COMPUTATIONAL PREDICTION OF PROTEIN SUBCELLULAR LOCALIZATION AND FUNCTION

by

MUTLU DOĞRUEL

Submitted to the Graduate School of
Engineering and Natural Sciences in
partial fulfillment of the
requirements for the degree of

Master of Science

Sabancı University

September 2002

COMPUTATIONAL PREDICTION
OF PROTEIN SUBCELLULAR
LOCALIZATION AND FUNCTION

APPROVED BY:

Assist. Prof. Dr. Uğur Sezerman ..……………………

(Thesis supervisor)

Prof. Dr. Hüveyda Başağa …………………….

Assist. Prof. Dr. Albert Levi ……………………..

DATE OF APPROVAL: ..……………………

ACKNOWLEDGEMENT

I am deeply indebted to my thesis supervisor Dr. Uğur Sezerman from Sabanci University, whose stimulating suggestions and encouragement helped me in all the time of research and for writing of this thesis. He provided a motivating, enthusiastic, and critical atmosphere during the many discussions we had. It was a great pleasure to me to conduct this thesis under his supervision.

I would like to express my gratitude to Prof. Hüveyda Başağa and Dr. Albert Levi who provided constructive comments on the prelimineary version of the dissertation.

I would like also to thank to Sabanci University's Academic Information Technology division coordinator Dr. Erkan Tekman for letting me to use their UNIX servers.

Finally, I want to express my special thanks to Ms. Elhan Özküçük for proof-reading my thesis.

To Sami Han, Melek, Figen, Nazlı, Semra, Serap and Yusuf Doğruel…

"Seviyorum; o halde varım!"

ABSTRACT

In this study, we present a computational approach in which it is possible to directly predict the protein functional categories from sequence and to identify the protein subcellular localization, which, in turn, is helpful for functional classification.

Subcellular protein locations and functions have been predicted basically from amino acid composition by using a machine learning approach. Expert systems based on Support Vector Machines have been designed to predict subcellular locations for proteins both in plants and nonplants, and function particularly for nonplants.

Four subcellular localization categories for plant and nonplant proteins have been identified by correct prediction accuracies of 95.4%, and 99.7% respectively. In addition to the three common categories mitochondrial, extracellular / secretory, and nuclear; the classes cytosolic for nonplants, and, chloroplast for plants are included.

Functional categories related to the subcellular compartments are predicted by using a similar approach applied for localization prediction. 92.9% of the 2321 protein sequences have been correctly assigned into the selected 10 functional categories.

Finally, the contribution of the data-mining of the MEDLINE papers to the function prediction is tested by another protein data set.

ÖZET

Bu çalışmada, proteinlerin hem fonksiyonlarının doğrudan bulunması, hem de fonksiyonlarının bulunmasında dolaylı olarak işe yarayan hücre içindeki yerlerlerinin saptanmasının mümkün olduğu iki işlemsel yaklaşım sunulmaktadır.

Proteinlerin hücre içindeki yerlerinin, içerdikleri amino asit oranları kullanılmak suretiyle, yapay zeka teknikleriyle saptanmaya çalışıldı. Bitkisel ve diğer proteinlerin hücre içindeki yerlerinin ve özellikle de ökaryotların fonksiyonlarının tespiti için Destek Vektörü Makineleri adı verilen yapay zeka uygulamasına dayanan uzman sistemler tasarlandı.

Bu sistemler kullanılarak, bitki ve diğer protein sınıfları için dörder hücre içi protein konumu, sırasıyla %95.4 ve %99.7 oranlarında doğru bir şekilde tahmin edilmiştir. Her iki grup için tahmin edilen mitokondri, hücredışı / sinyal ve nükleer sınıflarının yanı sıra, bitkiler için kloroplast, hayvanlar için ise sitozolik hücre konumları da sınıflandırmaya dahil edildiler.

Hücre içindeki organellerle ilgili faaliyet gösteren proteinlerin fonksiyonları, konum bulmada kullanılan yöntem kullanılarak tahmin edilmeye çalışıldı. 2321 protein dizisinin %92.3'ü, seçilmiş 10 fonksiyonel kategori içine doğru bir şekilde sınıflandı.

Son olarak, MEDLINE makalelerinin veri madenciliği ile analizinin fonksiyon tahminine katkı yapabileceği ayrı bir protein veri tabanı kullanılarak gösterildi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The complete genomes of several organisms have been determined in the last few years. There is a growing need for the rationalization of the available sheer mass of sequence information. The fundamental task of Bioinformatics is not only to derive more efficient means of data storage, but also to design more incisive analysis tools. Since subcellular location plays a crucial role in protein function, the availability of systems that can predict location from sequence will be essential to the full characterization of expressed proteins. Identifying the biological functions of these proteins is a key and challenging problem. Basically, the knowledge of subcellular localization of a protein greatly helps the biologist in determining its function. The presented computational technique will enable not only the prediction of subcellular localization, but also the identification of protein functions.

Prediction of subcellular localization sites of unannotated proteins is valuable in several ways. The knowledge of subcellular localization helps us to determine the possible processes with which a protein may be involved. Proteins and cellular functions, which, in turn, are determined by proteins can be related with specific cell compartments. For example, if a protein is located at the nucleus, its function is very likely to be related to nuclear organization and hence, to the DNA of the corresponding organism.

Beyond identifying possible functions of unknown or unannotated proteins, predicting the location information can alter the experimental approach to characterization a protein –e.g. purification [1]. It can also be used to screen candidate genes for drug discovery [2], and it enables us to automatically annotate the localization information for all hypothetical gene products identified in a genome (Eisenhaber and Bork, 1998).

The aim of this study can be summarized as predicting protein localization sites and functions by using their amino acid compositions and by using data-mining of MEDLINE scientific paper abstracts. The presented computational methods are based on an artificial intelligence learning tool called Support Vector Machines.

Section 2 starts with a very basic biological background. Support Vector Machines are introduced in the following section as well. The method and the computational experiments are mentioned in sections 3 and 4.

Section 3 has been devoted for protein localization prediction and Section 4 for functional classification. The thesis concludes with the results of these experiments and some discussions in the last section.

2. BACKGROUND INFORMATION

2.1 Biological Background

Proteins, the building blocks of all living organisms, are made up of the 20 types of amino acids. Proteins are synthesized within the cell from deoxyribonucleic acid (DNA) sequences which store all necessary genetic information. (Figure 2.1) The DNA alphabet is composed of four nucleotides, namely Adenine (A), Thymine (T), Guanine(G), and Cytosine (C). According to the central dogma of Microbiology, at first, the DNA sequences are replicated, then transcribed and finally, translated into amino acid sequences. (Figure 2.2) A group of three successive DNA nucleotides (called a codon) builds up a single amino acid. Due to the degeneracy in the genetic code, different codons may correspond to the same amino acid, as shown in Table 2.1.

All amino acids contain a carboxyl group (COOH) and an amino group (NH_2). The primary structure of a protein, that is, its amino acid sequence, is ordered from the amino group at the left hand side (called the 5' direction for DNA) toward the carboxyl group at right hand side (called the 3' direction for DNA). Other popular names are the N-terminus and the C-terminus for the start and the end regions, respectively. As it will be seen in the next sections, the analysis of the N-terminal sequences will be of great importance regarding the scope of this study.

The general chemical formula for an amino acid is given in Figure 2.3. In addition to these groups, a side chain (usually represented by "R") is attached to the central carbon atom. The nature of the side chain determines almost all important aspects of amino acids. Amino acids can be classified as nonpolar, polar & uncharged, and charged according to the chemical and physical characteristics of their side chains.

There are eight amino acids with nonpolar side chains. Glycine, alanine, and proline have small, nonpolar side chains and are all weakly hydrophobic (not "liking" water). Phenylalanine, valine, leucine, isoleucine, and methionine have larger side chains and are more strongly hydrophobic.

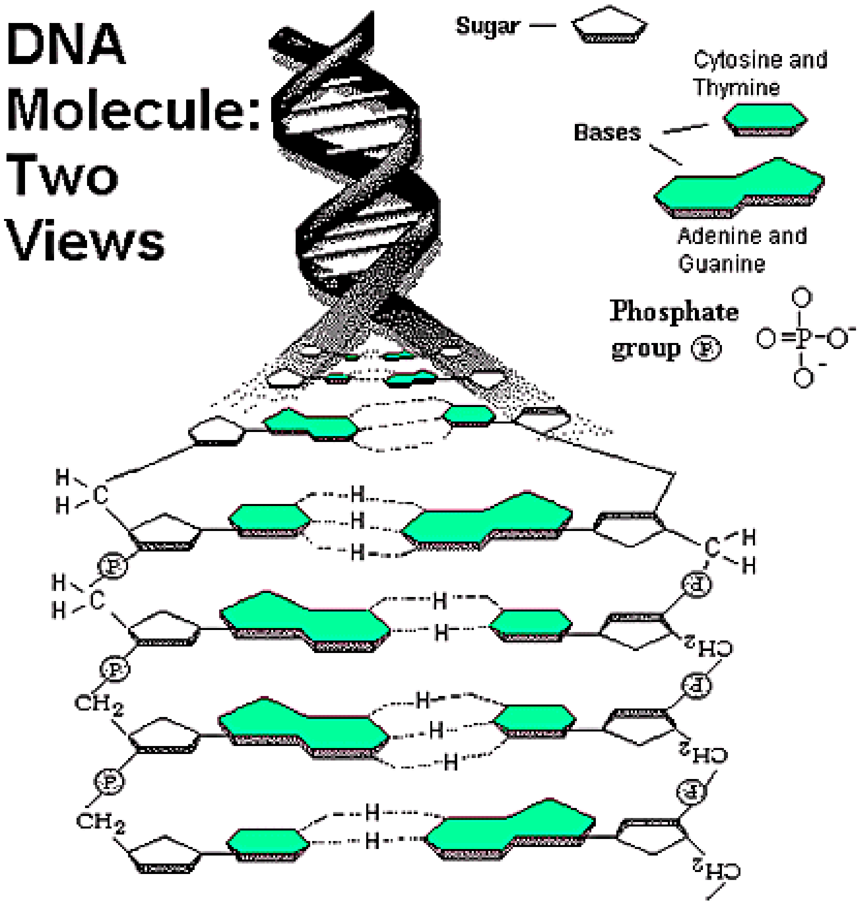


Figure 2.1 The structure of DNA and the double helix. (Source: National Health Museum, http://www.accessexcellence.org/AB/GG/)

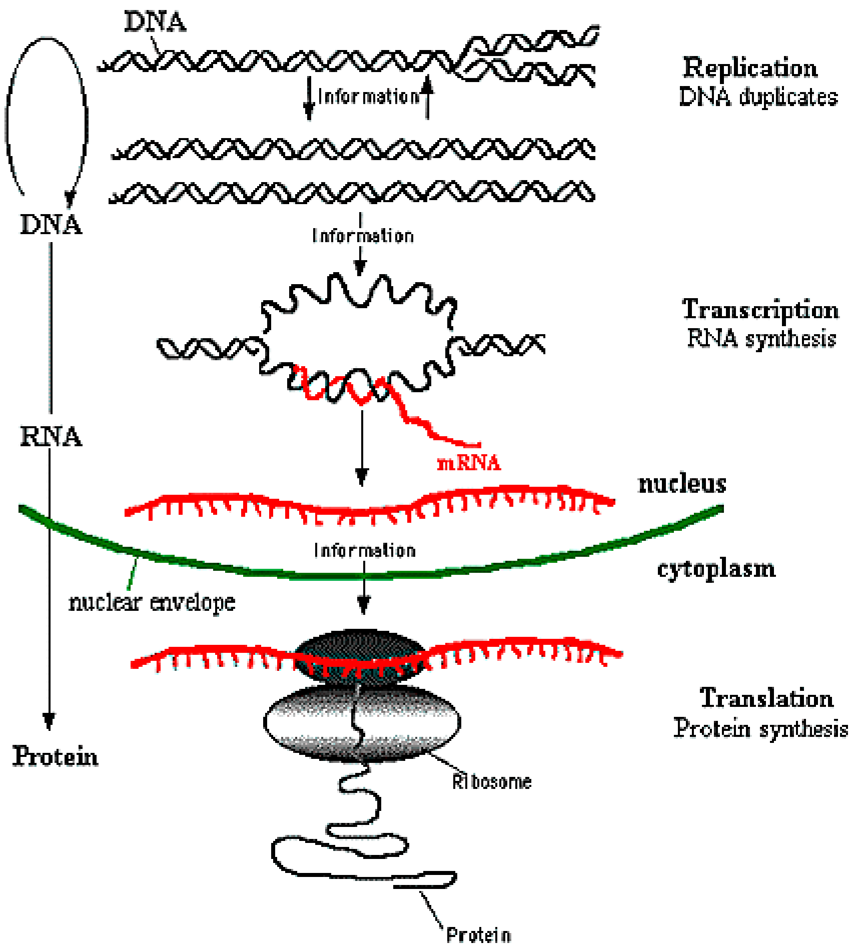


Figure 2.2 The Central Dogma of Molecular Biology (Source: National Health Museum, http://www.accessexcellence.org/AB/GG/)

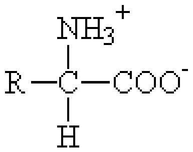


Figure 2.3 The general molecular formula for amino acids. R is the functional group (the side chain) of an amino acid.

Amino Acid	SLC	3LC	DNA codons
Isoleucine	I	ILE	ATT, ATC, ATA
Leucine	L	LEU	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	VAL	GTT, GTC, GTA, GTG
Phenylalanine	F	PHE	TTT, TTC
Methionine	M	MET	ATG
Cysteine	C	CYS	TGT, TGC
Alanine	A	ALA	GCT, GCC, GCA, GCG
Glycine	G	GLY	GGT, GGC, GGA, GGG
Proline	P	PRO	CCT, CCC, CCA, CCG
Threonine	T	THR	ACT, ACC, ACA, ACG
Serine	S	SER	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	TYR	TAT, TAC
Tryptophan	W	TRP	TGG
Glutamine	Q	GLN	CAA, CAG
Asparagine	N	ASN	AAT, AAC
Histidine	H	HIS	CAT, CAC
Glutamic acid	E	GLU	GAA, GAG
Aspartic acid	D	ASP	GAT, GAC
Lysine	K	LYS	AAA, AAG
Arginine	R	ARG	CGT, CGC, CGA, CGG, AGA, AGG
Stop codons			TAA, TAG, TGA

Table 2.1 The 20 amino acids with their single and 3 letter representations and the codons that make up them. Stop codons are listed as well.

There are also eight amino acids with polar, uncharged side chains. Serine and threonine have hydroxyl groups. Asparagine and glutamine have amide groups. Histidine and tryptophan have heterocyclic aromatic amine side chains. Cysteine has a sulfhydryl group. Tyrosine has a phenolic side chain. The sulfhydryl group of cysteine, phenolic hydroxyl group of tyrosine, and imidazole group of histidine all show some degree of pH-dependent ionization.

Finally, there are four amino acids with charged side chains. Aspartic acid and glutamic acid have carboxyl groups on their side chains. Each acid is fully ionized at pH 7.4. Arginine and lysine have side chains with amino groups. Their side chains are fully protonated at pH 7.4.

Living cells are usually categorized into two main groups -- prokaryotic and eukaryotic. This division is based on internal complexity. Organisms including animals and plants, which have more than one cell, are referred to as eukaryotes. The subcellular compartments of a typical eukaryotic species is displayed in Figure 2.4. These localization sites, called organelles, are the compartments where different vital functions of the cell are performed.

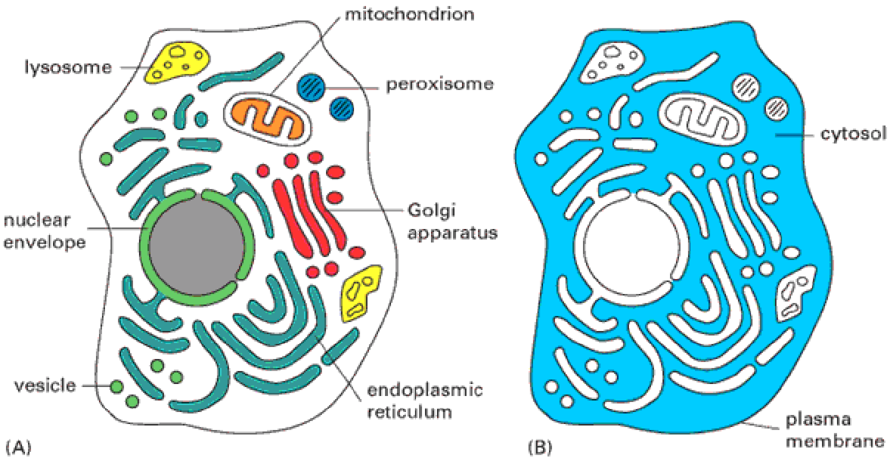


Figure 2.4 Eukaryotic cell. (Source: National Health Institute, http://www.accessexcellence.org/AB/GG/)

The major cell organelles and their primary functions (Figure 2.5) are as follows: Mitochondria are the cells' power sources. Energy-producing chemical reactions take place in Mitochondria. It also recycles and decomposes proteins, fats, and carbohydrates, and forms urea. The Golgi complex is the cells' packaging and shipping department. Endoplasmic reticulum is a tabular network fused into the nuclear membrane. It stores, separates, and serves as cell's transport system. The cytoskeleton supports cell and provides shape. It helps the movement of materials in and out of cells. Ribosomes are miniature protein factories. They compose one fourth of cell's mass. Vacuoles contain water solution. They store, digest and waste removal.

All these functions in the mentioned subcellular localizations are performed by specific proteins. It is very natural then to assume that proteins located in the Golgi apparatus, for instance, would most probably perform primarily some functions that are supposed to be carried out in the Golgi. Immediately after they are synthesized, proteins are sent to the relevant location where they would "perform" their functions. Therefore, the information needed for the proteins to be sent through some biological

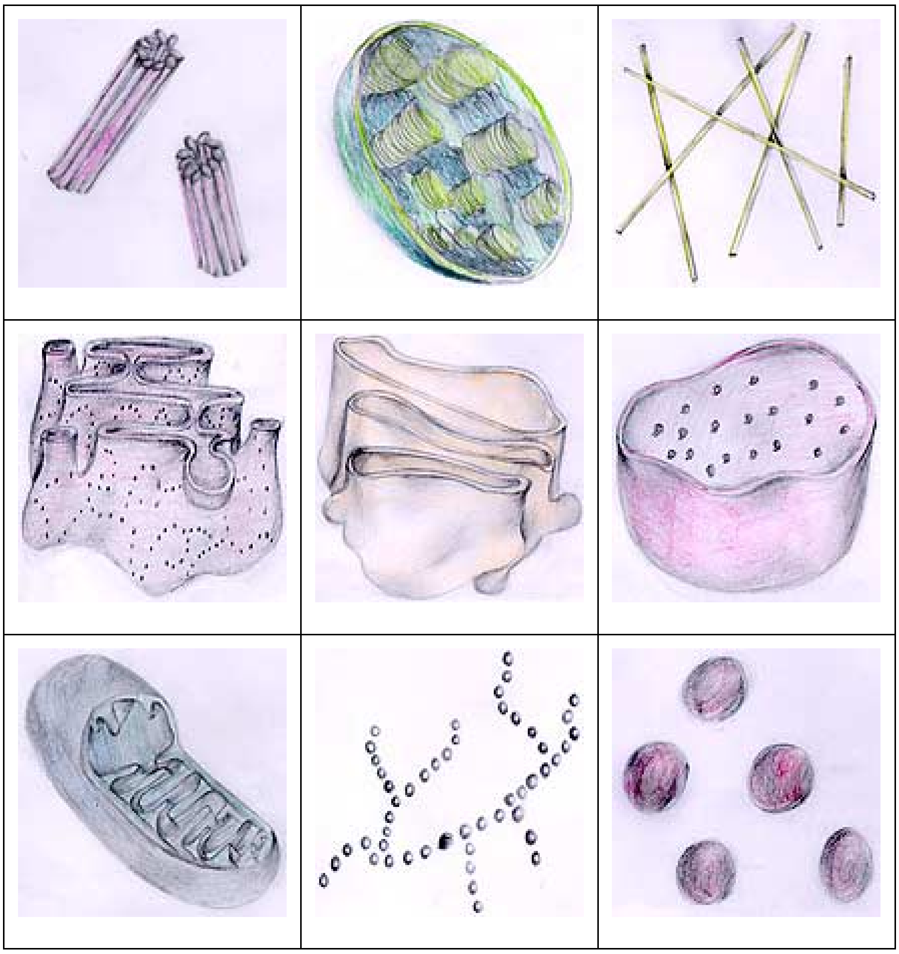


Figure 2.5 Organelles. From top left to bottom right: Centrioles, chloroplast, cytoskeleton, endoplasmic reticulum, Golgi apparatus, lysosome, mitochondria, ribosomes, and vacuoles. (Source: http://library.thinkquest.org/12413/structures.html)

pathways to the relevant location must somehow be encoded as an intrinsic signal on the protein.

Most of the proteins are synthesized in the cytosol [2] (Figures 2.4, 2.5, and 2.6). A small number of proteins are coded in the genomes of mitochondria and chloroplasts. Proteins need to be sorted to one or other subcellular compartment to perform their functions. Sorting usually relies on the presence of an N-terminal targeting sequence (Figure 2.7), which is proteolytically removed after entry [3]. For further sorting within the organelle, additional targeting information may be located in a secondary targeting sequence, either placed adjacent to the original targeting sequence or in other regions of the protein.

In most cases, a protein's subcellular localization is determined by some information encoded in its amino acid sequence [4]. Signal sequences (leader sequences) are the portion of the amino acid sequences that possesses the necessary encoded information to direct the protein toward or across the cytoplasmic membrane in prokaryotes, and the endoplasmic reticulum membrane in eukaryotes. Mitochondria targeting peptides (mTP), chloroplast transit peptides (cTP) for plants, and signal peptides (SP) are the typical N-terminal sorting signals. Contrary to what it has long been believed, recent studies have shown that there are several pathways for the translocation of proteins across the cytoplasmic / endosplasmic reticulum membrane.

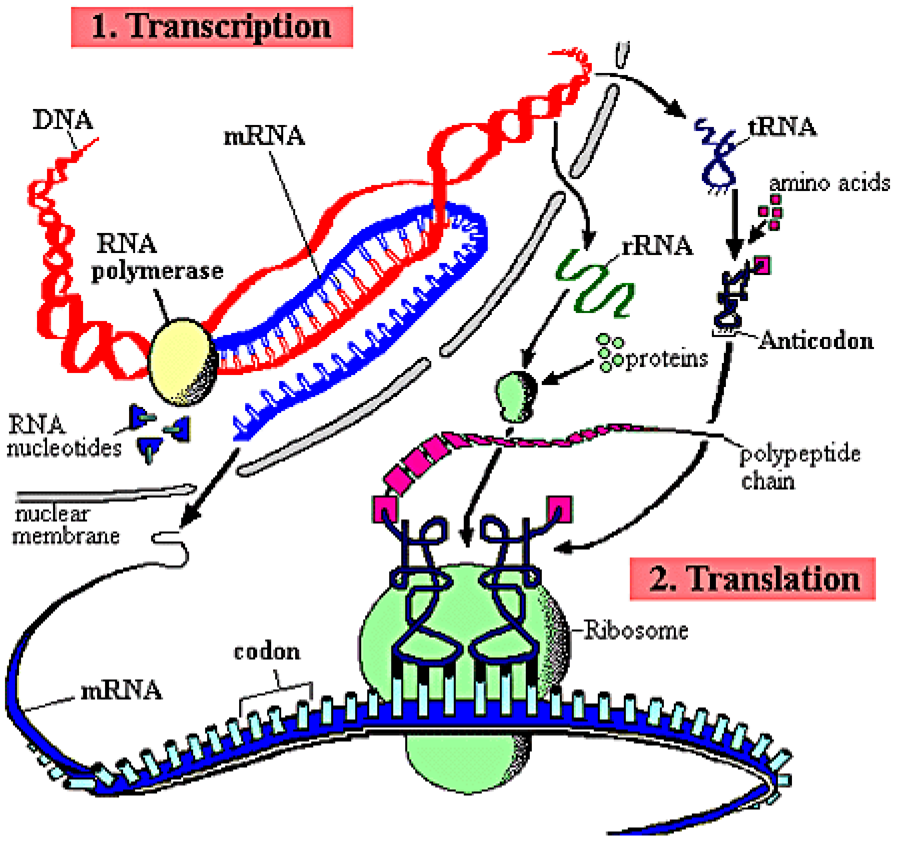
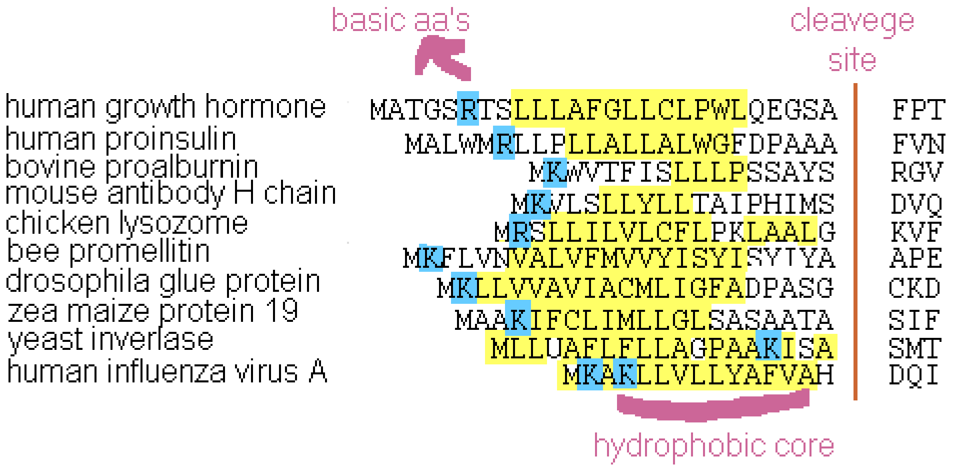


Figure 2.6 The protein synthesis process (National Health Museum, http://www.accessexcellence.org/AB/GG/)



human growth hormone	MATGSRTSLLLAFGLLCLPWLQEGSA	FPT
human proinsulin	MALWMRLLPLLALLALWGFDPAAA	FVN
bovine proalburnin	MKWVTFISLLLPSSAYS	RGV
mouse antibody H chain	MKVLSLLYLLTAIPHIMS	DVQ
chicken lysozome	MRSLLILVLCFLPKLAALG	KVF
bee promellitin	MKFLVNVALVFMVVYISYISYTYA	APE
drosophila glue protein	MKLLVVAVIACMLIGFADPASG	CKD
zea maize protein 19	MAAKIFCLIMLLGLSASAATA	SIF
yeast inverlase	MLLUAFLFLLAGPAAKISA	SMT
human influenza virus A	MKAKLLVLLYAFVAH	DQI

Figure 2.7 A set of N-terminal signal sequences, showing some common features.

Detection of these signal peptides is not straightforward, as there are many recognition factors. As a result, the molecular mechanisms related to signal peptides are rather complicated.

An important fact about protein translocation within the cell is that different protein types use different pathways to go to their final destinations. For example, most of the periplasmic and outer membrane proteins use the SecB-Dependent Pathway in which a cytosolic chaperone, SecB, first recognizes a target preprotein (i.e., a protein with a signal peptide) then with another helper protein, SecA (Schekman, 1994), the preprotein translocates across the cytoplasmic membrane through an aqueous gated pore [4]. The SRP-Dependent Pathway (SRP is a ribonucleoprotein complex) translocates mainly the inner membrane proteins in E. coli (Ulbrandt et al., 1997), the TAT-Dependent Pathway (TAT stands for twin-arginine translocation) is utilized by many proteins such as proteins binding iron-sulfer clusters (Berks, 1996). The features of each type of signals are summarized in Table 2.2.

In prokaryotes, the N-terminal sequence is usually 7 to 18 residues long and has a net charge of +1 to +7, but often with 1 or 2 negatively charged residues. Gram-negative eubacteria have the shortest N-terminal sequence while archaea organisms have the longest. [14]. The hydrophobic core in prokaryotes range in length from 9 to 18 residues. Usually the leucine content is high.

Obtaining protein sorting signals for eukaryotes is much more difficult than that of bacteria. Eukaryotic cells have an extensive internal membrane structure, as depicted in Figure 2.4. Unlike the prokaryotes, they have various membrane-bound compartments. The mitochondrion, for instance, has two membranes. These structures require simply more complicated pathways due to the need for the penetration of proteins through the target organelle membranes, which also makes it more difficult for us to discover these translocation pathways.

Eukaryotic signal sequences are similar in composition to prokaryotic signal sequences. However, they are longer than the prokaryotic signal sequences which have an average of 24 residues: they range from 18 to 80 residues. Endoplasmic reticulum has the shortest, and the stroma of the chloroplast in plants has the longest signal sequence [14,16].

Due to the insufficient motif data and the lack of information about the work processes of protein sorting pathways, alternative methods have been sought for. In 1994, Nakashima and Nishikawa suggested that intra- and extracellular proteins differ significantly in their amino acid composition and that these differences are strong enough to be used as the basis for a prediction method [5].

2.2 Support Vector Machines

Support Vector Machines (SVMs) are a very powerful machine learning class of algorithms, invented by Vladimir Vapnik and his co-workers, and were first introduced in 1992. However, the basic principles about SVMs have been already known and used in machine learning since the early sixties. SVMs have been used in Bioinformatics studies only for a couple of years.

SVM is an elegant tool for solving pattern-recognition and regression problems. Over the past few years, it has attracted many researchers from the neural network and mathematical programming community; the main reason for this being the ability of SVM to provide excellent generalization performance [8].

Support Vector Machines have several advantages over the traditional perception algorithm, as they will attempt to maximize the margin between data and therefore find a better generalizing solution. Figure 2.7 shows a comparison of a SVM unique solution to several solutions which could be produced by the perception algorithm.

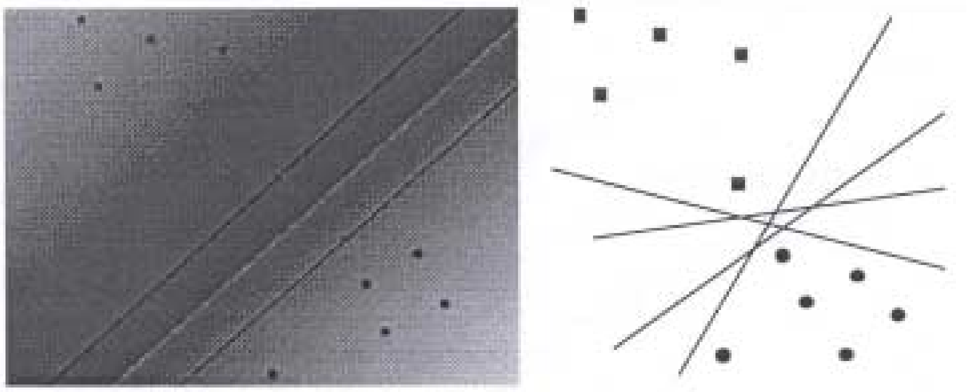


Figure 2.8 A Support Vector Machine Unique Solution (left); Several perception solutions (right).

2.3 A Simple Pattern Recognition Algorithm

The problem of assigning a new object into one of two classes is one of the fundamental problems of learning theory. This problem can be formalized as follows:

We are given some empirical data:

$$(x_1, y_1),...,(x_m, y_m) \in \chi \, x \, \{\pm 1\} \qquad (2.1)$$

Here, χ is a nonempty set from which the patterns x_i (also known as *cases, inputs, instances, or observations*) are taken, usually referred to as the *domain*; the y_i are called the *labels, targets,* or *outputs*. In this particular case, we have only two classes of patterns. They are labeled as +1 and -1. This case is referred to as *(binary) pattern recognition* or *(binary) classification*.

It is worthwhile to mention that the patterns could be just about anything, and that we have made no assumptions on χ other than being a set. For instance, the task might be to categorize sheep into the two classes *Karaman* (+1) or *Merinos* (-1), in which case x_i would be simply some features like color, geographic area, milk productivity, and so forth.

Signal	**Features**
SRP-dependent	18-26 amino acids (aa) in length; mostly positive N–region, hydrophobic h-region, and c-region harboring (-3,-1)
SRP-independent/ SecB-dependent	Similar to SRP-dependent, but length and/or hydrophobicity of h-region is smaller; also used at ER
TAT-Dependent	Longer at length (26-58aa); “twin-arginine” motif in n-region; not found at ER but at chloroplast
SPase II- dependent	Type II signal sequence; used for lipoproteins; “LA(G/A)C” motif for cleavage in c-region
Signal Anchor I	Few or no charges in n-region; longer h-region is favored than type II anchor
Signal Anchor II	(-3, -1) motif or longer h-region than signal peptides; positively charged n-region

Table 2.2 Examples of Nucleaocytoplasmic transport signals for Eukaryotes [4].

The studying of learning always involves generalizing to unseen data. In the case of sheep classification, this corresponds to finding $y \in \{\pm 1\}$, for a given new set of patterns $x \in \chi$. This is equivalent to estimating a function f, such that

$$f : \chi \rightarrow \{\pm 1\} \tag{2.2}$$

This is to mean that we choose y such that *(x,y)* is somewhat similar to the training examples (2.1). To this end, we need notions of similarity in χ. Nonetheless, determining this similarity measures is at the crux of machine learning studies.

Let us focus on the similarity measure of the form

$$\begin{aligned} &\mathbf{k} : \chi \,\mathbf{x}\, \chi \rightarrow \Re \\ &\quad (\mathbf{x},\mathbf{x}') \rightarrow \mathbf{k}(\mathbf{x},\mathbf{x}') \end{aligned} \tag{2.3}$$

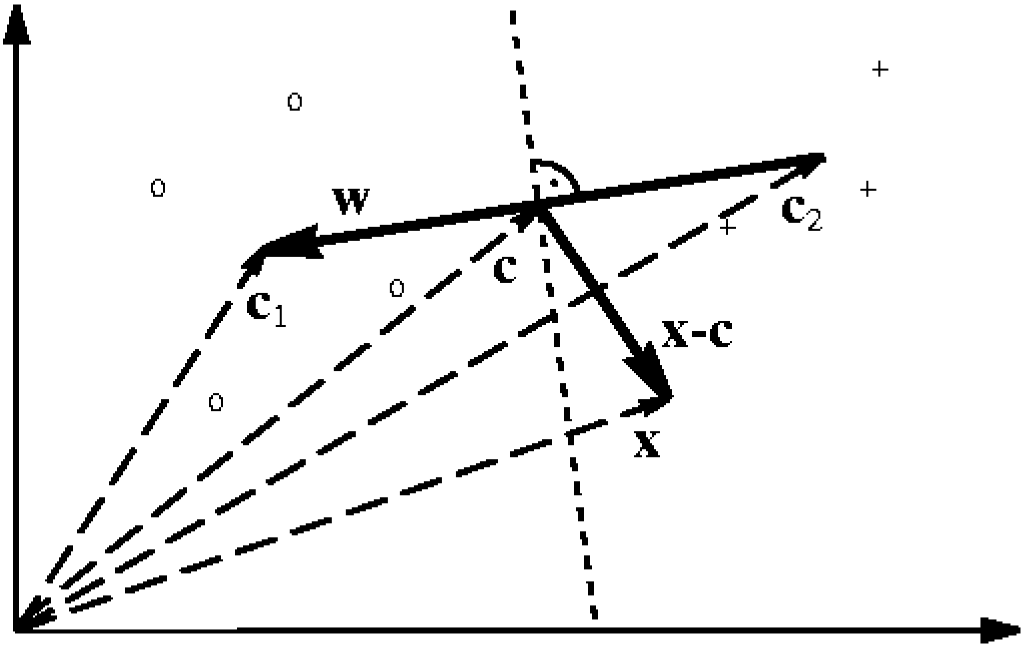


Figure 2.9 A simple geometric classification example: The decision boundary is orthogonal to w. (Schölkopf B., Smola A., Learning with Kernels, MIT press, 2002)

This is indeed a function returning the real number characterizing the similarity of the patterns x and x`. The function k is called a kernel. Since general similarity measures of this form are rather difficult, let us start from a particular simple case:

The new test pattern will be assigned to the class with closer mean. Therefore, we need to know the means of the classes:

$$c_{+} = \frac{1}{m_{+}} \sum_{\{i|y_i=+1\}} x_i \tag{2.4}$$

$$c_{-} = \frac{1}{m_{-}} \sum_{\{i|y_i=-1\}} x_i \tag{2.5}$$

m_+ and m_- show the number of examples with positive and negative labels, respectively. We assume that $m_+ >0$ and $m_- >0$. In order to assign a new point **x** to the class with the closest mean, we will employ the dot product to formulate this geometric construction. Let the vector c denote the half way between c_+ and c_- :

$$c = (c_{+} + c_{-})/2 \tag{2.6}$$

In fact, we check whether the vector *x-c* makes an angle with the vector

$$w = c_{+} - c_{-} \tag{2.7}$$

smaller than 90°, as shown in Figure 2.8. If the dot product of *x-c* and *w* is positive (or negative), then the test point ***x*** is said to be belonging to class +1 (or -1). This leads us to defining

$$\begin{aligned} y &= \operatorname{sgn}\langle (x-c), w\rangle \\ &= \operatorname{sgn}\langle (x-(c_{+}+c_{-})/2), (c_{+}-c_{-})\rangle \\ &= \operatorname{sgn}(\langle x, c_{+})\rangle - \langle x, c_{-}\rangle + b) \end{aligned} \tag{2.8}$$

,where the offset term is

$$b = \frac{1}{2}(\|c_{-}\|^2 - \|c_{+}\|^2) \tag{2.9}$$

with the norm being defined as

$$\|x\| = \sqrt{\langle x, x\rangle}\ . \tag{2.10}$$

If the means of the two classes have the same distance to the origin, then *b* will vanish [11].

Equation 2.8 corresponds to a linear decision boundary which has the form of a hyperplane, as shown in Figure 2.8. To obtain the decision function more explicitly, we substitute (2.4) and (2,5) into (2.3):

$$\begin{aligned} y &= \operatorname{sgn}\left(\frac{1}{m_+} \sum_{\{i|y_i=+1\}} \langle x, x_i \rangle - \frac{1}{m_-} \sum_{\{i|y_i=-1\}} \langle x, x_i \rangle + b \right) \\ &= \operatorname{sgn}\left(\frac{1}{\mathrm{m}_+} \sum_{\{i|y_i=+1\}} k(x, x_i) - \frac{1}{\mathrm{m}_-} \sum_{\{i|y_i=-1\}} k(x, x_i) + b \right) \end{aligned} \tag{2.11}$$

In a similar way, the offset can be rewritten as:

$$\mathrm{b} = \operatorname{sgn}\left(\frac{1}{\mathrm{m}_-^{\;2}} \sum_{\{(i,j)|y_i=y_j=-1\}} k(x_i, x_j) - \frac{1}{\mathrm{m}_+^{\;2}} \sum_{\{(i,j)|y_i=y_j=+1\}} k(x_i, x_j) \right) \tag{2.12}$$

The above example shows the use of a linear function as the kernel. For more complicated problems different kernel functions with different degrees may be used. The most commonly used kernel functions are listed in Table 2.3.

1. linear	$\mathrm{u'v}$
2. polynomial	$(\gamma \mathrm{u'v} + \mathrm{coef})^{\mathrm{degree}}$
3. radial basis	$\mathrm{e}^{-\gamma\lvert \mathrm{u}-\mathrm{v}\rvert^2}$
4. sigmoid	$\tanh(\gamma \mathrm{u'v} + \mathrm{coef})$

Table 2.3 The most commonly used kernel functions in SVMs

To summarize, for classification, SVMs operate by finding a hypersurface in the space of possible inputs. This hypersurface will attempt to split the positive examples from the negative examples. The split will be chosen to have the largest distance from the hypersurface to the nearest of the positive and negative examples. Intuitively, this makes the classification correct for testing data that is near, but not identical to the training data [12]. More information can be found in Burges' tutorial that is available at http://svm.research.bell-labs.com/SVMdoc.html or in Vapnik's famous book *Statistical Learning Theory*. (Vapnik, V., Wiley-Interscience, New York, 1998).

2.4 Previous Works

Cell fractionation, electron microscopy and fluorescence microscopy are the three main experimental techniques applied to determine the subcellular location. Nevertheless, these approaches are time consuming, subjective, and highly variable [3].

In fact, without sequence homology to some other proteins which have been already studied elaborately in terms of function and structure, predicting the function of a novel protein by computational tools is an extremely difficult task, if not merely impossible. There have been a lot of -several ongoing- attempts to predict the subcellular localization of proteins by incorporating some information like the known protein sorting signals (works of Nakai in 1991, 1999, and 2000), which turns out to be very helpful in determination of the function.

Since most of the “most information possessing” part about the sorting process of a protein is cleaved off after it is translocated within the cell, in recent years several prediction methods have been devised involving the analysis of the “remnants” of signal peptides. Günter Blobel proposed that "proteins have intrinsic signals that govern their transport and localization in the cell.", a discovery that brought him the Nobel Prize in Physiology/Medicine in 1999. Several other attempts have been made to discover and to make use of these signaling sequences throughout the “mature protein” sequence of a protein.

There are several internet sites on predicting the subcellular location of a protein. The most significant ones are listed in Table 2.4.

PSORT is an expert system that is used to predict subcellular location of eukaryotic proteins. In its first versions, signal sequence information is used. It is capable of predicting several different cellular localizations including the cytoplasm, nucleus, mitochondria sites, peroxisome, ER sites, Golgi, lysosome sties, plasma membrane sites, extracellular space, and chloroplast sites, with a general accuracy of 59.4%[18]. It is an expert system using many if-then rules and very good for its property of mimicking the actual molecular sorting process. On the other hand, PSORT is not regarded very effective due to the ever-increasing number of protein sequences that do not have signal sequences at all. Currently, there are three versions of PSORT: PSORT II for prediction of the cellular locations of animal/yeast proteins, old PSORT for bacterial and plant sequences, and iPSORT for the detection of N-terminal sorting signals. The new versions incorporating the amino acid composition are capable of predicting locations with accuracies 69.8 % for plants, and 83.2% for non-plants.

Neural networks have been used to predict location as well, as they are very convenient in that the network teaches itself during a training period and no preconceived model is required. Nonlinear correlations can be predicted and the quality of predictions can be improved with the use of new/updated databases of nonhomologous data sets.

TargetP is the most recent and the most successful neural network prediction system developed by Emanuelsson et al.[15]. Using N-terminal sequence information only, it discriminates between proteins destined for the mitochondrion, the chloroplast (for plants only), the secretory pathway, and “other” localizations. The used input data consist of mitochondrial targeting peptides (mTPs), signal peptides (SPs), chloroplast transit peptides (cTPs), and some nuclear and cytosolic sequences designated as “other”.

Name of Server	Website	Feature
PSORT	http://psort.nibb.ac.jp/	Sorting signal knowledge
TargetP	http://www.cbs.dtu.dk/services/TargetP/	By discriminating the individual targeting signal peptide
MitoProt	http://bioinformer.ebi.ac.uk/newsletter/archives/2/mitoprotii.html	By discriminating mitochondrial and chloroplast signal peptide
Predotar	http://www.inra.fr/Internet/Produits/Predotar	By discriminating mitochondrial, chloroplast signal peptide
NNPSL	http://predict.sanger.ac.uk/nnpsl	By amino acid composition
SobLoc	http://www.bioinfo.tsinghua.edu.cn/SubLoc	By amino acid composition
EukProL **ProProL**	http://tubic.tju.edu.cn	By more sequence information besides the amino acid composition

Table 2.4 Subcellular localization prediction web sites and their features

It is built from two layers, where the first one contains one dedicated network for each type of presequence used. For non-plant eukaryotes (i.e., yeast), the overall prediction accuracy is 90.0% for the three categories (mTP, SP, and “other”), and 85.3% for the four plant categories (cTP, mTP, SP, and “other”). TargetP prediction performance for plants and for non-plants is given in Tables 2.5, and 2.6 respectively.

Hua et al. [7] recently devised a prediction method that is based on amino acid composition. They attained a 91.4% prediction success rate for the three classes for nonplant proteins.

		Predicted				
True category	Number in category	cTP	mTP	SP	Other	Sensitivity
cTP	141 (140)	120 (119)	14 (14)	2 (2)	5 (5)	0.85 (0.85)
mTP	368 (140)	41 (18)	300 (109)	9 (3)	18 (10)	0.82 (0.78)
SP	269 (140)	2 (0)	7 (2)	245 (132)	15 (6)	0.91 (0.94)
other	162 (135)	10 (5)	13 (9)	2 (5)	137 (116)	0.85 (0.86)
Specificity		0.69 (0.84)	0.90 (0.81)	0.96 (0.93)	0.78 (0.85)	

Table 2.5 TargetP prediction performance for plants, in actual numbers, on redundancy reduced non-equalized (size-equalized in parenthesis) test sets [17]

		Predicted			
True category	Number in category	mTP	SP	Other	Sensitivity
mTP	371 (370)	330 (330)	9 (8)	32 (32)	0.80 (0.89)
SP	715 (370)	13 (6)	683 (354)	19 (10)	0.96 (0.96)
Other	715 (370)	152 (47)	49 (8)	1451 (315)	0.88 (0.85)
Specificity		0.67 (0.86)	0.92 (0.96)	0.97 (0.88)	

Table 2.6 TargetP prediction performance for non-plants, in actual numbers, on redundancy reduced non-equalized (size-equalized in parenthesis) test sets [17]

3. SUBCELLULAR LOCALIZATION PREDICTION BY SMVs

3.1 PREDICTION APPROACHES

There are two major approaches for predicting the localization sites: Either to search for some known motifs in N-terminal sorting signals or to use the amino acid composition rates of proteins. The known motifs tell us that prokaryotes and eukaryotes have different mechanisms of sorting signals. Beyond that, N-terminal signals vary from one species to the other, and they show variations over different protein families as well.

Although it can not be regarded as a concrete consensus sequence [4], the tripartite structure composed of an n-region, an h-region, and a c-region [13,14] is the classical feature seen in the N-terminal region (Figure 3.1). The n-region is positively charged, the h-region is hydrophobic and the c-region is mainly composed of polar amino acids [15].

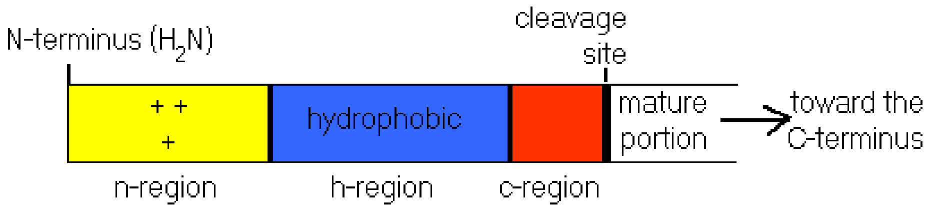


Figure 3.1 The three portions of N-terminal signal peptides (Adapted from Nakai 2000)

In mTPs, Arg, Ala and Ser are abundant while the negatively charged amino acid residues Asp and Glu are rarely seen [16].

In large genome analysis projects genes are usually automatically assigned and these assignments are often unreliable for the 5'-regions [6]. Usually, this results in the leader sequences to be missed or only partially included, which, in turn, causes problems for prediction algorithms depending solely on the N-terminal signals [6,7]. In other words, the prediction methods based on the recognition of the protein N-terminal sorting signals are strongly dependent on the quality of the gene 5'-region or protein N-terminal sequence assignment in databases.

This project focuses on a novel approach based on the combination of these techniques: the use of amino acid compositions of N terminal. These rates, along with the class labels (targets) are provided as input vectors to train the SVM using different parameters. Since N-terminal protein sequences have valuable motifs with the necessary information needed for the translocation of proteins, the amino acid composition of the first few sequences have been tested, and it has been seen that they provide sufficient information about the localization site prediction. We proved that the composition rates of the first 20 amino acids from the N terminal sequence is sufficient for providing an almost excellent prediction accuracy rates as high as 99.7% with the use of appropriate SVM parameters.

Apart from predicting the subcellular localizations, we tried to estimate the functional categories of proteins. Since the localization information is not a sufficient argument in prediction of functions, we included some structural information and key word frequency parameters extracted over some textual analyses of a set of related biological papers to increase the classification accuracy.

Two SVM applications have been used: BSVM 2.03 by Chih-Wei Hsu and Chih-Jen Lin (2002), which can be freely downloaded for academic use from http://www.csie.ntu.edu.tw/~cjlin/bsvm/ and, SVMlight by Joachims (1999), which is available at http://ais.gmd.de/~thorstes/svm-light/. BSVM, unlike SVMlight can do multiple classification. Some of its parameters to be adjusted by the user are given in Table 3.1.

-c cost	Set the parameter C of support vector machine (default 1)
-d degree	Set degree in kernel function (default 3)
-e epsilon	Set tolerance of termination criterion (default 0.001)
-g gamma	Set gamma in kernel function (default 1/k)
-h shrinking	whether to use the shrinking heuristics, 0 or 1 (default 1)
-m cachesize	Set cache memory size in MB (default 40)
-p epsilon	set the epsilon in loss function of support vector regression (default 0.1)
-q qpsize	set the sub-problem size for -s 0,1 and 3
-r coef0	Set coef0 in kernel function (default 0)

-s svm_type	set type of SVM (default 0) 0 -- two-class bound-constrained support vector classification 1 -- multi-class bound-constrained support vector classification 2 -- multi-class support vector classification from Cram and Singer 3 -- bound-constrained support vector regression
-t kernel_type	set type of kernel function (default 2)
0 – linear	U'*v
1 – polynomial	(gamma*u'*v + coef0)^degree
2 -- radial basis function	Exp(-gamma*|u-v|^2)
3 – sigmoid	tanh(gamma*u'*v + coef0)

Table 3.1 BSVM parameters

SVMligth, which is used widely in scientific research, is a very powerful binary classifier. Most of the current SVM studies are conducted by SVMLight. Almost all SVM approaches for different problems in Bioinformatics are tested by SVMLight.

3.2 LOCATION PREDICTION USING AMINO ACID COMPOSITION FOR NONPLANTS

We downloaded the protein sequences that Emanuelsson et al. [17] used in their work, from http://www.cbs.dtu.dk/services/TargetP/. They extracted the data from SWISS-PROT (Bairoch & Apweiler, 2000). To avoid problems related to redundant data during neural network training and testing, they removed the inappropriate sequences. In order to increase the size of the data sets as far as possible, sequences annotated as "potential", "by similarity", or "probable" were included as well.

To train the SVM, we typically used a radial-base kernel function that yielded the best performance in previous works. In the test phases, cross validation has been applied. In an n-cross validation test, the data is divided into n sets, and for each test, one portion is used for testing, while the others are used for training. All training sets and all test sets have equal number of randomly distributed samples from each class. Unless stated otherwise, in all different configurations, for all different input representations and for all SVM parameters we used, the prediction success has been tested over 5-fold cross validation sets. For each experiment, the mean of the correct prediction (true positives and true negatives) percentage for the 5-fold cross validation tests is given.

Experiment 1

As a first attempt, amino acid residue counts and motif information have been used. This would enable us to classify proteins more accurately as we would be merging the most common two important distinguishing features. To capture as many motifs as possible, we grouped amino acids into 6 classes according to some characteristics like hydrophobicity, physical properties etc., and computed the occurrence rates of all possible amino acid residue triplets (Table 3.2) to incorporate the neighborhood information as well. This way, we would be representing the abundance rates of amino acids, and presumably several motifs, like the mitochondrial intermediate peptidase cleavage site consensus sequence "RX | (F / L / I) X_2 (T / S / G)

X_4", among the several ones mentioned by Nakai [4]. Table 3.2 depicts the amino acid classification we used in Experiment 1.

Group number	Amino acid residues
1	ILVM
2	TSNQ
3	EDKRH
4	WFYP
5	C
6	GA

Table 3.2 Amino acid classes used in Experiment 1

A dataset of vectors consisting of 216 values and a location class label for the corresponding protein has been built. All occurrence values in a vector have been divided by the length of the protein. The first entry, for instance, is the normalized frequency of the amino acids represented by "1-1-1", and so forth. Table 3.3 shows the average amino acid residue number and the total number of sequences for each protein class we use. The average protein sequence lengths show significant variations from one class to another, and this difference can be used as a distinguishing feature. To test this observation, both a dataset with vectors of normalized occurrence (frequency / length) values and a dataset with vectors of pure occurrence counts have been used. For the training processes of both sets, the same SVM parameters and the same radial-base kernel function have been used.

The normalized set had a poor prediction accuracy of 60.5%. On the other hand, using the unnormalized set, out of the 2738 non-plant protein sequences, 69.8% were classified correctly by BSVM as "mTP", "SP", or "other". Experiment 1 showed us that the number of amino acids in a protein sequence could be a better measure than the composition rate, as it is somehow related to the length of the protein. Surprisingly this fact has not been noticed before, in any prediction method based on composition.

Protein class	Number of non-plant protein samples	Number of plant protein samples	Average sequence length for non-plant proteins	Average sequence length for plant proteins
mTP	371	368	1032.2	1033.6
SP	715	269	1110.1	782.6
Nuclear	1214	54	1524.7	971.7
Cytoplasmic	438	108	1476.5	1041.8
cTP	N/A	141	N/A	973.0

Table 3.3 Total number of protein sequences and average sequence lengths for plant and non-plant protein sets used

Experiment 2

This time, we used amino acid counting without any grouping. That is, for each of the 20 amino acid residues (Table 2.1), the occurrence times have been counted. This input scheme yielded a 70.5% prediction accuracy, slightly better than the previous configuration. The kernel function used is a radial-base kernel function with the ”gamma” parameter being 0.02 that gave the best performance. With the default gamma, which is the reciprocal of the true class numbers, the prediction accuracy read 70.4%. Altering the capacity parameter “C” did not result in any significant change. We repeated the experiment by the normalized counts, and saw that the accuracy dropped to 61.2% with the same SVM parameters. Setting the “s” parameter of BSVM to 2 (see Table 3.1), however, increased this figure up to 70.3%, which is still slightly less than the above prediction rate. This experiment once again verified that pure counts yield better prediction performance. Composition rates are not used in other experiments.

Experiment 3

Thomas and Dill [19] suggested several amino acid classification schemes based on contact potentials. In this experiment, we used their 5-class grouping (Table 3.4). The frequencies of all possible consecutive four-residue "words" were used for the formation of 625 dimensional input vectors. The composition of the quadruple strings, however, did not fulfill our expectations as the result was 70.4% at the best.

Group number	Amino acid residues
1	VILMFWYA
2	GPSTHQN
3	C
4	ED
5	RK

Table 3.4 Amino acid classes used in Experiment 3

Experiment 4

Thomas et al.'s [19] 10-class classification (Table 3.5) was more affirmative than the 5-class classification, when we worked out with three-letter words. The 1000 dimensional dataset gave a better result as high as 79.6%.

Group number	Amino acid residues
1	VILMF
2	HQN
3	C
4	ED
5	RK
6	A
7	G
8	WY
9	P
10	ST

Table 3.5 Amino acid classes used in Experiment 4

Experiments 5 and 6

The full-length sequences were analyzed in terms of amino acid composition according to only the amino acid groups in Table 3.4 and 3.5, with no neighborhood information supplied. The 10-dimensional sequence set produced a 70.4% success in assigning proteins into the correct class. This figure was improved to 70.7% by turning on the two-class bound-constrained support vector classification parameter of BSVM. The 5-dimensional input representation had almost the same correct prediction rate of 70.4%. Surprisingly, we were able to increase this figure up to 70.5% by the use of a sigmoid kernel (tanh(gamma u v + coef)), which have performed the worst in all other experiments.

Experiment 7

This study in effect aims to study the contribution of the N-terminal sequences in determination of cellular location of proteins. With this goal in mind, the previous experiments have been carried out for enabling us to do some comparisons among datasets in which full-length sequences were used and in which only a number of N-terminal sequences were used.

In this particular experiment, only the first 60 amino acid residues from the N-terminal were used. Just like Experiment 4, which gave a relatively good result, a 1000-dimensional dataset was used. Amino acids have been grouped into 10 classes as in Table 3.5. For each sequence, frequencies of all possible three-letter words have been recorded. As the protein sequence lengths are fixed in this case (the first 60 residues), there was no need to normalize the frequency terms at all. The prediction accuracy that is tested over 5 different datasets as usual, however, was not very promising: The 70.7% correct prediction rate was far from being close to what we got in Experiment 4. Thus, the 10-class amino acid grouping turned out to be not suitable when only the composition of the 60 residues was provided. However, it is worthwhile to note that even though only the first 60 amino acids were taken into account, the 70.7% accuracy was a better result than Experiment 7's 70,4%. Thus, the amino acid composition information coming from the N-terminal region is worth to analyze. Experiment 8 is designed for this purpose.

Experiment 8

It has been well known that eukaryotic signal sequence lengths range from 18 (shortest in ER) to 80 residues (in stroma of the chloroplast in plants) versus the average of about 24 in prokaryotes [14,16]. Emanuelsson et al. [17] have used the 100 N-terminal amino acids to feed their two-layer neural network system in different input window sizes for each protein class. Most probably, they chose to use especially the first 100 residues in order to cover as much as possible signal motifs whose lengths may be as long as 80 residues, as stated above. On the other hand, in all studies based on composition, the entire sequences have been considered. With the positive result we obtained in Experiment 7, we further reduced the size of sequences to see how the datasets consisting of fewer residues would perform. Only the first 30 amino acids from the N-terminus were considered. Frequencies of the amino acid groups in Table 3.2 were computed, that is the input vector length was 216 for each protein. The 5-fold cross validation method yielded a brilliant 80.6% true prediction rate, with the capacity term set to 1.5, gamma parameter 0.02, s 1, and the kernel function being a radial-base. All following experiments employ a radial base kernel.

Experiment 9

The 1000-dimensional input set used in this experiment, which is formed by 2738 protein sequences, is based on the amino acid classification given in Table 3.5. The result was slightly less than the previously obtained prediction performance: 80.3%.

Experiment 10

So far the best true prediction rate has been attained in this experiment. By using the 20 amino acid occurrence counts for the first 30 amino acid residues only, we were able to predict the three localization sites correctly by 88.2%. Therefore, the tendency of correct prediction accuracy to increase as the number of amino acid residue number is decreased, is proved to be still valid for the 30 residues from the N-terminus. Naturally, in the next experiments we seek if this trend continues.

Experiment 11

In this experiment, only the occurrence counts of the 6 amino acid groups given in Table 3.2 were used. The frequency terms were computed only for the first 20 residues, which constituted the training and test datasets of 216 dimensional input vectors. We were able to classify the protein sequences correctly into one of the three location classes by an 80.6% success rate, which is the same figure as what we obtained in Experiment 8, with only the first 30 residues being used to set up the 216 dimensional vectors. Using the same SVM parameters and the same input format, the protein sequences were classified into the four localization categories with an accuracy of 67.8%.

Experiment 12

The performance of the 1000-dimensional dataset with the first 20 residues used was recorded to be 79.2%.

Experiment 13

The best prediction scores have been obtained using directly the frequencies of the 20 amino acid types for the first 20 positions of protein sequences from the N-terminal region. Different results obtained by varying the BSVM parameters are shown in Table 3.6. All accuracy percentages shown are the mean of the 5-fold cross validation test results.

Clearly, the best score, 89.35%, was obtained with the use of a radial-base function of degree 30. However, the linear kernel function gave a very close result, 89.07%, suggesting that the. Another immediate result is that, obviously, using sigmoid kernel functions is not suitable for this particular classification problem. Albeit not presented, many other tests have been carried out with different capacity (C) parameters. In fact, altering the C-parameter did not improve the prediction rate any further, since in this problem it turned out that using smaller C (C=1) proved more effective for BSVM. Into the bargain, bigger C - parameters performed worse. For example, the mean of the 5-fold cross validation tests' accuracy rates for C=2000 was 85.5%, while that for C=10 was 86.2%. Note that, the C parameter in BSVM is not precisely identical to the C parameter in SVMlight, as it will be seen in the following attempts.

Kernel function	G	C	s	D	Prediction accuracy
Linear	1/3	1	0	N/A	88.51%
Linear	1/10	1	0	N/A	86.88%
Linear	1/100	1	0	N/A	88.92%
Linear	1/1000	1	0	N/A	87.26%
Linear	1/50	1	0	N/A	89.07%
Linear	3/100	1	0	N/A	88.89%
Linear	1/400	1	0	N/A	88.93%
Polynomial	1/50	1	0	2	88.93%
Polynomial	1/3	1	0	3	85.63%
Polynomial	1/3	1	0	9	83.92%
Radial-base	1/3	1	0	3	88.51%
Radial-base	1/50	1	0	3	89.08%
Radial-base	1/50	1	1	3	89.00%
Radial-base	1/50	1	1	9	89.13%
Radial-base	1/50	1	1	30	89.35%
Radial-base	1/3	1	2	3	89.26%
Radial-base	1/50	1	2	3	88.93%
Sigmoid	1/3	1	0	3	64.35%

Table 3.6 Results for Experiment 13

Experiment 14

This test suggested that using 10-dimensional (Table 3.5) input vectors formed by the 20 N-terminal residues, gives almost equally good results as Experiment 13. Different BSVM parameters led to a set of prediction rates ranging from 88.0% to 88.9%.

Experiment 15

With the frequency terms shrunk to 5 dimensions computed according to Table 3.4, we could still predict the localization sites correctly by up to 87.0%.

Experiment 16

We further limited the number of amino acids taken from the N-terminus down to 10. In fact, the composition information of the 20 types of amino acids in the short 10 N-terminal residue part went the prediction scores downhill to below 87.0%. No matter which amino acid grouping tried, the result has always got well below the above figure.

3.3 LOCATION PREDICTION USING AMINO ACID COMPOSITION AND STRUCTURAL INFORMATION FOR NONPLANTS

Experiment 17

According to some studies [2, 4, 17], it is believed that eukaryotic mitochondrial proteins form amphipathic alpha helices. The amphipathic helix motif is characterized by a repeating pattern of polar (P) and non-polar (N) side chains (Figure 2.3) that can be formalized as PxNPPNx. These clusters vary in length from 6 to 15 residues [20]. This can be used as a distinguishing feature to classify proteins, with an appropriate amphipathic helix predictor. Notwithstanding, no studies have been conducted to verify and use this observation directly in a location prediction method.

Profile score matrices can be used in order to gather some structural information. Particularly, as they can capture the periodic occurrences of certain amino acids with certain characteristics, they are very good alpha region predictors. The color codes for amphipatic alpha helices are given in Figure 3.2. The color represents frequency of occurrence. The key for the color scale is given in Table 3.7. These color codes have been converted to numerical values and used for assigning an amphipaticity score for each protein sequence in the dataset.

The best prediction accuracy score came from Experiment 13 in which the 20 N-terminal amino acid residues have been used. Upon this fact, a new dimension, the amphipaticity score, has been added to the input vectors to form the 21-dimensional dataset.

There is no available information on the distribution of these amphipatic alpha helices over the entire sequence. So, looking for the 6, or at most 15 residue long amphipatic helices through the 20 N-terminal window would not be effective. A helix formation starting at position 15, for example, would be missed. Therefore, in several experiments, the amphipatic helices have been searched in windows of different sizes. The maximum amphipaticity score obtained through the analyzed portion of the sequence has been added as the new dimension to the input data.

First of all, by scanning the full length protein sequences for an amphipatic helix in windows of length 17, a set of scores has been computed. With the gamma parameter set to 0.02, and the s parameter to 0, a prediction result of 88.8% has been obtained. Setting s to 1 improved the result up to 89.0%. Searching for the highest amphipaticity score in several different portions of the sequences yielded similar results ranging from 88.6 to 89.1%. However, by our trial-and-error approach, when we considered the 70 amino acid residue length portion following the first 20 N-terminal amino acids, the prediction accuracy increased only a several digits in a thousand.

Experiment 18

To boost the prediction rate as far as possible, this time we changed the 5-cross validation rule applied in all experiments so far. The jackknife method was applied for the dataset in Experiment 13, which can be summarized as follows: Out of the 2738 available protein sequences, 2737 ones have been used in the training procedure and the location class of the singled out sequence has been predicted according to the “model file” generated by BSVM. By this “1 vs. others” approach, iteratively for each individual sequence a single prediction has been done, while the remaining data were used as the training set. 89.44% of the protein sequences have been correctly classified as either “mTP”, “SP”, or “other”. The s parameter used was 2. With s set to 1, the result of the jackknife method produced a slightly worse performance of 89.15%. Probing the first 60 amino acids followed by the 20 N-terminal residues performed an 88.5% accuracy.

Red	>3 times greater than background
Orange	2-3 times background
Yellow	1-2 times background
Green	background frequency
Cyan	1-2 times less than background
Blue	2-3 times less than background
Dark blue	<3 times less than background

Table 3.7 The color scale for Figure 3.2

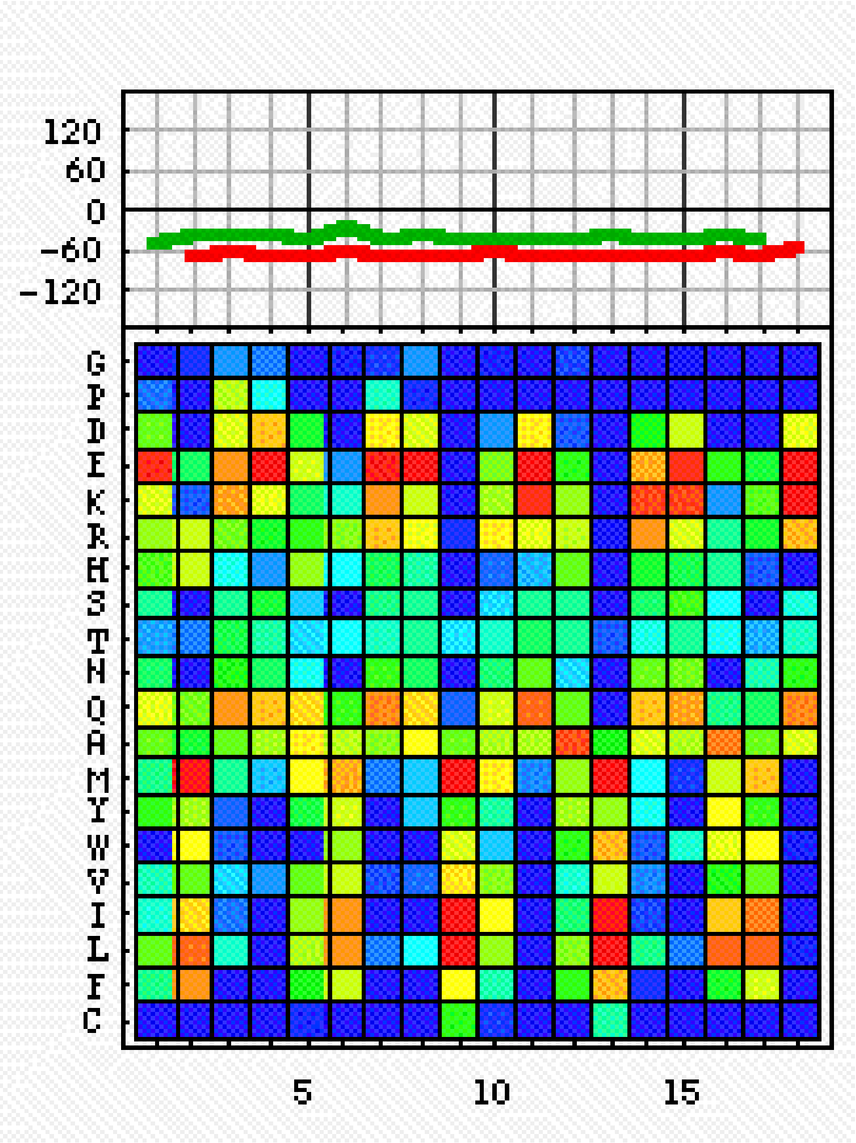


Figure 3.2 Color codes for amphipatic alpha helices. Color represents frequency of occurrence. The y-axis represents the amino acids, and the x-axis shows the position. Φ and Ψ are the backbone torsion angles [20].

Experiment 19

Experiments 17 and 18 use the occurrence frequencies of the 20 types of amino acids in the first 20 positions from the N-terminus, and additionally the best amphipaticity score found within the 70 amino acid residue long region following the first 20. In this experiment the contribution of the 30 N-terminal amino acids along with the amphipaticity scores obtained in the previously mentioned portions, were tested. Because, using only the first 30 residues, we obtained very close results, if not better, in Experiment 10 to those in Experiment 13. Yet, the prediction score could not exceed 88.1%.

Experiment 20

In this experiment we tested how secondary structure information may contribute to the determination of protein location. The secondary structure is a mapping of each amino acid residue into either as an alpha helix, a beta sheet, or a loop residue. We used the PSIPRED [21] prediction system, which is available on the world wide web at: http://bioinf.cs.ucl.ac.uk/psipred/, to calculate the secondary structure score for each amino acid in each sequence. We used the secondary structure composition rate per sequence length to build up a dataset of 23-dimensional input vectors. This resulted in still a similar result to Experiment 13's. We were able to reach the 89.3% limit we had previously obtained without any structural information, but could not exceed it. Moreover, using only the 3-dimensional input vectors representing the secondary structural information, a correct prediction rate of 64.5% has been attained. Another experiment which was conducted without normalizing the secondary structure abundance rates, that is by using the pure occurrence frequencies of the three secondary structures along with the amino acid composition, yielded a poor result of 74.3% success.

Experiment 21

In a 4-class classification, the normalized secondary structure composition information extracted out from the 20 N-terminal region performed a 75.0% prediction accuracy. This brought faintly a better performance than choosing not to make use of the structure information: 74.7% of the sequences have been assigned correctly into the 4 classes with the input configuration used in Experiment 13.

3.4 IMPROVING THE PREDICTION RESULTS BY AN COMMITTEE SYSTEM OF SVM

Experiment 22

The experiments carried out so far have served as to get an idea of what sort of input datasets would give us the best prediction results. Having tried many configurations and found that using the occurrence frequencies of the 20 N-terminal residues gives the best classification performance, it is now convenient to concentrate on improving this result, to make it comparable with Emanuelsson et al.'s 90%.

Localization prediction is a multi-class classification problem. We have 3 (mTP, SP, other) or 4 classes (mTP, SP, cyt, nuc) for the eukaryotes, and 4 classes (mTP, SP, chloroplast, nuc) for the plants.

Being one of the most widely used SVM software, SVMlight is a very good binary classifier. Its output is a real number between two user set limits, unlike BSVM, which outputs directly the class label by utilizing some multi classification techniques. Therefore, we need to construct n SMVs for a n-class classification. After being trained. each one will specialize in the prediction of a particular class. Each of these systems is called an expert.

The principle of divide and conquer states that we can solve a complicated computational task by dividing it into a number of computationally simpler tasks and then somehow combining the results. In machine learning, distributing the learning task into a number of individual predictor systems simplifies the problem. The combination of these expert systems constitutes a committee machine.

In this experiment, the outputs of each SVM, along with the correct class labels will be used to construct the n-dimensional "second-phase" training data for BSMV, which is capable of performing multi-class classification.

This is indeed a two-layered SVM system whose first layer is a committee machine formed by individual experts.

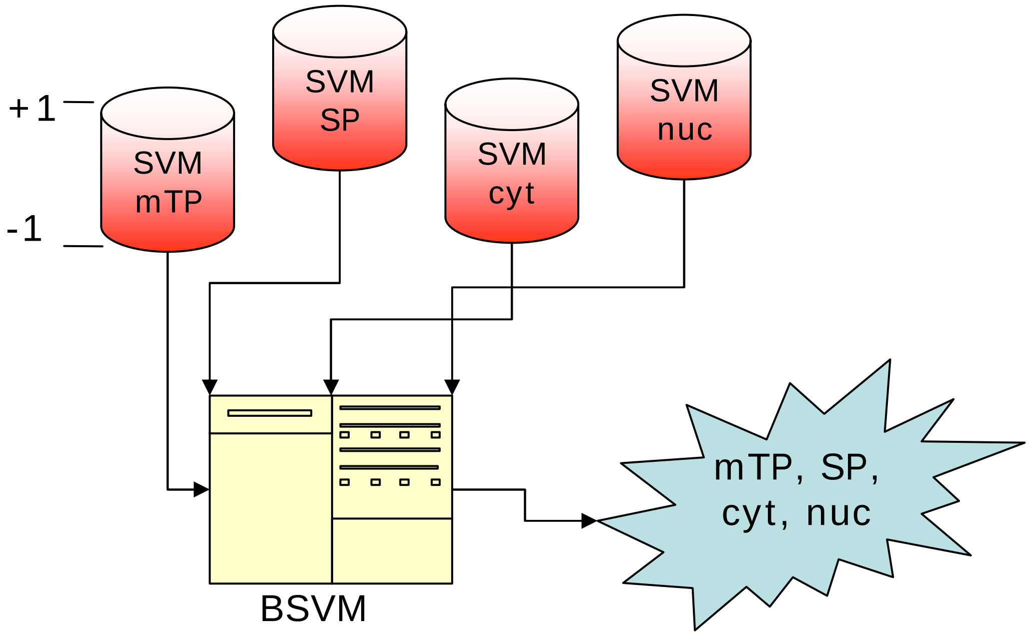


Figure 3.3 The architecture of the two-layer SVM expert system

The classification scheme used in the first layer is referred to as 1-v-r (one-versus-rest) SVM. For example, in the case of eukaryotic protein classification, the first SVM is trained with the mTP sequences being labeled as positive, and the rest with negative labels, and so forth. The probe sequence will be input to the SVM. If we were not using the second layer which has the BSVM, by following a “winner takes all” approach, the test protein sequence would be classified into the one class with the highest output score.

It is very probable that for some input, more than one expert system may output positive values. The second layer resolves problems occurring due to such conflicts resulting from the first layer. It learns cases in which more than one SVMLight produced a positive score, and for other similar test data, the sequence is classified into the correct class.

By using a radial-base kernel function the three SVMlight systems have been trained. The outputs are used to train the BSVM. The 5-fold cross validation testing approach yielded a 99.7% correct prediction accuracy for the 4-class classification. This validation data sets included all proteins. In each five attempts, while four fifth of the dataset was allocated for training, the remaining one fifth portion served as the test set.

If we were to use all four SVMlight expert systems by following a "winner takes all" approach, and without using a second layer, then the performance rate would be 90.5%. Obviously, using an integrative method involving both SVMlight and BSVM increased the accuracy rate remarkably. For BSVM, the problem becomes trivial when input data being initially 20-dimensions becomes 4-dimensional after it passes the first layer. In a sense, the first layer maps the input vector into a 4-dimensional one, converting the problem to a linearly separable one.

3.5 LOCATION PREDICTION BY AMINO ACID COMPOSITION FOR PLANTS

Having performed several experiments to find out the localization sites of eukaryotes, we have tested and applied the most successful ones for the prediction of plant proteins' locations. The approaches which yielded relatively good results for nonplants were also leading to good results for the plants. The best prediction performance has been achieved when the composition of the 20 types of 20 N-terminal amino acid residues were supplied as input to the two-level SVM expert system, just like the nonplants case. The overall prediction accuracy rate for the four location classes was 95.4%.

3.6 EVALUATION OF RESULTS

We evaluated the classification performances by precision / recall parameters. Precision and recall are defined as:

$$\mathbf{precision} = \frac{\mathbf{a}}{\mathbf{a+b}} \tag{3.1}$$

$$\mathbf{recall} = \frac{\mathbf{a}}{\mathbf{a+c}} \tag{3.2}$$

where; a, b, and c are the number of true positives, the number of false positives, and the number of false negatives, respectively.

The precision and recall values of the first layer of the expert system are given in Table 3.8 for nonplants, and in Table 3.9 for plants. Since the actual testing and evaluation have been performed by an expert system including both SVMlight and BSVM, the data provided in Table 3.8 and 3.9 serve only to give an idea of SVMlight's performance. Thus, the SMVlight is basically used in the mapping of input vectors of 20 dimensions to 4 dimensions.

In order to test the individual SVMlight systems, however, a part of the protein sequence list is used for training, and the remaining part for testing. The nonplant test set contains 100 sequences from each four classes. The "other" class, which is made up of nuclear and cytoplasmic sequences, contains 200 sequences. When we used a total of 300 sequences –100 for the "other" class as well, the recall decreased while the precision increased for the class "other".

Class	Accuracy	Incorrect/Total	Precision	Recall
mTP	91.00%	36 / 400	89.02%	73.00%
SP	94.50%	22 / 400	91.49%	86.00%
Other (100 nuc, 100 cyt)	91.25%	35 / 400	89.10%	94.00%
Other (50 nuc, 50 cyt)	91.00%	27 / 300	80.83%	96.04%

Table 3.8 Precision and recall parameters for the “first layer” SVMlight in nonplants

Class	Accuracy	Incorrect/Total	Precision	Recall
mTP	96.60%	32 / 940	95.41%	95.92%
SP	96.91%	29 / 940	97.46%	81.56%
Other (100 nuc, 100 cyt)	97.55%	23 / 940	98.43%	92.94%
Other (50 nuc, 50 cyt)	98.62%	13 / 940	98.06%	93.83%

Table 3.9 Precision and recall parameters for the “first layer” SVMlight in plants

4. AUTOMATIC FUNCTIONAL CATEGORIZATION OF PROTEINS BY SVMs

The MIPS comprehensive yeast genome database [22] presents information on the molecular structure and functional network of the entirely sequenced, well-studied model eukaryote, the budding yeast *Saccharomyces cerevisiae*. Having studied the protein subcellular location prediction techniques, in this section we focus our attention on the possibility of predicting protein functional categories mainly from primary structure information by similar approaches previously followed.

First of all, predicting function from sequence only is not a reliable method as there may exist proteins with similar sequence but dissimilar functions. Albeit it is believed that the 3-dimensional structure of a protein determines its function, this attempt serves as to see how far we may go in predicting the correct functional class by solely having amino acid composition as a distinguishing feature. Secondly, the functional categories are limited to those that involve the subcellular localizations, which is why we should expect this approach to work.

4.1 FUNCTION PREDICTION BY DATA-MINING OF MEDLINE ABSTRACTS

Experiment 23

Protein function prediction is among the most challenging problems in biology. In this experiment the contribution of MEDLINE paper abstracts in protein function prediction is evaluated. For each of a total of 254 protein sequences of which 154 were cytoplasmic and 100 cytoskeleton proteins, the MEDLINE database has been searched to find all the papers in which the corresponding protein name is cited. Then the cited protein-localization relations were analyzed by the data-mining of the abstracts.

In 2002, Stapley [1] et al. has used the IDF (Inverse Document Frequency) term which takes account of the number of MEDLINE documents relevant to a particular protein. The weight of localization term i for protein k is given by:

$$\log\left(1+\sum_j f_j(w_i)\right)-\log N(w_i)-\log(1+R_k) \tag{4.1}$$

where $f_j(w_i)$is the frequency of term i in document j. N is the number of documents containing term i, an R is the number of MEDLINE documents relevant to protein k.

The intuitive meaning of IDF is that terms which rarely occur over a collection of texts are valuable. The importance of each term is assumed to be inversely proportional to the number of texts that contain the term [23]. This way, unfortunately the importance of some term relations cited frequently can become as less significant as those that are rare, and vice versa. The very same relations mentioned by different papers should be rather strong and much more reliable as this information is strengthened by different sources.

In order improve the prediction score by somehow slotting in the *priori* knowledge that we have in biological paper abstracts into the SVM input, the following approach has been appropriated, instead of the IDF: For each protein sequence, if there is a cited location term throughout the all papers associated with that protein, no matter whether redundant relations exist, a “1” is written into the corresponding localization dimension in the input vector. Otherwise, that value is left null, meaning that no relation was found. Since in this particular experiment we have two protein classes, for each protein vector two new dimensions have been appended. However, it has been noticed that some papers had more than one localization term. For example, due to either the biological relations or the writing styles of authors, in a paper associated with a cytoplasmic protein, for instance, if the terms “cytoplasm” and the irrelevant term “endoplasmic reticulum” are both mentioned, and if this observation is present in some other papers as well, then it is probable that this type of indirect relations can improve the prediction performance significantly. The hypothesis proved to be working.

When we increased the number of localization terms to 4 to include the other commonly present localization sites, even while dealing with the prediction of only two types of protein classes, the correctly classified protein number increased by about 5% to be as high as 85.48%. To estimate the contribution of the MEDLINE parameters, the same protein sequence set has been trained and tested by using the 1-layer BSVM software without using the parameters obtained from data-mining. We were able to predict the correct functional class by 75.50% success.

4.2 FUNCTION PREDICTION BY AMINO ACID COMPOSITION

Experiment 24

To see how much we can do with our two-layer SVM system for the function prediction, we used 2321 protein sequences classified under 10 functional categories. The 10 functional categories used in prediction are listed in Table 4.1. The two-layer expert system yielded an excellent 92.86% prediction accuracy. This result is very important as it enables us to identify protein function very accurately from sequence only.

Once again, the SVM parameters used for this experiment are those that yielded the best prediction rates for the localization prediction. In addition, 20 N-terminal amino acid residues, without employing groping or neighborhood information have been used.

Class #	Functional category	Number of sequences
1	Organization of plasma membrane	147
2	Organization of cytoplasm	557
3	Organization of cytoskeleton	111
4	Organization of endoplasmic reticulum	156
5	Organization of Golgi	92
6	Nuclear organization	774
7	Mitochondrial organization	366
8	Peroxisomal organization	39
9	Vacuolar	58
10	Extracellular / secretory proteins	21

Table 4.1 Protein functional categories and number of sequences

5. DISCUSSIONS AND CONCLUSION

This study resulted in the best prediction scores obtained so far for protein subcellular localization both for eukaryotes and prokaryotes. Since our method is basically based on amino acid composition but not on the motif information, it is more reliable as the protein sequences are prone to some possible annotation errors.

In Experiment 1 and 2, the protein sequence length has been tested if it makes a contribution to the prediction accuracy. It was concluded that the increase in accuracy due to length information in Experiment 1 can be obtained by using different multiple classification schemes. In Experiments 3 and 4, the performances of other amino acid groupings have been compared. At the end, it was seen that using no amino acid classification has given the best results. After Experiments 5 and 6, it has become clear that incorporating neighborhood information into the input vectors does not give additional improvement. Quite the opposite, since the dimensionality was increased this way, the correct prediction percentages decreased dramatically.

From Experiment 7 through 22, we conclude that the composition of only 20 N-terminal amino acids sufficed to predict the location classes in very reasonable rates for both plant and nonplant proteins. Furthermore, using only the 20 N-terminal residues performed better than using the composition of the entire sequences. This means that the information encoded within the start region is sufficient to determine the localization. There are studies, however, showing that after removing part of the N-terminal region, it is still possible to get significant but slightly less accurate prediction results. This imparts that the entire primary structure is carrying information about the destination of a newly synthesized protein. The most informative part in a protein sequence turned out to be about the N-terminal region, however.

The subcellular localization categories for plant and nonplant proteins have been identified by correct prediction accuracies of 95.4%, and 99.7% respectively. As a future study, some effort could be put forward to study the few wrong predictions made by the SVM expert system individually.

The predicted localization categories are those involving mitochondria targeting peptides (mTP) pathway, signal peptides (SP) pathways, nucleus and cytosol for nonplants, and, mTP pathways, SP pathways, nucleus and chloroplast for plants. Since this study has given rise to very good prediction scores, the number of prediction classes can be extended to include more specific localization sites.

The best input representation for training the SVM has been to use the composition of all 20 types of amino acids without any grouping. Using the composition of 3 or 4-tuple words of consecutive amino acid residues in order to capture some signaling motifs and incorporate some neighborhood information into the input did not work better than the amino acid residue composition rates being used alone to form the 20 dimensional input set. Although mapping amino acids into several groups according to some common physical and chemical properties and then to make use of only the composition rates of these group letters in the training phase was a bright idea, the input data become scattered and thus more difficult to be learned by the SVM. In general, the learning performance of machine learning tools decreases while the number of input dimensions increases as it becomes more difficult to find the optimal hyperplane separating the classes.

This study was originally an attempt to find out a reliable automatic protein localization prediction tool. However, the scope has been extended to include the protein functional category prediction, upon coming across with the MIPS database in which proteins are classified according to their functions. Only the functional classes that are somehow related with the cellular locations, such as nuclear functions, proteins associated with organization of the cytoskeleton etc., have been considered and studied. With use of a two-layer expert SVM system, we predicted the protein functional category correctly for the 92.86% of the sequences. This result encourages us to apply the same method for functional categories that are not directly related with the organization of the localization sites.

Several protein features have been tested as to weather they contribute to the prediction rate. Among these protein properties, neither secondary structure nor the amphipaticity score information was helpful in the determination of subcellular location. On the other hand, they slightly improved the prediction rates for the protein functional categories.

Data mining of the MEDLINE abstracts greatly contributed to the functional categorization of proteins. It is also worth to note that some indirect term relations between localization terms and protein aliases were found. Relations of the form “Whenever an author mentions about a particular localization term along with a specific protein name in the same MEDLINE paper, he also mentions a particular localization term” are extracted. Considering the presence of some extra location terms that seem irrelevant at first sight improved the learning of SVM by almost 6 percent.

The SVM expert system consisting of two layers have performed the best throughout all experiments we made. It proved itself to be very effective. Unlike for SVMs, it is very common to use neural networks with more than one layer. This new two-layer SVM approach can be utilized in various applications including the other traditional application domains of machine learning such as pattern recognition.

BIBLIOGRAPHY

[1] Stapley, B. J. (2002) *Protein Functional Classification by Text Data-Miningonceptual* , unpublished article.

[2] Bannai, H. et. al. (2002) *Extensive feature detection of N-terminal protein sorting signals,* Bioinformatics, Vol. 18 no 2.

[3] Zhi-Ping Feng, (2002) *An overview on predicting the subcellular location of a protein. In Silico* Biology 2, 0027

[4] Nakai, K., (2000) *Protein sorting signals and prediction of subcellular localization*. Adv. Protein Chem., 54, 277-344

[5] Nakashima, H. and Nishikawa, K. (1994) J. Mol. Biol., 238. 54-61. Medline abstract

[6] Reinhardt, Hubbard T. (1998) *Using neural networks for prediction of the subcellular location of proteins*. Nucleic Acids Research, Oxford Univ. Press. 2230-2236.

[7] Hua S., Sun, Z. (2001) *Support Vector Machine approach for protein subcellular localization prediction. Bioinformatics*, Oxford Univ. Press., Vol 17, no 8, 721-728

[8] Murphy, R. F., Boland, M. V. and Velliste, M. (2000). *Towards A systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images.* Proc. Int. Conf. Intell. Syst. Mol. Biol. 8, 251-259.

[9] Shevade et al. (2001) *Improvements to SMO algorithm for SVM regression*

[10] Burges C. (1998) *A tutorial on Support Vector Machines for pattern recognition*. Data Mining and Knowledge Discovery 2, 121-167.

[11] Schölkopf B. and Smola A. (2002) *Learning with Kernels*, MIT press.

[12] John Platt (2001) *Support Vector Machines*, http://research.microsoft.com/~jplatt/svm.html, CCSP Group at Microsoft Research.

[13] von Heijne, G, (1983) *Patterns of amino acids near signal sequence cleavage sites*. Eur. J. Biochem. 133, 17-21.

[14] von Heijne, G, (1985) *Signal sequences: the limits of variation*. J. Mol. Biol. 184, 99-105.

[15] von Heijne, G, (1990) *The signal peptide*. J. Mol. Biol. 115, 195-201

[16] von Heijne, G, et al. (1989*) Domain structure of mitochondrial targeting peptides*. Eur. J. Biochem., 180, 535-545.

[17] Emanuelsson, O. et al. (2000) *Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.* H. Mol. Biol. 300, 1005-1016.

[18] Nakai, K. and Kanehisa M. (1992) *A knowledge base for predicting protein subcellular localization sites in eukaryotic cells.* Genomics. 14, 897-911.

[19] Thomas P., Dill K. (1996) *An iterative method for extracting energy-like quantities from protein structures*. Proc. Natl. Acad. Sci., Biophysics, 93, 11628-11633.

[20] "I-sites Library": http://isites.bio.rpi.edu/bystrc/Isites/amph_alpha.html

[21] Jones, D.T. (1999) *Protein secondary structure prediction based on*

position-specific scoring matrices. J. Mol. Biol. 292:195-202.

[22] "MIPS database", http://mips.gsf.de/proj/yeast/CYGD/db/index.html

[23] Tokunaga T., Iwayama M. (1994), *Text categorization based on weighted inverse document frequency.* Technical report, ISSN 0918-2802

INDEX