Development of a novel mitochondrial cleavage site predictor

Yoshinori Fukasawa

July 29, 2011

Abstract

A large fraction of mitochondrial proteins are cleaved upon entry into the mitochondria, but prediction of this cleavage is still challenging. In chapter 1, I summarize necessary background to this important problem. In chapter 2, I demonstrate that my system, Mitochondrial matrix targeting Signal Predictor, MoiraiSP, which is based on data from recent proteomic studies can correctly identified the cleavage site of MPP more than 75% accuracy in both plant and yeast dataset. In chapter 3, I introduce sequence divergence, LD(i), as a novel feature for sorting signal prediction, and show that prediction can be improved by LD(i) than random, especially with other famous features such as physico-chemical propensities. In chapter 4, I present that MoiraiSP can treat a related problem, predicting mitochondrial matrix targeting signal by using only N-terminal sequence information such as net-charge or LD(i). MoiraiSP discriminates between cleaved and non-cleaved mitochondrial proteins with a success rate of 97% (plant) or 91% (yeast) by cross validation. Finally, in chapter 5, I discuss some novel candidates of protease substrates, which came up during my work.

Acknowledgments

Two years ago when I started research at Computational Biology Research Center as a graduate student almost required skill and knowledge was not in me. Since I have had precious experience at this institute, I owe to the office space and environment CBRC provides.

I am indebted to my mentor, Prof. Paul Horton, for his help and advice throughout this research. His critical suggestions improved the result and revised the way of this work at several times.

Special thanks are due to Dr. Raymond Wan, Dr. Kenichiro Imai and Dr. Noriyuki Sakiyama for invaluable feedback and preparation of this work from the beginning. Also, Ross KK Leung kindly had many long discussion with me for MTS classification and helped sequence alignment of this work.

As a graduate student at the university of Tokyo, graduate students at CBRC have been always good responders and expressed profound patience to our discussion. I owe my basic knowledge and education of machine-learning method to the long discussion I have had with Dr. Szu-chin Fu and vital inspiration on this work to the discussion I have had with Junko Tsuji and Hiroki Ashida.

Finally, I want to thank my family for their support. Without their financial and moral aid, I would not continue my career.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Nuclear-encoded mitochondrial precursor proteins are generally recognized by receptors embedded in the mitochondrial membranes. At present, these proteins can be divided into two main groups in terms of their targeting signals: the amino-terminal signal (presequence) and non-cleavable internal targeting signals. Therefore, in the current mitochondrial research it is important to reveal those N-terminal and internal signals.

Matrix targeting signals (MTS), an amino-terminal presequence targeting proteins to the mitochondria, are usually eliminated in the mitochondrial matrix by Mitochondrial Processing Peptidase, a metallo-protease in the matrix, and other intermediate peptidases such as Oct1, which function after MPP in some cases [1]. Mitochondrial presequences are harmful for the function of mitochondrial membranes due to their amphiphilical property, as a result they dissipate membrane potential and uncouple respiration [2, 3]. To avoid such severe disturbances, MPP cleaves presequences and it

is reported that other metallo-protease degrades them after cleavage in *Arabidopsis thaliana* [4]. Although cleavage site prediction of signal peptides, a similar biological phenomenon to mitochondrial cleavage, has been successful, the sequence determinants of mitochondrial cleavage are still unclear and the prediction is challenging for bioinformatics.

Until recently, it has been said that MPP cleavage site contains at least four classes in terms of arginine position; namely, arginine at -10 position (R-10), at -3 position (R-3 motif), at -2 position (R-2 motif), and no arginine (R-none class) [1, 5]. R-10 motif implies two cleavages in the mitochondria: MPP and Oct1 [5, 6]. A similar specificity of MPP was also observed in plants as well, however, the R-10 motif has not been found [7]. In addition, a mitochondrion contains other proteases such as Pcp1, m-AAA, and IMP in its inter-membrane [8]. Compared to proteases in the matrix, their specificities are still obscure. For the above reasons, cleavage site prediction of mitochondrial proteins is still a hard problem even for MPP cleavage sites.

A recent proteomic study showed a novel intermediate protease named Icp55 can remove one amino acid from the N-terminal, and the relationship between this phenomena and the half-life of a protein determined by its N-terminal residue [9]. Almost all R-3 sites turn into R-2 sites due cleavage by Icp55. This discovery of Icp55 partly explains the complexity of mitochondrial cleavage. Schneider *et al.* reported that the amino acid frequency at the -1 position of the R-3 motif is dominated by tyrosine, and this amino acid (which has a destabilizing role) is cleaved by Icp55 [5, 9]. Taking Icp55 and Oct1 into account, mitochondrial cleavage can be explained MPP cleavage and intermediate cleavage after MPP (Figure 1.1).

The location of each proteases and relation among them are summarized in Figure 1.2.

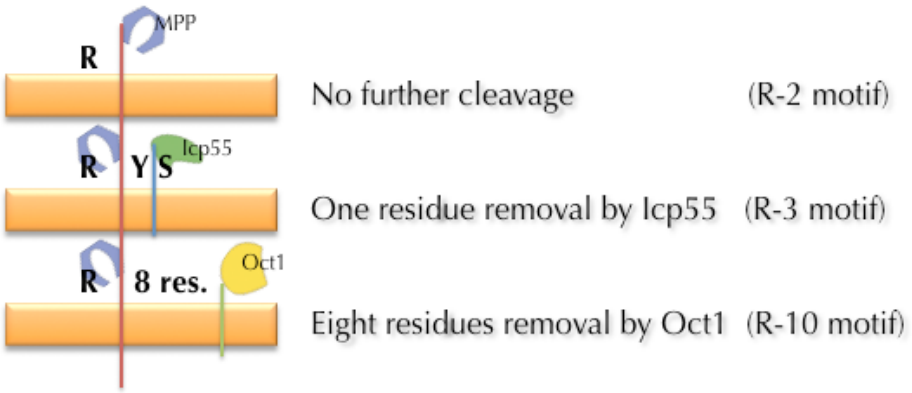

Figure 1.1: Description for mitochondrial cleavage.

1.2 Motivation and related work

In recent years, it has been revealed that cleavage in mitochondria is related not only to removal of signal sequence but also to quality control of mitochondria themselves. Since mitochondria is the power plant of a cell, it is said that dysfunction of mitochondria can lead to severe human diseases such as Parkinson's disease [10]. In fact, numerous neuronal diseases are related to the mitochodrial proteases [11]. Additionally, it has been revealed that a presequence cleaved by m-AAA supports the folding of MrpL32, a subunit of mitochondrial ribosomal complex, in the matrix [12]. Therefore, the importance to know cleavage site and its relevant proteases in mitochondria increases in the field of both medical and biological science. At present, there are two well known predictors for mitochondrial cleavage: TargetP [13] and MitoProtII [14]. These two systems depend on only the classical presequence signals and motifs, due to the shortage of available experimental data when they were developed. To accelerate research in this field, the need for improvement of prediction accuracy and integrating new insights such as existence of Icp55 or non-cleaved proteins have been discussed [7, 9]. At the very least, mitochondrial cleavage site predictors should treat plant and non-plant separately because the plant cleavage site motif does not show the typical Oct1 motif at present and the length of their presequences has a different distribution than those of yeast [7].

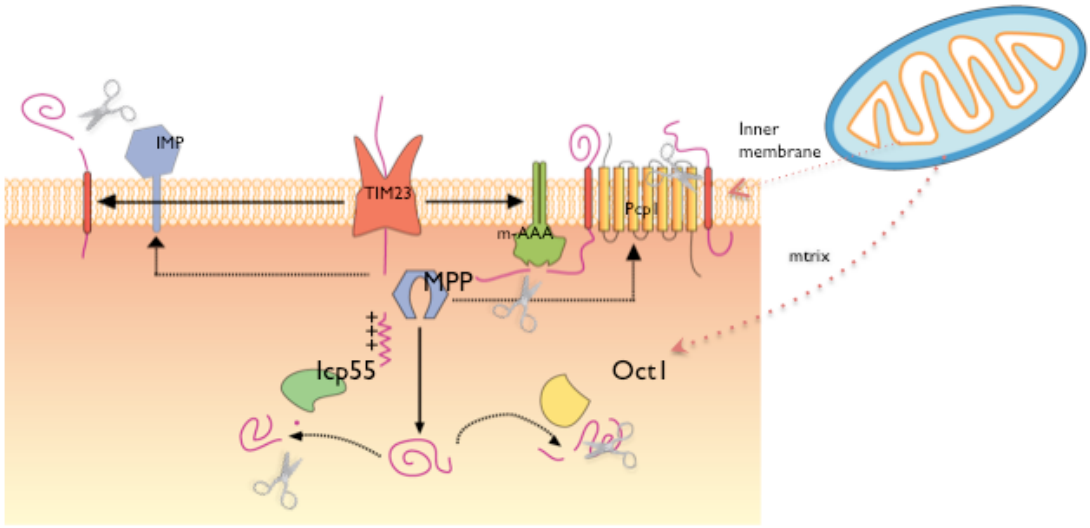

Figure 1.2: Proteases in a yeast mitochondrion.

Although a large portion of MTS's are cleavable N-terminal extended polypeptides, not all MTS's are necessarily eliminated by MPP in the matrix [15]. However, existing software does not consider this possibility. Because the MTS is recognized by TOM40 complex in the mitochondrial outer membrane [16], recognition of the R-2(3) motif by MPP seems to be independent from the MTS signal.

Here I present a novel cleavage site predictor based on recent proteomic experiments and a combination of simple profile Hidden Markov Model and Support Vector Machine. Our aim is accurate prediction of the cleavage site of mitochondrial proteins, especially related to MPP and related proteases, by incorporating recent finding in mitochondrial experimental research. This novel predictor consists of two layers predictor and predicts a cleavage site from only a query sequence, Figure 1.3.

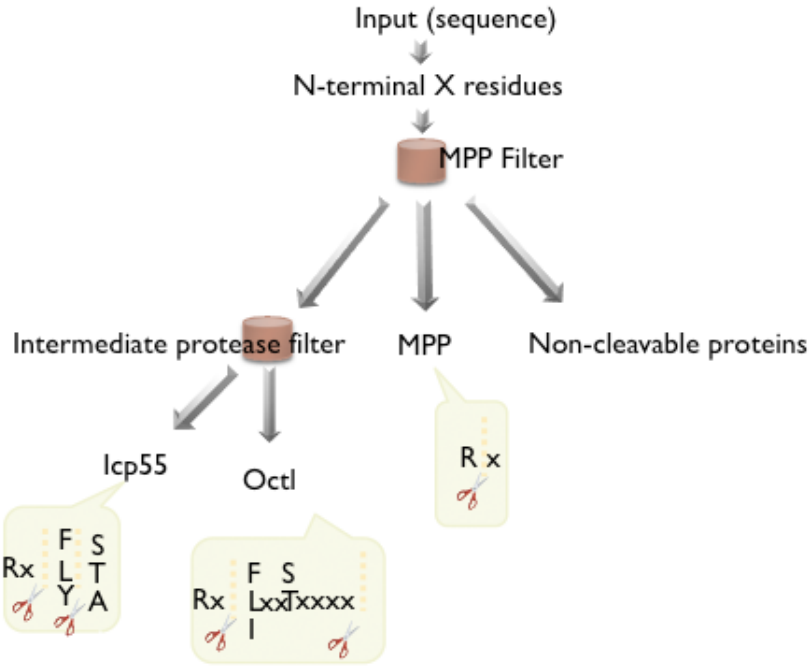

Figure 1.3: Flowchart of the prediction in case of the yeast.

Chapter 2

Prediction of MTS cleavage site

2.1 Results

2.1.1 Basis of dataset

All of the sequences were extracted from the proteomic analysis experiments for *Saccharomyces cerevisiae* (hereafter "*S.cere.*"), *Arabidopsis thaliana* and *Oryza satia* [7, 9], which we further curated. Following the annotation in the papers, I split up the data sets into the two categories: cleaved mitochondrial proteins and non-cleaved mitochondrial proteins.

To avoid incorrect motif finding and reduce noise, inappropriate sequences such as overlapping sequences in both the above two categories were excluded from the datasets. Also, sequences which have already been determined to localize outside the mitochondria are excluded. In particular, the original yeast dataset includes multiple cleavage sites for one protein in some cases, therefore, we selected the cleavage site identified most frequently as the representative site. After the above filtering, redundant sequences were removed.

As a result, the yeast data set has 245 sequences as cleaved proteins and 110 as non-cleaved

proteins. Arabidopsis and rice data set has 112 cleaved proteins and 37 non-cleaved proteins. Additionally, our literature search revealed numerous proteases explaining some of the cleaved sites in the yeast dataset such as i-AAA, m-AAA and IMP[17, 18], which do not contain arginine at the -2 position; therefore, at least the yeast dataset seem to contain diverse kinds of processed sites and the plant dataset might as well.

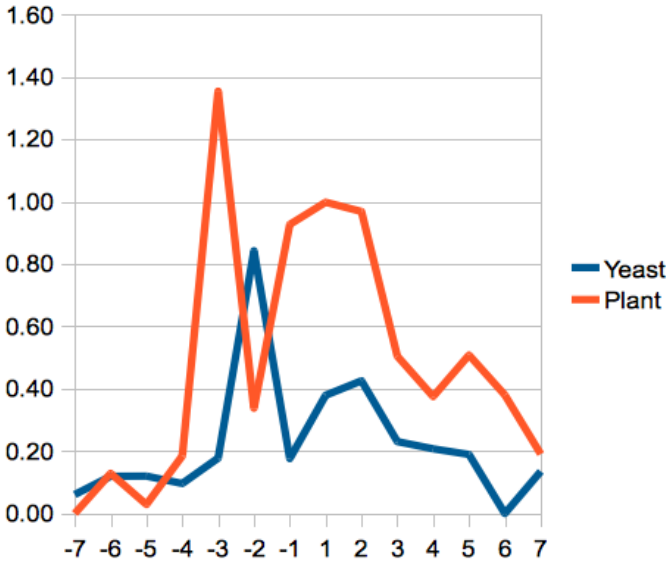

Figure 2.1: Entropy of yeast and plant.

2.1.2 R-2 and R-3 motif are main motifs around cleavage sites

To help understand our complex datasets, I calculated the entropy of each positions using 14 residues around the cleavage sites in the yeast data set.(Figure 2.1)

$$H(i) = -\sum_{j \in A} F(i,j) \lg F(i,j). \tag{2.1}$$

where i indicates position of a column and j stands for a kind of amino acid.

To make the plot more distinct, the entropy at each positions was subtracted from the maximum

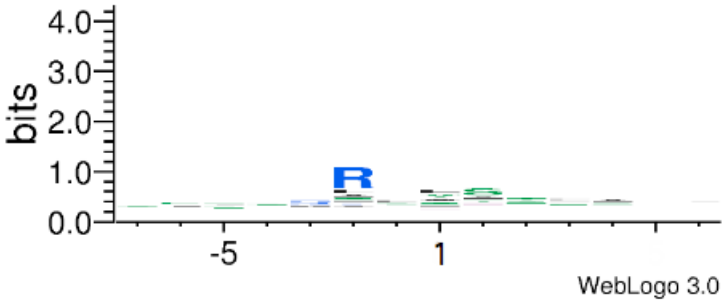

Figure 2.2: Sequence logo generated from the yeast data.

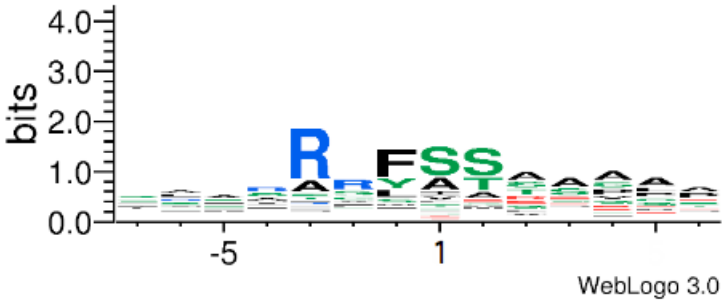

Figure 2.3: Sequence logo generated from the plant data.

value. -2 position of the yeast and -3 position of plant are most conserved, and conserved arginine at -2 positions is related to MPP cleavage as previously reported [1]. In fact, almost all substrates of both Icp55 and Oct1, 34 out of 37 for Icp55 and 12 out of 13 for Oct1, contain arginine at -2 position in their cleavage sites [9]. Additionally, the R-2 motif was recognized in the plant data set [7]. Figure 2.2, 2.3 which are generated from all of the sequences show conservation near cleavage site. In addition to this R-2 motif group, sequences which contain arginine not at -2 but at -3 or -10 positions are also detected in the yeast data set [9]. I confirmed that one of the R-10 motif sequences is recognized through its R-2 motif, yeast malate dehydrogenase [19]; however, MPP cleavage site of malate dehydrogenase was not detected by Vögtle *et al* [9]. Thus, these two groups might include more latent R-2 motif sequences. Also, MPP can recognize these un-canonical motifs other than R-2 [20]. This arginine enrichment near cleavage site was confirmed by chi square analysis as well; thus, arginine at -3, -2 and -1 positions were significantly overrepresented ($P \ll 0.05$) in the yeast.

However, +7 position also contains significantly numerous arginine (P=0.00001528), and reason for this is unknown. In the plant, a similar tendency regarding arginine enrichment was observed. At position -4, -3 and -2 arginine frequencies were significantly enriched, and position -6 as well ($P \ll 0.05$).

The number of sequences which do not contain R-2 motifs is not few, especially in the yeast data set, and I named this group "Others" (Figure 2.4). Since the research to clarify the specificity of cleavage by inner membrane proteases such as m-AAA or Pcp1 has recently started, mechanisms related to their cleavage site recognition are ambiguous at present [21]. All of proteins which were confirmed to be cleaved by m-AAA, i-AAA and IMP belong to the Others group; this diversity of proteases may explain why no pattern is evident in this group (Figure 2.5). Another possibility is that yeast dataset includes experimental errors. As an example, ATP6 is annotated as MTS containing protein; however, ATP6 is coded in mitochondrial genome. Thus, ATP6 dose not have to include MTS for transition. Although ATP6 is cleaved at position 10 as Vögtle *et al.* reported, this sequence is not MTS but so-called 'propeptide', another kind of cleavage before maturation. At present, it is obvious that annotations for unknown cleavage sites are in need for improvement. In this study, I did not use the Others group when training the HMM because it seems unlikely that they represent MPP cleavage sites. I removed four Icp55 and one Oct1 processed proteins because of their lacking arginine at -2 position. Although those five proteins have been experimentally confirmed their cleavage sites and related proteases, the number of those is too small to train profile HMM.

2.1.3 Architecture of profile Hidden Markov Model

Predicting cleavage site in the matrix uses a profile HMM as implemented by the HMMER 2 package (`http://hmmer.janelia.org/`). The three profiles were trained to discriminate MPP only,

MPP+Icp55, and MPP+Oct1 sites using ten residues, two residues, and four residues around the yeast cleavage site respectively and are visualized by HMMEditor in Figure 2.6 [22]. Window size for MPP profile was determined by cross validation (2.1).

Table 2.1: Window size and accuracies for cleavage site prediction with the size. Negative values indicate left border of the window and positive values for the right border in center of cleavage site.

Yeast	+1	+2	+3	+4	+5	+6	+7
-2	0.433	0.683	0.731	0.740	0.692	0.692	0.702
-3	0.587	0.740	0.779	0.760	0.779	0.760	0.740
-4	0.596	0.740	0.779	0.769	0.788	**0.798**	0.760
-5	0.587	0.769	0.779	0.769	0.760	0.769	0.769
-6	0.567	0.740	0.760	0.740	0.779	0.750	0.750
-7	0.567	0.712	0.721	0.712	0.712	0.731	0.740
Plant	+1	+2	+3	+4	+5	+6	+7
-3	0.333	0.457	0.580	0.741	0.765	0.728	0.765
-4	0.395	0.519	0.679	0.679	0.753	**0.778**	0.741
-5	0.432	0.494	0.704	0.691	0.741	0.741	0.728
-6	0.432	0.593	0.667	0.667	0.704	0.741	0.679
-7	0.420	0.469	0.593	0.630	0.679	0.753	0.741

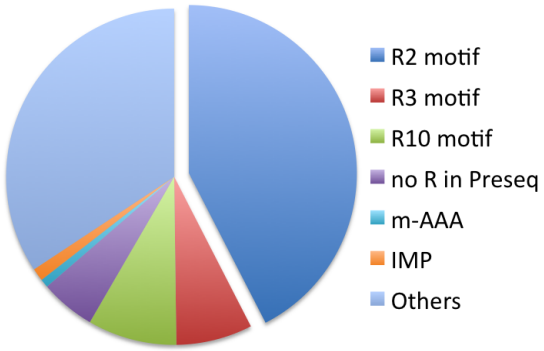

Figure 2.4: The diversity of yeast cleavage site.

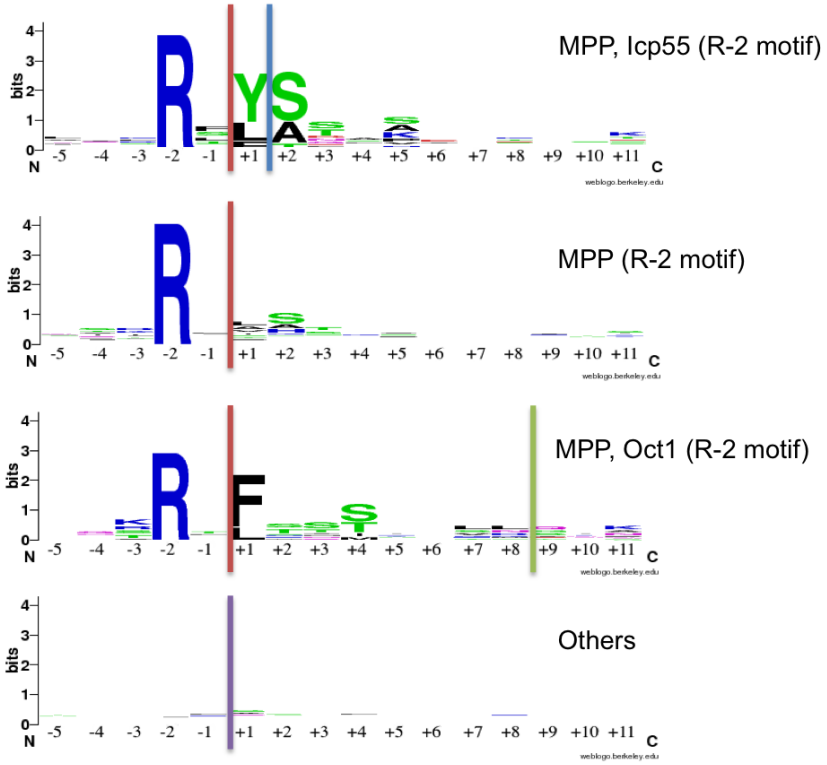

Figure 2.5: Sequence logos generated from the yeast data.

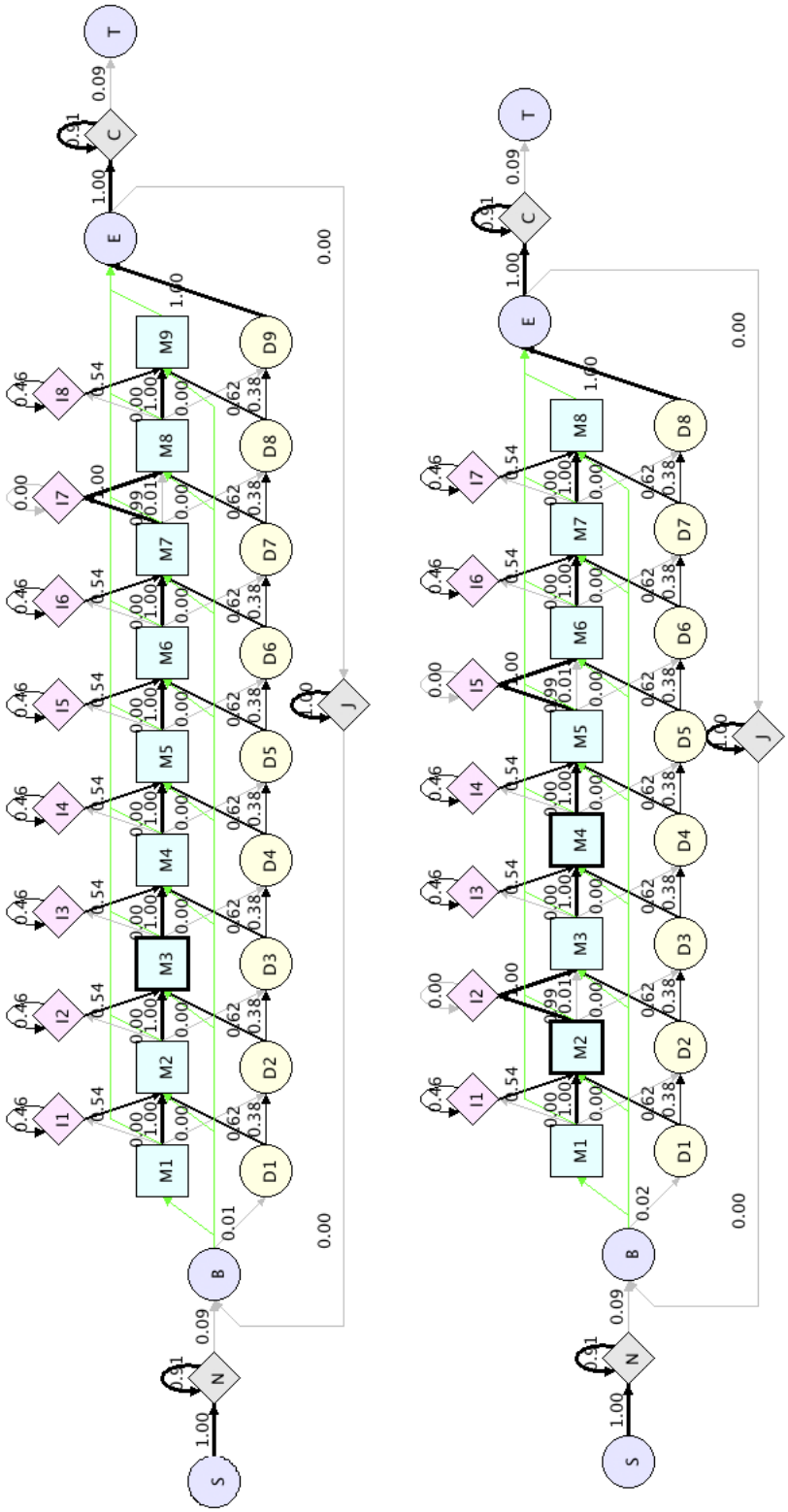

Figure 2.6: Architecture of profile HMMs. Top: yeast profile HMM for MPP, Bottom: plant profile HMM for MPP.

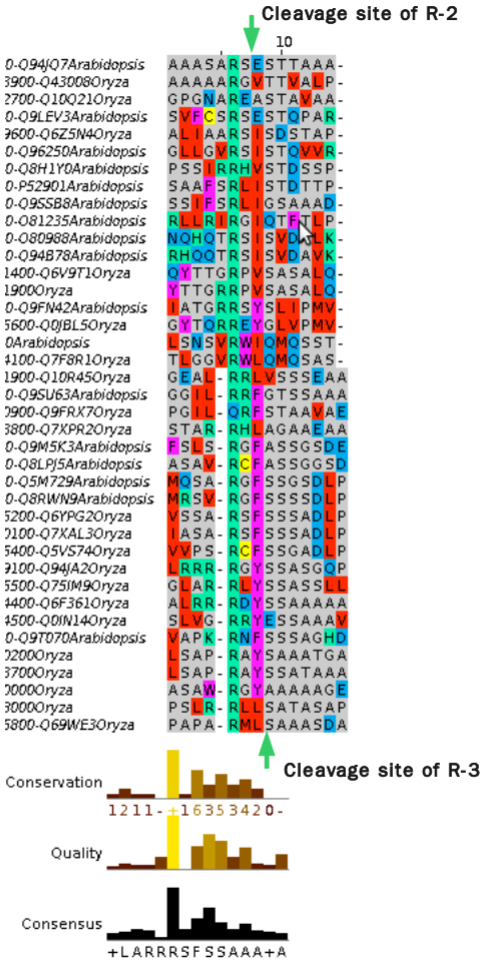

Figure 2.7: Alignment of R-2 and R-3.

These profiles were used to predict true cleavage site of the yeast and related proteases as our flow chart shows (Figure 1.3). Although the plant data set does not have annotation for intermediate proteases, two profiles similar to MPP only and MPP+Icp55 profiles for the yeast were trained to discriminate R-2 and R-3 using a multiple alignment (Figure 2.7) between these two motifs by MAFFT [23, 24]. Residues at the +1 position of the R-2 motif and -1 position of the R-3 motif

Table 2.2: Parameters of presequence length.

	mean length	min.	max.
Yeast	27.51 ± 15.68	7	86
Plant	43.51 ± 26.71	18	117

tend to be large and hydrophobic residues such as phenylalanine, leucine or isoleucine. We trained a profile HMM for MPP on the multiple alignment and one for a hypothesized counterpart of Icp55 in plants, on the R-3 group alone.

2.1.4 Hypothesized homologs of Icp55

In the plant dataset, a large fraction of the mature N-terminal sequences start three residues downstream of an arginine (the R-3 case), moreover in these sequences the residue immediately preceding the start of the mature N-terminal is non-random, thus it is natural to hypothesis that they may be the product of an additional cleavage of one residue by an Icp55 homolog after an initial MPP cleavage two residues downstream from the arginine. Conserved large and hydrophobic residues at position -1 of the R-3 motif proteins are argued as prokaryotic destabilizing residues by Vögtle *et al.* (Figure 2.8). Although conserved methionine at -1 position was not referred as a destabilizing residue [9], methionine was recently added to secondary destabilizing category as a novel N-degron [25]. Methionine does not function as a destabilizing amino acid in the eukaryotic N-end rule; therefore, the prokaryotic N-end rule seems to hold in plant mitochondria. Thus, hypothesized Icp55 homolog seems to cleave one destabilizing residue from N-terminal after MPP cleavage to make proteins stable. Blast search results in two homologs of Icp55 in *Arabidopsis thaliana*: At1g09300 and At4g29490. At1g09300 is annotated as a mitochondria localized protein and At4g29490 as a chloroplast protein. Interestingly, N-end rule-like degradation system in chloroplast was recently reported [26]. On the other hand, no homolog of Oct1 in Arabidopsis is annotated as a mitochondrial protein, and this is supported by absence of R-10 motif in plant, which is a typical motif for Oct1

Figure 2.8: Sequence logo generated from plant proteins whose cleavage site contain R-3 motif.

Table 2.3: Frequencies of amino acid combination nearby cleavage site in the plant dataset. D stands for destabilizing amino acid and S stands for stabilizing amino acid.

Type	D-D	D-S	S-D	S-S
R-3	0	49	0	2
R-2	1	6	0	23

in the yeast and mammals.

2.1.5 Analysis and distribution approximation for presequence length

The length of mitochondrial presequence was analyzed and summarized in Table 2.2. As Figure 2.9 shows, presequence length is not a uniform distribution in either the yeast nor the plant data set. Since distribution of presequence length can be a good feature, we validated two theoretical distributions, Gaussian distribution and Gamma mixture distribution, with their estimated parameters by goodness of fit test (Kolmogorov-Smirnov test). A 1-component Gamma distribution fits the yeast data the best, and a 3-component Gamma distribution fits the plant data set (Table 2.4). Although the 2-component Gamma distribution looks to fit the yeast histogram as well, the right-side component of bimodal distribution contains only four proteins. In this study, the unimodal distribution was applied as a theoretical distribution against presequence length of the yeast because of P-value and the small number of proteins in the right component. Applying the 3-component distribution of the plant presequence length to cleavage prediction is still in progress.

Table 2.4: Goodness of fit test.

	Distribution	P-value
Yeast	Gaussian	0.23
	Gamma 1-component	0.88
	Gamma 2-component	0.85
	Gamma 3-component	0.78
Plant	Gaussian	0.0
	Gamma 1-component	0.02
	Gamma 2-component	0.84
	Gamma 3-component	0.98

2.1.6 MTS cleavage site prediction

The prediction flow was precisely explained in the method section. Table 2.5, 2.6 show comparisons of cleavage site prediction of MoiraiSP and TargetP. Because MPP cleavage prediction of the yeast is new to this field, there is no comparisons about accuracy of MPP site prediction. The accuracy is relatively reliable comparing to that of final cleavage site prediction.

For reference, MCC of MPP prediction and that of cleavage site prediction by TargetP were plotted (Figure 2.10). Leaps of MCC as regards TargetP appeared at [0,1] and [7,8] in the yeast and [0,1] and [6,7] in the plant.

Discrimination between cleaved and non-cleaved proteins was measured by AUC of ROC curve: Yeast ROC AUC is 0.89 and plant AUC is 0.95. TargetP and MitoProt II predict all of proteins which are predicted to localize in mitochondria as cleavable, therefore, this aspect of MoiraiSP is also a novel.

At last, MoiraiSP also showed better performance in predicting final cleavage site including Oct1 site and Icp55 site against TargetP (Table 2.5, 2.6).

2.2 Dataset and Methods

2.2.1 Matthews correlation coefficient

The Matthews correlation coefficient, MCC [27], is a measure of performance for binary classification defined as follows:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{2.2}$$

where 'T' and 'F' stand for "true" and "false", while "N" and "P" stand for "negative" and "positive". Equivalently MCC can be defined as the Pearson's correlation coefficient of the binary vector of class labels compared to the binary vector of predicted class labels. MCC ranges from 1.0 for perfect prediction to -1.0 for perfect inverse prediction. Note that the MCC for the majority class classifier is identically zero, as is the expected value of MCC for random prediction.

2.2.2 Yeast data set construction

All data were extracted from the proteomic analysis experiments for *S.cere.*[9]. Vögtle *et al.* confirmed cleavage sites of Icp55 and Oct1 by knocking out of their genes and detecting the differences of protein cleavage between wild type and the mutants, and annotated them [9]. By their experiments, the authors discovered Icp55 cleaves one amino acid from the N-terminus when N-terminal residue is a destabilizing residue (leucine, lysine, tyrosine, phenylalanine, arginine and tryptophan) by the prokaryotic N-end rule. Without Icp55, substrates of Icp55 were degraded more rapidly. As a result of Icp55 cleavage, R-2 motif becomes R-3.

Vögtle *et al.* provides a list of mitochondrial proteins with their cleavage positions as supplemental data. From the data, we prepared two data sets; namely, cleaved and non-cleaved data sets. Because the data does not contain the full length of each protein, we extracted the sequence data

from SWISS-PROT release 57.9 using the ORF names in the paper. To reduce redundancy in the dataset, blastclust (unpublished, `http://www.ncbi.nlm.nih.gov/BLAST/docs/blastclust.html`) in BLAST 2.2.24 package was used. After removal of redundant sequences, suspicious entries were excluded from the data sets, namely, overlapping sequences between cleaved and non-cleaved data set or annotated signal sequences in the SWISS-PROT database. Although the original proteomic data shows multiple cleavage sites against a protein in many cases, only one cleavage site from a protein was extracted by its observed frequency to reduce experimental errors unless its multiple sites are annotated such as Icp55 cleavage. To avoid errors as much as possible when training HMM, we did not include proteins whose first cleavage site does not contain arginine at -2 position. As a result, the yeast data set contains 59 as annotated MPP only, 33 as MPP+Icp55, 12 as MPP+Oct1, and 110 as non-cleaved proteins. The point should be addressed here is that MPP+Icp55 and MPP+Oct1 are part of the groups which used to be called R-3 or R-10. Although almost R-3 and R-10 were proved to be cleaved twice (once by MPP and once by Icp55 or Oct1) by Vögtle *et al.*, there are still R-3 and R-10 motif proteins which have not been proven to be double digestion proteins.

2.2.3 Plant data set construction

Similarly to the yeast data set, all data were extracted from the proteomic analysis experiments for *Arabidopsis thaliana* and *Oryza satia* [7]. This proteomic data set includes full length sequences and Ordered Locus Names of each sequences. To gain general information about the sequences, I extracted annotations from SWISS-PROT 57.9 as OLNs are queries. At this step, OLNs of *Oryza sativa* were converted from MSU's locus ID to locus ID used in SWISS-PROT. Conversion was conducted on the table of RAP-DB(`http://rapdb.dna.affrc.go.jp/`) to make them match with IDs in SWISS-PROT. Some of sequences in the data set did not perfectly match with those of SWISS-PROT, and sequences in the data set were used for this research in these cases. Redundant

and suspicious entries were removed by the same flow for the yeast data set. Because Huang et al reported unique cleavage sites for each proteins, the unique sites were applied to train a predictor. While the data set contains several kinds of pattern around cleavage sites, we did not include proteins whose cleavage site does not contain arginine at -2 position or at -3 position to reduce errors in finding cleavage site motif. At last, the plant data set includes 81 MPP processing and 37 non-cleaved proteins.

2.2.4 Training and test set

Although the proteomic experiments identified many mitochondrial proteins, the number of the proteins is not high enough to separate the data set into independent test and training sets. Therefore, all results reported are based on ten-fold cross-validation. For the yeast prediction, the training and test data were separately partitioned in MPP only, MPP+Icp55, MPP+Oct1, and non-cleaved proteins based on the experimental results of Vögtle *et al.* and then randomly assigned folds for cross-validation. Since the plant data set does not have multiple cleavage site annotation, the training and test data were simply randomly divided into ten folds.

2.2.5 Profile HMM training

To train a profile HMM of MPP cleavage site with the HMMER implementation, residues within [-5,5] in center of cleavage site were used for MPP model training. As Figure 1.2 shows, MPP processing is followed by Oct1 or Icp55 processing when it is necessary. Due to this, training of MPP cleavage was conducted by cleavage sites of only MPP processing proteins and first cleavage sites of MPP+Icp55 and MPP+Oct1 annotated proteins. The other profiles for Icp55 or Oct1 cleavage site were trained on each training set. Icp55 training used residues within [+1,+2], and Oct1 used residues in [+1,+4]. We chose these intervals based on the results of previous work [9, 28, 1]. The HMMER2 package assigns similarity scores using logarithmic odds which is the log ratio of the probability

given by trained model to probability given by background model. To obtain better scores against query sequences, amino acid composition of mature mitochondrial protein and extension probability 10/11 was used as background model. Although by default HMMER2 uses 9-component Dirichlet mixture distribution for the prior distribution of amino acid composition at each columns, I use a 20-component mixture instead (`http://compbio.soe.ucsc.edu/dirichlets/`). Effective number of sequences is calculated to train model from multiple alignment by HMMER, however I turned off this option for training the Icp55 sub-filter. The purpose for calculation of effective number is to avoid overestimation of amino acid frequency at each position, which are used to compute *mean posterior estimates*. Because length of training data for Icp55 is only two residue long, almost sequences have close similarities to each other; as a result, HMMER determines too low effective sequence number. This results in too low frequency of appeared amino acids; thus, leading to abbreviated estimations for amino acid composition for Icp55. Effective sequence number option was applied to train MPP and Oct1 profiles.

2.2.6 Distribution parameters estimation

Gaussian distribution and Gamma distribution were applied to fit the histogram of presequence length in both yeast and plant data set. Bins of the histogram was determined by following formula:

$$\frac{Max - Min}{10} \tag{2.3}$$

Parameters of gaussian distribution were determined by simply calculating mean and standard deviation from the data. The shape and scale of Gamma distribution were estimated by using EM algorithm implemented in a package of the R [29, 30]. With these estimated parameters, discrepancies between the data and theoretical distribution was measured by Kolmogorov-Smirnov test as P-values. The best fit theoretical distribution was used as an attribute in the cleavage site

prediction.

2.2.7 Cleavage site prediction and its validation

Figure 1.3 is a flow chart of our prediction. Yeast and plant cleavage site predictions extract different length of N-terminal residues from queries, and the former uses 100 and the latter 130 at maximum due to each longest presequences. As the figure shows, MoiraiSP predicts MPP cleavage site first. The prediction uses a sliding window of 10 residues. The logarithmic probability of density function of Gamma distribution is added to the log odds score, using hmmpfam in the HMMER2 package, based on the window position. Since probability of density function equals to zero for a specific point, probability of density function for bin of position (Ref. 2.3) was calculated. This computation assumes the length of the presenquence is independent of the local context of cleavage sites.

The weighted scores are stored in an array and sites whose score is highest is predicted as MPP cleavage site. If the score of MPP site is higher than a threshold determined by comparison between cleaved and non-cleaved data set, it is predicted to be cleaved by MPP. In that case, the next step is to predict whether or not it is cleaved by intermediate peptidases. Otherwise, the site is predicted as a non-cleavable site. The intermediate peptidase cleavage prediction has two filters, Oct1 and Icp55. The Icp55 motif [YFL][STA] partially overlaps with the Oct1 motif [FLI]xx[ST], so the order of tests in our decision tree-like flowchart (Figure 1.3) can affect the results. We found that testing for Oct1 first works better. I trained the Oct1 and Icp55 profile-HMM's and used a one-versus-rest procedure to determine appropriate log odds score thresholds for each test. Plant has similar flow to yeast system; however the second step only includes a hypothesized analog to Icp55 (i.e. R-3) with no analog for Oct1.

Validation was conducted by comparison with TargetP. MoiraiSP does not predict whether or not a query localize in mitochondria as TargetP or MitoProt II. Therefore, result of the cleavage

prediction was compared only if TargetP predicted queries localize in the mitochondria. Also, MCC at residue level was used to compare the MPP prediction comparison with the result of TargetP.

2.3 Discussion

2.3.1 Limitation behind cleavage site prediction

As shown in Figure 2.10, leaps are observed in the prediction error (distance from true final cleavage site) of TargetP, which was trained on both yeast and plant proteins, thus leading to diverse but non-specific cleavage site prediction. The leap between 0 and 1 in both the yeast and plant data set should be based on mis prediction between R-2 and R-3 motif in the TargetP model, and leap occurred in [7,8] of the yeast or [6,7] of the plant seem to be affected by incorrect prediction between R-2 and R-10 or R-3 and R-10 motifs. The number of R-3 motif proteins in the plant data set is more numerous than that of R-2 ; this explains why the leap occurs between 6 and 7 rather than 7 and 8. My model can avoid this kind of error by combining three classes into one R-2 class at first, and this seems the reason of high MPP prediction result. In fact, R-2 motif proteins were sometimes mis predicted as R-3 in the plant data set as well. To avoid this bias in the plant data, alignment and combining of R-2 and R-3 motifs was conducted in the plant dataset. As a result, the accuracy of MPP prediction ranges from 75% of the plant to 78% of the yeast, and these scores are relatively reliable. However, plant system is still in progress; thus, improvement for the plant MPP prediction is likely to be obtained in the future.

2.3.2 Limitation for plant evaluation

Also, I have training data for the intermediate proteases Icp55 and Oct1 for the yeast data, but no training data for the hypothesized analog of Icp55 in the plant data. As a result, for the yeast data I can score the performance on MPP alone, but the plant performance measures correct prediction

of the R-2 and R-3 class combined. Even though existence of this limitation, plant cleavage site prediction also results in relatively reliable performance.

2.3.3 Icp55 homolog in plant

Interestingly, plant cleavage site which has arginine only at -2 position favors stabilizing residue such as isoleucine, glutamate or threonine. On the other hand, cleavage site contains arginine only at -3 position is likely to be cleaved between destabilizing and stabilizing residues as Icp55 does. The observation supports the hypothesis that plants possess an Icp55 analog. Indeed, Arabidopsis has At1G09300, a homolog of Icp55, and the protein shows the same domain structure as Icp55. Since neither of two homologs have not been confirmed their existence at protein level, I need to wait result of research in the future for further discussion.

Nevertheless, N-end rule like degradation system in chloroplast was recently reported [26], and the second most similar protein in Arabidopsis, At4g29490, is annotated as chloroplast protein. Setting true function of At1g09300 aside, N-end rule like degradation system is likely to exist in mitochondria as well.

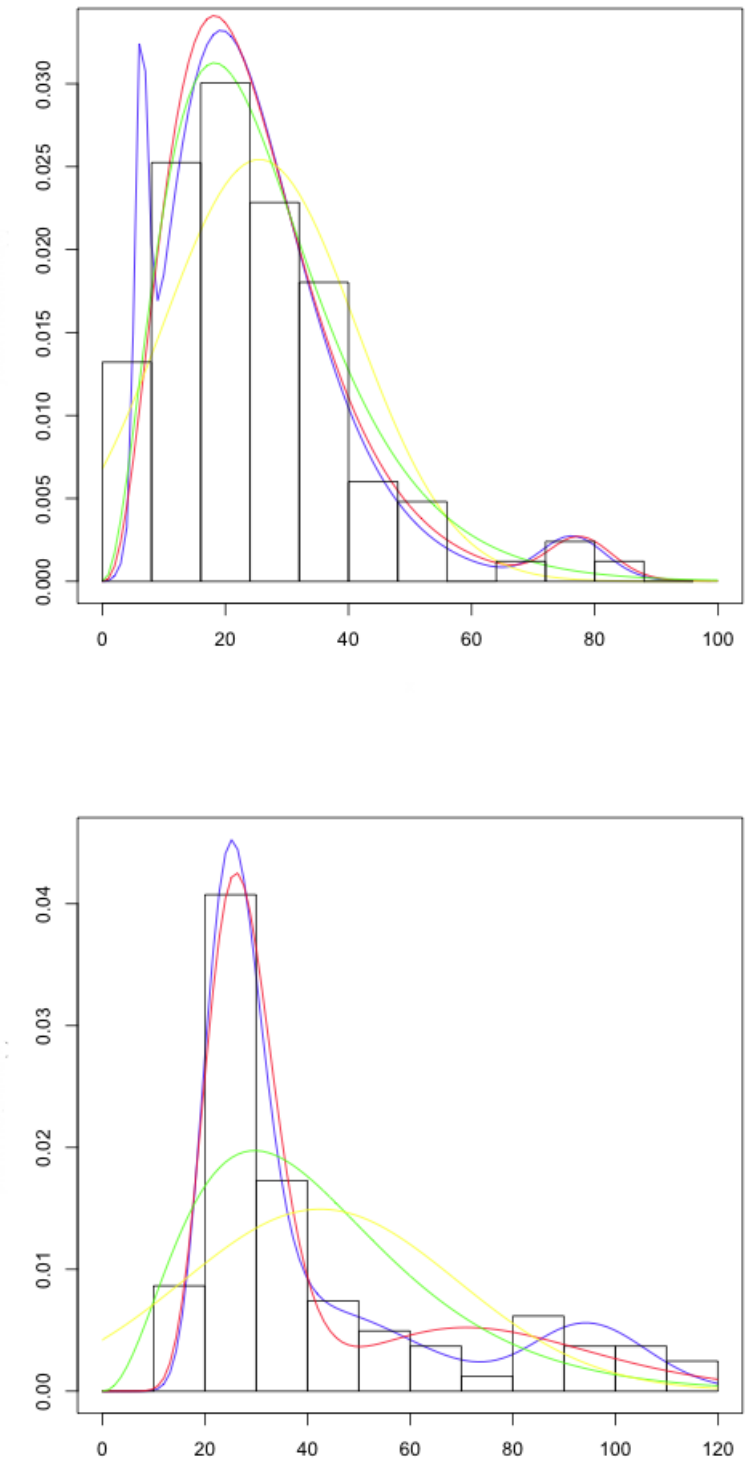

Figure 2.9: Distributions of the yeast and the plant. Up: Yeast, Down: Plant. Yellow line shows Gaussian distribution, green is Gamma unimodal, red is Gamma bimodal and blue is Gamma trimodal.

Table 2.5: Comparison of final cleavage site prediction with TargetP in the yeast dataset. Denominators show predicted numbers by TargetP if proteins contain MTS or not, and numerators indicate predicted number of cleavable proteins by TargetP in TargetP column or by MoiraiSP in MoiraiSP column.

	Predicted MTS-containing cleavable by TargetP					Predicted Non-MTS (non-cleavable) by TargetP	
	TargetP		MoiraiSP			TargetP	MoiraiSP
Noncleaved(110)	33/33		6/33			0/77	4/77
	Correct location	Incorrect location	Correct location	Incorrect location	Predicted non-cleavable		
Cleaved(104)	49/91	42/91	61/91	27/91	3/91	0/13	7/13

Table 2.6: Comparison of final cleavage site prediction with TargetP in the plant data set. Denominators and numerators mean the same as Table 2.5

	Predicted MTS-containing cleavable by TargetP					Predicted Non-MTS (non-cleavable) by TargetP	
	TargetP		MoiraiSP			TargetP	MoiraiSP
Noncleaved(37)	13/13		2/13			0/24	0/24
	Correct location	Incorrect location	Correct location	Incorrect location	Predicted non-cleavable		
Cleaved(81)	32/69	37/69	48/69	18/69	3/69	0/12	11/12

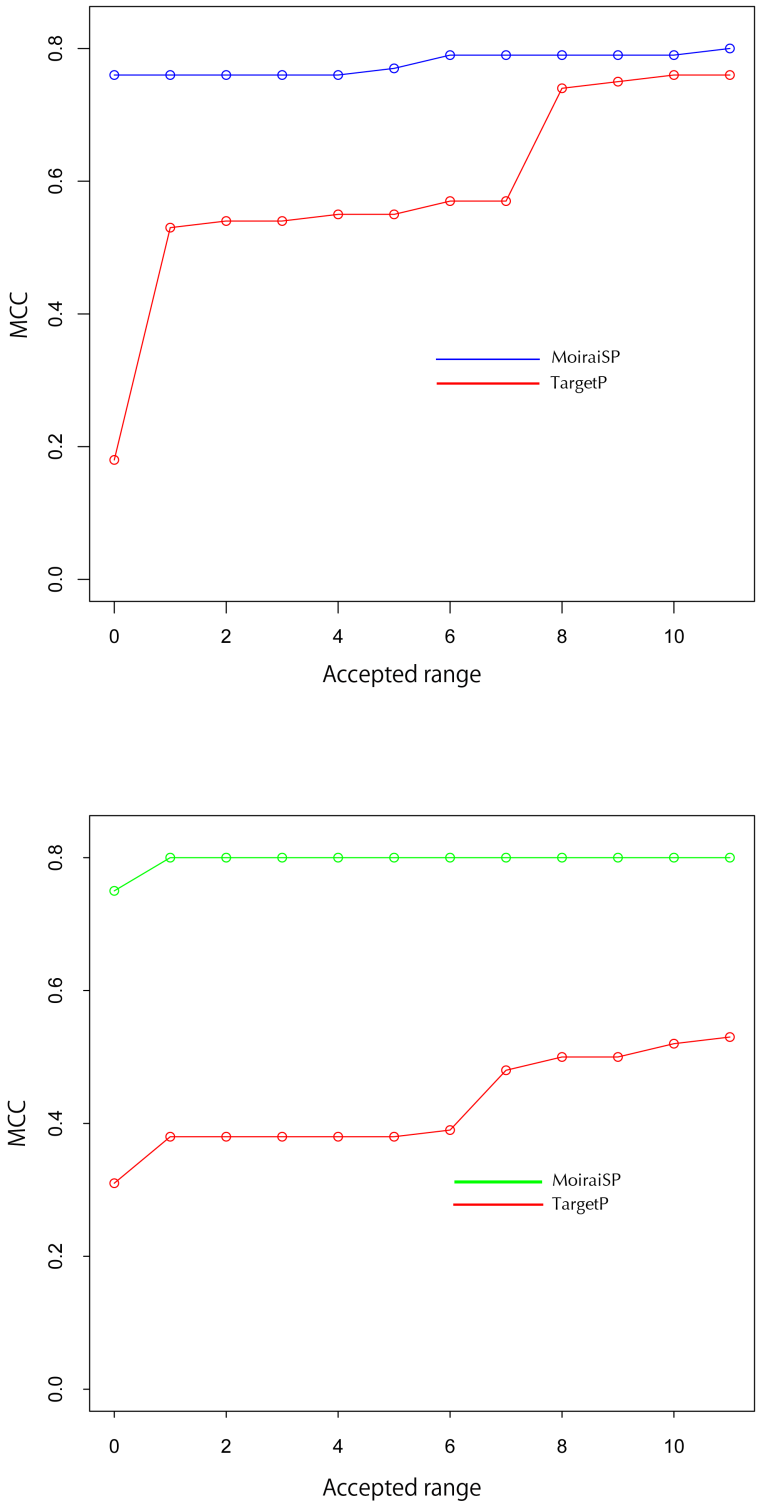

Figure 2.10: MPP prediction compared with TargetP. Up: Yeast MCC of MPP prediction, Down: Plant.

Chapter 3

Divergence as a novel feature

In this section I describe work to be presented at ISB 2011.

3.1 Sequence divergence in presequence

As table 2.2 shows, presequences length of mitochondrial proteins are highly divergent. In fact, there is not even consensus pattern in this region. Although some consensus motif has been reported for mitochondrial targeting signals [31, 32], it is information poor and produces too many false positives to be used for reliable prediction. Figure 3.1 is a typical example of the mitochondrial presequence. Clearly, presequence region has numerous gaps and infers higher mutation rate. To confirm whether or not this divergence can be a novel feature to predict subcellular localization of proteins, a simple experiment was conducted. This chapter describes the experiment and discusses its effect for the prediction.

3.2 Dataset

3.2.1 Proteins and their localization classes

This study focuses on the prediction of N-terminal sorting signals in the budding yeast *S.cere.*– the eukaryotic organism with the most complete annotation available regarding protein sub-cellular localization. It was focused on that the two most common N-terminal sorting signals, the "signal peptide" (which we abbreviate as "SP"), targeting proteins to the endoplasmic reticulum and the "MTS" (Matrix Targeting Signal) which targets proteins to the matrix (inner compartment) of the mitochondria. Although both of these signals reside near the N-terminus, they are though to be mutually exclusive, with different properties that are effectively discriminated by the cell. Although other types of N-terminal sorting signals exist, for example the PTS2 signal targeting the proteins to the peroxisome [33], the number of proteins using such signals is much smaller than those using the SP or MTS signals.

In this study we choose to leave these less common signals to future work and instead concentrate on three broad localization classes for proteins in *S.cere.*: 1) with SP's, 2) with MTS's, and 3) N-signal-less; of which we gathered 54, 182, and 462 examples respectively. We used UniprotKB/Swiss-Prot [34] to assign localization class labels, augmented by MTS containing proteins determined in the proteomics experiment of Vöglte et al. [9]. Because only a small number of SP's have been directly confirmed experimentally, we also included SP proteins whose label is "by similarity" or "potential" by curators of UniprotKB/Swiss-Prot. To reduce false positives, suspicious proteins were filtered out by prediction by SignalP [35] (see Discussion for a justification of using prediction results in our dataset). For N-signal-less proteins we used proteins which localize to the cytosol or nucleus (according to UniprotKB/Swiss-Prot annotation).

To avoid a bias in training and accuracy estimation, we used Blastclust 2.2.22 (`http://www.ncbinlmnih.gov/BLAST/`) to removed redundant sequences with a setting of 20% identity.

3.2.2 Orthologs and multiple alignment

We extracted orthologs from the Yeast Genome Order Browser [36]. YGOB includes curated ortholog sets from 11 fungi genomes (*S.cere.*, *S. castellii*, *S. kluyveri*, *K. waltii*, *A. gossypii*, *C.glabrata*, *K. lactis*, *Z. rouxii*, *K. thermotolerans*, *S. bayanus* and *K. polysporus*) by synteny of their genomes. For each *S.cere.* protein in our dataset, we obtained its *ortholog multiple sequence alignment* (orthoMSA) by aligning it to its orthologs with the MAFFT program [23]. We ran MAFFT using "LINSI", its most accurate mode.

3.3 Features for classification

3.3.1 Sequence evolutionary divergence score

Our study required assigning a divergence score to each position of each *S.cere.* protein, based on its orthoMSA.

Column entropy score

Several measures have been suggested for scoring evolutionary sequence conservation (or conversely divergence) [37, 38]. Here we adopt a simple Shannon entropy based score:

$$H(i) = -\sum_{j \in A} F(i,j) \lg F(i,j). \tag{3.1}$$

where i indicates position of a column and j stands for a kind of amino acid.

When multiple gaps are present in a column, we consider each to be a unique character instead of using gap penalty. For example, the entropy of an orthoMSA column '{`L`, `L`, `I`, `-`, `-`}' is computed as one character (the '`L`') with frequency 0.4 and three characters with frequency 0.2. We adopted this treatment of gap symbols so that the divergence of orthoMSA columns with many gaps would

Table 3.1: Smoothed entropy derived features are listed. Quantities shaded in grey were not used directly as features.

Feature name	Quantity
LD(i)	$\bar{H}_{i-10,i+10}$
$N_{\text{raw}}20$	$\bar{H}_{1,20}$
$N_{\text{raw}}40$	$\bar{H}_{1,40}$
$N_{\text{raw}}80\text{-}99$	$\bar{H}_{80,99}$
μ_w	Average of $\bar{H}_{\text{window}}$ for all length w windows
σ_w	Standard deviation of $\bar{H}_{\text{window}}$ for all length w windows
NCdiff	$N_{\text{raw}}20 - N_{\text{raw}}80\text{-}99$
$N20$	$\frac{(N_{\text{raw}}20-\mu_{20})}{\sigma_{20}}$ (z-score normalized)
$N40$	$\frac{(N_{\text{raw}}40-\mu_{40})}{\sigma_{40}}$ (z-score normalized)
$N80\text{-}99$	$\frac{(N_{\text{raw}}80\text{-}99-\mu_{20})}{\sigma_{20}}$ (z-score normalized)

be considered high. Since we use 11 species, the range of our column divergence score runs from 0 (perfect conservation) to 3.46 bits (maximally diverged).

Smoothed entropy score

For many orthoMSA's, the entropy often varies widely from column to column, therefore as a measure of divergence, we adopted a smoothed entropy score, $\bar{H}_{i,j}$, defined as the average entropy score for columns in the interval $[i, j]$.

Divergence based features

We employed several smoothed entropy score based features summarized in table 3.1.

3.3.2 Physico-chemical propensities

To explore the possibility of combining sequence divergence with standard features used in protein localization prediction, we defined three features computed from the first 30 N-terminal residues of each *S.cere.* protein: 1) the number of positively charged residues (#pos), 2) the number of negatively

charged residues (#neg), and 3) the average hydrophobicity as measured by the Kyte-Doolittle [39] index (Hphob).

3.4 Classifiers

3.4.1 Majority Class Classifier

The majority class classifier unconditionally predicts all examples to belong to the most common class. Its accuracy is equal to the fraction of examples belonging to the most common class.

3.4.2 J48

J48 is a version of the C4.5 decision tree induction algorithm of Quinlan [40], implemented in the Weka software package [41]. We used the default value of 0.25 for the confidence factor, which controls the complexity of the induced tree.

3.4.3 Support Vector Machine

The SVM [42] is perhaps the most popular classifier in current bioinformatics work. In its basic form it is a linear, binary classifier, but it has been extended to non-linear, multiclass classification. In this project, we used the LIBSVM implementation [43]. We used the Gaussian radial basis kernel function with default γ value (1.0 / # number of features). We also used the default value (1.0) for the SVM cost parameter C. In our study we conducted binary and 3-class classification. For multiclass discrimination LIBSVM adopts the "one-versus-one" method, in which a separate SVM is learned for each pair of classes, and majority voting amongst those SVM's is used when classifying examples.

3.4.4 Quantifying feature importance

We used the so called "F-score" to quantify the importance of each features. The F-score [44] is a simple measure of the predictive power of a feature in isolation (i.e. without consideration of its relationship to other features), defined as:

$$\frac{(\bar{x}^{(+)} - \bar{x})^2 + (\bar{x}^{(-)} - \bar{x})^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_k^{(+)} - \bar{x}^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_k^{(-)} - \bar{x}^{(-)})^2} \tag{3.2}$$

where $\bar{x}^{(+)}$, $\bar{x}^{(-)}$, and $\bar{x}$ are the mean values of the feature for the positive, negative and combined examples respectively; while $x_k^{(+)}$ and $x_k^{(-)}$ denote the value of the kth positive and negative examples respectively. A larger F-score indicates greater predictive power.

3.4.5 Classification performance evaluation

Accuracy is not always an effective measure of performance for skewed datasets (i.e. datasets with a very uneven number of examples from different classes) [45]. Therefore we use MCC (Ref. 2.2) as a measure to quantify the performance in addition to ROC AUC.

3.5 Results

3.5.1 Feature Analysis

N-terminal sorting signals are evolutionary divergent

It is well known that sorting signals, especially signal peptides, have very low sequence conservation [46]. As shown in Figure 3.1, this phenomenon is particularly clear for the mitochondrial heat shock protein, SSC1, in which main part of the protein is highly conserved but the N-terminal region is highly divergent. Figure 3.2 quantifies this trend for the proteins in our dataset.

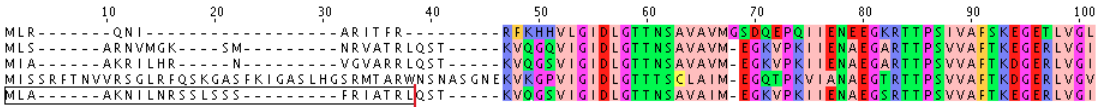

Figure 3.1: A multiple sequence alignment of the protein SSC1 (*S.cere.*Uniprot accession P12398) from five species of fungi. The red line shows the MPP cleavage site located at the end of the MTS. The conserved region is colored by Jalview.

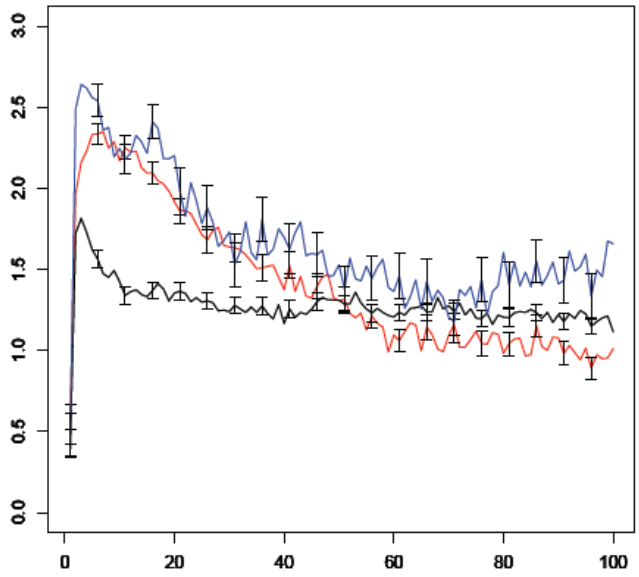

Figure 3.2: Divergence scores (entropy) are shown for the 100 residue N-terminal region for MTS containing (red), SP containing (blue), and N-signal-less (black) proteins. The error bars denote the standard error. For clarity, error bars are only shown for every fifth position.

Estimate of importance of each feature

As a rough estimate of feature importance, we computed the F-score for each feature (Figure 3.3). The two highest scoring features are the physico-chemical features #neg and Hphob, but the LD features near the N-terminus also show F-scores significantly greater than zero.

Sequence divergence is not redundant to physico-chemical trends

To be promising as a feature for prediction, it is desirable that evolutionary sequence diversity not be perfectly correlated with other useful features. To investigate this we plotted LD(13), the

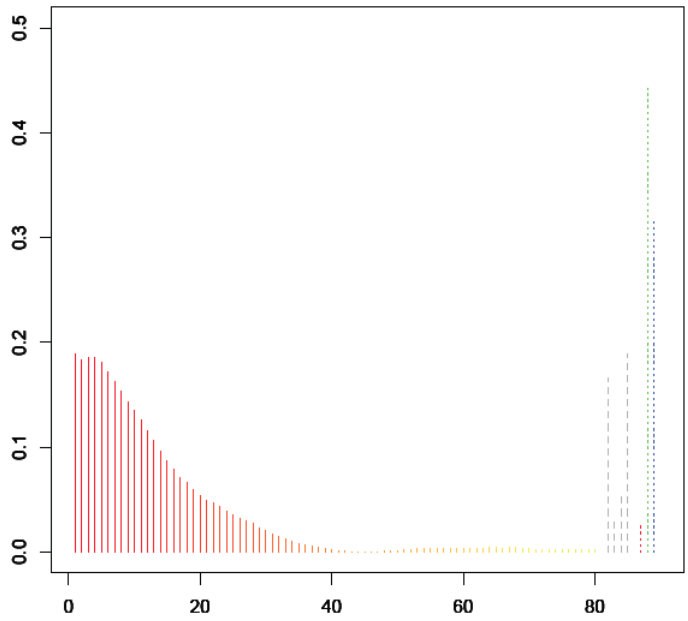

Figure 3.3: Importance of each attribute as estimated by F-score is shown. At left, the LD value for each position is shown by solid and heat colored lines. Gray dash lines denote $N20$, $N40$, $N80$-99 and NCdiff. Colored and dotted lines denote the N-terminal physico-chemical properties #pos, #neg and Hphob, respectively.

divergence feature with the highest F-score, against the two highest scoring physico-chemical features (Figure 3.4). Although it is difficult to discern the exact relationship, one can see that the feature pairs do not appear highly correlated.

3.5.2 Divergence predicts presence of N-terminal signal

We tested whether sequence divergence can be used to distinguish between proteins with an N-terminal localization signal (MTS or SP) and those with none. As shown in Table 3.2, for this binary classification task, sequence divergence *alone* allows for significantly higher prediction accuracy than randomized control experiments or the majority class fraction (66.2%).

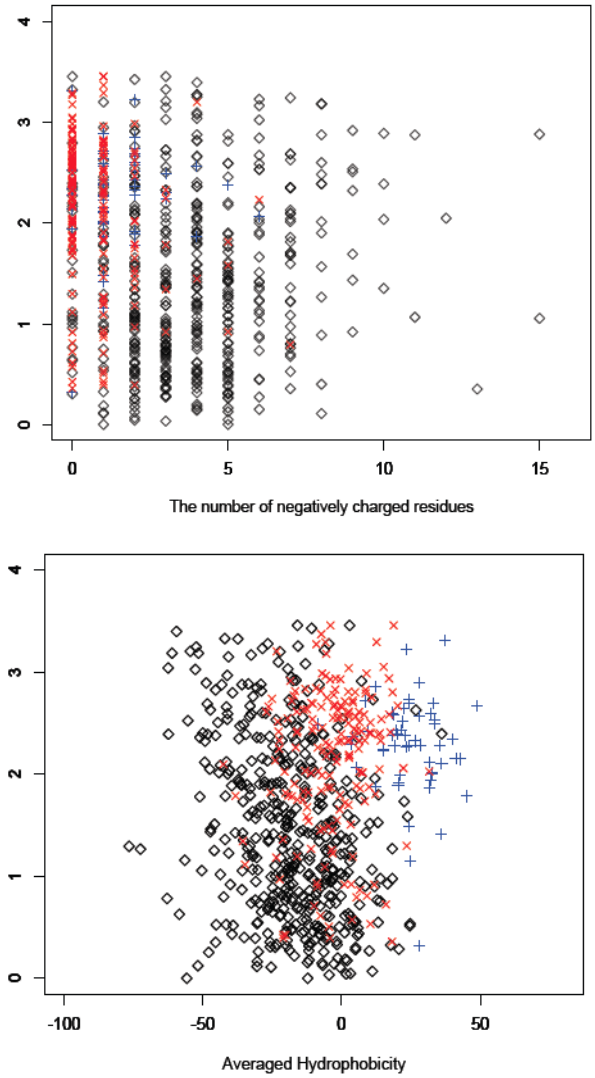

Figure 3.4: The scatter plot of LD(13) on the vertical axis *vs.* #neg (top) and Hphob (bottom) on the horizontal axis is shown. MTS, SP, and N-signal-less proteins are represented by red, blue and black dots, respectively.

3.5.3 Divergence distinguishes signal SP vs. MTS vs. N-signal-less

Although the sequence divergence profile of SP's and MTS's appear similar when averaged over proteins containing each signal (Figure 3.2), we found that sequence divergence is still somewhat effective for the three-way classification of SP *vs.* MTS *vs.* N-signal-free. As shown in Table 3.3 the

Table 3.2: Three classification performance measures are shown for the discrimination of N-signal containing and N-signal-less proteins. AUC denotes the area under the ROC curves. (randomized) indicates the values obtained with the localization class labels randomly shuffled 100 times. For each measure the average and standard deviation is shown over the 5 folds of the cross-validation, or 500 (5×100 trials) folds in the case of the randomized data.

	mean accuracy	mean AUC	mean MCC
J48	72.49 ± 3.30	$\mathbf{0.68} \pm 0.09$	$\mathbf{0.40} \pm 0.09$
- (randomized)	65.85 ± 0.66	0.50 ± 0.01	0.00 ± 0.03
SVM	$\mathbf{74.64} \pm 2.38$	$\mathbf{0.68} \pm 0.03$	$\mathbf{0.40} \pm 0.06$
- (randomized)	66.19 ± 0.09	0.50 ± 0.00	0.00 ± 0.01

Table 3.3: The 5-fold cross-validation performance of an SVM classifier using: divergence features only, physico-chemical features only, and the two combined; is shown for three-way classification on our entire dataset.

	Divergence		Physico-chemical features		Combination	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.65 ± 0.01	0.34 ± 0.03	0.81 ± 0.05	0.67 ± 0.09	$\mathbf{0.82} \pm 0.04$	$\mathbf{0.68} \pm 0.08$
SP	0.50 ± 0.00	0.00 ± 0.00	0.72 ± 0.05	0.57 ± 0.07	$\mathbf{0.86} \pm 0.06$	$\mathbf{0.72} \pm 0.08$
N-signal-free	0.64 ± 0.02	0.34 ± 0.06	0.79 ± 0.05	0.63 ± 0.10	$\mathbf{0.85} \pm 0.04$	$\mathbf{0.73} \pm 0.08$
% accuracy	71.06 ± 1.57		83.11 ± 3.44		$\mathbf{86.25} \pm 3.56$	

performance with divergence features is slightly better than the majority class fraction (66.2%) and also slightly improves the performance when added to the physico-chemical features.

The ratio of examples in our dataset is 8.56:3.37:1, for N-signal-less, MTS and SP containing proteins respectively. Skewed datasets are known to complicate both learning and performance evaluation [45]. Therefore we also measured performance on a dataset with uniform class occupancy, created by randomly discarding all but 54 proteins from each class. As shown in Table 3.4, in this experiment, classification based on the divergence features only, performance is much higher than the majority class fraction (0.33%); moreover the divergence features also contribute to the performance when combined with the physico-chemical features.

3.6 Discussion

First, this work must be considered as a proof of concept only with many limitations.

Table 3.4: The 5-fold cross-validation performance of an SVM classifier using: divergence features only, physico-chemical features only, and the two combined; is shown for three-way classification on a balanced dataset (54 proteins in each class).

	Profile		Physical features		Combination	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.65 ± 0.10	0.30 ± 0.19	$\mathbf{0.85} \pm 0.05$	$\mathbf{0.74} \pm 0.09$	0.81 ± 0.07	0.62 ± 0.12
SP	0.69 ± 0.05	0.40 ± 0.13	0.78 ± 0.09	0.59 ± 0.13	$\mathbf{0.90} \pm 0.04$	$\mathbf{0.82} \pm 0.07$
N-signal-less	0.73 ± 0.05	0.47 ± 0.11	0.77 ± 0.05	0.53 ± 0.10	$\mathbf{0.87} \pm 0.05$	$\mathbf{0.74} \pm 0.11$
% accuracy	58.56 ± 8.40		73.48 ± 4.41		$\mathbf{81.37} \pm 5.95$	

3.6.1 Measurement for evolutional divergence

Many sophisticated measures have been proposed to quantify the degree of sequence conservation [38]. Here we only present results using a simple entropy based measure which ignores the phylogenetic relationship of the species involved. For non-divergence features we used only three, reasonable but simple, physico-chemical based features. Since popular features such as amino acid composition were not tested, combination with other features should be conducted in the future.

3.6.2 Organisms and location defined for the prediction

We only evaluated our predictions on the well-studied fungi *S.cere.*. Although the mechanisms of sub-cellular localization are similar in principle in animals and plants (chloroplasts also import proteins via N-terminal signals), the details can be different [5, 47].

Although many predictors discriminate between 10 or more localization sites (e.g. WoLF PSORT [48]), we focused on only the two most common sorting signals.

3.6.3 Appropriateness of dataset

One weakness in this task, is that many of our SP proteins are not experimentally validated, but based only on prediction by sequence similarity and SignalP. This unfortunate circularity (predicting predictions) is unavoidable because: 1) only a handful of SP's have been experimentally verified,

and 2) the presence of SP's cannot be reliably inferred exclusively from localization site for most *S.cere.* proteins. It may be reasonable to assume that secreted proteins all have SP's, but *S.cere.* secretes very few proteins (the SWISS-PROT derived WoLF PSORT [48] dataset lists only six). Other SP containing proteins generally localize to the E.R. or Golgi body – but proteins annotated to localize to the E.R. or Golgi include non-SP containing proteins such as peripheral membrane proteins which localize to the outside of these organelles. In fact, proteins annotated as localizing to the E.R. or Goligi, but not SP containing did not contain hydrophobic region in their N-terminal; thus, they are unlikely to be SP containing proteins.

However, the risk of incorrect conclusion resulted from employing non-verified SP data is small. First, this problem only applies to the SP class, as recent proteomics data has provided direct measurement of many MTS's [9]. Second, given the intense study of *S.cere.* and the continued scrutiny of UniprotKB/Swiss-Prot by the research community, we find it unlikely that a large fraction of the SP proteins in our dataset are incorrectly labeled.

Chapter 4

Discrimination between MTS and non-MTS containing proteins

4.1 Factors related to import to the matrix

4.1.1 Positive charge is important for both matrix and MPP import

One of the main functions in mitochondria is oxidative phosphorylation; thus, leading to high density of proton outside the inner membrane and low density of proton in the matrix. To gain driving force from this membrane potential, N-terminal of mitochondrial proteins includes numerous positively charged residues. Additionally, a model was proposed that MPP also uses positive charge in N-terminal to import of their substrates into the cavity [49, 50]. Non-cleaved mitochondrial proteins such as outer membrane proteins do not need to contain positively charged N-terminal, because they do not penetrate inner membrane by electrically membrane potential. Therefore, charge in N-terminal can be a good feature to classify mitochondrial proteins into cleaved or non-cleaved

proteins.

4.1.2 Negatively charged residue in MTS region

It is known that amino acid composition of MTS region is different from mature region. For instance, MitoProtII defines that MTS region is upstream of continuous negatively charged residues [14]. In other words, at least two continuous negatively charged residues such as DD is regarded as an end of MTS in MitoProtII system. Because of low number of negatively charged and numerous positively charged residues, it is said that MTS can be enough positively charged to pass the inner membrane. In fact, negatively charged residue rarely appear in MTS regions; thus, even moderate number of negatively charged residue decreases a possibility of MTS.

4.1.3 Evolutional information

As I stated above, there is no consensus motif in the primary structure of MTS, which differs considerably even among orthologs [16]. In general, non-cleaved mitochondrial proteins includes not N-terminal but internal signal, which is poorly characterized at present [16]. Divergence in N-terminal signal can be a novel feature to classify proteins to N-terminal signal or N-signal-less proteins. Therefore, divergent scores in N-terminal, which is described in chapter 3, were also applied to this classification problem.

4.2 Features for classification

4.2.1 Log odds ratio of profile HMM

Profile HMM was trained on only ten residues around the cleavage sites; thus, the rest of the presequence sequence was not used. As I discuss in chapter 2, log odds ratio calculated by the

profile HMM alone can discriminates cleaved proteins and non-cleaved proteins. Therefore, HMM scores is given to the classifiers as an attribute.

As Probability for given length of presequence can be calculated by Gamma mixture with estimated parameters, calculated logarithmic probability for each length was added to the HMM score. Weighted HMM scores by Gamma mixture and non-weighted raw HMM scores are tested as independent models to each other.

4.2.2 Physico-chemical features

Since MTS region is positively charged, physico-chemical features (hereafter "F_{Phy}") are used to classify proteins into two categories. These features are defined by predicted cleavage site; namely, sequence from N-terminal to predicted cleavage position is used for this purpose. For example, the number of asparate from N-terminal to predicted cleavage position is given to a classifier. If the number of asparate is high, it is unlikely to be MTS, cleaved mitochondrial peptides.

4.2.3 Evolutional information

Basic idea is the same as LD(i) in chapter 3, namely, smoothed H(i).

$$H(i) = -\sum_{j \in A} F(i,j) \lg F(i,j). \tag{4.1}$$

where i indicates position of a column and j stands for a kind of amino acid.

The only difference is that LD(i) for this classification problem is normalized as a z-score. The number of orthologs varies from proteins to proteins; thus, score range must be normalized to be compared in this case.

Table 4.1: Features for the classification are listed. k is a index for (predicted) cleave position; thus, this equals to length of presequence. Quantities shaded in gray were not used directly as features. α is shape and β is scale parameters for gamma distribution. l is a bin to which position i belongs.

Feature name	Quantity
$\#_{\mathrm{D}}$	The number of D from N-terminal to kth position
$\#_{\mathrm{E}}$	The number of E from N-terminal to kth position
$\#_{\mathrm{H}}$	The number of H from N-terminal to kth position
$\#_{\mathrm{K}}$	The number of K from N-terminal to kth position
$\#_{\mathrm{R}}$	The number of R from N-terminal to kth position
$\#_{-}$	$\#_D + \#_E$
$\#_{+}$	$\#_R + \#_H + \#_K$
$\overline{Chg}_{net}$	$\frac{\#_+ - \#_-}{k}$
Sc_{hmm}	$log\frac{P(x\|model)}{P(x\|null)}$
$Sc_{hmm(distance)}$	$log\frac{P(x\|model)}{P(x\|null)} + log\{F(l;\alpha,\frac{1}{\beta}) - F(l-1;\alpha,\frac{1}{\beta})\}$
$\mathrm{LD}(i)$	$\bar{H}_{i-10,i+10}$
μ	Average of $\mathrm{LD}(i)$ $(1 \leq i \leq 80)$
σ	Standard deviation of $\mathrm{LD}(i)$ $(1 \leq i \leq 80)$
$\mathrm{LD}_{\mathrm{z}}(i)$	$\frac{\mathrm{LD}(i)-\mu}{\sigma}$ (z-score normalized)

4.3 Classifiers

4.3.1 Support Vector Machine

For classification of mitochondrial proteins, LIBSVM implementation was used [43]. I used the RBF kernel function with searched parameters γ value and the SVM cost parameter C. Grid search on the training data was conducted to determine γ and C values. To draw ROC (Receiver Operating Characteristic), probability was estimated by LIBSVM for each proteins [51].

4.3.2 Estimation for feature importance

F-score (Ref. 3.2) was used as a simple measurement of discriminative power.

4.3.3 Classification performance evaluation

MCC (Ref. 2.2) was used as a measure to quantify the performance in addition to AUC. Balanced accuracy (BAC) was also measured, because both classes have equal importance.

$$BAC = \frac{Sensitivity + Specificity}{2} \tag{4.2}$$

4.4 Dataset

The dataset was the same as the one which was described in the chapter 2. Divergent score in N-terminal was applied for only yeast dataset due to the limitation of the database.

4.4.1 Features for positive data

Since positive data set has cleavage positions for each proteins, F_{phy} were directly calculated from the cleavage position. Sc_{hmm} for SVM were calculated by leave-one-out (“loo”) training. Because

HMM score can be too high when training includes exact same sequence to test sequence, this score might disturb Sc_{hmm} distribution in SVM; therefore, loo training was conducted in the training dataset for Sc_{hmm}. For example, if kth-fold training data contains j sequences, j profile HMM models are built to calculate log odds scores for each j sequences.

4.4.2 Features for negative data

Non-cleaved data set lacks cleavage position, and at least one position should be defined to calculate F_{phy}. One approach to obtain F_{phy} is to define fixed length such as N-terminal 30 residues used in chapter 3. I used the non-cleaved proteins (extracted from papers of proteomic experiments [9, 7]) as negative data. More precisely, for each non-cleaved sequences, I computed the maximum HMM score in the 100 N-terminal residues in the yeast data and 120 residues in the plant as negative examples of cleavage sites.

4.5 Results

4.5.1 Non-cleaved proteins has relatively conserved N-region

As described in chapter 3, N-signal-less proteins such as cytosolic or nuclear proteins contain relatively conserved N-terminal (Figure 3.2). Similarly, non-cleaved mitochondrial proteins show similar tendency in very end of N-terminal (Figure 4.1).

4.5.2 Auxiliary attributes are better than pHMM score

In chapter 2, pHMM score and weighted score by Gamma mixture are described to classify mitochondrial proteins into two groups: cleaved and non-cleaved proteins. I tested if auxiliary features improves the result. Importance of each feature are were estimated by F-score (Figure 4.2). As

second highest dashed line is F-score of Sc_{hmm}, there are more important features such as highest dashed line, F-score of $\overline{Chg}_{net}$ or $\mathrm{LD}_z(i)$.

4.5.3 Performance comparisons with preceding systems

As characters of internal signal within non-cleaved proteins are poorly understood at present, needs for classification of non-cleaved proteins and cleaved proteins have been independently discussed [7, 9]. Preceding systems, TargetP and Mitoplot II, are referred as standard prediction systems; therefore, MoiraiSP was compared with the two systems. There are two kinds of profile HMM models; thus, raw HMM score (Table 4.2, 4.4) and weighted HMM score (Table 4.3, 4.5). In addition, evolutional divergence score was also used in the yeast data set (Table 4.2, 4.3).

As a result, combination of physico-chemical features, raw HMM score and evolutional divergent score was best amongst yeast prediction models (Table 4.2), and HMM scores weighted by Gamma mixture showed highest evaluation values in the plant data set (Table 4.5).

Table 4.2: Performances of yeast SVM models.

	MCC	AUC	BAC	Accuracy for cleavage
*HMM –base line	0.634 ± 0.149	0.855 ± 0.072	0.814 ± 0.070	54.8% (57/104)
HMM + F_{phy}	0.819 ± 0.121	0.945 ± 0.059	0.906 ± 0.060	64.4% (67/104)
HMM + F_{phy}+ $\mathrm{LD}_z(i)$	$\mathbf{0.828} \pm 0.125$	$\mathbf{0.957} \pm 0.048$	$\mathbf{0.913} \pm 0.063$	**65.4**% (68/104)
$\mathrm{LD}_z(i)$ only	0.595 ± 0.176	0.886 ± 0.066	0.794 ± 0.088	63.5% (66/104)
TargetP	0.582	-	0.788	47.1% (49/104)
MitoProtII	0.552	-	0.770	32.7% (34/104)

*: non SVM

Table 4.3: Performances of yeast SVM models when using Gamma mixture.

	MCC	AUC	BAC	Accuracy for cleavage
*HMM –baseline	0.739 ± 0.170	0.897 ± 0.075	0.865 ± 0.085	58.7% (61/104)
HMM + F_{phy}	0.780 ± 0.099	0.943 ± 0.050	0.887 ± 0.050	62.5% (65/104)
HMM + F_{phy}+ $\mathrm{LD}_z(i)$	0.792 ± 0.154	0.945 ± 0.055	0.893 ± 0.076	63.5% (66/104)

*: non SVM

Table 4.4: Performances of plant SVM models.

	MCC	AUC	BAC	Accuracy for cleavage
*HMM –base line	0.840 ± 0.130	0.957 ± 0.057	0.928 ± 0.066	72.8% (59/81)
HMM + F_{phy}	0.807 ± 0.182	0.954 ± 0.066	0.892 ± 0.100	**76.5**% (62/81)
TargetP	0.504	-	0.751	39.5% (32/81)
MitoProtII	0.676	-	0.806	16.0% (13/81)

*: non SVM

Table 4.5: Performances of plant SVM models when using Gamma mixture.

	MCC	AUC	BAC	Accuracy for cleavage
*HMM –base line	$\mathbf{0.933} \pm 0.116$	$\mathbf{0.979} \pm 0.041$	$\mathbf{0.969} \pm 0.053$	75.3% (61/81)
HMM + F_{phy}	0.827 ± 0.175	0.932 ± 0.098	0.905 ± 0.107	75.3% (61/81)

*: non SVM

4.5.4 Cleavage prediction under the best model

Since the yeast dataset includes not only R-2 motif proteins but also 18 and 21 proteins with R-3 and R-10 motif, respectively. In addition, Icp55 and Oct1 processed proteins contain a few R-none class proteins, which does not contain arginine at typical position -2, -3 or -10. In particular, R-3 and R-10 proteins might be candidate of Icp55 or Oct1 substrates; therefore, they were predicted by the most accurate model shown in Table 4.2, which uses HMM score, F_{phy} and normalized divergent scores (Table 4.6). MoiraiSP results in better performance overall; however, some categories showed worse results than the other two systems.

Table 4.6: Detailed cleavage prediction performances among systems. Denominators show number of predicted proteins as cleavable in each categories, and numerators indicate the number of correctly predicted proteins.

	Total	MPP only (R-2)	Icp55 (R-2)	Oct1(R-2)	R-3	R-10	R-none
MoiraiSP	**63.9***% (83/130)	**63.5***% (33/52)	**87.9***% (29/33)	50.0*% (6/12)	50.0% (6/12)	47.4% (9/19)	0% (0/2)
TargetP	52.9% (65/123)	34.7% (17/49)	80.0% (24/30)	66.7% (8/12)	**66.7**% (8/12)	44.4% (8/18)	0% (0/2)
MitoProtII	39.3% (48/122)	16.0% (8/50)	58.1% (18/31)	**72.7**% (8/11)	38.5% (5/13)	**52.9**% (9/17)	0% (0/1)

*: Validated by 10-fold cross-validation

4.6 Discussion

4.6.1 Limitation for the application of presequence length distribution

In chapter 2, weighting by estimated Gamma mixture for presequence length showed better performance than classification by raw HMM score. However, weighting changed the result for the worse in yeast SVM (Table 4.2,4.3). As I stated in the dataset section, Gamma mixture affects not only HMM score but also other physico-chemical features. Although F-score of Sc_{hmm} was improved by Gamma mixture, F-score of $\overline{Chg}_{net}$, which is estimated to be most important feature, decreased (Figure 4.3). At present, information about presequence length was used as Gamma mixture distribution and applied for weighting of HMM score. As a result, position of window whose HMM score was highest shifted to another point; thus leading to worse F-score of some important features and the result. To avoid changes by Gamma mixture, predicted position (or presequence length) was tested as a feature of SVM. However, result was not improved. In fact, F-score of position is 0.08 and indicates weak discriminative power. In the plant data set, however, weighting by Gamma mixture showed better performance. Since presequence distribution seems to be an important feature, different way of application might improve result for yeast as well.

4.6.2 Computation of divergent scores

Although this problem has been already discussed in chapter 3, same problem might lie behind classification between cleaved and non-cleaved proteins. For example, amino acids which have similar physical properties such as leucine and isoleucine are treated as distinct characters. However, the result shows good MCC for even a model using only divergent score; therefore, shannon entropy can work relatively appropriate measurement in this case.

4.6.3 Slightly divergent region in non-cleaved proteins

Unexpectedly, distribution of $LD_z(i)$ F-scores looks bimodal distribution (Figure 4.2). Figure 3.3 infers that N-terminal residues have discriminative power, but positions after 40 alone seems not to classify proteins correrctly. In fact, divergence looks mostly the same after 40 among three classes (Figure 3.2). On the other hand, N-signal-less mitochondrial proteins, non-cleaved proteins, might include relatively non-conserved region around position 50 to 60 where cleaved mitochondrial proteins are well conserved (Figure 4.1). Such characteristic two regions seem to lead to relatively good classification, even though no direct sequence information was used (Table 4.2). This character was not observed in chapter 3; therefore, non-cleaved proteins, N-signal-less proteins, might contain informative features.

Non-cleaved mitochondrial proteins posses internal signals for correct localization instead of cleavable N-terminal signal, and interact with other proteins within mitochondria [8]. Non-cleaved mitochondrial proteins tend to be membrane proteins; therefore, this region might be transmembrane domain. However, mean hydrophobicity plot shows low hydrophobicity around position 50 to 60 (Figure 4.4). Internal signal has not been well characterized, and search for them in experimental biology is in progress. Although the reason for this non-conservation part in the middle of N-terminal is elusive, this seems to contain some discriminative power.

4.6.4 Limitation about annotations of the dataset

As Table 4.6 shows, MoiraiSP has difficulty about cleavage prediction about R-3 or R-10 motif proteins. In fact, putative MPP cleavage sites for most of the R-3 or R-10 proteins locate one residue or eight residue upstream of the actual reported site. In this project, I used only annotated proteins as Ico55 or Oct1 processed to train sub-filter. However, almost R-3 or R-10 proteins' cleavage site look similar to those of Icp55 or Oct1 even though they were not annotated. Interestingly, some R-3

motif proteins has destabilizing residue such as methionine at position -1 and stabilizing residue at +1 position, even though these motif cannot match with known Icp55 motif [FLY][STA]. To label second cleavage peptidases to predicted site, I excluded these proteins. Nevertheless, training by them should have potential for improvement of the result, though it makes labeling of peptidases difficult. As a future work, to build two kinds of different models might be a point of compromise: a model which be less accurate but can show concrete labeling of intermediate proteases and more accurate model which shows only inferred labels . Here, I show the former model as result of my Master thesis research.

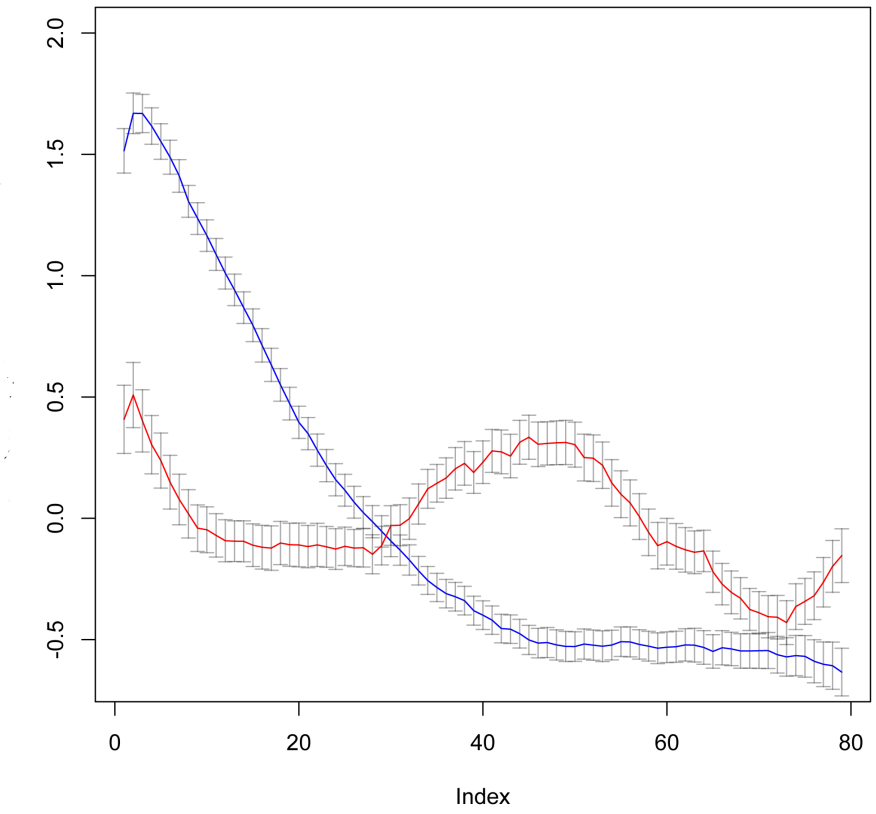

Figure 4.1: Normalized local divergence scores are shown for the 100 residue N-terminal region for cleaved MTS containing (blue) and non-cleaved mitochondrial (red) proteins in the yeast dataset. The error bars denote the standard error. y-axis shows $LD_z(i)$, and x-axis indicates start position of window.

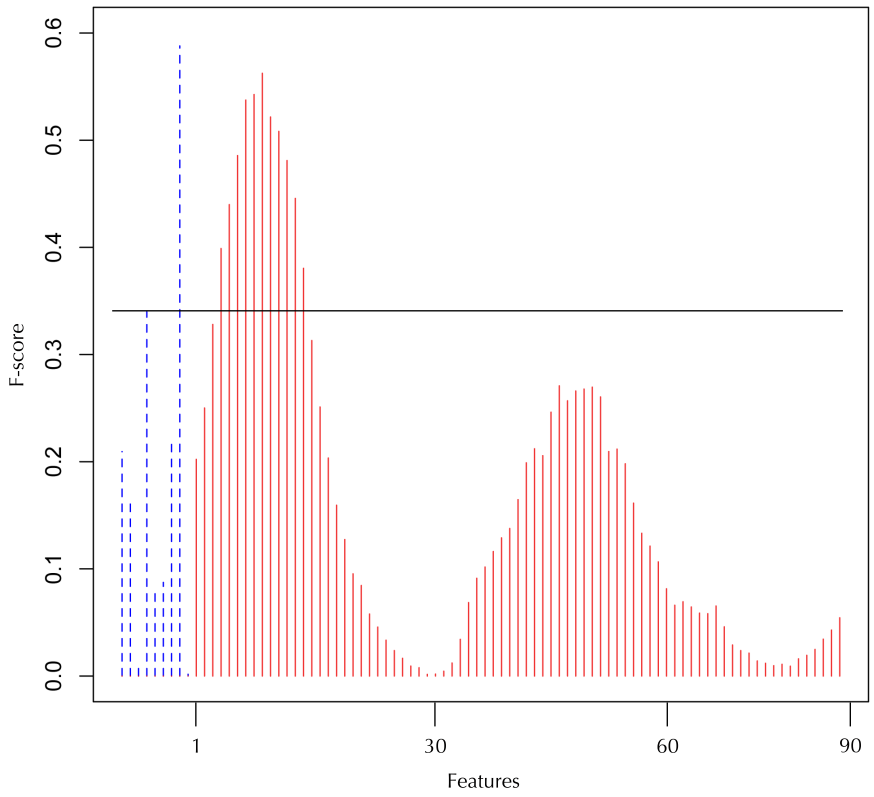

Figure 4.2: The estimated feature importance by F-score. Blue dashed lines denote physico-chemical and HMM score. Red lines show normalized local divergent scores in 100 N-terminal positions. Black horizontal line indicates F-score of HMM score. Blue dashed lines indicate F-scores of $\#_D, \#_E, \#_H, Sc_{hmm}, \#_R, \#_K, \#_-, \overline{Chg}_{net}, \#_{positive}$ from left to right.

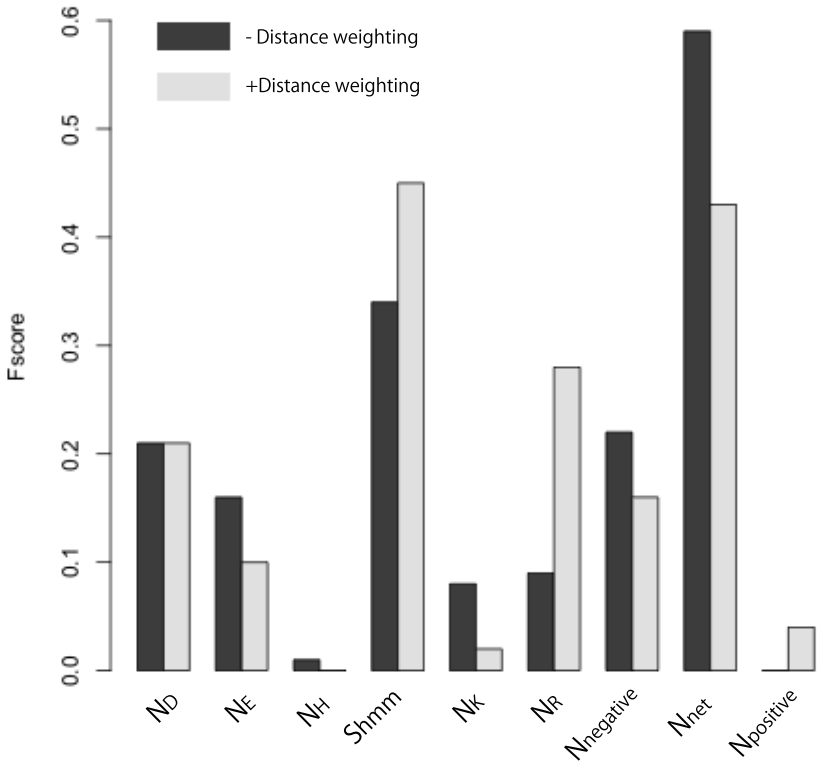

Figure 4.3: The estimated feature importance by F-score in both weighted HMM and raw HMM.

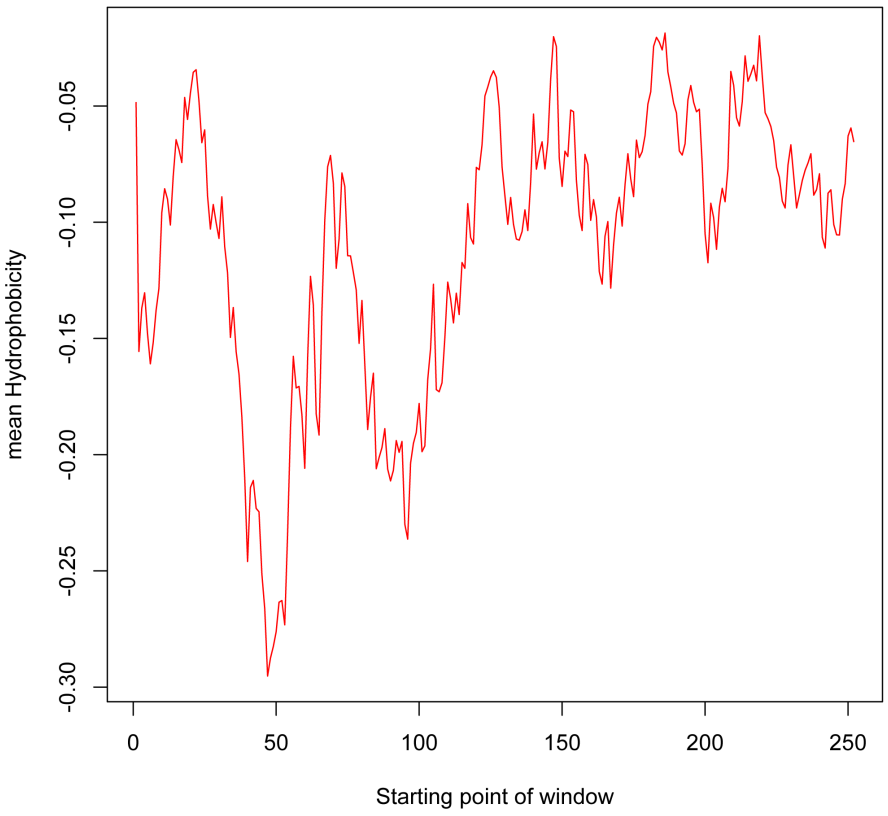

Figure 4.4: Mean hydrophobicity amongst non-cleaved proteins scaled by Kyte-Doolittle index. A window size of 9 was used for smoothing.

Chapter 5

Search for novel substrates

5.1 Results and Discussion

5.1.1 R-10 motif proteins are likely to be Oct1 substrates

Since proteomic data set provided by Vögtle *et al.* include 21 proteins whose cleavage sites look R-10 motif, which infers double digestion by MPP and Oct1. The problem of these proteins is difference from known Oct1 motif (Figure 5.1, 2.5). Although known R-10 Oct1 motif is RX|(F/L/I)XX(T/S/G)XXXX| [1, 28], position -8 and -5, position 3 and 6 in Figure 5.1, respectively, contain atypical residues such as tyrosine and tryptophan at -8 and leucine or aspartic acid at -5. There has not been strong authority that they are Oct1 substrate, so 21 R-10 motif proteins were separated from the training data.

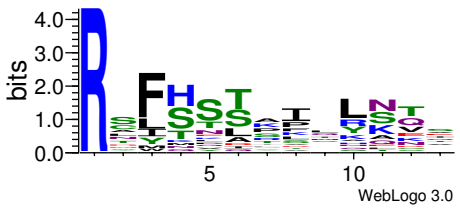

Figure 5.1: Sequence logo generated from cleavage site of R-10 motif proteins.

MoiraiSP results in good performance, especially for prediction of MPP cleavage site and classification, therefore, 21 proteins were tested for further discussion by the system , SVM model usign F_{phy} and $\mathrm{LD}_z(i)$.

Table 5.1: List of 21 R-10 motif proteins.

OLN	Probability	Distance from putative MPP site
YBR047W	0.93	8
YDR036C	0.9	8
YDR178W	0.94	16
YDR234W	0.88	8
YER078C	0.9	8
YGL107C	0.62	8
YHR147C	0.63	8
YIL070C	0.97	8
YJR080C	0.75	8
YJR100C	0.84	8
YLR069C	0.94	8
YLR295C	0.61	23
YML042W	0.95	8
YMR189W	0.95	8
YMR232W	Predicted non-cleavable	-
YNL177C	0.86	8
YNR036C	0.87	8
YOR354C	0.71	8
YPL224C	0.87	8
YPR001W	0.88	8
YPR025C	Predicted non-cleavable	-

MoiraiSP predicts 8 residue upstream of R-10 motif cleavage sites as putative MPP cleavage sites for 17 out of 19 predicted cleavable proteins. In other words, distance from 17 predicted MPP cleavage sites to reported cleavage sites are 8 residues, typical distance for Oct1 as its name indicates. 9 out of the 17 proteins are predicted as Oct1 substrates by MoiraiSP; thus, rest of them are not predicted as Oct1 substrates due to atypical residues in their sequence at position -8 or -5.

Recently, a novel Oct1 substrate was reported in yeast, and its sequence motif is unusual: position -10 is not either R or K but cysteine and -5 is leucine [52]. As figure 5.1 indicates, leucine is weakly conserved at position -5. Taking revised motif into account, Figure 5.1 does not contradict with the

statement that R-10 motif proteins are also Oct1 substrates.

Additionally, position -8 tends to prokaryotic destabilizing residue except two isoleucine, and position -7 includes numerous prokaryotic stabilizing residues except lysine. R-10 motif's destabilising-stabilising pattern is consistent with that of Icp55 or Oct1.

Finally, the distance of YDR178W from putative MPP site is 16. Because Oct1 function is still elusive, Oct1 may have the potential for double digestion after first Oct1 cleavage.

In conclusion, Oct1 is also related to mitochondrial protein turn over [52]; therefore, 21 proteins on this list are likely to be novel Oct1 substrates.

5.1.2 R-3 motif proteins might be potentially double digestioned proteins

As same as R-10 motif proteins, 18 R-3 motif proteins were also checked by MoiraiSP (Table 5.2). One third of the proteins were predicted as non-cleavable, and results of TargetP or MitiplotII are similar proportion. Although the reason is unknown, this category might include some experimental errors. Basic tendency is similar to that of R-10 proteins stated the above, and one interesting proteins is YBR026C. This contains methionine at -1 position and serine at +1 positon; thus, shows similar pattern to weakly conserved plant R-3 motif. Even if Icp55 does not cleave this protein, methionine peptidase may remove methionine after MPP processing in mitochondria. But so far, there is no report about MPP+methionine peptidase processing; therefore, this is a speculation at present.

5.1.3 Mcr1p might be not only IMP but also Pcp1 substrate

The yeast data set includes substrates for all of known mitochondrial proteases: MPP, Icp55, Oct1, m-AAA, i-AAA, IMP and Pcp1. As proteases other than the first three are related to regulation within mitochondria rather than N-terminal signal removal, importance for their substrate

Table 5.2: List of 18 R-3 motif proteins.

OLN	Probability	Distance from putative MPP site
YBR026C	0.91	1
YBR037C	0.91	8
YBR251W	0.84	1
YDL044C	0.66	1
YDR070C	Predicted non-cleavable	-
YDR298C	0.67	8
YDR494W	Predicted non-cleavable	-
YGL221C	Predicted non-cleavable	-
YJL180C	0.89	1
YJR003C	0.94	1
YKL134C	0.9	1
YKR063C	Predicted non-cleavable	-
YMR157C	0.86	0
YNL100W	Predicted non-cleavable	-
YOR298C-A	Predicted non-cleavable	-
YOR334W	0.96	1
YOR356W	0.79	1
YPL059W	0.81	18

search increases in recent years, especially in medical field. Amongst them, Pcp1 has been focused on because this is homolog of PARL, which is a rhomboid proteases of human mitochondria and related to Parkinson's disease through cleavage of PINK1 [53]. Rhomboid protease cleaves substrate within or nearby membrane, and cleavage motif for rhomboid was recently reported: [^WD][IMYFWLV][^WPD][^WF][AGCS]|[^P][FIMVACLTW] [54]. Since this motif was conserved from prokaryotes to eukaryotes, there should be possibility that yeast Pcp1 can also recognize the rhomboid motif. At present, there are two reported substrates for Pcp1, Mgm1 and Ccp1, and they do not match the motif [21]. In the yeast dataset, however, cleavage site of Mcr1p matches this motif completely, TVAIA|AA and the site locates within predicted transmembrane domain.

On the other hand, it was reported that Mcr1p is cleaved by IMP at different position [55]. In the yeast data set, IMP cleavage site was not observed, and only TVAIA|AA was reported.

Interestingly, Mcr1p locates in both outer membrane and inner membrane and accumulates only in outer membrane in the absence of membrane potential [56]. This reflects location shift of PINK1

which depends on rhomboid protease [53]. In addition, mutation of Mcr1p, which changed TVAIAAA to TVAIQQA, generates irregular 30-kDa fragment and localization of Mcr1p was disturbed [56].

Although size of shorter form is 32-kDa, theoretical molecular weight from reported IMP cleavage site is 29641.74 Da. On the other hand, theoretical size from rhomboid motif matched site is 31729.02 Da; thus, theoretical size also supports cleavage site between position 23 and 24.

As a summation, all known experimental result of Mcr1p is consistent with rhomboid cleavage for Mcr1p. Since N-terminal of Mcr1p in mature part has not been sequenced since 1994 [55], start point of Mcr1p seems unclear at present. IMP cleavage site is inconsistent with at least proteomic data [9]. Although the details are not yet clear, cleavage and regulation of Mcr1p is interesting in terms of both medical and biological research.

5.2 Methods

5.2.1 Cleavage site prediction

Classification and cleavage site prediction was conducted by MoiraiSP discussed in chapter 2 and 4. Physico-chemical features, HMM score and normalized divergence score were used, but Gamma mixture was not applied, as it did not seem to increase accuracy in preliminary trial.

5.2.2 Topology prediction

MEMSAT-SVM was used to discriminate membrane proteins from globular proteins and predict transmembrane domain [57]. Proteins predicted as globular proteins by MEMSAT-SVM were not further analyzed.

5.2.3 Molecular weight calculation

Compute pI/Mw (`http://web.expasy.org/compute_pi/`) was used to calculate molecular weight of shorter form of Mcr1p.

Chapter 6

Conclusion

An aim of this research is to develop a new predictor for cleavage site prediction of mitochondrial proteins. In the field of mitochondrial studies, the importance of prediction against various new kinds of cleavage site has emerged because of the relation between protease location and target protein function. Unfortunately, even prediction for MPP, the best known mitochondrial proteases, is not well solved. Moreover several membrane associated mitochondrial proteases exist for which only a handful of substrates data is available, thus prediction work focused on MPP due to the enrichment of the data and knowledge. Two software tools, MitoProtII and TargetP, are already existing to predict matrix targeting signals and can be considered industry standard tools in the field of mitochondrial research [14, 13], but their accuracies are far from ideal. In this thesis, I show that my predictor gives a good performance in this task.

In chapter 3, I discuss a novel application of evolutionary information. I defined LD(i), a simple measure of sequence divergence, and show that it correlates significantly and positively with the N-terminal sorting signals. Moreover, it can be combined with physico-chemical propensities for further increase of accuracy. In particular, using this divergence score for binary classification between

cleaved MTS and non-cleaved mitochondrial proteins results in quite good result and indicates interesting region in non-cleaved proteins, which was mentioned in chapter 4.

Since the yeast dataset does not contain annotation for proteases other than Icp55 and Oct1, I conducted a literature search. This work revealed numerous proteases related to the cleaved sites in the yeast dataset such as i-AAA, m-AAA [17, 18], and these sites do not contain arginine at -2 position; therefore, at least the yeast data set is likely to include substrates of diverse proteases in addition to MPP. Or, there is a possibility that the yeast data sets include experimental errors about cleavage sites. It is clear that systematic analysis of mitochondrial proteases is necessary to understand mitochondrial proteome. In this thesis, I believe that this thesis is a first step in that direction.

In summation, MoiraiSP can work to classify mitochondrial proteins into two groups at protein level, and predict cleavage position of MPP and intermediate proteases such as Oct1 and Icp55 or its analog in plant.

Bibliography

[1] O. Gakh, P. Cavadini, and G. Isaya. Mitochondrial processing peptidases. *Biochim Biophys Acta*, 1592(1):63–77, 2002.

[2] K. Nicolay, F. D. Laterveer, and W. L. van Heerde. Effects of amphipathic peptides, including presequences, on the functional integrity of rat liver mitochondrial membranes. *J Bioenerg Biomembr*, 26(3):327–334, 1994.

[3] D. Roise, S. J. Horvath, J. M. Tomich, J. H. Richards, and G. Schatz. A chemically synthesized pre-sequence of an imported mitochondrial protein can form an amphiphilic helix and perturb natural and artificial phospholipid bilayers. *EMBO J*, 5(6):1327–1334, 1986.

[4] P. Moberg, A. Stahl, S. Bhushan, S. J. Wright, A. Eriksson, B. D. Bruce, and E. Glaser. Characterization of a novel zinc metalloprotease involved in degrading targeting peptides in mitochondria and chloroplasts. *Plant J*, 36(5):616–628, 2003.

[5] Gisbert Schneider, Sara Sjöling, Erik Wallin, Paul Wrede, Elzbieta Glaser, and Gunnar von Heijne. Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. *PROTEINS*, 30:49–60, 1998.

[6] G. Isaya, F. Kalousek, and L. E. Rosenberg. Amino-terminal octapeptides function as recognition signals for the mitochondrial intermediate peptidase. *J Biol Chem*, 267(11):7904–7910, 1992.

[7] S. Huang, N. L. Taylor, J. Whelan, and A. H. Millar. Refining the definition of plant mitochondrial presequences through analysis of sorting signals, n-terminal modifications, and cleavage motifs. *Plant Physiol*, 150(3):1272–1285, 2009.

[8] O. Schmidt, N. Pfanner, and C. Meisinger. Mitochondrial protein import: from proteomics to functional mechanisms. *Nat Rev Mol Cell Biol*, 11(9):655–667, 2010.

[9] F. N. Vögtle, S. Wortelkamp, R. P. Zahedi, D. Becker, C. Leidhold, K. Gevaert, J. Kellermann, W. Voos, A. Sickmann, N. Pfanner, and C. Meisinger. Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. *Cell*, 139(2):428–439, 2009.

[10] E. Deas, H. Plun-Favreau, S. Gandhi, H. Desmond, S. Kjaer, S. H. Loh, A. E. Renton, R. J. Harvey, A. J. Whitworth, L. M. Martins, A. Y. Abramov, and N. W. Wood. PINK1 cleavage at position A103 by the mitochondrial protease parl. *Hum Mol Genet*, 20(5):867–879, 2011.

[11] P. Martinelli and E. I. Rugarli. Emerging roles of mitochondrial proteases in neurodegeneration. *Biochim Biophys Acta*, 1797(1):1–10, 2010.

[12] F. Bonn, T. Tatsuta, C. Petrungaro, J. Riemer, and T. Langer. Presequence-dependent folding ensures MrpL32 processing by the m-AAA protease in mitochondria. *EMBO J*, 30(13):2545–2556, 2011.

[13] 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*, 300(4):1005–1016, 2000.

[14] M.G. Claros and P. Vincens. Computational method to predict mitochondrially imported proteins and their targeting sequences. *European Journal of Biochemistry*, 241:779–786, 1996.

[15] V. Zara, F. Palmieri, K. Mahlke, and N. Pfanner. The cleavable presequence is not essential for import and assembly of the phosphate carrier of mammalian mitochondria but enhances the specificity and efficiency of import. *J Biol Chem*, 267(17):12077–12081, 1992.

[16] W. Neupert and J. M. Herrmann. Translocation of proteins into mitochondria. *Annu Rev Biochem*, 76:723–749, 2007.

[17] S. Augustin, M. Nolden, S. Muller, O. Hardt, I. Arnold, and T. Langer. Characterization of peptides released from mitochondria: evidence for constant proteolysis and peptide efflux. *J Biol Chem*, 280(4):2691–2699, 2005.

[18] K. Esser, B. Tursun, M. Ingenhoven, G. Michaelis, and E. Pratje. A novel two-step mechanism for removal of a mitochondrial signal sequence involves the mAAA complex and the putative rhomboid protease pcp1. *J Mol Biol*, 323(5):835–843, 2002.

[19] A. B. Taylor, B. S. Smith, S. Kitada, K. Kojima, H. Miyaura, Z. Otwinowski, A. Ito, and J. Deisenhofer. Crystal structures of mitochondrial processing peptidase reveal the mode for specific cleavage of import signal sequences. *Structure*, 9(7):615–625, 2001.

[20] T. Niidome, S. Kitada, K. Shimokata, T. Ogishima, and A. Ito. Arginine residues in the extension peptide are required for cleavage of a precursor by mitochondrial processing peptidase. demonstration using synthetic peptide as a substrate. *J Biol Chem*, 269(40):24719–24722, 1994.

[21] A. Schäfer, M. Zick, J. Kief, M. Steger, H. Heide, S. Duvezin-Caubet, W. Neupert, and A. S. Reichert. Intramembrane proteolysis of Mgm1 by the mitochondrial rhomboid protease is highly promiscuous regarding the sequence of the cleaved hydrophobic segment. *J Mol Biol*, 401(2):182–193, 2010.

[22] J. Dai and J. Cheng. HMMEditor: a visual editing tool for profile hidden Markov model. *BMC Genomics*, 9 Suppl 1:S8, 2008.

[23] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res*, 30(14):3059–3066, 2002.

[24] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton. Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.

[25] R. L. Ninnis, S. K. Spall, G. H. Talbo, K. N. Truscott, and D. A. Dougan. Modification of PATase by L/F-transferase generates a ClpS-dependent N-end rule substrate in Escherichia coli. *EMBO J*, 28(12):1732–1744, 2009.

[26] W. Apel, W. X. Schulze, and R. Bock. Identification of protein stability determinants in chloroplasts. *Plant J*, 63(4):636–650, 2010.

[27] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[28] S. S. Branda and G. Isaya. Prediction and identification of new natural substrates of the yeast mitochondrial intermediate peptidase. *J Biol Chem*, 270(45):27366–27373, 1995.

[29] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[30] Tatiana Benaglia, D. Chauveau, D.R. Hunter, and D. Young. mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*, 32(6):1–29, 2009.

[31] T. Saitoh, M. Igura, T. Obita, T. Ose, R. Kojima, K. Maenaka, T. Endo, and D. Kohda. Tom20 recognizes mitochondrial presequences through dynamic equilibrium among multiple bound states. *EMBO J*, 26(22):4777–4787, 2007.

[32] H. Yamamoto, N. Itoh, S. Kawano, Y. Yatsukawa, T. Momose, T. Makio, M. Matsunaga, M. Yokota, M. Esaki, T. Shodai, D. Kohda, A. E. Hobbs, R. E. Jensen, and T. Endo. Dual role of the receptor Tom20 in specificity and efficiency of protein import into mitochondria. *Proc Natl Acad Sci U S A*, 108(1):91–96, 2011.

[33] T. Tsukamoto, S. Hata, S. Yokota, S. Miura, Y. Fujiki, M. Hijikata, S. Miyazawa, T. Hashimoto, and T. Osumi. Characterization of the signal peptide at the amino terminus of the rat peroxisomal 3-ketoacyl-CoA thiolase precursor. *J Biol Chem*, 269(8):6001–6010, 1994.

[34] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, and A. Bairoch. UniProtKB/Swiss-Prot. *Methods Mol Biol*, 406:89–112, 2007.

[35] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, 340(4):783–795, 2004.

[36] K. P. Byrne and K. H. Wolfe. The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*, 15(10):1456–1461, 2005.

[37] I. Mayrose, D. Graur, N. Ben-Tal, and T. Pupko. Comparison of site-specific rate-inference methods for protein sequences: empirical bayesian methods are superior. *Mol Biol Evol*, 21(9):1781–1791, 2004.

[38] F. Johansson and H. Toh. A comparative study of conservation and variation scores. *BMC Bioinformatics*, 11:388, 2010.

[39] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132, 1982.

[40] J. Ross Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[41] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10, 2009.

[42] V.N.. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc, New York, 1999.

[43] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):1–27, 2011.

[44] Y.-W. Chen and C.-J. Lin. Combining svms with various feature selection strategies. Available from `http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf`, 2005.

[45] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21:1263–1284, September 2009.

[46] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*, 2(4):953–971, 2007.

[47] Maria Edman, Tanja Jarhede, Michael Sjöström, and Åke Wieslander. Different sequence patterns in signal peptides from mycoplasmas, other gram-positive bacteria, and escherichia coli: a multivariate data analysis. *PROTEINS: Structure, Function, and Genetics*, 35:195–205, 1999.

[48] P. Horton, K. J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, and K. Nakai. WoLF PSORT: protein localization predictor. *Nucleic Acids Res*, 35(Web Server issue):W585–W587, 2007.

[49] S. Kitada, E. Yamasaki, K. Kojima, and A. Ito. Determination of the cleavage site of the presequence by mitochondrial processing peptidase on the substrate binding scaffold and the multiple subsites inside a molecular cavity. *J Biol Chem*, 278(3):1879–1885, 2003.

[50] K. Kojima, S. Kitada, T. Ogishima, and A. Ito. A proposed common structure of substrates bound to mitochondrial processing peptidase. *J Biol Chem*, 276(3):2115–2121, 2001.

[51] T.F. Wu, C.J. Lin, and R.C. Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, 2004.

[52] F. N. Vögtle, C. Prinz, J. Kellermann, F. Lottspeich, N. Pfanner, and C. Meisinger. Mitochondrial protein turnover: role of the precursor intermediate peptidase Oct1 in protein stabilization. *Mol Biol Cell*, 22(13):2135–2143, 2011.

[53] S. M. Jin, M. Lazarou, C. Wang, L. A. Kane, D. P. Narendra, and R. J. Youle. Mitochondrial membrane potential regulates PINK1 import and proteolytic destabilization by PARL. *J Cell Biol*, 191(5):933–942, 2010.

[54] K. Strisovsky, H. J. Sharpe, and M. Freeman. Sequence-specific intramembrane proteolysis: identification of a recognition motif in rhomboid substrates. *Mol Cell*, 36(6):1048–1059, 2009.

[55] K Hahne, V Haucke, L Ramage, and G Schatz. Incomplete arrest in the outer membrane sorts NADH-cytochrome b5 reductase to two different submitochondrial compartments. *Cell*, 79(5):829–39, 1994.

[56] V. Haucke, C. S. Ocana, A. Honlinger, K. Tokatlidis, N. Pfanner, and G. Schatz. Analysis of the sorting signals directing NADH-cytochrome b5 reductase to two locations within yeast mitochondria. *Mol Cell Biol*, 17(7):4024–4032, 1997.

[57] T. Nugent and D. T. Jones. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, 10:159, 2009.