SIGNAL PEPTIDE PREDICTION IN THE SPACE-FREQUENCY DOMAIN

by

Ran Li

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering

Winter 2006

UMI Number: 1432421

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

SIGNAL PEPTIDE PREDICTION IN THE SPACE-FREQUENCY DOMAIN

by

Ran Li

Approved: _____

Javier Garcia-Frias, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____

Gonzalo R. Arce, Ph.D.
Chair of the Department of Electrical and Computer Engineering

Approved: _____

Eric W. Kaler, Ph.D.
Dean of the College of Engineering

Approved: _____

Conrado M. Gempesaw II, Ph.D.
Vice Provost for Academic and International Programs

ACKNOWLEDGEMENTS

It's a pleasant aspect that I have now such opportunity to express my gratitude for all the people which accompany and support me during these years at UD. First, I would like to thank my adviser, Dr.Javier Garcia-Frias, for his invaluable friendly guidance, trust and constant encouragement; I would also like to express my most sincere thanks to all the people in Office of Graduate Studies: Mary Martin, Susan Lee and their colleagues for their kindness and generous support; I must also thank my husband, Wei Zhou; If there were no caring, love, affection and continuous support of him, none of this would ever have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Bioinformatics makes use of applied mathematics, informatics, statistics and computer science to solve biological problems. With the explosion of biological data, the signal processing community is in a unique position to analyze the data using traditional and non-traditional signal processing methods.

Proteins are the building blocks of cells. A protein is a string of linked amino acids, each one of them represented as a symbol in a 20-letter alphabet. Analyzing the symbolic amino acid sequences helps us find the characteristics of the proteins and predict the secondary structures. However, in some cases, the characteristic patterns are too weak to be detected by analyzing the symbolic strings. Then, it is possible to assign chemical property indices to the amino acids to map the symbolic sequence to a numeric representation, so that the analysis is performed over this representation.

A signal peptide is a short stretch of amino acids found at the beginning of proteins. It is typically rich in hydrophobic amino acids. One of the major functionalities of the signal peptide is to localize proteins to specific regions within the cell. Therefore, knowledge of a specific signal peptide for a protein provides an important clue to its likely locations. In this thesis, we propose a new method to detect the presence of the signal peptide based on the space-frequency processing of the numeric amino acid sequences.

There are more than 400 index mappings available which map the symbolic amino acid sequences to numeric representations. Therefore, the selection of the index mapping becomes an important issue. In this thesis, we propose a method to

select the index mappings which is suitable for signal peptide detection. To detect the presence of the signal peptide in an amino acid sequence, the numeric sequence is transformed to the space-frequency domain by the Wigner-Ville transform. We observed that the amino acid sequence with signal peptide tends to have smaller variance in the space-frequency domain than the amino acid sequence without signal peptide. Based on this observation, we devised a new signal peptide detection algorithm which effectively differentiates between the amino acid sequences with the signal peptide and the amino acid sequence without the signal peptide.

To evaluate our method, we use a dataset of 210 protein sequences, half of which have signal peptides. We use half of the dataset to learn the optimal parameters of the detection algorithm. The other half of the dataset is used to test the algorithm. Experimental results show that our method detects the presence of the signal peptides at a promising rate.

Chapter 1

INTRODUCTION

1.1 Motivation

Bioinformatics is the subject of analyzing biological information using computational and statistical techniques. It develops and utilizes computer databases and algorithms to accelerate and enhance biological research. The history of bioinformatics can be traced back to 1865 when Gregor Mendel did experiments on the cross-fertilization of different colors of the same species and carefully recorded and analyzed the data. Since then, the understanding of genetics has advanced remarkably with the efforts of many scientists and researchers around the world. In 1990, the Human Genome Project was started, and at the end of 1998, scientists had completely determined the genome of one multicellular organism. As an increasing number of genomic sequences were determined, it become apparent that new methods are desirable to exploit the information content in the exploding raw data [4]. The new methods should record, annotate, store, analysis, and search/retrieve the biological data more easily and efficiently. The challenges of bioinformatics can be grouped into three major tasks:

- The development of new algorithms and statistics to assess relationships among members of large data sets;
- The analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures;

- The development and implementation of tools that enable efficient access and management of different types of information.

Genomes carry all the information from one generation to the next in every organism. A genome is the DNA in an organism which is made up of four similar chemicals abbreviated A, C, T, G. The particular order of As, Ts, Cs, and Gs is extremely important. The order underlies all diversity in life. It even dictates whether an organism is human or another species such as yeast, rice, or fruit fly, all of which have their own genomes [1]. Certain regions along DNA chains are called genes, which encode all the information to make a protein.

Proteins are chains of 20 different types of amino acids, which in principle can be joined together in any linear order. They are also called polypeptide chains. The sequence of the amino acids is known as primary structure, and it can be represented as a string of letters. The primary structure of a protein contains all the necessary information required for the manifestation of higher 3-D level structures. Analysis of protein sequence data provides insights of its functions, leading to the knowledge of biological active sites. Traditionally, the analysis is directly performed on the symbolic amino acid sequences. However, in some cases, the biological patterns are too weak to be detected by this method. An alternative method is to analyze the numeric amino acid sequence by assigning numeric index values to the amino acids according to certain chemical properties. Once the numeric indices are assigned to the protein, we can consider it as a signal, and existing signal processing methods such as frequency and time-frequency analysis can be applied to facilitate the analysis. Examples of DNA, protein and its numeric representation are as follows:

- DNA Sequence: A-T-T-C-C-G-A-C
- Protein Sequence: W-G-P-I-G-L-T-C

- Numeric representation of the protein sequence (Using the Kyte-Doolittle index mapping that we will introduce in Chapter 3): 5.4-9.0-8.0-5.2-9.0-4.9-8.6-5.5

A signal peptide is a short (15-60 amino acids long) peptide chain. It is widely believed that there are three regions (n-region, h-region and c-region) in all signal peptides that compose a recognition motif interpreted by the targeting machinery. One of the major functionalities of signal peptides is to direct the protein to its proper cellular and extracellular locations. For various purposes, it is desirable to identify signal peptides and their corresponding cleavage positions. There are three major methods used to predict the signal peptide: Neural Network(NN)[24] [20] [18], Hidden Markov Model (HMM) [21] and position weight matrix approach [14]. These algorithms, although predict relatively accurate results, are difficult to implement and very time-consuming. The goal of this thesis is to develop a new method to detect the presence of the signal peptide in an amino acid sequence more efficiently by performing space-frequency analysis on the numeric amino acid sequences.

1.2 Objectives and outlines

The objective of this thesis is to explore novel applications of traditional signal processing techniques to the explosive biological data. The major effort is focused on developing new algorithms for signal peptide detection.

The thesis begins with a molecular biology basics introduction in chapter 2. There are three basic molecules responsible for the functioning of all organism's cells: DNA, RNA and proteins. Chapter 2 reviews the three molecules' structures, functions and their relationships which are presented as the central dogma in biology.

Chapter 3 covers about amino acid indices. Each amino acid has various properties which are responsible for different functions. Scientists assign a numeric value to each amino acid according to its specific characteristic by theoretical and

experimental methods. Analyzing the numeric amino acid sequence, we can detect protein patterns which are too weak to be detected in their symbolic representations. There are more than 400 index mappings available which map the symbolic amino acid sequences to the numeric amino acid sequences. Therefore, the selection of index mapping becomes an important issue. In this chapter, we propose a method to select the index mappings which are suitable for the signal peptide detection.

Motivated by chapter 3, chapter 4 serves as a review of protein analysis. Analysis of protein sequence data not only provides insights to protein structures and functions, but also leads to finding its biological active sites. Mapping symbolic amino acid sequence to its equivalent numeric amino acid sequence, we can apply frequency and space-frequency analysis methods which are widely used in the signal processing community to perform pattern recognition in protein analysis. The Wigner-Ville transform is chosen because of its optimality in space-frequency resolution.

In chapter 5, we develop a new algorithm to detect the presence of signal peptides in an amino acid sequence. The numeric sequence is transformed to the space-frequency domain by the Wigner-Ville transform. We found that the amino acid sequence with signal peptide tends to have smaller variance in the space-frequency domain than the amino acid sequence without signal peptide. Based on this observation, we devised a new signal peptide detection algorithm which effectively differentiates the amino acid sequences with signal peptides and the amino acid sequences without signal peptides.

Finally, chapter 6 summarizes the contributions of this thesis and describes possible extensions to stimulate future research.

Chapter 2

MOLECULAR BIOLOGY BASICS

2.1 The molecular building blocks of life

The basic principle of life is reproduction. It transform the materials found in the environment of an organism into another organism. Despite nearly 4 billion years of evolution, the basic molecular objects still carry matter, energy and information. The basic units of matter are proteins, which subserve all of the structural and many of the functional roles in the cell; the basic unit of energy is a phosphate bond in the molecule adenosine triphosphate (ATP); and the units of information are four nucleotides, which are assembled together into DNA and RNA [15].

2.1.1 DNA

DNA is the building block of life. It contains the information of the cell required to synthesize proteins and to replicate itself. In other words, it is the storage repository for the information that is required for any cell to function. Watson and Crick discovered the structure of DNA in 1953. Basically DNA is made up by four nucleotide bases: Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). The DNA is arranged in a double helical structure, where each base has its complementary base: A has T as its complementary and similarly the complementary of G is C. The DNA is broken down into bits and is tightly wound into coils, which are called chromosomes; human beings have 23 pairs of chromosomes. These chromosomes are further broken down into smaller pieces of code called Genes. The genetic code in

Table 2.1: Comparison between DNA and RNA.

DNA	*RNA*
Double-Stranded	Single-Stranded
Has Thymine as a base	Has Uracil as a base
Deoxyribose as the sugar	Ribose as the sugar
Maintains protein-encoding information	Uses protein-encoding information

DNA is in triplets such as ATG. The 23 pairs of chromosomes contain about 70,000 genes and every gene has its own functions. A typical example of DNA sequence is as follows:

A-T-T-T-G-C-T-G-A-C-C-T-G

Its complementary sequence is:

T-A-A-A-C-G-A-C-T-G-G-A-C

2.1.2 RNA

RNA is similar to DNA; they are both nucleic acids of nitrogen-containing bases joined by a sugar-phosphate backbone. However, structural and functional differences distinguish RNA from DNA. Structurally, RNA is single-stranded, whereas DNA is double stranded. DNA has Thymine (T), whereas RNA has Uracil (U). RNA nucleotides include sugar ribose, rather than the Deoxyribose that is part of DNA. Functionally, DNA maintains the protein-encoding information, whereas RNA uses the information to enable the cell to synthesize a particular protein. Table 2.1 shows a comparison between DNA and RNA.

2.1.3 Proteins and Amino acids

It is the protein that performs most life functions and make up the majority of cellular structures. A protein has multiple levels of structures: primary, secondary, tertiary and quartenary. The primary structure of protein determines its higher level structures.

A protein's primary structure is a linear sequence of constituent molecules called amino acids which are encoded in the DNA by a triple of nucleotides. There are 64 possible triples of nucleotides (called condons) mapped into 20 different amino acids. Each amino acid contains two parts: one part is identical in all amino acids, and is used to link one amino acid to another to form the backbone of proteins; the other part is a unique side chain (also called R group) that determines the distinctive physical and chemical properties of the amino acids. Table 2.2 lists the names of the 20 amino acids, together with their 3-letter abbreviations and 1-letter standard symbols.

2.2 The Central Dogma of Molecular Biology

The central dogma of molecular biology is the transcription of DNA to RNA and its translation to protein. The information is carried by the genes (DNA which is transcribed to RNA), and expressed in proteins. Information flows from DNA to RNA to protein, but not the reverse. There are two major steps involved in this processing: transcription and translation.

Table 2.2: Amino acids and their symbols.

AminoAcid	$3-LetterCode$	$1-LetterCode$
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

2.2.1 Transcription

In the first step of protein synthesis, the two DNA strands in a gene unzip from each other. Similar to the way DNA replicates itself, a single strand of messager RNA (mRNA) is then made by pairing up mRNA bases with the exposed DNA nucleotide bases. The messager RNA (mRNA) also has four nucleotide bases - but in mRNA, the base uracil (U) replaces thymine (T).

2.2.2 Translation

After the mRNA is manufactured, it travels to a cellular organelle called ribosome. The ribosome reads the mRNA sequence and translates it into the amino acid sequence of the protein. The ribosome starts at the sequence AUG, then reads three nucleotides at a time. We call the three nucleotides a codon. Each codon specifies a particular amino acid. The stop codon (UAA, UAG and UGA) tells the ribosome that a protein is complete.

2.2.3 Signal peptide

Signal peptides play a central role in the targeting and translocation of nearly all secreted proteins and many integral membrane proteins in both prokaryotes and eukaryotes [10]. Signal peptides direct proteins to their proper cellular and extracellular locations [29]. It constitutes the N-terminal part of the whole protein sequence. It is commonly believed that there are three structurally, and possibly, functionally distinct regions that come together to form the signal peptide [10]:

- A basic N-terminal region (n-region).
- A central hydrophobic region (h-region).
- A more polar C-terminal region (c-region).

One of the major functionalities of the signal peptide is to direct the protein to its proper cellular and extracellular locations. One example is the translocation of

proteins across the cytoplasmic membranes via the well-established secrete pathway found in both eukaryotic and prokaryotic cells. In this secretorypathway, a protein designed to be exported from the cell is labeled by a N-terminal signal sequence. This signal sequence directs its protein to the secretion apparatus. The signal peptide sequence is usually cleaved from its protein once the protein export is underway. The structural determinants for this process seem to reside in the n-region and h-region, with positions one (-1) and three (-3) upstream of the cleavage site processing the most important amino acids [12]. The cleavage site for the signal peptide is located in c-region. The signal sequence conservation as well as its length and the cleavage site position vary from one protein to another. Moreover, there are major differences between different species. For various purposes, it is desirable to identify signal peptides and their corresponding cleavage positions. In this thesis, our goal is to develop a new method to detect the presence of the signal peptide in an amino acid sequence by performing space-frequency analysis on the numeric amino acid sequences.

2.3 An Example – E.coli

E. coli is the abbreviated name of the bacterium in the Family Enterobacteriaceae named Escherichia (Genus) coli (Species). The name Escherichia comes from the name of the person Escherich, who first isolated and characterized the bacterium in 1885. It is a inhabitant of the human intestine. It also lives in the intestine of many other animals, wild as well as domestic.

Escherichia coli is one of the most studied forms of life, which makes it a favorite organism for bioinformatists. In this thesis, we also use E.coli as our experimental data for the space-frequency processing, amino acid indexing and signal peptide detection.

Chapter 3

AMINO ACID INDEX

Amino acids are the building block of proteins. There are 20 amino acids in total, and their combinations correlate to different protein's characteristics in shape, size, and chemical reactivity among others. Studying the mapping of their symbolic representations into numerical sequences provides us more insights about the structural and functional patterns of the proteins. In this chapter, we review some examples of these indices and introduce a method of amino acid index selection.

3.1 Background

Each amino acid has different physicochemical and biochemical properties. An amino acid index is a set of 20 numerical values (each corresponding to a different amino acid) defined according to a biological property. For instance, The EIIP index which was proposed by Veljkovic [28] describes the average energy states of all valence electrons in a particular amino acid. In 1988, Nakai et. al. [19] collected 222 amino acid indices from public research papers and investigated the relationships utilizing hierarchical cluster analysis. In 1990, Tomii and Kanehisa [26] increased the size of the collection to include 402 indices and re-performed the clustering. A more complete collection is available in the AAindex database [16], which collected 434 amino acid indices. This collection can be divided into six major classes: α and turn properties, β propensity, amino acid composition, hydrophobicity, physicochemical properties, and other properties.

Table 3.1: Three examples of the index mappings: ZIMP, Kyte-Doolittle and GRAP.

$AminoAcid$	$ZIMP$	$Hydropathy(KD)$	$GRAP$
Alanine(A)	0.00	1.8	8.1
Arginine(R)	52.00	-4.5	10.5
Asparagine(N)	3.38	-3.5	11.6
Aspartic acid(D)	49.70	-3.5	13.0
Cysteine(C)	1.48	2.5	5.5
Glutamine(Q)	3.53	-3.5	10.5
Glutamic acid(E)	49.90	-3.5	12.3
Glycine(G)	0.00	-0.4	9.0
Histidine(H)	51.60	-3.2	10.4
Isoleucine(I)	0.13	4.5	5.2
Leucine(L)	0.13	3.8	4.9
Lysine(K)	49.50	-3.9	11.3
Methionine(M)	1.43	1.9	5.7
Phenylalanine(F)	0.35	2.8	5.2
Proline(P)	1.58	-1.6	8.0
Serine(S)	1.67	-0.8	9.2
Threonine(T)	1.66	-0.7	8.6
Tryptophan(W)	2.10	-0.9	5.4
Tyrosine(Y)	1.61	-1.3	6.2
Valine(V)	0.13	4.2	5.9

Table 3.1 lists three examples of these indices. The ZIMP represents the ZIMJ680103 polarity index introduced by [30], and is based on polarity properties of the amino acids. The KD means KYTJ820101 hydropathy index which was introduced by Kyte and Doolittle [17] in 1982, and is an index displaying the hydropathic character of a protein. The GRAP represents the GRAR740102 polarity index which incorporates weightings for molecular volume and molecular weight, amongst other ingredients [11]. The ZIMJ680103, KYTJ820101 and GRAR740102 are the entry names of the indices in the AAindex database [16].

Index selection is a key factor when applying space-frequency tools on protein

analysis. An appropriate index can lead to insights about the biological activities of the proteins, and not all the index mappings are always suitable for a specific protein analysis. Therefore, we need to find a numerical index mapping which highlights the properties of the proteins, and is related to the studied protein's functions. However, there are more than 400 indices available. Facing such a huge amino acid index database, how to choose an appropriate index for an analysis becomes an important issue. Concepts from information theory can provide some guidance in the amino acid index selection process.

3.2 Selection of Amino Acid Indices

Information theory is the mathematical theory of data communication and storage founded by Claude E. Shannon in 1948 [25]. In his classic paper "A Mathematical Theory of Communication", Shannon defined a measure of information content of a message: entropy H. The entropy is the measurement of the average number of bits required to encode a message. The entropy of an IID random variable M is defined as:

$$H(M) = -\sum_i p(m_i) log p(m_i) \tag{3.1}$$

where $p(m_i)$ is the probability of a outcome m_i. In information theory, entropy is a measurement of the uncertainty and ranges in value from 0 to $\log n$. When $p(m_i) = 1$, there is no uncertainty and $H = 0$; when $p(m_i) = 1/n$, where n is the number of possible outcomes, we have no information about the message, then the uncertainty reaches the maximum: $H = \log n$. Considering an indexed protein sequence as a message composed of a sequence of index numbers, we can apply the concepts of entropy to the problem of selecting an amino acid index [5].

As mentioned previously, there are 20 different amino acids in total. However, the index mapping is not one to one. As shown in table 3.1, different amino acid symbols may map to the same numerical number. In other words, the indexed protein sequence may contain fewer than 20 discrete values. In addition, we also have to consider the dynamic range of the amino acid indices. Based on these considerations, in [5], a method is developed to calculate the information content I of an indexed protein sequence. The method is briefly illustrated in the following paragraphs. For a more detail explanations, we refer readers to [5].

To deal with the problem of different dynamic ranges for different index mappings. The dynamic range of the index numbers is broken into 20 equally sized bins:

$$s = \frac{n_{max} - n_{min}}{20} \tag{3.2}$$

where s is the size of the bins. n_{max} and n_{min} are the maximum and minimum index numbers, respectively. Given an indexed protein sequence, for each bin, we count the number of the index numbers which fall in the bin range. Denote b_i as the number of index numbers which fall in bin range i, the information content I of an indexed protein sequence is calculated as follows:

$$I = \frac{\sum_{i=1}^{N} p(b_i) log p(b_i)}{log \frac{1}{N}} \tag{3.3}$$

where N is the number of possible index numbers in an amino acid sequence (N may less than 20). $p(b_i)$ is the probability of the amino acid index numbers which fall in bin range b_i.

As we know, lower values of I imply less random nature of the representation, which, therefore, should reveal more structural patterns associated with the

Table 3.2: The information content I of three different index mappings.

$AminoAcid$	$ZIMP$	$Hydropathy(K-D)$	$GRAP$
$\overline{I}_{SP}$	0.6139	0.8562	0.7723
$\overline{I}_{non-SP}$	0.6203	0.8688	0.7933

protein sequence. In this thesis, our goal is to find a way to efficiently differentiate the signal-peptide protein sequences from the non-signal-peptide protein sequences. Therefore, a suitable index mapping should be selected for this purpose. For a specific index mapping, we calculate the information contents I_{SP} and I_{non-SP} using both SP protein sequences and non-SP protein sequences, respectively. The SP protein sequences and non-SP protein sequences are taken from the Escherichia coli subset of the Gram-negative data in SignalP [20] [21] which is taken from SWISS-PROT version 29 [21] [3]. There are 105 SP protein sequences and 119 non-secretory protein sequences in the data set. For our purpose, we used 50 signal peptides and 50 non-secretory (non-SP) proteins. In both cases, we calculate the information content of the first 24 nucleotides (which corresponds to the average length of the signal peptide in our training group). The results are shown in table 3.2. ($\overline{I}_{SP}$ and $\overline{I}_{non-SP}$ are the averages over the 50 I_{SP}s and 50 I_{non-SP}s, respectively.)

From table 3.2, we notice that I_{SP} is less than I_{non-SP} for all the three index mappings. This coincides with our previous observations: SP proteins contain more structural information than non-SP proteins. For the purpose of differentiation between the two class of proteins, our selection rule is to select the index mapping which results in the largest difference between I_{SP} and I_{non-SP}. From table 3.2, we expect that GRAP would given the best classifications because the difference between I_{SP} and I_{non-SP} for GRAP is the largest among the three index mappings.

Chapter 4

PROTEIN ANALYSIS

With major advances in the field of molecular biology, coupled with improvements in genomic technologies, we are facing an explosive growth of various types of data. This deluge of genomic information not only has led to requirements for computerized database design but also requires various specialized tools to view and analyze the data. The protein's chemical properties, the chain conformation, the function of the protein, and its species specificity are determined by information carried in the amino acid sequences. Thus, finding the similarities between two or more amino acid sequences can provide insights into structure-function relationships of the active sites of a protein. This chapter introduces two analytical approaches being used to analyze protein sequences - frequency analysis and space-frequency analysis.

4.1 Frequency analysis of protein - Resonant Recognition Model

Generally speaking, a signal can be represented as a function of time (space) which shows how the signal magnitude changes over time (space). Alternatively, by performing the Fourier transform, the signal defined in the time (space) domain can be transformed into the frequency domain. Intuitively, this tells us how quickly the signal amplitude changes. Since, proteins are linear macromolecules made up of linked amino acids, we can treat an amino acid sequence as a signal and use signal processing tools to reveal the unknown biological functions and active sites in proteins.

The Resonant Recognition Model (RRM) is a physico-mathematical model that analyzes the interactions of proteins and their targets (other proteins, DNA regulatory segments or small molecules). The RRM assumes that the specificities of protein interactions are based on the resonant electromagnetic energy transfered at the specific frequency for each interaction. The main applications of this model are to predict protein functions and locations of its biological active sites using digital signal processing.

In this technique, proteins are first transformed into numerical sequences by assigning numerical values to each amino acid. These values correspond to certain properties of amino acids. The numerical sequences obtained in this way are then transformed into frequency domain by using the Discrete Fourier Transformation (DFT). The frequencies correspond to the distribution of protein's structural motifs, which are responsible for biological function of the amino acid sequence. When comparing proteins sharing the same biological or biochemical function, the technique allows detection of code/frequency pairs which are specific for their common biological properties.

The Discrete Fourier Transform (DFT) is defined as:

$$X(n) = \sum_{m=1}^{N-1} x(m) e^{-j(2/N)nm} \qquad n = 1, 2, 3, ..., N/2 \tag{4.1}$$

where $x(m)$ is the m'th component in the numerical sequence, N is the total sequence length, and $X(n)$ is the nth coefficient of the DFT. The coefficients describe the amplitudes, phases, and frequencies of sinusoids which make up the original signal. In the RRM, the signal is the numeric representation of the amino acid sequence and we assume that points in the numerical sequence are equally separated by a distance d=1. Then, the maximum frequency in the spectrum is F=$(1/2)d$=0.5.

The total number of points in the sequence influences the resolution of the spectrum only. Therefore, the resolution in the spectrum for N-point sequence is 1/N. The n'th point in the spectrum function corresponds to the frequency f=n/N.

In order to determine the correlation which reveals the biological function in a protein family, we need to know the common characteristics of the sequences which have the same biological functions. The cross-spectral function, S(n) determines the common frequency components of two signals:

$$S(n) = X(n)Y^*(n) \qquad n = 1, 2, 3, ..., N/2 \tag{4.2}$$

where $X(n)$ is the DFT coefficients for a signal $x(n)$, and $Y(n)$ is the complex conjugate DFT coefficients for another signal $y(n)$. The common frequency components in the two signals are determined by the peak frequencies in this cross-spectral function. Similarly, for a group of protein sequences, we can calculate the absolute values of multiple cross-spectral function coefficients M which are defined as follows to find the peak frequencies:

$$|S(n)| = |X_1(n)||X_2(n)||X_3(n)| \cdots |X_M(n)| \qquad n = 1, 2, 3, ..., N/2 \tag{4.3}$$

Next, we use two amino acid sequences: Arginine-binding periplasmic protein 1 Precursor and Arginine-binding periplasmic protein precursor to demonstrate this method. The two sequences are both from the bacterial solute-binding protein 3 family.

$> ARTI\_ESCO0$

$MKKVLIAALIAGFSLSATAAETIRFATEASYPPFESIDAN$

$NQIVGFDVDLAQALCKEIDATCTFSNQAFDSLIPSLKFRR$

$VEAVMAGMDITPEREKQVLFTTPYYDNSALFVGQQGKYTS$

$VDQLKGKKVGVQNGTTHQKFIMDKHPEITTVPYDSYQNAK$

$LDLQNGRIDGVFGDTAVVTEWLKDNPKLAAVGDKVTDKDY$

$FGTGLGIAVRQGNTELQQKLNTALEKVKKDGTYETIYNKW$

FQK

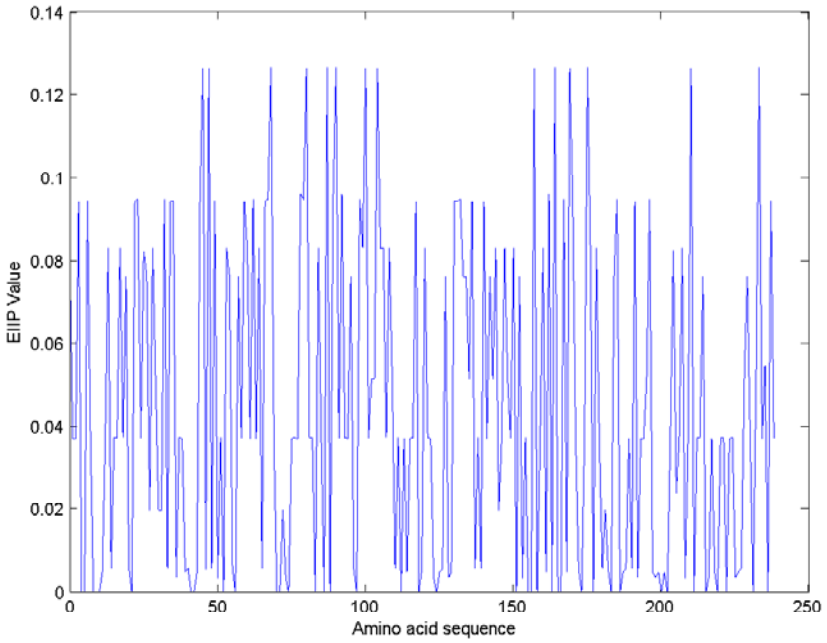

Figure 4.1: Arginine-binding periplasmic protein 1 precursor using EIIP index.

$>ARTI\_HAIN0$

$MKKTLLTAILLGASVAASAQELTFAMQPSYPPFETTNAKG$

$EIIGFDVDVTNAICQEIQATCKFKSETFDALIPNLKAKRF$

$DAAISAIDITDARAKQVLFSDAYYDSSASYVALKGKATLE$

$SAKNIGVQNGTTFQQYTVAETKQYSPKSYASLQNAILDLK$

$SGRIDIIFGDTAVLADMISKEPEIQFIGEKVTNKKYFGNG$

$LGIAMHKSNKDLAAQLNKGLAAIKANGEYQKIYDKWITK$

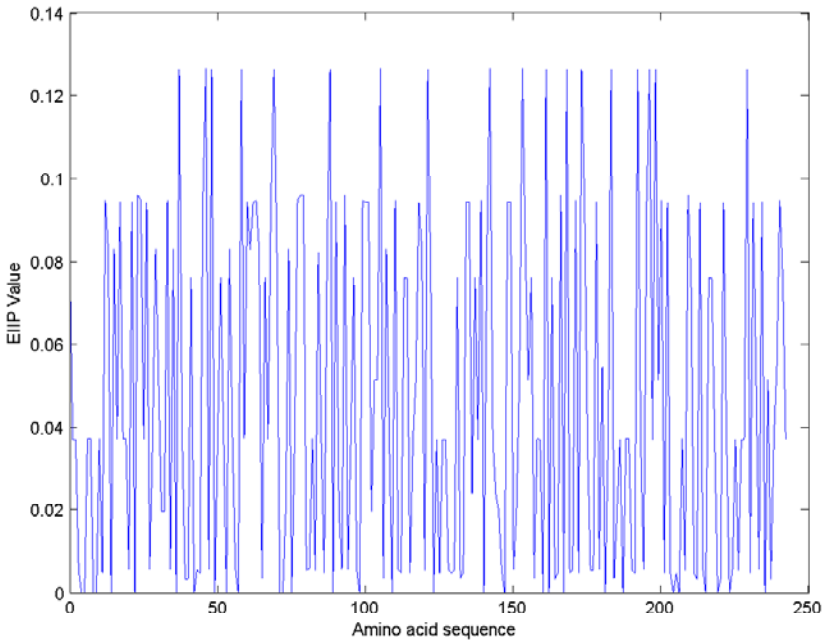

Figure 4.2: Arginine-binding periplasmic protein precursor using EIIP index.

Figure 4.1 and 4.2 show the symbolic and the numeric representations of the two protein sequences. The numeric values are indexed by the Electron-ion interaction potential (EIIP) index. The EIIP index was proposed by Veljkovic [28],

and describes the average energy states of all valence electrons in a particular amino acid. Their DFT representations are shown in Figure 4.3. Figure 4.4 shows the consensus spectrum which is the cross-spectral function of this group of proteins.

Notice that there is a prominent peak when examining the consensus spectrum of the two protein sequences. Previous studies [7] [6] [27] show that the significant presence of a peak frequency in consensus spectrum implies that all of the analyzed sequences within the group have one frequency component in common. This frequency is related to the biological function of the protein family. The peak frequency in our example corresponds to the biological function of this protein family. Once the RRM characteristic frequency of a particular biological function has been determined, it is possible to identify the amino acids that contribute mostly to the characteristic frequency and the biological function of the protein.

4.2 Space-Frequency Analysis

The aim of space-frequency analysis of a signal is to obtain information about how the spectral content of the signal changes along with space. The Wavelet transform is a relatively new signal processing tool for multiple-resolution analysis and local feature extraction of non-stationary signals [8]. It represents a signal in terms of a finite length or fast decaying oscillating waveform (known as the mother wavelet). This waveform is scaled and translated to match the input signal. Wavelet transforms are broadly classified into the discrete wavelet transform (DWT) and the continuous wavelet transform (CWT). In 1999, Fang and Cosic [9] proposed a method to analyze EEIP indexed protein sequences by using the continuous wavelet transform (CWT). The CWT is a tool used to decompose a signal into wavelets, which are highly localized in time. Its basis functions are scaled and shifted with the same resolution based on the space-localized mother wavelet. It can be used to construct a space-frequency representation of a signal that offers very good time and frequency localization. Although it can be chosen to localize individual events such

as an active site in protein sequences, it can not reveal the characteristic frequency component of the Resonant Recognition Model. In this thesis, we use another space-frequency analysis to address this weakness.

4.2.1 Wigner-Ville Distribution

The Wigner-Ville Distribution is a quadratic space-frequency representation. It satisfies a number of desirable mathematical properties and possesses optimal resolution in the space-frequency domain [2]. The Wigner-Ville Distribution (WVD) of a signal x(t) is defined as:

$$\boldsymbol{W}_x(t,f) = \int_\tau x(t+\tau/2)x^*(t-\tau/2)e^{-jw\tau}d\tau \tag{4.4}$$

The WVD can be considered as a Fourier transform of the signal's auto-correlation function with respect to the delay variable. It can also be thought of as a short-Time Fourier Transform (STFT) where the windowing function is a space-scaled, space-reversed copy of the original signal. In the signal processing community, it is a popular tool to extract subtle signal features. Figure 4.5 is an example performing such a transform using the Kyte-Doolittle index [17] on the Glucose-1-phosphatase protein in E.Coli.

Although the Wigner-Ville transform provides prominence in localizing a signal in both space and frequency domains, its application is limited by interference terms which are introduced by its quadratic nature. For example, the equation defined in (3.4) is a Wigner-Ville transform of x. When x is a sum of (a+b), the Wigner-Ville transform of such x will contain an interference term $2ab$ in addtion to desired quantity $(a^2 + b^2)$. These interference terms result in an increase noise level of the Wigner-Ville transform. How to reduce the cross term interferences

without destroying the useful properties of the WVD is very important to space-frequency analysis [23]. The nonlinear center affine filter introduced in [2] is a very promising tool to reduce the interference term and leave the desired quantity realtively unaffected. However, the performance in our study degrades when the center affine filter is used. Thus, all simulation results correspond to the case where no filter is utilized.

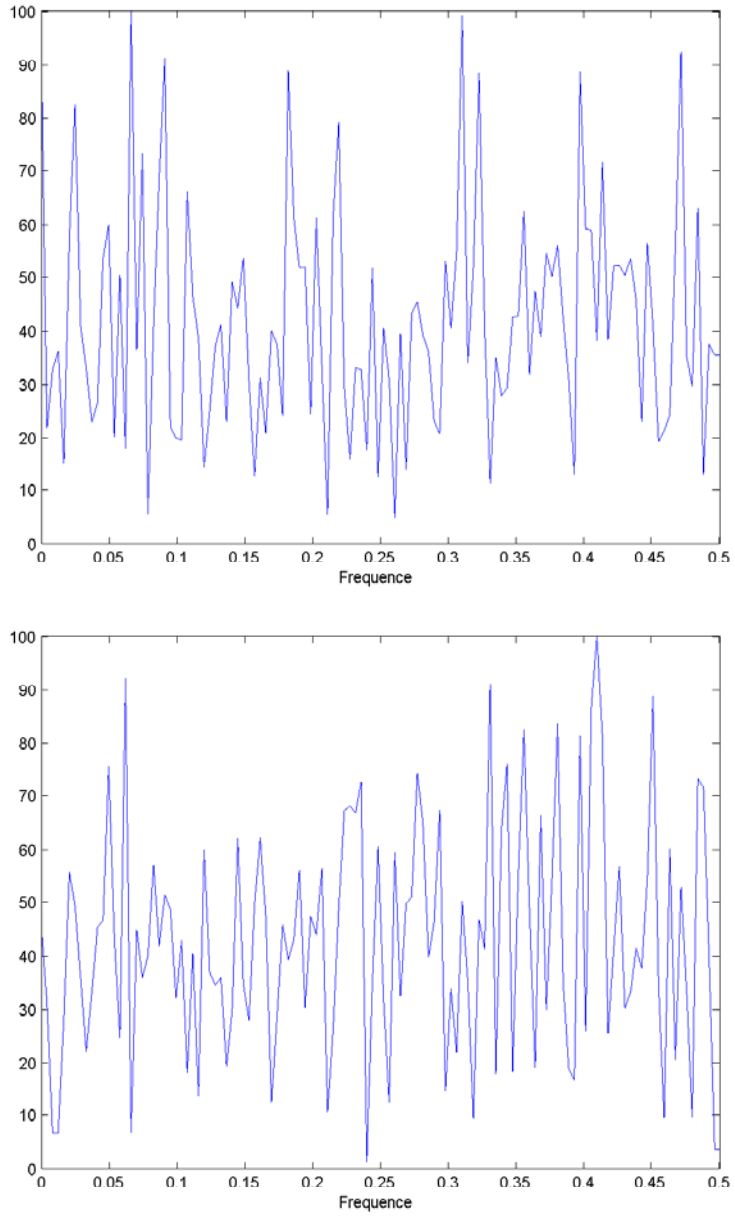

Figure 4.3: FFT representations of two proteins (Arginine-binding periplasmic protein 1 precursor and Arginine-binding periplasmic protein precursor).

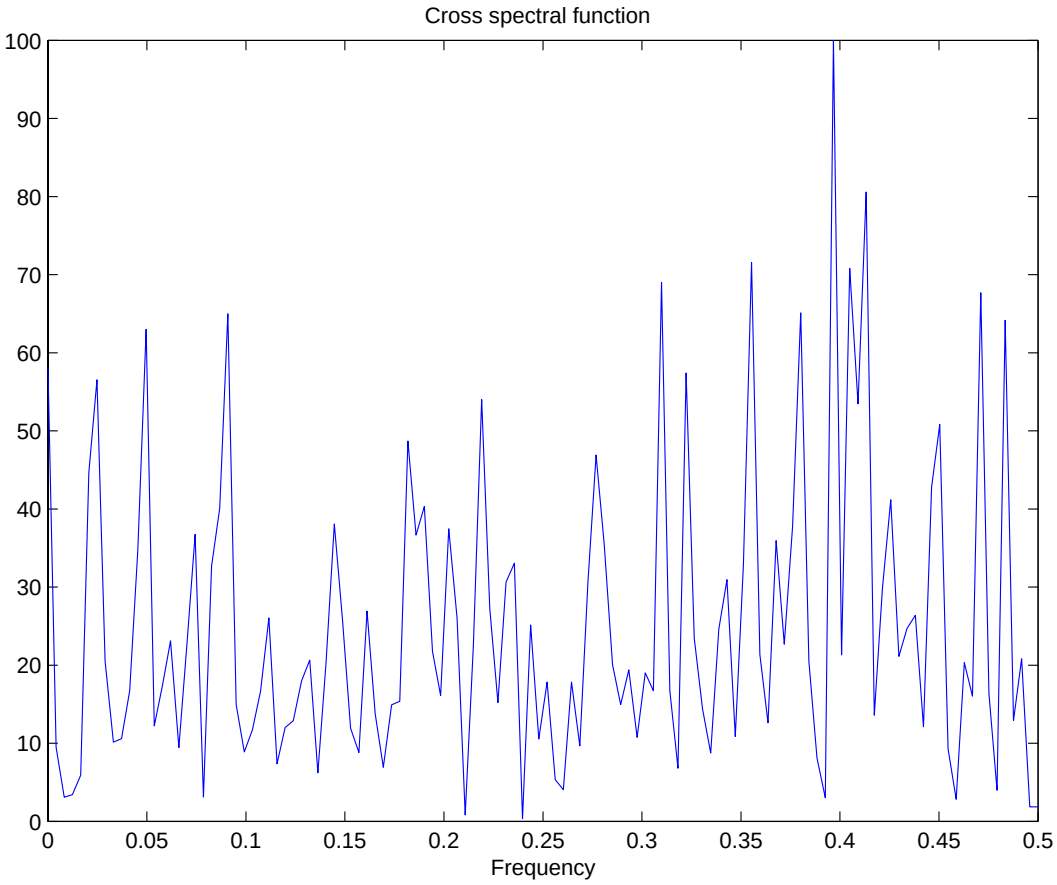

Figure 4.4: Cross spectrum of the two proteins shown in figure 3.3.

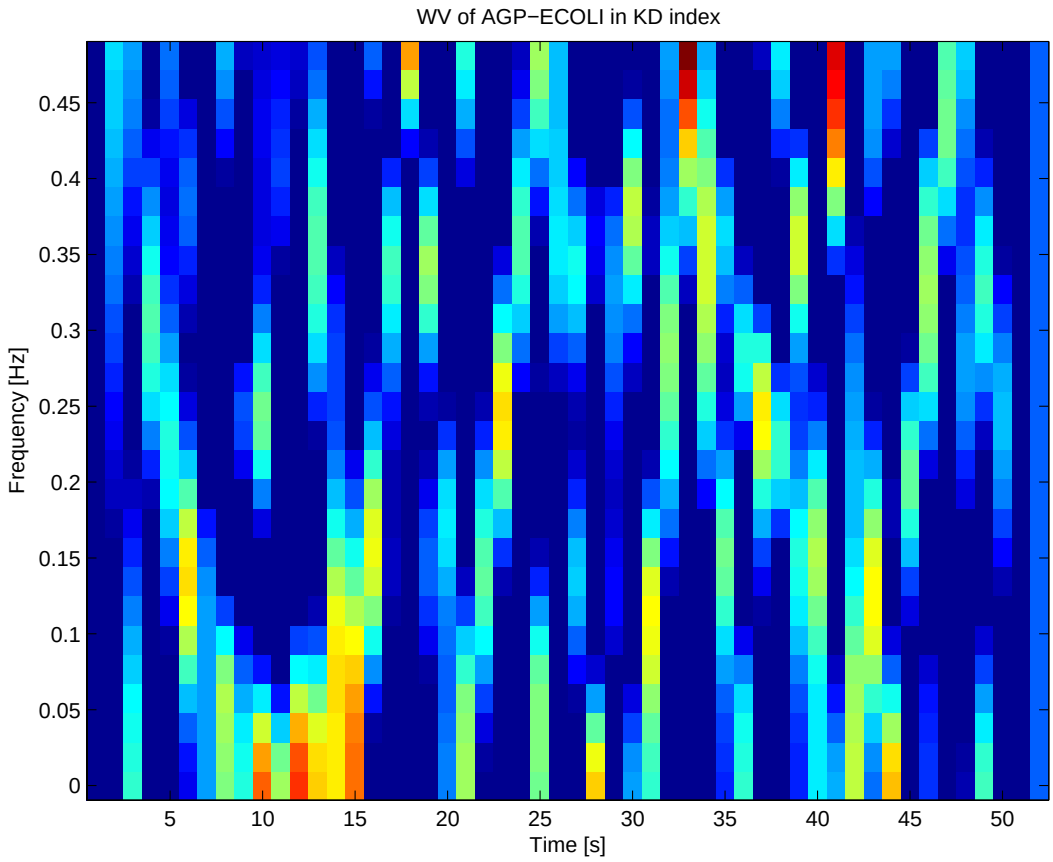

Figure 4.5: Wigner-Ville transform of the Glucose-1-phosphatase protein in E.Coli using Kyte-Doolittle index.

Chapter 5

SIGNAL PEPTIDE DETECTION

5.1 Protein targeting and Signal Peptides

As explained in Chapter 2, the central dogma of molecular biology is the transcription/translation of DNA to RNA to protein. Translation is one of the two processes of synthesizing proteins and consists of matching amino acids to corresponding sets of three nucleotides (codons) and linking them to form a protein. The site of translation is in the ribosome, which is a dynamic complex of serveral RNA molecules. In 1970, when Gunter Blobel conducted experiments on the translocation of proteins across membranes, he discovered that many proteins have a short amino acid sequence (signal sequence) at one end that functions like a postal code for the target organelle. The signal sequence of the protein is recognized by a signal recognition particle (SRP), which is a protein-RNA complex, while the protein is still being synthesized on the ribosome. The signal recognition particle (SRP) recognizes and transports specific proteins to the endoplasmic reticulum (ER) which is a membrane-bound organelle in eukaryotes and to the plasma membrane in prokaryotes. Protein transport across the ER membrane can occur either co- or post-translationally. In the cotranslational pathway, transport occurs while the polypeptide chain is synthesized on a membrane-bound ribosome; in the post-translational pathway, the polypeptide chain is completed before being transferred across the ER membrane.

A signal peptide is a short (15-60 amino acids long) peptide chain that directs the post translational transport of a protein. The amino acid sequences of signal

peptides direct proteins which are synthesized in the cytosol to certain organelles such as the nucleus, mitochondrial matrix, endoplasmic reticulum, chloroplast, and peroxisome. An example is the nuclear localization signal (NLS), which is a signal peptide directing to the nucleus and is often a unit consisting of plus-charged amino acids. The NLS is normally located inside the peptide chain. Almost all proteins that are transported to the endoplasmic reticulum have a sequence consisting of 5-10 hydrophobic amino acids on the N-terminus. Most of these proteins are transported from the endoplasmic reticulum to the Golgi apparatus. If these proteins have a particular 4-amino-acids sequence on the C-terminus, these proteins stay in the endoplasmic reticulum.

Although varying on length and sequence, almost all signal peptides comprise the N-terminal part of the amino acid chain and are cleaved off after the proteins are translocated through the membrane. The common structure of a signal peptide is described by three regions: the positively charged n-region, the hydrophobic h-region, and a neutral but polar c-region where the cleavage site occurs [10]. Since signal peptides control the entry of virtually all proteins to the secretory pathway, both in eukaryotes and prokaryotes [13], detection of signal peptides in proteins is an important aspect of bioinformatics. It helps not only in revealing mechanisms about how proteins are transported across membranes, so that organelles receive specific sets of proteins and achieve their different structures, but also in finding more effective vehicles for the production of proteins in recombinant systems for drug development. In this chapter, we proposed a new method to discriminate between proteins with Signal Peptides (SP) and proteins without Signal Peptides (non-SP).

5.2 Application of Wigner-Ville Transform

In chapter 3, we discussed how to assign chemical property indices to the elements of a protein sequence. By doing this, we map the symbolic representation

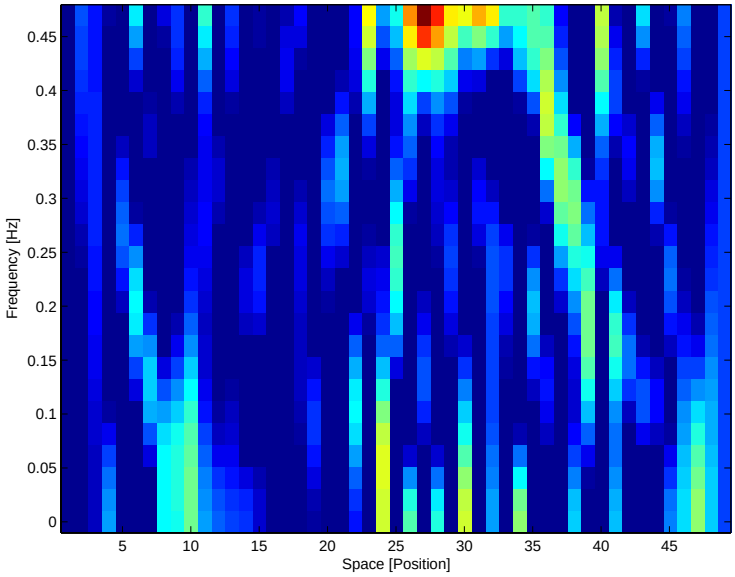

Figure 5.1: Wigner-Ville transform of an amino acid sequence with signal peptide in the x-interval [0,19] using the Kyte-Doolittle index.

of the protein sequence to a numerical representation. Representing the protein sequence in numerical form (signal) makes it possible to apply powerful numerical tools to perform the analysis. In order to discriminate between proteins with/without signal peptides, we utilize indices related to hydrophobicity and polarity, since, as explained before, these are the chemical property characteristic of signal peptides. Specifically, we utilize the ZIMP polarity index, the Kyte-Doolittle index, and the GRAP polarity index mappings, where ZIMP is a polarity index introduced by [30]; the Hydropathy (Kyte-Doolittle) index [17] is an index displaying the hydropathic character of a protein; and the GRAR polarity index is an index calculated from the physicochemical differences between the original and altered amino acid residues [11].

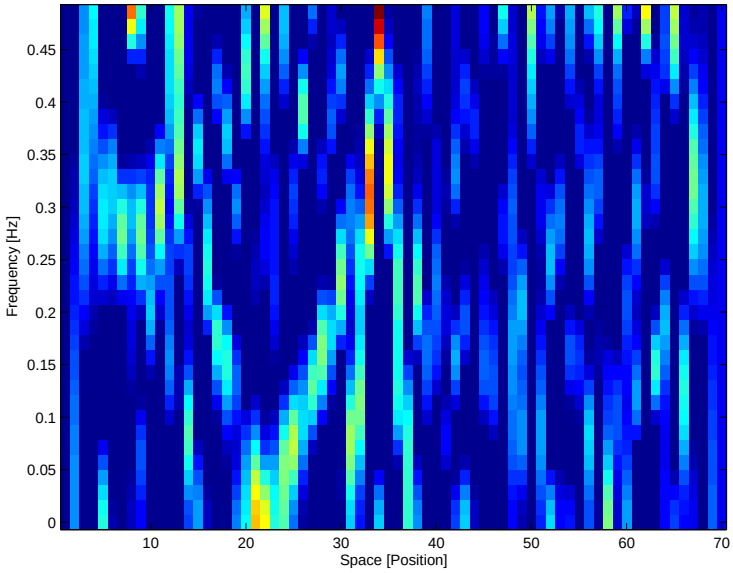

Figure 5.2: Wigner-Ville transform of an amino acid sequence without signal peptide in the x-interval [0,19] using the Kyte-Doolittle index.

The next step is to transform the numeric amino acid sequences to the space-frequency domain using Wigner-Ville transform. By inspecting sets of SP and non-SP proteins, we observed that the signal peptide tends to have less variations than the protein core. This observation is clearly manifested in Figures 5.1 and 5.2, where the Kyte-Doolittle-D index [17] is utilized. Figure 5.1 shows the WV transform of an SP protein sequence, and Figure 5.2 shows the WV transform of a non-SP protein sequence. Notice that in the x-interval [0,19], where the SP is located. Figure 5.2 shows more variations than Figure 5.1. The same observation is shown in Figures 5.3 and 5.4, where the GRAR polarity index [11] is utilized, and in Figures 5.5 in 5.6, where sequences are represented by the ZIMP polarity [30] index. The observation coincides with our expectation: the protein with signal peptide contains more information regarding the chemical properties characterized by the index than the protein without signal peptide, and more information correlates

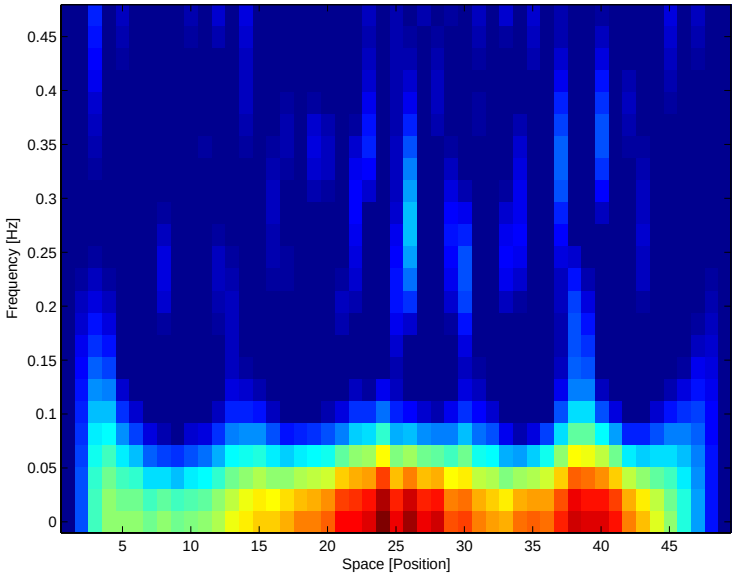

Figure 5.3: Wigner-Ville transform of an amino acid sequence with signal peptide in the x-interval [0,19] using the GRAP index.

with less variations in the space-frequency domain. Thus, the initial region of a SP protein (corresponding to the SP) will presents less variations than the initial region of a non-SP protein (where SP is not presented). Based on this observation, our intention is to find an appropriate statistic which can be automatically used to efficiently differentiate between the two classes of protein sequences.

5.3 Methodology

Denote the symbolic representation of a signal as a vector: $S_n = [s_1, s_2, \cdots, s_n]$, where n is the length of the vector. Utilizing one of the indices defined in section 2, the Kyte-Doolittle index [17], we map the symbolic representation of a signal to the numerical representation:

$$N_n = I(S_n) = [I(s_1), I(s_2), \cdots, I(s_n)] \tag{5.1}$$

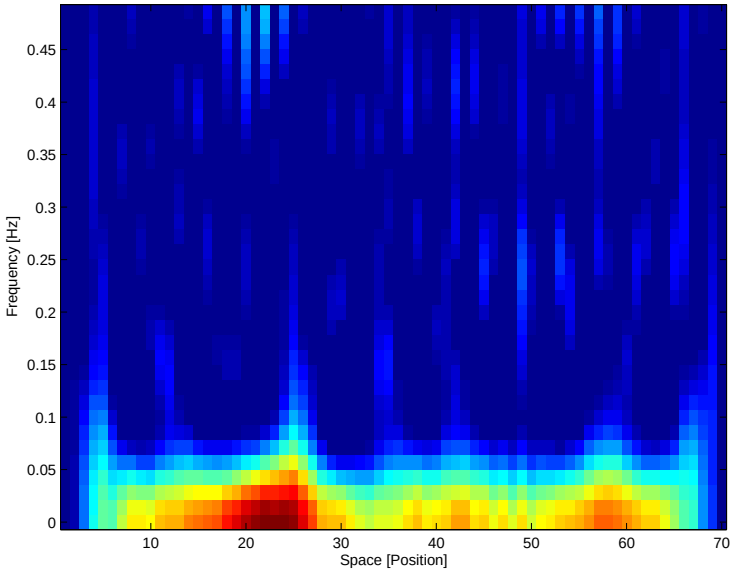

Figure 5.4: Wigner-Ville transform of an amino acid sequence without signal peptide in the x-interval [0,19] using the GRAP index.

where $I()$ is the index function which maps animo acid symbols to the corresponding index. For instance, using the hydrophobicity scale of Kyte and Doolittle [17], we have $I([a, c, i, n]) = [1.8, 2.5, 4.5, -3.5]$, which can be repeatedly used to obtain the numerical representations of the proteins.

To analyze the space-frequency features of the signal, we perform the Wigner-Ville wavelet transform:

$$W_{n,n} = WV(N_n), \tag{5.2}$$

where $W_{n,n}$ is the space-Frequency Representation (SFR) of the protein, and $WV()$ is the Wigner-Ville transform. Note that the Space-Frequency Representation has two dimensions: one is the space dimension and the other is the frequency dimension. Here we represent $W_{n,n}$ as an $n \times n$ matrix with each row representing a particular

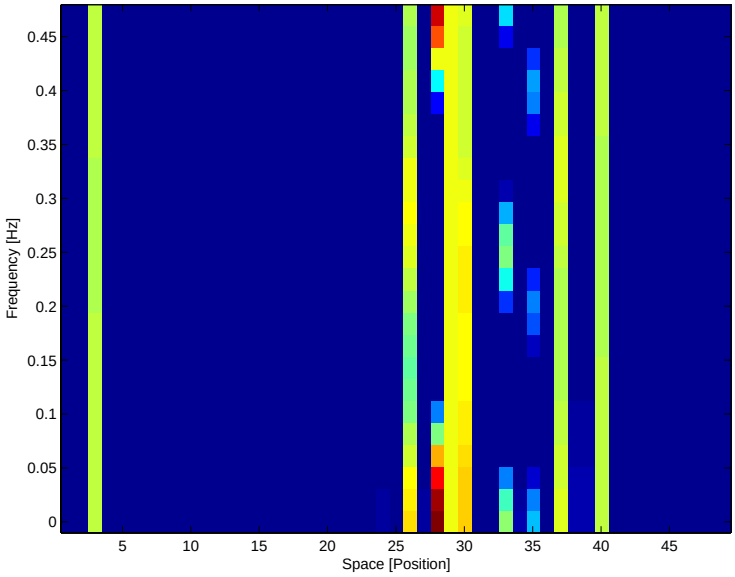

Figure 5.5: Wigner-Ville transform of an amino acid sequence with signal peptide in the x-interval [0,19] using the ZIMP index.

frequency, and each column representing a particular space instance:

$$W_{n,n} = \begin{bmatrix} w_{f1,t1} & w_{f1,t2} & \cdots & w_{f1,tn} \\ w_{f2,t1} & w_{f2,t2} & \cdots & w_{f2,tn} \\ \vdots & \vdots & \vdots & \vdots \\ w_{fn,t1} & w_{fn,t2} & \cdots & w_{fn,tn} \end{bmatrix}. \tag{5.3}$$

As we indicated before, the difference between SP proteins and non-SP proteins is the presence of the signal peptide header in the protein sequence. Therefore, to differentiate the two classes of proteins, we only need to compare the header parts of the two protein sequences. Suppose the length of the signal peptides is m, we cut the matrix $W_{n,n}$ in time domain to $W'_{n,m}$:

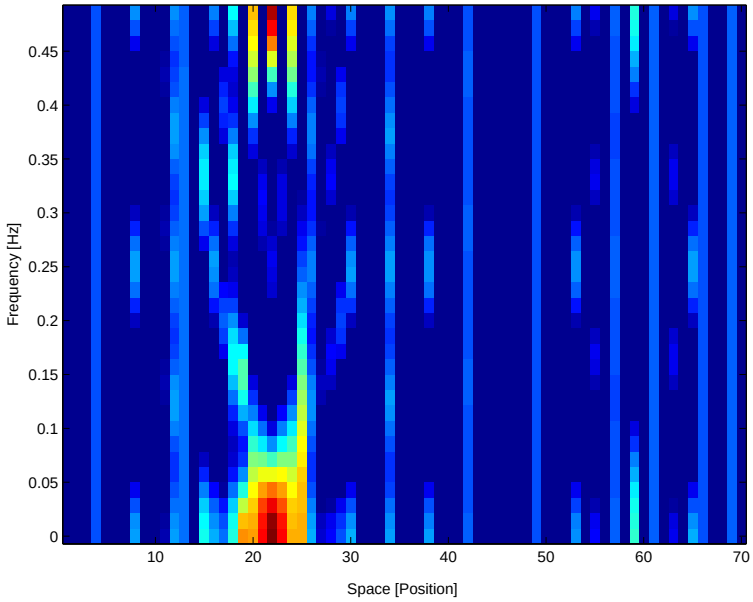

Figure 5.6: Wigner-Ville transform of an amino acid sequence without signal peptide in the x-interval [0,19] using the ZIMP index.

$$W'_{n,m} = \begin{bmatrix} w_{f1,t1} & w_{f1,t2} & \cdots & w_{f1,tm} \\ w_{f2,t1} & w_{f2,t2} & \cdots & w_{f2,tm} \\ \vdots & \vdots & \vdots & \vdots \\ w_{fn,t1} & w_{fn,t2} & \cdots & w_{fn,tm} \end{bmatrix}. \tag{5.4}$$

Denote the $i'th$ row of the matrix $W'_{n,m}$ as $W'_{fi} = [w_{fi,t1}, w_{fi,t2}, \cdots, w_{fi,tm}], i \in (1 \cdots n)$, and the j'th column of the matrix as $W'_{tj} = [w_{f1,tj}, w_{f2,tj}, \cdots, w_{fn,tj}]^T, j \in (1 \cdots m)$, we have:

$$W'_{n,m} = \begin{bmatrix} W'_{f1} \\ W'_{f2} \\ \vdots \\ W'_{fn} \end{bmatrix} \tag{5.5}$$

or

$$W'_{n,m} = \left[\begin{array}{cccc} W'_{t1} & W'_{t2} & \cdots & W'_{tm} \end{array} \right]. \tag{5.6}$$

As mentioned previously, we observed that, in the space-frequency domain, the SP regions tend to have more variations than the non-SP regions. Comparing the space-frequency variations, we expect that the two classes of the proteins can be effectively differentiated. In order to do so, we first calculate two variance vectors:

$$\Delta_f = \left[\begin{array}{c} \sigma_{f1} \\ \sigma_{f2} \\ \cdots \\ \sigma_{fn} \end{array} \right] \tag{5.7}$$

and

$$\Delta_t = \left[\begin{array}{cccc} \sigma_{t1} & \sigma_{t2} & \cdots & \sigma_{tm} \end{array} \right], \tag{5.8}$$

where $\sigma_{fi} = var(W'_{fi})$ and $\sigma_{tj} = var(W'_{tj})$. Then, we calculate a new variance matrix as:

$$\Delta_{n,m} = \Delta_f \times \Delta_t \tag{5.9}$$

$$= \left[\begin{array}{c} \sigma_{f1} \\ \sigma_{f2} \\ \cdots \\ \sigma_{fn} \end{array} \right] \times \left[\begin{array}{cccc} \sigma_{t1} & \sigma_{t2} & \cdots & \sigma_{tm} \end{array} \right] \tag{5.10}$$

$$= \begin{bmatrix} \sigma_{f1,t1} & \sigma_{f1,t2} & \cdots & \sigma_{f1,tm} \\ \sigma_{f2,t1} & \sigma_{f2,t2} & \cdots & \sigma_{f2,tm} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{fn,t1} & \sigma_{fn,t2} & \cdots & \sigma_{fn,tm} \end{bmatrix}, \tag{5.11}$$

where $\sigma_{fi,tj} = \sigma_{fi}\sigma_{tj}$. Therefore, we have $n \times m$ entries in matrix $\Delta_{n,m}$. To estimate the overall variation, we select the k biggest entries, where k is a design parameter, and calculate the average:

$$\begin{aligned} \Delta_k &= max(\Delta_{n,m}, k) \\ \overline{\sigma}_k &= \frac{\Sigma(\Delta_k)}{k} \end{aligned} \tag{5.12}$$

where $max(\Delta_{n,m}, k)$ is a function which takes a matrix $\Delta_{n,m}$, and returns a vector containing the k biggest entries of the matrix. Thus, given a protein, we calculate its average variance $\overline{\sigma}_k$. If it is smaller than a threshold $\overline{\sigma}_k^{th}$, we classify it as a SP protein. Otherwise, we classify it as a non-SP protein. The algorithm can be summarized as follows:

- Given a protein signal, map its symbolic amino-acid representation S_n to the numerical representation N_n. (Using one of the three index mentioned before: KD, ZIMP and GRAP).
- Calculate its Space-Frequency Representation (SFR) $W_{n,n}$.
- Cut the matrix $W_{n,n}$ in space domain to $W'_{n,m}$, where m is the length of the signal peptide.
- Calculate the variance matrix $\Delta_{n,m}$ using equation (5.11).
- Calculate the average variance $\overline{\sigma}_k$ using equation (5.12).

Table 5.1: Confusion Matrix of a two-class classifier.

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

- If $\overline{\sigma}_k$ smaller than a pre-defined threshold $\overline{\sigma}_k^{th}$, the protein is classified as an SP protein. Otherwise, it is classified as a non-SP protein.

Notice that there are three tunable parameters in our algorithm: m (the length of the signal peptide), k (the number of entries in the variance matrix which is used to calculate the average variance), and $\overline{\sigma}_k^{th}$ (the average variance threshold). Our goal is to find the optimal values of the three parameters, so that the algorithm provides the best classification results.

5.4 Performance measure

To evaluate the performance of the classifier for a given tuple (m, k, σ_k^{th}), we calculate the confusion matrix defined in [22]. Table 5.1 shows the confusion matrix for our two-class classifier. The entries in the table are defined as follows:

- a is the number of correct predictions that an instance is negative (non-SP predicted as non-SP),
- b is the number of incorrect predictions that an instance is negative (non-SP predicted as SP),
- c is the number of incorrect of predictions that an instance positive (SP predicted as non-SP),

- d is the number of correct predictions that an instance is positive (SP predicted as SP).

Based on the confusion matrix, we calculate four parameters: True Positive (TP) which is the proportion of the positive cases that are correctly identified; the False Positive (FP) which is the proportion of the negatives cases that are incorrectly classified as positive; Predicted Positive (PP) is the number of predicted positive cases, and Precision (P) is the proportion of the predicted positive cases that are correct.

$$TP = \frac{d}{c+d} \tag{5.13}$$

$$FP = \frac{b}{a+b} \tag{5.14}$$

$$P = \frac{d}{b+d} \tag{5.15}$$

$$PP = b+d \tag{5.16}$$

$$\tag{5.17}$$

Utilizing TP and FP, we draw a ROC graph of the results. A ROC graph is a way to examine the performance of a classifier. It is a plot with the false positive rate on the X axis and the true positive rate on the Y axis. In the ROC curve, the point (0,1) is the perfect classifier which classifies all positive and negative instances correctly. The point (1,0) is the classifier that is incorrect for all classifications. In many cases, a classifier has a parameter that can be adjusted to increase TP at the cost of an increase in FP, or to decrease FP at the cost of a decrease in TP. In our case, this parameter is σ_k^{th}, the threshold for the average variance. Each parameter setting (value of σ_k^{th}) provides a (FP, TP) pair and a series of such pairs can be used to plot an ROC curve. A non-parametric classifier is represented by a single ROC point corresponding to its (FP,TP) pair.

To evaluate a parameterized classifier, the area beneath the ROC curve can be used as a measure of classification accuracy. To compare non-parametric classifiers (points in the ROC curve), the Euclidean distance from the perfect classifier, point (0,1) on the graph, can be used. In that case, we can define AC_d as a distance-based performance measure for an ROC point and calculate it using the equation:

$$AC_d = 1 - \sqrt{\beta * (1 - TP^2) + (1 - \beta) * FP^2} \tag{5.18}$$

where β is a factor, ranging from 0 to 1. It is used to assign relative importance to false positives and false negatives ($\beta = 0.5$ in our simulation). AC_d ranges from 1 for the perfect classifier to 0 for a classifier that classifies all cases incorrectly.

5.5 Training

As mentioned previously, our classification algorithm has three tunable parameters: m (the length of the signal peptide), k (the number of entries in the variance matrix which is used to calculate the average variance), and σ_k^{th} (the variance threshold). The goal of the learning stage is to obtain the parameters so that the performance of the classifier is optimal. We can write the learning process as follows:

$$AC_d = \mathcal{C}_{m,k,\sigma_k^{th}}(training\ sequence) \tag{5.19}$$

where $\mathcal{C}_{m,k,\sigma_k^{th}}$ denotes the classifier with three parameters: m, k,and σ_k^{th}. Given a training sequence, we calculate the performance of the classifier for each of the possible parameter sets. The optimal parameter set is the one that gives the best performance measure AC_d.

We use the amino acid sequences of secretory signal peptides and N - terminal parts of sequences of non-secretory proteins as our training data. The data is an Escherichia coli subset of the Gram-negative data in SignalP [20] [21] which is taken

from SWISS-PROT version 29 [21] [3]. There are 105 signal peptides and 119 non-secretory proteins in the data set. In the training phase, we select 50 SP proteins and 50 non-SP proteins out of the set.

5.5.1 Learning the length of the signal peptide m

Notice that the length of the signal peptide m is known for all the elements of the training set. Thus, in this thesis, our learning process for m is very simple: we just use the average of the signal peptide length of the 50 training sequences as the value of the parameter m. In this case, $m = 24$.

5.5.2 Learning k and σ_k^{th}

To obtain the optimal values of k and σ_k^{th}, we calculate the performance of the classifier using the training data set. The optimal values of k and σ_k^{th} are the ones which provide the best classifications for this set.

Learning k and σ_k^{th} at the same time is a two-dimensional search problem. In this thesis, we propose to obtain the optimal value for k first. Then, with the learned value k, we estimate the optimal value of σ_k^{th}. Although the sequential search does not guarantee a global optimum, the learning process is significantly simplified.

By varying the parameter σ_k^{th}, our classifier becomes a parameterized classifier with parameter σ_k^{th}. As mentioned in the previous section, the performance of a parametric classifier can be measured using the area beneath the ROC curve.

The value of parameter k ranges from 1 to $n \times m$, where n is the length of the amino acid sequence, and m is the length of the signal peptide. From the previous section, m is set to 24, while the length of animo acid sequence n depends on the sequence. Therefore, the maximum value of k should be chosen to accommodate the shortest amino acid sequence in the date set. Investigating the data set, we set the maximum value of k to 600.

Table 5.2: Learning results with three index mapping: GRAP, ZIMP and KD.

Index mapping	GRAP	ZIMP	KD
Optimal $\overline{\sigma}_k^{th}$	3.31e+8	2.22e+12	3.17e+005
TP	0.92	0.92	0.82
FP	0.061	0.122	0.184
P	0.94	0.87	0.80
PP	49	52	50
AC_d	0.9288	0.8966	0.8182

To evaluate the parameter k, we fix the value of k to 10, 150, 300, 600. For each value of k, we evaluate the performance of the classifier with parameter σ_k^{th}. Figures 5.7 and 5.8 show the ROC curve and the area under ROC curve when the index mapping is KD. From the figures, we can see that $k = 600$ provides the largest ROC curve area, i.e. the best classification performance. Figure 5.9 and 5.10 show the ROC curve and the ROC curve area using the index mapping ZIMP. Again, as shown in the figures, $k = 600$ gives the best performance. The same result is also obtained for the index mapping GRAP which is shown in Figures 5.11 and 5.12.

Finally, we learn the parameter σ_k^{th} with k set to its optimal value: $k = 600$. In order to do so, we utilize the AC_d measure defined in equation (5.18) using the parameter $\beta = 0.5$. Figures 5.13, 5.14 and 5.15 plot the results using the index mappings KD, ZIMP and GRAP, respectively. These results are also shown in Table 5.2. From the table, we can see that the performance obtained with the GRAP index is the best among the three, which was expected, since, as we mentioned in chapter 3, the difference in information content between SP proteins and non-SP proteins is the largest when we utilize the GRAP index .

5.6 Testing results

To evaluate the performance of our classification algorithm, we use the remaining protein sequences from the data set to test our algorithm with the parameters obtained in the learning phase. The results are listed in Table 5.3. As shown in the table, our algorithm provides promising classification results, specially when utilizing the GRAP index mapping: $TP = 0.92$, $FP = 0.04$ and $AC_d = 0.9368$. We also show the results in a bar graph as Figure 5.16, where we can see that the performance of using the GRAP index mapping is the best among the three, which again supports our index mapping selection criterion.

Besides trying to select the most suitable index mapping, another way to utilize the index mapping is to combine them. One way to define this combination is to first learn the algorithm parameters utilizing each index mappings individually. Then, for each test protein sequence, we calculate the variances using each of the index mappings. Comparing the calculated variances with their corresponding variance thresholds, each index mapping will give a classification result. To combine these classification results, we assign each classification result with a probability, which is calculated based on the difference between the variance and the variance threshold. The final classification result is then obtained by combining the classification results for each index mapping by multiplying their corresponing probabilities. However, the simulations show that the results of the combination are similar to those obtained by simply choosing the best index mapping for all protein sequences. The reason is that the results coming from the classifiers using different indexes are highly correlated.

Table 5.3: Testing results for the three index mappings: GRAP, ZIMP and KD.

Index mapping	GRAP	ZIMP	KD
TP	0.92	0.80	0.78
FP	0.04	0.22	0.06
P	0.9583	0.7843	0.9286
PP	48	51	42
AC_d	0.9368	0.7898	0.8388

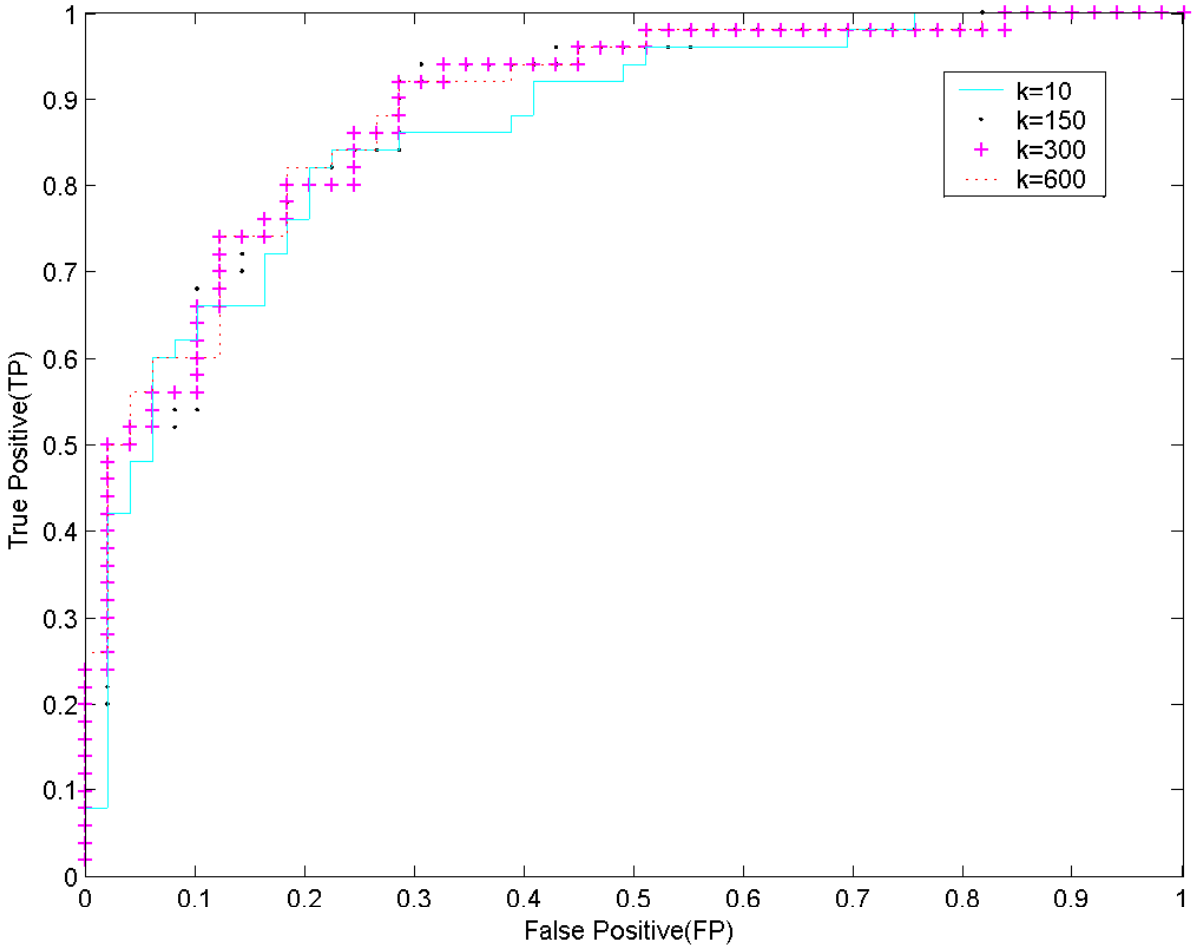

Figure 5.7: ROC curve when the KD index is used for the training set defined in section 5.5.

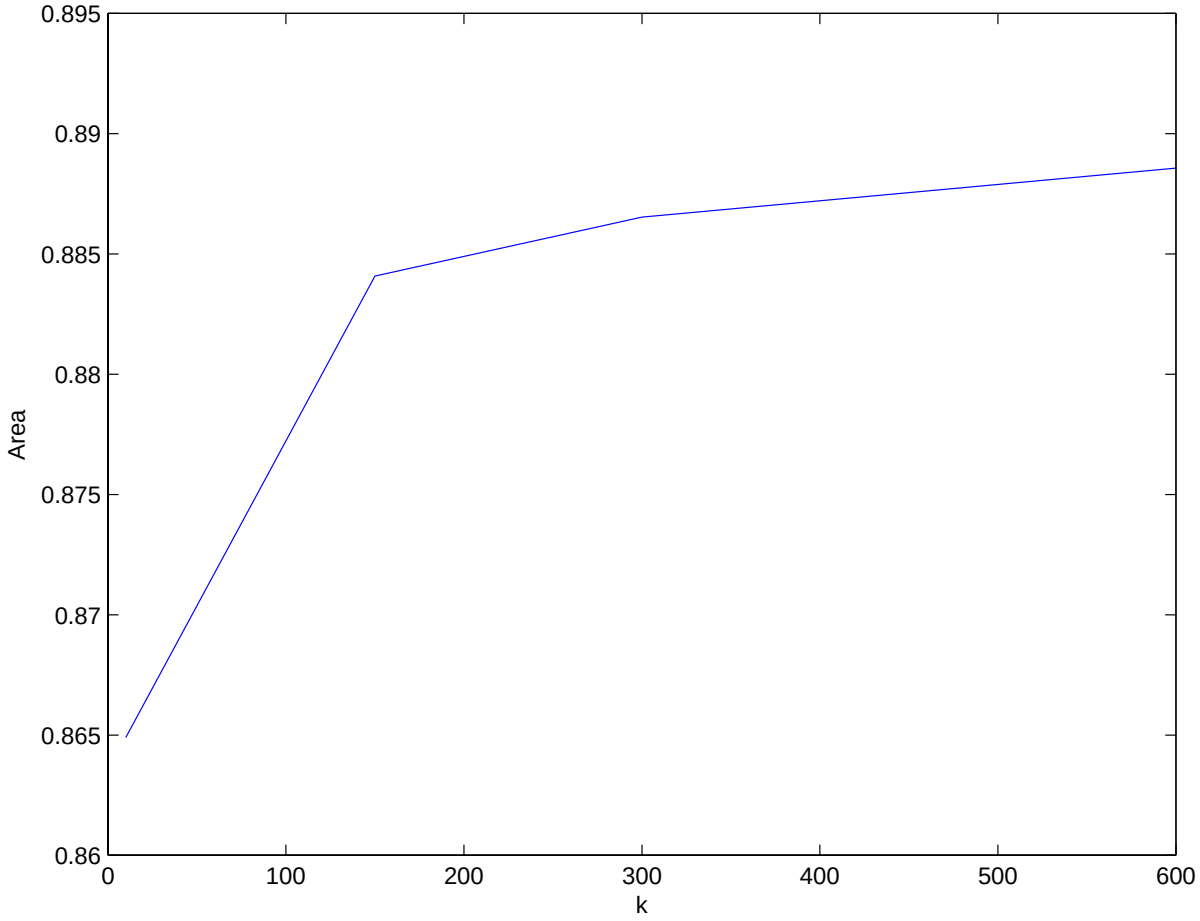

Figure 5.8: Area under ROC curve when the KD index is used for the training set defined in section 5.5.

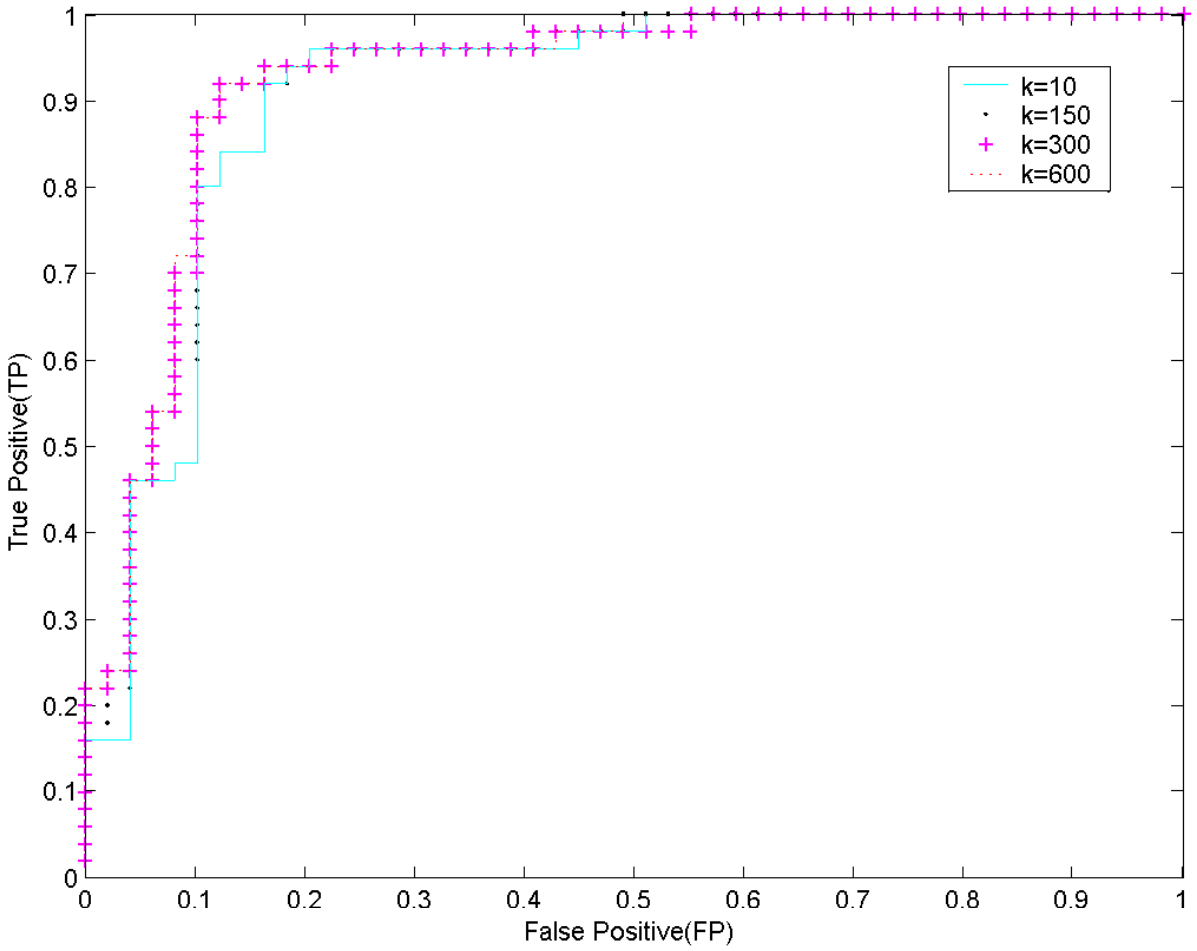

Figure 5.9: ROC curve when the ZIMP index is used for the training set defined in section 5.5.

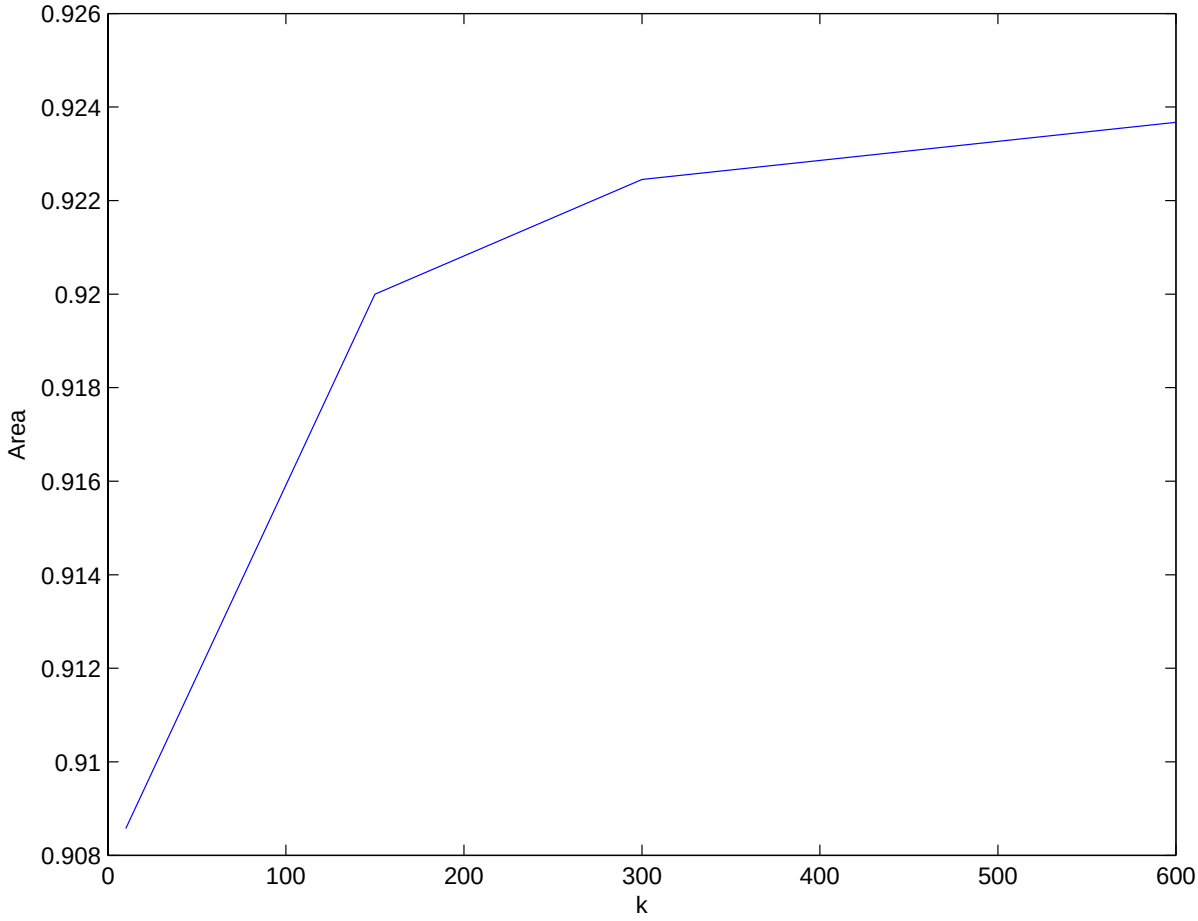

Figure 5.10: Area under ROC curve when the ZIMP index is used for the training set defined in section 5.5.

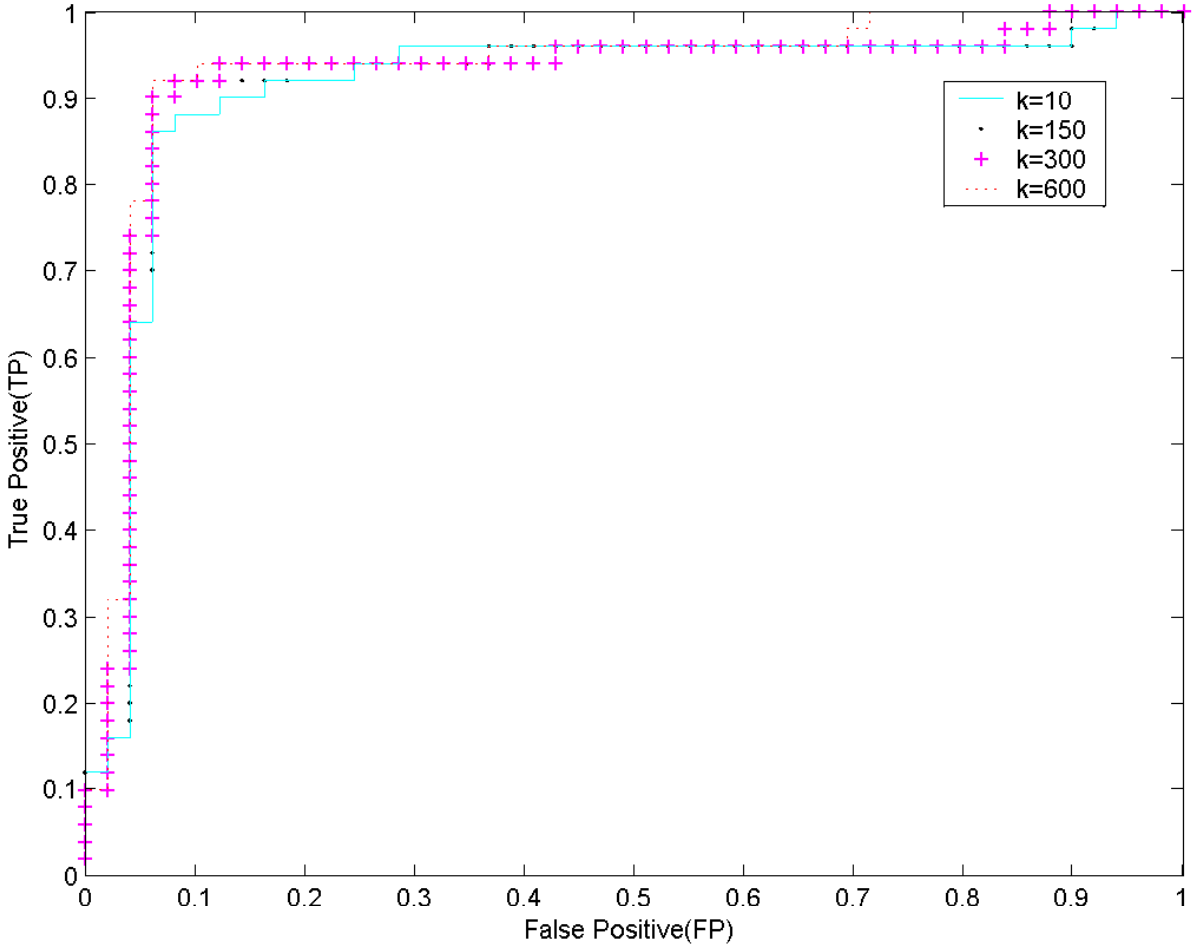

Figure 5.11: ROC curve when the GRAP index is used for the training set defined in section 5.5.

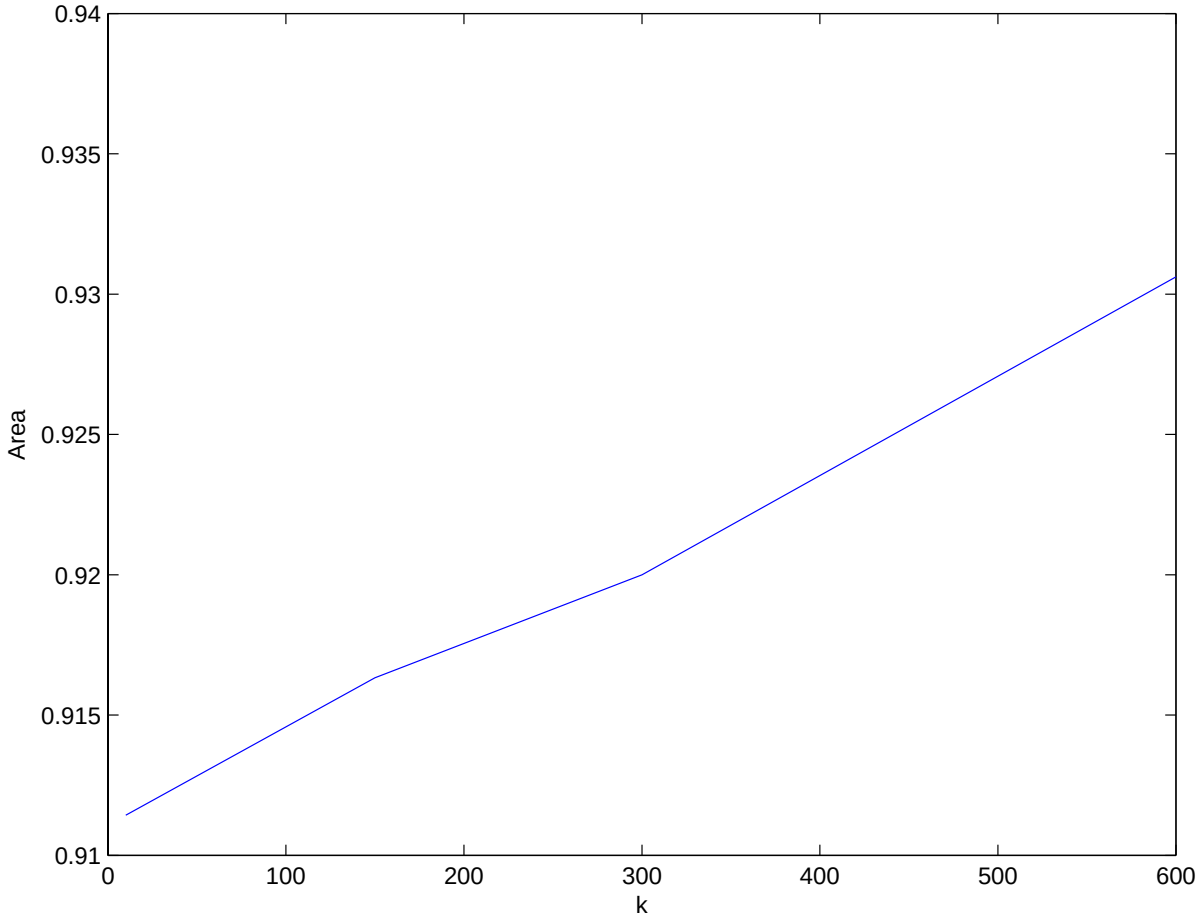

Figure 5.12: Area under ROC curve when the GRAP index is used for the training set defined in section 5.5.

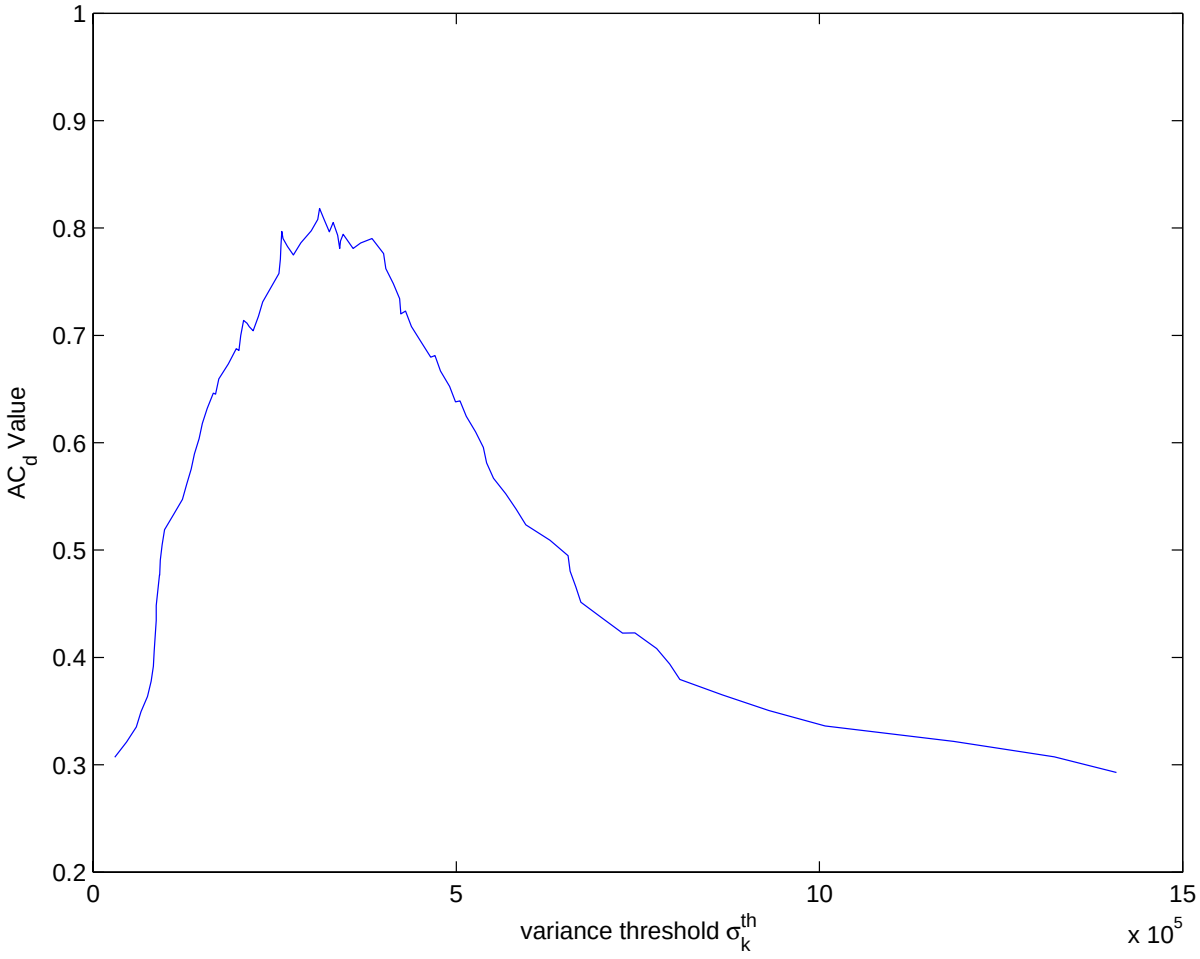

Figure 5.13: For the optimum $k = 600$, AC_d value as a function of the variance threshold σ_k^{th} when the KD index is used for the training set defined in section 5.5.

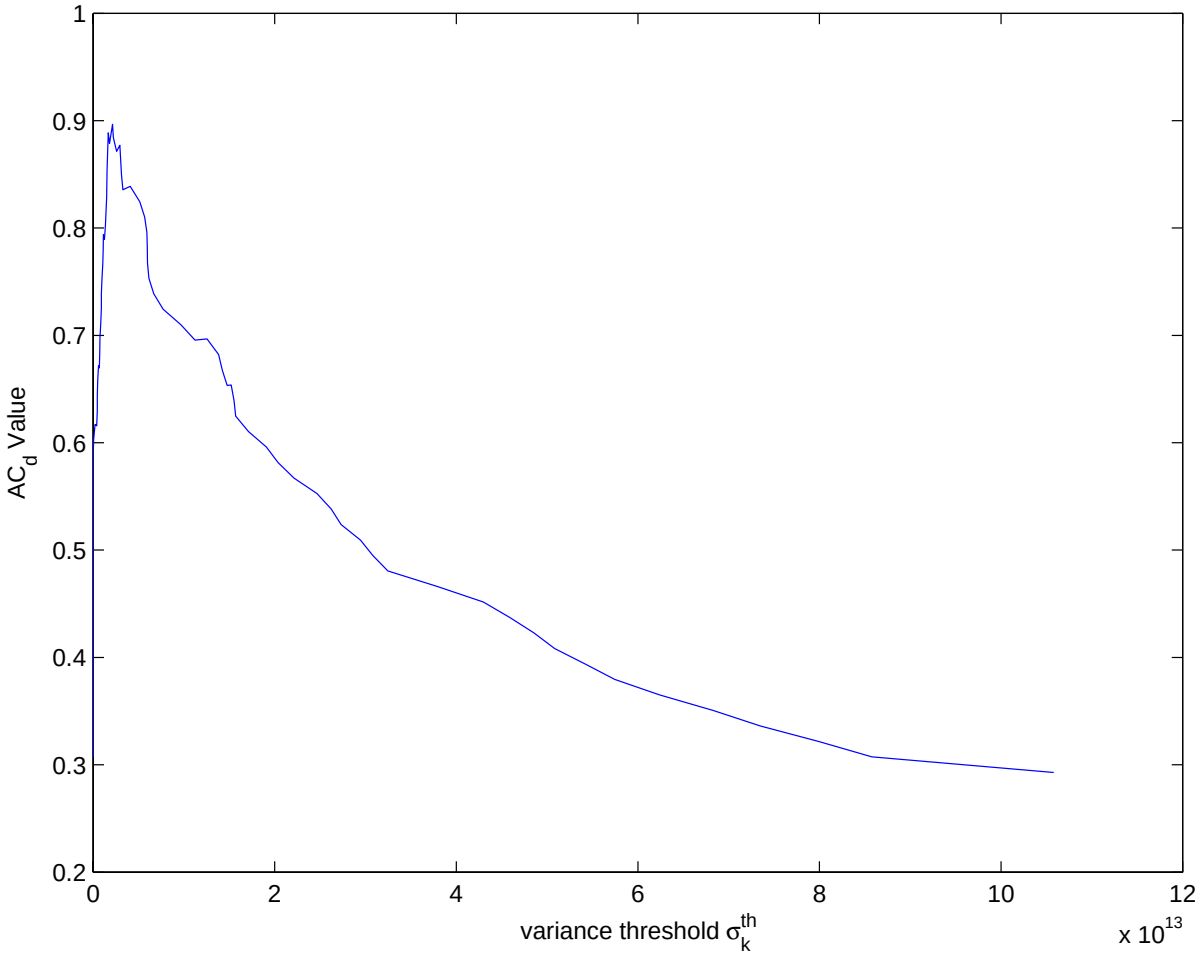

Figure 5.14: For the optimum $k = 600$, AC_d value as a function of the variance threshold σ_k^{th} when the ZIMP index is used for the training set defined in section 5.5.

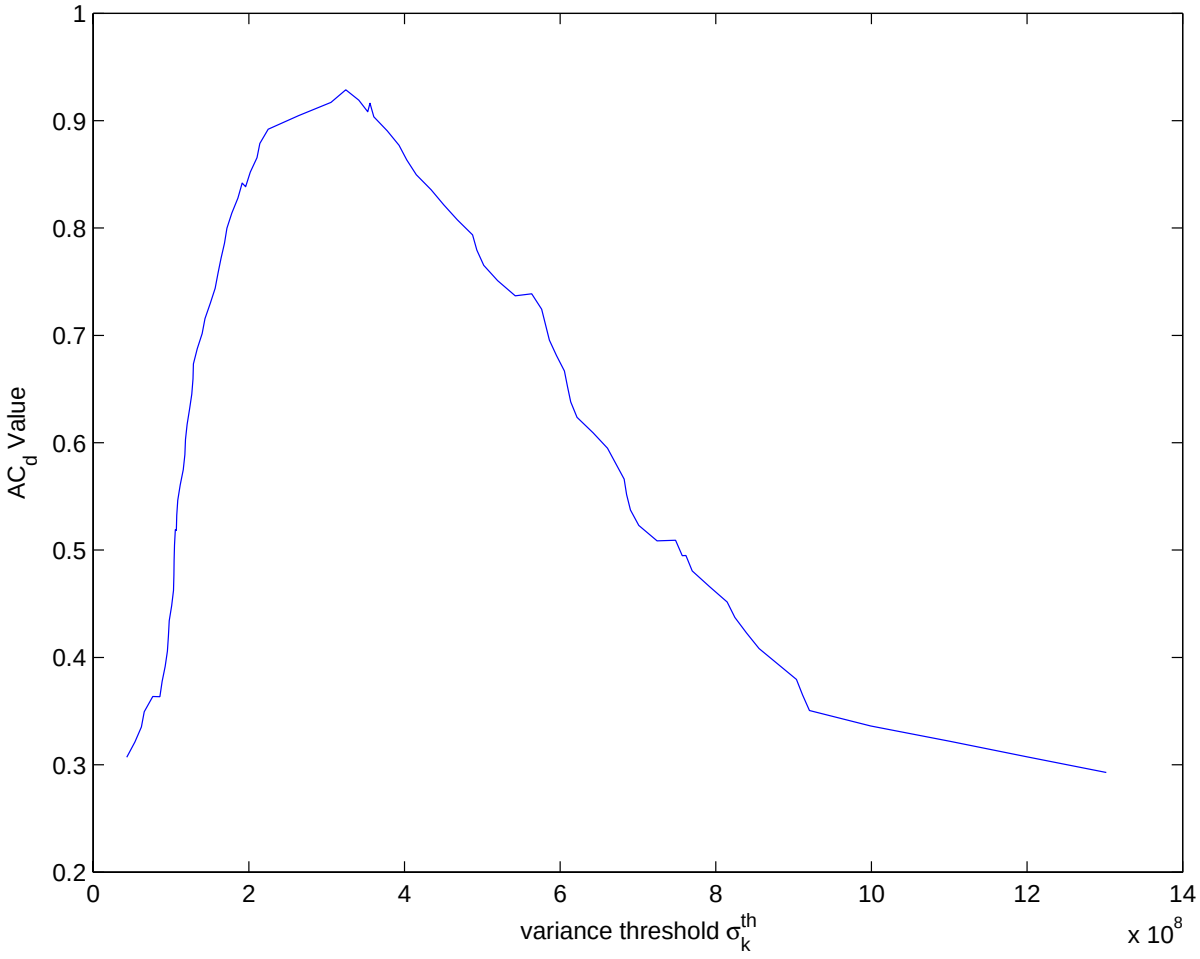

Figure 5.15: For the optimum $k = 600$, AC_d value as a function of the variance threshold σ_k^{th} when the GRAP index is used for the training set defined in section 5.5.

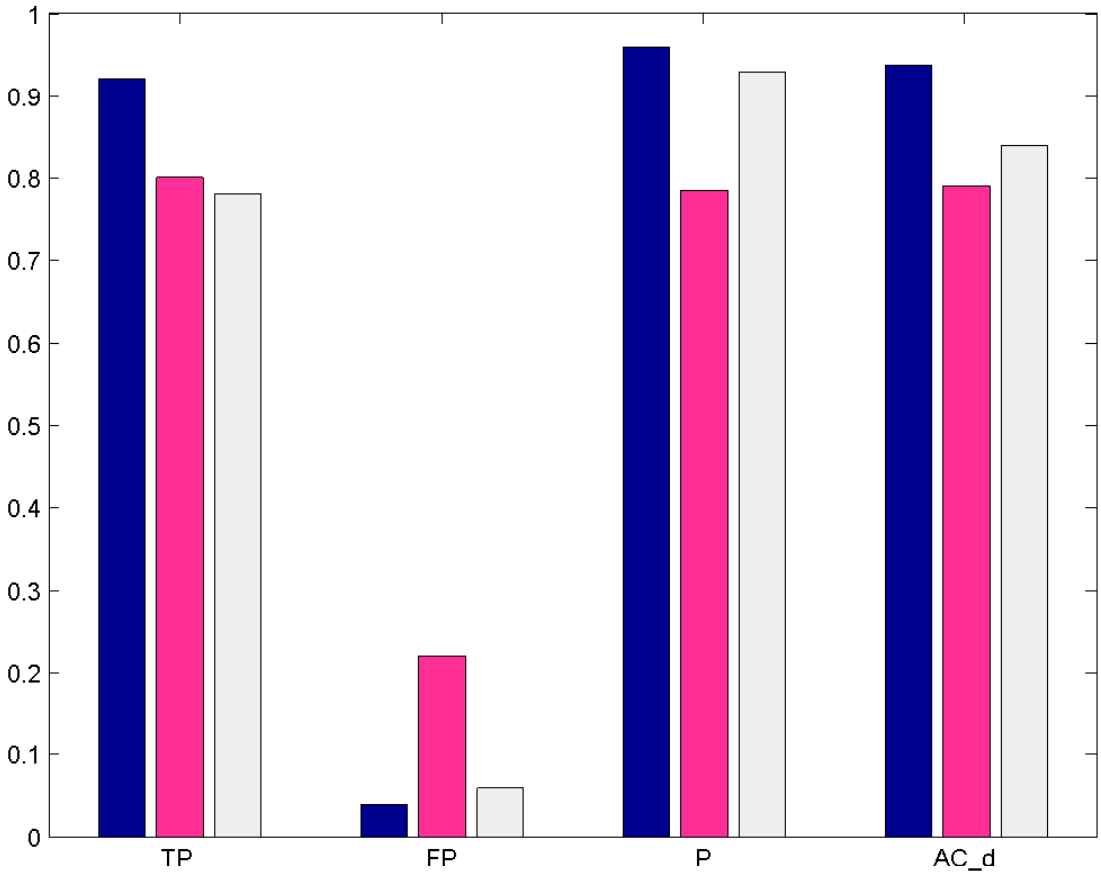

Figure 5.16: Testing results for index mapping: GRAP, ZIMP and KD (TP: True Positive, FP: False Positive, P: Precision, AC_d: performance measure of ROC point).

Chapter 6

CONCLUSION AND FUTURE WORK

Signal peptide detection is an important issue in the field of bioinformatics. The detection of the signal peptide helps localize the proteins to specific regions within the cell, and the knowledge of a specific signal peptide for a protein provides an important clue to its likely locations. In this thesis, we proposed a new method to detect the presence of signal peptides based on the space-frequency processing of the numeric amino acid sequences.

6.1 Contributions

- To analyze the protein sequence, we mapped the symbolic amino acid sequence to its equivalent numeric amino acid sequence. Analyzing the numeric amino acid sequence provides us the ability to detect the protein patterns which are too weak to be detected in their symbolic representations. In this work, we applied the space-frequency analysis method to perform pattern recognition in protein analysis. The Wigner-Ville transform is used to transform the signal from the space domain to the space-frequency domain.

- There are more than 400 index mappings available which map the symbolic amino acid sequences to the numeric amino acid sequences. Therefore, the selection of index mapping becomes an important issue. In this work, we proposed a method to select the index mappings which are suitable for the signal peptide detection.

- We developed a new technique to detect the presence of the signal peptide in an amino acid sequence. The numeric sequence is transformed to the space-frequency domain by the Wigner-Ville transform. We found that the amino acid sequence with signal peptide tends to have smaller variance in the space-frequency domain than the amino acid sequence without signal peptide. Based on this observation, we devised a new signal peptide detection algorithm which effectively differentiates the amino acid sequences with the signal peptide and the amino acid sequence without the signal peptide.

6.2 Future work

- In chapter 4, we proposed a method to select the index mapping which is suitable for signal peptide detection. To verify our selection criterion, we conducted the experiments with three different index mappings: GRAP, KD and ZIMP. In the future, we would like to do more experiments with more index mappings to test the robustness of our selection criterion.

- To learn the optimal parameters of the proposed algorithm, we used sequential search which may lead to a local optimal. In the future, we would like to investigate more sophisticated optimization algorithms to improve the performance of our method.

BIBLIOGRAPHY

[1] $http://www.ornl.gov/sci/techresources/human_genome/project/about.shtml$.

[2] G. R. Arce and S. R. Hasan. Elimination of interference terms of discrete wigner distribution using nonlinear filtering. 48(8):2321–2331, 2000.

[3] A. Bairoch and R. Apweiler. The swiss-prot protein sequence data bank and its supplement trembl. *Nucleic Acids Res.*, 25:31–36, 1997.

[4] S. A. Benner. Interpretive proteomics - finding biological meaning in genome and proteome database. *Advances in Enzyme Regulation*, 43:271–359, 2003.

[5] K. M. Bloch. Signal analysis methods for biological data. *Ph.D dissertation, Dept. of Electrical and Computer Engineering*, Spring, 2004.

[6] I. Cosic. The resonant recognition model of macromolecular bioactivity.

[7] I. Cosic. Macromolecular bioactivity: Is it resonant interaction between macromolecules? - theory and applications. 41(12):1101–1114, Dec 1994.

[8] I. Daubechies. Orthogonal bases of compactly supported wavelets. *Pure Appl. Math*, 41:909–996, 1988.

[9] Q. Fang and I. Cosic. Prediction of active sites of fibroblast growth factors using continuous wavelet transforms and resonant recognition models. In *The Inaugural Conference of the Victorian Chapter of IEEE EMBS*, pages 211–214, 1999.

[10] L. M. Gierasch. Signal sequences. *Biochemistry*, 28(3):923–930, 1989, Feb.

[11] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185:862–864, 1974.

[12] V. Heijne. Patterns of amino acids near signal-sequence cleavage sites. *Eur. J. Biochem.*, 133:17–21, 1983.

[13] V. Heijne. The structure of signal peptides from bacterial lipoprotein. *Protein. Eng.*, 2:531–534, 1989.

[14] K. Hiller, A. Grote, M. Scheer, R. Munch, and D. Jahn. Predisi: predition of signal peptides and their cleavage positions. *Nucleic Acids Research*, 32:375–379, 2004.

[15] L. Hunter. *Artificial Intelligence and Molecular Biology*. Illus. electronic text. ISBN 0-262-58115-9, 1993.

[16] S. Kawashima and M. Kanehisa. Aaindex: Amino acid index database. *Nucleic Acids Res*, 28(1):374, 2000.

[17] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1):105–32, 1982 May.

[18] I. Ladunga, F. Czako, I. Csabai, and T. Geszti. Improving signal peptide prediction accuracy by simulated neural network. *Bioinformatics*, 7:485–487, 1991.

[19] K. Nakai, A. Kidera, and M. Kanehisa. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.*, 2:93–100, 1988.

[20] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne. A neural network method for identification of prokaryotic and eukaryotic signal peptides and predition of their cleavage sites. *Int. J of Neural Syst.*, 8(5-6):581–599, 1997.

[21] H. Nielsen and A. Krogh. Prediction of signal peptides and signal anchors by a hidden markov model. In *Proc. Int. Conf. Intell. Syst. Mol. Biol. (ISMB 6)*, pages 122–130, 1998.

[22] F. J. Provost and R. Kohavi. Guest editors' introduction: On applied research in machine learning. *Machine Learning*, 30(2-3):127–132, 1998.

[23] S. Qian and D. Chen. Joint time-frequency analysis: Methods and applications. Prentice-Hall PTR, Upper Saddle River, NJ, 1996.

[24] M. Reczko, E. Staub, P. Fiziev, and A. Hatzigeorgiou. Finding signal peptides in human protein sequences using recurrent neural networks. In *Proceedings of the 2nd Int. workshop WABI 2002*, pages 60–67, Rome, Italy, 2002, September.

[25] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Report*, pages 27:379–423;623–656, 1948.

[26] K. Tomii and M. Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, 9:27–36, 1996.

[27] C. H. D. Trad, Q. Fang, and I. Cosic. Protein sequence comparison based on the wavelet transform approach. *Protein Engineering*, 15(3):193–2003, 2002.

[28] V. Veljkovic and I. Slavic. 'general models of pseudo-potentials.

[29] N. Zheng and L. M. Gierasch. Signal sequences: the same yet different. *Cell*, 86:849–852, 1996.

[30] J. M. Zimmerman, N. Eliezer, and R. Simha. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor Biol.*, 21(2):170–201, 1968 Nov.