博士論文

論文題目　Research on the amino sequence characteristics determining the transport, membrane topology and peptidase processing of mitochondrial proteins
(ミトコンドリア蛋白質の輸送、膜トポロジー及びペプチダーゼによる切断を決定するアミノ酸配列の特徴に関する研究)

氏　　名　深沢　嘉紀

Abstract

It is well known that mitochondria function as the essential power plants of most eukaryotic cells. They also make important contributions to many other vital cell functions such as lipid metabolism and calcium homeostasis. Moreover, mitochondrial dysfunction has been implicated in numerous medical conditions such as Parkinson's and Alzheimer's disease.

Although many mitochondrial proteins have been experimentally identified, a complete list is not available even for intensely studied model organisms. Thus bioinformatics tools to predict mitochondrial proteins from their amino acid sequences are widely used to complement experimental data; but their accuracy is far from perfect and they have not improved significantly for roughly a decade.

Existing prediction tools already employ sophisticated machine learning techniques so the key to progress seems to lie in the utilization of new proteomics data and the identification or refinement of sequence features that reflect the underlying molecular biology. Moreover, the development of such sequence features may be useful not only for more accurate prediction but also provide useful biological hints. Here I report features of local sequences in mitochondrial proteins which regulate their transport, membrane insertion, and peptidase processing.

In chapter 1, I summarize the necessary background to understand my thesis work. This chapter

overviews the known biology of mitochondrial proteins in terms of their import, cleavage and membrane spanning domains.

In chapter 2, I report on the sequence divergence of N-terminal sorting signals, and show that divergence is a promising novel feature for signal prediction. For yeast, mammal and plant datasets, evolutionary sequence divergence alone has significant power to identify sequences with N-terminal sorting sequences. First I utilized YGOB, a curated database for orthologs between budding yeast and its related species, for calculation of sequence divergence of yeast proteins. I then demonstrate that sequence divergence is nearly as effective when computed on automatically defined ortholog sets for yeast, mammal, and plant datasets as on the hand curated ones. Unfortunately, sequence divergence did not necessarily increase classification performance when combined with some traditional sequence features such as amino acid composition. However a post-hoc analysis of the proteins in which sequence divergence changes the prediction yielded some proteins with atypical (i.e. not MPP-cleaved) matrix targeting signals as well as a few misannotations.

In chapter 3, I introduce MitoFates, a prediction system for mitochondrial presequences (N-terminal regions cleaved upon translocation into the mitochondria) designed with the knowledge of mitochondrial intermediate peptidases in mind and trained on recent proteomics data. MitoFates achieves better performance in both signal and cleavage site prediction. To obtain this performance, I revisited classical features for predicting this signal and searched for novel specific sequence motifs in the mitochondrial N-terminal presequence. Among the classical features, I revisited a detector of local sequences with the potential to form an amphiphilic α-helix, with a hydrophobic and hydrophilic face, inspired by the structure of the presequence recognizer Tom20 and Tom22. In previous applications, this feature has not be very effective, but I noted that the formulation used did not distinguish between negative and positive charge. By introducing a new term rewarding

helices with positive charges opposite the hydrophobic face, I greatly increased the discriminant power of this feature. Employing recently developed techniques for sensitive multiple hypothesis testing, I discovered several novel and significant motifs from presequence, most of which show a positively charged amphiphilicity (possibily indicating recognition by TOM complex) or matching the consensus sequence of presequence cleavage sites. I also refined cleavage site of presequence by utilizing recent proteomics data and taking into consideration recent experimental results such as the discovery of Icp55. This leads to greatly improved performance of presequence cleavage site prediction (reducing misprediction of cleavage site position from $\approx 48\%$ to $\approx 29\%$, addressing the longstanding and often discussed lack of accurate tools for this task. In addition, in the light these refined and novel presequence features, I cluster and discuss classes of presequences.

In chapter 4, I present sequence features of transmembrane domains (TMD) of proteins in the mitochondrial inner membrane, improving the discrimination between those regions and spuriously similar regions in soluble cytosolic proteins. The difficulty of predicting the TMDs of mitochondrial membrane proteins has been noted anecdotally, but the distinct characteristics of mitochondrial TMDs had not been analyzed from the viewpoint of computational biology. Therefore, I analyzed the problem, starting with a previous model which calculates the free energy of TMD membrane insertion using positional amino acid profiles, based on parameters measured for the TMDs of E.R. membranes. As expected, TMDs of the mitochondrial inner membrane show characteristics distinct from either E.R. TMDs or spurious hydrophobic regions of cytosolic globular proteins. However, in terms of free energy distribution, the mitochondrial TMDs overlap with those two distributions, leading to difficult prediction. My statistical analysis surprisingly shows that glycine is significantly enriched in the center of mitochondrial TMDs and negatively charged residues show an asymmetric distribution, consistent with pioneering experimental work on mitochondrial TMDs.

I employ these different characteristics to discriminate mitochondrial TMDs from spurious regions of cytosolic proteins using the sequences of those proteins and their homologs in other organisms, leading to much improved prediction of mitochondrial TMDs in comparison to general predictors for TMDs. I examined the position of predicted TMDs in proteins from a mitochondrial presequence dataset, especially the presequences with atypical features discussed in the chapter 3, and found some interesting cases of non-annotated cleavage sites that locate downstream of TMDs.

Finally, in chapter 5, I summarize, discuss, and conclude my thesis work.

Acknowledgments

"Time flies like an arrow" – I remember this popular proverb now with lots of memory, which is sometimes funny, shameful, delightful, painful, and interesting. Five years ago when I started research at Computational Biology Research Center as a graduate student, I have no required skill and knowledge for a computational biologist. Due to the precious experiences at this institute, I am thankful to the office space, it is incredibly nice for me, and the all environment at 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 progress of this work several times. I owe him tons of drinks, and a comfortable bed in his room at a hotel nearby the Guangzhou airport.

Special thanks is due to Dr. Kenichiro Imai for invaluable feedback and preparation of this work from the beginning. Although I wrote this entire manuscript in the first person "I", Dr. Imai contributed directly to the work on presequence prediction. Also, Dr. Ross KK Leung shared many long discussion with me about MTS classification and helped with the beginning of the sequence divergence project. Prof. Kentaro Tomii is kindly accepted me as a team member at CBRC, and gave me his time for the research, especially team meeting. As a graduate student at the University of Tokyo, graduate students at CBRC have been always helpful and expressed kind patience.

Last but not the least, I would like to thank my family for their support. Without their aid, I

would be unable to continue my career. Above all, I have to appreciate my wife, Choungyoun Park. She is sometimes an intense teacher to revise my presentation, sometimes a most tender person to cheer me up, and almost she is my best friend who understands me the most. My research life would be miserable like a desert without her.

Contents

List of Figures

List of Tables

List of Abbreviations

AUC : Area Under the Curve

C.V. : Cross Validation

CTP : Chloroplast Transit Peptides

E.R. : Endoplasmic Reticulum

HMM : Hidden Markov Model

LD : Local Divergence

MCC : Matthews Correlation Coefficient

MPP : Mitochondrial Processing Peptidase

MSA : Multiple Sequence Alignment

MTS : Matrix Targeting Signal

PA : Positively charged Amphiphilicity

PCA : Principal (Primary) Component Analysis

PSSM : Position Specific Scoring Matrix

PWM : Position Weight Matrix

RBF : Radial Basis Function

RBH : Reciprocal Best Hits

ROC : Receiver Operation Characteristics

SP : Signal Peptides

SVM : Support Vector Machine

TMD : TransMembrane Domain

YGOB : Yeast Gene Order Browser

Chapter 1

Introduction

One character of a cell is that it comprises of multiple membranes to split itself to several compartments. Because of its multiple membranes, a cell can hold numerous different environments where complex and diverse chemical reactions occur. Relevant but functionally distinct internal compartments are called organelles, and it is said that some of them are acquired through engulfment by another organism. Mitochondria is one of such organelles, which is said that it originated from alphaproteobacteria [1], and their functions are vital for almost all of eukaryotes. In fact, mitochondrial dysfunction causes a wide variety of diseases such as muscle and neurodegenerative disease, cardiovascular disease, diabetes and cancer [2]. Obtaining a complete proteome list of mitochondrial proteins is biologically important and an essential step for medical purposes due to the above statement. Although mitochondria contain their own genome and key components are coded in them, almost mitochondrial proteins are coded in the nuclear genome and have to be transported to the mitochondria. In a cell, translocation is usually regulated by targeting signals.

At present, mitochondrial proteins can be divided into two groups in terms of their targeting signals: the amino-terminal signal (presequence) and internal targeting signals, which are recognized by receptors embedded in the mitochondrial membranes. Since proteome is not available at present and each protein's import pathway is not alway clear, it is difficult to know exact proportion of presequence pathway. However, recent proteomic analysis estimated that ∼70% of mitochondrial proteins depends on presequence pathway in the yeast [3]. Therefore, prediction improvement of presequence will lead to discover currently unknown mitochondrial proteins.

Presequences consist of 10-90 residues and are positively charged, and high net charge of presequence comes from skewed amino acid composition. It has been reported that presequence has high arginine and low negatively charged residues composition [4, 5]. Presequences are mostly translocated by the TOM and the TIM complex in the outer and the inner membrane, respectively [6, 7, 8]. The subunits of Tom20 and Tom22 in the TOM complex initiate import of the proteins by recognizing an amphiphilic helical feature consist of hydrophobic and positively charged faces of presequences [6, 9, 10]. Peptide library experiment revealed that 6 residue amphiphilic motif for Tom20 [9], however, this motif matches only 18% of recent proteomic data of yeast presequences [3]. An exhaustive motif search based on discriminative HMM has been performed against mitochondrial proteins, but significant amphiphilic motif was not found [11]. It has been reported in small dataset [12] or anecdotally known that N-terminal signal including presequence has higher mutation rate. Taking these into consideration, presequence seem to hold weakly conserved motif, which revealed experimentally but cannot be detected by distinct 20 amino acid letters, rather than global consensus pattern.

Upon import into mitochondria presequence is in most case eliminated by Mitochondrial Processing Peptidase (MPP) in the matrix, and other intermediate peptidases such as Oct1 in some

cases [13]. Mitochondrial presequences are harmful for the function of mitochondrial membranes due to their amphiphilic property, and as a result, they dissipate membrane potential and uncouple respiration [14, 15]. To avoid such severe disturbances, MPP cleaves presequences and it has been reported that other metallo-protease degrades cleaved presequence after cleavage in *Arabidopsis thaliana* [16]. In terms of the position of arginine from the cleavage site, three motifs of MPP cleavage sites have been reported: R-10 motif, R-3 motif and R-2 motif in which arginine located at -10 position, at -3 position and at -2 position from cleavage site, respectively [5, 13]. R-10 motif can be explained by twice cleavages by MPP and Oct1 [5, 17]. Vögtle *et al.* reported a novel intermediate protease Icp55 can remove one amino acid from the N-terminal after cleavage by MPP, and the relationship between this phenomenon and the half-life of a protein determined by its N-terminal residue [3]. This discovery of Icp55 explains why R-3 and R-2 motifs are found in the cleavage sites [5]. Although unsatisfactory accuracy of current cleavage site prediction is also argued [3, 18], this seems to be reasoned by lack of explicit consideration of these peptidases. In addition, a yeast mitochondrion contains other proteases such as Pcp1, m-AAA, and IMP in its inter-membrane [8]. A counterpart of Pcp1 in human is called PARL. 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.

As it is shown in Figure 1.2, such other proteases are located and function in inner membrane. In fact, many known substrates of those proteases are membrane proteins and contain α-helical transmembrane domain (TMD) (summarized in [8]). TMD analysis should be potentially beneficial to discover unknown substrates of these proteases, however, it is anecdotally said that prediction of mitochondrial TMD is more difficult than TMD in other membranes such as E.R. membrane or plasma membrane. Since proton densities in and out of mitochondrial inner membrane differ to

drive ATP synthase, its local environment has specific character and it can be expected that different TMD feature exists. One pioneering work has been conducted by Botelho and colleagues [19], and negatively charged residue rarely appeared inside of mitochondrial inner membrane. In general, charged residues show symmetric distribution in center of inserted TMD (described in Hessa's energy model [20]), and so-called "positive inside rue", which says that positively charged residues are abundant in loops of cytoplasmic side of membrane [21]. Apparently, mitochondrial inner membrane is located in different chemical condition, and asymmetric distribution may reflect this. Existing methods do not take these differences to their models due to the lack of annotated data and knowledge. Therefore, improvement on not only presequence prediction but also TMD prediction seems to be important in the field of mitochondrial biology. I should also note here that several substrates of inner membrane proteases have been reported in context of human diseases [22, 23, 24], therefore, refined TMD prediction in inner membrane lead to analysis this kind of potentially important proteins.

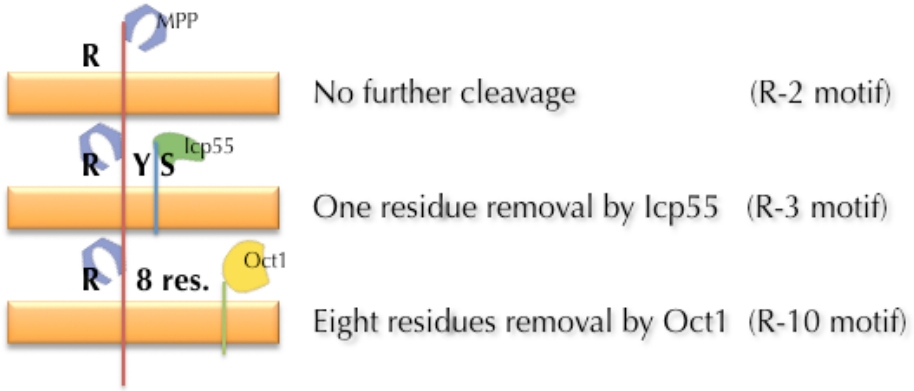


Figure 1.1: Description of mitochondrial cleavage.

The location of each protease and the relationships among them are summarized in Figure 1.2.

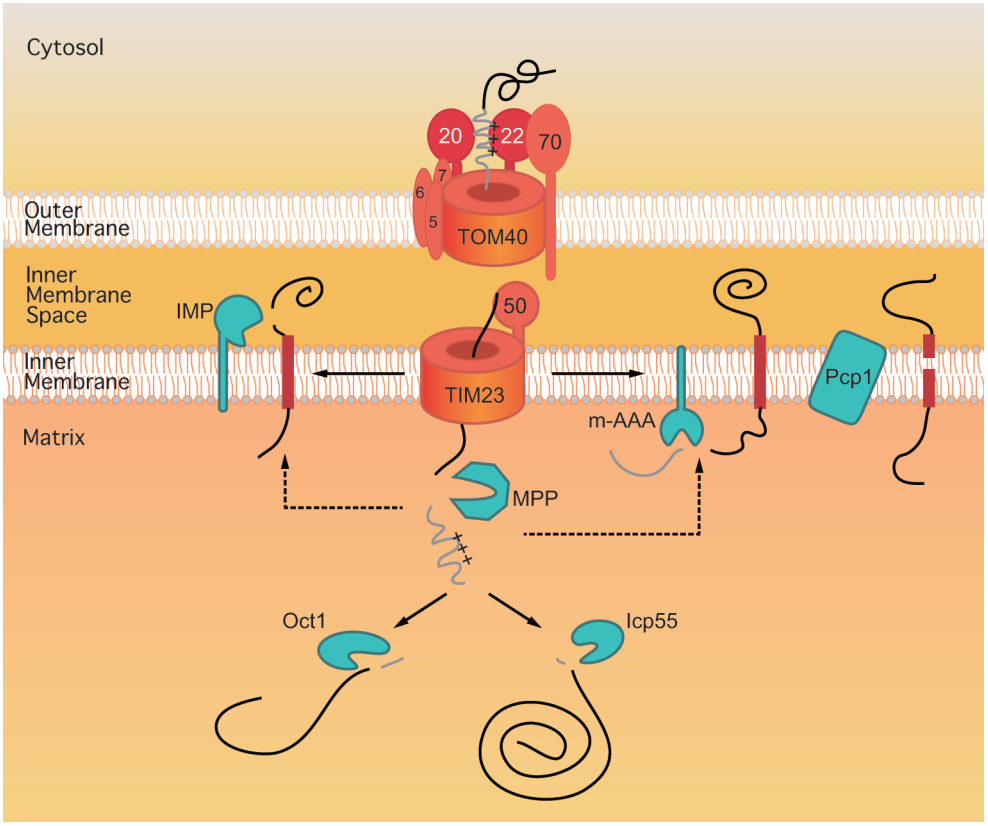


Figure 1.2: Proteases in a yeast mitochondrion.

Chapter 2

Sequence divergence of targeting signals

In this section I describe work published in BMC Genomics

2.1 Materials and methods

2.1.1 Sorting signal classes

I mainly focused on the two most common N-terminal sorting signals: *Signal Peptide*s (SP), targeting proteins to the endoplasmic reticulum and *Matrix Targeting Signal*s (MTS) which target proteins to the matrix (inner compartment) of the mitochondria. In the plant dataset, I also consider *Chloroplast Transit Peptide*s (CTP). All of these signals reside near the N-terminus but in

general have different properties and are effectively discriminated by the cell. In some cases however, the N-terminal "signal" can be ambiguous. In particular many examples are known in which the same amino acid sequence directs some copies of a protein to the mitochondria and others to the chloroplast [25, 26]. Nevertheless these examples still constitute only a small percentage of proteins and therefore I simplify the analysis by treating N-terminal sorting signal identification as a simple three- or four-way classification problem: {MTS, SP, (CTP), no signal}. Other types of N-terminal sorting signals exist, for example the PTS2 signal targeting proteins to the peroxisome [27], but the number of proteins using such signals is much smaller than those using the SP, MTS or CTP signals.

The sorting signal class labels I use in the datasets are partially based on direct experimental evidence. In the dataset of *S.cerevisiae*, I used UniProtKB/Swiss-Prot [28] to assign localization class labels, augmented by MTS containing proteins determined in the proteomics experiment of Vögtle et al. [3]. Because only a small number of SP's have been directly confirmed experimentally, I also included proteins whose SP is inferred in the database and predicted positive by SignalP [29]. I used proteins annotated to localize to the cytosol or nucleus as proteins without N-terminal signals. To reduce bias in training and accuracy estimation, I used BLASTClust 2.2.22 [30] to remove redundant sequences with a setting of 20% identity. For proteins in human and a few plant species I adopted the dataset of Predotar [31] and for plants augmented that small number by experimental proteomics data determined in the mass spectrometry experiment of Huang et al. [32].

2.1.2 Dataset

Organisms used

I gathered protein sequences from 11 relatively diverse and well annotated representative species of the three phylogenetic divisions: yeast, mammal and plant respectively (Table 2.1). The 11 mammal species and most of the plant species are annotated reference proteomes in UniProt, but a few of the plant species are only included in UniProt as complete, but not fully annotated, proteomes. Note that "plant" dataset contains the unicellular green algae *Chlamydomonas reinhardtii*, which is not a typical plant but is classified in the "viridiplantae" kingdom.

In each of the three divisions I designated one species as the "reference" species. I used information in proteins from the non-reference species only for computation of sequence divergence (via ortholog multiple sequence alignments). I chose *S.cere.*, *H. sapiens*, and *A. thaliana* as the reference species for yeast, animals and plants respectively, because they have the most complete annotation. However for plants even *A. thaliana* has rather limited annotation of SPs, so in order to increase the plant dataset size I used other species as the reference species in some cases.

Ortholog Determination

I performed some experiments on hand curated ortholog sets downloaded from the Yeast Gene Order Browser (YGOB) [33], but also computed ortholog sets for each of the three phylogenetic divisions.

Automatic identification of orthologs is a complex subject for which many sophisticated methods have been developed, the most suitable one being application dependent [34]. For this study, I adopted a simple procedure based on reciprocal best hits (RBHs) [35]. Formally, proteins P and P' from species S and S' respectively, are RBHs if P is more similar to P' than any other protein in

S' and P' is more similar to P than any other protein in S. I define the ortholog set of a reference species protein as all of its RBHs. When computing RBHs it is important that proteins from as many organisms as possible are included; but in the end I only have use for those ortholog sets in which the reference species is annotated, so in general I discarded the rest. However, in the case of plant, I attempted to rescue those discarded sequences by also trying *O. sativa*, *G. max* and *C. reinhardtii* in turn as the reference species.

In computing the similarity scores for RBH I chose to use global alignment rather than local alignment. Motivation for this was: 1) sorting signals often appear on the N- or C-terminal region of proteins, so differences in those regions may indicate a different localization of the "ortholog", and 2) for multiple domain proteins, strong similarity in one domain may not imply the same localization site (or signal). I used the heuristic but fast USEARCH [36] program with its default parameters to compute the global similarity scores. Table 2.2 summarizes the datasets.

Multiple Alignment

I computed multiple alignments for each of the 4 orthologs sets (1 curated and 3 automatic) by aligning with the MAFFT program [37], using "LINSI", its most accurate mode. Hereafter, I denote these alignments as "orthoMSA" in general, and as "autoOrthoMSA" when specifically referring to multiple alignments of automatically generated ortholog sets. The number of sequences in the automatically generated ortholog sets generally differs from the YGOB based sets, however, it seems that the distribution of the divergence score stabilizes when the number of sequences exceeds three (Figure 2.1), therefore I decided to include ortholog sets with at least four sequences.

S. cerevisiae	*H. sapiens*	*A. thaliana*
Saccharomyces castellii	*Gorilla gorilla*	*Glycine max*
Saccharomyces kluyveri	*Otolemur garnettii*	*Ricinus communis*
Kluyveromyces waltii	*Mus musculus*	*Populus trichocarpa*
Ashbya gossypii	*Oryctolagus cuniculus*	*Vitis vinifera*
Candida glabrata	*Sus scrofa*	*Sorghum bicolor*
Kluyveromyces lactis	*Ailuropoda melanoleuca*	*Brachypodium distachyon*
Zygosaccharomyces rouxii	*Myotis lucifugus*	*Oryza sativa*
Kluyveromyces thermotolerans	*Loxodonta africana*	*Selaginella moellendorffii*
Saccharomyces bayanus	*Sarcophilus harrisii*	*Physcomitrella patens*
Kluyveromyces polysporus	*Ornithorhynchus anatinus*	*Chlamydomonas reinhardtii*

Table 2.1: List of species used to define orthologs in each phylogenetic category. The species listed at top are the reference species used to determine the subcellular localization site class labels. In the case of plants, one of *G. max*, *O. sativa* and *C. reinhardtii* were used as the reference species for proteins for which no annotation was available in *A. thaliana*.

Localization class	*S.cere.* curated orthologs	*S.cere.* RBH	*H.sapiens* RBH	Plants RBH
MTS	179	219	81	61
SP	53	73	169	15
CTP	N/A	N/A	N/A	97
N-signal-free	450	560	415	99

Table 2.2: For each ortholog dataset, the number of ortholog sets in each localization class is listed. RBH orthologs are defined by the reciprocal best hit method.

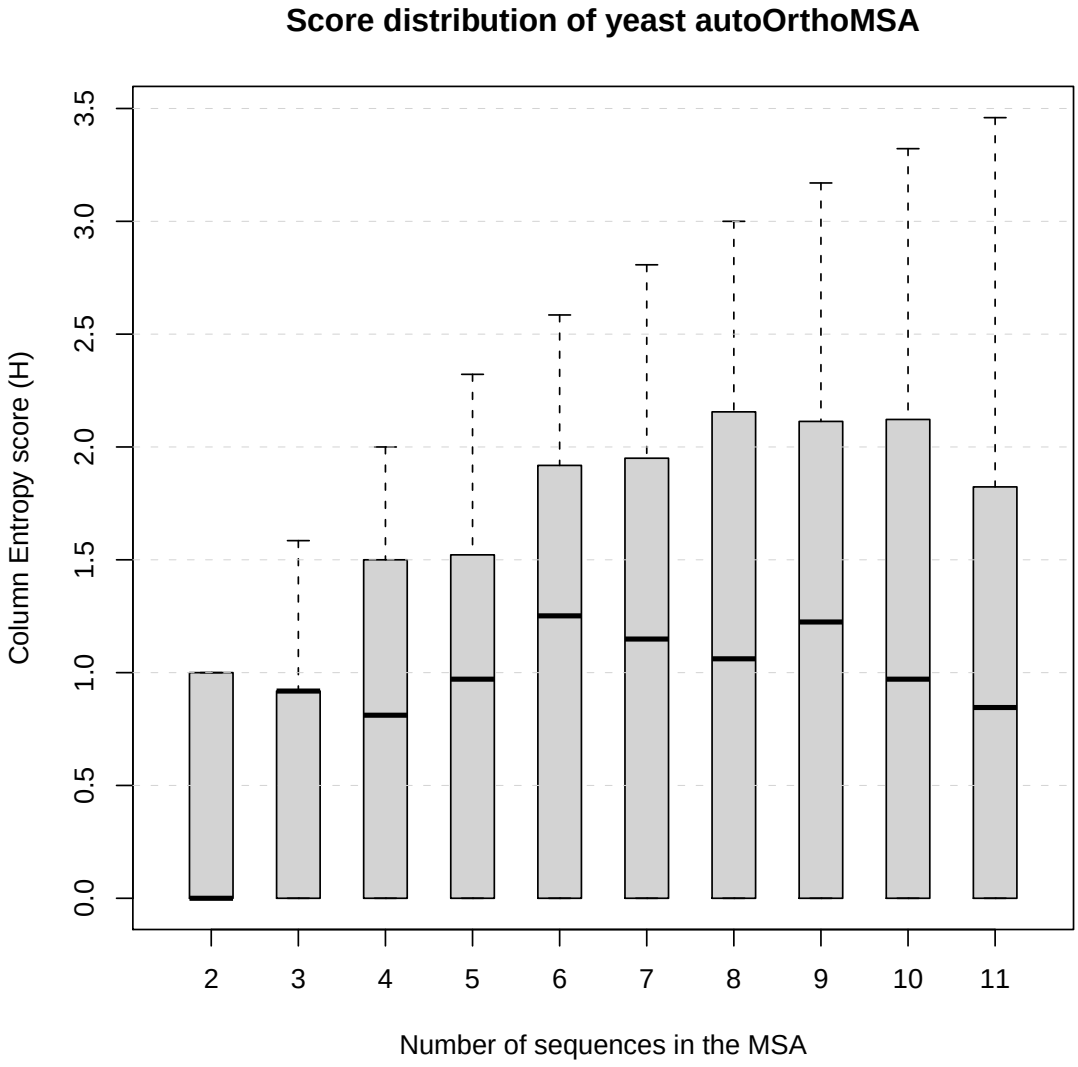


Figure 2.1: Relationship between mean divergence score and the number of sequence in MSA's. A box plot illustrating the mean, quartiles and range of the column entropy score for MSA's in the yeast autoOrthoMSA dataset partitioned by the number of sequences in the MSA is shown.

2.1.3 Features for classification

Column entropy score

Several measures have been suggested for scoring evolutionary sequence conservation (or conversely divergence) [38, 39]. Here I adopt a simple Shannon entropy based score. The Shannon entropy $H(i)$ of the ith column of an orthoMSA is defined as:

$$H(i) = -\sum_{j \in A} F(i,j) \log_2 F(i,j). \tag{2.1}$$

where A denotes the set of 20 amino acid characters plus gap characters, and $F(i,j)$ denotes the frequency of character j in column i of an orthoMSA. Note that when multiple gap characters are present in a column, I consider each to be a unique character. 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, because I treat the two '-' characters as distinct. I adopted this treatment of gap characters so that the divergence of orthoMSA columns with many gaps is considered high (I also tried using straight entropy, but the results, not shown, were slightly worse). The range of this divergence score runs from 0 to $\log_2 n$, where n is the number of sequences.

Divergence based features

For many orthoMSA's, the entropy often varies widely from column to column. Therefore, I defined a number of evolutionary divergence features based on a smoothed entropy score, $\bar{H}_{i,j}$, defined as the average entropy score for columns in the interval $[i,j]$. For example I define the local divergence (LD) of an orthoMSA at position k as $\bar{H}_{k-10,k+10}$. Another feature I defined is NCdiff, the average difference in divergence between in the first 20 residues and residues 80 to 99. Motivation for this

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)

Table 2.3: Smoothed entropy derived features are listed. Quantities shaded in grey were not used directly as features.

definition was the hope that subtracting the divergence from residues 80 to 99 would approximately normalize the feature when comparing proteins with different overall rates of evolution. These features are summarized in table 2.3.

Physico-chemical propensities

To explore the possibility of combining sequence divergence with standard features used in protein localization prediction, I defined three features computed from the first 20 or 40 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 [40] index (Hphob).

Amino acid composition

Amino acid composition is another standard feature for protein localization. I tested this feature computed on the first 20 residues, the first 40 residues, and the entire protein sequence.

2.1.4 Classifiers

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.

J48

J48 is a version of the C4.5 decision tree induction algorithm of Quinlan [41, 42], implemented in the Weka software package [43]. I used the default value of 0.25 for the confidence factor, which controls the complexity of the induced tree.

Support Vector Machine

The Support Vector Machine (SVM) [44] 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 (details is described in next paragraph). In this project, I used the LIBSVM implementation [45]. I used the Gaussian radial basis kernel function with default γ value (1.0 / # number of features). I used 50.0 for the SVM cost parameter C, because with the default cost parameter (1.0) prediction by RBF kernel failed for some features. In this study I 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 among those SVM's is used when classifying examples [46].

Hard margin case: SVM classifies given vectorized samples x into either binary class $C_{+1}(y \geq 0)$ and $C_{-1}(y < 0)$ by below formulation.

$$y = \texttt{sign}(\mathrm{w}^T \mathrm{x} - b) \tag{2.2}$$

Here, w is parameter for each feature, and b is an intercept. If given training data $s_i \in S$ ($s_i = \{\mathrm{x}_i, y_i\}$, x_i is a vector of sample and y_i is a label either +1 or -1) is assumed to be a linearly separable set, below formulation is considered.

$$y_i(\mathrm{w}^T \mathrm{x_i} - b) \geq 1 \tag{2.3}$$

$\mathrm{w}^T \mathrm{x_i} - b = 1$ is a hyperplane which includes support vectors, samples closest to discriminative hyperplane, of class C_{+1}, and $\mathrm{w}^T \mathrm{x_i} - h = -1$ is one which includes support vectors of class C_{-1}. In this conditions, margin can be written as a distance $1/\|\mathrm{w}\|$ between a discriminative hyperplane and planes on which support vectors exist. To maximize the margin, below objective function should be minimized about w.

$$\arg\min_{(\mathrm{w},b)} L(\mathrm{w}) = \frac{1}{2}\|\mathrm{w}\|^2 \tag{2.4}$$

subjected to (eq. 2.3) for $i = 1, ..., n$.

This objective function can be written as a dual form with Lagrange multipliers α, and maximize it about α.

$$\arg\max_{(\alpha_i)} L(\alpha) = \sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{2.5}$$

subject to $\alpha_i \geq 0$ for $i = 1, ..., n$.

Soft margin case: In practical situations such as this work, it is hard to assume complete linear

separation of S, therefore, soft-margin was suggested [44]. Non-negative slack variables, ξ_i, is introduced to measure the degree of misclassification for data s_i.

$$y_i(\mathrm{w}^T \mathrm{x_i} - b) \geq 1 - \xi_i \tag{2.6}$$

Then, the objective function (eq. 2.4) is written as below.

$$\arg\min_{(\mathrm{w},\xi_i,b)} L(\mathrm{w}) = \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{2.7}$$

subjected to (eq. 2.6) and $\xi_i \geq 0$ for $i = 1, ..., n$. C is a hyper-parameter so-called cost parameter, which determines tradeoff between a large margin and label error.

Similarly to the hard margin case, the objective function (eq. 2.7) can be written as a dual form with introducing Lagrange multipliers α, β for the subjectives.

$$\arg\max_{(\alpha_i)} L(\alpha) = \sum_{i=1}^{n}\alpha_i + \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j \tag{2.8}$$

subject to $C \geq \alpha_i \geq 0$ for $i = 1, ..., n$ and $\sum^n \alpha_i y_i = 0$.

Another technique in SVM to accept non-linear separable case is applying the kernel trick. Since feature vector x appears as only a dot product form in (eq. 2.5), this term is replaced by a non-linear kernel functions, $k(\mathrm{x}_i, \mathrm{x}_j)$. This enables the algorithm to fit the maximum-margin hyperplane in a transformed feature space, usually high dimensional and non-linear. Gaussian radial basis kernel function is below.

$$k(\mathrm{x}_i, \mathrm{x}_j) = \exp(-\gamma\|\mathrm{x}_i - \mathrm{x}_j\|^2) \tag{2.9}$$

γ is a non-negative kernel parameter.

Measuring the influence of divergence features: As reported in the results section, I performed a post-hoc analysis of proteins for which the divergence features greatly influenced the prediction outcome. To do this I needed to compare 6 numbers (three SVM scores {MTS vs SP, MTS vs none, SP vs none} each computed with and without the divergence features) into a measure of how much the divergence features influenced the prediction. Because the SVM scores are not given directly as probabilities and each individual SVM addresses a different subset of classes, it is not trivial to derive a well-principled way to do this. As described in more detail in the appendix, I chose to define this in terms of exponential loss-based decoding [47]. I do not claim that this is necessarily the best measure, but it appears to give reasonable results. Fortunately, for my purposes it is enough that truly large differences are assigned in a roughly suitable order.

2.1.5 Quantifying feature importance

I used the so called "information gain" to quantify the importance of each feature. Information gain 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:

$$I(C, F) \quad = \quad H(C) - H(C|F). \tag{2.10}$$

where C and F denote class and feature respectively. $H(C)$ the denotes information theoretic entropy of the overall distribution of the class labels, while $H(C|F)$ denotes the conditional entropy of the class label when feature F is given. A larger information gain indicates greater predictive power. Because the divergence based features have a large number of possible values, I first binned those values into a smaller number by the method of Fayyad & Irani [48].

2.1.6 Classification performance evaluation

Accuracy is not always the most meaningful measure of performance for skewed datasets (i.e. datasets with a very uneven number of examples from different classes) [49]. Therefore I report several measures in addition to accuracy.

Matthews correlation coefficient

The Matthews correlation coefficient, MCC [50, 51], is a measure of performance for binary classification defined as follows:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{2.11}$$

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 of the majority class classifier is identically zero, as is the expected value of MCC under random prediction.

Area under the ROC curve

The Area under the curve (AUC) for a receiver operating characteristics (ROC) graph is a widely used metric to evaluate binary classification accuracy [52]. The usual way to generate an ROC plot is to rank instances by their predicted scores with increasing threshold values, plotting true positive rate (y-axis) versus false positive rate (x-axis). AUC ranges from 0 to 1.0, with perfect prediction yielding 1.0 and perfectly wrong prediction 0.0. AUC can be interpreted as the probability that a classifier is able to distinguish a randomly chosen positive example from a randomly chosen negative

example [53]. For this task, the majority class classifier gives no information over coin flipping and therefore can be considered to yield an AUC of 0.5.

2.2 Results

2.2.1 Feature Analysis

N-terminal sorting signals are evolutionary divergent

It is well known that N-terminal sorting signals exhibit relatively low sequence conservation [12]. As shown in Figure 2.2, this phenomenon is particularly clear for the mitochondrial heat shock protein, mtHSP70, in which the main part of the protein is highly conserved but the N-terminal region is highly divergent. Figure 2.3 quantifies this trend for the proteins in the YGOB ortholog set.

Estimate of importance of each feature

As a rough estimate of feature importance, I computed the information gain for each feature (Figure 2.4). The two highest scoring features are the physico-chemical features #neg and Hphob, but the LD features near the N-terminus also show information gain significantly greater than zero.

Sequence divergence is not redundant to physico-chemical trends or amino acid composition

To be promising as a feature for prediction, it is desirable that evolutionary sequence diversity not be perfectly correlated with other features. To investigate this I plotted LD(13), the divergence feature with the highest information gain, against Hphob, #neg and the arginine composition (the

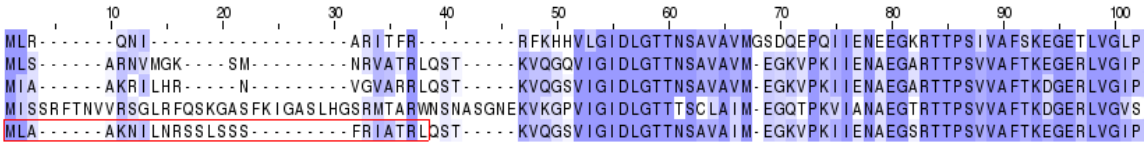


Figure 2.2: An example of a divergent MTS. A multiple sequence alignment of the protein mtHSP70 (UniProt accession P0CS90) and its orthologs from five species of yeast is shown. The red box indicates the cleaved MTS in *S.cere.*. Conserved positions are colored by Jalview.

	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 ± 0.00
Majority class fraction	65.98%	N/A	N/A

Table 2.4: Three classification performance measures when using only divergence features are shown for the discrimination of N-signal containing and N-signal-free proteins (yeast curated ortholog sets). 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.

three highest scoring standard features in the 40 residue N-terminal region) (Figure 2.5). Although there may be some relationship, the feature pairs do not appear highly correlated.

2.2.2 Divergence predicts the presence of N-terminal signals

I 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 2.4, for this binary classification task, sequence divergence *alone* allows for significantly higher prediction accuracy than randomized control experiments or the majority class fraction (66.0%) in the yeast dataset.

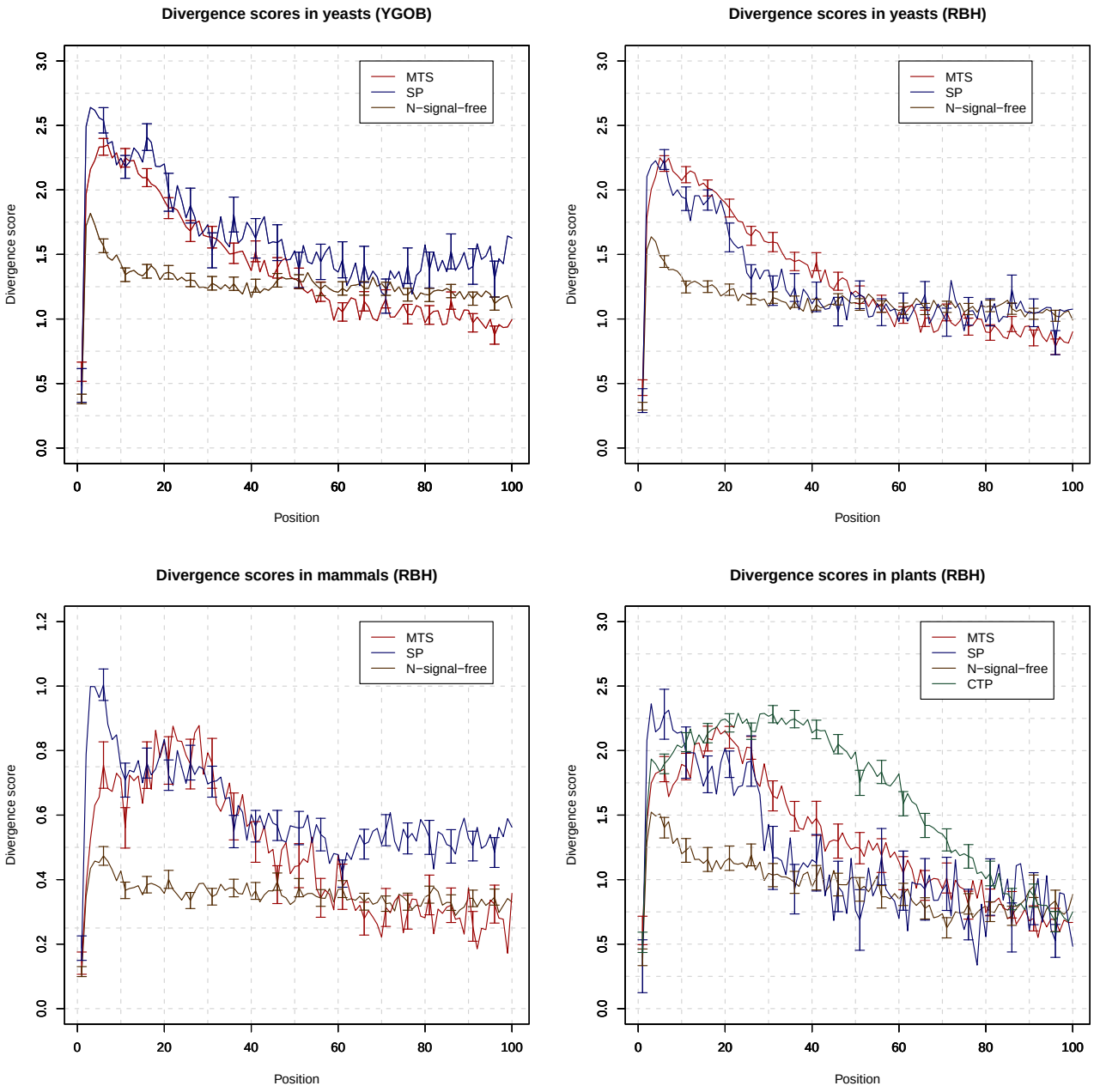


Figure 2.3: Local divergence score over N-terminal region. Average local divergence scores are shown for the 100 residue N-terminal region of: MTS containing, SP containing, and N-signal-free proteins. Top left panel is calculated from orthologs of yeast curated dataset, and the others from automatically collected orthologs. For the plant dataset, CTP containing proteins are also shown. The error bars denote standard error. For clarity, error bars are only shown for every fifth position.

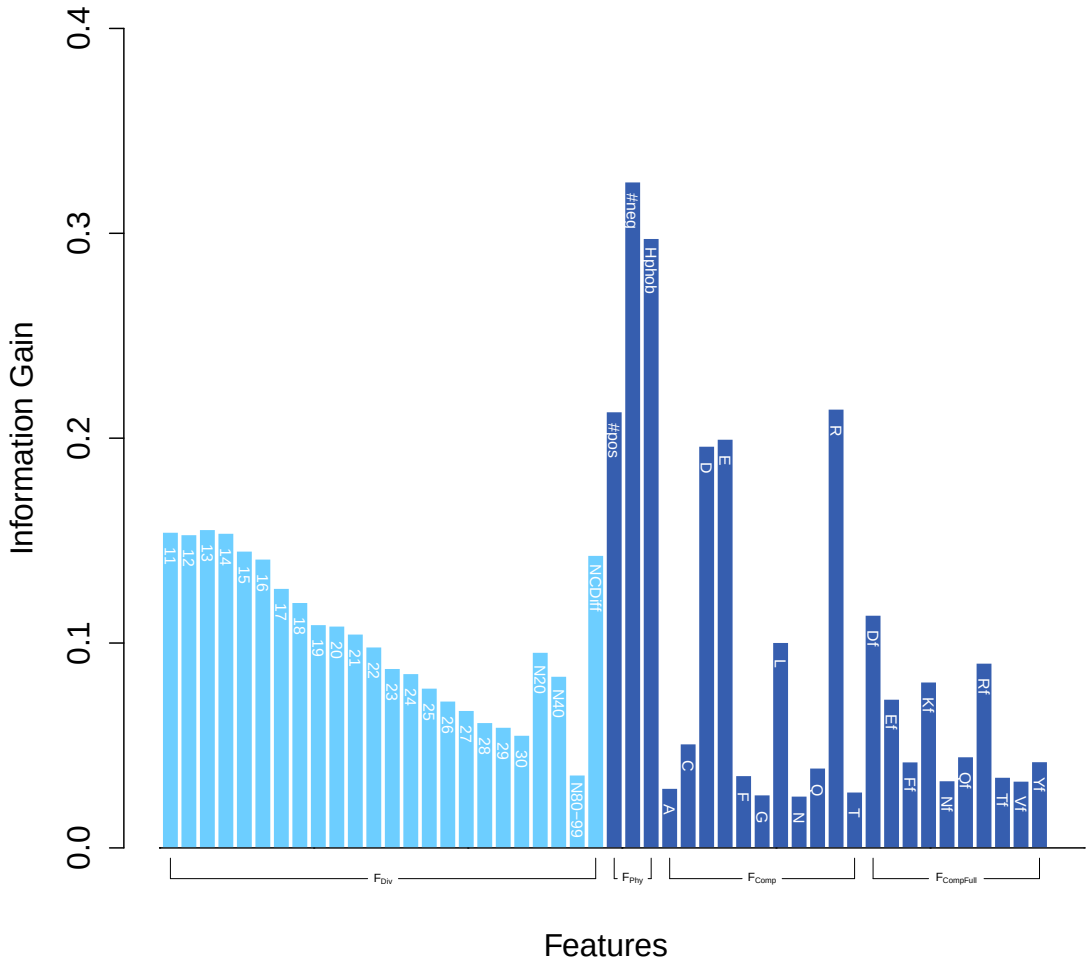


Figure 2.4: Importance of each feature. The importance of each feature as estimated by information gain is shown for the YGOB ortholog set. At left, the divergence related scores are shown by light blue color lines. For local divergence features LD(i), only the residue number i is listed. Dark blue colored lines denote standard features of the N-terminal 40 residues such as physico-chemical properties or amino acid composition. The suffix "f" denotes amino acid composition from the full length of the protein.

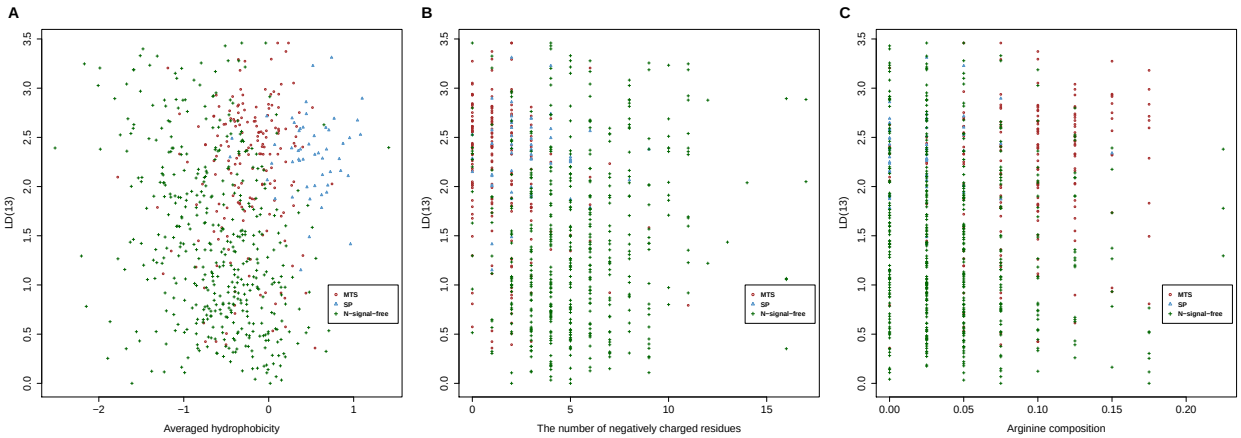


Figure 2.5: Scatter plots between divergence score and standard features. Scatter plots of LD(13) (on the vertical axis) *vs.* #neg, Hphob and arginine composition on the horizontal axis are shown for the YGOB ortholog set. MTS proteins are shown in red, SP in blue and N-signal-free proteins in green.

2.2.3 Divergence distinguishes SP vs. MTS vs. N-signal-free

Although the sequence divergence profile of SP's and MTS's appear similar when averaged (Figure 2.3), I 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 2.5 the performance with divergence features is slightly better than the majority class fraction (66.0%) and also slightly improves the performance when added to the physico-chemical features in N-terminal 40 residues or amino acid composition in either N-terminal 40 or full length (appendix).

The ratio of examples in the dataset is 8.5 : 3.4 : 1, for N-signal-free, MTS and SP containing proteins respectively. Skewed datasets are known to complicate both learning and performance evaluation [49]. Therefore I also measured performance on a dataset with uniform class occupancy, created by randomly discarding all but 53 proteins from each class. As shown in Table 2.6, in this experiment the divergence feature only performance (63%) is much higher than the majority class fraction (33%), and the divergence features also contribute more to the performance when combined

	Divergence		Classical features		Combination	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.67 ± 0.03	0.36 ± 0.06	$\mathbf{0.87} \pm 0.03$	0.76 ± 0.05	$\mathbf{0.87} \pm 0.03$	$\mathbf{0.77} \pm 0.03$
SP	0.50 ± 0.00	0.00 ± 0.00	0.81 ± 0.08	0.70 ± 0.11	$\mathbf{0.90} \pm 0.06$	$\mathbf{0.83} \pm 0.07$
N-signal-free	0.66 ± 0.02	0.36 ± 0.03	0.85 ± 0.03	0.72 ± 0.05	$\mathbf{0.87} \pm 0.02$	$\mathbf{0.77} \pm 0.03$
% accuracy	70.82 ± 1.61		87.24 ± 1.86		$\mathbf{89.30} \pm 0.66$	

Table 2.5: The 5-fold cross-validation performance of an SVM classifier using: divergence features only, classical features only, and the two combined; is shown for three-way classification on the yeast curated ortholog dataset. Classical features are computed based on the N-terminal 40 residues.

with the standard features (Table 2.6).

I further tested the prediction power of divergence features when combined with classical features computed on a 20 residue N-terminal instead of 40 (which might be too long for the SP class). In this experiment, divergence features improved the performance only slightly when combined with other standard features (Table 2.7). I also computed the confusion matrix for this dataset (Table 2.8) and the other datasets investigated in the study (appendix, tables A.14–A.25).

	Divergence		Classical features		Combination	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.67 ± 0.10	0.35 ± 0.20	0.84 ± 0.07	0.68 ± 0.13	$\mathbf{0.88} \pm 0.05$	$\mathbf{0.78} \pm 0.09$
SP	0.71 ± 0.09	0.41 ± 0.16	0.92 ± 0.05	0.85 ± 0.10	$\mathbf{0.94} \pm 0.01$	$\mathbf{0.88} \pm 0.03$
N-signal-free	0.79 ± 0.07	0.60 ± 0.13	0.78 ± 0.09	0.57 ± 0.18	$\mathbf{0.86} \pm 0.07$	$\mathbf{0.74} \pm 0.13$
% accuracy	62.86 ± 5.84		79.92 ± 5.54		$\mathbf{86.19} \pm 4.67$	

Table 2.6: The 5-fold cross-validation performance of an SVM classifier using: divergence features only, classical features only, and the two combined; is shown for three-way classification on a balanced dataset (53 proteins from each class, yeast curated orthologs).

	Divergence		Classical features		Combination	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.67 ± 0.03	0.36 ± 0.06	$\mathbf{0.89} \pm 0.02$	0.80 ± 0.02	$\mathbf{0.89} \pm 0.01$	$\mathbf{0.81} \pm 0.02$
SP	0.50 ± 0.00	0.00 ± 0.00	0.97 ± 0.03	0.92 ± 0.07	$\mathbf{0.98} \pm 0.03$	$\mathbf{0.97} \pm 0.04$
N-signal-free	0.66 ± 0.02	0.36 ± 0.03	$\mathbf{0.90} \pm 0.01$	0.81 ± 0.02	$\mathbf{0.90} \pm 0.01$	$\mathbf{0.83} \pm 0.02$
% accuracy	70.82 ± 1.61		91.49 ± 1.26		$\mathbf{92.23} \pm 1.25$	

Table 2.7: The 5-fold cross-validation performance of an SVM classifier using: divergence features only, classical features only, and the two combined; is shown for three-way classification on the entire yeast curated ortholog dataset. Classical features are calculated from N-terminal 20 amino acids.

	Divergence			Classical features			Combination		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	83	0	96	148	1	30	144	0	35
SP	16	0	37	0	50	3	1	51	1
N-signal-free	50	0	400	20	4	426	15	1	434

Table 2.8: Confusion matrix of the 5-fold cross-validation performance of an SVM classifier using: divergence features only, classical features only, and the two combined; is shown for three-way classification on the entire yeast curated ortholog dataset. Classical features are calculated from N-terminal 20 amino acids.

2.2.4 Divergence computed from automatically generated ortholog sets is consistent with the hand curated dataset.

Although the YGOB based dataset convincingly demonstrates that the divergence score has discriminative power for N-terminal signal prediction, it covers only 11 yeast species and requires hand curation. Thus as described in the Methods section, in this work I adopted a simple procedure based on reciprocal best hit relationships to obtain automatically generated ortholog sets as well (Table 2.2).

In yeast, the average divergence score at each positions is similar to the score from the YGOB ortholog set, and the overall tendency looks similar for animals and plants (Figure 2.3). Interestingly, CTP shows a high and longer region of elevated divergence, consistent with previous observations

that CTPs tend to be longer than MTSs [32]. Additionally, I note that the score range of the human autoOrthoMSA's is significantly different from those of yeast or plants. This is expected because divergence amongst yeast sequences is at least as large as that of the chordates [54], so divergence in mammals should be smaller.

2.2.5 Divergence computed from autoOrthoMSA also predicts N-terminal signals

First, I confirmed whether or not divergence features can be applied to a simple binary classification: discrimination between N-terminal signal containing proteins and N-signal-free proteins. Although the ratio of positive to negative examples in each dataset differs, the result of prediction by divergence features alone is higher than majority class classifier for all datasets (Table 2.9).

Next, I tested the predictive power of divergence in three-way classification on a dataset balanced to have equal class frequency (Table 2.10). It is evident that on balanced datasets, divergence also shows significant predictive power in distinguishing between the two different kinds of N-terminal signals, even for the relatively closely related mammal species.

In plants, the divergence score can also discriminate between the three possible kinds of N-terminal signals better than random. However, there are only 15 experimentally validated SPs in this phylogenetic category (Table 2.2). Since this small sample size leads to a high statistical variance, I also computed the performance on balanced 3-way classification of MTS vs CTP vs N-signal-free (Table 2.11).

In the appendix I list cross-validated performance estimates on various combinations of datasets and features. From these I draw two conclusions: in most cases divergence features slightly improve prediction when combined with standard features and in general computing standard features on

Yeast dataset	mean accuracy	mean AUC	mean MCC
J48	71.47 ± 5.00	0.67 ± 0.07	0.36 ± 0.12
SVM	$\mathbf{75.35} \pm 3.49$	$\mathbf{0.71} \pm 0.04$	$\mathbf{0.44} \pm 0.08$
Majority class fraction	65.23%	N/A	N/A
Human dataset			
J48	69.32 ± 4.10	$\mathbf{0.72} \pm 0.07$	$\mathbf{0.43} \pm 0.09$
SVM	$\mathbf{72.28} \pm 5.95$	$\mathbf{0.72} \pm 0.06$	$\mathbf{0.43} \pm 0.12$
Majority class fraction	62.41%	N/A	N/A
Plant dataset			
J48	79.41 ± 6.03	0.75 ± 0.06	0.55 ± 0.13
SVM	$\mathbf{83.47} \pm 4.01$	$\mathbf{0.79} \pm 0.04$	$\mathbf{0.64} \pm 0.09$
Majority class fraction	63.60%	N/A	N/A

Table 2.9: Three classification performance measures when using only divergence features are shown for the discrimination of N-signal containing and N-signal-free proteins on automatically collected orthologs. AUC denotes the area under the ROC curves. For each measure the average and standard deviation is shown over the 5 folds of the cross-validation.

the N-terminal 20 residues leads to higher accuracy than computing on 40 residues.

	F_{Div} Yeast (73)		F_{Div} Human (81)	
	AUC	MCC	AUC	MCC
MTS	0.65 ± 0.09	0.31 ± 0.18	0.66 ± 0.05	0.31 ± 0.11
SP	0.60 ± 0.07	0.19 ± 0.14	0.70 ± 0.08	0.40 ± 0.15
N-signal-free	0.66 ± 0.08	0.35 ± 0.15	0.69 ± 0.06	0.39 ± 0.11
% accuracy	51.63 ± 7.21		57.61 ± 4.71	

Table 2.10: The 5-fold cross-validation performance of an SVM classifier using divergence features is shown for three-way classification on the automatically generated ortholog dataset for yeasts and mammals. The number of examples is given in parenthesis at top.

2.2.6 Post-hoc analysis of proteins for which divergence strongly influences the prediction result

In this section I discuss proteins for which the use of divergence features strongly affects the results.

The ortholog MSA's of all proteins mentioned in this section are available in Table S_MSAs.

	F_{Div} Plant 4 classes (15)		F_{Div} Plant 3 classes (61)	
	AUC	MCC	AUC	MCC
MTS	0.62 ± 0.11	0.24 ± 0.21	0.66 ± 0.08	0.35 ± 0.14
SP	0.78 ± 0.11	0.58 ± 0.23	N/A	N/A
CTP	0.73 ± 0.16	0.43 ± 0.31	0.77 ± 0.12	0.51 ± 0.23
N-signal-free	0.80 ± 0.14	0.72 ± 0.20	0.81 ± 0.09	0.67 ± 0.13
% accuracy	60.00 ± 9.13		66.22 ± 10.11	

Table 2.11: The 5-fold cross-validation performance of an SVM classifier using divergence features is shown for three-way classification on balanced sets of (automatically generated) plant orthologs with or without the SP class. The number of examples is given in parenthesis at top.

Divergence features may help flag misannotation

Prior to this work, evolutionary divergence has not been applied systematically to N-terminal signal prediction. However I expected that it might be able to capture interesting examples not revealed by other features. To investigate this, I ranked instances whose SVM prediction changes drastically depending on whether or not divergence features are used. Because of its rich annotation, I focused on *S.cere.*, using the automatically defined ortholog set. The prediction result of 43 proteins changed depending on whether divergence features were added to conventional features. For these 43 proteins, I used the SVM numerical scores to rank the size of the effect as explained in the appendix (ranked list in Table A.1). In general, prediction differences are observed between the MTS and N-signal-free classes. The most highly affected protein is mitochondrial alanine tRNA ligase, ALA1 (P40825), which is predicted to have an MTS when sequence divergence features are used. Upon closer inspection I discovered that the sequence I used for this protein should in fact have been labeled as an MTS containing protein, but the dataset based on an earlier version of UniProtKB/Swiss-Prot contained mistaken annotation which holds for an alternative translation start site. Thus in this case sequence divergence yields the correct answer.

PTP1 (P25044) is another protein whose prediction changes from N-signal-free to MTS when

divergence is considered. Following UniProtKB/Swiss-Prot, I treated it as a cytoplasmic protein, but there is no reference given for this annotation. Moreover PTP1 is identified as a mitochondrial protein by two large-scale experiments. This is suggestive that it may have a mitochondrial localization, although even in that case it would not necessarily have an MTS. Hopefully future work will clarify if this is another case in which divergence features flagged misannotations in the dataset.

Divergence features may help detect mitochondrial proteins with non-classical MTS signals

FMP52 (P40008) is a protein included in the dataset for which the SVM with standard features predicts an MTS but the SVM with divergence features predicts N-signal-free. As shown in Figure 2.6, FMP52's N-terminal region is not divergent like typical MTS's, especially very near the N-terminus. FMP52 is indeed a mitochondrial protein, but upon closer scrutiny I discovered a previous report that it strongly associates with the outer membrane [55] — and therefore is unlikely to have a matrix targeting MTS. Moreover, FMP52 is one of the non-MTS containing proteins in the yeast proteomic analysis [3]. Swiss-Prot does annotate FMP52 with an MTS (1–44), but I could not find a reference or supporting information for this MTS annotation; therefore, I conclude that it is unlikely to have MTS. CYM1 (P32898) is another interesting example which has been reported to localize in the intermembrane space and not to be processed by mitochondrial proteases [56]. Since MTS is a cleavable targeting signal for the matrix, the intermembrane space localization and lack of proteolytic cleavage of CYM1 suggests its N-terminal signal is not a typical classical MTS.

MrpL19 (P53875) is another case in which sequence divergence features highlight a ribosomal mitochondrial protein which does not appear to have a classical MTS signal. According to both

UniProtKB/Swiss-Prot annotation and a large-scale proteomics experiment [3] MrpL19 has an MTS, but the annotated "MTS" is unusually long and lacks an arginine in position -2, which is normally observed in MPP cleavage sites [13]. Moreover the N-terminal sequence of MrpL19 is very well conserved not only in yeasts but even in bacteria. Indeed the three dimensional structure of rplK, a homolog of MrpL19 in *E.coli*, has been solved and it is evident that the two proteins have a similar structured N-terminal. Taken together the evidence suggests that MrpL19 may not have an N-terminal mitochondrial localization signal, but rather be imported via an alternative pathway.

On the other hand, I also observed ribosomal mitochondrial proteins whose N-terminal is poorly conserved. One example is MrpL32 (P25348), which cannot be predicted as having an MTS by standard tools such as TargetP [57] or Predotar [31], nor by SVM's trained without divergence features. MrpL32 shows a high divergence in its N-terminal region (Figure 2.7 and is predicted to have an MTS by SVM when using divergence features. A literature search revealed that MrpL32 does indeed have an MTS, but it is unusual in the sense that it is cleaved by the protease m-AAA [23, 58] instead of MPP. Mrp7 (P12687) is a similar case. Like MrpL32, Mrp7 is also a component of a large ribosomal subunit and is not predicted to have an MTS by TargetP, Predator, nor by SVM without divergence features, but is predicted to have an MTS when divergence features are used. In UniProtKB/Swiss-Prot, Mrp7 is annotated as having an MTS, and indeed the processing of Mrp7 by MPP has been reported multiple times [59, 3]. So in this case high sequence divergence allows an MTS to be correctly predicted.

Another case worth discussing is IMO32 (P53219), which has recently been reported to be processed by the intermediate protease Oct1 (after MPP) in the matrix [60]. It is unusual in that its inferred MPP cleavage site represents a rare exception to the almost invariant presence of arginine at the -2 position. IMO32 is predicted as an MTS by Predator [31] and SVM when I use

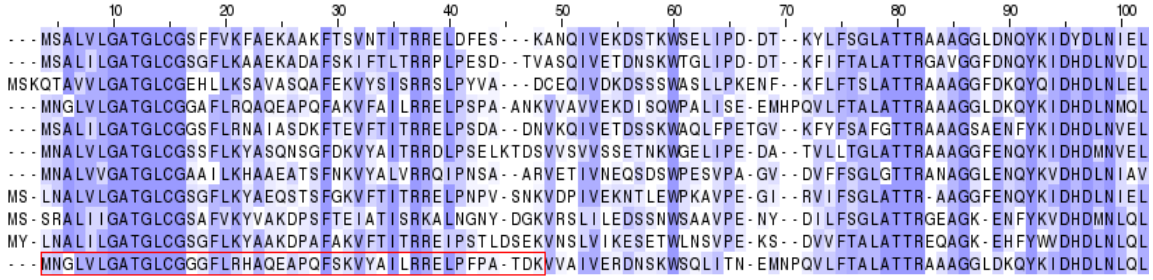


Figure 2.6: Multiple sequence alignment of FMP52 in *S.cere.*and its orthologs in 10 other yeast species. The red boxed region shows the annotated MTS of FMP52. Conserved positions are colored by Jalview.

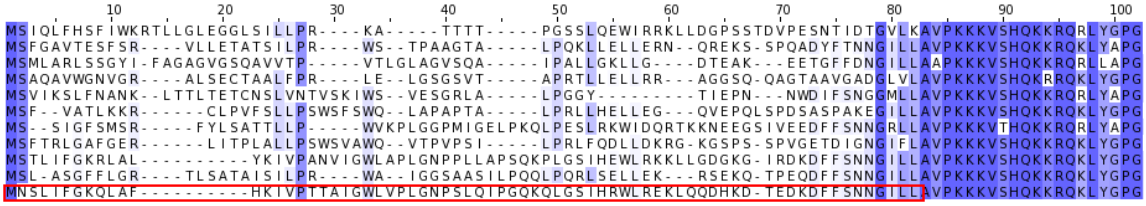


Figure 2.7: Multiple sequence alignment of MrpL32 in *S.cere.*and its orthologs in other 10 other yeast species. The red boxed region shows the MTS of MrpL32. Conserved positions are colored by Jalview.

divergence, but not by SVM without divergence features, nor by TargetP [57].

2.3 Discussion

Although strong sequence similarity is a widely used indicator of co-localization, characteristically low sequence conservation in signal sequence regions has not been utilized for prediction. Other authors have noted the low sequence conservation of N-terminal sorting signals such as MTS sequences [61], but this work reported here is the first investigation of the utility of sequence divergence as a predictive feature for N-terminal sorting signals.

The method reported here requires defining an ortholog set for each gene. The YGOB curated dataset for 11 yeast species is a reliable way to obtain orthologs, but this kind of database is not

available for most species. I show that a simple reciprocal best hit method identified orthologs with sufficient reliability for the purposes of computing sequence diversity. One avenue for future research is to relax the requirement of global alignment reciprocal best hit designed to find orthologs, and simply use for (possibly paralogous) homologous sequences. In this study I chose to focus on orthologs because paralogs often have distinct localization sites. For example, Rosso et al. [62] describe the interesting case of the human glutamate dehydrogenases GLUD1 and GLUD2. These paralogs result from a gene duplication event, but GLUD1 localizes to both the cytosol and the mitochondria while GLUD2 localizes exclusively to the mitochondria. Interestingly, the N-terminal region of GLUD2, which functions as an MTS, has evolved faster than GLUD1 [62].

Since I made a few somewhat arbitrary choices when defining divergence features, I performed an *post hoc* analysis to see if simply tuning those parameters would significantly affect the prediction accuracy. Namely, I investigated the effect of the changing the window length and position of the downstream normalizing window used to define NCdiff, but found that prediction accuracy is not strongly dependent on the exact value of these parameters (Figures A.1, A.2). Another potential weakness of this method is the simple entropy based definition I used for sequence divergence, which ignores the phylogenetic relationship of the species involved. Many sophisticated measures have been proposed to quantify the degree of sequence conservation [39]. I did experiment with some of them, such as the Jensen-Shannon divergence [63] to try to improve prediction, but without success (results not shown). However I did not extensively explore the possibilities and believe that the simple entropy score employed here probably can be improved upon.

On the other hand, I did provide quantitative evidence that the entropy divergence score has considerable predictive power by itself. The examples ALA1 and FMP52 show that divergence can flag proteins (typically mitochondrial ones) with misannotated MTS information and give a hint

regarding which compartment of the mitochondria they localize to. Examples like MrpL32, show that when the predictions of standard predictors are inconsistent with the degree of sequence divergence, non-typical MTS's, processing proteases or alternative mitochondrial localization pathways may be indicated.

One weakness in the datasets is that many of SP proteins are not experimentally validated, but rather annotated as SP proteins due to UniProtKB/Swiss-Prot annotation and prediction from amino acid sequence with SignalP [29] in the yeast dataset. 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 [64] dataset lists only six). Proteins which localize to the E.R. or Golgi body generally posses SPs, but many proteins annotated as E.R. or Golgi are non-SP containing peripheral membrane proteins, which localize to the periphery of these organelles. 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 [3, 32]. Second, given the intense study of *S.cere.* and the continued scrutiny of UniProtKB/Swiss-Prot by the research community, I find it unlikely that a large fraction of the SP proteins in the dataset are incorrectly labeled. Third, this argument is not completely circular. SignalP prediction is based on physico-chemical features but not divergence (or conservation) for prediction, and the results shown in Figure 2.5 suggest physico-chemical features do not correlate very closely with sequence divergence.

Chapter 3

Prediction of presequence and its cleavage site

3.1 Materials and methods

3.1.1 Training and test dataset

Presequence prediction

The positive training dataset contains 759 mitochondrial protein sequences with a presequence which are extracted from UniProtKB/Swiss-Prot [28] ver. 2012_10. The mitochondrial protein data includes recent presequence proteome data [3, 32] , the dataset of TargetP , and that of Predotar. For negative examples I used 6310 non-mitochondrial with clear Swiss-Prot annotation of subcellular localization and 108 non-cleaved yeast mitochondrial proteins [3]. No pair shared more than 80% mutual sequence identity in each positive and negative dataset. To compare the prediction

performance, I prepared an independent test dataset consists of 78 mitochondrial proteins possessing a presequence and 8934 non-mitochondrial proteins; the sequence identity between training and test datasets are less than 25%. Also, there is no pair sharing more than 25% sequence identity in each positive and negative test dataset.

Cleavage site prediction

Cleavage site were extracted from the proteomic analysis experiments for *S.cerevisiae* [3] and *A.thaliana* and *O.sativa* [32]. To reduce redundancy, sequences were extracted from their N-terminal up to three residues after cleavage site as same as TargetP [65], and redundant sequences were reduced with 40% identity in each taxonomic groups. Although the original proteomic data for the yeast shows multiple cleavage sites on a protein in some cases, most frequently observed sites were extracted. To extract cleavage sites which are processed by MPP, I exclude proteins whose first cleavage site does not contain arginine at -2 position (in plant dataset, -3 position is also taken into consideration due to unreported hypothetical Icp55 equivalent sites). Although a few proteins which do not contain arginine at -2 position are annotated that they are cleaved by MPP and other intermediate proteases, they were not used for training but tested. Test was conducted by 10-fold cross validation. Negative dataset is prepared by extracting sequences which matches X{2}RX{6} at non-cleaved position of N-terminal in the positive dataset.

3.1.2 Training MitoFates

The flow of MitoFates consists of two parts: presequence prediction and its cleavage site prediction. Both prediction bases on Support Vector Machine (SVM) implemented in LIBSVM 3.0 with RBF-kernel [45]. Details of SVM is described in materials and methods section of chapter 2. SVM with

polynomial kernel or random forest were also applied, SVM with RBF kernel shows best performance (detail is described in the appendix). The features used in the presequence predictor are: the presequence frequent 6-mer motif hits, the improved hydrophobic moment score, weighted position weight matrix (PWM) of MPP cleavage site, physicochemical features of the N- and C- terminal region and amino acid compositions (the details are described below). Prediction performance was evaluated by Precision-Recall curve. Precision and Recall are defined by below equations. Similarly, features for cleavage site prediction are: PWM, physico-chemical properties and amino acid composition from N-terminal to each candidate sites.

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \end{aligned}$$

, where TP, FP, FN are True Positive, False Positive and False Negative, respectively.

3.1.3 Mitochondrial presequence frequent motif finding

For motif finding, I used N-terminal 90 residues of 317 mitochondrial proteins with a presequence and 3897 non-mitochondrial proteins. These sequences are a subset of training dataset and share less than 25% sequence identity. I reduced 20 amino acids to 5 letters based on physicochemical properties; hydrophobic ϕ (L, F, I, V, W, Y, M, C, A), basic β (R, K, H), acidic α (E, D), polar σ (S, T, N, D) and secondary structure breaker γ (P, G). Then, I partitioned the N-terminal 90 residues into three blocks of 30 residues and I counted protein match to the 6-mer motif comprised of the 5 letters in each blocks of the mitochondrial and non-mitochondrial proteins. With these counts, I used a highly sensitive multiple-testing method LAMP [66] to conduct Fisher's exact test

for the 6-mer. All statistically significant 6-mer (p-value $< 10^{-5}$) are found in the first block (see results section), thus I used the fourteen 6-mer with p-value less than 10^{-5} found in N-terminal 30 residues for the prediction. The score of each 6-mer hit is defined by $-\log_{10}(p-value)$.

3.1.4 Revised hydrophobic moment for presequence (μ_N)

$$\mu N \equiv \frac{1}{n}\{\sqrt{(\sum_i H_i cos(\delta i))^2 + (\sum_i H_i sin(\delta i))^2}$$
$$- \quad rcos\theta\sqrt{(\sum_i C_i cos(\delta i))^2 + (\sum_i C_i sin(\delta i))^2}\}$$

, where r is a ratio parameter to balance between hydrophobic moment and charge moment, and θ is a degree consists of vectors of hydrophobic and charge moments. Here, H_i indicates Aboderin hydrophobicity scale [67] and C_i is charge index ([R, K, H] and [D, E] are valued at +1 and -1, respectively and other residues are 0). μ_N is normalized by length of window n. Parameters, degree of helix and balancing parameters, were optimized by using the training dataset to maximize discriminative power measured by F-score. Best degree and balancing parameters are 96° and 8.5, respectively.

3.1.5 Distance from N-terminal considered Position Weight Matrix

Local sequence around cleavage site is important for MPP recognition as known as "R-2 rule" [13]. PWM was generated from residues between the -4 position and the +5 position around cleavage site. Here, negative value indicates upstream of the cleavage site and positive value does downstream. Thus, cleavage position is indicated by -1 and +1. Observed frequencies at each positions are smoothed by 20 component Dirichlet mixtures [68], and these 20 values are divided by amino acid composition of mature region in mitochondrial cleaved proteins as background frequencies. PWM

score is calculated as log-odds ratio between trained cleavage site and background. In addition, presequence length seems to be informative for the prediction; therefore, scores at each position were weighted by distribution of presequence length, $f(t)$, according to distance from N-terminal to the window, i.

$$\log_2 \frac{P(x|model)}{P(x|null)} + \log_2 \int_{i-1}^{i} f(t)dt \tag{3.1}$$

The length distribution was fitted to Gamma mixture and parameters were estimated by EM algorithm [69].

3.1.6 Amino acid composition

It is well known that protein secondary structure correlates with amino acid composition and invokes correlation between nearby residues. Therefore I adopted amino acid composition and dipeptide composition in N-terminal 30 residues. Moreover I skip two dipeptide composition, defined as AxxB, where A, B are fixed amino acid residues, and x is any residue. When this 4-mer forms a helix, A is close to B in the helix. Similarly, amino acid composition up to candidate site k were applied to cleavage site prediction. In the latter case, length k can be very small; therefore, composition is smoothed and transformed to posterior probability by 20 component Dirichlet mixture prior [68] and transformed 20 dimension vector as features in SVM.

3.1.7 Physico-chemical propensities

Presequence prediction:

Proteins bound for the endoplasmic reticulum usually and peroxisome often possess predictable sorting signals in their N- and C-terminals, respectively. To distinguish between mitochondrial presequence and such signal sequences, I partition the N- and C- terminal 90 residues into 6 blocks

of 15 residues, and then compute the average hydrophobicity [67], α-helical and β-strand periodicity scores [70, 71], and the density of basic (K, R, H), acidic (D, E), small polar (S, T), aromatic (W, Y, F) and secondary structure breaker (P, G) residues for each block. Clearly these features are also relevant to structural motifs and in fact I also include them computed over the entire sequence in the feature set. In addition, four signal peptide related features are used. The four signal peptide related features are the same in a previous study [71].

Cleavage site prediction:

As it is known that positive charge importance for protein import to mitochondria, a similar hypothesis was proposed that MPP also uses positive charge in N-terminal to import their substrates into the cavity [72, 73]. To quantify those features in cleavage site model, averaged net charge and averaged hydrophobicity were used. In addition, the number of characteristic charged residues were used as inputs. Within presequence region, negatively charged residues rarely appeared. Thus, increasing of the number of such residues can be a sign for the end of presequence.

3.1.8 Discrimination of intermediate proteases

In yeast and metazoa models, some mitochondrial proteins are cleaved by intermediate proteases in the matrix after MPP cleavage; namely, Oct1 and Icp55 (in metazoa, Icp55 is still hypothetical). To predict correct position of cleavage site, MitoFates classifies Oct1 substrates, Icp55 substrates and proteins which are not cleaved twice by simply applying PWM profiles of Oct1 and Icp55. Oct1 profile was trained on four residue-long sequences after MPP cleavage site in Oct1 substrates. Similarly, Icp55 profile was generated from two residue-long sequences in Icp55 substrates. If residues after predicted MPP cleavage site shows higher score than threshold, it is predicted as

double digestion proteins. Threshold for two profiles were determined by highest MCC value in training data set. To classify a query into MPP+Oct1, MPP+Icp55 and MPP only, MitoFates predicts MPP+Oct1 class first and MPP+Icp55 class if its score is lower than threshold of Oct1 profile. When two profiles do not match a sequence, it is predicted as MPP only class. In plant model, I take only hypothetical Icp55-like protease into my account.

3.1.9 Examination of effective features

I used the so called "F-score" to quantify the importance of each feature. The F-score [74] 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:

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

, 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.

3.1.10 Clustering of yeast presequence

Clustering analysis of presequence in the yeast dataset was conducted with selected features to make interpretation and clustering easier. Selected features are length, averaged net-charge, improved H-moment score (μ_N), MPP cleavage score (PWM score) around cleavage site, compositions of charged residues (Arg, Lys, Asp and Glu), and amino acid conservation. Since conservation scores fluctuate by column to column, first 36, average length of entire presequences in the yeast proteomic analysis [3], and its half 18 positions are averaged, respectively. Clustering is conducted by

application of Gaussian mixture model, and model parameters are estimated by EM algorithm [69] implemented in Weka [43].

Given data $x \in R^d$, Gaussian mixture model can be defined with parameters, namely mean μ_k and covariance Σ_k for each Gaussian density of K-component Gaussian mixture. Define the number of samples is N and that of components is K. Each component is a multivariate Gaussian density:

$$p_k(x|\theta_k) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}} \exp(-\frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)) \tag{3.2}$$

, where $\theta_k = \{\mu_k, \Sigma_k\}$. p_k is a k-th component of a mixture model, and p is defined below:

$$p(x|\theta) = \sum_{k=1}^{K} \alpha_k p_k(x|z_k, \theta_k) \tag{3.3}$$

, where $\alpha_k (0 < \alpha_k < 1)$ are the mixture weights, and $\sum_{k=1}^{K} \alpha_k = 1$. z_k are latent variable vectors that determines the components from which each sample originates. With Bayes rule and current parameters Θ, posterior probability is written as below.

$$p(z_{ik}|x_i, \Theta) = \frac{p(z_{ik}|x_i, \theta_k) \cdot \alpha_k}{\sum_{m=1}^{K} p(z_{im}|x_i, \theta_m) \cdot \alpha_m} \tag{3.4}$$

EM algorithm for Gaussian mixture model comprises of E-step and M-step. E-step: Calculate $p(z_{ik}|x_i, \Theta)$ with current parameters Θ for all samples $x_i (i = 1, ..., N)$ and all components $k = 1, ..., K$. M-step: Compute new parameters as below.

$$\alpha_k^{new} = \frac{\sum_{i=1}^{N} p(z_{ik}|x_i, \Theta)}{N}, 1 \leq k \leq K \tag{3.5}$$

With these new mixture weights, update θ_k.

$$\mu_k^{new} = \left(\frac{1}{\sum_{i=1}^{N} p(z_{ik}|x_i,\Theta)}\right)\sum_{i=1}^{N} p(z_{ik}|x_i,\Theta)x_i, 1 \leq k \leq K \tag{3.6}$$

$$\Sigma_k^{new} = \left(\frac{1}{\sum_{i=1}^{N} p(z_{ik}|x_i,\Theta)}\right)\sum_{i=1}^{N} p(z_{ik}|x_i,\Theta)(x_i-\mu_k^{new})(x_i-\mu_k^{new})^t, 1 \leq k \leq K \tag{3.7}$$

Then, go back to M-step with θ_k^{new}. Until convergence, these E-step and M-step are iterated. To decide convergence or not, log-likelihood function is defined:

$$\log l(\Theta) = \sum_{i=1}^{N}(\log \sum_{k=1}^{K} \alpha_k p_k(x|z_k,\theta_k)) \tag{3.8}$$

Decision on the number of clusters is one of key points, and in this work it was determined by following default greedy and ad-hoc algorithm of the implementation simply. Thus, given data is randomly split into 10 to conduct 10-fold C.V and computes log-likelihood 10 times. If averaged log-likelihood is increased with increasing cluster number by 1, iterate this step. Otherwise, stop and returns the number of clusters as a most likely model to the given data.

3.2 Results

3.2.1 Prediction performance of MitoFates

Presequence prediction

Three computational tools are widely applied to presequence prediction at present: MitoProt, TargetP, and Predotar as I described [65, 75, 31]. To compare performances fairly against all

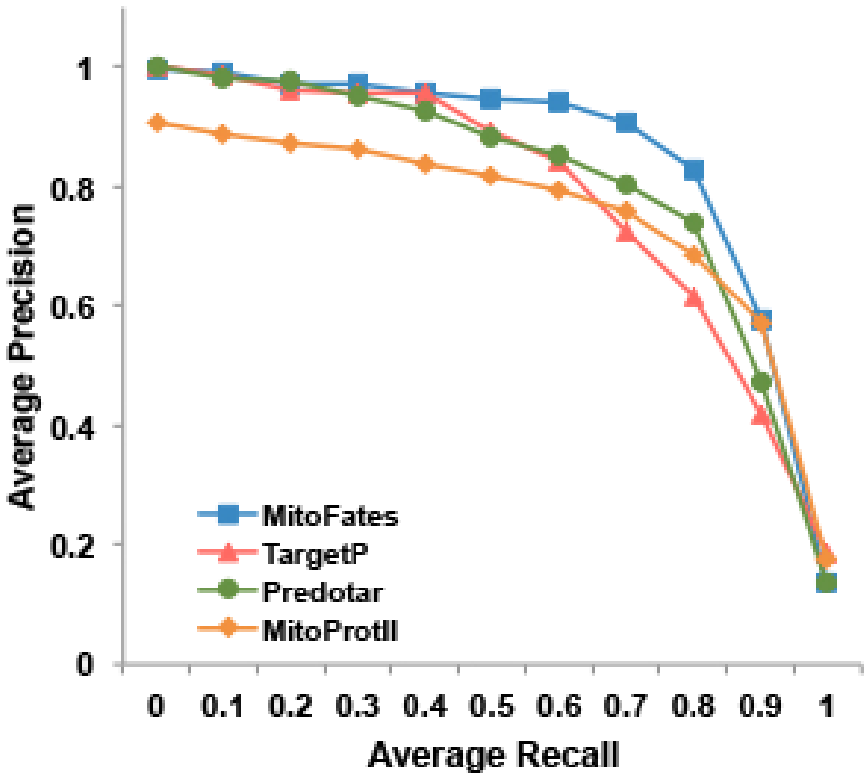


Figure 3.1: Precision-Recall curves for MitoFates, Predotar, TargetP, and MitoProtII. Since ratio between presequence containing proteins and negative test dataset is very skewed, negative dataset is randomly split into a set which includes 500 sequences. Shown points are averages of 10 iterations.

including predictor described here (MitoFates), independent test dataset is prepared (detail is described in the materials and methods). Figure 3.1 shows the 11 point precision-recall curve (PR-curve) of each predictor by testing 10 times with keeping the positive test data to avoid imbalance affect between the positive and negative datasets. The curve of Mitofates locates at most upper-right on the PR space, outperforming TargetP, Predotar and MitoProtII. MitoFates attains superior precision than other predictors at recall 50-90%. The improvement should leads to higher chance to identify of undiscovered mitochondrial proteins.

There are two problems in the evaluation: small number of positive test dataset and unrealistically skewed ratio between the positive and the negative test datasets. In fact, recall is almost saturated at about 4% false positive rate for MitoFates and Predotar maybe due to the limited number of positive test dataset, 78. To avoid these problems in statistical test, McNemar's test, one of paired tests, is applied to the positive dataset with controlling false positive rate. In this manner, two predictors' performance is compared in the positive dataset whether prediction of a query is correct or not, namely score is higher than the threshold determined in the negative dataset. Therefore, false positive rate is tried to change from 1% to 15%, then test at each false positive rate. As shown in Figure 3.2, recall of MitoFates is entirely higher, especially in low false positive rate ($< 2\%$), and within this very low false positive rate, MitoFates is significantly accurate in the positive dataset. Predotar and TargetP have other predictor in the system for SP and N-Signal-free, and final prediction is determined by comparing prediction scores from all models. Under this condition it is difficult to estimate recall at high false positive rate region, so comparing with scores of only mitochondrial model is also conducted (in this condition, scores of other models and predicted labels for queries are ignored). Dotted lines show result of this model only considering condition. Although the differences between MitoFates and the others are small in high false positive rate region, recall of MitoFates is still highest amongst them. In high false positive rate, prediction of the positive dataset is almost saturated, and this can be observed as p-value of the statistical test (Figure 3.2 top).

Cleavage site prediction

The other target of MitoFates is to predict MPP and intermediate proteases' cleavage site of mitochondrial presequence. Since cleavage site of presequence does not contain much information,

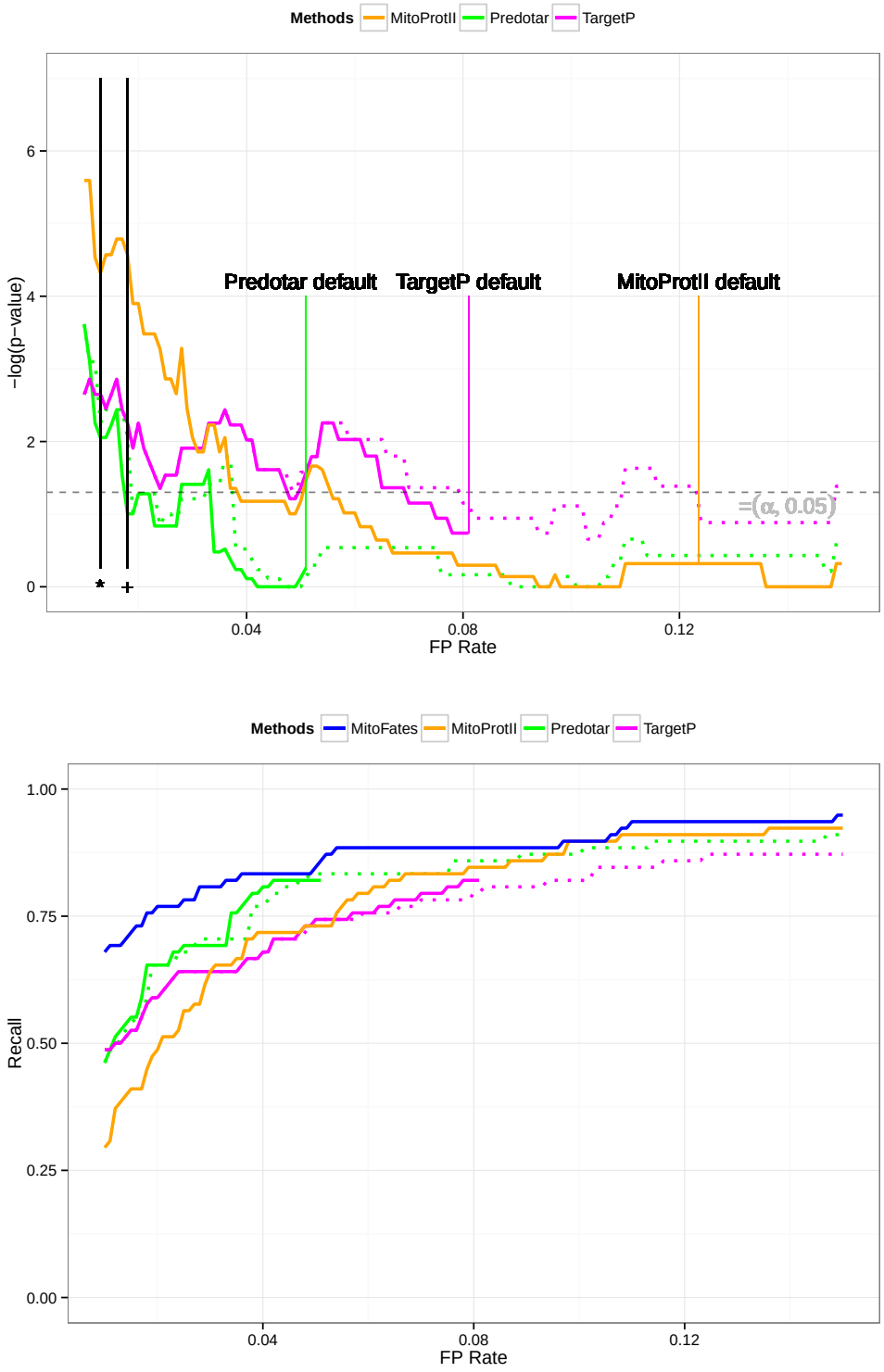


Figure 3.2: Statistical test on performance of MitoFates with other predictors. Dotted lines show prediction result in which only mitochondrial model is considered. (Top) P-values of McNemar's test on the positive test dataset is plotted. Threshold on the positive is controlled by false positive rate in the negative dataset. x-axis is false positive rate in the negative dataset, and y-axis is -log(p-value) McNemar's test at given false positive rate. * shows MitoFates default threshold automatically determined by the model learning, and + shows user adjustable threshold (set to the threshold at which recall is 80% in the training dataset). (Bottom) Recall, or sensitivity, of four predictors is plotted. x-axis is the same as the top figure, and y-axis is recall.

prediction of cleavage site is not at satisfactory level at present. Arg nearby cleavage site is a key feature and it has been experimentally confirmed that Arg at -2 position interacts with negatively charged residues in MPP [76]. SVM model for cleavage site which includes local information as PWM and other surrounding features searches MPP cleavage site at first, since intermediate proteases function after cleavage of MPP. Improvement of MPP cleavage prediction leads to better performance in total. Presequences cleaved by MPP and intermediate proteases were validated by ten-fold cross validation on yeast proteomic dataset [3]. Since two other methods, MitoProtII and TargetP, are widely used, accuracies are compared with those methods (Figure 3.3). In some cases, cleavage site prediction is not available in both two methods, proteins which has cleavage site prediction for the two methods are compared. MitoFates relatively stably predicts cleavage site of presequence, however, MitoProtII or TargetP result show some leaps between 0 and 1 or around 7 in terms of difference with actual cleavage site. This can be observed in plant dataset as well (Figure 3.4). Since presequences of plants show different length distribution and lack of R-10 motif, MitoFates takes such differences into account; thus, only length distribution was trained from plant dataset and Oct1 profiles is ignored to follow this hypothesis. Although other parameters were trained from yeast dataset, such simple adjustment improved accuracy in case of plant presequence. In addition, leaps between actual cleavage site and prediction reflects specificity for intermediate proteases. For instance, Phe at position +1 is a representative amino acid for Oct1, however, it is a dominant residue for R-3 motif in plant at position -1 (hypothetical recognized residue for plant Icp55). Another difference for plant presequence is weakly observed Met at position -1, and Icp55 profile trained on yeast cannot discriminate this residue. I should note here that the leap between 0 and 1 in plant dataset is result of such difference (Figure 3.4).

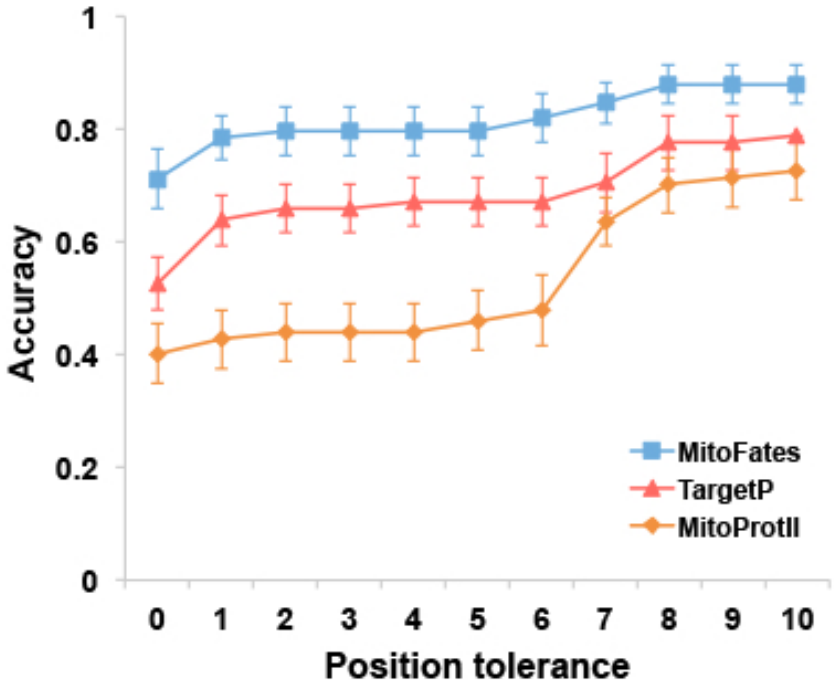


Figure 3.3: Evaluation of cleavage site prediction conducted by 10-fold C.V. in the yeast dataset. Error bar shows S.E. X-axis shows tolerance level, which defines how much extent difference between prediction and experimental annotation is accepted. Y-axis shows accuracy at each tolerance.

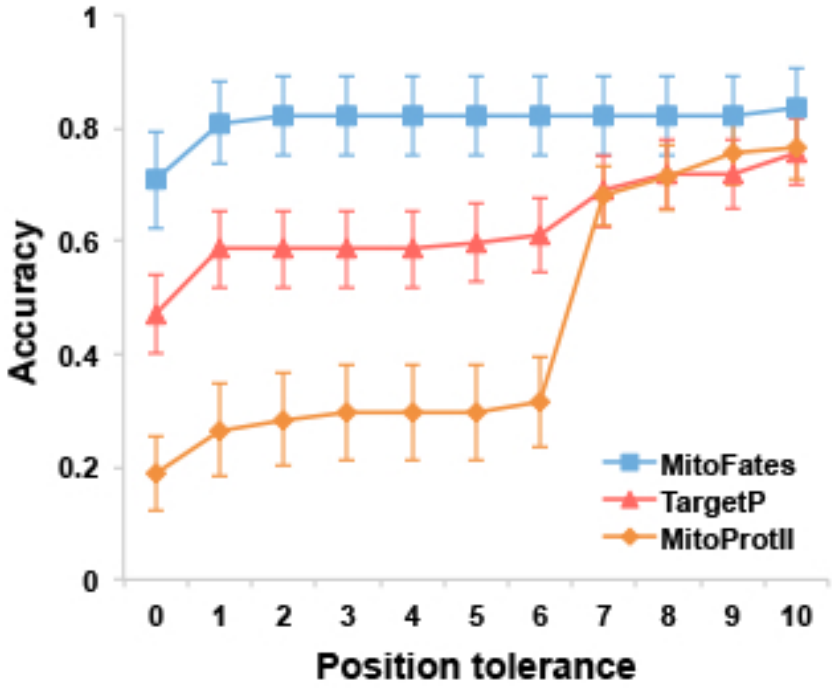


Figure 3.4: Evaluation of cleavage site prediction conducted by 10-fold C.V. in the plant dataset. In each fold, only length distribution is learn from the dataset. Error bar shows S.E. X-axis shows tolerance level, which defines how much extent difference between prediction and experimental annotation is accepted. Y-axis shows accuracy at each tolerance.

Name of feature	F-score
Cleavage score	0.245
R composition	0.217
Motif score	0.159
LR composition	0.126
Moment score	0.126

Table 3.1: Features used in MitoFates, and listed five highest ranked features among them. Features in italics are suggested in this thesis. To rank them, discriminative power is measured by F-score.

Name of feature	ρ
Cleavage score	0.429
R composition	0.398
Motif score	0.360
Moment score	0.334
LR composition	0.310

Table 3.2: Features used in MitoFates, and listed five highest ranked features among them. Features in italics are suggested in this thesis. To rank them, discriminative power is measured by Spearman's correlation coefficient (ρ).

3.2.2 Feature analysis

I calculate F-scores (eq. 3.2) for each prediction feature to examine the effective features (Table 3.1). The best features is the score of cleavage site (F-score = 0.25). The second-fifth best are the composition of Arg in N-terminal 30 residues (0.22), the total score of 6-mer motifs match in N-terminal 30 residues (0.16), dipeptide composition of Leu-Arg (0.13) and improved hydrophobic helical moment score (0.13), respectively. These F-scores suggested newly integrated three presequence features well contribute to the improvement of predictions. With another measurement (by Spearman's rank correlation coefficient), top five features are kept (Table 3.2).

3.2.3 Characteristic features for cleavage site

Length of mitochondrial presequence is relatively diverse. Gamma distribution were applied to fit presequence length in both yeast and plant dataset (Figure3.5). The shape and scale parameters

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

Table 3.3: Results of goodness of fit tests.

of Gamma distribution were estimated by using the EM algorithm implemented as a package for R [69, 77]. With these estimated parameters, fitness between the data and theoretical distribution was measured by a Kolmogorov-Smirnov test (Table 3.3). The best-fit theoretical distribution were Gamma unimodal for yeast data and Gamma trimodal distribution for plant data set. These distributions were used as attributes in the cleavage site scoring. Characteristic difference between two distributions is very short (<10 amino acid) yeast presequence. Minimum length for yeast presequence is 6, however that for plant presequence is 18. Due to this difference, different weighting distributions applied to sequences.

It has argued that cleavage site of the presequence contain three distinct motifs with regards to arginine position: R-2, R-3 and R-10 motifs as described in the introduction. Because of similarity between R-2 and R-3 motifs, Schneider and colleagues predicted a putative protease which cleaves hydrophobic residue at -1 position in cleavage site which holds Arg at -3 position [5]. In fact, the putative protease was discovered and named Icp55 [3]. With the discovery of Icp55 and their annotations, I could develop a more appropriate profile for MPP with taking Icp55 and Oct1 into account (Figure 3.6). Profiles for Icp55 and Oct1 were also developed from yeast proteomic analysis (Figure 3.6), and their length are 2-mer and 4-mer, respectively. Although Oct1 cleaves typically eight amino acids after MPP [13], only first four residue are characteristic.

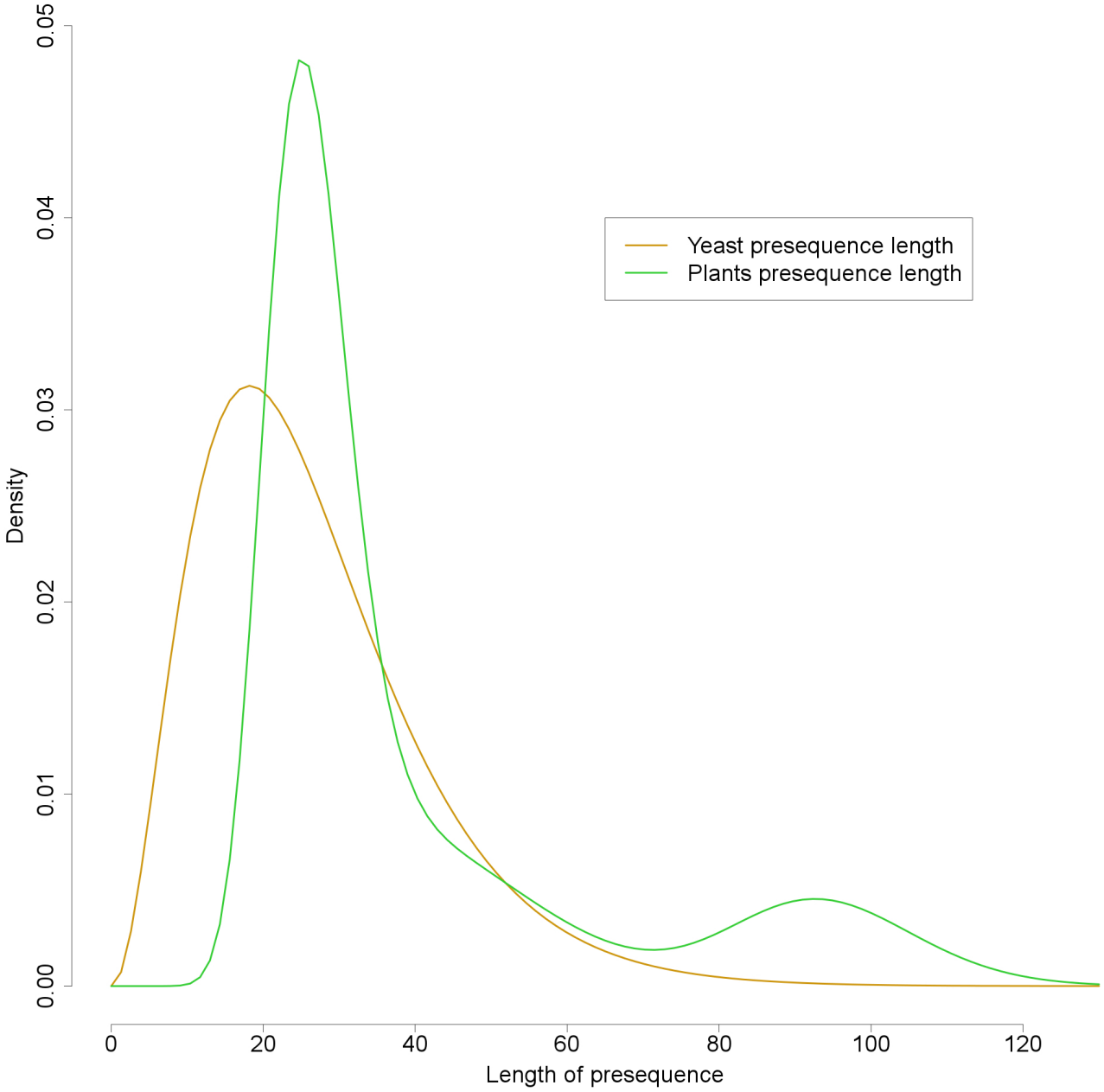


Figure 3.5: Length distribution learned from the yeast and plant dataset. Gamma mixture is fitted to the actual presequence length data. For the yeast, unimodal distribution was selected, and trimodal distribution for the plant dataset.

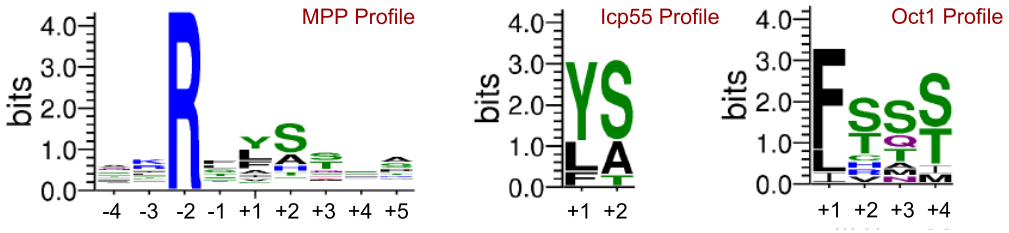


Figure 3.6: Sequence logo diagrams for MPP, Icp55, and Oct1 profiles, respectively from the left.

Although Oct1 is a conserved protease from budding yeast to human, R-10 motif has not been observed in plant presequence [32]. In fact, *A.thaliana* has one homolog for Oct1, At5G51540, however, a large scale analysis reported the protein localized in chloroplast [78]. Therefore, only profile for hypothetical Icp55 was applied to predict hypothetical Icp55 like processing. I should note here that dominant residue is Phe at hypothetical Icp55 cleavage site; however, equivalent position favors Tyr in Icp55 substrates. These difference might reflect absence of Oct1 homolog in plant mitochondria. Icp55 can cleave Phe in yeast, therefore, Icp55 might complement Oct1 like processing in plant.

3.2.4 Refinement of scoring amphipathic α-helix in presequence

Mitochondrial presequence have 10-90 residues of length and the potential to form positively charged amphiphilic helices. Import receptor Tom20 and Tom22 is assumed to recognize an amphiphilic helical feature consist of hydrophobic and positively charged faces of presequences [6, 9, 10]. Although amphiphilic α-helix has been reported as one of the characteristics for presequence, applying this feature is not successful in practical situation. MitoProtII calculates maximum 18 residue long hydrophobic moment in N-terminal as a feature of presequence, however, Predotar discusses that parameters related to amphiphilic helix formation failed to improve predictions [75, 31]. TargetP

does not depend on this feature explicitly [65]. Thus, I investigated the distributions of maximum hydrophobic moment score in N-terminal 30, 60, 90 residues regions against both presequence containing proteins and negative dataset. The best discrimination was shown by the maximum hydrophobic moment score in N-terminal 30 residues, but the distributions of the score looked overlapped to each other (Figure 3.7). Because the the original formula of hydrophobic moment is based on hydrophobicity index, the formula are unable to distinguish positively charged, negatively charged, or polar residues. Thus the original formula does not discriminate positively charged amphiphilic helix and general amphipathic helix.

To overcome this drawback of the original formula, I determined the improved hydrophobic moment formulation by integrating charge moment (μ_N). Detail of the formula is described in the method section. As a result of the adjustment, the improved hydrophobic moment μ_N shows better discrimination of the presequence containing proteins and negative examples (Figure 3.8). Note that this improved hydrophobic moment can predict only two, but all known Tom20 binding sites in the presequence (Su9 of *N.crassa* [79] and ALDH2 of *R.norvegicus* [80]).

3.2.5 Novel motif finding in presequence

The sequences of presequences are often differ between orthologs, thus there is thought to be no consensus in the primary sequence. Meanwhile, a peptide library experiment actually revealed that 6-mer amphiphilic motif $\theta\phi\chi\beta\phi\phi$ (where θ, β, ϕ and χ represent a hydrophilic, hydrophobic, basic and any residue) for Tom20 [9]. However the motif covered only 18% and 19% of yeast proteomic presequences data [3] and the presequence data in this study, respectively. A motif finding based on discriminative hidden Markov model (HMM) has been performed against mitochondrial proteins. In that motif finding, only a few 4-mer motif candidates were found [11]. However,

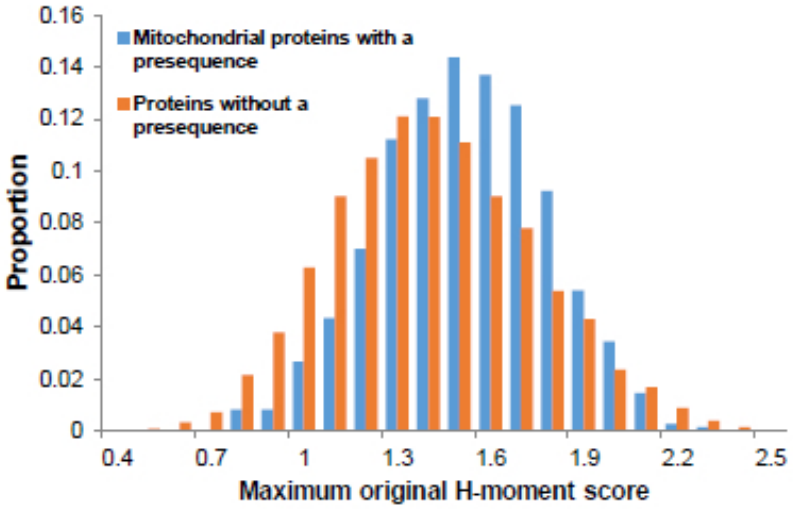


Figure 3.7: Distributions of classical hydrophobic moment scores for presequence containing proteins and proteins without the presequence.

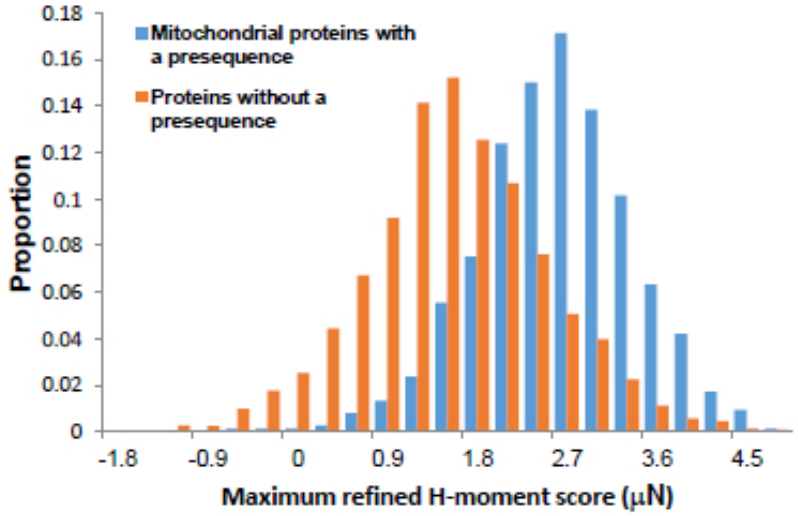


Figure 3.8: Distributions of refined hydrophobic moment scores μ_N for presequence containing proteins and proteins without the presequence.

that was done without considering whether proteins have a presequence or not. Taken together, there is still possibility for finding undiscovered presequence-specific motifs. Thus, I attempted to find statistically significant 6-mers for presequences by using five reduced amino acid letters; hydrophobic ϕ (L, F, I, V, W, Y, M, C, A), basic β (R, K, H), acidic α (E, D), polar σ (S, T, N, D) and secondary structure breaker γ (P, G). By using highly sensitive multiple-testing method LAMP [66], motif finding were performed in N-terminal 90 residues, which is portioned into three blocks of 30 residues.

As a result, fourteen statistically significant 6-mers (p-value $< 10^{-5}$ as compared to negative examples) are detected in Figure 3.9. Interestingly, the significant 6-mers were found only in the first N-terminal block, first 30 residue-long region. All of 6-mers has at least three hydrophobic residues and one basic residue. In addition, no acidic residues occurred in the 6-mers. To clear whether the 6-mer are caused from amino acid composition of presequences or not, I also performed same motif finding against the scrambled sequences of presequences. However the fourteen 6-mers and also other significant 6-mers were not found in them. Thus the fourteen 6-mer is not influenced by amino acid composition of presequences. Interestingly, the most of helical wheels of the significant 6-mers seems to have an amphiphilic helical feature consist of hydrophobic and positively charged faces. As shown in Figure 3.9, the maximum refined H-moment scores (μ_N) of 6-mers in N-termini 30 residues of proteins possessing a presequence tend to be higher than those of non-presequence proteins (the sensitivity and specificity versus cutoff value is 2.56). In the 11 of 14 significant 6-mers, the average of μ_N is greater than 2.56, meaning the 6-mers have higher amphiphilicity. The found positively charged amphiphilic 6-mers might include novel motif candidates for Tom20 and Tom22 recognition. The best 6-mer was $\phi\phi\sigma\beta\phi\phi$ (p-value: 5.7×10^{-13}) and the 6-mer was found in 13% of mitochondrial proteins. This 6-mer is similar to the motif given by peptide library study [9].

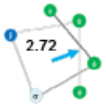
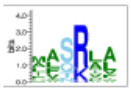
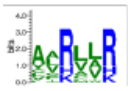
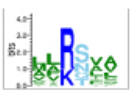
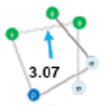
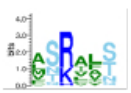
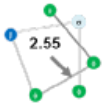
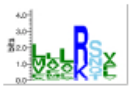
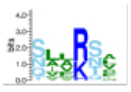
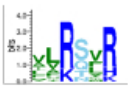
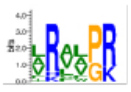
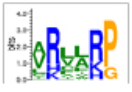
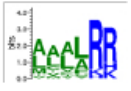
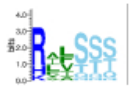
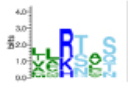
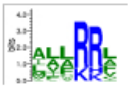
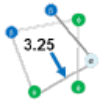
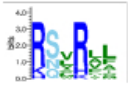
Rank	6-mer	P-value	Coverage of Positive (%)	Coverage of Negative (%)	Helical wheel	Sequence logo	Rank	6-mer	P-value	Coverage of Positive (%)	Coverage of Negative (%)	Helical wheel	Sequence logo
1	ϕϕσβϕϕ	5.7E-13	12.9	2.2			8	ϕϕβϕϕβ	8.9E-07	6.3	0.8		
2	ϕϕβσϕϕ	1.2E-11	11.4	1.9			9	ϕσβϕϕσ	9.7E-07	7.6	1.3		
3	ϕϕϕβσϕ	1.1E-09	9.5	1.5			10	σϕϕβσϕ	1.2E-06	6.6	0.9		
4	ϕϕβσϕβ	1.8E-09	6.3	0.5			11	ϕβϕϕγβ	1.7E-06	3.8	0.2		
5	ϕϕβσϕγ	6.1E-09	4.4	0.1			12	ϕϕϕϕββ	4.9E-06	5.9	0.8		
6	βϕϕσσσ	9.3E-08	5.7	0.5			13	ϕϕβσϕσ	6.4E-06	6.3	0.9		
7	ϕϕϕββϕ	1.2E-07	7.6	1.1			14	βσϕβϕϕ	9.3E-06	4.7	0.4		

Figure 3.9: Fourteen motifs are listed. Header describes each content of a column in the table. Sequence logos were generated by WebLogo.

However the first resides is different between the two; the first residue is hydrophilic in the motif while that is hydrophobic in the best 6-mer. In ignoring last position of the motif, 8th and 9th best 6-mers ($\phi\phi\beta\phi\phi\beta$ and $\phi\sigma\beta\phi\phi\sigma$) also match to the motif. The $\phi\chi\beta\phi\phi$ might be a core sequence for recognition by Tom20.

Any of the fourteen 6-mers found in 55% of N-terminal 30 residues of mitochondrial proteins with a presequence while found in only 10% of that of negative examples. So the fourteen 6-mers should be useful for discrimination of presequences. Thus I used the matching of the fourteen 6-mers in N-terminal 30 residues for the prediction as features.

3.2.6 Clustering of mitochondrial presequences

MitoFates attains better performances than the existing other predictors however I still failed to predict a number of presequences. The fails mostly are caused by presequences with less positively charged or less score of MPP cleavage, indicating that there is a number of non-classical presequence. Characterizing non-classical presequences is necessary for further improvement of the prediction.

Yeast mitochondrial presequences were most experimentally identified in studied organisms [3]. Thus I tried to classify 243 yeast presequences by using EM algorithm with Gaussian mixture model employing the 9 features; length, averaged net-charge, improved H-moment score (μ_N), MPP cleavage score (PWM score) around cleavage site, compositions of charged residues (Arg, Lys, Asp and Glu) and amino acid conservation (see the materials and methods). Similar to signal peptides for E.R, the amino acid conservation of presequences tend to be poor [12]. The poor conservation is one of features of presequence, I therefore used the conservation in the cluster analysis. Degree of conservation is measured by symmetric Jensen-Shannon divergence [63] and orthologs are extracted from Yeast Gene Order Browser [33].

The cluster analysis results in at least three clusters of presequences (see Table B.3 in the appendix). In Figure 3.10A, I mapped the three presequence groups in 3D subspace of primary component analysis (PCA). As shown Figure 3.10A, largest cluster (cluster I) consists of 144 presequences which are positively charged (almost no negatively charged residues), weak conservation, have moderate length (the average is 25 residues), higher μ_N and significantly high MPP cleavage scores. Classical presequences are categorized into the cluster I since the properties of the presequences are consistent with those of typical presequences. The characters of the remaining 99 presequences in the other two clusters differ from the cluster I; these are "non-classical presequences".

The second largest cluster (cluster II) comprised of 64 presequences. Similarly to the cluster I, the presequences in the cluster II are weak amino acid conservation and their μ_N are comparable to the cluster I. However, the length of the presequences is longer (the average is 60 residues) and also the net-charge is lower. The reasons for lower net-charge are lower composition of Arg and higher composition of negatively charged residues (Fig 3.10B). Besides, the most of MPP cleavage scores are significantly lower than cluster I, meaning that their cleavage sites do not match with the cleavage pattern by MPP. The observation intimates that the cluster II includes the presequences which are cleaved by other proteases. In fact, the substrates of inner membrane proteases such as m-AAA and Imp are classified into the cluster II; Ccp1, MrpL32, Cyt1 and Gut2 (summarized in [8]). The averages of length and Arg composition of presequences of the four substrates are 62.8 residues and 0.06, respectively. The longer presequences having less Arg composition may tend to be cleaved by proteases other than MPP. In addition, Imo32, which cleaved by MPP and Oct1 [60], but the MPP cleavage site does not match with MPP cleavage site motif, is included in the cluster II. The presequence length of Imo32 is 38 residues, and has only two Args. Such less Arg composition may influence abnormal cleavages of MPP. Like Imo32, presequences, which are cleaved by unusual manner of MPP, would be included in the cluster II. So, the cluster II might be classified into subgroups.

The 35 presequences in the third cluster (cluster III) are elusive. Similarly to cluster II, the averaged net-charge and MPP cleavage scores of the presequences are lower than cluster I (net-charge of 40% of the presequences in the cluster is below zero, and also Arg composition is low). Unlike the other clusters, the conservation is higher. Moreover the matching of the fourteen presequence frequent 6-mers with N-terminal 30 residues of presequences is quite lower than the others; the coverage of the total hit of fourteen 6-mers for the cluster I, II and III are 56.3%, 42.2% and 14.3%,

respectively. 13 of the 35 presequences in cluster III are derived from the mitochondrial protein annotated as dual localization or non-mitochondrial localization in Swiss-Prot. The features of the 13 presequences represent the above-described cluster III features; high conservation, low averaged net-charge, low MPP cleavage score and low Arg composition. In addition, H-moment score μ_N of the 13 presequences is lower. Low averaged net-charge and μ_N are consistent with previous reported the features of dual-localized mitochondrial proteins [81]. The cluster III includes some presequences with higher averaged net-charge (but, less Arg composition) which is comparable to the net-charge of cluster I presequences. Almost of all the presequences with such higher net-charge is 6 ribosomal protein presequences. The presequences of ribosomal proteins are also confirmed in Cluster I and II (23 and 6 ribosomal presequences in Cluster I and II, respectively). The presequences of ribosomal proteins which having higher averaged net-charge, but low MPP cleavage score and high conservation, are classified into the cluster III. Thus, Remaining anomalous presequences may be gathered in the cluster III, and the presequences could be subdivided.

Finding features which are effective to discriminate the non-classical presequences (cluster II and III) is essential to further improvement of the prediction.

3.3 Discussion

In this study, the prediction for a mitochondrial presequence and its MPP related cleavages site was improved by updating dataset and modeling presequence containing proteins with features. So far, it is said that 50-70% of all the known mitochondrial protein possesses a presequence [3]. MitoFates predicts 52% of annotated mitochondrial proteins; 557 of 1183 human mitochondrial proteins are predicted in Mitocarta [82], in which I updated the sequences and removed duplicated sequences. The coverage is consistent with the previous estimation.

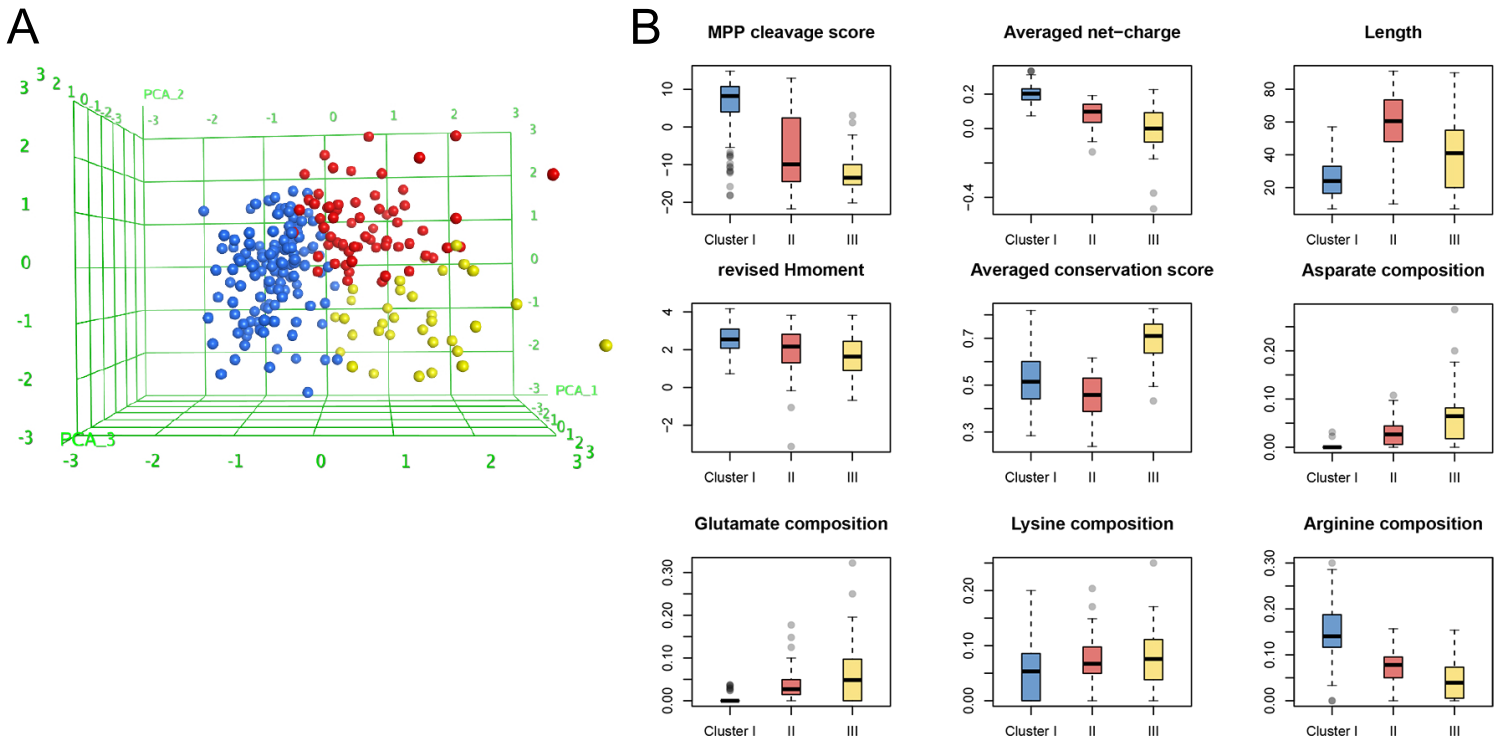


Figure 3.10: Blue, red, and yellow indicate cluster I, II, and III, respectively. (A) To visualize, clustering result is mapped to three dimensional space with principal component scores of PCA. (B) Distributions of each feature for three clusters are summarized in whisker plots. Light gray dots show outliers in each cluster.

The recent presumption estimates that few hundreds of undiscovered mitochondrial proteins exist in human [83]. Other than mitochondrial protein possessing a presequence, mitochondrial protein having non-cleavable internal targeting signal are also probably included in the undiscovered proteins. Besides, a couple of examples of mitochondrial protein having a C-terminal cleavable targeting signal are identified [84, 85]. Taken together, number of undiscovered mitochondrial proteins would be larger than the presumption.

Mitochondrial dysfunction causes diverse diseases such as muscle and neurodegenerative disease, cardiovascular disease, diabetes and cancer [2]. To understand the role of mitochondria in health and disease, comprehending of the protein composition of the organelle is essential.

One of the feature of presequence is amphiphilic helical feature consisting of hydrophobic and positively charged hydrophilic faces which is assumed to recognize by Tom20 and Tom22. In most of the fourteen presequence frequent 6-mers, the helical wheal shows a positively charged amphiphilic helix. These helical wheels indicate the possibility that the 6-mers is novel motif candidates for Tom20 and Tom22 recognition. I should note here that positive data and negative data which were compared in the task of motif finding contain different amino acid composition, especially N-terminal end. Conservative way to avoid such composition variance is using scrambled data of the positive dataset, which assumes such short significant motif is conserved. Although some of fourteen motifs does not appear in the list of significant hit of scrambled tests, the rest was detected by such test (see Table B.2 in the appendix). Since taking high evolutionary rate of N-terminal signal into account, I cannot deny potential function of such vanished motifs (there is a possibility that randomly appeared short motif might function). However, at least the rest repeatedly detected motifs even in the shuffled tests seem not to be an artifact of this work.

Recently, a long presequence pSu9 was reported to contain two distinct Tom20-binding elements;

the Tom20 binding element in N-terminal half and the efficient element for protein import in C-terminal half [79]. The report suggests other long presequences also have two Tom20-binding elements. To explore motif candidate of the C-terminal element in long presequence, presequences are classified into two groups by length; short group (40 amino acids or less, 215 mitochondrial proteins) and long group (more than 40 amino acids, 102 mitochondrial proteins), and then searched for statistically significant 6-mer in N-terminal 90 residues divided by three blocks of 30 residues. 5 highly significant 6-mer (p-value $< 10^{-5}$) are detected but all the 6-mers were found in only first block regardless of presequence length four and one significant 6-mers are found in N-terminal 30 residues in short and long presequences, respectively). Even if a higher p-value is set (p-value < 0.05), I could not find any statistically significant 6-mer in second and third blocks. The C-terminal binding with Tom20 of long presequences is efficient for import, but the binding element might not be distributed widely in long presequences. Since the statistically significant 6-mers are found in only N-terminal 30 residues regardless of presequence length, the important features for protein import into mitochondria might be organized within N-terminal 30 residues. Although it is difficult to analysis due to the number, mixing presequences from a variety of taxonomic groups might be negatively affect the result. Since plant presequences are usually longer (about 10 aa shift in Figure 3.5), long presequence from green plant might be different. In other words, 40 might be not long for this taxonomic groups. However, the number of presequences from green plant is small, this effect seems to be not so strong in the analysis.

The other purpose of MitoFates is to predict cleavage site of MPP and intermediate proteases. At present, the most popular cleavage site predictors are TargetP and MitoProtII [65, 75]. TargetP predicts cleavage site by sliding window with scoring matrices for R-2, R-3 and R-10 motifs [65], and MitoProtII uses sequence patterns for the three motifs [75]. These motifs can be explained

by intermediate proteases functioning after MPP, Icp55 and Oct1, and numerous substrates are annotated by the yeast proteomic analysis [3].

Chapter 4

Discrimination of mitochondrial membrane spanning regions

4.1 Materials and methods

4.1.1 Dataset

To analyze mitochondrial innner membrane proteins with a single transmembrane domain (TMD), I collected such proteins whose topology and structure are known from the OPM database [86]. Since the number of those proteins is limited, I also collected proteins whose localization and insertion mechanism were experimentally verified in a recent study [19].

In addition, I also added mitochondrial proteins annotated as single spanning inner membrane proteins in Swiss-Prot to the dataset. As a negative dataset, I used a collection of cytosolic proteins whose localization has been verified in *S.cerevisiae* and *H.sapiens*. Since it is easy to discriminate

non-TMD sequences without plausible TMD regions, I only included sequences which have regions with $\Delta G_{app}^{pred} < 3$ in the negative dataset. I prepared the datasets with sequence redundancy reduced to 40% identity in both full length and local TMD region (including 10 flanking residues in both N-terminal and C-terminal direction). To compare with TMDs inserted into the E.R. membrane, I also used sequence datasets from prior research [87]. Considering the possibility that insertion mechanism or features differ among taxonomic groups, I initially prepared separate datasets for fungi and vertebrates. However I could not find significant differences and therefore I prepared a redundancy-reduced merged dataset by the procedure described above. The results I report here are based on the merged dataset.

4.1.2 Sec61 translocon insertion model

T.Hessa and colleagues have reported an energy model for TMD inserted by Sec61 translocon [20]. The advantage of their model is its capability of explicit length correction, modeled by c_1, c_2, and c_3 below, and better discriminative parameters against plausible region in globular proteins, modeled by $\Delta G_{app}^{aa(i)}$. Negative values of ΔG_{app}^{pred} indicate spontaneously inserted TMDs.

$$\Delta G_{app}^{pred} \equiv \sum_{i=1}^{l} \Delta G_{app}^{aa(i)} + c_0 \sqrt{(G_{app}^{aa(i)} \sin(100^\circ i))^2 + (G_{app}^{aa(i)} \cos(100^\circ i))^2} + c_1 + c_2 l + c_3 l^2. \quad (4.1)$$

The values of c_0, c_1, c_2, c_3 (0.27, 9.3, 0.65, and 0.0082, respectively) are experimentally optimized [20]. $\Delta G_{app}^{aa(i)}$ assumes a symmetric Gaussian function (Figure 4.3). For the 18 amino acids other than Trp and Tyr, a single Gaussian is defined with two parameters.

$$\Delta G_{app}^{aa(i)} = a_0^{aa} \exp(-a_1^{aa} i^2) \quad (4.2)$$

where i indicates position within a window ($i = 0$ is the window center).

Trp and Tyr have two Gaussian model with four parameters.

$$\Delta G_{app}^{aa(i)} = a_0^{aa} \exp(-a_1^{aa} i^2) + a_2^{aa}(\exp(-a_3^{aa}(i - a_4^{aa})^2) + \exp(-a_3^{aa}(i - a_4^{aa})^2)) \tag{4.3}$$

4.1.3 Statistical free energy calculation

Eq. 4.1 includes positional energy parameters for the 20 standard amino acids at 19 positions relative to the TMD center, $\Delta G_{app}^{aa(i)}$ (shown in Figure 4.3). Although these 20×19 parameters have been measured experimentally, it is also possible to estimate these parameters empirically from multiple sequences via the following statistical mechanics relationship: [20].

$$G_{stat}^{aa(ij)} = -RT \log \frac{P_{ij}}{Q_j} \tag{4.4}$$

where R is the gas constant, T is set to room temperature (300K), i indicates the relative position within a 19 residue window, and j the amino acid. For the background probability distribution Q_j, I used the amino acid composition of all registered proteins in UniProtKB/Swiss-Prot release 2013_04.

4.1.4 Evolutionary information

Since evolutionary relevant homologs give clearer information regarding TMDs, I extracted evolutionary information in the form of PSSMs generated by DELTA-Blast [88]. According to [88], DELTA-Blast without iterative calculation achieves at comparable sensitivity to that of PSI-Blast [89] with three iterations, leading to practical computational time. I utilized the information from PSSMs in two ways: weighting parameters for positional parameters in the above model, $\Delta G_{app}^{aa(i)}$

and simple PSSM amino acid compositions in candidate regions to be predicted. I adopted the former model [90] with minor modification as described below.

$$\begin{aligned}\Delta G_{MSA} \equiv & \sum_{i=1}^{l} \sum_{j \in AA} f(aa(j)) \Delta G_{app}^{aa(ij)} \\ & + c_0 \sqrt{(\sum_{i=1}^{l} \sum_{j \in AA} f(aa(ij)) \Delta G_{app}^{aa(ij)} \sin(100^\circ i))^2 + (\sum_{i=1}^{l} \sum_{j \in AA} f(aa(ij)) \Delta G_{app}^{aa(ij)} \cos(100^\circ i))^2} \\ & + c_1 + c_2 l + c_3 l^2 \end{aligned} \tag{4.5}$$

where $f(aa(ij))$ is amino acid composition for 20 amino acids at position i. The calculation of $f(aa(ij))$ is described below.

4.1.5 Amino acid composition

I extracted two kinds of amino acid composition: simple composition from a query sequence and PSSM composition from a PSSM of the query generated by DELTA-Blast. As described in the result section, positional difference of several amino acids is observed between different regions of TMDs. To extracting these difference, for queries complete with PSSM I sectioned the TMDs into three regions: N-terminal 5 residue, C-terminal 5 residue, and the rest in the middle. I did not section single query sequence due to the sparseness of data which would result in esimating three sets of 20 amino acid frequencies. PSSM composition, C_j^{PSSM}, is simply defined from the PSSM matrix values (log-odds scores) over each section of length L. The log-odds score of amino acid j at position i, s_{ij}, is calculated from the background probability of amino acid j. s_{ij} is scaled by

the sigmoid function.

$$S_j = \sum_{i=1}^{L} \frac{1}{1+\exp(-s_{ij})}$$
$$C_j^{PSSM} = \frac{S_j}{\sum_{k \in AA} S_k} \tag{4.6}$$

The definition of f(aa(ij)) is almost the same, but to avoid the sampling error effect of rare {amino acid, position} combinations, combinations with negative values of s_{ij} are ignored when calculating ΔG_{app}^{pred}.

4.1.6 Predictor Architecture and Training

I employed a two-layered predictor architecture to discriminate TMDs from non-TMD regions; both layers use SVM classifiers but the first layer focuses on individual positions, computing a TMD/non-TMD score based on the PSSM centered at each position, while the second layer uses the first layer as input in addition to other input such as ΔG to delineate the TMD boundaries. The Support Vector Machine (SVM) implemented in LIBSVM 3.1 with the RBF-kernel [45] are applied to both layer models.

Calculation of TMD candidate

(In addition to other features) each SVM in the second layer receives one feature summarizing a span from the first layer. Roughly speaking this span is a local minimum of ΔG. More precisely, candidate regions are searched by combination of ΔG_{MSA} and ΔG_{app}^{pred}. First, a query is scanned by ΔG_{MSA} with changing its window size from 15 to 30. The region with the minimum score is selected, and the rest region is recursively searched by the same way until no region has lower than 3. In some cases, ΔG_{MSA} misses TMDs, so ΔG_{app}^{pred} without PSSM is run in the region where

ΔG_{MSA} is higher than 3.

4.1.7 Classification performance evaluation

Matthews correlation coefficient

The detail of the Matthews correlation coefficient (MCC) is described in the materials and methods section of chapter 2.

Sensitivity and Specificity

Sensitivity equals to a measurement so-called recall. Both measurements are defined below.

$$\begin{aligned} Sensitivity &= \frac{TP}{TP+FN} \\ Specificity &= \frac{TN}{TN+FP} \end{aligned} \tag{4.7}$$

, where TP, FP, TN, and FN are True Positive, False Positive, True Negative, and False Negative, respectively.

4.2 Results

4.2.1 Mitochondrial TMDs tends to be short and less hydrophobic

At first, free energy distributions of single spanning TMD of mitochondrial inner membrane and E.R. are analyzed with Sec61 translocon model (Figure 4.1). To compare with membrane proteins, globular cytosolic proteins are used as negative samples. Pseudo-TMD region is calculated by scanning entire sequences for them, and region with lowest energy is selected. As expected, almost all of single spanning TMDs of E.R. membranes have energy less than 0 kcal/mol, which should

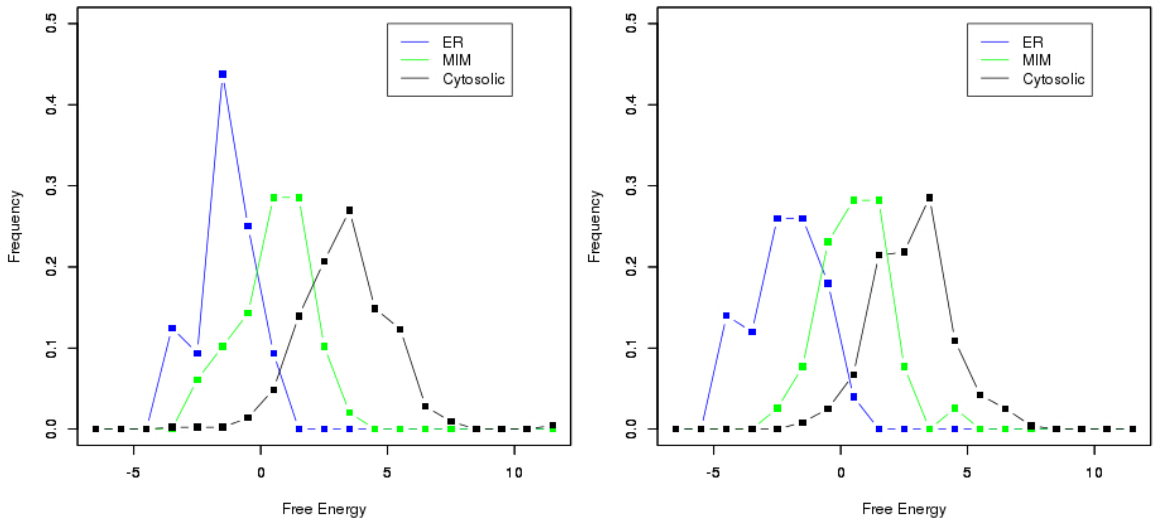


Figure 4.1: Free energy distribution of single spanning TMD measured by Sec61 translocon model. (Left) TMDs are extracted from fungi. (Right) TMDs are extracted from vertebrates.

be inserted spontaneously, and cytosolic proteins rarely have region whose energy is less than 0 kcal/mol. Discrimination between these two groups seems to be done well by the model, however, mitochondrial TMDs show ambiguous distribution. These tendencies are common between fungi and vertebrate (Figure 4.1), and prediction difficulty of mitochondrial TMDs is partly explained by this result. Half of them have strong enough TMD characters, however, the rest is overlapped with negative samples. Relation between length and energy is also analyzed (Figure 4.2). Although it is known that ΔG_{app}^{pred} is correlated length of the TMD segment, approximation of the model can reflect actual length effect [20]. Basically, mitochondrial TMDs have shorter hydrophobic segments than that of E.R. membranes. Since the model is optimized by experiment of Sec61 translocon insertion, higher free energy might simply reflect different insertion mechanism in the mitochondria.

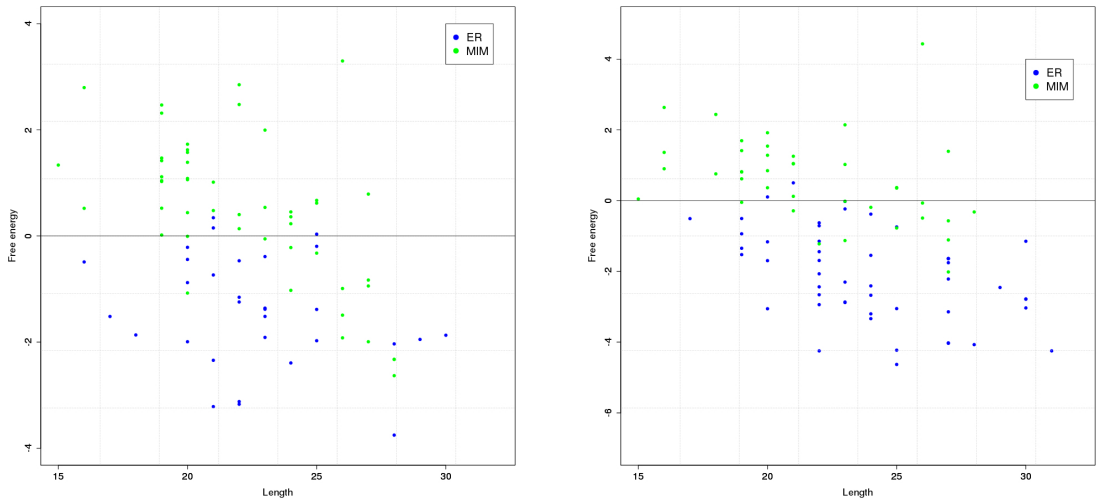


Figure 4.2: Scatter plot of free energy versus length of single spanning TMD measured by Sec61 translocon model. (Left) TMDs extracted from fungi. (Right) TMDs extracted from vertebrates.

4.2.2 Differences in amino acid composition

Since ΔG_{app}^{pred} depends on positional parameters for 20 amino acid, different energy distribution intimates either unfavorable order of amino acids such as charged residue in the middle of TMD or different amino acid composition, or both. To check composition variance, amino acid composition of mitochondrial TMDs and E.R. TMDs are compared. Since distribution of amino acid composition is not normally distributed, the Mann-Whitney U test was applied. Leu, Ile or other characteristic amino acids for TMD such as Trp, Tyr, or Arg are not significantly different in both fungi or vertebrate. Surprisingly, Gly is significantly different in both fungi and vertebrate. In fact, Gly is unfavored amino acid residue at any position in the Sec61 translocon model, high density of Gly in TMD can lead to higher free energy. Composition lacks information of relative position within TMD, therefore, positional preference was analyzed. Although one problem to make positional amino acid profile of TMD empirically is length variation, simple linear interpolation was applied to fit 19 amino acid long profile. Considering general similarities between fungi and vertebrate,

AA	Fungi	Vertebrate
A	0.01454	0.14678
C	**0.00012**	0.09667
D	**0.00000**	0.01735
E	0.00501	0.03654
F	0.00621	0.08144
G	**0.002591**	**0.00032**
H	0.24780	0.24305
I	0.02454	0.42331
K	0.70420	0.72505
L	0.00384	0.00268
M	0.09394	0.50796
N	0.63880	0.16080
P	0.36790	0.67532
Q	0.05545	0.00396
R	0.08034	0.42878
S	0.44990	0.00945
T	0.14830	0.71128
V	0.15740	0.04809
W	0.13940	0.18680
Y	0.02199	0.18219

Table 4.1: P-values of the Mann-Whitney test are listed. Amino acid compositions in TMD region in fungi or vertebrate are compared. Entries significant at the 0.05 confidence level after Holm-Bonferroni correction for multiple hypothesis testing are shown in bold.

both dataset merged into one after redundancy reduction. Figure 4.3 shows statistically calculated free energy profile of 20 amino acid and fitted energy profile for Sec61 translocon model by Gaussian function (detail is described in materials and method section). Prior work conducted by Botelho and colleagues revealed asymmetric distribution of negatively charged residue [19], and this asymmetry is observed, especially in Asp. Although empirical estimation with limited dataset can be doubtful, this empirical distribution seems to be consistent with actual profile of TMD. Significant Gly is generally observed in the middle of mitochondrial inner membrane TMDs, and small residue such as Ala or Ser look slightly abundant. Taken together, higher free energy seems to be explained by these positional difference of amino acid observation.

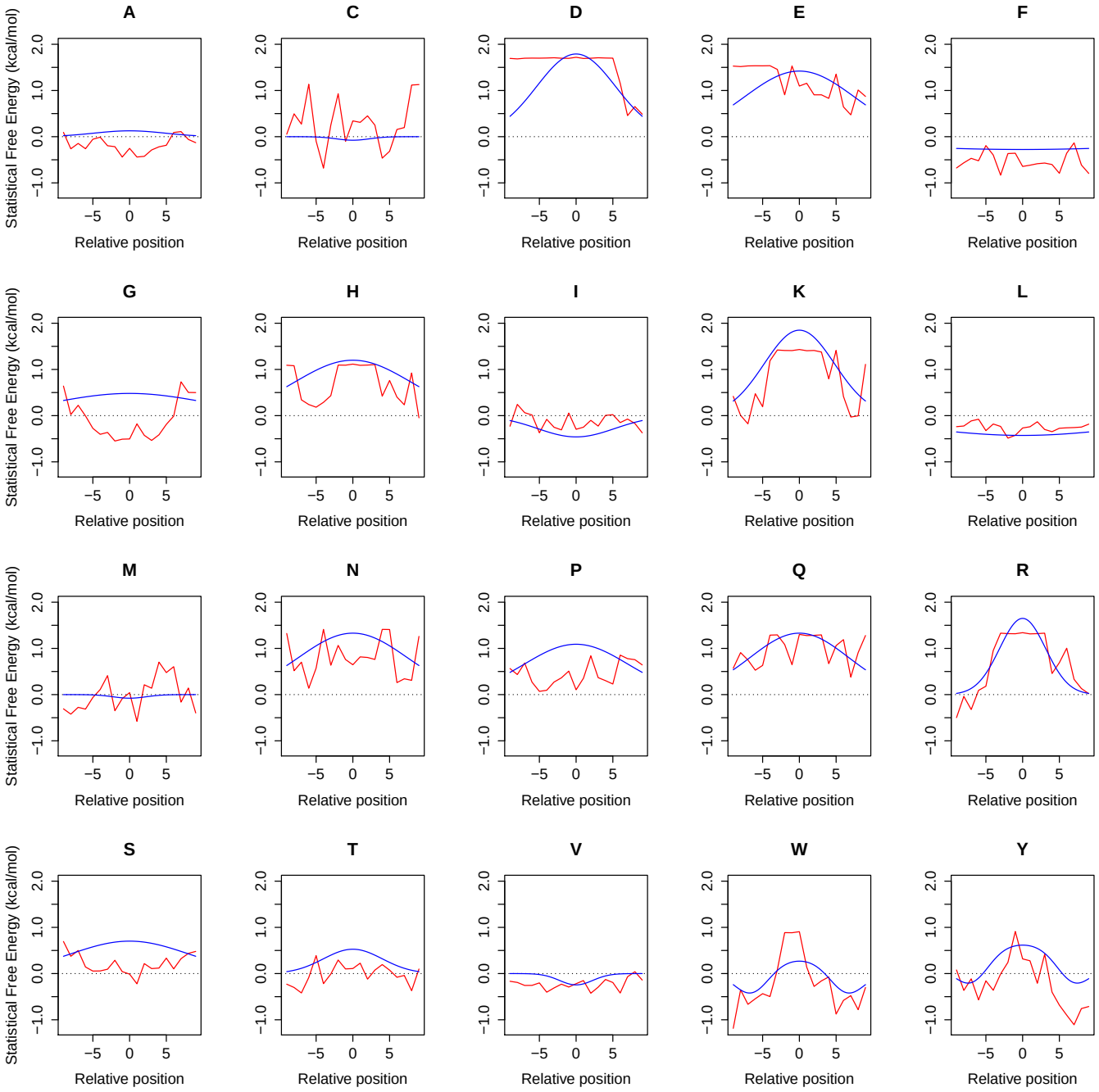


Figure 4.3: Statistical free energy for mitochondrial membrane insertion. The red line shows statistically calculated free energy, and the blue line the Gaussian function is used in the Sec61 translocon model. Negative positions indicate the mitochondrial matrix side.

Taxonomy	AUC(- PSSM composition)	AUC(+PSSM composition)
Fungi	0.8752	0.9354
Vertebrate	0.8883	0.9390

Table 4.2: ROC AUC is listed for jack-knife test in fungi and vertebrate dataset.

4.2.3 Evolutionary improves TMD region prediction

Next, application of evolutionary information is considered in addition to amino acid composition of each query or variation of free energy distribution. If homologous sequences retain conserved TMD features, evolutionary conservation seems to be informative for discrimination with plausible TMDs in globular proteins. To discuss whether taxonomy affects actual prediction result, fungi and vertebrate dataset was independently tested. First, SVM model which discriminates candidate region of query is trained with simple features, length, ΔG_{app}^{pred}, averaged hydrophobicity in GES-scale [92], and amino acid composition. Candidate TMD region is defined by Sec61 translocon model, and features such as amino acid composition is normalized within each candidate region. The result is measured by ROC AUC, which is described in chapter 2. Table 4.2 shows this model can discriminates mitochondrial TMD from plausible region relatively well.

Evolutionary information summarized in 60 dimension vector is added and tested to discuss whether or not it is informative. As a result, performance is improved as measured in AUC (Table 4.2). In addition, there seems to be no or ignorable difference among taxonomy as reported above in actual prediction.

Since homologous sequences retain discriminative information, application in different ways is also considered. One application is explicit weighting in ΔG_{app}^{pred} as weights for $\Delta G_{app}^{aa(ij)}$ [90]. This improves membrane region prediction from 0.78 to 0.91 (measured by MCC) in OPM small dataset as well.

4.2.4 Two layer predictor and benchmark

The other way is extracting features from PSSM more precisely. Evolutionary pattern might be characteristic for mitochondrial TMD, therefore, evolutionary patterns surrounding individual residues of mitochondrial TMD and those of plausible region are discriminated by SVM model. In this case, individual residues in positive dataset and negative dataset are treated positive and negative instances, respectively. In this model, features are extracted PSSM log-odds score s_{ij}, and surrounding w positions are also considered, namely $20 \times 2w$ features. With window size 33, thus surrounding w equals 16, the model was most accurate at 93.4% accuracy. Because this individual residue model returns predicted score for each residue, these scores are averaged in candidate region selected by ΔG_{MSA} or ΔG_{app}^{pred} to integrate with coarse grained model, which is described above.

To compare this model with current existing predictors, 10-fold cross validation was conducted. Although jack knife test reflects actual performance in general, the number of fold is reduced to 10 due to the computational time. Since the coarse grained predictor at upper layer depends on the predictor for individual residues at lower layer, nested cross validation is run. Benchmarked result is summarized in Table 4.3. As widely used predictors, SOSUI [93], TMHMM [94], Phobius [95], and Octopus [91] are compared. SCAMPI [90] uses $\Delta G_{app}^{aa(i)}$ internally, therefore, this predictor is also included. In general, predictors below return only predicted region not quantitative score, so measurements for binary prediction are used to evaluate each of them. TMHMM shows high specificity, and Octopus returns best sensitivity. Predictor in this work shows second best in both sensitivity and specificity, leading to best MCC.

	Sensitivity	Specificity	MCC
SOSUI	39.74%	90.55%	0.3377
TMHMM ver. 2.0	63.16%	97.46%	0.6868
Phobius	76.32%	91.27%	0.6579
SCAMPI	67.11%	82.91%	0.4592
Octopus(topcons)	96.05%	60.36%	0.4651
Predictor in this work	84.21%	95.27%	0.7911

Table 4.3: Evaluations as a binary problem was summarized in the above table. Positive examples without any TMD region or negative examples with predicted TMD were treated mistakes.

4.2.5 Feature analysis

Integration layer and individual residue layer have related but different feature, so discriminative features are listed separately (Table 4.4, 4.5). Averaged score of individual residue model within candidate region is by far the best feature in integration model. Following it, free energy values calculated from PSSM or single query are second or third best. PSSM composition of C-terminal segment (five residue long), such as C_W or C_Y, are also good features. In fact, these amino acids are also discriminative features in individual residue model. As expected, averaged hydrophobicity in GES scale or Gly composition within an entire candidate region have discriminative power and are 5th or 6th best, respectively.

Table 4.5 summarizes top thirty features of individual residue model. F-score of each feature is low and close to each other, so it seems to me that feature space is highly complex and model discriminates in high dimensional feature space. Even if taking this into consideration, high ranked features are reasonable. Trp in downstream regain tend to be high ranked, and this is consistent with general TMD characteristics. Although data is not shown, Trp in upstream region such as Trp at -7 or -8 are ranked 38th or 39th. The other point is that Gly in middle region tend to be highly ranked. If multiple sequence alignment which generates PSSM is assumed to be enough accurate, Gly, which is enriched in TMD region, is evolutionary conserved. Pro is also fairly high

Name of feature	F-score
Averaged score of individual residue model	2.066903
ΔG_{MSA}	0.24367
ΔG_{app}^{pred}	0.149455
C_W^{PSSM} in C-terminal 5	0.135296
Averaged hydrophobicity	0.134169
G composition	0.114176
C_Y^{PSSM} in C-terminal 5	0.1055
C_G^{PSSM} in the middle region	0.098735
C_H^{PSSM} in the middle region	0.080938
C_R^{PSSM} in the middle region	0.075786

Table 4.4: Features used in the integration layer, and listed ten highest ranked features among them. To rank them, discriminative power is measured by F-score.

ranked, thus, mitochondrial TMD might favors slightly unfolded helix. However, scoring matrix for sequence alignment has high score between Pro and Gly in general, so I cannot deny a possibility that this observation is biased by Gly enrichment.

4.2.6 Prediction on yeast presequence dataset

One relevant application of this work is to find a substrate candidate of inner membrane proteases. Presequence dataset used in chapter 2 and 3 provided by Vögtle *et al.* are scanned by this predictor. If presequence C-terminal end locates after TMD region, it might be cleaved by inner membrane space by a protease there. All known substrates in the dataset passed this criterion, and four other unannotated proteins were found (Table 4.6). Although it is still hard to predict which protease is relevant to unannotated proteins processing, this list provides putative substrates.

AA	relative position	F-score
W	9	0.045024
W	6	0.044292
G	0	0.042942
W	5	0.042535
W	8	0.042404
W	7	0.042307
W	4	0.040908
W	10	0.040874
G	-1	0.040341
W	3	0.038599
G	-2	0.038273
G	1	0.03652
W	11	0.036416
G	2	0.035511
W	2	0.035272
G	3	0.033509
W	1	0.032825
G	-3	0.031178
P	-1	0.030676
P	-3	0.030556
P	-2	0.030346
P	0	0.03020 9
G	-4	0.030018
G	4	0.029424
Y	9	0.029354
Y	10	0.02878
W	12	0.028127
Y	11	0.026866
Y	6	0.026678
G	5	0.026623

Table 4.5: Features used in the individual residue layer, and listed thirty highest ranked features among them. To rank them, discriminative power is measured by F-score.

OLN	UniProt AC	Cluster	Annotation
YBL095W	P38172	II	
YGR174C	P37267	III	
YIL155C	P32191	II	Gut2, IMP1
YKL150W	P36060	I	Mcr1, IMP1
YKR066C	P00431	II	Ccp1, Pcp1
YML081C-A	P81450	III	
YOR065W	P07143	II	Cyt1, IMP2
YPL103C	Q02883	II	

Table 4.6: Candidate substrate list. Already known substrates are emphasized in annotation column with gene name and protease name. Cluster number discussed in chapter 3 is also added.

4.3 Discussion

Recently, proteases in the mitochondrial inner membrane has been focused due to the discovery of PINK1 processing by PARL [24]. PARL is classified as a rhomboid like protease, which processes within or nearby TMD, and PINK1 is also cleaved within TMD at 103 by this protease. TMD prediction has been important in the context of structural biology, however, it seems to be also important as an application at least in mitochondrial processing. Problem relevant to the prediction is weak hydrophobicity of mitochondrial membrane proteins as illustrated in Figure 4.1. In fact, TMDs of PINK1 or another substrate of PARL, PGAM5 [96] have relatively higher free energy measured in ΔG_{app}^{pred}, 0.90 and 1.03, respectively. Detailed analysis of these proteins revealed that Gly located in the middle of TMD contributed higher energy, and this is consistent with statistical analysis. The reason why Gly is favored in mitochondrial TMD is elusive. Although its average frequency at the single sequence level is not statistically significantly different, when PSSMs are used the log odds ratio of Pro is discriminant (attains a high F-score). Moreover, mitochondrial TMDs are generally less hydrophobic. Taken together, the mitochondrial inner membrane might favor weak TMDs due to its membrane composition.

One weakness of predictor in this work is that this has not been optimized to the multi spanning membrane proteins because of lack of the dataset. Distribution of free energy for mitochondrial single spanning membrane proteins is similar to that of multi spanning membrane proteins [20]. It is still unknown whether TMD of multi spanning mitochondrial membrane proteins is weaker than general multi spanning TMD.

At present, topology prediction depends on positive-inside rule. As observed in Figure 4.3, negatively charged residue does not appear at matrix side of TMD. Although it is difficult to test applicability of asymmetric distribution due to the lack of dataset, this asymmetric distribution

might be informative to the topology prediction, especially for multi spanning membrane proteins in mitochondria.

Chapter 5

Conclusion

Translocation into and within mitochondria is finely regulated by several mechanisms. The most intensely researched mechanism is the presequence dependent pathway. Although perfect delineation of the features of presequences is not yet possible, several features such as evolutionary sequence divergence, positively charged amphiphilicity, and characteristic motifs including cleavage site motifs are analyzed in detail and refined as discriminative features in this work.

Feature related to the sequence divergence sheds light on usefulness of weakly conserved regions of proteins, which are usually discarded as non-informative region. As discussed in the chapter 2, N-terminal signal is tend to be diverged among orthologous proteins, and this divergence is a novel feature for signal prediction. Defining ortholog is a key and not easy task for this kind of analysis due to gene or genome duplication, however, simple reciprocal best hit approach shows enough accurate in our signal prediction. Although combination of sequence divergence and the classical features did not gain significantly better accuracy, sequence divergence alone has discriminative power for N-terminal signals. In fact, this novel feature finds a few mitochondrial signal, which are

difficult to be predicted, and misannotated proteins in the curated database.

Among N-terminal signals, mitochondrial presequence is an important in terms of scientific and industrial applications, however, prediction accuracy seems to be not enough with the classical features such as used in the chapter refchap:Divergence. To develop a competitive tool in this field of study, I refined and introduced several features of the presequence. Quantification of positively charged amphiphilicity is improved by taking positive charge in the opposite side against hydrophobic surface into the consideration, and several novel motifs are also discovered. In particular, refinement of cleavage site motifs archives the highest discriminative power in signal prediction features, and this is also confirmed by cleavage site prediction accuracy. Low accuracy of cleavage site prediction of mitochondrial presequence has been argued in mitochondrial biology, it seems that MitoFates is a first tool which shows practical performance. In summary, I achieved significant improvement over previous predictors in discrimination of classical presequences and in particular their cleavage sites by MPP and its relevant intermediate proteases.

However, as shown in the clustering results reported in chapter 3, some presequence are still difficult to predict, namely cluster II and III, although some of them cluster II proteins, such as MrpL32 or Imo32, can be predicted based on evolutionary sequence divergence. Cluster II includes several proteins which are cleaved by proteases located in the mitochondrial inner membrane and thus further elucidation of this class of proteases and their substrates is a promising line of future work. The work on mitochondrial Transmembrane domain (TMD) prediction I present in chapter 4 will provide a basis for this; providing significantly improved TMD prediction by explicit modeling of the differences between mitochondrial inner membrane other membranes.

The difficulty of predicting the TMDs of mitochondrial membrane proteins has been discussed anecdotally, it has been unknown which character affects prediction difficulty. With the model for

E.R. membrane insertion, I argued that about half of mitochondrial TMDs has shorter and different amino acid content leading to higher free energy. There are two key points in terms of sequence characteristics: high proportion of glycine residue in center of TMDs and asymmetric observation for negatively charged residues. Since these tendencies are conserved among evolutionary relevant sequences, mitochondrial membrane proteins seem to be exposed to different evolutionary pressure. Explicit modeling of mitochondrial TMDs led to better discriminative performance, therefore, I can conclude that difficulty of TMD prediction for mitochondrial proteins can be explained by such differences.

Appendix A

Appendix for sequence divergence of targeting signals

Measure of influence of divergence features:

As reported in the results section, I performed a post-hoc analysis of proteins for which the divergence features greatly influenced the prediction outcome. This requires a concrete, quantitative measure of that influence, which I chose to define in terms of a numerical score known as exponential loss-based decoding [1].

For each protein, and each of two feature sets (with and without divergence features), I compute a probability vector P estimating the probability that the protein is a member of each of the three sorting classes {SP, MTS, N-signal-free}. I then use the Jensen-Shannon divergence as a quantitative measure of how much the two probability vectors (predictions with and without divergence features) differ.

The Jensen-Shannon divergence is a standard measure of distance between two probability distributions. The definition is:

$$JSD(P_{\text{div}}||P_{\text{nodiv}}) \;=\; \frac{1}{2}D(P_{\text{div}}||M) + \frac{1}{2}D(P_{\text{nodiv}}||M) \tag{A.1}$$

where $M = \frac{1}{2}(P_{\text{div}} + P_{\text{nodiv}})$ and $D(P||Q)$ indicates the Kullback-Leibler divergence:

$$D_{\text{KL}}(P||Q) \;=\; \sum_i P(i)\log_2\frac{P(i)}{Q(i)} \tag{A.2}$$

The precise method I used to compute P, the probability vector over classes for a given protein, is somewhat involved. I first used all of the yeast YGOB data to train three binary SVM classifiers {0:1, 0:2, 1:2}, where the integers {0,1,2} to denote the three classes {SP, MTS, N-signal-free}. For each protein instance, each SVM classifier outputs a score related to the classification margin, which roughly reflects the confidence of its prediction. Let s_{ij} denote the score for the SVM discriminating between classes i and j, so for example s_{12} denotes the score of the SVM discriminating between MTS and N-signal-free, with a large positive value indicating a strong prediction of MTS, and a large negative value a strong prediction of N-signal-free. Following the exponential loss function described in [1], I define P as:

$$P \;\propto\; e^{s_{01}} + e^{-s_{01}} \;+\; e^{s_{02}} + e^{-s_{02}} \;+\; e^{s_{12}} + e^{-s_{12}} \tag{A.3}$$

I compute P with this equation and then linearly normalize so that its elements sum to one.

Rank	UniProt AC	JSD	Annotation	Prediction$_{all}$	Prediction$_{nodiv}$
1	P40825—SYA (Ala1)	0.088	Non-Signal	MTS	Non-Signal
2	P13099—RL10	0.063	Non-Signal	Non-Signal	MTS
3	P32504—CBF3A	0.053	Non-Signal	MTS	Non-Signal
4	P53875—RM19(MrpL19)	0.048	MTS	Non-Signal	MTS
5	P32523—PRP19	0.048	Non-Signal	Non-Signal	MTS
6	P47123—MOG1	0.038	Non-Signal	MTS	Non-Signal
7	P53219—IMO32	0.037	MTS	MTS	Non-Signal
8	P40957—MAD1	0.034	Non-Signal	MTS	Non-Signal
9	Q02792—XRN2(RAT1)	0.032	Non-Signal	Non-Signal	MTS
10	P38228—TCM62	0.032	MTS	Non-Signal	MTS
11	P12687—RM02(MRP7)	0.032	MTS	MTS	Non-Signal
12	P32324—EF2(EFT1)	0.03	Non-Signal	Non-Signal	MTS
13	Q12019—MDN1	0.029	Non-Signal	MTS	Non-Signal
14	Q01163—RT23(RSM23)	0.026	MTS	SP	MTS
15	P53727—BUD17	0.026	Non-Signal	Non-Signal	MTS
16	P46672—G4P1(ARC1)	0.025	Non-Signal	MTS	Non-Signal
17	P16862—K6PF2(PFK2)	0.019	Non-Signal	Non-Signal	MTS
18	P09620—KEX1	0.017	SP	Non-Signal	SP
19	P25044—PTP1	0.016	Non-Signal	MTS	Non-Signal
20	P32333—MOT1	0.015	Non-Signal	Non-Signal	MTS
21	P25039—EFGM(MEF1)	0.015	MTS	MTS	Non-Signal
22	Q12428—PRPD(PDH1)	0.013	MTS	MTS	Non-Signal
23	P10663—RT02(MRP2)	0.012	MTS	Non-Signal	MTS
24	P41338—THIL(ERG10)	0.012	Non-Signal	Non-Signal	MTS
25	P25348—RM32(MRPL32)	0.012	MTS	MTS	Non-Signal
26	Q03691—ROT1	0.011	SP	SP	Non-Signal
27	P39927—PTI1	0.01	Non-Signal	Non-Signal	MTS
28	Q12031—ACEB(ICL2)	0.009	MTS	MTS	Non-Signal
29	P40008—FMP52	0.008	MTS	Non-Signal	MTS
30	P28007—GAR1	0.007	Non-Signal	MTS	Non-Signal
31	P32898—CYM1	0.006	MTS	Non-Signal	MTS
32	P00958—SYMC(MES1)	0.006	Non-Signal	Non-Signal	MTS
33	P39735—SAW1	0.006	Non-Signal	Non-Signal	MTS
34	P36046—MIA40	0.005	MTS	MTS	Non-Signal
35	P35189—TAF14	0.005	Non-Signal	MTS	Non-Signal
36	P40018—RSMB(SMB1)	0.004	Non-Signal	Non-Signal	MTS
37	P43605—ECO1	0.003	Non-Signal	MTS	Non-Signal
38	P61830—H3(HHT[12])	0.003	Non-Signal	Non-Signal	MTS
39	P41805—RL10	0.003	Non-Signal	Non-Signal	MTS
40	P00447—SODM(SOD2)	0.002	MTS	SP	MTS
41	P36517—RM04(MRPL4)	0.002	MTS	MTS	Non-Signal
42	P38719—DBP8	0.001	Non-Signal	Non-Signal	MTS
43	P08524—FPPS(ERG20)	0.001	Non-Signal	MTS	Non-Signal

Table A.1: Ranking of proteins whose prediction score is affected by divergence feature addition. Selected examples are discussed in discussion section of the main text.

A.1 Divergence score combined with standard features in N-terminal 40 residues

A.1.1 *S. cerevisiae*, curated orthologs (N_{40})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.67 ± 0.03	0.36 ± 0.06	0.86 ± 0.02	0.72 ± 0.04	0.86 ± 0.05	0.73 ± 0.07
SP	0.50 ± 0.00	0.00 ± 0.00	0.77 ± 0.02	0.62 ± 0.07	0.78 ± 0.11	0.62 ± 0.21
N-signal-free	0.66 ± 0.02	0.36 ± 0.03	0.84 ± 0.02	0.69 ± 0.03	0.85 ± 0.05	0.71 ± 0.07
% accuracy	70.82 ± 1.61		85.19 ± 1.36		85.77 ± 3.15	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.67 ± 0.02	0.42 ± 0.06	0.84 ± 0.03	0.71 ± 0.05	0.86 ± 0.04	0.77 ± 0.07
SP	0.67 ± 0.11	0.50 ± 0.22	0.90 ± 0.05	0.82 ± 0.07	0.88 ± 0.04	0.79 ± 0.08
N-signal-free	0.67 ± 0.02	0.42 ± 0.04	0.86 ± 0.02	0.73 ± 0.03	$\mathbf{0.87} \pm 0.03$	$\mathbf{0.77} \pm 0.05$
% accuracy	74.78 ± 1.78		87.39 ± 0.95		89.15 ± 1.93	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.80 ± 0.03	0.65 ± 0.06	$\mathbf{0.87} \pm 0.05$	0.75 ± 0.08	0.84 ± 0.03	0.69 ± 0.06
SP	0.78 ± 0.07	0.66 ± 0.11	0.79 ± 0.12	0.64 ± 0.22	0.82 ± 0.05	0.72 ± 0.10
N-signal-free	0.79 ± 0.02	0.63 ± 0.04	0.85 ± 0.04	0.72 ± 0.07	0.83 ± 0.03	0.68 ± 0.06
% accuracy	82.99 ± 1.66		86.50 ± 3.20		85.04 ± 2.73	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	$\mathbf{0.87} \pm 0.03$	$\mathbf{0.78} \pm 0.04$	0.86 ± 0.04	0.76 ± 0.07	0.86 ± 0.01	0.74 ± 0.02
SP	0.80 ± 0.08	0.70 ± 0.09	0.90 ± 0.07	0.79 ± 0.08	0.89 ± 0.06	0.82 ± 0.06
N-signal-free	0.85 ± 0.03	0.74 ± 0.04	$\mathbf{0.87} \pm 0.02$	$\mathbf{0.77} \pm 0.04$	$\mathbf{0.87} \pm 0.02$	0.75 ± 0.03
% accuracy	87.97 ± 1.25		89.15 ± 1.91		88.27 ± 1.29	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	$\mathbf{0.87} \pm 0.03$	0.76 ± 0.05	$\mathbf{0.87} \pm 0.03$	0.77 ± 0.03	$\mathbf{0.87} \pm 0.03$	0.77 ± 0.03
SP	0.81 ± 0.08	0.70 ± 0.11	$\mathbf{0.91} \pm 0.06$	$\mathbf{0.85} \pm 0.06$	0.90 ± 0.06	0.83 ± 0.08
N-signal-free	0.85 ± 0.03	0.72 ± 0.05	$\mathbf{0.87} \pm 0.02$	$\mathbf{0.77} \pm 0.03$	$\mathbf{0.87} \pm 0.02$	$\mathbf{0.77} \pm 0.02$
% accuracy	87.24 ± 1.86		$\mathbf{89.44} \pm 1.12$		89.30 ± 0.66	

Table A.2: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast curated ortholog dataset.

A.1.2 *S. cerevisiae*, RBH orthologs (N_{40})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.65 ± 0.04	0.34 ± 0.08	0.85 ± 0.03	0.72 ± 0.08	0.87 ± 0.03	0.75 ± 0.05
SP	0.50 ± 0.00	0.00 ± 0.00	0.81 ± 0.04	0.66 ± 0.06	0.85 ± 0.04	0.75 ± 0.04
N-signal-free	0.64 ± 0.04	0.33 ± 0.10	0.85 ± 0.02	0.71 ± 0.05	0.87 ± 0.02	0.75 ± 0.04
% accuracy	70.06 ± 3.05		85.44 ± 2.83		87.79 ± 1.83	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.66 ± 0.03	0.41 ± 0.07	0.85 ± 0.04	0.74 ± 0.08	0.87 ± 0.03	0.77 ± 0.06
SP	0.74 ± 0.08	0.65 ± 0.13	0.85 ± 0.05	0.76 ± 0.09	0.88 ± 0.03	0.80 ± 0.06
N-signal-free	0.68 ± 0.01	0.45 ± 0.03	0.87 ± 0.02	0.77 ± 0.05	0.88 ± 0.03	0.79 ± 0.06
% accuracy	75.94 ± 1.12		88.14 ± 2.03		89.43 ± 2.47	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.78 ± 0.04	0.61 ± 0.06	0.87 ± 0.03	0.76 ± 0.07	0.86 ± 0.02	0.72 ± 0.05
SP	0.82 ± 0.08	0.75 ± 0.14	0.86 ± 0.03	0.75 ± 0.02	0.87 ± 0.05	0.79 ± 0.06
N-signal-free	0.80 ± 0.04	0.65 ± 0.07	0.87 ± 0.03	0.76 ± 0.06	0.85 ± 0.03	0.71 ± 0.07
% accuracy	83.45 ± 3.23		88.15 ± 2.64		86.62 ± 2.85	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.87 ± 0.05	0.77 ± 0.09	0.87 ± 0.03	0.76 ± 0.06	0.86 ± 0.05	0.75 ± 0.08
SP	0.90 ± 0.04	0.85 ± 0.06	0.88 ± 0.04	0.78 ± 0.07	0.91 ± 0.03	0.87 ± 0.05
N-signal-free	0.87 ± 0.05	0.77 ± 0.09	0.89 ± 0.03	0.79 ± 0.06	0.88 ± 0.04	0.79 ± 0.06
% accuracy	89.44 ± 3.81		89.20 ± 2.52		89.67 ± 2.72	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.87 ± 0.04	0.76 ± 0.07	0.88 ± 0.02	$\mathbf{0.80} \pm 0.04$	$\mathbf{0.89} \pm 0.03$	$\mathbf{0.80} \pm 0.05$
SP	0.89 ± 0.03	0.84 ± 0.07	$\mathbf{0.93} \pm 0.02$	$\mathbf{0.90} \pm 0.02$	$\mathbf{0.93} \pm 0.02$	0.89 ± 0.06
N-signal-free	0.87 ± 0.04	0.76 ± 0.09	$\mathbf{0.90} \pm 0.02$	0.82 ± 0.04	$\mathbf{0.90} \pm 0.02$	$\mathbf{0.83} \pm 0.04$
% accuracy	89.08 ± 3.48		91.31 ± 1.63		$\mathbf{91.67} \pm 1.63$	

Table A.3: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast automatically collected dataset.

A.1.3 Human, RBH orthologs (N_{40})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.52 ± 0.03	0.10 ± 0.18	0.79 ± 0.06	0.61 ± 0.16	0.84 ± 0.02	0.69 ± 0.06
SP	0.65 ± 0.05	0.29 ± 0.10	0.82 ± 0.04	0.65 ± 0.07	0.86 ± 0.03	0.74 ± 0.06
N-signal-free	0.66 ± 0.05	0.35 ± 0.09	0.87 ± 0.03	0.74 ± 0.06	0.89 ± 0.03	0.80 ± 0.07
% accuracy	65.11 ± 3.55		83.01 ± 4.17		87.07 ± 3.25	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.67 ± 0.07	0.43 ± 0.15	0.81 ± 0.03	0.65 ± 0.07	0.82 ± 0.05	0.69 ± 0.08
SP	0.75 ± 0.06	0.52 ± 0.12	0.87 ± 0.03	0.75 ± 0.07	0.89 ± 0.03	0.79 ± 0.07
N-signal-free	0.75 ± 0.06	0.52 ± 0.12	0.88 ± 0.02	0.78 ± 0.06	$\mathbf{0.91} \pm 0.03$	0.82 ± 0.07
% accuracy	74.89 ± 5.49		86.32 ± 3.25		88.42 ± 3.39	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.73 ± 0.07	0.53 ± 0.11	0.84 ± 0.02	0.69 ± 0.06	0.84 ± 0.05	0.69 ± 0.09
SP	0.81 ± 0.05	0.62 ± 0.06	0.87 ± 0.04	0.75 ± 0.07	0.86 ± 0.05	0.73 ± 0.08
N-signal-free	0.81 ± 0.04	0.64 ± 0.05	0.90 ± 0.04	0.80 ± 0.08	0.89 ± 0.04	0.79 ± 0.09
% accuracy	79.85 ± 1.35		87.22 ± 3.80		86.77 ± 4.47	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.85 ± 0.02	0.72 ± 0.07	0.83 ± 0.05	0.70 ± 0.08	0.84 ± 0.04	0.69 ± 0.09
SP	0.86 ± 0.04	0.73 ± 0.08	0.89 ± 0.03	0.79 ± 0.07	0.88 ± 0.05	0.77 ± 0.07
N-signal-free	0.88 ± 0.04	0.78 ± 0.08	$\mathbf{0.91} \pm 0.03$	$\mathbf{0.83} \pm 0.08$	0.90 ± 0.03	0.81 ± 0.07
% accuracy	86.92 ± 3.86		88.87 ± 3.62		87.67 ± 3.59	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	$\mathbf{0.87} \pm 0.02$	$\mathbf{0.76} \pm 0.07$	0.84 ± 0.05	0.72 ± 0.06	0.86 ± 0.04	0.74 ± 0.05
SP	0.87 ± 0.03	0.75 ± 0.06	$\mathbf{0.90} \pm 0.04$	$\mathbf{0.80} \pm 0.06$	0.89 ± 0.03	0.79 ± 0.06
N-signal-free	0.90 ± 0.04	0.80 ± 0.09	$\mathbf{0.91} \pm 0.04$	$\mathbf{0.83} \pm 0.08$	$\mathbf{0.91} \pm 0.04$	$\mathbf{0.83} \pm 0.08$
% accuracy	88.12 ± 3.58		89.17 ± 3.51		$\mathbf{89.32} \pm 3.25$	

Table A.4: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the mammal automatically collected dataset.

A.1.4 Plant model organisms, RBH orthologs (N_{40})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.61 ± 0.05	0.30 ± 0.15	0.70 ± 0.11	0.40 ± 0.20	$\mathbf{0.81} \pm 0.06$	0.61 ± 0.09
SP	0.50 ± 0.00	0.00 ± 0.00	0.56 ± 0.08	0.12 ± 0.18	0.69 ± 0.19	0.44 ± 0.44
CTP	0.78 ± 0.08	0.54 ± 0.15	0.78 ± 0.05	0.55 ± 0.09	0.85 ± 0.05	0.68 ± 0.10
N-signal-free	0.80 ± 0.05	0.60 ± 0.09	0.85 ± 0.03	0.70 ± 0.05	0.86 ± 0.04	0.73 ± 0.05
% accuracy	65.05 ± 6.00		69.10 ± 4.46		76.81 ± 3.53	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.58 ± 0.06	0.19 ± 0.15	0.65 ± 0.10	0.32 ± 0.23	0.79 ± 0.08	0.58 ± 0.14
SP	0.69 ± 0.14	0.48 ± 0.29	0.66 ± 0.17	0.36 ± 0.36	0.76 ± 0.25	0.50 ± 0.47
CTP	0.68 ± 0.08	0.35 ± 0.15	0.80 ± 0.05	0.58 ± 0.10	0.84 ± 0.04	0.67 ± 0.09
N-signal-free	0.66 ± 0.05	0.31 ± 0.08	$\mathbf{0.90} \pm 0.02$	$\mathbf{0.82} \pm 0.04$	0.88 ± 0.02	0.79 ± 0.03
% accuracy	53.33 ± 7.70		72.40 ± 4.28		77.93 ± 2.41	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.63 ± 0.04	0.32 ± 0.11	0.80 ± 0.08	0.59 ± 0.12	0.68 ± 0.11	0.39 ± 0.23
SP	0.70 ± 0.22	0.48 ± 0.46	0.69 ± 0.19	0.41 ± 0.41	0.83 ± 0.20	0.63 ± 0.36
CTP	0.78 ± 0.05	0.53 ± 0.11	0.83 ± 0.04	0.65 ± 0.08	0.75 ± 0.07	0.48 ± 0.14
N-signal-free	0.82 ± 0.06	0.63 ± 0.10	0.87 ± 0.04	0.76 ± 0.06	0.85 ± 0.01	0.72 ± 0.02
% accuracy	67.30 ± 4.19		76.08 ± 3.11		69.51 ± 6.82	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.77 ± 0.09	0.53 ± 0.15	0.79 ± 0.08	0.57 ± 0.10	0.69 ± 0.04	0.42 ± 0.07
SP	$\mathbf{0.86} \pm 0.21$	$\mathbf{0.68} \pm 0.38$	0.76 ± 0.26	0.52 ± 0.51	0.76 ± 0.25	0.53 ± 0.49
CTP	0.83 ± 0.05	0.66 ± 0.09	0.85 ± 0.05	0.68 ± 0.09	0.77 ± 0.06	0.53 ± 0.11
N-signal-free	0.86 ± 0.02	0.73 ± 0.05	0.89 ± 0.02	0.79 ± 0.04	0.89 ± 0.05	0.79 ± 0.10
% accuracy	76.09 ± 5.59		78.29 ± 3.16		72.76 ± 4.43	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.76 ± 0.08	0.52 ± 0.13	0.79 ± 0.07	0.62 ± 0.12	0.80 ± 0.06	$\mathbf{0.63} \pm 0.09$
SP	$\mathbf{0.86} \pm 0.21$	$\mathbf{0.68} \pm 0.38$	0.80 ± 0.22	0.61 ± 0.41	0.83 ± 0.24	0.65 ± 0.44
CTP	0.83 ± 0.05	0.65 ± 0.09	$\mathbf{0.86} \pm 0.03$	$\mathbf{0.70} \pm 0.07$	$\mathbf{0.86} \pm 0.03$	$\mathbf{0.70} \pm 0.06$
N-signal-free	0.86 ± 0.03	0.72 ± 0.05	$\mathbf{0.90} \pm 0.06$	0.81 ± 0.11	0.89 ± 0.03	0.81 ± 0.05
% accuracy	75.73 ± 4.63		80.11 ± 5.87		$\mathbf{80.50} \pm 3.93$	

Table A.5: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the plant automatically collected dataset.

A.1.5 *S. cerevisiae*, curated orthologs – classes balanced (N_{40})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.67 ± 0.10	0.35 ± 0.20	0.85 ± 0.07	0.67 ± 0.15	0.81 ± 0.07	0.61 ± 0.12
SP	0.71 ± 0.09	0.41 ± 0.16	0.88 ± 0.08	0.75 ± 0.15	0.88 ± 0.05	0.76 ± 0.10
N-signal-free	0.79 ± 0.07	0.60 ± 0.13	0.78 ± 0.10	0.60 ± 0.20	0.76 ± 0.11	0.54 ± 0.22
% accuracy	62.86 ± 5.84		78.02 ± 8.75		75.54 ± 7.94	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.75 ± 0.10	0.49 ± 0.18	$\mathbf{0.90} \pm 0.07$	$\mathbf{0.80} \pm 0.13$	0.83 ± 0.08	0.69 ± 0.15
SP	0.80 ± 0.05	0.61 ± 0.10	0.92 ± 0.03	0.84 ± 0.07	0.93 ± 0.03	0.84 ± 0.07
N-signal-free	0.70 ± 0.06	0.40 ± 0.13	0.86 ± 0.07	0.72 ± 0.14	0.85 ± 0.12	0.70 ± 0.23
% accuracy	66.69 ± 7.71		85.56 ± 6.42		82.44 ± 8.37	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.80 ± 0.06	0.61 ± 0.14	0.80 ± 0.07	0.60 ± 0.14	0.86 ± 0.09	0.72 ± 0.16
SP	0.85 ± 0.04	0.70 ± 0.08	0.89 ± 0.03	0.77 ± 0.07	0.89 ± 0.03	0.78 ± 0.06
N-signal-free	0.81 ± 0.08	0.63 ± 0.12	0.76 ± 0.12	0.52 ± 0.23	0.76 ± 0.08	0.56 ± 0.16
% accuracy	76.13 ± 6.37		75.52 ± 7.95		78.65 ± 5.90	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.84 ± 0.07	0.68 ± 0.13	0.84 ± 0.07	0.72 ± 0.12	0.86 ± 0.08	0.73 ± 0.15
SP	0.91 ± 0.06	0.82 ± 0.08	0.93 ± 0.03	0.86 ± 0.07	$\mathbf{0.94} \pm 0.04$	$\mathbf{0.89} \pm 0.06$
N-signal-free	0.78 ± 0.09	0.57 ± 0.16	$\mathbf{0.87} \pm 0.10$	0.73 ± 0.20	0.85 ± 0.10	0.69 ± 0.17
% accuracy	79.27 ± 4.61		84.31 ± 7.59		84.31 ± 6.18	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.84 ± 0.07	0.68 ± 0.13	0.86 ± 0.04	0.74 ± 0.08	0.88 ± 0.05	0.78 ± 0.09
SP	0.92 ± 0.05	0.85 ± 0.10	0.93 ± 0.03	0.85 ± 0.06	$\mathbf{0.94} \pm 0.01$	0.88 ± 0.03
N-signal-free	0.78 ± 0.09	0.57 ± 0.18	0.85 ± 0.08	0.72 ± 0.15	0.86 ± 0.07	$\mathbf{0.74} \pm 0.13$
% accuracy	79.92 ± 5.54		84.29 ± 4.35		$\mathbf{86.19} \pm 4.67$	

Table A.6: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast balanced dataset of curated orthologs.

A.1.6 *S. cerevisiae*, RBH orthologs – classes balanced (N_{40})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.65 ± 0.09	0.31 ± 0.18	0.85 ± 0.05	0.70 ± 0.09	0.82 ± 0.06	0.65 ± 0.11
SP	0.60 ± 0.07	0.19 ± 0.14	0.97 ± 0.03	0.94 ± 0.06	0.95 ± 0.04	0.89 ± 0.07
N-signal-free	0.66 ± 0.08	0.35 ± 0.15	0.86 ± 0.05	0.74 ± 0.10	0.84 ± 0.04	0.69 ± 0.08
% accuracy	51.63 ± 7.21		85.87 ± 3.29		82.67 ± 5.16	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.70 ± 0.06	0.39 ± 0.11	$\mathbf{0.88} \pm 0.03$	0.76 ± 0.03	0.86 ± 0.04	0.74 ± 0.08
SP	0.81 ± 0.09	0.63 ± 0.18	0.98 ± 0.04	0.96 ± 0.06	0.97 ± 0.03	0.93 ± 0.06
N-signal-free	0.69 ± 0.04	0.40 ± 0.10	$\mathbf{0.90} \pm 0.03$	$\mathbf{0.80} \pm 0.07$	0.88 ± 0.03	0.75 ± 0.07
% accuracy	64.40 ± 7.56		89.04 ± 1.05		86.78 ± 3.68	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.78 ± 0.03	0.55 ± 0.07	0.83 ± 0.06	0.68 ± 0.10	0.86 ± 0.05	0.71 ± 0.09
SP	0.87 ± 0.05	0.76 ± 0.10	0.96 ± 0.04	0.92 ± 0.08	0.98 ± 0.02	0.96 ± 0.04
N-signal-free	0.81 ± 0.04	0.62 ± 0.08	0.87 ± 0.03	0.74 ± 0.07	0.84 ± 0.05	0.71 ± 0.07
% accuracy	75.81 ± 4.63		84.96 ± 5.17		85.87 ± 3.29	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.84 ± 0.04	0.69 ± 0.05	0.87 ± 0.03	$\mathbf{0.77} \pm 0.06$	0.87 ± 0.07	0.74 ± 0.13
SP	$\mathbf{0.99} \pm 0.01$	$\mathbf{0.98} \pm 0.03$	0.98 ± 0.03	0.97 ± 0.04	0.98 ± 0.02	0.97 ± 0.03
N-signal-free	0.84 ± 0.04	0.69 ± 0.07	$\mathbf{0.90} \pm 0.04$	0.78 ± 0.09	0.89 ± 0.05	0.78 ± 0.10
% accuracy	85.39 ± 2.57		$\mathbf{89.06} \pm 3.32$		88.58 ± 5.80	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.83 ± 0.04	0.67 ± 0.06	0.87 ± 0.06	0.74 ± 0.10	$\mathbf{0.88} \pm 0.05$	0.76 ± 0.08
SP	0.98 ± 0.02	0.96 ± 0.04	0.98 ± 0.03	0.96 ± 0.04	0.98 ± 0.03	0.96 ± 0.04
N-signal-free	0.84 ± 0.03	0.69 ± 0.04	0.87 ± 0.07	0.75 ± 0.12	0.88 ± 0.06	0.76 ± 0.11
% accuracy	84.47 ± 2.51		87.67 ± 4.70		88.14 ± 4.34	

Table A.7: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast balanced dataset of automatically collected orthologs.

A.1.7 Human, RBH orthologs – classes balanced (N_{40})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.66 ± 0.05	0.31 ± 0.11	0.85 ± 0.05	0.70 ± 0.08	0.86 ± 0.07	0.71 ± 0.14
SP	0.70 ± 0.08	0.40 ± 0.15	0.82 ± 0.05	0.64 ± 0.10	0.82 ± 0.06	0.66 ± 0.11
N-signal-free	0.69 ± 0.06	0.39 ± 0.11	0.86 ± 0.04	0.74 ± 0.09	0.88 ± 0.05	0.78 ± 0.07
% accuracy	57.61 ± 4.71		79.42 ± 2.96		80.68 ± 4.13	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.81 ± 0.07	0.61 ± 0.13	0.87 ± 0.05	0.72 ± 0.11	0.87 ± 0.06	0.72 ± 0.11
SP	0.72 ± 0.08	0.43 ± 0.15	0.86 ± 0.06	0.73 ± 0.11	0.86 ± 0.05	0.73 ± 0.08
N-signal-free	0.77 ± 0.07	0.54 ± 0.11	0.87 ± 0.07	0.76 ± 0.13	0.88 ± 0.06	0.80 ± 0.10
% accuracy	68.36 ± 6.49		82.30 ± 6.31		82.70 ± 4.04	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.77 ± 0.05	0.54 ± 0.09	0.87 ± 0.08	0.73 ± 0.14	0.86 ± 0.03	0.73 ± 0.06
SP	0.74 ± 0.06	0.48 ± 0.12	0.83 ± 0.07	0.67 ± 0.14	0.87 ± 0.04	0.73 ± 0.08
N-signal-free	0.77 ± 0.06	0.55 ± 0.10	0.89 ± 0.05	0.79 ± 0.08	0.88 ± 0.05	0.78 ± 0.07
% accuracy	67.92 ± 4.08		81.90 ± 4.64		82.73 ± 3.04	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.86 ± 0.04	0.72 ± 0.07	0.87 ± 0.06	0.72 ± 0.11	$\mathbf{0.89} \pm 0.05$	$\mathbf{0.77} \pm 0.09$
SP	0.83 ± 0.07	0.67 ± 0.12	0.87 ± 0.04	0.75 ± 0.07	$\mathbf{0.90} \pm 0.03$	$\mathbf{0.81} \pm 0.05$
N-signal-free	0.88 ± 0.03	0.78 ± 0.05	0.88 ± 0.07	0.78 ± 0.11	$\mathbf{0.90} \pm 0.04$	0.82 ± 0.07
% accuracy	81.07 ± 2.66		83.12 ± 4.48		$\mathbf{86.44} \pm 3.37$	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.87 ± 0.05	0.73 ± 0.10	0.88 ± 0.03	0.76 ± 0.05	0.88 ± 0.03	0.76 ± 0.05
SP	0.85 ± 0.04	0.71 ± 0.08	0.87 ± 0.04	0.75 ± 0.09	0.88 ± 0.05	0.76 ± 0.09
N-signal-free	0.88 ± 0.03	0.78 ± 0.04	$\mathbf{0.90} \pm 0.05$	0.82 ± 0.05	$\mathbf{0.90} \pm 0.05$	$\mathbf{0.83} \pm 0.07$
% accuracy	82.32 ± 2.99		84.78 ± 3.40		85.20 ± 3.90	

Table A.8: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the mammal balanced dataset of automatically collected orthologs.

A.1.8 Plant model organisms, RBH orthologs – classes balanced (N_{40})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.66 ± 0.08	0.35 ± 0.14	0.78 ± 0.07	0.55 ± 0.13	0.76 ± 0.05	0.51 ± 0.09
CTP	0.77 ± 0.12	0.51 ± 0.23	0.79 ± 0.04	0.59 ± 0.10	0.80 ± 0.07	0.61 ± 0.16
N-signal-free	0.81 ± 0.09	0.67 ± 0.13	0.83 ± 0.10	0.69 ± 0.17	0.85 ± 0.05	0.72 ± 0.09
% accuracy	66.22 ± 10.11		73.27 ± 5.63		73.80 ± 4.73	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.62 ± 0.04	0.23 ± 0.08	0.76 ± 0.09	0.53 ± 0.17	$\mathbf{0.87} \pm 0.03$	0.73 ± 0.07
CTP	0.64 ± 0.03	0.28 ± 0.06	0.77 ± 0.11	0.53 ± 0.22	$\mathbf{0.84} \pm 0.10$	$\mathbf{0.68} \pm 0.18$
N-signal-free	0.66 ± 0.05	0.34 ± 0.09	0.90 ± 0.04	0.81 ± 0.08	0.91 ± 0.05	0.84 ± 0.10
% accuracy	51.94 ± 3.16		74.37 ± 8.44		83.14 ± 6.59	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.69 ± 0.07	0.38 ± 0.14	0.76 ± 0.05	0.52 ± 0.09	0.75 ± 0.08	0.50 ± 0.14
CTP	0.77 ± 0.09	0.54 ± 0.16	0.79 ± 0.08	0.58 ± 0.18	0.74 ± 0.02	0.48 ± 0.04
N-signal-free	0.79 ± 0.07	0.61 ± 0.12	0.86 ± 0.05	0.74 ± 0.11	0.83 ± 0.07	0.69 ± 0.13
% accuracy	66.74 ± 7.68		73.81 ± 6.02		69.95 ± 5.44	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.76 ± 0.06	0.51 ± 0.10	$\mathbf{0.87} \pm 0.03$	$\mathbf{0.74} \pm 0.06$	0.73 ± 0.07	0.46 ± 0.10
CTP	0.79 ± 0.07	0.57 ± 0.14	$\mathbf{0.84} \pm 0.10$	$\mathbf{0.68} \pm 0.18$	0.76 ± 0.09	0.52 ± 0.18
N-signal-free	0.78 ± 0.09	0.59 ± 0.17	$\mathbf{0.92} \pm 0.05$	$\mathbf{0.85} \pm 0.10$	0.88 ± 0.05	0.76 ± 0.12
% accuracy	69.97 ± 7.29		$\mathbf{83.68} \pm 6.20$		71.64 ± 4.58	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.76 ± 0.03	0.52 ± 0.05	0.82 ± 0.04	0.64 ± 0.08	0.81 ± 0.03	0.61 ± 0.06
CTP	0.78 ± 0.06	0.56 ± 0.13	0.78 ± 0.05	0.57 ± 0.10	0.78 ± 0.05	0.56 ± 0.11
N-signal-free	0.79 ± 0.08	0.62 ± 0.15	0.88 ± 0.06	0.78 ± 0.09	0.89 ± 0.05	0.79 ± 0.07
% accuracy	70.51 ± 5.43		77.09 ± 4.31		76.53 ± 3.91	

Table A.9: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the plant balanced dataset of automatically collected orthologs.

A.2 Divergence score combined with standard features in the N-terminal 20 residues

A.2.1 *S. cerevisiae*, curated orthologs (N_{20})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.67 ± 0.03	0.36 ± 0.06	0.87 ± 0.03	0.73 ± 0.06	0.87 ± 0.04	0.75 ± 0.07
SP	0.50 ± 0.00	0.00 ± 0.00	0.89 ± 0.04	0.85 ± 0.02	0.94 ± 0.04	0.88 ± 0.05
N-signal-free	0.66 ± 0.02	0.36 ± 0.03	0.89 ± 0.02	0.78 ± 0.03	0.89 ± 0.03	0.80 ± 0.06
% accuracy	70.82 ± 1.61		88.71 ± 1.75		89.88 ± 2.34	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.67 ± 0.02	0.42 ± 0.06	0.84 ± 0.03	0.70 ± 0.06	0.87 ± 0.03	0.77 ± 0.04
SP	0.67 ± 0.11	0.50 ± 0.22	0.89 ± 0.04	0.85 ± 0.04	0.91 ± 0.02	0.88 ± 0.04
N-signal-free	0.67 ± 0.02	0.42 ± 0.04	0.87 ± 0.03	0.77 ± 0.04	$\mathbf{0.90} \pm 0.02$	$\mathbf{0.83} \pm 0.04$
% accuracy	74.78 ± 1.78		87.98 ± 2.17		90.91 ± 1.34	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.80 ± 0.03	0.65 ± 0.06	0.87 ± 0.04	0.76 ± 0.06	0.88 ± 0.03	0.75 ± 0.05
SP	0.78 ± 0.07	0.66 ± 0.11	0.95 ± 0.03	0.89 ± 0.05	0.94 ± 0.04	0.91 ± 0.03
N-signal-free	0.79 ± 0.02	0.63 ± 0.04	$\mathbf{0.90} \pm 0.03$	0.80 ± 0.05	0.88 ± 0.03	0.77 ± 0.04
% accuracy	82.99 ± 1.66		90.18 ± 1.76		89.44 ± 1.71	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	$\mathbf{0.89} \pm 0.02$	0.80 ± 0.03	0.87 ± 0.02	0.77 ± 0.03	0.88 ± 0.03	0.77 ± 0.04
SP	0.96 ± 0.02	0.92 ± 0.07	0.90 ± 0.04	0.87 ± 0.05	0.95 ± 0.05	0.94 ± 0.07
N-signal-free	$\mathbf{0.90} \pm 0.02$	0.81 ± 0.03	$\mathbf{0.90} \pm 0.02$	0.83 ± 0.02	0.89 ± 0.02	0.79 ± 0.04
% accuracy	91.20 ± 1.58		90.91 ± 0.40		90.47 ± 1.38	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	$\mathbf{0.89} \pm 0.02$	0.80 ± 0.02	$\mathbf{0.89} \pm 0.02$	0.80 ± 0.02	$\mathbf{0.89} \pm 0.01$	$\mathbf{0.81} \pm 0.02$
SP	0.97 ± 0.03	0.92 ± 0.07	$\mathbf{0.98} \pm 0.03$	$\mathbf{0.97} \pm 0.04$	$\mathbf{0.98} \pm 0.03$	$\mathbf{0.97} \pm 0.04$
N-signal-free	$\mathbf{0.90} \pm 0.01$	0.81 ± 0.02	$\mathbf{0.90} \pm 0.01$	0.82 ± 0.03	$\mathbf{0.90} \pm 0.01$	$\mathbf{0.83} \pm 0.02$
% accuracy	91.49 ± 1.26		91.93 ± 1.58		$\mathbf{92.23} \pm 1.25$	

Table A.10: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast curated ortholog dataset.

A.2.2 *S. cerevisiae*, RBH orthologs (N_{20})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.65 ± 0.04	0.34 ± 0.08	0.86 ± 0.04	0.73 ± 0.08	0.87 ± 0.04	0.77 ± 0.08
SP	0.50 ± 0.00	0.00 ± 0.00	0.93 ± 0.04	0.88 ± 0.06	$\mathbf{0.97} \pm 0.03$	0.92 ± 0.05
N-signal-free	0.64 ± 0.04	0.33 ± 0.10	0.90 ± 0.03	0.80 ± 0.06	0.90 ± 0.03	0.81 ± 0.05
% accuracy	70.06 ± 3.05		89.32 ± 2.79		90.84 ± 2.87	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.66 ± 0.03	0.41 ± 0.07	0.85 ± 0.03	0.72 ± 0.06	0.87 ± 0.03	0.78 ± 0.06
SP	0.74 ± 0.08	0.65 ± 0.13	0.93 ± 0.04	0.86 ± 0.09	0.94 ± 0.03	0.91 ± 0.03
N-signal-free	0.68 ± 0.01	0.45 ± 0.03	0.89 ± 0.04	0.79 ± 0.08	0.91 ± 0.03	0.84 ± 0.05
% accuracy	75.94 ± 1.12		89.20 ± 2.23		91.43 ± 2.34	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.78 ± 0.04	0.61 ± 0.06	0.87 ± 0.05	0.76 ± 0.08	0.88 ± 0.03	0.74 ± 0.03
SP	0.82 ± 0.08	0.75 ± 0.14	0.96 ± 0.02	0.91 ± 0.04	0.95 ± 0.04	0.92 ± 0.06
N-signal-free	0.80 ± 0.04	0.65 ± 0.07	0.90 ± 0.03	0.81 ± 0.06	0.89 ± 0.02	0.77 ± 0.03
% accuracy	83.45 ± 3.23		90.61 ± 3.10		89.32 ± 1.28	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.89 ± 0.04	0.79 ± 0.07	0.87 ± 0.03	0.77 ± 0.06	0.87 ± 0.04	0.76 ± 0.07
SP	0.96 ± 0.02	0.95 ± 0.02	0.94 ± 0.03	0.91 ± 0.03	0.96 ± 0.03	0.94 ± 0.02
N-signal-free	0.91 ± 0.04	0.82 ± 0.07	0.91 ± 0.03	0.83 ± 0.05	0.90 ± 0.03	0.81 ± 0.06
% accuracy	91.79 ± 2.90		91.31 ± 2.22		90.73 ± 2.54	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.89 ± 0.05	0.80 ± 0.07	$\mathbf{0.90} \pm 0.03$	$\mathbf{0.81} \pm 0.06$	0.89 ± 0.04	$\mathbf{0.81} \pm 0.06$
SP	$\mathbf{0.97} \pm 0.03$	$\mathbf{0.96} \pm 0.03$	0.96 ± 0.02	0.93 ± 0.03	$\mathbf{0.97} \pm 0.03$	0.95 ± 0.02
N-signal-free	0.91 ± 0.04	0.83 ± 0.07	$\mathbf{0.92} \pm 0.03$	0.84 ± 0.05	$\mathbf{0.92} \pm 0.03$	$\mathbf{0.85} \pm 0.05$
% accuracy	92.02 ± 2.85		92.49 ± 2.25		$\mathbf{92.61} \pm 2.14$	

Table A.11: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast automatically collected dataset.

A.2.3 Human, RBH orthologs (N_{20})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.52 ± 0.03	0.10 ± 0.18	0.81 ± 0.06	0.64 ± 0.08	0.82 ± 0.08	0.65 ± 0.11
SP	0.65 ± 0.05	0.29 ± 0.10	0.87 ± 0.03	0.76 ± 0.05	0.89 ± 0.02	0.78 ± 0.04
N-signal-free	0.66 ± 0.05	0.35 ± 0.09	0.90 ± 0.01	0.81 ± 0.02	0.93 ± 0.02	0.88 ± 0.05
% accuracy	65.11 ± 3.55		87.37 ± 1.71		89.32 ± 2.09	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.67 ± 0.07	0.43 ± 0.15	0.81 ± 0.06	0.67 ± 0.08	0.82 ± 0.04	0.68 ± 0.07
SP	0.75 ± 0.06	0.52 ± 0.12	0.89 ± 0.03	0.79 ± 0.05	0.90 ± 0.03	0.80 ± 0.06
N-signal-free	0.75 ± 0.06	0.52 ± 0.12	0.92 ± 0.02	0.85 ± 0.05	0.94 ± 0.04	0.88 ± 0.08
% accuracy	74.89 ± 5.49		89.02 ± 2.74		90.08 ± 2.09	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.73 ± 0.07	0.53 ± 0.11	0.83 ± 0.04	0.67 ± 0.07	0.82 ± 0.05	0.64 ± 0.03
SP	0.81 ± 0.05	0.62 ± 0.06	0.89 ± 0.03	0.79 ± 0.05	0.88 ± 0.02	0.78 ± 0.02
N-signal-free	0.81 ± 0.04	0.64 ± 0.05	0.93 ± 0.02	0.88 ± 0.04	0.91 ± 0.03	0.82 ± 0.06
% accuracy	79.85 ± 1.35		89.47 ± 1.84		87.82 ± 1.71	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.81 ± 0.07	0.66 ± 0.09	$\mathbf{0.84} \pm 0.02$	$\mathbf{0.72} \pm 0.02$	$\mathbf{0.84} \pm 0.03$	0.70 ± 0.06
SP	0.89 ± 0.03	0.77 ± 0.06	0.90 ± 0.04	0.80 ± 0.07	$\mathbf{0.91} \pm 0.04$	$\mathbf{0.81} \pm 0.05$
N-signal-free	0.93 ± 0.04	0.87 ± 0.09	0.94 ± 0.04	$\mathbf{0.89} \pm 0.09$	0.93 ± 0.03	0.85 ± 0.06
% accuracy	89.32 ± 3.12		$\mathbf{90.83} \pm 3.42$		89.62 ± 1.95	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.83 ± 0.07	0.69 ± 0.07	0.83 ± 0.06	0.71 ± 0.08	$\mathbf{0.84} \pm 0.06$	0.70 ± 0.07
SP	0.89 ± 0.02	0.78 ± 0.03	$\mathbf{0.91} \pm 0.02$	0.80 ± 0.04	0.89 ± 0.02	0.79 ± 0.03
N-signal-free	0.94 ± 0.04	0.88 ± 0.07	0.94 ± 0.04	$\mathbf{0.89} \pm 0.08$	$\mathbf{0.95} \pm 0.04$	$\mathbf{0.89} \pm 0.09$
% accuracy	89.77 ± 1.73		90.68 ± 1.73		90.38 ± 1.87	

Table A.12: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the mammal automatically collected dataset.

A.2.4 Plant model organisms, RBH orthologs (N_{20})

	F_{Div}		F_{Phy}		F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.61 ± 0.05	0.30 ± 0.15	$\mathbf{0.88} \pm 0.07$	0.72 ± 0.10	$\mathbf{0.88} \pm 0.02$	0.76 ± 0.05
SP	0.50 ± 0.00	0.00 ± 0.00	0.76 ± 0.09	0.59 ± 0.16	0.83 ± 0.17	0.72 ± 0.25
CTP	0.78 ± 0.08	0.54 ± 0.15	0.82 ± 0.05	0.65 ± 0.10	0.88 ± 0.05	0.75 ± 0.10
N-signal-free	0.80 ± 0.05	0.60 ± 0.09	0.86 ± 0.08	0.73 ± 0.13	0.88 ± 0.04	0.75 ± 0.08
% accuracy	65.05 ± 6.00		78.29 ± 6.40		83.06 ± 4.52	

	$F_{CompFull}$		F_{Div} & F_{Phy}		F_{Div} & F_{Comp}	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.58 ± 0.06	0.19 ± 0.15	0.85 ± 0.09	0.69 ± 0.16	0.87 ± 0.04	0.75 ± 0.08
SP	0.69 ± 0.14	0.48 ± 0.29	0.66 ± 0.17	0.40 ± 0.39	0.80 ± 0.22	0.65 ± 0.39
CTP	0.68 ± 0.08	0.35 ± 0.15	0.88 ± 0.03	0.76 ± 0.07	$\mathbf{0.90} \pm 0.05$	$\mathbf{0.78} \pm 0.10$
N-signal-free	0.66 ± 0.05	0.31 ± 0.08	$\mathbf{0.94} \pm 0.05$	$\mathbf{0.88} \pm 0.10$	0.92 ± 0.05	0.86 ± 0.09
% accuracy	53.33 ± 7.70		83.81 ± 6.23		85.64 ± 4.66	

	F_{Div} & $F_{CompFull}$		F_{Phy} & F_{Comp}		F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.63 ± 0.04	0.32 ± 0.11	$\mathbf{0.88} \pm 0.02$	$\mathbf{0.78} \pm 0.07$	$\mathbf{0.88} \pm 0.06$	0.76 ± 0.08
SP	0.70 ± 0.22	0.48 ± 0.46	0.83 ± 0.12	0.70 ± 0.17	$\mathbf{0.93} \pm 0.09$	$\mathbf{0.83} \pm 0.13$
CTP	0.78 ± 0.05	0.53 ± 0.11	0.89 ± 0.07	0.76 ± 0.13	0.84 ± 0.05	0.68 ± 0.10
N-signal-free	0.82 ± 0.06	0.63 ± 0.10	0.87 ± 0.04	0.75 ± 0.06	0.87 ± 0.06	0.75 ± 0.10
% accuracy	67.30 ± 4.19		83.41 ± 5.93		81.60 ± 4.40	

	F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Phy} & F_{Comp}		F_{Div} & F_{Phy} & $F_{CompFull}$	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.85 ± 0.04	0.73 ± 0.05	$\mathbf{0.88} \pm 0.06$	$\mathbf{0.78} \pm 0.09$	0.84 ± 0.07	0.71 ± 0.10
SP	0.86 ± 0.14	0.72 ± 0.20	0.83 ± 0.20	0.66 ± 0.39	0.86 ± 0.21	0.67 ± 0.40
CTP	0.88 ± 0.04	0.75 ± 0.08	$\mathbf{0.90} \pm 0.05$	$\mathbf{0.78} \pm 0.11$	0.87 ± 0.07	0.72 ± 0.14
N-signal-free	0.87 ± 0.04	0.75 ± 0.06	0.92 ± 0.04	0.85 ± 0.09	0.93 ± 0.03	0.86 ± 0.06
% accuracy	82.33 ± 3.71		$\mathbf{86.36} \pm 5.87$		83.79 ± 6.66	

	F_{Phy} & F_{Comp} & $F_{CompFull}$		F_{Div} & F_{Comp} & $F_{CompFull}$		ALL	
	AUC	MCC	AUC	MCC	AUC	MCC
MTS	0.87 ± 0.05	0.77 ± 0.07	0.85 ± 0.06	0.72 ± 0.07	0.87 ± 0.05	0.77 ± 0.06
SP	0.86 ± 0.14	0.72 ± 0.20	0.86 ± 0.14	0.81 ± 0.16	0.86 ± 0.14	0.81 ± 0.16
CTP	0.88 ± 0.05	0.75 ± 0.09	0.88 ± 0.04	0.75 ± 0.07	0.89 ± 0.05	$\mathbf{0.78} \pm 0.10$
N-signal-free	0.87 ± 0.04	0.76 ± 0.07	0.90 ± 0.06	0.81 ± 0.10	0.91 ± 0.04	0.83 ± 0.06
% accuracy	83.06 ± 4.50		83.80 ± 4.11		85.64 ± 3.64	

Table A.13: The 5-fold cross-validation performance of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the plant automatically collected dataset.

A.3 *Post hoc analysis* for NCDiff parameters

Since I made an arbitrary choice when defining divergence features, parameters for NCDiff have been searched with in the yeast curated dataset: window length and normalization start position in C-terminal from 40 to 80 with or without classical N-terminal features of the first 20 amino acids. In the case with the classical features, average accuracy for parameter space is 91.82%, and best accuracy is 92.52% with either a combination of window size 15 and start position at 50 or combination 16 and 49. Similarly, in the case without the classical features, average accuracy is 70.45% and best accuracy is 71.99% when window size is 40 and start position is 48. Because of multiple test, this difference does not seem to be significant; however, analysis result is summarized in Figure A.1 and A.2.

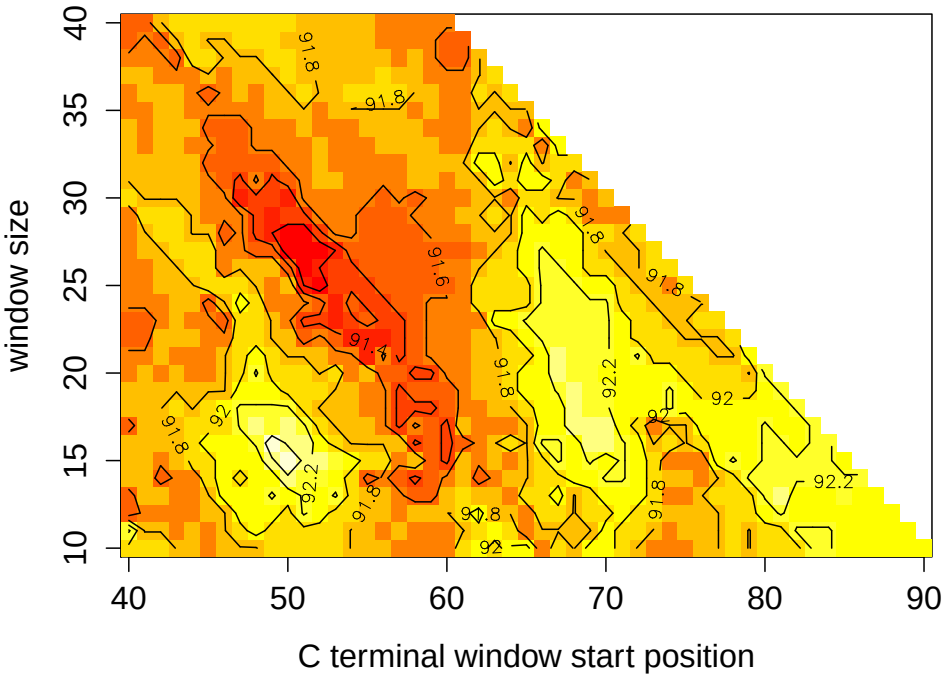


Figure A.1: Heat map for two parameters of NCDiff in terms of accuracy based on yeast curated dataset using divergence and classical features in N-terminal 20 residues.

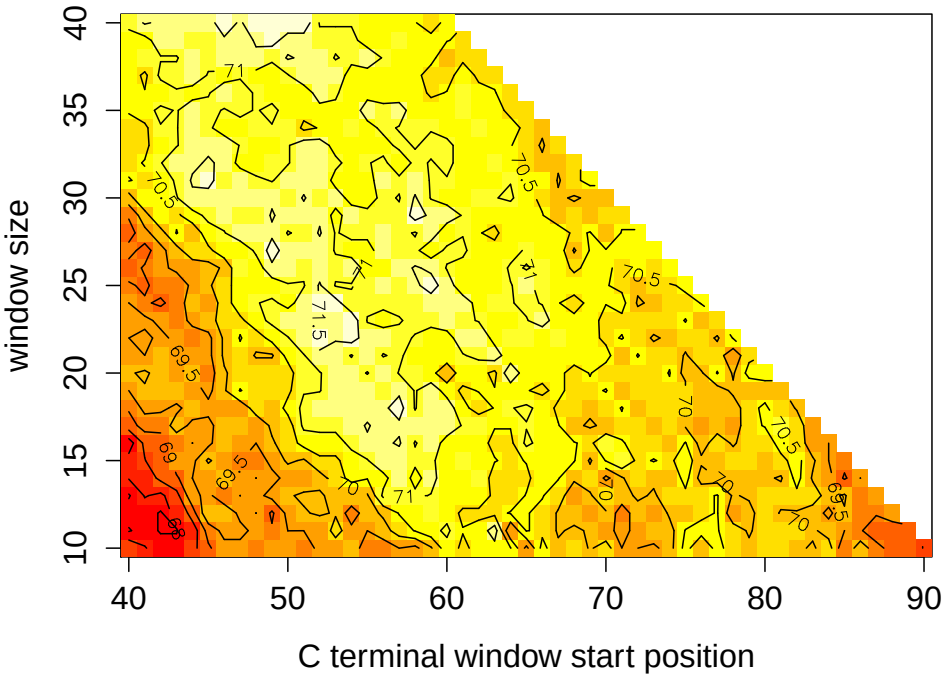


Figure A.2: Heat map for two parameters of NCDiff in terms of accuracy based on yeast curated dataset using only divergence features.

A.4 Divergence score combined with standard features in N-terminal 40 residues

A.4.1 *S. cerevisiae*, curated orthologs (N_{40})

	F_{Div}			F_{Phy}			F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	83	0	96	141	2	36	140	2	37
SP	16	0	37	6	30	17	6	31	16
N-signal-free	50	0	400	30	10	410	26	10	414

	$F_{CompFull}$			F_{Div} & F_{Phy}			F_{Div} & F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	73	1	105	132	2	45	133	2	44
SP	4	19	30	4	43	6	3	41	9
N-signal-free	29	3	418	23	6	421	10	6	434

	F_{Div} & $F_{CompFull}$			F_{Phy} & F_{Comp}			F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	120	1	58	144	2	33	136	2	41
SP	3	31	19	4	32	17	3	35	15
N-signal-free	28	7	415	26	10	414	35	6	409

	F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Phy} & F_{Comp}			F_{Div} & F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	141	1	37	133	3	43	137	2	40
SP	2	32	19	3	43	7	3	42	8
N-signal-free	18	5	427	11	7	432	22	5	423

	F_{Phy} & F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Comp} & $F_{CompFull}$			ALL		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	140	1	38	137	1	41	137	1	41
SP	2	34	17	2	44	7	2	43	8
N-signal-free	22	7	421	16	5	429	15	6	429

Table A.14: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast curated ortholog dataset.

A.4.2 *S. cerevisiae*, RBH orthologs (N_{40})

	F_{Div}			F_{Phy}			F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	91	0	128	168	4	47	172	3	44
SP	18	0	55	8	47	18	7	52	14
N-signal-free	54	0	506	33	14	513	27	9	524

	$F_{CompFull}$			F_{Div} & F_{Phy}			F_{Div} & F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	82	1	136	164	3	52	171	3	45
SP	4	35	34	9	52	12	7	56	10
N-signal-free	29	1	530	19	6	535	19	6	535

	F_{Div} & $F_{CompFull}$			F_{Phy} & F_{Comp}			F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	138	4	77	174	2	43	172	3	44
SP	6	48	19	6	53	14	2	54	17
N-signal-free	34	1	525	25	11	524	43	5	512

	F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Phy} & F_{Comp}			F_{Div} & F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	172	0	47	171	4	44	168	3	48
SP	3	58	12	9	56	8	5	61	7
N-signal-free	24	4	532	19	8	533	23	2	535

	F_{Phy} & F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Comp} & $F_{CompFull}$			ALL		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	173	0	46	175	1	43	176	2	41
SP	3	57	13	3	63	7	4	63	6
N-signal-free	26	5	529	18	2	540	16	2	542

Table A.15: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast automatically collected dataset.

A.4.3 Human, RBH orthologs (N_{40})

	F_{Div}			F_{Phy}			F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	3	35	43	50	16	15	58	10	13
SP	1	84	84	15	124	30	13	132	24
N-signal-free	2	67	346	10	27	378	8	18	389

	$F_{CompFull}$			F_{Div} & F_{Phy}			F_{Div} & F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	31	14	36	53	10	18	54	13	14
SP	6	102	61	11	134	24	7	143	19
N-signal-free	13	37	365	10	18	387	8	16	391

	F_{Div} & $F_{CompFull}$			F_{Phy} & F_{Comp}			F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	40	14	27	58	10	13	58	13	10
SP	8	121	40	13	135	21	9	129	31
N-signal-free	11	34	370	9	19	387	12	13	390

	F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Phy} & F_{Comp}			F_{Div} & F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	58	11	12	55	13	13	59	9	13
SP	8	130	31	7	142	20	12	136	21
N-signal-free	7	18	390	7	14	394	12	15	388

	F_{Phy} & F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Comp} & $F_{CompFull}$			ALL		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	63	9	9	58	11	12	60	11	10
SP	9	133	27	7	142	20	8	140	21
N-signal-free	8	17	390	9	13	393	8	13	394

Table A.16: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the mammal automatically collected dataset.

A.4.4 Plant model organisms, RBH orthologs (N_{40})

	F_{Div}				F_{Phy}				F_{Comp}			
Predicted →	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free
MTS	17	0	35	9	30	0	25	6	44	0	12	5
SP	3	0	4	8	2	2	5	6	1	6	5	3
CTP	3	0	90	6	13	1	78	7	10	2	83	4
N-signal-free	6	0	21	70	7	4	8	78	11	2	8	76

	$F_{CompFull}$				F_{Div} & F_{Phy}				F_{Div} & F_{Comp}			
Predicted →	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free
MTS	16	0	27	18	23	1	31	6	40	0	17	4
SP	0	6	1	8	3	5	5	2	1	8	5	1
CTP	11	1	67	20	11	1	85	2	9	2	85	3
N-signal-free	12	2	27	56	2	2	9	84	8	2	8	79

	F_{Div} & $F_{CompFull}$				F_{Phy} & F_{Comp}				F_{Phy} & $F_{CompFull}$			
Predicted →	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free
MTS	21	0	30	10	43	0	15	3	28	0	26	7
SP	2	6	2	5	1	6	6	2	0	10	3	2
CTP	8	0	81	10	10	3	82	4	17	2	74	6
N-signal-free	8	0	14	75	11	2	8	76	4	1	15	77

	F_{Comp} & $F_{CompFull}$				F_{Div} & F_{Phy} & F_{Comp}				F_{Div} & F_{Phy} & $F_{CompFull}$			
Predicted →	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free
MTS	41	0	13	7	40	0	16	5	29	0	28	4
SP	1	11	3	0	2	8	4	1	0	8	5	2
CTP	13	2	78	6	9	2	84	4	15	0	78	6
N-signal-free	13	1	6	77	8	1	7	81	4	1	9	83

	F_{Phy} & F_{Comp} & $F_{CompFull}$				F_{Div} & F_{Comp} & $F_{CompFull}$				ALL			
Predicted →	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free
MTS	39	0	15	7	40	0	17	4	41	0	17	3
SP	1	11	3	0	1	9	5	0	1	10	4	0
CTP	11	2	79	7	6	1	87	5	6	1	87	5
N-signal-free	12	1	7	77	8	1	6	82	8	1	7	81

Table A.17: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the plant automatically collected dataset.

A.4.5 *S. cerevisiae*, curated orthologs – classes balanced (N_{40})

	F_{Div}			F_{Phy}			F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	31	15	7	46	2	5	40	3	10
SP	16	34	3	6	45	2	4	46	3
N-signal-free	9	9	35	12	8	33	11	8	34

	$F_{CompFull}$			F_{Div} & F_{Phy}			F_{Div} & F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	37	3	13	45	1	7	39	4	10
SP	8	38	7	2	48	3	2	49	2
N-signal-free	13	9	31	4	6	43	6	4	43

	F_{Div} & $F_{CompFull}$			F_{Phy} & F_{Comp}			F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	38	7	8	39	4	10	45	0	8
SP	5	43	5	3	47	3	3	46	4
N-signal-free	8	5	40	12	7	34	10	9	34

	F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Phy} & F_{Comp}			F_{Div} & F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	43	1	9	40	3	10	43	0	10
SP	2	46	5	2	49	2	3	48	2
N-signal-free	11	5	37	5	3	45	7	3	43

	F_{Phy} & F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Comp} & $F_{CompFull}$			ALL		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	43	0	10	43	3	7	44	2	7
SP	2	48	3	2	49	2	1	50	2
N-signal-free	11	6	36	7	4	42	6	4	43

Table A.18: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast balanced dataset of curated orthologs.

A.4.6 *S. cerevisiae*, RBH orthologs – classes balanced (N_{40})

	F_{Div}			F_{Phy}			F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	36	28	9	59	2	12	55	4	14
SP	18	42	13	3	70	0	4	69	0
N-signal-free	11	27	35	13	1	59	13	3	57

	$F_{CompFull}$			F_{Div} & F_{Phy}			F_{Div} & F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	47	7	19	61	1	11	57	2	14
SP	13	53	7	3	70	0	2	71	0
N-signal-free	23	9	41	9	0	64	8	3	62

	F_{Div} & $F_{CompFull}$			F_{Phy} & F_{Comp}			F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	55	4	14	56	3	14	61	1	11
SP	12	57	4	4	69	0	1	72	0
N-signal-free	16	3	54	11	1	61	16	2	55

	F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Phy} & F_{Comp}			F_{Div} & F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	58	0	15	59	0	14	59	1	13
SP	1	72	0	2	71	0	2	71	0
N-signal-free	15	1	57	7	1	65	9	0	64

	F_{Phy} & F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Comp} & $F_{CompFull}$			ALL		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	57	1	15	61	0	12	61	0	12
SP	2	71	0	2	70	1	2	70	1
N-signal-free	15	1	57	11	1	61	10	1	62

Table A.19: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast balanced dataset of automatically collected orthologs.

A.4.7 Human, RBH orthologs – classes balanced (N_{40})

	F_{Div}			F_{Phy}			F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	45	19	17	66	11	4	69	9	3
SP	21	53	7	11	63	7	14	60	7
N-signal-free	18	21	42	7	10	64	7	7	67

	$F_{CompFull}$			F_{Div} & F_{Phy}			F_{Div} & F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	61	12	8	69	6	6	71	7	3
SP	16	51	14	11	66	4	13	65	3
N-signal-free	7	20	54	8	8	65	9	7	65

	F_{Div} & $F_{CompFull}$			F_{Phy} & F_{Comp}			F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	57	13	11	70	9	2	67	9	5
SP	16	52	13	12	62	7	9	68	4
N-signal-free	11	14	56	7	7	67	7	8	66

	F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Phy} & F_{Comp}			F_{Div} & F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	66	10	5	70	8	3	71	5	5
SP	11	63	7	10	68	3	8	69	4
N-signal-free	5	8	68	10	7	64	7	4	70

	F_{Phy} & F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Comp} & $F_{CompFull}$			ALL		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	68	8	5	71	6	4	71	6	4
SP	11	65	5	11	66	4	11	67	3
N-signal-free	6	8	67	6	6	69	6	6	69

Table A.20: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the mammal balanced dataset of automatically collected orthologs.

A.4.8 Plant model organisms, RBH orthologs – classes balanced (N_{40})

	F_{Div}			F_{Phy}			F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	30	26	5	47	11	3	42	12	7
SP	9	50	2	13	43	5	13	45	3
N-signal-free	11	9	41	13	4	44	8	5	48

	$F_{CompFull}$			F_{Div} & F_{Phy}			F_{Div} & F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	34	14	13	41	16	4	51	10	0
SP	22	31	8	15	43	3	8	49	4
N-signal-free	17	14	30	4	5	52	5	4	52

	F_{Div} & $F_{CompFull}$			F_{Phy} & F_{Comp}			F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	36	16	9	43	13	5	42	14	5
SP	13	44	4	14	44	3	16	41	4
N-signal-free	13	6	42	8	5	48	7	9	45

	F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Phy} & F_{Comp}			F_{Div} & F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	42	13	6	51	10	0	39	15	7
SP	9	46	6	8	49	4	17	40	4
N-signal-free	12	9	40	4	4	53	6	3	52

	F_{Phy} & F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Comp} & $F_{CompFull}$			ALL		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	42	14	5	48	11	2	46	12	3
SP	9	46	6	13	44	4	13	43	5
N-signal-free	11	9	41	4	8	49	4	6	51

Table A.21: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the plant balanced dataset of automatically collected orthologs.

A.5 Divergence score combined with standard features in the N-terminal 20 residues

A.5.1 *S. cerevisiae*, curated orthologs (N_{20})

	F_{Div}			F_{Phy}			F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	83	0	96	149	1	29	140	2	37
SP	16	0	37	8	42	3	6	47	0
N-signal-free	50	0	400	34	2	414	20	4	426

	$F_{CompFull}$			F_{Div} & F_{Phy}			F_{Div} & F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	73	1	105	133	2	44	140	1	38
SP	4	19	30	9	42	2	8	44	1
N-signal-free	29	3	418	24	1	425	13	1	436

	F_{Div} & $F_{CompFull}$			F_{Phy} & F_{Comp}			F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	120	1	58	142	2	35	146	0	33
SP	3	31	19	5	48	0	2	47	4
N-signal-free	28	7	415	21	4	425	30	3	417

	F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Phy} & F_{Comp}			F_{Div} & F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	146	1	32	141	1	37	145	0	34
SP	0	49	4	8	43	2	2	48	3
N-signal-free	20	3	427	13	1	436	25	1	424

	F_{Phy} & F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Comp} & $F_{CompFull}$			ALL		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	148	1	30	144	0	35	144	0	35
SP	0	50	3	0	51	2	1	51	1
N-signal-free	20	4	426	17	1	432	15	1	434

Table A.22: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast curated ortholog dataset.

A.5.2 *S. cerevisiae*, RBH orthologs (N_{20})

	F_{Div}			F_{Phy}			F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	91	0	128	174	6	39	173	3	43
SP	18	0	55	7	64	2	3	69	1
N-signal-free	54	0	506	36	1	523	23	5	532

	$F_{CompFull}$			F_{Div} & F_{Phy}			F_{Div} & F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	82	1	136	165	5	49	171	2	46
SP	4	35	34	9	63	1	8	65	0
N-signal-free	29	1	530	25	3	532	15	2	543

	F_{Div} & $F_{CompFull}$			F_{Phy} & F_{Comp}			F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	138	4	77	173	4	42	180	0	39
SP	6	48	19	5	68	0	4	66	3
N-signal-free	34	1	525	25	4	531	41	4	515

	F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Phy} & F_{Comp}			F_{Div} & F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	179	0	40	171	2	46	173	1	45
SP	3	68	2	8	65	0	4	67	2
N-signal-free	23	2	535	16	2	542	26	1	533

	F_{Phy} & F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Comp} & $F_{CompFull}$			ALL		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	179	0	40	180	2	37	180	2	37
SP	2	69	2	3	68	2	2	69	2
N-signal-free	23	1	536	17	3	540	19	1	540

Table A.23: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the yeast automatically collected dataset.

A.5.3 Human, RBH orthologs (N_{20})

	F_{Div}			F_{Phy}			F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	3	35	43	54	14	13	54	18	9
SP	1	84	84	12	133	24	14	140	15
N-signal-free	2	67	346	12	9	394	7	8	400

	$F_{CompFull}$			F_{Div} & F_{Phy}			F_{Div} & F_{Comp}		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	31	14	36	53	18	10	54	17	10
SP	6	102	61	8	143	18	12	144	13
N-signal-free	13	37	365	8	11	396	5	9	401

	F_{Div} & $F_{CompFull}$			F_{Phy} & F_{Comp}			F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	40	14	27	57	16	8	55	13	13
SP	8	121	40	15	139	15	12	138	19
N-signal-free	11	34	370	8	8	399	13	11	391

	F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Phy} & F_{Comp}			F_{Div} & F_{Phy} & $F_{CompFull}$		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	52	21	8	57	17	7	58	15	8
SP	11	142	16	11	144	14	8	146	15
N-signal-free	5	10	400	4	8	403	12	11	392

	F_{Phy} & F_{Comp} & $F_{CompFull}$			F_{Div} & F_{Comp} & $F_{CompFull}$			ALL		
Predicted →	MTS	SP	N-signal-free	MTS	SP	N-signal-free	MTS	SP	N-signal-free
MTS	56	19	6	56	19	6	57	19	5
SP	13	140	16	9	147	13	12	143	14
N-signal-free	5	9	401	5	10	400	5	9	401

Table A.24: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the mammal automatically collected dataset.

A.5.4 Plant model organisms, RBH orthologs (N_{20})

	F_{Div}				F_{Phy}				F_{Comp}			
Predicted →	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free
MTS	17	0	35	9	52	0	8	1	49	0	6	6
SP	3	0	4	8	5	8	1	1	1	10	2	2
CTP	3	0	90	6	10	2	77	10	5	1	85	8
N-signal-free	6	0	21	70	5	2	14	76	5	1	9	82

	$F_{CompFull}$				F_{Div} & F_{Phy}				F_{Div} & F_{Comp}			
Predicted →	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free
MTS	16	0	27	18	47	0	12	2	48	0	11	2
SP	0	6	1	8	6	5	4	0	3	9	3	0
CTP	11	1	67	20	6	1	88	4	3	1	91	4
N-signal-free	12	2	27	56	3	1	5	88	5	0	7	85

	F_{Div} & $F_{CompFull}$				F_{Phy} & F_{Comp}				F_{Phy} & $F_{CompFull}$			
Predicted →	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free
MTS	21	0	30	10	49	0	6	6	50	0	9	2
SP	2	6	2	5	1	10	2	2	0	13	1	1
CTP	8	0	81	10	3	2	87	7	7	2	80	10
N-signal-free	8	0	14	75	5	1	10	81	5	1	12	79

	F_{Comp} & $F_{CompFull}$				F_{Div} & F_{Phy} & F_{Comp}				F_{Div} & F_{Phy} & $F_{CompFull}$			
Predicted →	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free
MTS	46	1	7	7	49	0	9	3	45	0	13	3
SP	1	11	2	1	2	10	3	0	2	11	2	0
CTP	4	2	86	7	2	2	90	5	7	2	85	5
N-signal-free	5	1	10	81	4	0	7	86	2	1	7	87

	F_{Phy} & F_{Comp} & $F_{CompFull}$				F_{Div} & F_{Comp} & $F_{CompFull}$				ALL			
Predicted →	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free	MTS	SP	CTP	N-signal-free
MTS	48	0	7	6	46	0	11	4	48	0	9	4
SP	2	11	1	1	3	11	1	0	3	11	1	0
CTP	3	3	87	6	4	1	88	6	3	1	89	6
N-signal-free	4	1	12	80	5	0	9	83	3	0	9	85

Table A.25: Confusion matrix of the 5-fold cross-validation of an SVM classifier, using various feature set combinations as listed above each column, is shown for three-way classification on the plant automatically collected dataset.

Appendix B

Appendix for prediction of presequence and its cleavage site

B.1 Classifiers for presequence prediction

Other than SVM with RBF kernel I tried SVM with polynomial kernel and random forest to predict presequence. Implementation of SVM with polynomial kernel is the same package as that of RBF kernel, and I used randomForest package in R [97]. Random forest is an ensemble classifier and usually shows good performance. Learning procedure depends on mainly two steps: bootstrap sampling B times from training data and learning multiple decision trees from the sampled dataset with randomly choosing m features from M total features. For classification task random forest predicts a query by voting of B trees. To calculate AUC, the number of trees which predict positive is used as a score by dividing it with B.

Polynomial kernel in this work is defined below:

$$k(x_i, x_j) = (1 + \gamma x_i \cdot x_j)^D \tag{B.1}$$

, where γ and D are kernel parameters.

Benchmarking is conducted by five-fold cross validation in the training dataset, and result is summarized in Table B.1. Although difference between SVM with two different kernel functions is small, in this work SVM with RBF kernel was applied.

Classifier	ROC AUC
SVM (RBF)	0.955
SVM (Polynomial)	0.953
Random forest	0.946

Table B.1: Evaluations for different classifiers.

B.2 Motif analysis

Although the reported fourteen motifs are significantly discriminative motif between presequence containing and non-containing proteins, this result might be biased due to the amino acid difference between these two sets. Scramble test was conducted 100 times by shuffling amino acids in the positive dataset after segmentation: first 30, middle 30, and last 30 in N-terminal 90 residues. Only first methionine was kept its position. Assumption of this test is that short motif is conserved its order within 6-mer window if the motif is not randomly mutated. The number of detected times in this shuffle test is summarized in the Table B.2. Roughly speaking, half of motifs are detected repeatedly, and the most significant motif, HHPBHH, is perfectly detected in all shuffle tests. Since Arg composition is remarkably high in the positive dataset, I tried shuffled test with

Arg as a distinct letter. This condition leads to much more tests due to the higher number of candidate motifs, and detected times are generally reduced. Surprisingly, the most HHPBHH is stably observed in both conditions; therefore, Arg might be conserved at least within this motif.

Rank	Motif	P-value	#Observations in 100 times scramble test
1	HHPBHH	5.715E-013	100 (100)
2	HHBPHH	1.183E-011	88 (23)
3	HHHBPH	1.063E-009	0 (0)
4	HHBPHB	1.838E-009	34 (18)
5	HBHHBb	6.131E-009	43 (0)
6	BHHPPP	0.000000093	73 (11)
7	HHHBBH	1.208E-007	0 (0)
8	HHBHHB	8.935E-007	0 (0)
9	HPBHHP	9.676E-007	29 (0)
10	PHHBPH	0.000001229	3 (0)
11	HBHHbB	0.000001742	2 (1)
12	HHHHBB	0.000004929	0 (0)
13	HHBPHP	0.000006382	0 (0)
14	BPHBHH	0.000009312	0 (0)

Table B.2: Fourteen detected motifs are listed. Right most column indicates how many times each motif is detected in 100 scrambled tests, and a number in parenthesis is an observed number of each motif in different scramble test where Arg is distinct amino acid from basic residues group.

B.3 Clustering result

Detail of clustering result is summarized in Table B.3.

OLN	UniProtAC	MPP	NetCharge	Length	Hmoment	$Cons_{half}$	Cons	D	E	K	R	Cluster
YAL008W	P18411	-10.15	0.16	69.00	3.82	0.48	0.43	0.00	0.00	0.07	0.09	II
YAL019W	P31380	-15.10	-0.04	74.00	-0.17	0.54	0.53	0.07	0.05	0.05	0.03	II
YAL039C	P06182	-10.24	-0.06	62.00	0.40	0.78	0.80	0.06	0.05	0.03	0.02	III
YAL044C	P39726	6.68	0.23	22.00	1.52	0.47	0.58	0.00	0.00	0.05	0.18	I
YAL054C	Q01574	-15.91	0.01	69.00	1.25	0.52	0.61	0.03	0.09	0.10	0.03	II
YBL022C	P36775	14.83	0.24	37.00	3.05	0.50	0.47	0.00	0.00	0.05	0.19	I
YBL038W	P38064	8.00	0.11	37.00	0.99	0.43	0.42	0.00	0.03	0.05	0.08	II
YBL045C	P07256	10.67	0.24	17.00	3.23	0.52	0.57	0.00	0.00	0.12	0.12	I
YBL064C	P34227	-2.42	0.17	30.00	2.91	0.69	0.71	0.00	0.00	0.10	0.07	I
YBL090W	P38175	7.72	0.31	16.00	2.75	0.38	0.45	0.00	0.00	0.06	0.25	I
YBL095W	P38172	-12.43	0.03	59.00	3.13	0.61	0.58	0.05	0.02	0.05	0.05	II
YBL099W	P07251	13.47	0.20	35.00	3.14	0.58	0.57	0.00	0.00	0.03	0.17	I
YBR026C	P38071	12.85	0.22	9.00	2.92	0.39	0.52	0.00	0.00	0.11	0.11	I
YBR037C	P23833	8.51	0.18	40.00	2.63	0.33	0.33	0.00	0.00	0.05	0.13	I
YBR039W	P38077	10.23	0.13	32.00	2.15	0.46	0.49	0.00	0.00	0.03	0.09	I
YBR044C	P38228	-12.16	0.24	17.00	3.53	0.54	0.58	0.00	0.00	0.12	0.12	I
YBR047W	P38231	11.25	0.18	22.00	1.68	0.57	0.61	0.00	0.00	0.09	0.09	I
YBR084W	P09440	-6.74	0.18	34.00	1.75	0.42	0.44	0.00	0.00	0.03	0.15	I
YBR104W	P38087	-19.30	-0.38	16.00	0.12	0.67	0.68	0.13	0.25	0.00	0.00	III
YBR111C	Q01976	8.79	0.25	16.00	3.48	0.41	0.50	0.00	0.00	0.00	0.25	I
YBR122C	P36531	13.92	0.30	10.00	1.63	0.50	0.59	0.00	0.00	0.20	0.10	I
YBR146W	P38120	5.70	0.21	24.00	2.17	0.44	0.38	0.00	0.00	0.00	0.21	I
YBR176W	P38122	-11.88	0.25	24.00	2.41	0.34	0.41	0.00	0.00	0.17	0.08	I
YBR185C	P38300	-19.54	0.19	47.00	2.58	0.40	0.46	0.02	0.04	0.15	0.11	II
YBR221C	P32473	10.59	0.18	33.00	3.02	0.44	0.52	0.00	0.00	0.00	0.18	I
YBR251W	P33759	10.09	0.23	13.00	2.31	0.58	0.63	0.00	0.00	0.08	0.15	I
YBR263W	P37292	6.27	0.21	19.00	1.53	0.51	0.60	0.00	0.00	0.05	0.16	I
YBR282W	P36526	-11.54	0.19	54.00	3.82	0.65	0.71	0.02	0.02	0.17	0.06	III
YCL009C	P25605	7.17	0.17	24.00	2.55	0.58	0.67	0.00	0.00	0.00	0.17	I
YCL017C	P25374	-8.56	0.07	60.00	2.47	0.49	0.44	0.05	0.02	0.05	0.08	II
YCR003W	P25348	-7.21	0.03	71.00	2.29	0.43	0.41	0.06	0.03	0.08	0.03	II
YCR028C-A	P32445	9.90	0.22	9.00	3.77	0.67	0.72	0.00	0.00	0.00	0.22	I
YCR046C	P25626	6.04	0.20	15.00	3.23	0.37	0.49	0.00	0.00	0.00	0.20	I
YCR071C	P25642	-9.95	0.08	80.00	2.23	0.45	0.41	0.01	0.05	0.08	0.06	II
YCR083W	P25372	-14.06	0.11	66.00	3.37	0.32	0.27	0.03	0.02	0.09	0.06	II
YDL004W	Q12165	10.60	0.24	21.00	3.12	0.78	0.69	0.00	0.00	0.10	0.14	I
YDL044C	P10849	5.48	0.17	23.00	3.93	0.49	0.44	0.00	0.00	0.04	0.13	I

Table B.3: Clustering result for the yeast presequence data. Used features are rounded at the second decimal place. Length of presequence are extracted the proteomic analysis.

OLN	UniProtAC	MPP	NetCharge	Length	Hmoment	$Cons_{half}$	Cons	D	E	K	R	Cluster
YDL130W-A	P01098	12.68	0.23	22.00	2.40	0.45	0.49	0.00	0.00	0.00	0.23	I
YDL178W	P46681	-3.52	0.20	35.00	2.14	0.42	0.39	0.00	0.00	0.06	0.14	I
YDL181W	P01097	11.16	0.19	21.00	1.95	0.45	0.49	0.00	0.00	0.00	0.19	I
YDL202W	P36521	11.01	0.13	30.00	1.82	0.38	0.40	0.00	0.03	0.07	0.10	I
YDL203C	Q07622	-14.65	0.00	15.00	0.86	0.58	0.52	0.07	0.00	0.07	0.00	III
YDR036C	P28817	7.58	0.16	31.00	2.60	0.43	0.40	0.00	0.00	0.10	0.06	I
YDR070C	Q12497	-0.13	0.24	29.00	3.60	0.51	0.46	0.00	0.00	0.03	0.21	I
YDR116C	Q04599	-5.41	0.18	28.00	2.56	0.48	0.46	0.00	0.00	0.11	0.07	I
YDR148C	P19262	3.21	0.17	71.00	1.84	0.21	0.34	0.00	0.01	0.11	0.07	II
YDR175C	Q03976	2.52	0.23	22.00	2.19	0.42	0.47	0.00	0.00	0.14	0.09	I
YDR178W	P37298	8.11	0.23	31.00	2.53	0.54	0.63	0.00	0.00	0.10	0.13	I
YDR194C	P15424	-14.47	0.13	53.00	2.18	0.36	0.36	0.04	0.00	0.02	0.15	II
YDR234W	P49367	12.04	0.17	18.00	1.29	0.57	0.66	0.00	0.00	0.00	0.17	I
YDR298C	P09457	9.74	0.18	17.00	2.85	0.59	0.68	0.00	0.00	0.00	0.18	I
YDR337W	P21771	7.35	0.16	25.00	1.05	0.45	0.52	0.00	0.00	0.04	0.12	I
YDR347W	P10662	8.62	0.33	12.00	3.72	0.23	0.38	0.00	0.00	0.08	0.25	I
YDR376W	P48360	10.42	0.13	8.00	3.60	0.55	0.65	0.00	0.00	0.00	0.13	I
YDR405W	P32387	8.53	0.12	33.00	2.46	0.39	0.42	0.00	0.03	0.09	0.06	I
YDR430C	P32898	12.16	0.29	7.00	2.30	0.71	0.72	0.00	0.00	0.00	0.29	I
YDR462W	P36527	-13.92	0.18	44.00	1.69	0.45	0.54	0.00	0.02	0.11	0.09	II
YDR494W	Q03430	2.88	0.23	13.00	0.72	0.58	0.62	0.00	0.00	0.00	0.23	I
YDR508C	P48813	-0.97	-0.08	79.00	1.06	0.46	0.46	0.04	0.18	0.06	0.08	II
YDR511W	Q04401	8.16	0.25	12.00	1.56	0.40	0.52	0.00	0.00	0.08	0.17	I
YDR513W	P17695	12.97	0.03	29.00	1.38	0.35	0.41	0.03	0.03	0.03	0.07	II
YEL024W	P08067	8.03	0.23	22.00	2.61	0.51	0.59	0.00	0.00	0.14	0.09	I
YEL052W	P32317	10.74	0.20	25.00	2.07	0.32	0.33	0.00	0.00	0.04	0.16	I
YER015W	P39518	-14.56	-0.18	17.00	1.59	0.61	0.62	0.18	0.06	0.00	0.06	III
YER017C	P39925	-10.50	0.17	29.00	3.21	0.33	0.34	0.00	0.00	0.03	0.14	I
YER020W	P10823	-13.51	0.04	91.00	1.02	0.55	0.48	0.03	0.04	0.09	0.03	II
YER069W	Q01217	9.99	0.19	57.00	3.42	0.56	0.49	0.00	0.00	0.12	0.07	I
YER073W	P40047	11.81	0.23	22.00	3.20	0.54	0.64	0.00	0.00	0.00	0.23	I
YER078C	P40051	6.33	0.15	26.00	3.36	0.41	0.46	0.00	0.00	0.04	0.12	I
YER080W	P40053	11.36	0.23	35.00	2.95	0.53	0.45	0.00	0.00	0.09	0.14	I
YER087W	P39965	1.93	0.11	56.00	2.51	0.45	0.48	0.00	0.04	0.05	0.09	II
YER140W	P40085	-21.74	-0.06	72.00	2.15	0.17	0.27	0.10	0.13	0.10	0.07	II
YER141W	P40086	-13.03	0.18	66.00	3.18	0.51	0.49	0.00	0.02	0.08	0.12	II
YER182W	P40098	10.10	0.25	12.00	1.73	0.49	0.51	0.00	0.00	0.08	0.17	I
YFL018C	P09624	8.88	0.31	13.00	3.03	0.45	0.57	0.00	0.00	0.08	0.23	I

Table B.3: Clustering result for the yeast presequence data. Used features are rounded at the second decimal place. Length of presequence are extracted the proteomic analysis.

OLN	UniProtAC	MPP	NetCharge	Length	Hmoment	$Cons_{half}$	Cons	D	E	K	R	Cluster
YFL046W	P43557	-17.65	0.11	37.00	3.14	0.51	0.60	0.03	0.03	0.05	0.11	II
YFR033C	P00127	-20.14	-0.47	90.00	-0.13	0.71	0.69	0.20	0.32	0.04	0.01	III
YFR049W	P19955	2.95	0.13	8.00	2.07	0.66	0.64	0.00	0.00	0.00	0.13	I
YGL059W	P53170	-9.86	0.02	64.00	2.67	0.44	0.57	0.02	0.06	0.05	0.05	II
YGL107C	P53140	9.79	0.10	51.00	2.75	0.53	0.48	0.02	0.02	0.04	0.10	II
YGL119W	P27697	-7.72	0.22	41.00	3.00	0.51	0.50	0.00	0.00	0.07	0.15	I
YGL125W	P53128	-10.84	0.00	73.00	-1.06	0.48	0.59	0.05	0.05	0.05	0.05	II
YGL129C	Q01163	-7.32	0.15	41.00	1.80	0.49	0.56	0.00	0.02	0.05	0.12	I
YGL187C	P04037	6.31	0.24	17.00	3.85	0.62	0.58	0.00	0.00	0.06	0.18	I
YGL221C	P53081	0.48	0.22	9.00	3.84	0.62	0.69	0.00	0.00	0.00	0.22	I
YGL229C	P53036	-18.63	-0.14	76.00	0.42	0.71	0.63	0.08	0.14	0.03	0.05	III
YGR031W	P53219	-9.14	0.11	38.00	1.43	0.52	0.54	0.00	0.00	0.05	0.05	II
YGR033C	P53220	10.04	0.17	41.00	1.49	0.46	0.43	0.00	0.00	0.02	0.15	I
YGR084C	P12686	-17.07	0.12	67.00	3.13	0.46	0.53	0.03	0.04	0.10	0.09	II
YGR150C	P48237	-16.15	0.08	66.00	1.21	0.41	0.43	0.03	0.06	0.12	0.05	II
YGR174C	P37267	-12.80	0.06	66.00	2.81	0.70	0.73	0.02	0.08	0.08	0.08	III
YGR193C	P16451	-7.94	0.17	30.00	3.78	0.48	0.53	0.00	0.00	0.13	0.03	I
YGR244C	P53312	7.33	0.20	30.00	2.76	0.48	0.63	0.00	0.00	0.07	0.13	I
YHL021C	P23180	8.40	0.23	22.00	2.57	0.51	0.52	0.00	0.00	0.00	0.23	I
YHL035C	P38735	-9.25	0.00	48.00	1.69	0.54	0.55	0.06	0.02	0.00	0.08	II
YHR008C	P00447	-5.06	0.17	18.00	2.61	0.54	0.67	0.00	0.00	0.17	0.00	I
YHR024C	P11914	9.99	0.22	9.00	2.77	0.49	0.58	0.00	0.00	0.00	0.22	I
YHR037W	P07275	11.03	0.27	15.00	3.60	0.60	0.64	0.00	0.00	0.13	0.13	I
YHR051W	P00427	8.51	0.23	39.00	2.48	0.66	0.56	0.00	0.00	0.05	0.18	I
YHR147C	P32904	11.74	0.19	16.00	1.75	0.64	0.64	0.00	0.00	0.00	0.19	I
YHR183W	P38720	-10.15	-0.03	61.00	2.35	0.72	0.75	0.07	0.03	0.05	0.02	III
YHR199C	P38885	-1.23	0.15	27.00	3.85	0.39	0.41	0.00	0.04	0.07	0.11	I
YHR208W	P38891	12.07	0.25	16.00	3.42	0.27	0.39	0.00	0.00	0.13	0.13	I
YIL022W	Q01852	2.49	0.17	42.00	1.56	0.42	0.50	0.00	0.00	0.00	0.17	I
YIL042C	P40530	-11.83	0.06	77.00	2.34	0.34	0.31	0.03	0.06	0.08	0.08	II
YIL066C	P21672	-13.74	0.15	13.00	2.78	0.82	0.80	0.08	0.08	0.15	0.15	III
YIL070C	P40513	9.40	0.20	46.00	3.36	0.47	0.48	0.00	0.00	0.07	0.13	I
YIL094C	P40495	11.55	0.21	14.00	2.45	0.56	0.65	0.00	0.00	0.00	0.21	I
YIL125W	P20967	3.38	0.20	35.00	1.51	0.58	0.48	0.00	0.00	0.09	0.11	I
YIL155C	P32191	-17.99	0.08	37.00	0.99	0.46	0.46	0.03	0.00	0.00	0.11	II
YIR024C	P40576	7.50	0.19	26.00	2.30	0.48	0.50	0.00	0.00	0.04	0.15	I
YJL082W	P47031	-8.07	-0.04	72.00	2.30	0.55	0.63	0.08	0.08	0.10	0.03	III
YJL104W	P42949	6.05	0.07	27.00	2.31	0.77	0.76	0.00	0.04	0.04	0.07	I

Table B.3: Clustering result for the yeast presequence data. Used features are rounded at the second decimal place. Length of presequence are extracted the proteomic analysis.

OLN	UniProtAC	MPP	NetCharge	Length	Hmoment	$Cons_{half}$	Cons	D	E	K	R	Cluster
YJL109C	P42945	-12.35	-0.08	13.00	1.03	0.77	0.79	0.08	0.00	0.00	0.00	III
YJL131C	P47015	2.41	0.14	50.00	0.37	0.41	0.43	0.06	0.02	0.12	0.10	II
YJL133C-A	Q3E7A3	6.67	0.13	8.00	1.91	0.56	0.54	0.00	0.00	0.00	0.13	I
YJL180C	P22135	9.35	0.22	23.00	1.86	0.32	0.33	0.00	0.00	0.09	0.13	I
YJR003C	P47084	11.03	0.18	17.00	3.28	0.51	0.45	0.00	0.00	0.06	0.12	I
YJR080C	P47127	7.35	0.14	22.00	2.71	0.44	0.45	0.00	0.00	0.00	0.14	I
YJR100C	P47140	7.94	0.17	24.00	3.08	0.51	0.50	0.00	0.00	0.04	0.13	I
YJR101W	P47141	8.95	0.29	7.00	2.02	0.52	0.58	0.00	0.00	0.14	0.14	I
YJR144W	P32787	7.40	0.23	22.00	2.03	0.45	0.41	0.00	0.00	0.14	0.09	I
YKL003C	P28778	-7.16	0.08	71.00	2.67	0.68	0.72	0.03	0.06	0.10	0.07	III
YKL029C	P36013	12.95	0.21	38.00	1.70	0.40	0.40	0.00	0.00	0.00	0.21	I
YKL040C	P32860	7.33	0.19	21.00	3.76	0.40	0.51	0.00	0.00	0.14	0.05	I
YKL085W	P17505	11.29	0.33	9.00	2.09	0.64	0.71	0.00	0.00	0.11	0.22	I
YKL106W	Q01802	4.84	0.21	14.00	2.82	0.33	0.43	0.00	0.00	0.00	0.21	I
YKL132C	P36001	-15.23	-0.14	7.00	2.55	0.63	0.68	0.29	0.00	0.00	0.14	III
YKL134C	P35999	10.88	0.18	28.00	2.57	0.29	0.36	0.00	0.00	0.07	0.11	I
YKL141W	P33421	-6.52	0.16	50.00	1.58	0.38	0.39	0.00	0.00	0.08	0.08	II
YKL148C	Q00711	9.18	0.25	20.00	1.48	0.46	0.46	0.00	0.00	0.15	0.10	I
YKL150W	P36060	-10.79	0.13	23.00	2.35	0.64	0.58	0.00	0.00	0.04	0.09	I
YKL155C	P36056	-11.64	0.17	24.00	2.46	0.33	0.28	0.00	0.00	0.08	0.08	I
YKL192C	P32463	-4.40	0.16	37.00	3.28	0.50	0.48	0.00	0.00	0.00	0.16	I
YKL195W	P36046	10.75	0.26	31.00	2.29	0.40	0.38	0.00	0.00	0.00	0.26	I
YKL196C	P36015	-13.95	-0.02	45.00	1.04	0.73	0.76	0.02	0.11	0.04	0.07	III
YKR036C	P36130	5.84	0.04	45.00	3.35	0.48	0.47	0.02	0.04	0.07	0.04	II
YKR063C	P36146	5.63	0.02	87.00	1.36	0.61	0.59	0.06	0.07	0.07	0.08	II
YKR065C	P36147	11.89	0.11	36.00	1.70	0.47	0.44	0.00	0.00	0.00	0.11	I
YKR066C	P00431	-11.17	0.10	67.00	3.01	0.59	0.58	0.00	0.00	0.04	0.06	II
YLR059C	P54964	7.53	0.24	25.00	2.40	0.55	0.46	0.00	0.00	0.04	0.20	I
YLR069C	P25039	11.79	0.17	42.00	2.22	0.39	0.34	0.00	0.02	0.07	0.12	I
YLR089C	P52893	11.86	0.14	77.00	3.32	0.28	0.35	0.00	0.01	0.08	0.08	II
YLR090W	P39102	-17.38	0.00	50.00	-0.68	0.48	0.60	0.14	0.04	0.10	0.08	III
YLR091W	Q12393	-12.81	0.04	51.00	2.97	0.31	0.41	0.02	0.04	0.10	0.00	II
YLR163C	P10507	13.63	0.33	15.00	2.49	0.50	0.59	0.00	0.00	0.07	0.27	I
YLR203C	P32335	8.36	0.14	35.00	3.73	0.41	0.58	0.00	0.00	0.03	0.11	I
YLR239C	Q06005	-1.02	0.14	28.00	1.49	0.42	0.39	0.00	0.00	0.00	0.14	I
YLR259C	P19882	10.92	0.30	20.00	3.06	0.52	0.64	0.00	0.00	0.00	0.30	I
YLR295C	Q12349	3.70	0.16	32.00	2.00	0.45	0.52	0.00	0.00	0.00	0.16	I
YLR304C	P19414	10.01	0.27	15.00	2.76	0.58	0.67	0.00	0.00	0.07	0.20	I

Table B.3: Clustering result for the yeast presequence data. Used features are rounded at the second decimal place. Length of presequence are extracted the proteomic analysis.

OLN	UniProtAC	MPP	NetCharge	Length	Hmoment	$Cons_{half}$	Cons	D	E	K	R	Cluster
YLR312W-A	P36523	2.06	0.11	28.00	3.28	0.45	0.57	0.00	0.04	0.00	0.14	I
YLR355C	P06168	4.30	0.21	47.00	1.93	0.69	0.66	0.00	0.00	0.04	0.17	I
YLR395C	P04039	2.28	0.22	27.00	2.83	0.59	0.59	0.00	0.00	0.07	0.15	I
YLR419W	Q06698	-15.17	0.10	40.00	1.63	0.57	0.55	0.08	0.08	0.25	0.00	III
YML025C	P51998	10.22	0.25	16.00	3.45	0.44	0.44	0.00	0.00	0.13	0.13	I
YML042W	P32796	9.49	0.13	30.00	0.96	0.36	0.35	0.03	0.00	0.03	0.13	II
YML078W	P25719	11.34	0.25	12.00	3.02	0.62	0.65	0.00	0.00	0.08	0.17	I
YML081C-A	P81450	-12.12	0.14	29.00	3.78	0.69	0.68	0.00	0.00	0.10	0.03	III
YMR072W	Q02486	-8.56	0.04	26.00	2.59	0.49	0.45	0.00	0.04	0.04	0.04	II
YMR108W	P07342	-9.71	0.21	28.00	3.31	0.52	0.48	0.00	0.00	0.07	0.14	I
YMR115W	Q04472	8.54	0.18	34.00	1.55	0.43	0.39	0.00	0.00	0.06	0.12	I
YMR157C	Q03798	2.80	0.21	39.00	2.82	0.34	0.38	0.00	0.00	0.08	0.13	I
YMR177W	Q03218	-14.98	0.16	44.00	0.99	0.40	0.35	0.00	0.05	0.11	0.09	II
YMR186W	P15108	-14.13	-0.07	27.00	1.82	0.81	0.83	0.00	0.11	0.04	0.00	III
YMR188C	Q03246	-2.11	0.22	41.00	1.29	0.76	0.78	0.00	0.05	0.17	0.10	III
YMR189W	P49095	12.10	0.16	37.00	1.27	0.42	0.40	0.00	0.00	0.00	0.16	I
YMR192W	Q04322	-17.66	-0.14	81.00	-3.13	0.37	0.38	0.09	0.15	0.05	0.05	II
YMR193W	P36525	-18.24	0.23	35.00	2.98	0.50	0.59	0.00	0.03	0.11	0.14	I
YMR232W	Q05670	1.15	0.03	32.00	0.20	0.60	0.63	0.13	0.00	0.13	0.03	III
YMR267W	P28239	-14.83	0.07	76.00	1.93	0.51	0.55	0.03	0.05	0.08	0.07	II
YMR282C	P22136	-2.03	0.15	67.00	3.29	0.33	0.38	0.01	0.01	0.07	0.10	II
YMR287C	P39112	-9.57	0.18	51.00	2.51	0.45	0.49	0.04	0.04	0.10	0.16	II
YMR302C	P32843	0.76	0.14	22.00	2.13	0.35	0.34	0.00	0.00	0.00	0.14	I
YNL005C	P12687	6.24	0.04	27.00	2.31	0.42	0.57	0.04	0.00	0.04	0.04	II
YNL037C	P28834	12.90	0.30	10.00	2.56	0.62	0.70	0.00	0.00	0.10	0.20	I
YNL052W	P00424	7.35	0.20	20.00	2.16	0.73	0.74	0.00	0.00	0.00	0.20	I
YNL071W	P12695	12.75	0.22	27.00	3.71	0.60	0.68	0.00	0.00	0.00	0.22	I
YNL073W	P32048	5.70	0.29	28.00	2.13	0.43	0.44	0.00	0.00	0.07	0.21	I
YNL100W	P50945	2.39	-0.06	54.00	0.13	0.28	0.27	0.06	0.09	0.06	0.04	II
YNL104C	P06208	-9.71	0.00	9.00	1.48	0.54	0.64	0.00	0.11	0.11	0.00	III
YNL137C	P27929	-14.42	0.23	53.00	3.23	0.78	0.78	0.02	0.02	0.17	0.09	III
YNL169C	P39006	5.47	0.22	54.00	1.42	0.51	0.55	0.00	0.00	0.06	0.17	I
YNL177C	P53881	10.40	0.12	34.00	3.43	0.42	0.49	0.00	0.00	0.00	0.12	I
YNL185C	P53875	-11.40	0.07	58.00	2.22	0.76	0.79	0.02	0.02	0.09	0.02	III
YNL213C	P40156	-16.90	0.18	82.00	3.12	0.47	0.51	0.02	0.05	0.17	0.09	II
YNL239W	Q01532	-17.56	0.09	77.00	2.88	0.19	0.24	0.05	0.01	0.09	0.06	II
YNL284C	P36520	7.24	0.11	57.00	2.67	0.38	0.44	0.02	0.02	0.05	0.09	II
YNL315C	P32453	11.70	0.18	34.00	1.92	0.44	0.50	0.00	0.00	0.06	0.12	I

Table B.3: Clustering result for the yeast presequence data. Used features are rounded at the second decimal place. Length of presequence are extracted the proteomic analysis.

OLN	UniProtAC	MPP	NetCharge	Length	Hmoment	$Cons_{half}$	Cons	D	E	K	R	Cluster
YNR001C	P00890	11.03	0.14	36.00	2.06	0.39	0.39	0.00	0.00	0.06	0.08	I
YNR002C	P32907	-18.63	-0.13	23.00	1.91	0.46	0.43	0.09	0.13	0.04	0.04	III
YNR036C	P53732	12.35	0.11	28.00	2.53	0.44	0.52	0.00	0.00	0.00	0.11	I
YNR037C	P53733	-18.06	0.17	24.00	2.94	0.76	0.77	0.00	0.00	0.04	0.13	I
YNR041C	P32378	-13.55	0.13	54.00	2.10	0.30	0.26	0.00	0.02	0.06	0.09	II
YOL008W	Q08058	4.44	0.23	30.00	3.99	0.46	0.52	0.00	0.00	0.10	0.13	I
YOL021C	Q08162	-15.45	0.19	26.00	0.61	0.60	0.70	0.04	0.00	0.08	0.15	III
YOL071W	Q08230	7.94	0.11	35.00	2.99	0.31	0.32	0.00	0.00	0.06	0.06	I
YOR022C	Q12204	-14.86	0.10	61.00	3.36	0.48	0.51	0.05	0.00	0.07	0.08	II
YOR037W	P38909	7.90	0.25	16.00	4.17	0.41	0.42	0.00	0.00	0.06	0.19	I
YOR040W	Q12320	7.91	0.20	10.00	2.50	0.77	0.76	0.00	0.00	0.10	0.10	I
YOR065W	P07143	-6.96	0.05	76.00	2.80	0.52	0.54	0.01	0.03	0.07	0.03	II
YOR108W	Q12166	-11.07	0.11	9.00	1.20	0.54	0.64	0.00	0.00	0.11	0.00	III
YOR136W	P28241	12.23	0.29	14.00	3.10	0.60	0.69	0.00	0.00	0.00	0.29	I
YOR142W	P53598	8.76	0.25	16.00	3.72	0.57	0.65	0.00	0.00	0.13	0.13	I
YOR187W	P02992	10.51	0.20	30.00	2.34	0.47	0.45	0.00	0.00	0.07	0.13	I
YOR196C	P32875	-15.82	0.21	33.00	2.35	0.84	0.82	0.00	0.00	0.00	0.21	I
YOR215C	Q12032	9.40	0.21	19.00	2.44	0.40	0.50	0.00	0.00	0.00	0.21	I
YOR227W	Q12276	-12.53	0.11	83.00	0.88	0.46	0.51	0.04	0.02	0.11	0.06	II
YOR232W	P38523	9.31	0.21	43.00	2.44	0.50	0.46	0.00	0.00	0.05	0.16	I
YOR285W	Q12305	-12.38	-0.02	59.00	2.68	0.49	0.50	0.05	0.07	0.07	0.03	II
YOR286W	Q08742	9.86	0.17	24.00	2.77	0.41	0.52	0.00	0.00	0.04	0.13	I
YOR298C-A	O14467	3.07	0.07	29.00	0.87	0.74	0.75	0.07	0.00	0.00	0.14	III
YOR334W	Q01926	8.28	0.22	32.00	3.11	0.44	0.42	0.00	0.00	0.03	0.19	I
YOR354C	Q08818	11.59	0.10	29.00	2.86	0.45	0.36	0.03	0.00	0.03	0.10	II
YOR356W	Q08822	6.36	0.15	41.00	2.00	0.33	0.36	0.02	0.02	0.07	0.12	II
YOR374W	P46367	11.46	0.17	23.00	2.97	0.49	0.61	0.00	0.00	0.04	0.13	I
YPL040C	P48526	-18.85	0.08	84.00	2.62	0.28	0.48	0.02	0.10	0.13	0.07	II
YPL059W	Q02784	3.48	0.21	29.00	2.31	0.41	0.49	0.00	0.00	0.07	0.14	I
YPL097W	P48527	8.66	0.11	36.00	2.52	0.63	0.60	0.00	0.03	0.03	0.11	I
YPL103C	Q02883	-8.21	0.16	73.00	1.39	0.39	0.41	0.00	0.03	0.05	0.14	II
YPL132W	P19516	11.19	0.16	44.00	2.00	0.36	0.38	0.02	0.00	0.05	0.14	I
YPL135W	Q03020	7.61	0.18	34.00	2.81	0.41	0.56	0.00	0.00	0.03	0.15	I
YPL137C	Q03016	-7.93	0.08	48.00	0.79	0.46	0.51	0.06	0.02	0.10	0.06	II
YPL155C	P28743	5.90	0.15	86.00	2.84	0.58	0.58	0.01	0.00	0.02	0.14	II
YPL224C	Q08970	11.62	0.11	56.00	1.60	0.40	0.35	0.02	0.00	0.04	0.09	II
YPL226W	Q08972	-15.86	0.02	45.00	1.55	0.71	0.69	0.11	0.00	0.13	0.00	III
YPL231W	P19097	-13.46	-0.10	51.00	2.49	0.73	0.76	0.04	0.14	0.04	0.04	III

Table B.3: Clustering result for the yeast presequence data. Used features are rounded at the second decimal place. Length of presequence are extracted the proteomic analysis.

OLN	UniProtAC	MPP	NetCharge	Length	Hmoment	$Cons_{half}$	Cons	D	E	K	R	Cluster
YPL262W	P08417	5.41	0.25	24.00	2.66	0.45	0.56	0.00	0.00	0.13	0.13	I
YPL270W	P33311	-10.90	0.15	79.00	2.11	0.43	0.40	0.00	0.01	0.04	0.13	II
YPL271W	P21306	-9.79	0.07	55.00	1.92	0.74	0.73	0.02	0.02	0.05	0.05	III
YPR001W	P43635	7.68	0.17	23.00	1.91	0.48	0.55	0.00	0.00	0.04	0.13	I
YPR002W	Q12428	-14.49	0.15	54.00	1.73	0.43	0.45	0.04	0.04	0.20	0.02	II
YPR004C	Q12480	-0.56	0.23	22.00	2.46	0.44	0.55	0.00	0.00	0.18	0.05	I
YPR006C	Q12031	7.38	0.22	32.00	2.89	0.50	0.56	0.03	0.00	0.16	0.09	I
YPR011C	Q12251	-5.12	-0.08	25.00	2.01	0.63	0.71	0.08	0.08	0.08	0.00	III
YPR024W	P32795	0.98	0.09	11.00	2.37	0.42	0.43	0.00	0.00	0.09	0.00	I
YPR025C	P37366	4.85	0.02	65.00	0.48	0.47	0.62	0.11	0.05	0.12	0.05	II
YPR047W	P08425	11.06	0.25	16.00	2.34	0.41	0.48	0.00	0.00	0.06	0.19	I
YPR067W	Q12425	-2.67	0.17	35.00	2.54	0.42	0.43	0.00	0.00	0.06	0.11	I
YPR113W	P06197	-7.23	0.04	55.00	2.40	0.73	0.72	0.00	0.04	0.05	0.02	III
YPR134W	P08593	-4.96	0.00	10.00	1.88	0.41	0.50	0.00	0.10	0.00	0.10	II
YPR155C	Q12374	-17.18	-0.11	46.00	0.93	0.58	0.49	0.04	0.20	0.02	0.11	III
YPR166C	P10663	-18.84	0.19	16.00	2.85	0.78	0.78	0.00	0.00	0.13	0.06	III

Table B.3: Clustering result for the yeast presequence data. Used features are rounded at the second decimal place. Length of presequence are extracted the proteomic analysis.

Bibliography

[1] J. C. Thrash, A. Boyd, M. J. Huggett, J. Grote, P. Carini, R. J. Yoder, B. Robbertse, J. W. Spatafora, M. S. Rappé, and S. J. Giovannoni, "Phylogenomic evidence for a common ancestor of mitochondria and the sar11 clade," *Sci Rep*, vol. 1, p. 13, 2011.

[2] M. R. Duchen and G. Szabadkai, "Roles of mitochondria in human disease," *Essays Biochem.*, vol. 47, pp. 115–137, 2010.

[3] F. Vögtle, S. Wortelkamp, R. 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*, vol. 139, no. 2, pp. 428–439, 2009.

[4] G. von Heijne, "Mitochondrial targeting sequences may form amphiphilic helices," *EMBO J*, vol. 5, pp. 1335–42, Jun 1986.

[5] G. Schneider, S. Sjöling, E. Wallin, P. Wrede, E. Glaser, and G. von Heijne, "Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides," *PROTEINS*, vol. 30, pp. 49–60, 1998.

[6] A. Chacinska, C. M. Koehler, D. Milenkovic, T. Lithgow, and N. Pfanner, "Importing mitochondrial proteins: machineries and mechanisms," *Cell*, vol. 138, no. 4, pp. 628–644, 2009.

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

[8] O. Schmidt, N. Pfanner, and C. Meisinger, "Mitochondrial protein import: from proteomics to functional mechanisms.," *Nat Rev Mol Cell Biol*, vol. 11, no. 9, pp. 655–667, 2010.

[9] Y. Abe, T. Shodai, T. Muto, K. Mihara, H. Torii, S. Nishikawa, T. Endo, and D. Kohda, "Structural basis of presequence recognition by the mitochondrial protein import receptor Tom20," *Cell*, vol. 100, no. 5, pp. 551–560, 2000.

[10] K. Yamano, Y. Yatsukawa, M. Esaki, A. E. Hobbs, R. E. Jensen, and T. Endo, "Tom20 and Tom22 share the common signal recognition pathway in mitochondrial protein import," *J. Biol. Chem.*, vol. 283, no. 7, pp. 3799–3807, 2008.

[11] T. H. Lin, R. F. Murphy, and Z. Bar-Joseph, "Discriminative motif finding for predicting protein subcellular localization," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 8, no. 2, pp. 441–451, 2011.

[12] E. J. Williams, C. Pal, and L. D. Hurst, "The molecular evolution of signal peptides," *Gene*, vol. 252, no. 2, pp. 313–322, 2000.

[13] O. Gakh, P. Cavadini, and G. Isaya, "Mitochondrial processing peptidases," *Biochim Biophys Acta*, vol. 1592, no. 1, pp. 63–77, 2002.

[14] 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*, vol. 26, no. 3, pp. 327–334, 1994.

[15] 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*, vol. 5, no. 6, pp. 1327–1334, 1986.

[16] P. Moberg, A. Ståhl, 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*, vol. 36, no. 5, pp. 616–628, 2003.

[17] G. Isaya, F. Kalousek, and L. E. Rosenberg, "Amino-terminal octapeptides function as recognition signals for the mitochondrial intermediate peptidase.," *J Biol Chem*, vol. 267, no. 11, pp. 7904–7910, 1992.

[18] A. Candat, P. Poupart, J. P. Andrieu, A. Chevrollier, P. Reynier, H. Rogniaux, M. H. Avelange-Macherel, and D. Macherel, "Experimental determination of organelle targeting-peptide cleavage sites using transient expression of green fluorescent protein translational fusions," *Anal. Biochem.*, vol. 434, no. 1, pp. 44–51, 2013.

[19] S. C. Botelho, M. Osterberg, A. S. Reichert, K. Yamano, P. Björkholm, T. Endo, G. von Heijne, and H. Kim, "Tim23-mediated insertion of transmembrane α-helices into the mitochondrial inner membrane," *EMBO J*, vol. 30, pp. 1003–11, Mar 2011.

[20] T. Hessa, N. M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S. H. White, and G. von Heijne, "Molecular code for transmembrane-helix recognition by the sec61 translocon," *Nature*, vol. 450, pp. 1026–30, Dec 2007.

[21] E. Wallin and G. von Heijne, "Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms," *Protein Sci*, vol. 7, pp. 1029–38, Apr 1998.

[22] G. Casari, M. De Fusco, S. Ciarmatori, M. Zeviani, M. Mora, P. Fernandez, G. De Michele, A. Filla, S. Cocozza, R. Marconi, A. Dürr, B. Fontaine, and A. Ballabio, "Spastic paraplegia and oxphos impairment caused by mutations in paraplegin, a nuclear-encoded mitochondrial metalloprotease," *Cell*, vol. 93, pp. 973–83, Jun 1998.

[23] M. Nolden, S. Ehses, M. Koppen, A. Bernacchia, E. I. Rugarli, and T. Langer, "The m-AAA protease defective in hereditary spastic paraplegia controls ribosome assembly in mitochondria," *Cell*, vol. 123, no. 2, pp. 277–289, 2005.

[24] 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*, vol. 20, no. 5, pp. 867–879, 2011.

[25] O. Yogev and O. Pines, "Dual targeting of mitochondrial proteins: mechanism, regulation and function," *Biochim. Biophys. Acta*, vol. 1808, no. 3, pp. 1012–1020, 2011.

[26] C. Christopher and I. Small, "A reevaluation of dual-targeting of proteins to mitochondria and chloroplasts," *Biochim. Biophys. Acta*, vol. 1833, no. 2, pp. 253–259, 2013.

[27] 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*, vol. 269, no. 8, pp. 6001–6010, 1994.

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

[29] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, "Improved prediction of signal peptides: SignalP 3.0.," *J Mol Biol*, vol. 340, no. 4, pp. 783–795, 2004.

[30] I. Dondoshansky, *Blastclust (NCBI software development toolkit)*, 2002.

[31] I. Small, N. Peeters, F. Legeai, and C. Lurin, "Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences," *Proteomics*, vol. 4, no. 6, pp. 1581–1590, 2004.

[32] 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 Physiology*, vol. 150, no. 3, pp. 1272–1285, 2009.

[33] 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*, vol. 15, no. 10, pp. 1456–1461, 2005.

[34] A. M. Altenhoff and C. Dessimoz, "Inferring orthology and paralogy," in *Evolutionary Genomics: Statistics and Computational Methods* (M. Anisimova, ed.), Methods in Molecular Biology, pp. 259–277, USA: Humana Press, 2012.

[35] R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev, "The use of gene clusters to infer functional coupling," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, no. 6, pp. 2896–2901, 1999.

[36] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010. USEARCH.

[37] 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*, vol. 30, no. 14, pp. 3059–3066, 2002.

[38] 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*, vol. 21, no. 9, pp. 1781–1791, 2004.

[39] F. Johansson and H. Toh, "A comparative study of conservation and variation scores.," *BMC Bioinformatics*, vol. 11, p. 388, 2010.

[40] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J Mol Biol*, vol. 157, no. 1, pp. 105–132, 1982.

[41] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

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

[43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, p. 10, 2009.

[44] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[45] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[46] C. Hsu and C. Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, 2002.

[47] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *The Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2001.

[48] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *International Joint Conference on Artificial Intelligence*, pp. 1022–1027, Morgan Kaufmann, 1993.

[49] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. on Knowl. and Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

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

[51] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[52] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[53] S. Argarwal, T. Graepel, R. Harbrich, S. Har-Peled, and D. Roth, "Generalization bounds for the area under the ROC curve," *J. Mach. Learn. Res.*, vol. 6, pp. 393–425, 2005.

[54] B. Dujon, "Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution," *Trends Genet*, vol. 22, no. 7, pp. 357–387, 2006.

[55] R. P. Zahedi, A. Sickmann, A. M. Boehm, C. Winkler, N. Zufall, B. Schönfisch, B. Guiard, N. Pfanner, and C. Meisinger, "Proteomic analysis of the yeast mitochondrial outer membrane reveals accumulation of a subclass of preproteins," *Mol. Biol. Cell*, vol. 17, no. 3, pp. 1436–1450, 2006.

[56] M. Kambacheld, S. Augustin, T. Tatsuta, S. Muller, and T. Langer, "Role of the novel metallopeptidase Mop112 and saccharolysin for the complete degradation of proteins residing in different subcompartments of mitochondria," *J. Biol. Chem.*, vol. 280, no. 20, pp. 20132–20139, 2005.

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

[58] F. Bonn, T. Tatsua, C. Petrungaro, J. Riemer, and T. Langer, "Presequence-dependent folding ensures MrpL32 processing by the m-AAA protease in mitochondria," *EMBO J*, vol. 30, no. 13, pp. 2545–2556, 2011.

[59] L. Grohmann, H. R. Graack, V. Kruft, T. Choli, S. Goldschmidt-Reisin, and M. Kitakawa, "Extended N-terminal sequencing of proteins of the large ribosomal subunit from yeast mitochondria," *FEBS Lett.*, vol. 284, no. 1, pp. 51–56, 1991.

[60] 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*, vol. 22, no. 13, pp. 2135–2143, 2011.

[61] S. R. Doyle, N. R. Kasinadhuni, C. K. Chan, and W. N. Grant, "Evidence of evolutionary constraints that influences the sequence composition and diversity of mitochondrial matrix targeting signals," *PLoS ONE*, vol. 8, no. 6, p. e67938, 2013.

[62] L. Rosso, A. C. Marques, A. S. Reichert, and H. Kaessmann, "Mitochondrial targeting adaptation of the hominoid-specific glutamate dehydrogenase driven by positive darwinian selection," *PLoS genetics*, vol. 4, no. 8, p. e1000150, 2008.

[63] J. A. Capra and M. Singh, "Predicting functionally important residues from sequence conservation," *Bioinformatics*, vol. 23, no. 15, pp. 1875–1882, 2007.

[64] 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*, vol. 35, no. Web Server issue, pp. W585–W587, 2007.

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

[66] A. Terada, M. Okada-Hatakeyama, K. Tsuda, and J. Sese, "Statistical significance of combinatorial regulations," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 110, no. 32, pp. 12996–13001, 2013.

[67] A. Aboderin, "An empirical hydrophobicity scale for α-amino-acids and some of its applications," *International Journal of Biochemistry*, vol. 2, no. 11, pp. 537–544, 1971.

[68] R. Hughey and A. Krogh, "Hidden Markov models for sequence analysis: extension and analysis of the basic method," *Comput. Appl. Biosci.*, vol. 12, no. 2, pp. 95–107, 1996.

[69] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[70] K. Imai and S. Mitaku, "Mechanisms of secondary structure breakers in soluble proteins," *Biophysics*, vol. 1, no. 0, pp. 55–65, 2005.

[71] K. Imai, N. Fujita, M. M. Gromiha, and P. Horton, "Eukaryote-wide sequence analysis of mitochondrial β-barrel outer membrane proteins," *BMC Genomics*, vol. 12, p. 79, 2011.

[72] 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*, vol. 278, no. 3, pp. 1879–1885, 2003.

[73] K. Kojima, S. Kitada, T. Ogishima, and A. Ito, "A proposed common structure of substrates bound to mitochondrial processing peptidase.," *J Biol Chem*, vol. 276, no. 3, pp. 2115–2121, 2001.

[74] 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.

[75] M. G. Claros and P. Vincens, "Computational method to predict mitochondrially imported proteins and their targeting sequences," *Eur. J. Biochem.*, vol. 241, no. 3, pp. 779–786, 1996.

[76] 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*, vol. 9, no. 7, pp. 615–625, 2001.

[77] T. Benaglia, D. Chauveau, D. Hunter, and D. Young, "mixtools: An r package for analyzing finite mixture models," *Journal of Statistical Software*, vol. 32, no. 6, pp. 1–29, 2009.

[78] J. B. Peltier, A. J. Ytterberg, Q. Sun, and K. J. van Wijk, "New functions of the thylakoid membrane proteome of Arabidopsis thaliana revealed by a simple, fast, and versatile fractionation strategy," *J. Biol. Chem.*, vol. 279, pp. 49367–49383, 2004.

[79] 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*, vol. 108, no. 1, pp. 91–96, 2011.

[80] T. Saitoh, M. Igura, Y. Miyazaki, T. Ose, N. Maita, and D. Kohda, "Crystallographic snapshots of tom20-mitochondrial presequence interactions with disulfide-stabilized peptides," *Biochemistry*, vol. 50, pp. 5487–96, Jun 2011.

[81] M. Dinur-Mills, M. Tal, and O. Pines, "Dual targeted mitochondrial proteins are characterized by lower mts parameters and total net charge," *PLoS One*, vol. 3, no. 5, p. e2161, 2008.

[82] D. J. Pagliarini, S. E. Calvo, B. Chang, S. A. Sheth, S. B. Vafai, S. E. Ong, G. A. Walford, C. Sugiana, A. Boneh, W. K. Chen, D. E. Hill, M. Vidal, J. G. Evans, D. R. Thorburn, S. A. Carr, and V. K. Mootha, "A mitochondrial protein compendium elucidates complex I disease biology," *Cell*, vol. 134, no. 1, pp. 112–123, 2008.

[83] C. Meisinger, A. Sickmann, and N. Pfanner, "The mitochondrial proteome: from inventory to function," *Cell*, vol. 134, no. 1, pp. 22–24, 2008.

[84] C. M. Lee, J. Sedman, W. Neupert, and R. A. Stuart, "The dna helicase, hmi1p, is transported into mitochondria by a c-terminal cleavable targeting signal," *Journal of Biological Chemistry*, vol. 274, no. 30, pp. 20937–20942, 1999.

[85] R. Ieva, A. K. Heißwolf, M. Gebert, F.-N. Vögtle, F. Wollweber, C. S. Mehnert, S. Oeljeklaus, B. Warscheid, C. Meisinger, M. van der Laan, *et al.*, "Mitochondrial inner membrane protease promotes assembly of presequence translocase by removing a carboxy-terminal targeting sequence," *Nature communications*, vol. 4, 2013.

[86] M. A. Lomize, A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg, "Opm: orientations of proteins in membranes database," *Bioinformatics*, vol. 22, pp. 623–5, Mar 2006.

[87] H. J. Sharpe, T. J. Stevens, and S. Munro, "A comprehensive comparison of transmembrane domains reveals organelle-specific properties," *Cell*, vol. 142, pp. 158–69, Jul 2010.

[88] G. M. Boratyn, A. A. Schäffer, R. Agarwala, S. F. Altschul, D. J. Lipman, and T. L. Madden, "Domain enhanced lookup time accelerated blast," *Biol Direct*, vol. 7, p. 12, 2012.

[89] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, pp. 3389–402, Sep 1997.

[90] A. Bernsel, H. Viklund, J. Falk, E. Lindahl, G. von Heijne, and A. Elofsson, "Prediction of membrane-protein topology from first principles," *Proc Natl Acad Sci U S A*, vol. 105, pp. 7177–81, May 2008.

[91] H. Viklund and A. Elofsson, "Octopus: improving topology prediction by two-track ann-based preference scores and an extended topological grammar," *Bioinformatics*, vol. 24, pp. 1662–8, Aug 2008.

[92] D. M. Engelman, T. A. Steitz, and A. Goldman, "Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins," *Annu Rev Biophys Biophys Chem*, vol. 15, pp. 321–53, 1986.

[93] T. Hirokawa, S. Boon-Chieng, and S. Mitaku, "Sosui: classification and secondary structure prediction system for membrane proteins," *Bioinformatics*, vol. 14, no. 4, pp. 378–9, 1998.

[94] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden markov model: application to complete genomes," *J Mol Biol*, vol. 305, pp. 567–80, Jan 2001.

[95] L. Käll, A. Krogh, and E. L. Sonnhammer, "A combined transmembrane topology and signal peptide prediction method.," *J Mol Biol*, vol. 338, no. 5, pp. 1027–1036, 2004.

[96] S. Sekine, Y. Kanamaru, M. Koike, A. Nishihara, M. Okada, H. Kinoshita, M. Kamiyama, J. Maruyama, Y. Uchiyama, N. Ishihara, K. Takeda, and H. Ichijo, "Rhomboid protease parl mediates the mitochondrial membrane potential loss-induced cleavage of pgam5," *J Biol Chem*, vol. 287,

pp. 34635–45, Oct 2012.

[97] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.