P02221 | P90703 | Q9VUQ5 | Q6NH04 | O64422 | O64422 | Q92YD2 | Q88IU4 |
| P32428 | P05978 | P07274 | Q9ZWQ8 | O86222 | O86222 | Q5WV02 | Q97Q59 |
| Q01957 | P02163 | Q9ZRB2 | P0AAS7 | O61199 | O61199 | P25020 | P83571 |
| Q27975 | Q9P6C8 | Q45670 | Q23400 | P64673 | P64673 | Q8W3K3 | P46152 |
| P08897 | Q7WGI4 | O60235 | Q6ENY5 | P96787 | P96787 | P07441 | P21680 |
| Q9Y718 | Q80LP4 | P19120 | O35077 | P22781 | P22781 | P02164 | Q5XDM4 |
| P29142 | Q95177 | P25974 | P46029 | Q92JM0 | Q92JM0 | Q8PTN8 | P12585 |
| O77022 | P0C0L0 | Q81VT8 | P30557 | O27670 | O27670 | P50413 | P60689 |
| Q5ARI5 | P24006 | P00705 | P30694 | P65954 | P65954 | P55915 | Q8TTA9 |
| P15722 | O65049 | Q9XF88 | P08944 | Q8DE73 | Q8DE73 | Q02096 | P33669 |
| O76264 | Q5PDJ5 | Q4JQI4 | O50224 | P21889 | P21889 | Q48K64 | Q56195 |
| Q5R632 | Q8WWQ8 | P52289 | P37941 | Q8PQ74 | Q8PQ74 | Q29491 | P25308 |
| P09350 | Q9LTZ0 | P33687 | Q8Z157 | Q43460 | Q43460 | Q9ZZ51 | P01634 |
| Q9UUE4 | P82716 | Q9UU78 | P50162 | Q5F8Y9 | Q5F8Y9 | Q8VHC5 | Q82DN7 |
| P0A0L9 | Q967Y9 | Q8FKI8 | P61618 | P06019 | P06019 | P29139 | Q9PQQ2 |
| Q51853 | P80059 | P61629 | O26337 | P83987 | P83987 | P26741 | Q9RVY1 |
| P08017 | P84290 | O42918 | Q47VQ7 | Q6ELV4 | Q6ELV4 | Q8UFH1 | Q4P9E5 |
| O59925 | P0A2F4 | P81370 | Q5X3Q5 | P39108 | P39108 | P07442 | Q8PD71 |
| Q9BUF5 | Q5FHC4 | P07370 | Q96WJ0 | Q8ZYE5 | Q8ZYE5 | Q90686 | Q57I25 |
| Q9Y6F8 | O30970 | Q05609 | Q83HM3 | O27216 | O27216 | P52575 | Q7WMU9 |
| P00567 | P54375 | P81711 | Q9KAD9 | Q9BH10 | Q9BH10 | P53655 | P83146 |
| P59435 | P50267 | Q5LN53 | P51207 | P15075 | P15075 | P98064 | Q8UFY5 |
| O23758 | P09455 | P00302 | Q9WVA8 | Q5P1Z7 | Q5P1Z7 | O69268 | P22335 |
| P49575 | O93532 | Q87DY6 | O44786 | Q9A434 | Q9A434 | Q5UQ49 | Q5HHK8 |
| Q6F6N3 | Q4ZUW2 | Q7NAY0 | Q9CJR2 | Q92259 | Q92259 | Q9V429 | Q9DBG9 |
| P08157 | Q6L8H2 | Q5XLD3 | P18005 | O69475 | O69475 | P13105 | Q5SHR6 |
| O49818 | P91958 | O74660 | Q6EW36 | P11755 | P11755 | Q95P08 | P44643 |
| Q17005 | Q9WVC1 | P01313 | Q87HS0 | Q9PDL4 | Q9PDL4 | P12703 | Q6Q972 |
| P22074 | Q02028 | P09756 | Q8XE76 | P94845 | P94845 | Q5E985 | Q38898 |
| Q94A16 | Q6YPM1 | Q95NI4 | P09618 | P57478 | P57478 | Q91041 | P72505 |
| Q96PK2 | P49874 | Q9ZJH5 | Q99MR1 | P33579 | P33579 | P18418 | P03933 |
| P48644 | Q40227 | P16418 | P0A080 | P67908 | P67908 | P41260 | Q8EVF5 |
| P84612 | Q6ADR6 | P06708 | Q664I5 | Q9DF69 | Q9DF69 | P16121 | P10073 |
| O04386 | Q37643 | Q5WDK9 | Q9SEE5 | Q57TC9 | Q57TC9 | P38013 | P11510 |
| P55097 | O33012 | Q60432 | Q14703 | Q60347 | Q60347 | O35679 | Q6F8H3 |
| Q3IUI0 | Q8CDC0 | Q9Z2P5 | O05139 | P66368 | P66368 | O47679 | P97798 |
| Q9PV90 | O43597 | Q8T0W5 | Q8EKS1 | P57125 | P57125 | Q63556 | P38724 |
| Q60173 | P25776 | Q40680 | P83414 | P30296 | P30296 | P24775 | P08315 |
| P59865 | Q5WWX9 | P08221 | O07923 | P67672 | P67672 | P36494 | Q4WHG1 |
| P30722 | Q6EU76 | P19417 | P57629 | Q47RV4 | Q47RV4 | P02541 | P72394 |
| Q9B6D5 | P30101 | P27657 | P46675 | Q82ZD6 | Q82ZD6 | Q7TYY6 | P30822 |
| P16476 | P80352 | O54759 | Q8Z9U4 | P30840 | P30840 | Q8TUV6 | Q8EIG9 |
| P42555 | Q7MH47 | Q9LN94 | Q6GH45 | Q4USF7 | Q4USF7 | P29510 | Q9PVW8 |
| Q8IBS5 | P02597 | Q5U8Z7 | O30509 | P23487 | P23487 | Q9K0N4 | Q5LRU0 |
| P05019 | P24369 | P42895 | Q49YB1 | P35644 | P35644 | P67970 | P28495 |
| Q9SNW7 | P92120 | P81942 | Q83F35 | O88454 | O88454 | Q99757 | P29777 |
| Q7U3C4 | Q9SVE5 | P19378 | P22523 | Q8EEF1 | Q8EEF1 | P05390 | Q8G0U3 |
| P29511 | P11343 | P00749 | P68490 | P06637 | P06637 | Q8FKI9 | P51691 |
| P52014 | P54423 | Q9YGI6 | Q89VC7 | P20973 | P20973 | Q6B345 | P56199 |
| Q9TUN1 | O93724 | O22476 | Q9V1Z8 | P75285 | P75285 | P12244 | Q6D001 |
| Q39786 | Q27177 | P08570 | Q62725 | P46892 | P46892 | Q6RG04 | Q65G93 |
| O66778 | Q505F5 | P18944 | Q9VCR7 | Q03142 | Q03142 | P27420 | Q6GIC6 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Q8I030 | P07477 | Q95086 | P58326 | Q8DPL5 | Q8DPL5 | Q9Z315 | P52685 |
| Q9LEA7 | Q03685 | P36953 | P82085 | P33203 | P33203 | P15193 | Q58064 |
| O83246 | Q07473 | Q9JLD2 | Q9JL15 | Q9PN51 | Q9PN51 | Q91516 | Q9PQX7 |
| O22862 | Q00398 | P02125 | P49425 | P21996 | P21996 | Q492B9 | P07095 |
| Q5AQL0 | P0C0I1 | O75690 | Q8PW40 | Q7NAN0 | Q7NAN0 | O93450 | Q7WC23 |
| Q61233 | P06607 | P29291 | O25626 | Q6A6Q8 | Q6A6Q8 | P16973 | O65352 |
| P56203 | P19825 | P68382 | Q8EXA3 | Q83G66 | Q83G66 | P40903 | Q92E89 |
| P09801 | P00635 | Q5JI69 | O46036 | Q8YHN7 | Q8YHN7 | P61889 | Q82CI4 |
| P25704 | P0AGD6 | Q8NKC2 | O35799 | P08948 | P08948 | P29865 | Q9Y5S1 |
| Q8HXP8 | Q7ZTA0 | O97162 | P04955 | Q8YN62 | Q8YN62 | P29063 | P18496 |
| Q9FLC0 | P04284 | P50673 | P08254 | O13747 | O13747 | P35035 | Q9M571 |
| P51408 | P48247 | P02625 | Q88L02 | Q32B26 | Q32B26 | Q9U4X3 | P47410 |
| Q43767 | Q7YT83 | P02604 | O78440 | Q9Y5M8 | Q9Y5M8 | Q723V9 | Q5R6T1 |
| P11589 | Q8CPY3 | P52270 | Q9HQD7 | P28154 | P28154 | P38507 | Q31E63 |
| Q898R0 | P77968 | Q8HXP6 | Q47EH3 | O78467 | O78467 | Q6G554 | Q99J56 |
| Q26422 | P04373 | P80691 | Q3J252 | Q4A9M9 | Q4A9M9 | P09582 | Q58737 |
| Q07954 | P13909 | P22973 | P39071 | O17800 | O17800 | Q9Y5Q5 | Q9YGL6 |
| P28023 | P80246 | Q9LVL1 | Q35101 | Q7YTU4 | Q7YTU4 | Q8D3K3 | P77569 |
| P48495 | P26728 | P52265 | Q7MQJ5 | P59761 | P59761 | P04072 | P31176 |
| Q9STX5 | Q9C0N4 | P29091 | P15770 | P28936 | P28936 | O95347 | Q89HW8 |
| Q9PMQ6 | Q7N9A4 | O59651 | Q75ET6 | Q99PA3 | Q99PA3 | Q6URB0 | Q8XX54 |
| Q9BDL1 | P43646 | P80579 | P47980 | Q9PK77 | Q9PK77 | Q6IE51 | P16213 |
| Q5PAB8 | O82246 | P12470 | Q3J5W1 | Q7M3C4 | Q7M3C4 | P98072 | Q90ZS6 |
| O02389 | Q7U0U6 | P47735 | P95806 | P40757 | P40757 | O97507 | Q7VQN1 |
| Q80W65 | O88799 | Q86SG5 | P0ADK4 | O60044 | O60044 | P07947 | P13788 |
| P60882 | P04354 | P16562 | P55166 | P12636 | P12636 | P09802 | Q5VV43 |
| Q3JP10 | P14533 | P05419 | Q5PDG2 | Q5LR15 | Q5LR15 | Q9HTJ2 | Q4JQI1 |
| Q65JE7 | P42322 | P30406 | P60375 | Q8ZHF1 | Q8ZHF1 | Q9UYX0 | Q8TH25 |
| Q91BH1 | Q63617 | P62184 | Q8R0N9 | P19002 | P19002 | P38660 | P91374 |
| P81423 | P07053 | P28768 | Q9NVF7 | P36395 | P36395 | Q65CZ5 | P41894 |
| O73955 | Q28506 | Q10427 | P55294 | Q8NH73 | Q8NH73 | P59035 | P25118 |
| P32070 | P22275 | Q6ITB0 | P35280 | Q9SXU1 | Q9SXU1 | Q99895 | Q9RLB6 |
| P49236 | Q5KA96 | Q8IJN7 | Q49647 | P10676 | P10676 | Q9CD33 | P07297 |
| Q8UW25 | Q38W93 | P0A8G8 | Q9PC77 | Q9NVD7 | Q9NVD7 | P05619 | P42218 |
| O46644 | P46573 | P00758 | Q9NQX5 | Q6F1W4 | Q6F1W4 | Q48CM6 | P17778 |
| P24669 | P37869 | Q08752 | Q45480 | O81395 | O81395 | P80067 | P57303 |
| Q9R1T3 | Q5R875 | P53440 | O93884 | P17614 | P17614 | Q23858 | Q12857 |
| Q8BXZ1 | Q9PYY5 | P04443 | P67032 | P43340 | P43340 | P05581 | Q89AL6 |
| P38669 | P06240 | O60760 | Q9KUM8 | Q9P4P9 | Q9P4P9 | P06733 | Q7VKB6 |
| P42636 | P20363 | P25304 | Q00664 | Q6CZV1 | Q6CZV1 | P68992 | Q8UDN3 |
| Q8BZ10 | Q9PUH3 | P17751 | Q9Z6U6 | P04445 | P04445 | P05570 | Q7MK12 |
| Q9QZE3 | Q7VW79 | Q63207 | Q9JKA5 | P65821 | P65821 | Q5R143 | P46066 |
| Q9H3N1 | P15964 | P83487 | O96008 | P40831 | P40831 | O76003 | P24336 |
| P78010 | P24032 | P29621 | P51731 | Q9T390 | Q9T390 | P02205 | Q8K941 |
| P11679 | Q86VQ3 | P27666 | Q9I3F3 | P44368 | P44368 | P23285 | Q7N119 |
| Q8G1Z9 | P62146 | O88947 | Q49XC1 | P04114 | P04114 | P07205 | P69406 |
| Q8WR63 | Q9HFQ3 | P05456 | Q6FDQ8 | P94364 | P94364 | P19352 | Q5WZ43 |
| Q9BS26 | Q8K9V4 | P20961 | Q5H185 | P25737 | P25737 | P50678 | P81433 |
| Q6VAG0 | P12543 | Q87LQ0 | Q9XAD5 | Q89WC9 | Q89WC9 | P21859 | P63255 |
| Q922U2 | Q4JXX6 | P11424 | P02365 | Q9X7A1 | Q9X7A1 | P18520 | Q9CXV1 |
| Q10057 | Q6A555 | P08553 | Q53FV1 | Q9K708 | Q9K708 | Q00871 | Q6EV70 |
| O93866 | Q28451 | P25778 | Q6DDL7 | P59175 | P59175 | P50719 | P84078 |
| Q7U804 | Q03700 | O34241 | P19334 | Q7W9J6 | Q7W9J6 | Q8BGI7 | Q9YEL4 |
| Q8RH05 | P50666 | P23605 | P34942 | Q5QYL5 | Q5QYL5 | P0AB73 | Q4QMG7 |
| Q39527 | P04189 | Q14651 | Q63T22 | Q5UNX2 | Q5UNX2 | P0AGD1 | P00827 |
| Q09430 | Q9FI31 | Q28983 | Q9P427 | Q822G7 | Q822G7 | O93609 | Q3J8S1 |
| P52406 | Q8LKZ1 | P42637 | Q6A6K4 | Q5GS47 | Q5GS47 | Q8SRI6 | O07145 |
| P22929 | Q9SP22 | P91798 | Q9Y678 | Q5XBZ4 | Q5XBZ4 | P35556 | Q04236 |
| P26730 | P54714 | Q8L899 | P0A8E7 | P49307 | P49307 | Q8IT89 | P03022 |
| O80622 | Q8V5U0 | P04784 | Q6P0I6 | Q6FXJ8 | Q6FXJ8 | P10591 | Q8YPK5 |
| Q01806 | P29498 | P37570 | Q11103 | Q99497 | Q99497 | Q6FRI3 | P09961 |
| Q9Z2Q3 | Q89AR7 | P62046 | P19024 | P03660 | P03660 | P04070 | Q884I1 |
| P42210 | P21848 | Q06805 | Q8Z234 | Q9Z9A7 | Q9Z9A7 | Q96VB9 | Q9Z3Z6 |
| O74864 | P79847 | P97861 | P43566 | Q928A5 | Q928A5 | P05617 | P40723 |
| Q6VAF6 | Q9JXT8 | P02227 | P16797 | P75156 | P75156 | P29143 | P22765 |
| Q08652 | Q4FQH3 | Q9RR60 | Q9BJK5 | P0AAR0 | P0AAR0 | P02206 | P60510 |
| P02616 | Q6HPR6 | P13539 | Q7VQQ0 | Q9AEP7 | Q9AEP7 | Q9UPN3 | P34730 |
| P69046 | Q9N1Q9 | P13869 | P0A997 | Q96DX7 | Q96DX7 | Q88DV4 | Q9KCB0 |
| Q47TI0 | P05208 | P29429 | Q09005 | P08021 | P08021 | Q27775 | P69004 |
| P06906 | P68372 | Q8YH68 | Q38162 | P16744 | P16744 | Q5HBT2 | Q5LRI7 |
| Q8LEK4 | Q37706 | P15455 | O84335 | P0A5S1 | P0A5S1 | P29516 | P23190 |
| P04259 | P26795 | P24364 | P33268 | Q8KFL5 | Q8KFL5 | Q90326 | P14332 |
| Q8KML6 | P34755 | Q61W58 | Q04687 | Q6CXK7 | Q6CXK7 | P17183 | Q8U9L4 |

Table E.3: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Q4URD2 | P31693 | Q9TUI5 | P70039 | P77624 | P77624 | P20730 | P55304 |
| P29421 | Q73HQ2 | P17820 | Q5SL87 | Q13642 | Q13642 | O97578 | Q5LXR8 |
| Q61759 | P10708 | P00330 | Q8Y347 | Q8PYQ1 | Q8PYQ1 | P04810 | Q8KTR0 |
| P53644 | P48722 | Q9Y7Q2 | P03865 | Q9P1Q0 | Q9P1Q0 | O75443 | Q7VKJ6 |
| Q08115 | P00442 | Q39043 | P0AER7 | P14434 | P14434 | P58515 | P0A9P9 |
| Q92PU3 | P29026 | P18291 | Q8BUR3 | Q608S3 | Q608S3 | Q8T115 | Q7WJ91 |
| Q9BEA9 | P27740 | P19647 | Q5VWZ2 | P04182 | P04182 | Q9D6F9 | P32529 |
| P09465 | O15540 | Q9Y3C6 | Q62I10 | Q58097 | Q58097 | Q8NU97 | Q9RUS2 |
| P00408 | P83363 | P28762 | P62867 | Q3J8T8 | Q3J8T8 | P07288 | P82284 |
| P54107 | O96347 | P11146 | Q8PK34 | Q6GE96 | Q6GE96 | Q4A0Q1 | Q8ES61 |
| Q9I6C8 | Q7VRP7 | Q40024 | Q6LVI5 | Q68X64 | Q68X64 | P79345 | P11667 |
| Q12548 | P51545 | O24332 | P35074 | P0AF23 | P0AF23 | Q9UKF2 | P01350 |
| P81637 | Q9BTN0 | Q9HGV0 | Q16928 | Q9GMA3 | Q9GMA3 | P07596 | Q9EQG3 |
| Q8F5I5 | O97176 | Q6UWN8 | Q9ZQ77 | P26459 | P26459 | P11147 | Q5RJK8 |
| P14851 | P43508 | Q8W474 | O83142 | Q8Y3P9 | Q8Y3P9 | P81558 | Q8FU05 |
| P81176 | P05584 | Q71U34 | Q9QZY7 | Q9T443 | Q9T443 | P14297 | P29535 |
| Q6GJC6 | P29001 | Q40374 | P07193 | O55060 | O55060 | P04159 | Q46560 |
| P02161 | P39688 | Q9SHD1 | Q4N4T9 | O14682 | O14682 | P62560 | P18101 |
| Q43804 | P65224 | Q6D0B7 | Q13495 | P17345 | P17345 | Q9S8P4 | O08400 |
| P56076 | P45853 | P16413 | Q60HE7 | Q10167 | Q10167 | P52253 | Q8G768 |
| O54763 | P15233 | Q98PG4 | P52000 | P02742 | P02742 | P07913 | P58374 |
| P17180 | P0A616 | Q39692 | Q7V9U7 | Q03222 | Q03222 | P02771 | P0ABV9 |
| Q92I65 | P83958 | Q9SJZ2 | Q5LWF3 | P37079 | P37079 | Q8LDI5 | Q9V3A4 |
| Q967W0 | P15950 | P05594 | P17796 | Q8WEW3 | Q8WEW3 | Q83B44 | P26583 |
| Q9Y927 | Q05000 | P13115 | Q589A4 | Q3K407 | Q3K407 | P15456 | Q97HE5 |
| P12036 | P41827 | O13401 | Q9GK30 | Q8DK13 | Q8DK13 | O16116 | Q8PNR4 |
| Q03699 | Q86WD7 | P46711 | P15358 | Q88MG3 | Q88MG3 | P51555 | Q493D9 |
| Q9C8G4 | P54407 | P46252 | Q5KU05 | Q94516 | Q94516 | Q60HG6 | Q5F5K6 |
| P80079 | P0A0K5 | Q9F2Q3 | Q5JJ91 | Q6BSY5 | Q6BSY5 | Q62472 | P16020 |
| O18879 | Q7LZM3 | Q5R5B6 | Q9HMY5 | P25740 | P25740 | Q9SZA7 | Q96CJ1 |
| P02634 | Q9L4K0 | Q43594 | Q72LS1 | Q5N3U9 | Q5N3U9 | Q9SI20 | P40006 |
| P07389 | Q9S5A4 | O01812 | P65827 | Q748X2 | Q748X2 | Q43495 | Q53462 |
| Q09151 | Q92IA0 | P98021 | Q09556 | O83792 | O83792 | Q5SME1 | P11440 |
| P69951 | Q74F05 | P35031 | Q5PAI9 | Q921D9 | Q921D9 | P25417 | Q6LUJ4 |
| Q883Y9 | P05601 | P00409 | P0AA08 | P0AEC9 | P0AEC9 | Q90325 | Q57QH7 |
| Q8G0F7 | Q9CDE9 | P82007 | Q8P2K1 | Q8X5V2 | Q8X5V2 | Q64057 | P10205 |
| O01359 | P69049 | P12815 | Q58827 | Q68RJ7 | Q68RJ7 | P02124 | P40689 |
| P81859 | P68873 | Q60BA1 | O29734 | Q9Y2K1 | Q9Y2K1 | P51977 | P0C069 |
| Q06248 | P04187 | Q43304 | P41678 | Q7A5Y4 | Q7A5Y4 | Q95191 | O29534 |
| P37073 | Q06445 | O48818 | P07774 | P41267 | P41267 | P26726 | P97297 |
| P80249 | Q9SI17 | Q62177 | Q52698 | Q8XV35 | Q8XV35 | P98035 | O67623 |
| Q9FJR1 | P07851 | P02687 | O83220 | Q8XLK2 | Q8XLK2 | Q9NY65 | Q9KVD6 |
| Q8PAK9 | Q9FL80 | Q74IT6 | Q933K8 | Q00578 | Q00578 | O65768 | P52663 |
| P07901 | Q9WVB4 | P46697 | Q81WJ6 | Q722H1 | Q722H1 | P14625 | P41207 |
| P53691 | Q8A0U2 | P04692 | Q21815 | P50359 | P50359 | Q16772 | Q72NU6 |
| Q9U9J4 | P18254 | Q9M4S8 | P0A4J9 | P21932 | P21932 | P19228 | P0A2R6 |
| P08414 | Q9NJP0 | Q05927 | Q5WVI1 | P10715 | P10715 | Q4IBK5 | Q57IW5 |
| O24174 | P18292 | P21107 | P07398 | P21342 | P21342 | Q9Z0J0 | P42321 |
| Q05981 | Q8S4P5 | P02691 | Q8UBS0 | Q9X9J5 | Q9X9J5 | Q9PUH4 | P12235 |
| Q7XJ02 | O78682 | Q7S2Z9 | Q9Y5I2 | O96004 | O96004 | P02589 | Q93EU6 |
| P68015 | P09942 | O64392 | P44050 | Q50634 | Q50634 | P26643 | Q5HM84 |
| P45852 | P29451 | P48060 | Q8BDD8 | Q9BY12 | Q9BY12 | O22348 | P14184 |
| P21797 | Q9SVQ6 | P41979 | Q6HA24 | P08278 | P08278 | P22010 | P54286 |
| Q60767 | Q99M73 | Q39366 | P57654 | Q64PY6 | Q64PY6 | Q4W1I9 | O51933 |
| P25050 | Q7V7D2 | Q37419 | P78411 | P19192 | P19192 | P44499 | P43853 |
| O74286 | P50258 | P47771 | Q7M8E0 | Q9JZI5 | Q9JZI5 | P51546 | Q6DE96 |
| Q39KH5 | P22851 | Q9V3G3 | P61604 | Q3KLI5 | Q3KLI5 | Q8Z2F4 | P84826 |
| Q9P0G3 | P05049 | P02782 | Q64QR7 | P41378 | P41378 | P08009 | Q7UWN8 |
| P04185 | P17641 | Q940Q1 | P22130 | P66410 | P66410 | Q8ERF4 | Q5HW83 |
| Q7YS85 | O96102 | Q6BZH1 | P43626 | Q72HS8 | Q72HS8 | P31726 | P35449 |
| Q9SKP6 | Q9FJK8 | Q9NFZ6 | Q8W9F5 | P27387 | P27387 | P50690 | P10865 |
| P41887 | P00524 | Q7MA77 | P27033 | P27599 | P27599 | P84527 | P04407 |
| Q9ZP06 | P92721 | Q4WCR3 | Q99YE2 | Q9CLR1 | Q9CLR1 | Q5PXY7 | O70274 |
| O61308 | P81881 | Q90WW4 | P27680 | P66353 | P66353 | P43652 | Q9XPP0 |
| P35338 | Q45FF9 | P34932 | P51280 | Q8N488 | Q8N488 | P79815 | P69095 |
| Q5FNN5 | P29373 | P48595 | Q9TTE2 | Q8ZJ89 | Q8ZJ89 | Q7XBQ9 | Q4FR05 |
| P02669 | P18165 | Q08863 | P28670 | Q8Z9E1 | Q8Z9E1 | Q06931 | P41070 |
| Q6ICZ8 | O16797 | P22302 | P41606 | P57046 | P57046 | Q9SEU6 | P69770 |

# E. LIST OF ALLERGEN AND NON-ALLERGEN PROTEIN SEQUENCES

Table E.4: Uniprot IDs for Non-Allergens Training Set - Allerhunter vs Uniprot Database

| ID | ID | ID | ID | ID | ID | ID | ID |
|---|---|---|---|---|---|---|---|
| P31947 | Q9Y2D5 | P54253 | P07711 | O43633 | O43633 | P40394 | P49454 |
| P27348 | Q99996 | Q8WWM7 | Q9UBR2 | Q9UQN3 | Q9UQN3 | P00334 | Q13352 |
| P63104 | Q12802 | Q99700 | Q03135 | Q9BY43 | Q9BY43 | Q15848 | Q96BT3 |
| P30443 | Q9ULX6 | Q96GD4 | P56539 | Q9H444 | Q9H444 | Q60994 | Q71F23 |
| P01892 | P39010 | Q96247 | Q95RV2 | Q9Y3E7 | Q9Y3E7 | Q9H2P0 | Q20646 |
| P01891 | Q8INB9 | P02701 | Q39253 | Q12114 | Q12114 | Q94BT6 | Q53EZ4 |
| P01889 | P31749 | P31414 | P17106 | Q9GZX3 | Q9GZX3 | Q9NRN7 | P00450 |
| P30464 | Q38898 | O15169 | P33322 | P01555 | P01555 | P07248 | O48946 |
| P03989 | P31751 | Q9UPN4 | P08185 | P01556 | P01556 | P07550 | Q941L0 |
| P30685 | Q9Y243 | P00282 | Q13191 | P00794 | P00794 | Q16186 | Q9SWW6 |
| P30475 | Q96B36 | Q07817 | P22681 | Q99828 | Q99828 | P12235 | O13297 |
| Q04826 | P02768 | O43521 | P22682 | O75339 | O75339 | P00257 | P41208 |
| P18464 | P05091 | P61769 | P00730 | Q9P2M7 | Q9P2M7 | Q640N1 | P11597 |
| P10321 | P04075 | P02730 | P15085 | Q15642 | Q15642 | P55197 | P00751 |
| Q14738 | P00883 | P08037 | P48052 | Q9LDI3 | Q9LDI3 | P55196 | P00746 |
| Q13362 | P05062 | P15291 | Q92793 | P47001 | P47001 | Q9UHB7 | P08603 |
| Q38845 | P15121 | P56817 | P16152 | P0ABH7 | P0ABH7 | P06280 | P05156 |
| P30153 | P07764 | Q9Y5Z0 | P35520 | Q96SN8 | Q96SN8 | Q9UPQ3 | O15519 |
| P30154 | P0AB71 | P02945 | P83916 | Q8WWK9 | Q8WWK9 | Q99490 | P13569 |
| P04229 | P14540 | O75531 | P83917 | P25859 | P25859 | P52594 | P01233 |
| P01912 | Q9UM73 | Q99933 | Q13185 | Q7Z460 | Q7Z460 | O04379 | P0A6F5 |
| P13760 | Q96Q42 | O95817 | O00257 | O75122 | O75122 | Q9UKV8 | P10809 |
| Q30134 | P02760 | P46379 | P45973 | Q9NBD7 | Q9NBD7 | O00253 | O14646 |
| Q95IE3 | Q99217 | Q9UQB8 | P92973 | Q9UJ71 | Q9UJ71 | Q8N157 | P32657 |
| Q5Y7A7 | P46883 | Q94F62 | Q9H6F5 | Q9H2X3 | Q9H2X3 | Q9C5U2 | Q12873 |
| Q9GIY3 | Q01433 | Q92560 | Q16204 | Q9BXN2 | Q9BXN2 | Q9C5U1 | Q14839 |
| P01911 | P50579 | Q99728 | P13500 | Q9UQC9 | Q9UQC9 | Q9C5U0 | Q9P2D1 |
| Q29974 | P15144 | P35613 | P10147 | P37019 | P37019 | Q09666 | Q9HCK8 |
| P26439 | P30533 | P18572 | P13236 | P35523 | P35523 | P38013 | Q3L8U1 |
| HSD3B2 | P54144 | Q07812 | P13501 | P51795 | P51795 | P35869 | P07363 |
| P78314 | P69681 | Q9NRL2 | P14635 | P51798 | P51798 | P30561 | P0AE67 |
| P29372 | P04745 | Q9UIG0 | P24385 | Q00610 | Q00610 | P55008 | Q96EP1 |
| P11171 | P0C1B3 | Q9UIF9 | P30281 | O00299 | O00299 | O95831 | Q9UNE7 |
| Q13541 | P10538 | Q96RK4 | P24864 | Q9Y696 | Q9Y696 | Q12904 | O14757 |
| Q9NRA8 | P04746 | P50895 | P51946 | P30622 | P30622 | O43918 | Q6DE87 |
| P08195 | P00690 | P56945 | O75909 | P49759 | P49759 | P14550 | O96017 |
| P28335 | P06278 | Q9P287 | Q9UK58 | P49761 | P49761 | Q04828 | P70232 |
| P46098 | P39687 | O95999 | O60563 | Q13286 | Q13286 | P52895 | P06276 |
| Q9H0P0 | P54802 | P10415 | Q8ND76 | P63284 | P63284 | P42330 | P00183 |
| P02763 | P10275 | Q86UU0 | P00431 | Q9HAW4 | Q9HAW4 | P17516 | P17927 |
| P01009 | P01160 | Q9NYF8 | P31384 | P10909 | P10909 | Q02952 | P20023 |
| P08697 | P03950 | Q6W2J9 | P51681 | O80809 | O80809 | P15848 | P43320 |
| P01023 | P01019 | P11274 | Q6YHK3 | P05059 | P05059 | P18440 | P16220 |
| P05067 | P16157 | Q9Y276 | Q9UQ88 | P10645 | P10645 | P11245 | Q03060 |
| P12023 | Q01484 | P35817 | P21127 | P09543 | P09543 | Q13510 | P07315 |
| P08592 | Q8C8R3 | P23560 | P08571 | O81445 | O81445 | Q9Y294 | P07320 |
| Q15758 | Q8IWZ3 | A6H8Y1 | P06126 | Q15021 | Q15021 | P32447 | P46108 |
| P01011 | Q9V4P1 | O76090 | P29016 | Q15003 | Q15003 | Q9VW15 | Q64010 |
| Q2M2I8 | Q9NQW6 | P00722 | P15813 | Q9BPX3 | Q9BPX3 | Q9UBL3 | P0ACJ8 |
| Q9UGJ0 | Q99873 | P16278 | P15812 | P42695 | P42695 | P17405 | P02741 |
| Q13131 | O14744 | P16442 | Q9NNX6 | Q6IBW4 | Q6IBW4 | P00805 | Q64735 |
| P54646 | Q96LA8 | Q15582 | Q5ZPR3 | Q16281 | Q16281 | P20933 | Q6UUV9 |
| Q9NY61 | P20594 | P37702 | P10747 | Q14028 | Q14028 | Q8IZT6 | Q53ET0 |
| P00509 | O75179 | P08236 | P42771 | P34972 | P34972 | Q13625 | Q6UUV7 |
| Q9ZR72 | P01008 | Q9UHR4 | Q8N726 | P02452 | P02452 | P00966 | Q96318 |
| Q9C8G9 | P02833 | P53004 | Q64364 | P11087 | P11087 | P05023 | P82279 |
| Q94FB9 | Q9H6X2 | Q9Y6D5 | Q9Y5K6 | P02454 | P02454 | P16615 | Q43125 |
| Q9XIE2 | P58335 | O00499 | O95400 | P02467 | P02467 | P20020 | O77059 |
| Q9HC16 | Q38914 | P12995 | P06729 | P08123 | P08123 | O43520 | Q96524 |
| APOBEC-3G | P04083 | P12996 | P16671 | P02458 | P02458 | Q6PL18 | P02489 |
| O95477 | P07355 | P13000 | P28907 | P06681 | P06681 | P15336 | P02511 |
| P78363 | P12429 | O15392 | P07766 | P02461 | P02461 | P17544 | O75534 |
| O95342 | P08758 | Q96CA5 | P20963 | P01024 | P01024 | P53104 | Q9Y2W7 |
| Q09428 | P08133 | P62593 | P29965 | P01027 | P01027 | Q9Y4P1 | Q9QXT8 |
| P33897 | Q16853 | Q01532 | P16070 | P02462 | P02462 | P38182 | P07333 |
| Q8NE71 | P21397 | P43583 | P15379 | P02463 | P02463 | Q8WXF7 | P09603 |
| P45844 | P27338 | P54132 | Q08722 | P08572 | P08572 | Q13315 | P15509 |
| Q9UNQ0 | P35631 | P30043 | P01730 | Q01955 | Q01955 | P38110 | Q99062 |
| Q8IZP0 | P17427 | P35226 | P13987 | P53420 | P53420 | P54259 | P01243 |
| Q8CBW3 | P63010 | Q13873 | P11912 | P29400 | P29400 | P0AB98 | P68400 |
| A0MES8 | Q96CW1 | P36894 | P11911 | P0C0L4 | P0C0L4 | P00846 | P67870 |
| Q9SJN0 | P84092 | P41832 | P40259 | P0C0L5 | P0C0L5 | Q04656 | Q8WXE0 |

Table E.4: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P00519 | P35632 | Q8NFC6 | P48960 | P01029 | P01029 | P35670 | O14936 |
| P00520 | O14727 | Q14137 | P21926 | P20908 | P20908 | P00829 | P41240 |
| P42684 | Q9H1A4 | Q53HL2 | Q96GN5 | P05997 | P05997 | P68699 | Q13098 |
| O14639 | P25054 | Q00587 | P32458 | P01031 | P01031 | Q8WXE1 | Q9UNS2 |
| Q8K4G5 | P27695 | P50747 | P32797 | P12109 | P12109 | O75882 | Q92905 |
| P15891 | Q9BZZ5 | P00974 | Q13042 | P12110 | P12110 | P46100 | P13611 |
| Q13085 | P02647 | Q12830 | Q12834 | P12111 | P12111 | Q9VXG8 | Q00657 |
| Q5SWU9 | P02652 | P15056 | P26309 | P13671 | P13671 | Q13535 | P00183 |
| Q00955 | P06727 | P38398 | Q9UJX2 | Q02388 | Q02388 | P38111 | P17927 |
| P11310 | Q6Q788 | P51587 | P16522 | P07357 | P07357 | Q76LX8 | P20023 |
| P49748 | P04114 | P46736 | P30260 | P07360 | P07360 | P00918 | P43320 |
| Q5T8D3 | P02655 | O95696 | P06704 | P02748 | P02748 | P07451 | P16220 |
| P07108 | P05090 | P25440 | Q16543 | Q03692 | Q03692 | P22748 | Q03060 |
| P55926 | P02649 | Q15059 | P60953 | P13942 | P13942 | Q16790 | P07315 |
| Q96AP0 | P02749 | O60885 | P60766 | O43405 | O43405 | P01258 | P07320 |
| Q9BYF1 | O95445 | Q9NPI1 | P25694 | P23528 | P23528 | P62157 | P46108 |
| P0A9G6 | Q7Z2E3 | Q9H0E9 | P07834 | Q9H9E3 | Q9H9E3 | P62152 | Q64010 |
| P22303 | P07741 | Q5VTR2 | Q99459 | Q9UMD9 | Q9UMD9 | P62158 | P0ACJ8 |
| P04058 | P29972 | Q9NXR7 | Q6P1J9 | O04197 | O04197 | P62161 | P02741 |
| Q10714 | P41181 | O22476 | O00311 | P39060 | P39060 | P27797 | Q64735 |
| P12821 | P60844 | P55201 | Q69YH5 | P39061 | P39061 | P27824 | Q6UUV9 |
| P32297 | P10398 | Q10589 | Q9H5V8 | P38432 | P38432 | P35564 | Q53ET0 |
| P02708 | P25098 | Q06187 | P0ABF6 | P01190 | P01190 | Q08AD1 | Q6UUV7 |
| Q9UKV3 | P63345 | P06129 | Q9NYV4 | P01189 | P01189 | P20807 | Q96318 |
| Q07912 | P84077 | O60566 | Q14004 | P46946 | P46946 | Q86VP6 | P82279 |
| P53396 | P62330 | O43683 | O94921 | Q8N668 | Q8N668 | Q01518 | Q43125 |
| P21399 | Q8N6T3 | P47136 | Q00536 | P49747 | P49747 | P20160 | O77059 |
| P0A6A8 | Q9NP61 | P48524 | P06493 | P21964 | P21964 | P00864 | Q96524 |
| Q9LX29 | P93022 | P23293 | P24941 | P43254 | P43254 | P13773 | P02489 |
| P31224 | P05089 | P10845 | Q00526 | P53621 | P53621 | P0A6F1 | P02511 |
| Q01574 | Q92888 | P15538 | P11802 | P00395 | P00395 | P00968 | O75534 |
| P07830 | Q92974 | P19099 | Q00535 | O43303 | O43303 | Q9WVG6 | Q9Y2W7 |
| P62736 | Q15052 | Q86VB7 | P50613 | P05093 | P05093 | P02666 | Q9QXT8 |
| P60709 | Q14155 | Q07021 | P50750 | Q24478 | Q24478 | O15234 | P07333 |
| P12814 | Q9ES28 | P02746 | P24100 | P11511 | P11511 | Q8NG31 | P09603 |
| P35609 | O15085 | P00736 | O76039 | P04798 | P04798 | P02668 | P15509 |
| O43707 | Q9NZN5 | P09871 | P38936 | P05177 | P05177 | Q14511 | Q99062 |
| P05095 | Q5SW96 | P11586 | P46527 | Q16678 | Q16678 | P29466 | P01243 |
| P68133 | O14497 | Q6PHS9 | Q06850 | P08686 | P08686 | P42574 | P68400 |
| P60010 | Q8NFD5 | Q9NZU7 | Q38872 | P11509 | P11509 | Q14790 | P67870 |
| P37023 | Q99856 | O00555 | Q9H211 | P20813 | P20813 | P55211 | Q8WXE0 |
| P45381 | Q4LE39 | Q13936 | Q9Y232 | P10632 | P10632 | Q92851 | O14936 |
| Q13444 | P36404 | O60840 | Q7Z7A1 | P11712 | P11712 | P41180 | P41240 |
| P35348 | Q9H0F7 | Q13698 | Q5SW79 | P33261 | P33261 | P04040 | Q13098 |
| P08913 | P04424 | Q02641 | O15078 | P10635 | P10635 | P07858 | Q9UNS2 |
| Q13443 | P77398 | Q9H251 | Q5VT06 | P05181 | P05181 | P53634 | Q92905 |
| P56658 | P0A4Z2 | P12830 | P13688 | P08684 | P08684 | P07339 | P13611 |
| P06134 | Q8LGE3 | P09803 | P06731 | P20815 | P20815 | P14091 | Q00657 |
| P00813 | P17870 | Q9BY67 | Q9XWD6 | Q02928 | Q02928 | Q9UBX1 | |
| P03958 | P49407 | Q8R5M8 | Q92879 | Q6FI81 | Q6FI81 | P08311 | |
| P35611 | P29066 | Q13111 | O95319 | O43809 | O43809 | P43235 | |
| Q9UEY8 | P32121 | Q13112 | P19835 | P31327 | P31327 | P07199 | |
| P00325 | P29067 | Q24572 | P49450 | P50416 | P50416 | Q03188 | |
| P00330 | P15289 | P00915 | P36012 | P23786 | P23786 | Q02224 | |

# E. LIST OF ALLERGEN AND NON-ALLERGEN PROTEIN SEQUENCES

Table E.5: Uniprot IDs for Allergens Independent Data Set

| ID | ID | ID | ID | ID | ID | ID | ID |
|---|---|---|---|---|---|---|---|
| O46206 | O49860 | O18535 | P18632 | AAB30829 | Q9XF37 | Q9U3U5 | CAA59340 |
| Q6QHU2 | Q8T379 | Q8GSC5 | P02672 | P38949 | Q5TIW1 | P32936 | P01552 |
| 1G5U_B | BAA09633 | CAA27571 | P43184 | P18153 | O81092 | Q9XG86 | O97192 |
| O23752 | Q84ZX5 | Q5EZB4 | JC5475 | Q39430 | P49822 | Q8GT40 | Q9NAS5 |
| JE0227 | P31757 | Q9HF12 | AAB32842 | Q94508 | Q13845 | 2122374B | Q9U1G2 |
| Q9LLB7 | P49370 | Q9TWR0 | Q23939 | Q40280 | S13614 | P49455 | Q42661 |
| P02754 | Q9SCI1 | H44583 | Q8GT39 | P43237 | AAA34257 | S65144 | P15322 |
| Q39431 | Q84RR5 | P14292 | CAA39880 | Q8WQK5 | P39675 | AAA68882 | AAA20067 |
| AAB23303 | AAA34258 | AAA34275 | Q941R0 | P15252 | Q9LEI9 | Q5EZB0 | CAA44345 |
| P82946_2 | P58171 | AAD47382 | AAB31957 | Q5EZA0 | Q9M7M8 | AAB32652 | P12993 |
| E53240 | P35775 | P46075 | CAA74694 | P00698 | Q9BIX2 | CAA45777 | Q84MM5 |
| Q5EZA8 | P23472 | AAB35084 | O18530 | CAA54819 | P08835 | P24627 | Q8MVU3 |
| Q9JJH9 | Q7M1X8 | Q42499 | O82015 | O24172 | Q6VPT9 | Q9T0M8 | |
| Q90YK9 | Q41260 | Q6QHU1 | Q40962 | P35081 | AAF65313 | P28296 | |
| O04725 | Q8TCD2 | Q01940 | P07380 | P82616_2 | O46212 | P68436 | |
| P81241 | Q5XWE1 | CAA27588 | P82946 | P80207 | P81826 | Q8J1L4 | |
| AAC49648 | Q9FPK2 | AAT80664 | Q9XF42 | AAB23464 | CAA54485 | Q7M3Y8_3 | |
| Q5EZB2 | P46419 | P27631 | Q6VPU6 | Q5EZB6 | Q9U5P1 | Q5EZ76 | |

Table E.6: Uniprot IDs for Non-Allergens Independent Data Set

| ID | ID | ID | ID | ID | ID | ID | ID |
|---|---|---|---|---|---|---|---|
| P02167 | P37228 | Q9UXR0 | P81517 | O59045 | Q9MUM2 | P68132 | Q6HF98 |
| P42269 | P02601 | P29293 | Q8LDP4 | Q83AJ2 | Q4QNR3 | P05087 | P53686 |
| P25871 | Q8D228 | P53513 | Q9GKX7 | Q5L0N0 | P08463 | Q5FVI3 | Q4WT34 |
| P33624 | Q6MB26 | P54108 | P27491 | P14652 | P14243 | P96377 | P26448 |
| Q5ASQ0 | P68080 | P80031 | Q9C566 | Q8GH68 | Q9RWC7 | P08058 | P61223 |
| Q8IUD2 | Q9DGJ1 | P00997 | P12069 | Q9L6P1 | P11986 | O52633 | O82827 |
| Q9SZY1 | Q9L9I0 | Q8K1K6 | Q86RN8 | Q57JA2 | P08022 | Q92249 | Q9UBT2 |
| P02047 | P51879 | P12329 | Q9NSB2 | Q8P550 | P48419 | P08217 | Q8MXI4 |
| P81660 | Q7KQL8 | Q25088 | P51781 | P81755 | O83770 | P69888 | P29755 |
| P16291 | Q93WF1 | Q37718 | P67944 | Q9M290 | Q2MI83 | Q7M3V3 | Q6R2R2 |
| P66829 | P31108 | Q9M8Z8 | Q9LT77 | Q8G2L7 | Q6GEJ0 | P05577 | P17984 |
| O73872 | Q8G0G3 | P03953 | P16627 | P68547 | Q9PE46 | P19527 | P65111 |
| P26413 | Q66DP8 | P22792 | P09243 | P17730 | Q9XSB8 | P23904 | P66980 |
| P08905 | P10649 | Q03902 | Q7SXW3 | Q9D168 | Q8MJ47 | P02059 | Q90523 |
| P68082 | P35795 | Q3BVB8 | P42373 | P35070 | Q9QZ25 | P06606 | P03998 |
| P29245 | P42374 | O51401 | Q4P235 | Q39011 | P33968 | P27522 | P75275 |
| P34930 | Q02200 | Q9LRM5 | Q41112 | P56297 | Q9V726 | P24722 | Q8N7C0 |
| P41148 | P69753 | P04727 | P26456 | P62143 | Q5HIQ6 | Q40665 | P64854 |
| Q02157 | P50691 | P81214 | P06741 | Q6L201 | P37508 | P07444 | P35391 |
| P49237 | P04347 | P30412 | P19236 | Q7N589 | P63713 | Q21355 | P55549 |
| P14841 | Q9Z520 | P08019 | P00198 | Q59060 | Q89AE0 | O35453 | Q5UR58 |
| Q9BN10 | P27521 | O60575 | Q6FMG7 | Q8Y7B6 | P59920 | P16394 | Q65ZZ4 |
| P05156 | O43109 | P05937 | Q88AE7 | P57801 | Q5YY83 | P09212 | Q8FJB7 |
| Q06548 | P06750 | P09542 | Q7M8C4 | P16649 | P47228 | P51902 | Q9CG33 |
| P19849 | P19171 | Q82TV8 | P12411 | Q83HC7 | P46218 | P51647 | P0AG61 |
| Q4PBY6 | Q9LDI3 | Q9ZEE0 | Q70Q35 | O14958 | Q49WX5 | P04988 | P67630 |
| P92133 | P52013 | P93447 | O80327 | Q28019 | Q9UYR5 | P17444 | P56383 |
| Q9QYM9 | Q8HZ59 | Q9CM49 | Q8CFM6 | Q8NH93 | Q66G58 | P82778 | O43819 |
| P06959 | P10246 | Q6HP91 | P00762 | P92564 | Q8FD50 | Q9C401 | P57258 |
| P20289 | P11827 | P82188 | P55737 | P03903 | Q19375 | Q42460 | P75238 |
| P98031 | P08331 | P83594 | P43373 | Q91605 | P42748 | Q8VWY6 | O58687 |
| P21783 | Q4QJW4 | P01090 | Q9QZ76 | Q9WZQ8 | P23103 | P01319 | Q62L77 |
| Q89KU3 | Q3LI67 | P48493 | Q4L919 | Q7SBR3 | Q82XA0 | P68279 | Q8D2J1 |
| Q9UBX1 | P42279 | P29517 | Q9MZF4 | Q7WME6 | Q9PKX1 | Q5U907 | Q60FY0 |
| Q9N2G9 | P24988 | Q43358 | P18241 | Q85FW2 | Q8YXH5 | O06653 | Q96I25 |
| O80995 | Q9LXW3 | P09006 | Q5GZV7 | Q4UMR5 | Q91316 | Q8K9B2 | P37958 |
| Q29092 | P37743 | Q01595 | P52260 | Q8YZQ3 | P0A621 | P98094 | Q65RX0 |
| P08144 | P35174 | P17505 | Q09675 | P31782 | Q67TJ2 | P04906 | P25734 |
| P22998 | P49935 | P09107 | P17156 | Q87XL0 | Q720A1 | P35835 | Q3BAH6 |
| P12437 | Q60AK3 | Q28222 | P54826 | Q7MM54 | Q9F0R1 | Q96460 | P19142 |
| P54202 | P04110 | Q24789 | Q57H69 | Q8EQZ8 | Q5HF87 | P05109 | Q9Y6M5 |
| Q43875 | Q99405 | Q42952 | Q9FL76 | P25600 | Q8G6J8 | Q24798 | Q12622 |
| P04957 | Q9ZMM2 | P54628 | P11684 | Q9I013 | Q8KBF4 | Q6BXM5 | Q56254 |
| P84346 | O30974 | Q02970 | P19859 | Q98A73 | Q6ZLA3 | P13436 | P68056 |
| Q74H59 | Q7XA40 | P10787 | P16545 | P04659 | Q02193 | P07171 | P16519 |
| Q9Z319 | Q9XBR5 | P12863 | Q5R6F7 | Q9VLA1 | Q04372 | P14750 | Q571F8 |

Table E.6: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P51459 | Q90339 | Q26563 | Q9TLT1 | P0AEE3 | P0A0Y7 | Q9W7R2 | Q9I2W7 |
| P16053 | Q9UNI1 | Q18688 | Q05423 | P64191 | O54862 | Q39445 | P00648 |
| P05941 | Q8IWT6 | P98030 | O15393 | Q8CP01 | P67767 | Q09092 | Q6LU53 |
| Q9NFZ7 | P81993 | P09967 | Q69AB1 | Q9BDN4 | Q5GRY9 | P78504 | P08493 |
| P55325 | Q17967 | Q9Z9M2 | Q65VZ7 | P41717 | P43332 | P41951 | P64764 |
| Q28944 | O08800 | P83578 | Q8JGT9 | Q99FY3 | Q5WHM7 | Q9SY33 | Q6GJI3 |
| P07214 | P13744 | P02585 | Q00446 | Q5RCA4 | Q05587 | Q8ZJK9 | Q9ZGW3 |
| P46369 | P25775 | P59671 | Q9ES45 | O95070 | Q8KCS2 | P82724 | Q22918 |
| O42724 | Q8EPE2 | P29500 | P32937 | Q8BHL8 | Q7Z494 | P36304 | Q64255 |
| Q9XSC6 | O74187 | Q5HV33 | Q8E0H0 | Q80DT0 | Q732N3 | P50423 | Q6PI62 |
| P32733 | Q43735 | P29531 | Q9NY56 | O42354 | P12404 | P28783 | O48410 |
| P08306 | P05934 | P23239 | Q81LW0 | Q5NHV4 | Q06069 | P17066 | Q3B6M7 |
| Q9FYF7 | P41184 | P15090 | P25307 | P0A5E2 | P09915 | P39451 | Q24524 |
| P05964 | P34931 | P29109 | O96790 | Q4I061 | P37363 | P35415 | P01178 |
| O35660 | Q9NQ38 | P02224 | Q9LLL8 | Q5RFN3 | Q71UG0 | Q8CX68 | P02826 |
| Q89CK8 | P13601 | Q83BQ3 | P09856 | O29548 | P75109 | O47428 | P66635 |
| Q8ETY8 | Q7ZZN9 | P30151 | Q42431 | Q8Y3T4 | Q92JM8 | P66943 | Q9P378 |
| O77811 | P68390 | O05700 | O88181 | Q944H0 | P12673 | P16252 | Q7MAZ9 |
| O60087 | P98044 | Q9FH83 | O94273 | Q96YA4 | Q9R1T7 | Q9ZRB0 | Q23529 |
| O15041 | P48740 | Q9NS15 | Q28794 | P47027 | Q87QV1 | P09224 | O35385 |
| P52855 | Q43779 | P28760 | Q4UDU8 | Q5WTF5 | Q7LBR1 | P32877 | Q75AH9 |
| Q5R8E8 | Q8PPG7 | Q9TSP2 | O87777 | Q9KP47 | Q9RCA1 | Q43534 | O55102 |
| P00406 | P25843 | O23237 | Q5H186 | P11292 | Q7V568 | Q9LF54 | P39247 |
| Q12794 | O04996 | Q5PK93 | P21199 | O51931 | Q98KC1 | P24789 | P53252 |
| P09653 | P00984 | Q9Z9C4 | Q52RN5 | Q88YP9 | Q90660 | Q01490 | P37002 |
| Q4PCH8 | P27518 | P31393 | Q7ZT99 | Q7MNN7 | Q33800 | Q06016 | Q6GGD5 |
| P18868 | P23883 | O88281 | P68392 | Q4FS54 | Q10164 | Q9ZNY3 | Q6PZD9 |
| Q46YS9 | Q7N0P4 | P23490 | P05615 | P51776 | Q98FG0 | P81709 | Q9PJM1 |
| Q9LHB9 | P25326 | P08106 | P29446 | Q20932 | O13712 | P67799 | P01236 |
| Q8DR29 | P09605 | O65399 | P21567 | Q8KER4 | O35568 | P19666 | P0ABS6 |
| Q8CTD5 | Q9NWZ3 | Q28923 | P52244 | Q9EQ28 | Q881U0 | Q24560 | Q9JXF5 |
| O77020 | P09733 | Q8HXS3 | P30112 | P57809 | Q87FT1 | Q8P5D8 | O66772 |
| O47675 | P09655 | Q37604 | Q9P926 | O74315 | P0A966 | Q10717 | Q8CSX7 |
| P39085 | P0A4L1 | P01034 | Q6CL78 | O30175 | Q8RT67 | Q23763 | P17310 |
| P34368 | Q92044 | P68508 | Q8R4Y4 | P96069 | Q7VH57 | Q9P7P7 | Q06010 |
| P35045 | P36952 | Q9K5Z5 | Q9YGJ2 | O34967 | P39342 | P28161 | P55959 |
| Q02844 | Q29145 | Q707X3 | P46587 | Q5LNU7 | P55780 | Q9CQ19 | Q92JI7 |
| Q06655 | P22133 | P10782 | Q88VW2 | Q3JWH7 | P77700 | Q6QNF4 | Q9PH36 |
| P83595 | P01001 | Q43873 | P18856 | P01621 | Q81W16 | P18172 | O95922 |
| Q4ING3 | P91253 | P80646 | Q04432 | Q692W3 | Q8G8Y2 | Q9Y4C1 | P54484 |
| Q03376 | P91902 | Q29426 | P05687 | P37700 | P94632 | P0C088 | P44477 |
| P01337 | Q9Z2H4 | P02204 | P25765 | P0A4A6 | Q7TV16 | Q98MZ3 | P25514 |
| P11948 | P48673 | P07436 | P03520 | Q8U261 | Q49YH9 | P09324 | Q9RS39 |
| Q9HTJ1 | Q7VVY2 | P02854 | P48720 | Q5KWJ2 | Q9CMY2 | Q9FRX4 | P13591 |
| Q9JHJ7 | Q96522 | P02154 | P24031 | O54751 | Q966L8 | Q8G4G9 | Q9EXQ1 |
| Q8ZL52 | P27463 | Q27450 | O43790 | Q9KP97 | Q9CWM2 | P32590 | P60502 |
| P32938 | P15714 | Q9SYQ8 | Q9ZDX9 | P11279 | Q23280 | Q9W6G6 | P77234 |
| Q9X2T1 | P10974 | Q65PB5 | P34460 | Q5XCQ3 | P19807 | Q95NR9 | P0AG16 |
| P26792 | P46368 | P20233 | P05592 | P75363 | P68028 | Q9PAZ2 | P32995 |
| Q05511 | Q9SZ67 | Q88VM0 | Q83MH5 | Q6CKU6 | P57692 | P28524 | Q5PKK7 |
| P30404 | Q01173 | P27337 | Q9Y4L1 | Q9Y6H1 | P17673 | Q04902 | Q9Z2D3 |
| Q91Z98 | Q43019 | Q9NYK1 | P02591 | P82935 | Q9CRB6 | P10917 | P09849 |
| Q02942 | P48671 | P11480 | P41011 | P60183 | Q8F4W2 | P14210 | P12236 |
| P11588 | Q9Y5X9 | Q5R1W3 | Q862Z5 | P32669 | O32160 | Q67LB8 | P64325 |
| P00798 | Q40302 | Q04736 | Q6C2T9 | Q43133 | Q5R6P6 | Q9SLY8 | Q48VS4 |
| P49223 | P09871 | P68530 | P09466 | Q5WHP3 | P44903 | P61184 | P42851 |
| P35336 | P56202 | O46427 | Q5BJE1 | P98175 | P18670 | P23286 | Q710D7 |
| Q9EQT5 | Q37369 | Q03975 | Q9JJZ2 | Q9Z7T3 | P54976 | P33157 | P0AAN8 |
| P32822 | O24585 | P81902 | O24581 | Q42539 | Q83KD8 | P16635 | Q71KN3 |
| Q6X9Z5 | Q986N6 | Q03044 | P50291 | Q4K8H3 | P07471 | P56626 | P22485 |
| P05167 | P30842 | P56625 | Q9BIR7 | Q8Z4R0 | Q94694 | Q91195 | Q5UQL2 |
| P79819 | Q6BM74 | P14614 | Q9UBX7 | Q58926 | Q9K620 | P98043 | P46389 |
| O35684 | Q29150 | Q93572 | P92131 | Q8BWN8 | Q8YAC4 | Q6GDP4 | Q815J4 |
| P01041 | P08055 | P30563 | Q9M263 | P51229 | O82134 | Q9JM71 | Q8PZI1 |
| Q06331 | P36401 | P25784 | P24303 | O28355 | Q9D3G2 | P67978 | O04487 |
| P79139 | P02195 | P02773 | P11965 | Q9F0D4 | O14198 | Q89A93 | Q89AT8 |
| P50343 | P32589 | P68407 | P00986 | Q6SEH4 | P36547 | Q8YA20 | P26142 |
| Q39799 | P05611 | Q8K3H7 | Q8BJR6 | Q8NGR8 | P0A7S2 | P92996 | O94544 |
| Q9K9B2 | P29860 | O87712 | P28758 | Q52562 | P63294 | Q9FIC6 | Q71Z80 |
| P48674 | Q9LNU3 | Q6CJR7 | P21589 | Q8YQ78 | Q62GM2 | Q5KAW8 | P34666 |
| P05939 | Q3SB11 | P16641 | Q8P1W3 | Q899M3 | P13273 | Q3SZ10 | Q6BSE7 |
| P34827 | P12399 | Q5JJ64 | O76284 | Q8Y4H1 | Q822J3 | P01088 | P0C1B8 |
| Q9NR96 | P52269 | Q91233 | P35033 | Q5FJP6 | Q9ZHF6 | P09734 | Q8R6F5 |

Table E.6: (continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P93329 | P01000 | Q7VDY0 | P05943 | P52467 | Q3A6M1 | P10822 | O51578 |
| O77814 | Q9H4G4 | P18294 | P04248 | Q9Y935 | Q8RGF0 | Q01705 | P83260 |
| Q9L8F6 | Q57JA9 | Q4IPB3 | Q68XI2 | Q6DIX1 | P31085 | Q9FL89 | Q5LIK4 |
| O08789 | P01325 | P04247 | P13918 | Q69ZB8 | Q93JF1 | P92999 | Q00901 |
| Q07663 | P35034 | Q9Y7S9 | Q8YHF5 | Q3JZ19 | O15374 | Q9LLR6 | P11934 |
| P52232 | Q28987 | Q9SM64 | Q98QA8 | Q9HQ97 | O26147 | P35590 | Q8CQF4 |
| P48871 | Q5HER0 | Q4P555 | P0A343 | Q7NWN7 | O66756 | Q9FRV0 | P84247 |
| P42236 | P42088 | Q7Z410 | P61701 | Q8P1E3 | Q5NRQ1 | Q4WIF3 | Q38J84 |
| Q38865 | Q27666 | P12333 | P36454 | Q6FD81 | Q9Z7P2 | P73920 | P53191 |
| Q4A0E7 | P40001 | Q8YFY0 | Q9Y473 | P84005 | Q9BBR4 | O04151 | Q8ZH66 |
| P34648 | P02606 | Q9XSM2 | P66671 | Q4K6V0 | Q8K0W9 | P02196 | Q9RDV8 |
| Q16937 | Q3IYM7 | Q03301 | Q3ASF8 | O29573 | Q890Q7 | P05438 | P62170 |
| P25036 | P93338 | P00566 | P54007 | P23910 | P57476 | Q5X3L4 | O66805 |
| O19092 | P35004 | P36369 | Q48AW1 | P46314 | Q58442 | Q9LXG3 | P42094 |
| Q46XI7 | P26455 | P29447 | P36023 | Q37601 | P58603 | P06472 | P42252 |
| P00791 | P05120 | P11503 | Q16960 | P33109 | P68118 | Q8KEA4 | P00369 |
| Q9LE15 | Q5RCH2 | P82978 | P0A1S6 | Q08908 | Q8ZIX3 | Q8JFR1 | P50876 |
| Q42517 | Q7AH91 | Q5R8A4 | Q8WMX8 | Q8BJL1 | P0C026 | Q98ME7 | Q9VCQ3 |
| O75830 | Q9D267 | P17333 | Q05962 | P50228 | Q9QXZ6 | Q8Z9R1 | Q08583 |
| P10184 | P09761 | Q80ZD8 | Q49161 | O35024 | P49460 | P98047 | Q861U8 |
| P27435 | P13087 | Q06827 | Q8U4M7 | Q9VEX0 | P62982 | P05994 | O76093 |
| P04789 | P41975 | P05787 | Q8G767 | P16011 | Q9UHC1 | Q3ZYV1 | Q8CHG5 |
| Q94570 | P10039 | P47767 | Q9PQB5 | Q8VSR1 | P17526 | P00996 | P42703 |
| P17670 | P30436 | P15638 | Q6N5R5 | P0A2T3 | P63110 | Q6G6C0 | Q9CQ20 |
| P35038 | O64654 | Q6P698 | P43098 | P57497 | Q7U7I2 | Q62587 | Q57JQ5 |
| P02159 | Q3KIA0 | Q04691 | Q8X5W4 | Q87VB6 | P16181 | P96744 | Q8Z991 |
| Q28372 | Q00002 | Q3BS21 | Q6NH13 | P61899 | P04447 | Q74D53 | Q51575 |
| Q9SZE7 | P07728 | P58283 | Q872I5 | P34057 | Q9HSL7 | Q9GKY0 | O00468 |
| Q8EW32 | P30231 | P21250 | Q9A1V8 | Q8EB78 | Q8BMN3 | O08976 | P04474 |
| Q8T6B3 | P42688 | Q9LDJ3 | Q9XGH7 | Q9VW26 | P51506 | Q87RX3 | P53696 |
| P25700 | Q7NDH1 | P31017 | Q6APV5 | Q88YP8 | P43223 | P13392 | Q9NY64 |
| Q6CGE6 | Q6LLW2 | P24565 | Q95189 | Q9HQG3 | Q6N9F2 | P98074 | Q9U6V7 |
| P58478 | O66686 | P13540 | Q86Y34 | Q60416 | Q8VEE1 | P17647 | Q6GC09 |
| P46633 | P09488 | P42899 | Q6ME32 | Q890P0 | P41585 | Q03211 | P0A5X4 |
| P12257 | Q83SE6 | Q63081 | Q95MP1 | Q11010 | O74543 | P0AGF4 | |
| O82399 | Q8L3R3 | Q65NN2 | Q9KUS1 | Q9NZ42 | O15439 | Q93NE2 | |
| Q88CW7 | P26447 | O22711 | Q8YM52 | P25615 | P47484 | Q8DY73 | |

# Appendix F

# Published Papers