Fast Subcellular Localization by Extracting Informative Regions of Protein Sequences for Profile Alignment

Wang Wei

supervised by

Dr. M.W. MAK

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Master of Philosophy

Department of Electronic and Information Engineering

The Hong Kong Polytechnic University

September 2010

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

WANG Wei _____(Name of student)

Abstract

Fast Subcellular Localization by Extracting Informative Regions of Protein Sequences for Profile Alignment

The determination of protein subcellular localization is vital for the understanding of the functions of proteins and for the design of drugs. However, the experimental methods of subcellular localization are expensive and time-consuming. On the other hand, computational methods provide the potential to annotate large protein datasets in a cost effective and time efficient manner. With the ever increasing amount of sequenced proteins, the gap between the newly found protein sequences and the knowledge of their subcellular localization has widened rapidly. Thus, it is imperative to speedup the subcellular localization algorithms.

In this thesis, a cascaded fusion of cleavage site prediction and subcellular localization prediction is developed to alleviate the computational burden of homolog-based prediction methods. Specifically, the informative region (signal peptides or transit peptides) of a protein sequence is first determined by a cleavage site predictor. Then, only the informative segment is applied to a homology-based predictor for the determination of subcellular locations. A cleavage site predictor based on conditional random fields (CRFs) is developed. It was found that CRFs outperform neural networks and hidden Markov models in the prediction of cleavage site positions. To minimize the training and classification time of the subcellular localization predictors, a kernel Fisher discriminator is proposed. Specifically, the profile of the informative segment of a protein sequence is first generated by PSI-BLAST. The profile is then vectorized by computing the profile-alignment scores between the profile and all of the training profiles. The resulting vector is projected onto a low-dimensional space by using a new form of kernel discriminant analysis called kernel perturbation discriminant analysis.

The vector in the low-dimensional space is then classified by a support-vector-machine classifier. It was found that the reduction in dimension leads to further computation saving when compared with the direct classification of profile-alignment vectors.

The proposed method was evaluated on a newly created redundancy-removed data set using five-fold cross validations. Results show that the method can attain accurate localization while reducing the computational time substantially when compared to some start-of-the-art methods. In particular, it was found that truncating the sequences at their cleavage sites can reduce the profile creation time (by PSI-BLAST) as compared to truncating the profiles. A sensitivity analysis suggests that subcellular localization accuracy is inversely proportional to the discrepancy of the truncation positions with respect to the ground-truth cleavage sites. It was also found that the subcellular localization accuracy of chloroplast transit peptides (cTP) is highly dependent on the correct prediction of their cleavage site, suggesting further investigation is necessary to improve the cleavage site prediction of cTP.

ACKNOWLEDGMENTS

I would like to express sincere gratitude to various bodies from The Hong Kong polytechnic University, where I have the opportunity to study with. My major debt is to my Supervisor Dr. M. W. Mak, whose expertise, understanding, and patience, added considerably to my graduate experience. I appreciate his vast knowledge and skill in many areas, and his assistance in writing papers and this dissertation. I have learned a lot of things from him. Without his help, this study could not be completed.

Besides, I would like to express my appreciation to all the professors who have taught me during in my master study. The countless discussions with my teachers and their enthusiastic disabusing have proved to be fruitful and inspiring. I would also like to thank all members of staff of the department of Electronic and Information Engineering and the clerical staff in the General Office. They have created a creative environment for me to study in.

It is my pleasure to acknowledge the Research and Postgraduate Studies Office of The Hong Kong Polytechnic University for its generous support over the past two years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Knowing the subcellular locations of proteins is the first step towards understanding their functions. Subcellular localization is a problem of predicting which part in a cell a protein will be transported to, given the amino acid sequence (i.e., string data) of the protein. This problem is very important because localizations and functions depend on each other, i.e., the localization of proteins provides clues about their role in a cell when other information is not available. Thus, localization prediction can offer numerous insights that assist the prioritization of proteins for downstream analysis. Because of the rapidly increase in the number of sequenced genomes, it is highly desirable to develop effective prediction methods so that the newly found proteins can be effectively used in drug development.

1.1 Subcellular Localization and Organelles

It is well known that cells are the most basic structural and functional units of life. Organisms whose cells have a nucleus are called eukaryotes. The functions of a eukaryotic cell are mainly performed by the proteins in the cell. Protein molecules reside in many different compartments or organelles of a cell. The cells of eukaryotic organisms are divided into functionally distinct compartments, most of which are enclosed by internal membranes. Some major organelles of eukaryotic cells are: extracellular space, mitochondria, chloroplast, cytoplasm, nucleus, endoplasmic reticulum (ER), Golgi apparatus, peroxisome, vacuoles, cytoskeleton, nucleoplasm and

ribosomes. Cytoplasm, a jelly-like material, takes up most of the cell volume. The nucleus is a large, round body in the middle of a eukaryotic cell and contains molecules of DNA which encode the genetic information of the organism. The mitochondrion is the generator of chemical energy for the cell. It makes use of the energy from the oxidation of food molecules, such as sugars, to produce chemical fuel that powers cell's activities. Every subcellular compartment contains specific proteins, including enzymes.

1.2 Sorting Signals

Most eukaryotic proteins are synthesized in the cytoplasm and translocated into proper locations. A newly created protein will either transported to an organelle of a cell or secreted outside the cell through a secretary pathway [2]. The destination information can be found in a short segment of the amino acid sequence of the protein, which is in some way analogous to the IP address of a TCP/IP packet in data communication. These short segments are generally known as sorting-signals, targeting sequences, or signal peptides. After the protein is translocated across the cell membrane, the signal peptide will be *cleaved* off by an extracellular signal peptidase. The location at which the cleave off occurs is called the cleavage site.

The mechanism by which a cell transports a protein to its target location within or outside the cell is called the protein sorting process. Defects in the sorting process can cause serious diseases. Therefore, identifying signal peptides and their cleavage sites have both scientific and commercial values. For instance, to produce recombinant secreted proteins or receptors, it is important to know the exact cleavage sites of signal peptides. The information of signal peptides also allows pharmaceutical companies to manipulate the secretory pathway of a protein by attaching a specially designed tag to it. This ability has opened up opportunity for the design of better drugs.

Many proteins contain cleavable peptides at the N-terminus. The peptides contain

information (address) that allows the protein to be transported to either the secretory pathway (in which case they are called signal peptides) or to mitochondria and chloroplast (in which case they are called transit peptides).

The secretory signal peptide (SP) is an N-terminal peptide and acts like a "zip code" of the nascent secretory protein. It targets a protein for translocation across the endoplasmic reticulum (ER) membrane in eukaryotes [3]. It typically contains 15–30 amino acids and will be cleaved off during the translocation of the protein across the membrane. There is no simple consensus sequence for SPs, but they typically show three distinct compositional zones: an N-terminal region (n-region) which often contains positively charged residues, a hydrophobic region (h-region) of at least six residues and a C-terminal region (c-region) of polar uncharged residues with some conservation at the -3 and -1 positions relative to the cleavage site [4].

The targeting peptides of chloroplasts and mitochondria are also N-terminal peptides [5], and similar to SPs. These transit peptides (presequences) are cleaved off when entering into their final compartment. Their sequence features are less well characterized and the reported sequence motifs are even less conserved than those of the secretory SP.

The chloroplast transit peptide (cTP), which directs nuclear-encoded proteins into the chloroplast, is rich in hydroxylated residues, in particular Ser, and rarely has acidic residues [6]. Like SPs, cTPs have been characterized as having a three-domain structure. But the signal is relatively weak. The most conserved residue is an Ala directly after the N-terminal methionine. cTPs from different proteins vary considerably in length (20–100 residues).

The mitochondrial targeting peptide (mTP), which directs nuclear-encoded proteins into the mitochondria, tends to be rich in Arg, Ser and Ala, whereas negatively charged residues are rare [6]. The sequence conservation around the cleavage site is low, with an Arg in position -2 or -3 relative to the cleavage site as the most common motif [7]. The length of reported mTPs spans from 8 residues up to 122.

1.3 Motivation of Computational Prediction of Subcellular Localization

Proteins must be transported to the correct organelles of a cell and folded into correct 3-D structures to properly perform their functions. Therefore, knowing the subcellular localization is one step towards understanding its functions. Accurate prediction of subcellular locations can also assist the prioritization of proteins for downstream analysis and the identification of drug targets.

Experimental high-throughput approaches have been applied to determine protein localization in Arabidopsis thaliana [8]. However these techniques cannot be generally applied to all the eukaryotic cells and determination of subcellular localization via experimental means is often time-consuming and laborious. The number of newly found proteins is increasing rapidly and it is expected that the gap between the newly found protein sequences and the knowledge of their subcellular localization will widen continuously [9]. As a result, efficient and reliable prediction methods are needed in order to screen the huge amount of data derived from genome projects.

Protein subcellular localization prediction involves the computational prediction of where a protein resides in a cell. By identifying one or more of the signals that are known to influence protein targeting or by extracting features that correlate with a specific cellular compartment, a protein's probable cellular location can be deduced using protein sequence information.

Typically, as mentioned earlier, the information about a protein's organelle destination can be found within a short segment of sorting-signal sequences in N-terminus. When the final destination is the mitochondria, the chloroplast, or the secretory pathway, sorting usually relies on the presence of an N-terminal targeting sequence that is recognized by the translocation machinery [4]. Whereas the complete range of signals influencing the targeting of a protein is not completely clear yet, it seems that these signals are amenable to computational identification.

In recent years, impressive progress has been made in the computational prediction

Table 1.1: Length of secretory pathway signal peptide (SP), mitochondrial targeting peptide (mTP), and chloroplast transit peptide (cTP).

Peptide	Length (No. of Amino Acids)
SP	15–30
mTP	8–122
cTP	20–100

of subcellular localization. A number of approaches have also been proposed in the literature. These methods can be generally divided into the four categories, including predictions based on sorting signals [4, 10–18], global sequence properties [19–28], homology [29–35] and other information in addition to sequences [9, 36–41]. Methods based on sorting signals are very fast, but they typically suffer from low prediction accuracy. Homology-based methods are more accurate, but they are very slow. Therefore, *fast* and *reliable* predictions of subcellular localization still remain a challenge.

1.4 Our Proposal for Addressing the Limitations

The computation burden of homology-based methods is mainly due to the alignment of the whole sequences. Because localization information is not evenly spread over the whole sequence (otherwise the signal-based method will perform poorly), potential computation saving can be achieved by aligning the portion of the sequences that contains most of the localization information. For this, the signal-based methods can provide a good solution because these methods scan the whole sequence to look for the signal peptide (i.e., informative region).

In fact, the length of chloroplast transit peptide (cTP), mitochondrial targeting peptide (mTP), and secretory pathway signal peptide (SP) is under 122 amino acids only [18], as illustrated in Table 1.1.

We propose using cleavage site prediction to determine the most informative region

for alignment. Experiments on a data set extracted from a recent release of Swiss-Prot show that the computation time of homology-based subcellular localization can be substantially reduced by aligning profiles up to the cleavage site positions of signal peptides, mitochondrial targeting peptides, and chloroplast transit peptides [42]. While this method can reduce the profile alignment time by as much as 20 folds, it cannot reduce the computation time spent on creating the profiles. Therefore, we further propose an approach that can reduce both the profile creation time and profile alignment time. In the new approach, instead of cutting the profiles, we shorten the sequences by cutting them at the cleavage site locations. The shortened sequences are then presented to PSI-BLAST to compute the profiles [43].

Experimental results and analysis of profile-alignment score matrices suggest that both profile creation time and profile alignment time can be reduced without sacrificing subcellular localization accuracy. Once a pairwise profile-alignment score matrix has been obtained, a one-vs-rest SVM classifier can be trained. To further reduce the training and recognition time of the classifier, we propose a kernel perturbation discriminant analysis (KPDA) technique. It was found that KPDA enjoys a short training time as compared to the conventional SVM.

The remainder of this dissertation is organized as follows. In Chapter 2, the methods of subcellular localization are reviewed. The computational methods for cleavage cites prediction are described in Chapter 3. In Chapters 4 and 5, the cascaded fusion of cleavage site detection and homology-based approaches for subcellular localization are described. In Chapter 6, we describe the experiments dataset and the assessment criterion of performance. In Chapter 7, we compare the performance of localization predictors using different schemes. Finally, Chapter 8 presents our conclusions.

Chapter 2

LITERATURE REVIEW

Knowing the subcellular locations of proteins is important for elucidating the proteins' interactions with other molecules and for understanding their biological functions. Many papers have been published describing a number of approaches to solving this problem. These methods can be generally divided into four categories and they will be briefly explained in this chapter.

2.1 Prediction Based on Sorting Signals

Signal-based methods predict the localization via the recognition of N-terminal sorting signals in amino acid sequences. One of the first attempts at predicting subcellular localization used a weight matrix for detecting secretory SPs [10]. A weight matrix is a simple sequence profile built from an ungapped multiple sequence alignment, where the weights are calculated from the counts of each amino acid at each position in a window around the site of interest. PSORT, proposed by Nakai in 1991 [11, 12], is another early predictor that uses sorting signals for protein's subcellular localization. PSORT and its extensions – WoLF PSORT [16, 17] – use various sequence-derived features such as the presence of sequence motifs and amino acid compositions.

In the late 1990's, researchers started to investigate the application of neural networks [44] to recognize the sorting signals. In a neural network, input patterns are presented to one or more layers of artificial "neurons" that compute a weighted sum of their inputs and apply a nonlinear function to the sum. Because amino acid sequences are of variable length, the input to the neural network is extracted from a

short window sliding over the amino acid sequence. TargetP [13, 14] is a well-known predictor that uses neural networks. It requires the N-terminal sequence as an input into two layers of artificial neural networks, utilizing the other two binary predictors, SignalP [45] and ChloroP [46].

The Hidden Markov Model (HMM) method [15] can be considered as an extension of the weight matrix approach. In this method, gaps are allowed to exist in the alignment such that motifs of varying length can be represented. Like neural network, HMM is data driven, that is, it adjusts the free parameters gradually by repeated presentation of a data set, and thereby learn to generalize from the data it has been trained on.

Compared to other categories, prediction based on sorting signals is closer to mimicking the actual information processing in cells.

2.2 Prediction by Global Sequence Properties

Another type of approaches relies on the fact that proteins of different subcellular compartments differ in global properties, such as their amino acid composition. It has been shown that the characteristic in amino-acid composition is due almost entirely to surface residues [47].

Nakashima and Nishikawa [19] pioneered the prediction of protein sorting by using a simple odds-ratio statistics to discriminate between soluble intracellular and extracellular proteins on the basis of amino-acid composition and residue-pair frequencies. One popular predictor based on amino acid composition is SubLoc [20]. In SubLoc, a query sequence is converted to 20-dim amino acid composition vector for classification by SVM. Recently, there are several new approaches that use global sequence properties as features for subcellular localization. For example, in [21, 22], an AdaBoost learner is proposed to predict the subcellular localization of proteins based on their amino acid composition. In SubCellProt [23], the scores of a K-nearest neigh-

bor classifier and a probabilistic neural-network classifier were combined to classify unknown sequences into one of 11 subcellular localizations. Xu et al. [24] proposed a semi-supervised learning technique (a kind of transductive learning) that makes use of unlabelled test data to boost the classification performance of SVMs. These studies suggest that features, such as amino acid composition, derived from primary sequence and physicochemical properties can be used to predict subcellular localization with a reasonably high accuracy.

The advantage of using global sequence properties is that the prediction of localizations for which the sorting signals are not sufficiently defined is possible. One limitation of composition-based methods is that information about the sequence order is not easy to represent. Some authors proposed using amino-acid pair compositions (dipeptide) [26–28] and pseudo amino-acid compositions [25] to enrich the representation power of the extracted vectors.

2.3 Prediction by Homology

The homology-based methods use the query sequence to search protein databases for homologs [29,30] and predict the subcellular location of the query sequence as the one to which the homologs belong. This kind of method can achieve very high accuracy when homologs of experimentally verified sequences can be found in the database search [48]. A number of homology-based predictors have been proposed. For example, Proteome Analyst [31] uses the presence or absence of the tokens from certain fields of the homologous sequences in the Swiss-Prot database as a means to compute features for classification. In Kim et al. [32], an unknown protein sequence is aligned with every training sequences (with known subcellular locations) to create a feature vector for classification. Mak et al. [33] proposed a predictor called PairProSVM that uses profile alignment to detect weak similarity between protein sequences. Given a query sequence, a profile is obtained from PSI-BLAST search [49]. The profile is

then aligned with every training profile to form a score vector for classification by SVMs. More recently, the fact that a protein sequence can have multiple subcellular locations has motivated the development of new algorithms that can deal with this multiple-localization situation. For example, in [34], a similar-peptide knowledge base is constructed to store the local similarity of highly dissimilar sequences (with sequence identity less than 25%). To deal with the low sequence similarity, the authors apply the transitivity concept [35] to peptide fragments instead of using sequence alignment. During prediction, a set of peptide fragments similar to the query protein is determined by PSI-BLAST. Then, the peptide fragments are searched against the knowledge base to determine the subcellular localization of the query sequence.

2.4 Prediction Using Other Information in addition to Sequences

Some predictors not only use amino acid sequences as input but also require extra information such as Gene Ontology (GO) entries [9,36,37], lexical context in database entries [38] or PubMed abstracts [39, 40] as input. More recently, SherLoc2 [41] integrates several sequence-based features, GO entries as well as phylogenetic profiles [50] and achieves fairly good performance. Although studies have shown that this type of method can outperform sequence-based methods, the performance has only been measured on data sets where all sequences have the required additional information. Thus, the applicability is limited.

2.5 Protein Cleavage Site Prediction

Although signal sequences that direct proteins to their target location differ in length and contents, common features that make the sequences to act like signals still exist, as exemplified in Fig. 2.1. For example, all signal sequences have a long central region (the h-region) that is highly hydrophobic. These properties allow the cleavage sites to be predicted computationally.

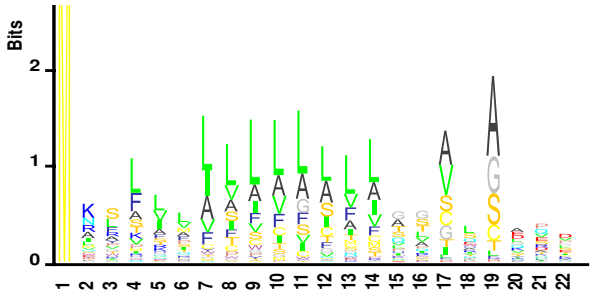


Figure 2.1: Logo diagram of 179 signal peptides with cleavage site between Positions 19 and 20. Positions preceding to the cleavage site are rich in hydrophobic (e.g. A and L) and polar (e.g. G and S) residues. The taller the letter, the more often the corresponding amino acid appears in the signal peptides.

The earliest approach to cleavage site prediction is to compute a weight matrix based on the position-specific amino acid frequencies of aligned signal peptides (aligned at the cleavage site) [10]. To predict the cleavage site of an unknown sequence, the matrix is scanned against the sequence to find the position of highest sum of weights. A recent implementation based on this approach is the PrediSi [51]. The weight matrix approach is very efficient, but the performance is inferior to more advanced approaches discussed below.

In SignalP 1.1 [13], a sliding window is applied to scan over an amino acid sequence. For each subsequence within the window, a numerically encoded vector is presented to a neural network for detecting whether the current window contains a cleavage site. An advantage of this approach is that a wide range physicochemical properties can be selected as network inputs. However, the prediction accuracy is dependent on the encoding methods [52]. In SignalP 2.0 and 3.0 [53, 54], an amino acid sequence is thought of as generated from a Markov process that emits amino acids according to some probability distributions when transiting probabilistically from state to state. To predict the cleavage site of an unknown sequence, the most likely transition path is found and the amino acid that aligns with the cleavage site node is considered as the cleavage site. One advantage of using this approach is that biological knowledge can

be easily incorporated into the models. Another advantage is that symbolic inputs can be naturally accommodated, and therefore numerical encoding as in the neural network approach is not required.

Recent research has demonstrated that conditional random fields are capable of predicting the cleavage site locations of signal peptides, and their performance is comparable to that of SignalP—a state-of-the-art predictor based on hidden Markov models and neural networks [55].

2.6 Limitations of Existing Approaches

Among all these methods, the signal-based and homology-based methods have attracted a great deal of attention, primarily because of their biological plausibility and robustness in predicting newly discovered sequences. Comparing these two approaches, the signal-based methods seem to be more direct, because they determine the localization from the sequence segments that contain the localization information. However, this type of method is typically limited to the prediction of a few subcellular locations only. For example, the popular TargetP [4, 18] can only detect three localizations: chloroplast, mitochondria, and secretory pathway. The homology-based methods, on the other hands, can in theory predict as many localizations as available in the training data. The downside, however, is that the whole sequence is used for the homology search or pairwise alignment, without considering the fact that some segments of the sequence are more important or contain more information than the others. Moreover, the computation requirement will be excessive for long sequences. The problem will become intractable for database annotation where hundreds of thousands of proteins are involved.

Chapter 3

FUSION OF CONDITIONAL RANDOM FIELDS AND SIGNALP FOR CLEAVAGE SITE PREDICTION

As will be discussed in Chapter 5, accurate prediction of protein cleavage sites is an important step in our proposed subcellular localization method. This chapter describes the Conditional Random Fields (CRFs) based cleavage site predictor and investigates the degree of complementarity between CRFs-based predictors and SignalP and proposes using the complementary properties to fuse the two predictors.

In previous investigation [55], Mak et al. have shown that conditional random fields (CRFs) [56] are capable of predicting cleavage site locations and that the prediction accuracy of CRFs is comparable to that of SignalP. We extend this work in two fronts: (1) we investigate the degree of complementarity between CRFs-based predictors and SignalP and propose a new fusion scheme based on the complementary information; and (2) we attempt to improve the prediction accuracy of CRFs by using spatially dispersed amino acids to construct the state features of the CRFs.

Evaluation based on the signal peptides extracted from the Swiss-Prot database shows that about 40% of the sequences that are incorrectly predicted by SignalP can be correctly predicted by CRF, and that about 30% of the sequences that are incorrectly predicted by CRFs can be correctly predicted by SignalP. This suggests that SignalP and CRFs posses significant complementary information, leading to better prediction performance when this information is exploited in the fusion process. This chapter also shows that the performance of CRFs can be further improved by constructing the state features from spatially dispersed amino acids in the training

Word	This	has	increased	the	risk	of	the	government
POS	DT	VBZ	VBN	DT	NN	IN	DT	NN
Chunk ID	B-NP	O	O	B-NP	I-NP	O	B-NP	I-NP

Table 3.1: An example sentence with a part-of-speech (POS) tag and a chunk identifier (in IOB2 format) for each word.

sequences.

3.1 Conditional Random Fields

CRFs were originally designed for sequence labeling tasks such as Part-of-Speech (POS) tagging, as exemplified in Table 3.1. Given a sequence of observations, a CRFs finds the most likely label for each of the observations. CRFs have a graphical structure consisting of edges and vertices in which an edge represents the dependency between two random variables (e.g., two amino acids in a protein) and a vertex represents a random variable whose distribution is to be inferred. Therefore, CRFs are undirected graphical models, as opposed to directed graphical models such as HMMs. Also, unlike HMMs, the distribution of each vertex in the graph is conditioned on the whole input sequence.

3.1.1 Formulation

Denote

$$\mathbf{x} = \{x_1, \ldots, x_T\} \quad \text{and} \quad \mathbf{y} = \{y_1, \ldots, y_T\}$$

as an observation sequence and the associated sequence of labels, respectively. In the case of cleavage site prediction,

$$\mathbf{x} \in \mathcal{A} \qquad \text{and} \qquad \mathbf{y} \in \mathcal{L} \equiv \{\mathrm{S}, \mathrm{C}, \mathrm{M}\},$$

where $\mathcal{A}$ is the set of 20 amino acid letters, and S, C, and M stand for the signal part, cleavage site, and mature part of a protein sequence, respectively. The cleavage site is located at the transition from C to M in $\mathbf{y}$.

Generative models such as HMMs model the joint distribution $p(\mathbf{x}, \mathbf{y})$ and computes the likelihood $p(\mathbf{x}|\mathbf{y})$ by assuming that the state y_t is only responsible for generating the observation x_t. In other words, when predicting the label at position t, HMMs cannot directly use information other than x_t. The independence assumption of x_t's restricts HMMs from capturing long-range dependence between $\mathbf{x}$ and $\mathbf{y}$. For example, standard HMMs cannot model explicitly the dependence between x_{t-d} and x_t where $d > 1$ or between x_{t-d} and y_t where $d \neq 0$. Most biological sequences, however, have such long-range dependence [57, 58]. Fig. 3.1 shows the correlation of amino acids at different positions relative to the cleavage site. Evidently, there is significant correlation between amino acids at non-adjacent positions. In particular, the correlation is fairly strong between amino acids at positions -6 and -14, which are 8 positions apart.

In fact, to predict the labels $\mathbf{y}$ given $\mathbf{x}$, the only distribution needs to be modeled is $p(\mathbf{y}|\mathbf{x})$. CRFs [56] are discriminative models that directly evaluate $p(\mathbf{y}|\mathbf{x})$:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \frac{F(\mathbf{x}, \mathbf{y})}{Z(\mathbf{x})} \\ &= \frac{\prod_{t=1}^{T} \exp\left\{ \sum_{i=1}^{|\mathcal{L}|} \sum_{j=1}^{|\mathcal{L}|} \alpha_{ij} f_{ij}(y_{t-1}, y_t) + \sum_{j=1}^{|\mathcal{L}|} \sum_{k=1}^{|\mathcal{P}|} \beta_{jk} g_{jk}(\mathbf{x}, y_t) \right\}}{Z(\mathbf{x})} \end{aligned} \tag{3.1}$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$ is a normalization factor, α_{ij} and β_{jk} are model parameters, $f_{ij}(\cdot)$ are transition-feature functions, $g_{jk}(\cdot)$ are state-feature functions, $\mathcal{P}$ is a set of amino acid patterns (see Section 4.3 for an example), and $|\mathcal{L}|$ is the cardinality of the set $\mathcal{L}$. Therefore, in CRFs, the relationship between adjacent states (y_{t-1}, y_t) is modelled as a Markov random field conditioned on the whole input sequence $\mathbf{x}$.

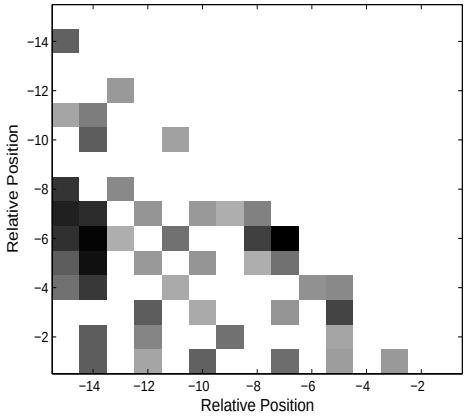


Figure 3.1: Correlation of hydrophobicity of 1695 protein sequences at different positions relative to their cleavage site. Entries in gray mean that the correlation between the hydrophobicity at the corresponding relative positions are statistically significant (p-value < 0.05). The grayness is proportional to the degree of correlation. Correlation at identical relative positions, which gives maximum correlation, is not shown for clarity of display.

3.1.2 Feature Functions

The definitions of feature functions depend on the application. In fact, one advantage of CRFs is the freedom of choosing suitable feature functions for modeling. This allows investigators to incorporate domain knowledge into the model.

To facilitate presentation in the sequel, let's denote $\mathcal{L}_i$ as the i-th label in $\mathcal{L}$, e.g., $\mathcal{L}_1 \equiv$ S. A similar notation is also applied to $\mathcal{P}$. The feature functions are typically boolean functions of the form:

$$f_{ij}(y_{t-1}, y_t) = \begin{cases} 1 & \text{if } y_{t-1} = \mathcal{L}_i \text{ and } y_t = \mathcal{L}_j \\ 0 & \text{Otherwise} \end{cases} \tag{3.2}$$

$$g_{jk}(\mathbf{x}, y_t) = \begin{cases} 1 & \text{if } y_t = \mathcal{L}_j \text{ and } b(\mathbf{x}, t) = \mathcal{P}_k \\ 0 & \text{Otherwise} \end{cases} \tag{3.3}$$

where $1 \leq i, j \leq |\mathcal{L}|$, $1 \leq k \leq |\mathcal{P}|$, and $b(\mathbf{x}, t)$ is a function that depends on the

amino acids in $\mathbf{x}$ around position t. One possibility is to use n-grams of the amino acid alphabet as $\mathcal{P}$ and the residues near position t as $b(\mathbf{x}, t)$. More formally, we have

$$\mathcal{P} = \text{n-gram}(\mathcal{A}) \quad \text{and} \quad b(\mathbf{x}, t) = x_{t-d_1} x_{t-d_2} \cdots x_{t-d_n}, \tag{3.4}$$

where $d_1 > d_2 > \cdots > d_n$. A large d_i enables the CRFs to capture the long-range dependence among the amino acids in the input sequence.

The operation of the feature functions can be explained via a simple example. Consider the amino acid sequence and its labels in Table 3.2. At $t = 5$, we have $y_4 = \text{S}$ and $y_5 = \text{C}$. Because $\mathcal{L}_1 = \text{S}$, $\mathcal{L}_2 = \text{C}$, and $\mathcal{L}_3 = \text{M}$, we have $f_{1,2}(y_4, y_5) = 1$. Assume that bi-gram is used for generating $\mathcal{P}$, i.e.,

$$\mathcal{P} = \{\texttt{AA}, \texttt{AC}, \ldots, \texttt{WA}, \ldots, \texttt{YY}\},$$

and that $d_1 = 1$ and $d_2 = 0$. Assume further that the amino acid pair `WA` occupies position k in $\mathcal{P}$, i.e., $\mathcal{P}_k = \texttt{WA}$. Then, we have $b(\mathbf{x}, 5) = \texttt{WA} = \mathcal{P}_k$ and therefore $g_{2,k}(\mathbf{x}, y_5) = 1$.

3.1.3 Advantages of CRFs

The CRFs enjoy several advantages over the HMMs.

1. *Avoid computing likelihood.* Because CRFs are discriminative models that compute the conditional probability $p(\mathbf{y}|\mathbf{x})$, it is not necessary to compute the likelihood of the input observation. It has been shown that discriminative models are usually superior to the generative models [59] because computing the probability of the observation is avoided.
2. *Model long-range dependence.* CRFs can model long-range dependence between the labels and observations without making the inference problem intractable, making it particularly useful for text processing [56] and bioinformatics [60].

3. *Guarantee global optimal.* The global normalization in Eq. 3.1 means that the global optimal solution can always be found.

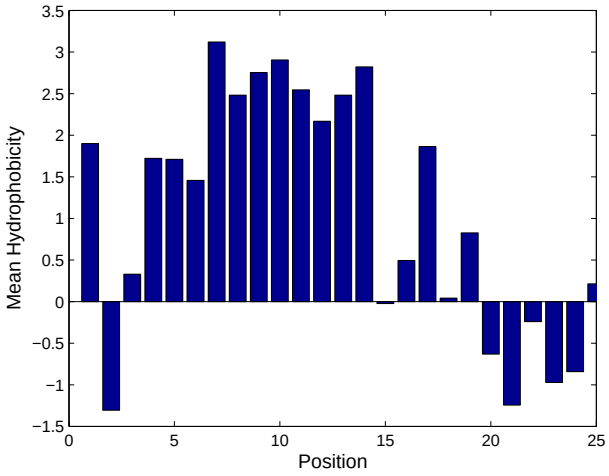
4. *Alleviate label-bias problem.* Many discriminative models, such as the maximum entropy Markov model, are prone to the label-bias problem (preferring states with fewer outgoing transitions) [56]. Because CRFs use global normalization, they possess the advantages of discriminative models but without suffering from the label bias problem.

3.2 CRFs for Cleavage Site Prediction

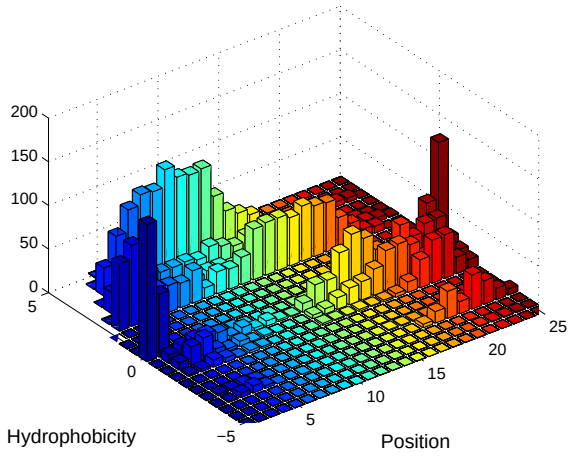
To use CRFs for cleavage site prediction, the prediction problem is formulated as a sequence labelling task in which amino acid sequences are treated as observations and each amino acid in the sequences is labelled as either "`S`ignal", "`C`leavage", or "`M`ature", e.g., `SSSSSSCMMMMMM`, as illustrated in Fig. 3.3. The cleavage site is located at the transition between `C` and `M`. Similar to the POS tagging task [61] in Table 3.1 where words are categorized as different types, amino acids of similar properties can be categorized as sub-groups.[1] We propose to divide the 20 amino acids according to their hydrophobicity and charge/polarity as shown in Table 3.3. These properties are used because the h-region of signal peptides is rich in hydrophobic residues and the c-region is dominated by small, non-polar residues [63], as illustrated in Fig. 2.1. Moreover, as illustrated in Fig. 3.2, the degree of hydrophobicity is also very different at different positions. It is believed that different sets of alphabets can complement each other in finding significant conserved regions along the amino acid residues. In case several alphabet sets indicate the same conserved region, that region is also likely to be of functionally important to the protein.

Table 3.2 shows an example amino acid sequence together with its hydrophobicity sequence and charge/polarity sequence. Note that either amino acid, hydrophobicity,

[1] This is called alphabet indexing [62] in the literature.



(a)



(b)

Figure 3.2: (a) The mean and (b) the histograms of hydrophobicity of 179 signal peptides at different sequence positions. The cleavage site of these sequences is between Positions 19 and 20.

charge/polarity, or their combinations can be used as observations to train a CRF.

To facilitate researchers to use CRFs for cleavage site prediction, a web server called CSitePred [55, 64] was developed.[2] CSitePred allows users to submit amino

[2]http://158.132.148.85:8080/CSitePred/faces/Page1.jsp

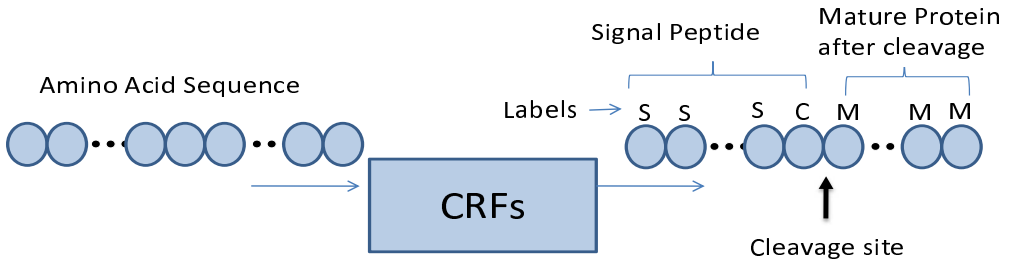


Figure 3.3: CRFs for Cleavage Cite Prediction. Given a sequence of observations, each amino acid in the sequences is labelled as either "Signal", "Cleavage", or "Mature", e.g., SSSSSSCMMMMMM. The cleavage site is located at the transition between C and M.

AA Sequence ($\mathbf{x}$)	T	–	Q	–	T	–	W	–	A	–	G	–	S	–	H	–	S
Hydrophobicity ($\mathbf{x}$)	H_2	–	H_1	–	H_2	–	H_3	–	H_3	–	H_2	–	H_2	–	H_2	–	H_2
Charge/Polarity ($\mathbf{x}$)	C_3	–	C_3	–	C_3	–	C_4	–	C_4	–	C_3	–	C_3	–	C_2	–	C_3
Label ($\mathbf{y}$)	S	–	S	–	S	–	S	–	C	–	M	–	M	–	M	–	M

Table 3.2: An example amino acid sequence with the corresponding hydrophobicity sequence and charge/polarity sequence. The 2nd and 3rd rows represent the hydrophobicity and charge/polarity groups shown in Table 3.3.

acid sequences by either coping-and-pasting FASTA format sequences into a window or uploading a FASTA file containing a large number of sequences. The web server returns the most likely cleavage site locations and their corresponding prediction scores of the submitted sequences to the user. Therefore, prediction on individual sequences or whole datasets are supported.

3.3 Fusion of CRFs and SignalP

We noticed from the outputs of SignalP and CRFs that for some sequences, when CRFs made a wrong decision, SignalP made a correct one. Similarly, there are also sequences whose cleavage sites are incorrectly predicted by CRFs but correctly pre-

Property	**Group**
Hydrophobicity	H1={D,E,N,Q,R,K} H2={C,S,T,P,G,H,Y} H3={A,M,I,L,V,F,W}
Charge/Polarity	C1={R,K,H} C2={D,E} C3={C,T,S,G,N,Q,Y} C4={A,P,M,L,I,V,F,W}

Table 3.3: Grouping of amino acids according to their hydrophobicity and charge/polarity [1].

dicted by SignalP. This suggests a potential performance improvement by fusing the decisions of CRFs and SignalP. To fuse the two decisions, some kinds of reliability scores need to be determined. For CRF, we used the probability of the best viterbi path, and for SignalP, we used the $C_{\max}$ scores. Hereafter, we refer to these scores as CRFs scores and SignalP scores, respectively.

Table 3.4 (a) shows the number of sequences with CRFs scores smaller than some pre-defined thresholds, below which the predicted sites are deemed untrustworthy. The table shows that less than 40% of these untrustworthy decisions are correct, suggesting that CRFs has difficulty in predicting the cleavage sites of these sequences. On the other hand, among these sequences, over 60% of them can be correctly predicted by SignalP. The situation is reversed in Table 3.4 (b). In particular, while SignalP can only predict the difficult sequences at a rate of 54%–69%, the CRFs achieves 97% accuracy on these sequences.

Based on these observations, we implemented the fusion as follows.

Step 1 Given a query sequence $\mathbf{x}$, present it to the CRFs and SignalP to obtain a CRFs score (denoted $\texttt{crf}(\mathbf{x})$) and a SignalP score (denoted $\texttt{snp}(\mathbf{x})$), respectively.

Step 2 Perform z-norm independently on these two scores to obtain the z-norm scores, namely $\texttt{crfn}(\mathbf{x})$ and $\texttt{snpn}(\mathbf{x})$.

Step 3 Determine the cleavage site position according to

$$\mathtt{p}(\mathbf{x}) = \begin{cases} \text{SignalP's decision} & \text{if } \mathtt{snpn}(\mathbf{x}) > \mathtt{crfn}(\mathbf{x}) - \epsilon \text{ or } \mathtt{crfn}(\mathbf{x}) < \eta \\ \text{CRF's decision} & \text{otherwise} \end{cases}$$

where ϵ and η are predefined constants that can be be determined from training data. In this work, $\epsilon = 0.8$ and $\eta = -2$. A positive ϵ means that the cleavage site position is based on CRFs only when the normalized CRFs score is significantly higher than the normalized SignalP scores. There are two main causes that contribute to the complementary performance of CRF and SignalP. Firstly, the prediction in SignalP is based on a neural network that uses a short sliding window over the query amino-acid sequence as input, whereas the prediction in CRF is based on the optimal Viterbi path along the whole query sequence. Therefore, the CRF-based prediction can make better use of the global information in the sequence than SignalP. The second cause is due to the difference in the training data applied to the two programs.

Score Threshold	No. of seqs below threshold (deemed untrustworthy)	No. of seqs correctly predicted by CRFs	No. of seqs correctly predicted by SignalP
0.60	94	32 (34.0%)	64 (68.1%)
0.65	125	44 (35.2%)	81 (64.8%)
0.70	156	61 (39.1%)	96 (61.5%)

(a)

Score Threshold	No. of seqs below threshold (deemed untrustworthy)	No. of seqs correctly predicted by CRFs	No. of seqs correctly predicted by SignalP
0.60	444	426 (96.0%)	243 (54.7%)
0.70	668	647 (97.0%)	412 (61.7%)
0.80	917	893 (97.4%)	628 (68.5%)

(b)

Table 3.4: The complement between CRFs and SignalP. The results were obtained by a two-stage process. In the first stage, either (a) CRFs or (b) SignalP was used for finding the sequences whose cleavage sites cannot be reliably predicted (i.e., below a predefined threshold in the first column of the tables). The number of sequences belonging to these categories are listed in the second column. Then, in the second stage, the cleavage sites of these sequences were predicted by CRFs and SignalP; the number of sequences with correctly predicted cleavage sites are listed in the 3rd and 4th columns.

Chapter 4

SUBCELLULAR LOCALIZATION PREDICTION BY KERNEL METHODS

4.1 Pairwise Profile Alignment

Kernel techniques based on profile alignment have been used successfully in detecting remote homologous proteins [65] and in predicting subcellular locations of eukaryotic protein [33]. Instead of extracting feature vectors directly from sequences, this method trains an SVM classifier by using the scores of local profile alignment.

A profile is a matrix in which elements in a column specify the frequency of that amino acid appears in the corresponding position. Given a sequence, a profile can be derived by aligning it with a set of similar sequences. The similarity score between a known and an unknown sequence can be computed by aligning the profile of the known sequence with that of the unknown sequence [65]. Because the comparison involves not only two sequences but also their closely related sequences, the score is more sensitive to detecting weak similarity between protein families.

Practically, the profile of a sequence can be obtained by using the sequence as a seed to search against a protein database (e.g., Swiss-Prot) for homologous sequences using the PSI-BLAST program [49]. The homolog information pertaining to the aligned sequences are represented by two matrices (profiles): position-specific scoring matrix (PSSM) and position-specific frequency matrix (PSFM). Each entry of a PSSM represents the log-likelihood of the residue substitutions at the corresponding positions in the query sequence. The PSFM contains the weighted observation frequencies of each position of the aligned sequences.

4.2 Local alignment-based kernels

Let us denote the operation of PSI-BLAST search given the query sequence $S^{(i)}$ of length n_i as,

$$\phi^{(i)} \equiv \phi(S^{(i)}) : S^{(i)} \rightarrow \left\{ \boldsymbol{P}^{(i)}, \boldsymbol{Q}^{(i)} \right\} \tag{4.1}$$

where $\boldsymbol{P}^{(i)}$ and $\boldsymbol{Q}^{(i)}$ are the PSSM and PSFM of $S^{(i)}$, respectively. Using the profile alignment algorithm specified in [33], we obtain the profile alignment scores $\rho(\phi(S^{(i)}), \phi(S^{(j)}))$. Then, the following normalized alignment scores are obtained:

$$\zeta(\phi^{(i)}, \phi^{(j)}) = \frac{\rho(\phi(S^{(i)}), \phi(S^{(j)}))}{\sqrt{\rho(\phi(S^{(i)}), \phi(S^{(i)}))\rho(\phi(S^{(j)}), \phi(S^{(j)}))}}. \tag{4.2}$$

Given N training sequences, the scores $\{\zeta(\phi^{(i)}, \phi^{(j)})\}_{i,j=1}^{N}$ constitute a symmetric matrix $\boldsymbol{Z}$ whose columns can be considered as N-dimensional vectors:

$$\boldsymbol{\zeta}^{(j)} = [\zeta(\phi^{(1)}, \phi^{(j)}) \quad \ldots \quad \zeta(\phi^{(N)}, \phi^{(j)})]^{\mathsf{T}} \quad j = 1, \ldots, N. \tag{4.3}$$

This means that there are N feature vectors with dimension equal to the training set size.

When $K(\cdot)$ is a linear kernel, we have

$$\begin{aligned} K(\phi(S), \phi(S^{(j)})) &= < \boldsymbol{\zeta}, \boldsymbol{\zeta}^{(j)} > \\ &= \sum_{n=1}^{N} \zeta(\phi(S^{(n)}), \phi(S)) \zeta(\phi(S^{(n)}), \phi(S^{(j)})). \end{aligned} \tag{4.4}$$

4.3 Multi-Classification using SVM

SVMs was first introduced by Vapnik [66] and are now broadly used in classification tasks. An SVM maps an input pattern onto a high-dimensional space and then defines an optimal separating hyperplane in that space. The hyperplane classifies the patterns into two categories and maximizes their distance from the hyperplane.

The multi-class problem can be solved by the one-vs-rest approach. Specifically, for a C-class problem (here C is the number of subcellular locations), C independent SVM classifiers are constructed. Given an unknown sequence S, the output of the c-th SVM is computed as:

$$f_c(S) = \sum_{i \in S_c} y_{c,i} \alpha_{c,i} K(\phi(S^{(i)}), \phi(S)) + b_c \tag{4.5}$$

where $\mathcal{S}_c$ is a set composed of the indexes of the support vectors, $y_{c,i} \in \{-1, +1\}$ is the label of the i-th training sequence, and $\alpha_{c,i}$ is the i-th Lagrange multiplier of the c-th SVM. The predicted class of S is given by

$$y(S) = \arg\max_c f_c(S), c = 1, ..., C. \tag{4.6}$$

The prediction process is illustrated in Figure 4.1. The profile of the query sequence is aligned with all of the protein sequences in the training set. The resulting profile-alignment scores are then used to form a feature vector for classification by the SVM classifier.

4.4 Kernel Discriminant Analysis for Efficient Classification

To further reduce the training and recognition time of the classifier, we propose to use kernel perturbation discriminants as an alternative option of SVMs. This section derives the formulation of kernel perturbation discriminant analysis (KPDA) and explains how it can be applied to multi-class problems such as subcellular localization. The key idea of KPDA lies on the equivalency between the optimal projection vectors in the Hilbert space, spectral space and empirical space. A more in-depth treatment in KPDA can be found in [67].

4.4.1 Input, Hilbert, Spectral, and Empirical Spaces

Denote the mapping from an input space $\mathcal{X}$ into a Hilbert space $\mathcal{H}$ as:

$$\overrightarrow{\phi} : \mathcal{X} \to \mathcal{H} \quad \text{such that} \quad \boldsymbol{x} \mapsto \overrightarrow{\phi}(\boldsymbol{x}).$$

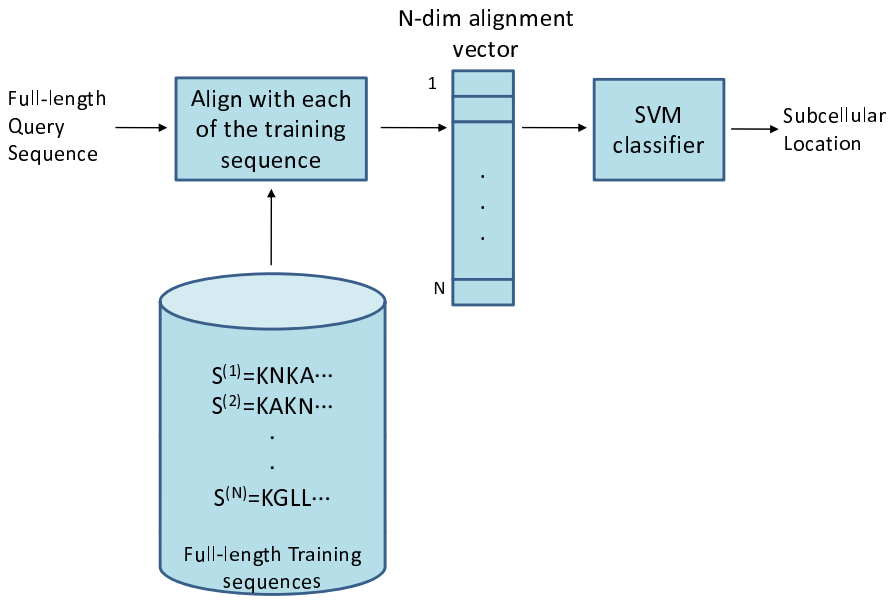


Figure 4.1: Flow chart of pairwise profile alignment SVM for classification

In bioinformatics, $\mathcal{X}$ is a vectorial space for microarray data and a sequence space for DNA or protein sequences. Given a training dataset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ in $\mathcal{X}$ and a kernel function $K(\boldsymbol{x}, \boldsymbol{y})$, an object can be represented by a vector of similarity with respect to all of the training objects [68]:

$$\overrightarrow{\boldsymbol{k}}(\boldsymbol{x}) \equiv [K(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, K(\boldsymbol{x}_N, \boldsymbol{x})]^T.$$

This N-dim space, denoted by $\mathcal{K}$, will be named empirical space. The associate kernel matrix is defined as

$$\boldsymbol{K} = \left[\overrightarrow{\boldsymbol{k}}(\boldsymbol{x}_1), \ldots, \overrightarrow{\boldsymbol{k}}(\boldsymbol{x}_N)\right].$$

The construction of the empirical space for vectorial and non-vectorial data are quite different. For the former, the elements of $\boldsymbol{K}$ are a simple function of the corresponding pair of vectors in $\mathcal{X}$. For the latter, the elements in $\boldsymbol{K}$ are similarities between the corresponding pairs of objects.

The kernel matrix $\boldsymbol{K}$ can be factorized with respect to the basis functions in $\mathcal{H}$: $\boldsymbol{K} = \boldsymbol{\Phi}^T\boldsymbol{\Phi}$, where $\boldsymbol{\Phi} = [\overrightarrow{\boldsymbol{\phi}}(\boldsymbol{x}_1), \ldots, \overrightarrow{\boldsymbol{\phi}}(\boldsymbol{x}_N)]$. Alternatively, it can be factorized via spectral decomposition: $\boldsymbol{K} = \boldsymbol{U}^T\boldsymbol{\Lambda}\boldsymbol{U} = \boldsymbol{U}^T\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U} = (\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U})^T(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}) = \boldsymbol{E}^T\boldsymbol{E}$, where $\boldsymbol{E} = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}$.

Denote the i-th row of $\boldsymbol{E}$ as $\boldsymbol{e}^{(i)} = \left[e^{(i)}(\boldsymbol{x}_1), \ldots, e^{(i)}(\boldsymbol{x}_N)\right]$. Because $\boldsymbol{E}\boldsymbol{E}^T = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\Lambda}^{\frac{1}{2}} = \boldsymbol{\Lambda}$, the rows of $\boldsymbol{E}$ exhibit a vital orthogonality property:

$$\boldsymbol{e}^{(i)}\boldsymbol{e}^{(j)T} = \begin{cases} 0 & \text{if } i \neq j \\ \lambda_i & \text{if } i = j, \end{cases}$$

where λ_i is the i-th element of the diagonal of $\boldsymbol{\Lambda}$.

For any positive-definite kernel function $K(\boldsymbol{x}, \boldsymbol{y})$ and training dataset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ in $\mathcal{X}$, there exists a (nonlinear) mapping from the original input space $\mathcal{X}$ to an N-dim spectral space $\mathcal{E}$:

$$\overrightarrow{\boldsymbol{e}} : \mathcal{X} \to \mathcal{E} \quad \text{such that} \quad \boldsymbol{x} \mapsto \overrightarrow{\boldsymbol{e}}(\boldsymbol{x}) \equiv \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{U}\,\overrightarrow{\boldsymbol{k}}(\boldsymbol{x}).$$

Many kernel-based machine learning problems involve finding optimal projection vectors in the spaces $\mathcal{H}$, $\mathcal{E}$, and $\mathcal{K}$, which will be respectively denoted as $\boldsymbol{w}$, $\boldsymbol{v}$, and $\boldsymbol{a}$. It can be shown [67] that the projection vectors are linearly related as follows:

$$\boldsymbol{w}^T\overrightarrow{\boldsymbol{\phi}}(\boldsymbol{x}) = \boldsymbol{v}^T\overrightarrow{\boldsymbol{e}}(\boldsymbol{x}) = \boldsymbol{a}^T\overrightarrow{\boldsymbol{k}}(\boldsymbol{x}), \tag{4.7}$$

where we have used the relationships $\boldsymbol{w} = \boldsymbol{\Phi}\boldsymbol{a}$ and $\boldsymbol{v} = \boldsymbol{E}\boldsymbol{a}$.

Assume that the dimension of $\mathcal{H}$ is M. When $M \geq N$, all of the N training vectors $\{\overrightarrow{\boldsymbol{\phi}}(\boldsymbol{x}_i); i = 1, \ldots, N\}$ will fall on an $(M-1)$-dim *data hyperplane*. Mathematically, the data-hyperplane is represented by its normal vector $\boldsymbol{p}$ such that $\boldsymbol{\Phi}^T\boldsymbol{p} = \boldsymbol{1}$. The optimal decision-hyperplane in $\mathcal{H}$ (represented by $\boldsymbol{w}$) must be orthogonal to the data-

hyperplane:

$$\boldsymbol{w}^T\boldsymbol{p} = 0 \quad \Rightarrow \quad \boldsymbol{a}^T\boldsymbol{\Phi}^T\boldsymbol{p} = 0 \quad \Rightarrow \quad \boldsymbol{a}^T\mathbf{1} = 0.$$

4.4.2 Kernel Fisher Discriminant Analysis (KFDA)

The objective of KFDA [69] is to determine an optimal discriminant function (linearly) expressed in the Hilbert space $\mathcal{H}$:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\overrightarrow{\boldsymbol{\phi}}(\boldsymbol{x}) + b,$$

where b is a bias. The discriminant function may be equivalently expressed in the N-dim spectral space $\mathcal{E}$:

$$f(\boldsymbol{x}) = \boldsymbol{v}^T\overrightarrow{\boldsymbol{e}}(\boldsymbol{x}) + b.$$

The finite-dimensional space $\mathcal{E}$ facilitates our analysis and design of optimal classifiers. In fact, the optimal projection vector $\boldsymbol{v}_{\text{opt}}$ in $\mathcal{E}$ can be obtained by applying conventional FDA to the column vectors $\{\overrightarrow{\boldsymbol{e}}(\boldsymbol{x}_i)\}$. To derive the objective function of KFDA, let us define

$$\boldsymbol{d} = \frac{2}{d_+ + d_-}\left(d_+\mathbf{1}_+ - d_-\mathbf{1}_-\right), \tag{4.8}$$

where $d_+ = \sqrt{\frac{N_-}{NN_+}}$ and $d_- = \sqrt{\frac{N_+}{NN_-}}$; $\mathbf{1}_+$ and $\mathbf{1}_-$ contain 1's in entries corresponding to Classes $\mathcal{C}_+$ and $\mathcal{C}_-$, respectively, and 0's otherwise; and N_+ and N_- are the number of training samples in classes $\mathcal{C}_+$ and $\mathcal{C}_-$, respectively. It can be shown that the objective function of KFDA is:

$$J_{\text{KFDA}}(\boldsymbol{v}) = \frac{\boldsymbol{v}^T\boldsymbol{S}_b^{\mathcal{E}}\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{S}_w^{\mathcal{E}}\boldsymbol{v}} = \frac{\boldsymbol{v}^T\boldsymbol{E}\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{E}^T\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{E}\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{N}\right)\boldsymbol{E}^T\boldsymbol{v}}, \tag{4.9}$$

where $\mathbf{1}$ is an N-dim vector with all elements equal to 1 and $\boldsymbol{S}_b^{\mathcal{E}} = \boldsymbol{E}\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{E}^T$ and $\boldsymbol{S}_w^{\mathcal{E}} = \boldsymbol{E}(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{N})\boldsymbol{E}^T$ are between-class and within-class covariance matrices in $\mathcal{E}$ space, respectively.

4.4.3 Perturbed Discriminant Analysis (PDA)

The FDA and KFDA are based on the assumption that the observed data are perfectly measured. It is however crucial to take into account the inevitable perturbation of training data. For the purpose of designing practical classifiers, we can adopt the following perturbed discriminant analysis (PDA).

It is assumed that the observed data is contaminated by additive white noise in the spectral space. Denote the center-adjusted matrix of $\boldsymbol{E}$ as $\bar{\boldsymbol{E}}$ and the uncorrelated noise as $\boldsymbol{N}$, then the perturbed scattered matrix is

$$(\bar{\boldsymbol{E}}+\boldsymbol{N})(\bar{\boldsymbol{E}}+\boldsymbol{N})^T \approx \bar{\boldsymbol{E}}\bar{\boldsymbol{E}}^T + \rho\boldsymbol{I} = \boldsymbol{E}\left(\boldsymbol{I}-\frac{\mathbf{1}\mathbf{1}^T}{N}\right)\boldsymbol{E}^T + \rho\boldsymbol{I},$$

where ρ is a parameter representing the noise level. Its value can sometimes be empirically estimated if the domain knowledge is well established a priori. Under the perturbation analysis, the kernel Fisher score in Eq. 4.9 is modified into the following perturbed variant:

$$J_{\mathrm{KPDA}}(\boldsymbol{v}) = \frac{\boldsymbol{v}^T\boldsymbol{E}\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{E}^T\boldsymbol{v}}{\boldsymbol{v}^T\left[\boldsymbol{E}\left(\boldsymbol{I}-\frac{\mathbf{1}\mathbf{1}^T}{N}\right)\boldsymbol{E}^T+\rho\boldsymbol{I}\right]\boldsymbol{v}}. \tag{4.10}$$

By taking the derivative of $J_{\mathrm{KPDA}}(\boldsymbol{v})$ with respect to $\boldsymbol{v}$, the optimal solution to Eq. 4.10 can be obtained as:

$$\boldsymbol{v}_{\mathrm{opt}} = \left[\boldsymbol{E}\left(\boldsymbol{I}-\frac{\mathbf{1}\mathbf{1}^T}{N}\right)\boldsymbol{E}^T+\rho\boldsymbol{I}\right]^{-1}\boldsymbol{E}\boldsymbol{d},$$

and using the Sherman-Morrison-Woodbury identity it can be shown that

$$\boldsymbol{v}_{\mathrm{opt}} = \left(\boldsymbol{E}\boldsymbol{E}^T+\rho\boldsymbol{I}\right)^{-1}\boldsymbol{E}(\boldsymbol{d}-\eta\mathbf{1}) = \left(\boldsymbol{\Lambda}+\rho\boldsymbol{I}\right)^{-1}\boldsymbol{E}(\boldsymbol{d}-\eta\mathbf{1}) \tag{4.11}$$

where η is a scaler whose value can be determined through the optimal solution in $\mathcal{K}$

space as follows.

Recall from Eq. 4.7 that dot-products in the three spaces are equivalent. Therefore, the discriminant function in $\mathcal{K}$ space can be written as:

$$f(\boldsymbol{x}) = \boldsymbol{a}^T \overrightarrow{\boldsymbol{k}}(\boldsymbol{x}) + b. \tag{4.12}$$

Given the optimal solution $\boldsymbol{v}_{\text{opt}}$ in the $\mathcal{E}$ space, the corresponding optimal solution in the $\mathcal{K}$ space is[1]

$$\begin{aligned} \boldsymbol{a}_{\text{opt}} &= \boldsymbol{E}^{-1}\boldsymbol{v}_{\text{opt}} \\ &= \boldsymbol{U}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{\Lambda}+\rho\boldsymbol{I})^{-1}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}(\boldsymbol{d}-\eta\mathbf{1}) \\ &= (\boldsymbol{K}+\rho\boldsymbol{I})^{-1}(\boldsymbol{d}-\eta\mathbf{1}), \end{aligned} \tag{4.13}$$

where we have used $\boldsymbol{K} = \boldsymbol{U}^T\boldsymbol{\Lambda}\boldsymbol{U}$ and $\boldsymbol{E} = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}$. Note that unlike Eq. 4.11, Eq. 4.13 does not require spectral decomposition, thus offering a fast close-form solution. Now using the orthogonal hyperplanes principle [67], we have

$$\begin{aligned} \boldsymbol{a}_{\text{opt}}^T\mathbf{1} = (\boldsymbol{d}^T - \eta\mathbf{1}^T)(\boldsymbol{K}+\rho\boldsymbol{I})^{-1}\mathbf{1} = 0 \\ \Rightarrow \quad \eta = \frac{\boldsymbol{d}^T(\boldsymbol{K}+\rho\boldsymbol{I})^{-1}\mathbf{1}}{\mathbf{1}^T(\boldsymbol{K}+\rho\boldsymbol{I})^{-1}\mathbf{1}}. \end{aligned} \tag{4.14}$$

The value of b can be obtained by using the relationship [67]:[2] $(\boldsymbol{y}-b\mathbf{1}) = (\boldsymbol{d}-\eta\mathbf{1})$, which gives

$$b = y_i - (d_i - \eta) \quad \text{for any } i = 1, \ldots, N. \tag{4.15}$$

---

[1] Eq. 4.7 suggests that $\boldsymbol{a}^T\boldsymbol{K} = \boldsymbol{v}^T\boldsymbol{E}$. Therefore, we have $\boldsymbol{a}^T = \boldsymbol{v}^T\boldsymbol{E}\boldsymbol{K}^{-1} = \boldsymbol{v}^T\boldsymbol{E}(\boldsymbol{E}^T\boldsymbol{E})^{-1} = \boldsymbol{v}^T\boldsymbol{E}^{-T}$, which suggests that $\boldsymbol{a} = \boldsymbol{E}^{-1}\boldsymbol{v}$.

[2] Note that our definition of $\boldsymbol{d}$ in Eq. 4.8 and that of [67] differ by a proportional constant.

4.4.4 Application of KPDA to Multi-Class Problems

A C-class problem can be formulated as C binary classification problems in which each problem is solved by a one-versus-rest binary classifier. Here, we propose two approaches to applying KPDA to solve multi-class problems.

One-vs-Rest KPDA Classifier

Given the training samples of C classes, we train C KPDA score functions as follows:

$$f_i(\boldsymbol{x}) = \boldsymbol{a}_i^T \overrightarrow{\boldsymbol{k}}(\boldsymbol{x}) + b_i, \qquad i = 1, \ldots, C,$$

where $\boldsymbol{a}_i$ and b_i are obtained by using Eq. 4.13 and Eq. 4.15, respectively. Then, given a test sample $\boldsymbol{x}$, the class label is obtained by

$$l = \arg\max_i f_i(\boldsymbol{x}).$$

Cascaded Fusion of KPDA and SVM

Because of the dependence in $d_i, i = 1, \ldots, C$, the rank of matrix $[\boldsymbol{d}_1, \ldots, \boldsymbol{d}_C]$ is $C-1$. Therefore, there are $C-1$ independent sets of KPDA parameters:

$$\begin{aligned}\hat{\boldsymbol{A}} &= [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{C-1}] \\ &= (\boldsymbol{K} + \rho \boldsymbol{I})^{-1}([\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{C-1}] - \boldsymbol{1}[\eta_1, \ldots, \eta_{C-1}]).\end{aligned}$$

During recognition, an unknown sample $\boldsymbol{x}$ is projected onto a $(C-1)$-dim KFDA space spanned by $[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{C-1}]$ using

$$\boldsymbol{g}(\boldsymbol{x}) = \hat{\boldsymbol{A}}^T \boldsymbol{k}(\boldsymbol{x}) + [b_1, \ldots, b_{C-1}]^T.$$

Then, $\boldsymbol{g}(\boldsymbol{x})$ is classified by one-vs-rest SVMs. In the sequel, we refer to this cascaded fusion as KPDAproj+SVM.

Chapter 5

SPEEDING UP PROFILE ALIGNMENT BY EXTRACTING INFORMATIVE REGION

The computation burden of homology-based methods is mainly due to the profile creation and alignment of the whole sequences. This chapter explains our proposed method to mitigate the computation burden.

5.1 Truncation of Profiles for Fast Profile Alignment

Generally, the length of signal peptide is less than 100 amino acids. Given the fact that the majority of proteins in the Swiss-Prot database have about a few hundred amino acids and that some proteins could have length longer than 5,000 amino acids, tremendous computational saving can be achieved by aligning the pre-sequence region (from the N-terminus to the cleavage site) for those proteins containing a signal or targeting sequence.

For profile alignment, this amounts to aligning the pre-profile region, i.e., the PSSM and PSFM in Section 4.1 are truncated at the column corresponding to the cleavage site before carrying out profile alignment.

The above observation suggests that the computation burden can be largely alleviated by a cascaded fusion of signal-based and homology-based methods. The fusion has three steps (Fig. 5.1):

1. *Cleavage site detection.* The cleavage site (if any) of a query sequence is determined by a signal-based method.

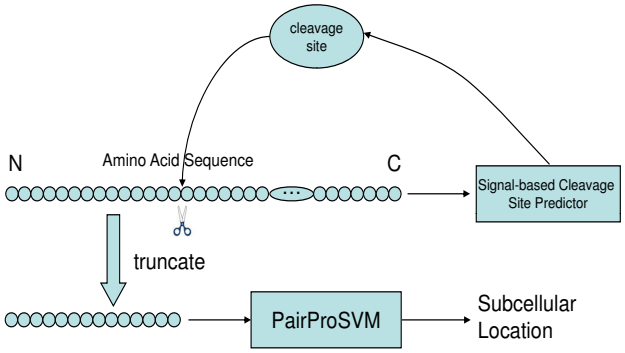


Figure 5.1: Cascaded fusion of signal-based and homology-based methods. The signal-based cleavage site predictor, such as TargetP [18] and CSitePred [70], is used as a pre-processor that reduces the sequence length for the computationally expensive homology-based method such as PairProSVM [33].

2. *Pre-profile selection.* The pre-profile of the query is obtained by selecting from the N-terminus up to the cleavage site.

3. *Pairwise alignment.* The pre-profile is aligned with each of the training pre-profiles to form an N-dim vector, which is fed to a one-vs-rest SVM classifier for prediction.

The cleavge site preditor can be based on TargetP and CRFs. TargetP [4,18] is one of the most popular signal-based subcellular localization predictors and cleavage site predictors. Given a query sequence, TargetP can determine its subcellular localization and will also invoke SignalP [54], ChloroP [46], or a program specialized for mTP to determine the cleavage site of the sequence. TargetP requires the N-terminal sequence of a protein as input. During prediction, a sliding window scans over a query sequence; for each segment within the window, a numerically encoded vector is presented to a neural network to compute the Y-score of the segment. The cleavage site is determined by finding the position at which the Y-score is maximum. The cleavage site prediction accuracy of SignalP on Eukaryotic proteins is around 70% [45] and that of ChloroP

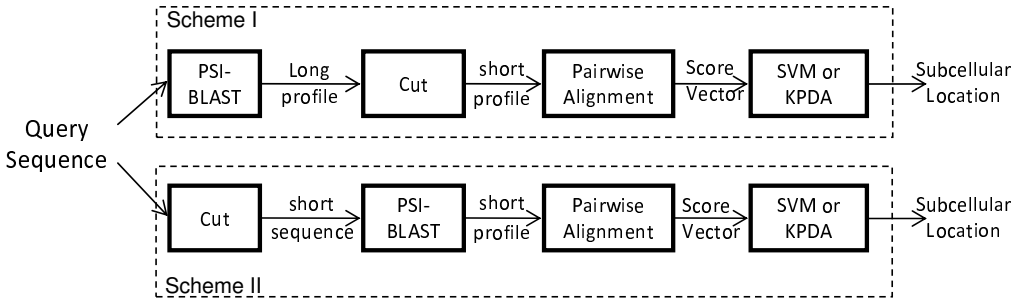


Figure 5.2: Two schemes for reducing the computation of the subcellular localization process. In Scheme I, a full-length query sequence is presented to PSI-BLAST for computing a full-length profile; then the profile is truncated at the predicted cleavage site. The truncated profile is then aligned with all of the truncated training profiles to produce a profile-alignment score vector for classification. In Scheme II, the query sequence is truncated at the predicted cleavage site before inputting to PSI-BLAST for computing the profile. The cleavage sites are predicted by CSitePred [55] or TargetP [4].

on cTP is 60% ($\pm$2 residues) [46], suggesting that there is room for improvement.

During the training phase, N training pre-profiles are obtained by truncating at the columns corresponding to the cleavage sites. Pairwise alignments are then performed to create an $N \times N$ symmetric score matrix whose column vectors are used to train a one-vs-rest SVM classifier.

5.2 Truncation of Protein Sequences for Fast Profile Acquirement

As will be demonstrated in the Section 7.6, by using the method described in the Section 5.1, the computation time of subcellular localization based on profile alignment SVMs can be substantially reduced. Although 20-fold reduction in total computation time (including alignment, training and recognition time) has been achieved, the method fails to reduce the profile creation time, which will become a substantial part of the total computation time when the database becomes large. We further propose a new approach that can reduce both the profile creation time and profile alignment

time [42]. In the new approach, instead of cutting the profiles, we shorten the sequences by cutting them at the cleavage site locations. The shortened sequences are then presented to PSI-BLAST to compute the profiles. Fig. 5.2 shows the difference between these two approaches.

Chapter 6

EXPERIMENTS

6.1 Materials and Procedures

6.1.1 Data Set Construction

Protein sequences with experimentally annotated subcellular locations were collected from Swiss-Prot Release 57.5 according to the following criteria [71].

1. Only the entries of Eukaryotic species were included, which were annotated with "Eukaryota" in the OC (Organism Classification) fields in Swiss-Prot.

2. A large amount of sequences in Swiss-Prot are annotated with ambiguous words, such as "probable", "by similarity" and "potential". These entries were excluded because of the lack of experimental evidence.

3. Sequences annotated with "fragment" were excluded.

4. Sequences that have 25% or higher sequence identity to any other sequences are excluded.

5. For signal peptides, mitochondria, and chloroplast, only sequences with experimentally annotated cleavage sites are included.

Sequence quality is of primary importance for the development of good prediction methods. To this end, all training sequences should have experimental evidence and should not be inferred by similarity or existing prediction methods. Otherwise, it can

Table 6.1: Breakdown of the eukaryotic dataset used in this work. The data were extracted from Swiss-Prot Release 57.5, with sequence identity less than 25%.

Subcellular Location	No. of Sequences
Extracellular	693
Mitochondria	167
Chloroplast	74
Cytoplasm/Nucleus	1,617
All	2,552

lead to circular prediction in which methods reproduce each other's predictions. For this reason we built a non-redundant dataset comprising proteins sharing less than 25% sequence identity. Table 6.1 shows the breakdown of the number of sequences in each class.

Sprenger et al. [72] compare the performance of five subcellular localization methods that are capable of predicting at least nine locations. It was concluded that none of the five methods had a sufficient level of sensitivity that would allow reliable prediction of hypothetical proteins. Therefore, we consider four subcellular compartments shown in Table 6.1: Extracellular, mitochondria, chloroplast and others (including cytoplasm and nucleus). We decided not to predict other locations because the number of annotated proteins with less than 25% sequence identity is very small, which do not allow us to train a predictor with good generalization capability.

6.2 Procedures

6.2.1 Cleavage Site Prediction

To assess the performance of different cleavage site predictors, TargetP and CSitePred (a CRFs-based predictor [55, 70]) were compared for the prediction accuracy of the cleavage site for SP, mTP and cTP. During prediction, the subcellular locations of the test sequences were assumed to be unknown. For TargetP, the subcellular location

of a test sequence was first determined by presenting the sequence to TargetP using either the 'Plant' or 'Non-plant' option of the predictor. Based on the subcellular location, TargetP will then determine the cleavage site of the sequence by invoking SignalP, ChloroP (for plant), or a program specialized in predicting the cleavage sites of mTP. For CSitePred, given a query sequence, the CRF (corresponding to either SP, mTP, or cTP) with the maximum Viterbi-search score was first identified. Then, the cleavage site was obtained from the optimal Viterbi search path of this maximum-scoring CRF.

The property set $\mathcal{P}$ for the state-feature function $f_{jk}(\cdot)$ contains n-grams of amino acids, where $n = 1, \ldots, 5$, and bi-gram of hydrophobicity groups and polarity/charge groups.

To investigate the effect of the maximum allowable offset for indexing amino acids in a sequence on prediction accuracy, various values of $\max\{d_n\}$ in Eq. 3.4 were tried.

6.2.2 Effect of Incorrect Cleavage Site Prediction on Subcellular Localization

To evaluate the effect of incorrect cleavage site prediction on the accuracy of subcellular localization, sensitivity analysis was performed by using the N-terminal signal peptides cleaved at the ground-truth cleavage sites or plus/minus several positions of the ground-truths. The sequence cut-off positions are 16, 8, 2 amino acids upstream or 2, 16, 32, 64 amino acids downstream from the ground-truth cleavage site. For comparison, another experiment was done in which the cleaved-off position was set to 170, i.e., none of the profiles (or sequences) have length exceeding 170.

6.2.3 Subcellular Localization

The performance of subcellular localization prediction by the proposed cascaded fusion method was evaluated and compared with two state-of-the-art subcellular localization predictors: SubLoc [20], TargetP [4] and PairProSVM [33]. The performance of SubLoc and TargetP were obtained by presenting the sequences of the dataset

to their webserver. We used TargetP and CSitePred for cleavage site detection and used PairProSVM for classification of pre-profiles (or pre-sequences). Sequences with cleavage site include extracellular, mitochondria, and chloroplast. Sequences without a cleavage site include cytoplasm and nucleus. We measured the computation time to (1) create a profile, (2) perform profile alignment, (3) train an SVM, and (4) recognize a profile-alignment vector based on an Intel Core 2 Duo 3.16 GHz CPU. For PSI-BLAST, parameters `h` and `j` were set to 0.001 and 3, respectively. The Spider Toolbox[1] was used to implement the SVM classifiers, and CRF++[2] was used for implementing CSitePred.

6.2.4 Assessment of Prediction Performance

We used 5-fold cross validation to evaluate the performance. In this technique, the original dataset was divided randomly into 5 sets consisting of nearly equal number of sequences. In each fold, one subset was singled out as a testing set, and the remaining ones were merged as the training set; this process was repeated five times.

The overall prediction accuracy, the accuracy for each subcellular location, and the Matthew's correlation coefficient (MCC) [73] were used to quantify the prediction performance. MCC allows us to overcome the shortcoming of accuracy (Acc) on unbalanced data [73].

---

[1] http://www.kyb.mpg.de/bs/people/spider/

[2] http://crfpp.sourceforge.net/

Chapter 7

RESULTS AND DISCUSSIONS

7.1 Effect of Indexing Offsets

Table 7.1 shows the performance of CRFs at different value of $\max\{d_n\}$ in Eq. 3.4. Evidently, varying the maximum allowable offset affects the prediction performance. The superiority of large offset seems to suggest that signal sequences exhibit long-range dependency. However, this conjecture needs to be confirmed biologically.

7.2 Fusion of CRFs and SignalP

Table 7.2 suggests that fusing the decisions of SignalP and CRFs can increase the prediction accuracy. In particular, the fusion strategy adopted in Section 3.3 achieves an even higher performance than the one we used in [55].

7.3 Histograms of Sequence Length

As shown in Fig. 7.1, the majority of proteins in the datasets have a few hundred amino acids. The average sequence length is 469 amino acids and some proteins have length up to 5,560 amino acids. Fig. 7.2 shows the histograms of the length of signal peptides, mitchohondrial transit peptides, and chloroplast transit peptides. It is obvious that the lengths of the three types of peptides are rather short (ranging from 6 to 100), with cTP longer than mTP and SP on average. The length distribution of SP has a relatively narrow peak, whereas that of the cTP and mTP spread over a wider range.

Maximum Allowable Offset	Prediction Accuracy
5	80.54%
6	81.41%
7	82.40%
8	83.17%
9	**83.32**%
10	83.12%
11	82.71%
12	82.40%

Table 7.1: Accuracy of CRFs-based predictors at different maximum AA position offsets, i.e., $\max\{d_n\}$ in Eq. 3.4.

Cleavage Site Predictor	**Accuracy**
SignalP [54]	81.88%
PrediSi [51]	77.06%
CRF5 [55]	79.71%
CRF5 + SignalP [55]	83.12%
CRF9	83.32%
CRF9 + SignalP	**85.03**%

Table 7.2: Accuracy of different cleavage site predictors and the fusion of CRFs and SignalP. CRF5 and CRF9 stand for CRFs with window size of 5 and 9 amino acids, respectively.

7.4 Sensitivity Analysis

The results of sensitivity analysis were displayed in Table 7.3 and Fig. 7.3. When profiles were cut at the ground-truth cleavage sites (denoted as "p" in Table 7.3 (a)), the overall accuracy reaches 98.47%. The prediction accuracy for Ext and Cyt/Nuc is above 98%. Despite the relatively weak signal, 80% of chloroplasts were correctly predicted. It is obvious that the localization performance degrades when the cut-off position drifts away from the ground-truth cleavage site. But the overall accuracy can be maintained at above 95% even if the drift is as large as -16 and $+64$ positions from the ground-truth.

It is obvious that mTP and cTP are more sensitive to the error of cleavage site

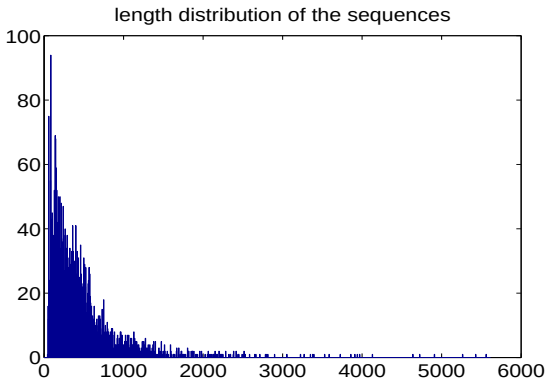
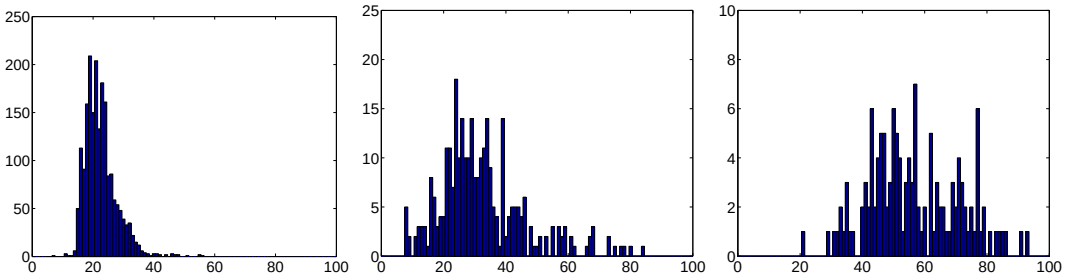


Figure 7.1: The histogram of the length of the sequences in our dataset. Vertical axis: number of occurrences; horizontal axis: sequence length.



(a) (b) (c)

Figure 7.2: The histograms of length of (a) secretory pathway signal peptides, (b) mitochondrial targeting peptides, and (c) chloroplast transit peptides. Vertical axis: number of occurrences; horizontal axis: sequence length.

prediction, which agrees with the fact that the signals of mTP and cTP are weaker. For comparison, another experiment was done in which the cleaved-off position was set to 170, i.e., none of the sequences (or profiles) have length exceeding 170. The prediction performance using fixed-length pre-profile alignment is shown in the last row of Table 7.3. Evidently, cutting the profiles/sequences at the cleavage sites can achieve a higher accuracy than cutting them at a fixed position. It is observed that a forward drift of 64 positions from the ground truth cleavage site leads to a higher overall accuracy when compared to that of a backward drift of 16 positions, which

Pro. Cutoff Position	Accuracy of Individual Class (%)				Overall Accuracy(%)
	Ext	Mit	Chl	Cyt/Nuc	
$p-16$	94.95	51.50	74.32	100.00	94.71
$p-8$	98.56	86.23	77.03	99.94	98.00
$p-2$	98.70	85.63	79.73	99.94	98.08
p	98.85	90.42	82.43	99.88	**98.47**
$p+2$	98.99	88.62	85.14	99.88	98.47
$p+16$	99.28	88.62	70.27	99.69	98.00
$p+32$	99.28	86.83	64.86	99.51	97.61
$p+64$	98.99	77.25	54.05	99.01	96.28
Fix-length(170)	91.92	53.89	28.38	97.28	90.98

(a)

Seq. Cutoff Position	Accuracy of Individual Class (%)				Overall Accuracy(%)
	Ext	Mit	Chl	Cyt/Nuc	
$p-16$	94.23	55.09	74.32	99.94	94.71
$p-8$	98.27	84.43	75.68	99.94	97.77
$p-2$	98.41	86.83	81.08	100.00	98.16
p	98.70	89.82	78.38	99.94	**98.31**
$p+2$	98.70	86.83	77.03	99.94	98.08
$p+16$	98.99	91.62	71.62	99.69	98.16
$p+32$	99.99	88.62	66.22	99.51	97.69
$p+64$	98.27	81.44	66.22	98.82	96.59
Fix-length(170)	91.62	56.89	30.32	96.98	90.46

(b)

Table 7.3: Sensitivity of subcellular localization accuracy with respect to the cut-off position. p is the ground-truth cleavage site. For "Cyt/Nuc" proteins, p is set to 170. (a) Full-length profiles were truncated, followed by profile alignment(see Scheme I in Fig. 5.2). (b) Full-length sequences were cut, followed by PSI-BLAST search(see Scheme II in Fig. 5.2).

suggests that cutting sequences before their cleavage sites may lose useful information in the signal peptides while including extra (may be irrelevant) information by cutting sequences after their cleavage sites is not detrimental to subcellular location accuracy. Table 7.3 shows that the performance of subcellular localization does not rely significantly on the precision of cleavage site prediction as long as the predicted sites are not too far away from the ground-truth.

Table 7.4: Cleavage-site prediction accuracies achieved by TargetP and CSitePred. For TargetP, (P) and (N) means using the 'Plant' and 'Non-plant' option of the predictor, respectively. TargetP will invoke SignalP, ChloroP, or a program specialized in predicting mTP for cleavage site prediction. CSitePred is based on conditional random fields.

Cleavage Site Predictor	Cleavage Site Prediction Accuracy (%)			
	SP	mTP	cTP	Overall
TargetP(P)	71.49	44.04	8.82	64.55
TargetP(N)	84.63	46.69	2.21	75.28
CSitePred	79.40	39.40	31.62	71.73

7.5 Performance of Cleavage Site Prediction

As demonstrated in the Section 7.4, the accuracy of subcellular localization depends on the positions at which the protein sequences are cut. Therefore, the performance of cleavage site predictor has significant effect on the performance of subcelluar localization predictor, especially for mTP and cTP.

Table 7.4 and Fig. 7.4 show the cleavage site prediction accuracy of TargetP and CSitePred (a CRFs-based predictor). It is shown that CSitePred is better than TargetP(P) in terms of predicting the cleavage sites of signal peptide (Ext) but is worse than TargetP(N). The results also suggest that while CSitePred is slightly inferior to TargetP in predicting the cleavage sites of mitochondria, it is significantly better than TargetP in predicting the cleavage site of chloroplasts. Note that the overall accuracies depend heavily on the Ext class because of the large number of signal peptides in the dataset (see Table 6.1).

Note that the prediction accuracy of chloroplasts in our experiments is significantly lower than that of [46]. There are two reasons for this difference: (1) our dataset has sequence identity lower than that of [46] and (2) we consider the prediction of the exact ground-truth position as a correct prediction whereas [46] consider a prediction within ± 2 of the ground-truth as a correct prediction. In fact, if we relaxed the

Seq. Cutoff position	Alignment Time for Each Sequence (sec.)	Accuracy of each Sequence Class (%)				Overall Accuracy (%)
		Ext	Mit	Chl	Cyt/Nuc	
Full length	34.7	95.15	51.94	32.22	97.14	91.64
170	4.7	91.92	53.89	28.38	97.28	90.98
Ground-truth	1.9	99.28	90.29	90.00	99.89	98.77
Determined by TargetP(P)	**1.8**	90.48	71.86	58.11	95.98	89.08
Determined by TargetP(N)	**1.7**	97.11	69.46	41.89	96.23	93.14
Determined by CSitePred	**1.9**	93.36	62.87	41.89	96.10	91.61

Table 7.5: Subcellular localization accuracy and computation time for different cutoff positions for sequences with and without cleavage sites. Computation time for alignment is the time taken to create a profile-alignment score matrix. In the first column, "Full length" means there is no cutoff for sequences, i.e., the whole sequences will be directly processed by PairProSVM. "TargetP(P)" and "TargetP(N)" mean that the cutoff position is determined by TargetP using the "Plant" option and "Non-plant" option, respectively. CSitePred is a cleavage site predictor based on conditional random fields.

criterion of correct prediction to ± 2 ground-truth positions, the prediction accuracy on chloroplasts achieved by TargetP increases to 47.06%.

7.6 Performance of Cascaded Fusion

Fig. 7.5 shows that the computation time for full-length profile alignment is striking — nearly thirty-five seconds per sequence, which suggests that full-length alignment is computationally prohibitive for most practical applications. Therefore, it is imperative to limit the length of the sequences or profiles before alignment. It is shown in Table 7.5 that our method leads to nearly a 20 folds reduction in computation time, which is consistent with the computational complexity of Smith-Waterman Alignment as shown in Table 7.6. This is because the signal segment can be found in the N-terminus, and removing the amino acids beyond the cleavage site helps the alignment

	Average Length (Amino Acids)	Alignment Time (seconds/protein)
Before Truncation	469	34.2
After Truncation	108	1.8

Table 7.6: The average length of protein sequences and alignment time before and after truncation. The computational complexity of Smith-Waterman Alignment is $O(N^2)$.

focus on the relevant features in the sequences and disregard noise.

7.7 Comparing Profile Creation Schemes

Fig. 7.6 shows the score matrices obtained by the two profile creation schemes (see Fig. 5.2). The figure shows that the two alignment score matrices exhibit a similar pattern, suggesting that classifiers based on these matrices will produce similar classification accuracy. This argument is confirmed by Table 7.7, which shows that cutting the sequences at cleavage sites before inputting to PSI-BLAST can reduce the profile creation time by 6 times without significant reduction in subcellular localization accuracy.

7.8 SVM versus KPDA

Table 7.8 shows that the training time of KPDA and KPDAproj+SVM are only one-fifth of that of SVM. However, the accuracy of KPDA and KPDAproj+SVM are lower than that of SVM. It is worth mentioning that the classification time is much less than that spent on profile creation and profile alignment. So the reduction of the classification time does not have profound effect on the whole computation time.

7.9 Compared with State-of-the-Art Predictors

We compared the accuracy of our cascaded fusion method with SubLoc [20], TargetP [4] and PairProSVM [33]. Table 7.9 shows that the overall accuracy of the proposed method (the 5th row) is 5.2% higher than that of TargetP (3rd row) and is significantly better than that of SubLoc (1st row). Our method outperforms TargetP in Ext and Cyt/Nuc prediction while performing worse than TargetP in predicting Mit and Chl. One limitation of TargetP is that users need to select either "Plant" or "Non-plant". If the former is selected, the performance of Ext and Cyt/Nuc degrade significantly, leading to a low overall accuracy; if the latter is selected, none of the chloroplast proteins can be correctly predicted. The cascaded fusion of cleavage site prediction and PairProSVM, on the other hand, can classify all four classes with fairly high accuracy, leading to a higher overall accuracy.

The prediction accuracy and MCC of the proposed methods (Rows 4–10 in Table 7.9) are comparable to PairProSVM (Row 4 in Table 7.9). The main improvement, as discussed in Section 7.7, is on time reduction.

Because ChloroP is weak in predicting the cleavage sites of chloroplasts (see Table 7.4), it is not a good candidate for assisting PairProSVM. This is evident by the low subcellular localization accuracy of chloroplasts in Table 7.9 when TargetP is used as a cleavage site predictor. However, TargetP is fairly good at predicting the subcellular location of chloroplasts when it is used as a localization predictor.

Among the four classes in Table 7.9, the subcellular localization accuracies of mitochondria and chloroplasts are generally lower than that of Ext and Cyt/Nuc. The reason may be that these transit peptides are less well characterized and their motifs are less conserved than those of secretary SP [18].

Table 7.9 also suggests that the TargetP(N) is very effective in assisting PairProSVM, leading to the highest prediction accuracy (92.6%) among all subcellular localization predictors. In particular, TargetP combining with PairProSVM can sur-

Scheme	Input to PSI-BLAST	Profile Creation Time (sec.)	Subcellular Localization Accuracy
I	Full-length sequences	30.5	91.69%
II	Sequences truncated at predicted cleavage sites	4.7	91.61%

Table 7.7: Average computation time to create a profile by PSI-BLAST using sequences of different length as input. In Scheme I, full-length sequences were presented to PSI-BLAST and the resulting profiles were truncated at the predicted cleavage sites. In Scheme II, truncation was applied to the sequences before presenting to PSI-BLAST. In both cases, CRFs (CSitePred) were used to predict the cleavage sites.

Classification Method	Training Time (sec.)	Classification Time (sec.)	SubLoc Acc.
SVM	51.4	0.7	91.45%
KPDA	9.9	1.9	90.24%
KPDAproj+SVM	8.9	0.1	89.97%

Table 7.8: The computation time and performance of different classifiers in the subcellular localization task. The classification time is the time to classify a profile-alignment score vector with dimension equal to the number of training vectors. The training time is time required to train a classifier, given a profile-alignment score matrix. In KPDAproj+SVM, KPDA was applied to project the samples in the input space to a $(C-1)$-dim space ($C = 4$ here); the projected vectors were then classified by RBF-SVMs.

pass the other methods in subcellular localization accuracy and MCC in all categories, except in predicting Chl (worse than TargetP(P)).

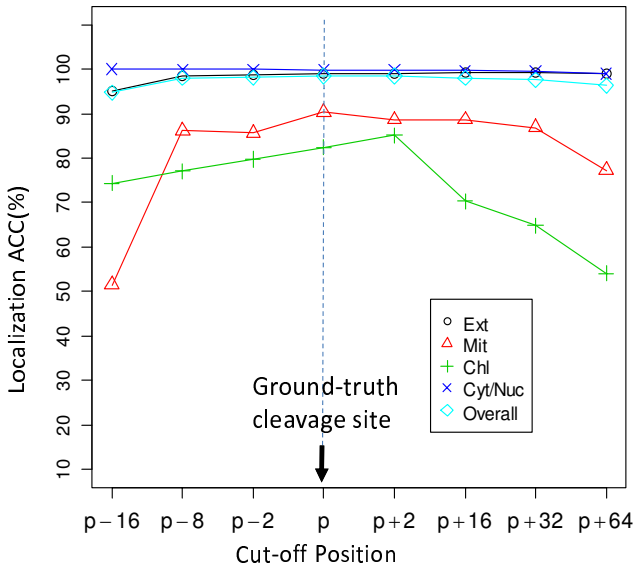
Row	Cleavage Site Predictor	Localization Predictor	Classification Accuracy (%)				
			Ext	Mit	Chl	Cyt/Nuc	Overall
1	—	SubLoc [20]	51.4	55.8	—	77.9	66.8
2	—	TargetP (P)	79.1	88.0	89.2	69.6	73.9
3	—	TargetP (N)	97.4	89.2	0.0	87.9	88.0
4	—	SVM	95.2	52.0	32.2	97.1	91.6
5	TargetP(N)	SVM	97.3	67.1	36.5	95.9	92.6
6	TargetP(N)	KPDA	97.6	61.7	6.8	95.6	91.3
7	TargetP(N)	KPDAproj + SVM	97.3	65.3	37.8	93.6	91.1
8	CRFs	SVM	94.5	63.5	28.4	95.9	91.5
9	CRFs	KPDA	94.8	59.4	1.4	95.6	90.2
10	CRFs	KPDAproj + SVM	94.7	63.5	25.7	93.6	90.0

(a)

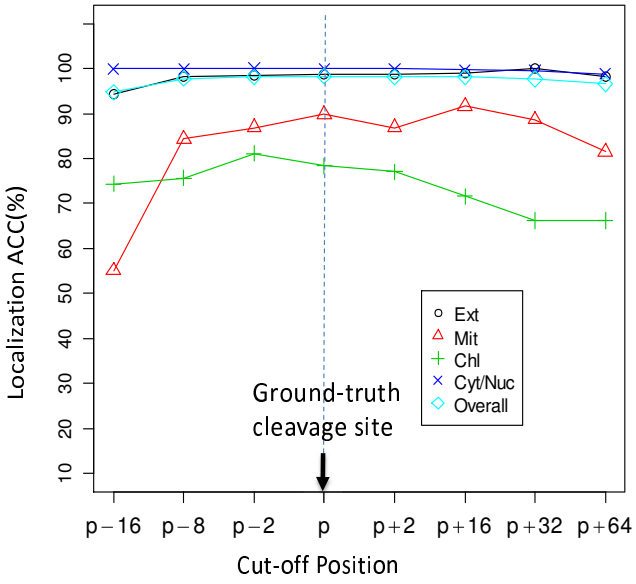
Row	Cleavage Site Predictor	Localization Predictor	Matthew's correlation coefficient (MCC)				
			Ext	Mit	Chl	Cyt/Nuc	Overall
1	—	SubLoc [20]	—	—	—	—	—
2	—	TargetP (P)	0.79	0.49	0.79	0.64	0.65
3	—	TargetP (N)	0.93	0.58	0.00	0.81	0.84
4	—	SVM	0.92	0.62	0.50	0.85	0.89
5	TargetP(N)	SVM	0.93	0.70	0.53	0.86	0.90
6	TargetP(N)	KPDA	0.91	0.68	0.26	0.84	0.88
7	TargetP(N)	KPDAproj + SVM	0.93	0.64	0.50	0.83	0.88
8	CRFs	SVM	0.90	0.68	0.45	0.84	0.89
9	CRFs	KPDA	0.88	0.67	0.11	0.82	0.81
10	CRFs	KPDAproj + SVM	0.90	0.60	0.41	0.82	0.87

(b)

Table 7.9: Subcellular localization performance achieved by different classifiers: (a) classification accuracy, (b) Matthew's correlation coefficient (MCC). For each table, the second column specifies the the cleavage site predictors that were used for determining the positions at which the amino sequences were truncated. Notice that TargetP can perform both cleavage site prediction and subcellular localization. For Rows 5–7, TargetP was used as a cleavage site predictor, where "TargetP(N)" mean selecting non-plant option in TargetP. For Rows 8–10 "CRFs" means that conditional random fields were used for cleavage site prediction.



(a)



(b)

Figure 7.3: Sensitivity of subcellular localization accuracy with respect to the (a) profile cut-off positions and (b) sequence cut-off positions. p is the groundtruth cleavage site. For "Cyt/Nuc" proteins, p is set to 170.

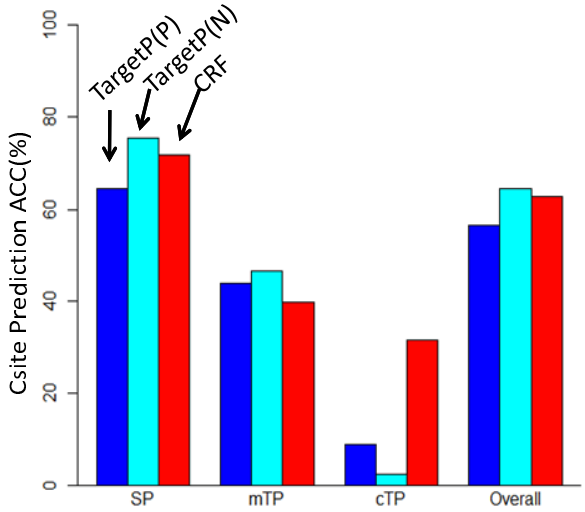


Figure 7.4: Cleavage-site prediction accuracies achieved by TargetP and CSitePred.

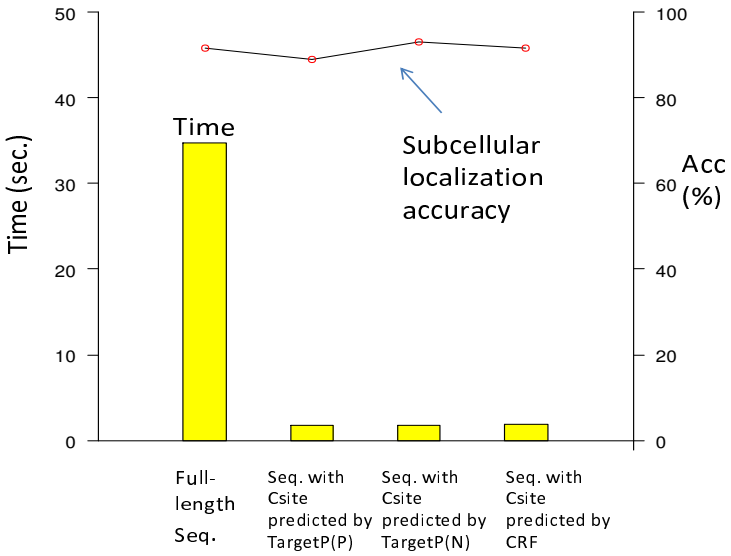
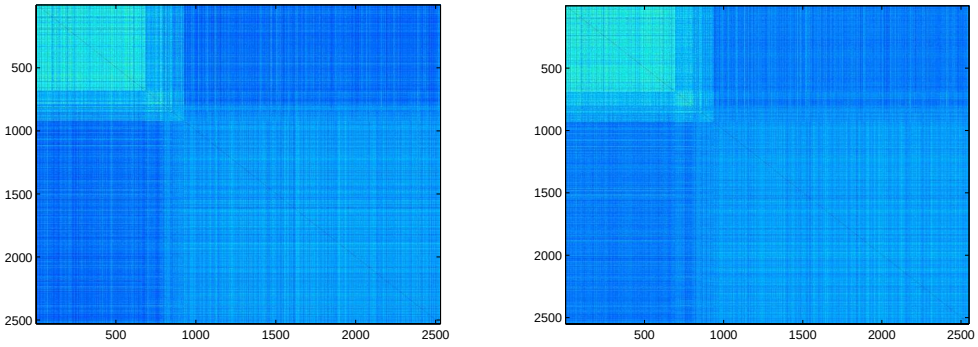


Figure 7.5: The performance of cascaded fusion. The computational time and localization prediction accuracies for profile alignment is shown for different length of sequences: the full-length sequence, the cut-off sequence predicted by TargetP and CRFs. The bars represent the computation time for alignment and the circles represent the overall prediction accuracies.



(a) (b)

Figure 7.6: Profile-alignment score matrices produced by (a) Scheme I and (b) Scheme II in Fig. 5.2.

Chapter 8

CONCLUSIONS

The research presented in this thesis has demonstrated that there is a high degree of complementarity between CRF-based predictors and SignalP, and that this complementary information can be easily exploited to fuse the two types of predictors in a protein cleavage site prediction task. Our work also shows that the CRF can be further enhanced by constructing state features from more spatially dispersed amino acids along the peptide chain.

We proposed a novel subcellular-localization-prediction method that is based on the cascaded fusion of signal-based and homology-based methods. Our work shows that homology-based subcellular localization can be speeded up by reducing the length of the query amino acid sequences. Because shortening an amino acid sequence will inevitably throw away some information in the sequence, it is imperative to determine the best truncation positions. We found that these positions can be determined by cleavage site predictors such as TargetP and CSitePred. Our work also shows that as far as localization accuracy is concerned, it does not matter whether we truncate the sequences or truncate the profiles. However, truncating the sequence has computation advantage because this strategy can save the profile creation time by as much as 6 folds.

This thesis has three key findings: (1) cTP is more sensitive to the error of cleavage site prediction; (2) cutting the profiles or sequences at the cleavage sites can achieve a significant reduction in computation time; and (3) discrepancy of cleavage site prediction is inversely proportional to the subcellular localization prediction accuracy. We hope that this in-silico method can be complementary to experimental subcellular

localization techniques.

There are still open challenges for the development of new methods, such as increasing the classification accuracy across many cellular compartments, allowing for multiple predictions per protein and stabilizing performance across many species. The recent availability of large protein-protein interaction networks provides the possibility to partially address these challenges. Integrated analysis of protein-protein interaction data suggests that interaction may serve as an indicator for co-localization. It is also found that protein networks can be applied to predict localization of proteins and markedly improve the prediction accuracy in higher eukaryotes [74]. It is therefore of interest to investigate how protein interaction networks can provide further insight into the prediction of subcellular localization.

BIBLIOGRAPHY

[1] C. H. Wu and J. M. McLarty, *Neural Networks and Genome Informatics*, Elsevier Science, 2000.

[2] L. M. Gierasch, "Signal sequences," *Biochemistry*, vol. 28, pp. 923930, 1989.

[3] G. von Heijne, "The signal peptide," *Journal of Membrane Biology*, vol. 115, no. 3, pp. 195–201, 1990.

[4] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence," *J. Mol. Biol.*, vol. 300, no. 4, pp. 1005–1016, 2000.

[5] G. Schatz and B. Dobberstein, "Common principles of protein translocation across membranes," *Science*, vol. 271, no. 5255, pp. 1519, 1996.

[6] G. Heijne, J. Steppuhn, and R.G. Herrmann, "Domain structure of mitochondrial and chloroplast targeting peptides," *European journal of biochemistry*, vol. 180, no. 3, pp. 535–545, 1989.

[7] O. Emanuelsson, G. von Heijne, and G. Schneider, "Analysis and prediction of mitochondrial targeting peptides," *Methods in Cell Biology*, vol. 65, pp. 175–187, 2001.

[8] T. Kleffmann, D. Russenberger, A. von Zychlinski, W. Christopher, K. Sjolander, W. Gruissem, and S. Baginsky, "The Arabidopsis thaliana chloroplast proteome reveals pathway abundance and novel protein functions," *Current Biology*, vol. 14, no. 5, pp. 354–362, 2004.

[9] K.C. Chou and H.B. Shen, "Recent progress in protein subcellular location prediction," *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.

[10] G. von Heijne, "A new method for predicting signal sequence cleavage sites," *Nucleic Acids Research*, vol. 14, no. 11, pp. 4683–4690, 1986.

[11] K. Nakai and M. Kanehisa, "Expert system for predicting protein localization sites in gram-negative bacteria," *Proteins: Structure, Function, and Genetics*, vol. 11, no. 2, pp. 95–110, 1991.

[12] K. Nakai and M. Kanehisa, "A knowledge base for predicting protein localization sites in eukaryotic cells," *Genomics*, vol. 14, no. 4, pp. 897–911, 1992.

[13] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, "A neural network method for identification of prokaryotic and eukaryotic signal perptides and prediction of their cleavage sites," *Int. J. Neural Sys.*, vol. 8, pp. 581–599, 1997.

[14] H. Nielsen, J. Engelbrecht, S. Brunak, and G. Von Heijne, "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites," *Protein Engineering Design and Selection*, vol. 10, no. 1, pp. 1, 1997.

[15] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.

[16] P. Horton, K. J. Park, T. Obayashi, and K. Nakai, "Protein subcellular localization prediction with WoLF PSORT," in *Proc. 4th Annual Asia Pacific Bioinformatics Conference (APBC06)*, 2006, pp. 39–48.

[17] P. Horton, K.J. Park, T. Obayashi, N. Fujita, H. Harada, CJ Adams-Collier, and K. Nakai, "WoLF PSORT: protein localization predictor," *Nucleic acids research*, vol. 35, no. Web Server issue, pp. 585–587, 2007.

[18] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, "Locating proteins in the cell using TargetP, SignalP, and related tools," *Nature Protocols*, vol. 2, no. 4, pp. 953–971, 2007.

[19] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *J. Mol. Biol.*, vol. 238, pp. 54–61, 1994.

[20] S. J. Hua and Z. R. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, pp. 721–728, 2001.

[21] Y. Jin, B. Niu, K.Y. Feng, W.C. Lu, Y.D. Cai, and G.Z. Li, "Predicting subcellular localization with AdaBoost Learner," *Protein and Peptide Letters*, vol. 15, no. 3, pp. 286–289, 2008.

[22] B. Niu, Y.H. Jin, K.Y. Feng, W.C. Lu, Y.D. Cai, and G.Z. Li, "Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins," *Molecular diversity*, vol. 12, no. 1, pp. 41–45, 2008.

[23] A. Garg, M Bhasin, and G. P. S. Raghava, "SVM-based method for subcellular localization of human proteins using amino acid compositions, their order and similarity search," *J. of Biol. Chem.*, vol. 280, pp. 14427–14432, 2005.

[24] X. Qian, H. Derek, X. Hong, Y. Weichuan, and Y. Qiang, "Semi-supervised protein subcellular localization," *BMC Bioinformatics*, vol. 10, 2009.

[25] K. C. Chou, "Prediction of protein cellular attributes using pseudo amino acid composition," *Proteins: Structure, Function, and Genetics*, vol. 43, pp. 246–255, 2001.

[26] Z. Yuan, "Prediction of protein subcellular locations using Markov chain models," *FEBS Letters*, vol. 451, no. 1, pp. 23–26, 1999.

[27] K. J. Park and M. Kanehisa, "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs," *Bioinformatics*, vol. 19, no. 13, pp. 1656–1663, 2003.

[28] Y. Huang and Y. Li, "Prediction of protein subcellular locations using fuzzy k-NN method," *Bioinformatics*, vol. 20, no. 1, pp. 21–28, 2004.

[29] R. Mott, J. Schultz, P. Bork, and C.P. Ponting, "Predicting protein cellular localization using a domain projection method," *Genome research*, vol. 12, no. 8, pp. 1168–1174, 2002.

[30] M.S. Scott, D.Y. Thomas, and M.T. Hallett, "Predicting subcellular localization via protein motif co-occurrence," *Genome research*, vol. 14, no. 10a, pp. 1957–1966, 2004.

[31] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner, "Predicting subcellular localization of proteins using machine-learned classifiers," *Bioinformatics*, vol. 20, no. 4, pp. 547–556, 2004.

[32] JK Kim, GPS Raghava, SY Bang, and S Choi, "Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine," *Pattern Recognition Letters*, vol. 27, no. 9, pp. 996–1001, 2006.

[33] M. W. Mak, J. Guo, and S. Y. Kung, "PairProSVM: Protein subcellular localization based on local pairwise profile alignment and SVM," *IEEE/ACM Trans. on Computational Biology and Bioinformatics*, vol. 5, no. 3, pp. 416 – 422, 2008.

[34] L. Hsin-Nan, C. Ching-Tai, S. Ting-Yi, H. Shinn-Ying, and H. Wen-Lian, "Protein subcellular localization prediction of eukaryotes using a knowledge-based approach," *BMC Bioinformatics*, vol. 10, 2009.

[35] E. Bolten, A. Schliep, S. Schneckener, D. Schomburg, and R. Schrader, "Clustering protein sequences–structure prediction by transitive homology," *Bioinformatics*, vol. 17, no. 10, pp. 935–941, 2001.

[36] Z. Lei and Y. Dai, "Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction," *BMC bioinformatics*, vol. 7, no. 1, pp. 491, 2006.

[37] W.L. Huang, C.W. Tung, S.W. Ho, S.F. Hwang, and S.Y. Ho, "ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization," *Bmc Bioinformatics*, vol. 9, no. 1, pp. 80, 2008.

[38] R. Nair and B. Rost, "Inferring sub-cellular localization through automated lexical analysis," *Bioinformatics*, vol. 18, pp. S78–S76, 2002.

[39] S. Brady and H. Shatkay, "EpiLoc: a (working) text-based system for predicting protein subcellular location," in *Pac Symp Biocomput*. Citeseer, 2008, vol. 604, p. 15.

[40] A. Fyshe, Y. Liu, D. Szafron, R. Greiner, and P. Lu, "Improving subcellular localization prediction using text classification and the gene ontology," *Bioinformatics*, vol. 24, no. 21, pp. 2512, 2008.

[41] S. Briesemeister, T. Blum, S. Brady, Y. Lam, O. Kohlbacher, and H. Shatkay, "SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins," *J. Proteome Res*, vol. 8, no. 11, pp. 5363–5366, 2009.

[42] W. Wang, M. W. Mak, and S. Y. Kung, "Speeding up subcellular localization by extracting informative regions of protein sequences for profile alignment," in *Proc. Computational Intelligence in Bioinformatics and Computational Biology*, Montreal, May 2010, pp. 147–154.

[43] M. W. Mak, W. Wang, and S. Y. Kung, "Truncation of protein sequences for fast profile alignment with application to subcellular localization," in *IEEE BIBM 2010*, Hong Kong, Dec 2010.

[44] P. Baldi and S. Brunak, *Bioinformatics : The Machine Learning Approach*, MIT Press, 2 edition, 2001.

[45] H. Nielsen, S. Brunak, and G. von Heijne, "Machine learning approaches for the prediction of signal peptides and other protein sorting signals," *Protein Eng.*, vol. 12, no. 1, pp. 3–9, 1999.

[46] O. Emanuelsson, H. Nielsen, and G. von Heijne, "Chlorop, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites," *Protein Science*, vol. 8, pp. 978–984, 1999.

[47] M. A. Andrade, S. I. O'Donoghue, and B. Rost, "Adaptation of protein surfaces to subcellular location," *Journal of Molecular Biology*, vol. 276, no. 2, pp. 517–525, 1998.

[48] R. Nair and B. Rost, "Sequence conserved for subcellular localization," *Protein Science*, vol. 11, pp. 2836–2847, 2002.

[49] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, 1997.

[50] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, R. Grothe, and T.O. Yeates, "Assigning protein functions by comparative genome analysis protein phylogenetic profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4285–4288, 1999.

[51] K. Hiller, A. Grote, M. Scheer, R. Munch, and D. Jahn, "PrediSi: Prediction of signal peptides and their cleavage positions," *Nucleic Acids Research*, vol. 32, pp. 375–379, 2004.

[52] S. R. Maetschke, M. Towsey, and M. B. Boden, "BLOMAP: An encoding of amino acids which improves signal peptide cleavage site prediction," in *3rd Asia Pacific Bioinformatics Conference*, Y. P. Phoebe Chen and L. Wong, Eds., Singapore, 17-21 Jan 2005, pp. 141–150.

[53] H. Nielsen and A. Krogh, "Prediction of signal peptides and signal anchors by a hidden Markov model," in *Proc. Sixth Int. Conf. on Intelligent Systems for Molecular Biology*, J. Glasgow et al., Ed. 1998, pp. 122–130, AAAI Press.

[54] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, "Improved prediction of signal peptides: Signalp 3.0," *J. Mol. Biol.*, vol. 340, pp. 783–795, 2004.

[55] M. W. Mak and S. Y. Kung, "Conditional random fields for the prediction of signal peptide cleavage sites," in *Proc. ICASSP*, Taipei, April 2009, pp. 1605–1608.

[56] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. on Machine Learning*, 2001.

[57] O. Weiss and H. Herzel, "Correlations in protein sequences and property codes," *J. theor. Biol*, vol. 190, pp. 341–353, 1998.

[58] C. Hemmerich and S. Kim, "A study of residue correlation within protein sequences and its application to sequence classification," *EURASIP J. Bioinformatics Syst. Biol.*, vol. 2007, no. 1, pp. 9–9, 2007.

[59] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing 14*, Cambridge, MA, 2002, MIT Press.

[60] K. Sato and Y. Sakakibara, "RNA secondary structural alignment with conditional random fields," *Bioinformatics*, vol. 21, pp. 237–242, 2005.

[61] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Proc. of the Third Workshop on Very Large Corpora*, Cambridge, MA, 1995.

[62] S. Shimozono, "Alphabet indexing for approximating features of symbols," *Theor. Comput. Sci.*, vol. 210, no. 2, pp. 245–260, 1999.

[63] G. von Heijne, "Patterns of amino acids near signal-sequence cleavage sites," *Eur J Biochem.*, vol. 133, no. 1, pp. 17–21, Jun 1983.

[64] M. W. Mak, W. Wang, and S. Y. Kung, "Fusion of conditional random field and SignalP for protein cleavage site prediction," in *Proc. APSIPA'09*, Supporo, Oct. 2009, pp. 716–721.

[65] H. Rangwala and G. Karypis, "Profile-based direct kernels for remote homology detection and fold recognition," *Bioinformatics*, vol. 21, no. 23, pp. 4239–4247, 2005.

[66] V.N. Vapnik, *The nature of statistical learning theory*, Springer Verlag, 2000.

[67] S. Y. Kung, "Kernel approaches to unsupervised and supervised machine learning," in *Proc. PCM*, P. Muneesawang, et al., Ed. 2009, LNCS 5879, pp. 1–32, Springer-Verlag.

[68] K. Tsuda., "Support vector classifier with asymmetric kernel functions," in *Proceedings ESANN*, Brussels, 1999, pp. 183–188.

[69] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*, Y. H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds., 1999, pp. 41–48.

[70] http://158.132.148.85:8080/CSitePred/faces/Page1.jsp, ," .

[71] K. M. L. Menne, H. Hermjakob, and R. Apweiler, "A comparison of signal sequence prediction methods using a test set of signal peptides," *Bioinformatics*, vol. 16, pp. 741–742, 2000.

[72] J. Sprenger, J.L. Fink, and R. Teasdale, "Evaluation and comparison of mammalian subcellular localization prediction methods," *BMC bioinformatics*, vol. 7, no. Suppl 5, pp. S3, 2006.

[73] B. W. Matthews, "Comparison of predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[74] K.Y. Lee, H.Y. Chuang, A. Beyer, M.K. Sung, W.K. Huh, B. Lee, and T. Ideker, "Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species," *Nucleic Acids Research*, vol. 36, no. 20, pp. e136, 2008.