Learning Cellular Sorting Pathways Using Protein Interactions and Sequence Motifs

Tien-ho Lin

CMU-10-021

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:

Ziv Bar-Joseph (Carnegie Mellon University, Chair)
Robert F. Murphy (Carnegie Mellon University, Chair)
Jaime Carbonell (Carnegie Mellon University)
David Heckerman (Microsoft Research)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technologies

Keyword: Protein Sorting Pathways, Subcellular Localization, Protein-Protein Interactions, Sequence Motifs, Proteomics

Abstract

Proper subcellular localization is critical for proteins to perform their roles in cellular functions. Proteins are transported by different cellular sorting pathways, some of which take a protein through several intermediate locations until reaching its final destination. The pathway a protein is transported through is determined by carrier proteins that bind to specific sequence motifs. This thesis introduces new computational methods that extract these sequence motifs and carrier proteins, and learn the sorting pathways.

We first develop a system that utilizes the known cellular sorting pathways to learn sequence motifs and predict locations. We proposed a discriminative motif finding method that identifies potential targeting motifs. Our method utilizes a tree structure mimicking the known targeting pathways. Using these motifs we were able to improve localization prediction on a benchmark dataset of yeast proteins. The motifs identified are more conserved than the average protein sequence. Using our motif-based predictions we were also able to correct annotation errors in public databases for the location of some of the proteins.

Furthermore we present a new method that integrates sequence, motif and protein interaction data to model how proteins are sorted through the sorting pathways with a hidden Markov model (HMM). Using data for yeast, we show that our model leads to accurate prediction of subcellular localization. We also show that the pathways learned by our model recover many known sorting pathways and correctly assign proteins to the path they utilize. We extend this model to support alternative splicing and multiple cell types in higher organisms. Using our method we performed the first systematic discovery of targeting pathways in the human proteome based on confocal microscopy images on HPA. We show that our pathways structure improves localization prediction, and the learned structure resembles our basic understanding of cellular sorting mechanism.

Acknowledgements

First and foremost, I would like to thank my advisers Ziv Bar-Joseph and Robert F. Murphy for their support and guidance. Both the professional and the personal relation with them are invaluable. Their encouragement is the most important reason I can finish my graduate study. I learned a lot from how Ziv choose and define research problems, the way he approaches problems, what he emphasizes and what he avoids. Bob started me in the field that becomes my research focus and inspired me to seek biological problems with true impact. His experience taught me a great deal about how to do research in computational biology.

I would also like to thank my other committee members, Jaime Carbonell and David Heckerman. They gave me insightful comments, idea and suggestions that benefit this thesis greatly. I was also fortunate to intern at David's group on the summer of 2009. David has been such a wonderful mentor, and I learn to be involved in a larger collaboration project. I want to thank everyone there for a rewarding experience.

This work would not be possible without many people's help. Particularly, Jennifer Bakal provided excellent programming support. Jieyue Li in the Murphy lab helped processing the Human Protein Atlas data. I appreciate how they took care of all my requests timely even under tight schedule.

I have many other people to acknowledge and thank. Current and former members of the systems biology group and the Murphy lab: Anthony Gitter, Guy Zinman, Hai-Son Le, Shan Zhong, Marcel Schulz, Peter Huggins, Jason Ernst, Yanjun Qi, Yong Lu, Yanxin Shi, Tao Peng, Armaghan Naik, Luis Coelho, Taraz Buck, Aabid Shariff, Joshua Kangas, and Aparna Kumar. Former coauthors for their collaboration, which while not included in this dissertation, is still an important part of my training. All the students, faculty, and staff of Language Technology Institute and Lane Center of Computational Biology for creating a stimulating environment to pursue scientific knowledge.

Contents

Chapter 1

Introduction

1.1 Motivation

An important challenge in systems biology is to build detailed models of cell organization that provide accurate predictions of cell behaviors. Many (if not all) of the proteins expressed by a given cell require proper subcellular localization in order to make their contributions to those behaviors. The location of a protein provides information about its function and interacting partners [1,2]. Aberrant localization has a role in certain diseases, including cancer [3,4], Alzheimer's disease [5], hyperoxaluria [6] and cystic fibrosis [7]. The effect of some drugs depends on their role in changing protein locations [8]. The knowledge about protein sorting will help us understand such diseases and the drug effect. Hence a proteome-wide understanding of subcellular localization is critical in understanding the protein behaviors within a cell. Our aim is to find out where a protein is transported in the cell, the path it passes through, and the mechanism that determines its path.

Extensive work has been done on proteome-scale determination of location from yeast to human at various levels of resolution, both by fractionation [9,10] and by microscopy [1,11–14]. Automatic prediction of location based on microscopy images is now very effective [15]. Databases containing localization information curated from the literature are also available, including SGD [16], FlyBase [17] and UniProt [18]. With these resources we believe it is now possible to study the cellular sorting mechanism using computational analysis. This is particularly important for the less-understood *sub-compartments*, considering that many proteins are found in only a specific region of an organelle. In addition, proteins are sorted through several compartments until reaching the destination, with each transport along the

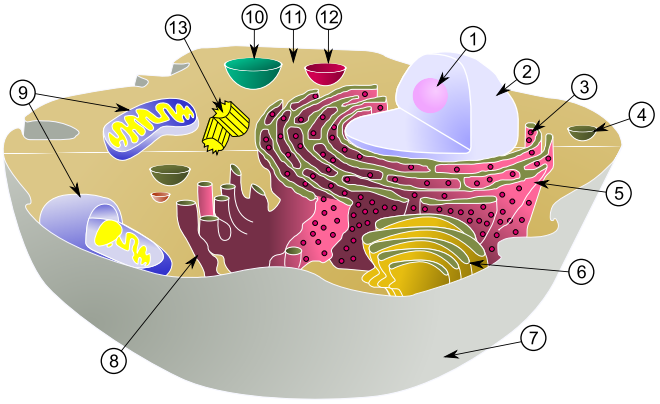


Figure 1.1: The major organelles within a typical animal cell: (1) nucleolus, (2) nucleus, (3) ribosome, (4) vesicle, (5) rough ER, (6) Golgi apparatus, (7) cytoskeleton, (8) smooth ER, (9) mitochondrion, (10) vacuole, (11) cytosol, (12) lysosome, and (13) centriole. Image is from http://en.wikipedia.org/wiki/File:Biological_cell.svg.

path determined by a carrier-protein binding to targeting signals. Several such targeting pathways have been widely studied, like the secretory pathway, but it is believed that there are more non-conventional pathways.

To have greater biomedical impact we would like to support higher organism, especially human, as well as unicellular organism like yeast. In higher organism protein locations may vary between cell types or within the same cell type under different conditions. For example, changes in protein subcellular location are associated with differentiation [19]. There are tens of thousands of proteins for potentially hundreds of cell or tissue types under many conditions, and collecting information for all combinations is infeasible. We need a system that infers location changes in cell types or conditions. If the rules are interpretable, it could shed light on how changes in location are regulated and will greatly benefit our understanding of protein targeting in human.

1.2 Background

1.2.1 Major Cellular Compartments

The major cellular compartments, or organelles, are displayed in Figure 1.1. The cell is enclosed by the plasma membrane (PM), and the interior is the cytosol which contains the organelles, including the nucleus that stores the DNA in which proteins are encoded. The

mitochondrion, sometimes called the "power plant" in a cell, is the organelle that generates energy. The peroxisome breaks down fatty acids and amino acids for reuse and also rid the cell of toxic peroxides. The secretory pathway is the most common procedure to transport a protein out of the cell, and enzymes residing in related organelles are also sorted this way. Most proteins are synthesized in the cytosol, and those with an N-terminal signal peptides will be bound by the protein complex signal recognition particle (Srp) and taken to the endoplasmic reticulum (ER) during synthesis. Then proteins with a C-terminal ER retention signal sequence KDEL (bound by the KDEL receptor) will remain in the ER, the reminder transported to the Golgi apparatus; from whence proteins are sorted to the lysosome or the plasma membrane or outside the cell. The lysosome in animal cells or vacuole in yeast and plant cells is described as the "garbage disposal" system of the cell, responsible for digestion of macromolecules. A good introduction is given in [20].

1.2.2 Cellular Sorting Pathways

Proteins are transported to the compartments by cellular sorting pathways. A sorting pathway sorts a protein through a series of unobserved intermediate locations until reaching the final destination. For example, in the secretory pathway proteins are transported through ER and Golgi, reaching either lysosome, PM, or secreted outside the cell. The decision of moving from one intermediate location to the next is determined by carrier-proteins that recognize specific sequence motifs. Such sorting motifs are typically short and sometimes called signal sequences. For example, Srp binds to the ER signal peptides, taking proteins into the secretory pathway; Erd2 binds to KDEL motif so proteins remain in ER.

Besides the secretory pathway, other sorting pathways also use such mechanism. A protein is imported to the nucleus if the Importin binds to nuclear localization signal (NLS). It may be further imported to the nucleulus given the nucleolar localization motif. Similarly a protein enters the mitochondria if the translocase of the outer membrane (TOM) complex binds to the mitochondrial targeting signal (MTS). To enter the peroxisome a protein either contains the peroxisomal targeting signals 1 (PTS1) which is recognized by Pex5, or PTS2 which is recognized by Pex7.

These classical pathways are believed to be conserved from the simplest to the most complex eukaryote. So far the discovery of cellular sorting pathways relies on manual investigation and experiments focusing on a specific topic, protein, or motif. On the other hand, there are non-classical or alternative pathways that are followed by a minor fraction

of proteins or that differ from the first discovered pathway. Non-classical pathways include leaderless secretion pathway and cytoplasm-to-vacuole targeting (CVT) pathway. In both cases a protein does not pass through ER and Golgi as in the classical secretory pathway, but directly moves to the destination (secreted or vacuole, respectively). One can verify whether a protein belongs to a non-classical pathway by inhibiting classical ones experimentally.

1.3 Related Work

Fluorescent microscopy imaging technique has provided proteome-wide data on localization in yeast [1]. Automated determination of localization from images is also accurate [15]. The Human Protein Atlas (HPA) team has collected confocal microscopy images using antibodies in human [14], and automated analysis has also been performed [13, 21]. However image-based analysis does not provide causal insight about the mechanism.

Most of the previous computational biology research on protein sorting focus on predicting the location, e.g. WoLF PSort [22], TargetP [23], LOCtree [24], PSLT2 [25] and DC-kNN [26]. Some are based on the current (partial) knowledge of protein sorting (e.g. signal sequences), making the decision rules interpretable. Some learn from data but the decision rules are not interpretable (e.g. amino acid composition and support vector machine). Although the classical sorting pathways play an important role in protein sorting, most predictors do not utilize any structure among the compartments; only a small number utilize the established pathway structure (e.g. LOCtree) and show improvement. Very few previous methods try to extract novel sorting motifs that explain the localization. We are not aware of any previous work that identifies novel protein carriers as well (DC-kNN utilizes the protein network but does not provide insight on the mechanism). With the availability of more protein localization resources, it is important to have computational tools that extract novel sorting motifs from sequences, which is the standard mechanism for protein sorting. In Chapter 2 these predictors are discussed in more detail.

We are not aware of prior research on learning novel sorting pathways from data. Some methods learn decision trees for predicting subcellular localization, including PSLT2 [25] and YimLOC [27]. While the decision trees generated by these methods are often quite accurate, they are not intended to reflect sorting pathways, and they utilize features that, while useful for classification, are not related to the biochemical process of protein sorting. There are computational methods that predict whether a protein goes through a specific

pathway or not (for example, SignalP [28] and SecretomeP [29]). However, these methods rely on the pathway as an input and cannot be used to infer new pathways.

1.4 Overview of Thesis

The overarching goal of this thesis is to study the cellular sorting mechanism by modeling targeting pathways, in which the path is determined by carrier-proteins and sequence motifs. In Chapter 2 we first present a system that utilizes the known targeting pathways to learn motifs and predict locations. We use a tree structure to mimic the targeting pathways. Motifs are represented as profile hidden Markov models (HMM) which allow insertions and deletions of variable-length. The HMMs are learned by a novel discriminative motif finding method. These models search for motifs that are present in a compartment but absent in other, nearby, compartments by utilizing a hierarchical structure that mimics the protein sorting mechanism. This method predicts the localization at least as good as the state-of-the-art system based on known motifs on a benchmark dataset of yeast proteins. Both discriminative motif finding and the hierarchical structure improve the performance in prediction. The motifs identified can be mapped to known targeting motifs and they are more conserved than the average protein sequence. Using our motif-based predictions we can identify potential annotation errors in public databases for the location of some of the proteins.

Besides relying on the established knowledge of protein targeting pathways, we aim to discover novel pathways from sequence, interaction, and localization data. In Chapter 3 we developed a new method that integrates sequence, motif and protein interaction data to model how proteins are sorted through targeting pathways. We use a hidden Markov model (HMM) to represent targeting pathways. The model is able to determine intermediate sorting states and to assign carrier proteins and motifs to the sorting pathways. In simulation studies, we show that the method can accurately recover an underlying sorting model. Using data for yeast, we show that our model leads to accurate prediction of subcellular localization. We also show that the pathways learned by our model recover many known sorting pathways and correctly assign proteins to the path they utilize. The learned model identified new pathways and their putative carriers and motifs and these may represent novel protein sorting mechanisms.

Although proteome information is more abundant in yeast, it is of more importance to

understand targeting pathways in higher organisms, especially human. In Chapter 4, we address challenges in higher organisms in order to support this human localization data. With the availability of large amount of location proteomic data based on confocal microscopy images using antibodies from Human Protein Atlas (HPA) [14], we extend our targeting pathway model from yeast to human. The method supports alternative splicing which is common in higher organisms. Furthermore we can utilize localization data in multiple cell types and conditions to examine common and condition-specific carriers, motifs, and pathways. Using the extended model, we performed the first systematic discovery of targeting pathways in the human proteome based on confocal microscopy images on HPA. By comparing to a classifier without using a structure we show that incorporating the targeting pathway leads to more accurate prediction of the destinate compartment.

Chapter 2

Motifs Based on Predefined Sorting Pathways[1]

In this chapter we present how to extract targeting motifs based on predefined targeting pathways, and how to predict localization using the extracted motifs. In the next chapter we will discuss how to infer pathways from protein sequence, interaction, and localization data.

Here we develop and apply a discriminative motif finding algorithm which utilizes HMMs that are constructed to optimize a discriminative criteria, the conditional likelihood of the sequences given the motifs. We used maximal mutual information estimate (MMIE), a technique that was initially applied to speech recognition, to train these HMMs discriminatively. Our models select motifs that are unique to the different compartments. In addition to their use for classification they may also provide information about the function of the proteins in each compartment or the mechanisms involved in targeting these proteins to their cellular locations.

A hierarchical structure or a tree has been used to represent targeting pathways in predicting subcellular localization, and accuracy improves [24,31,32]. We apply such structures to motif discovery, rather than only prediction, by searching for discriminative motifs at every split (internal nodes) on the hierarchical compartment structure in Figure 2.1. This allows us to take advantage of current biological knowledge regarding the organization of compartments within a cell.

[1]The content of this chapter is based on the paper [30]

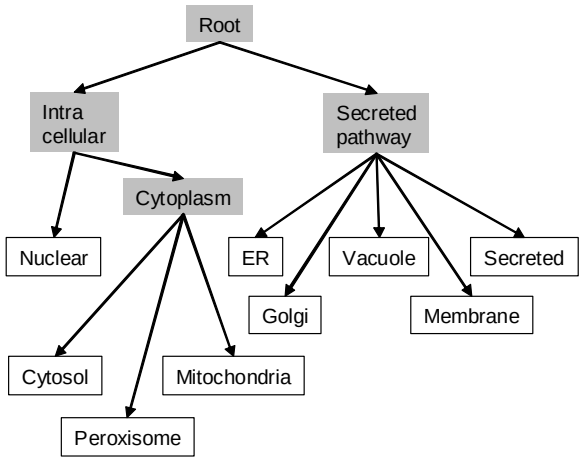


Figure 2.1: Hierarchical structure of compartments based on cellular sorting.

For subcellular compartment classification, our discriminative HMM method that does not utilize any prior motif information improves upon methods that use a list of known motifs. We also show that incorporating the protein sorting hierarchy results in better prediction on average. Our method was able to recover known motifs and to suggest new motifs for various compartments. These new motifs are more conserved than average amino acids in agreement with their predicted role in protein localization. Using our predicted motifs we were also able to reassign a number of proteins to new compartments, correcting what we believe are errors in current annotation databases.

2.1 Related Work

2.1.1 Classifying Subcellular Localization

A number of methods have been proposed for using sequence information to predict localization. These include WoLF PSort [22], TargetP [23], LOCtree [24], PSLT2 [25], TBpred [33], and iPSORT [34]. While useful, some of these methods (e.g. LOCtree and WoLF PSort) are based on general sequence characteristics (GC content etc.) and thus it is hard to interpret the sequence features that lead to accurate classification in terms of localization mechanism. Some (e.g. TargetP and WoLF PSort) are based on known motifs, making it hard to correctly classify proteins that lack the known motifs. PSLT2 considers all motifs in the InterPro database [35], but does not try to search for novel targeting motifs. Beside known

motifs TargetP also uses the motif finder MEME [36] to characterize manually curated special cases of the mitochondrial targeting signal [23]. This procedure is necessary because no well defined sequence motif has been previously found for the mitochondrial targeting signal, but it cannot discover novel targeting motifs. TBpred uses MEME to identify one motif overrepresented in each of the four subcellular locations of mycobacterial proteins [33], which is arguably not enough to explain all targeting pathways. Such a motif could also appear in other locations and therefore may not be associated with localization. TBpred made no attempt to examine or interpret the biological meaning of the identified motif. iPSORT can discover two type of features, amino acid properties and pattern matching. However amino acid properties (e.g. hydrophobic or hydrophilic) may be a result of the biochemical characteristics of the compartments, and do not provide as much information on the protein sorting mechanism as motifs do. iPSORT can discover patterns, but the patterns are not as expressive as common motif representation like regular expression.

2.1.2 Generative Motif Finding

Protein sequence motifs are represented by profile hidden Markov models (HMMs), which models local alignments using match, insert, and delete states [37]. Profile HMMs have been successfully utilized to model protein families and domains, and they are used to represent domains in the Pfam database [38]. Unlike position weight matrices (PWMs) (for example, those used by MEME [36]), profile HMMs allow for variable length insertions and deletions that are common in protein motifs, for example the nucleoplasmin nuclear location sequence [39] and the sequence targeting proteases to the food vacuole in *P. falciparum* [40]. Unlike regular expressions, which have also been used to represent such motifs, profile HMMs can assign different frequencies to each amino acid and are thus more expressive.

Traditional motif finding algorithms start by assembling a subset of sequences (for example, all proteins in the same compartment) and then searching for motifs in those sequences. These methods typically utilize generative models that attempt to model the process by which the motifs were generated based on simplifying assumptions. Generative motif finding methods and models for proteins include MEME [36] and NestedMICA [41] using PWMs, and HMMER [37] using profile HMMs, among others.

2.1.3 Discriminative Motif Finding

While generative motif finding is useful, these methods do not use important information about the negative set (sequences that are assigned to other compartments) when constructing the models. Such information may be useful for building refined models of the differences between similar compartments. Relatively little work has focused on a different approach: discriminative learning of probabilistic motif models. Discriminative methods search for motifs that are present in one class (positive set) but absent in other classes (negative set). Most of such methods focus on DNA motifs, usually for transcription factor binding sites. There are string-based methods, for example DWE [42], but probabilistic models like PWM are more expressive. Representing motifs as PWM, Segal *et al.* [43] established the framework of discriminative motif finding that optimizes the conditional likelihood by conjugate gradient ascent [44], as part of a regulatory network inference system that combines sequence and gene expression. The positive set is assumed to have exactly one occurrence per sequence (OOPS) [45]. The initial motif parameters for gradient ascent is derived from the exhaustive string search as in [46]. Sharan*et al.* [47] extended this framework to allow positive sequences containing no motif, while negative sequences are still not allowed to contain the motif. This model is named noisy OOPS (NOOPS) in [48].

DME (Discriminative Matrix Enumerator) [49] uses a different approach for discriminative learning of a PWM. A global, enumerative search on a discrete space of PWM. The objective function is the log likelihood ratio instead of the conditional likelihood. DME-X [50] generalizes DME by incorporating a weight for each sequence, so that there is no strict distinction between the positive and negative set. Multivariate regression is used to evaluate candidate motifs.

DIPS (Discriminative PWM Search) [51] proposed a different objective function, the difference between the average motif occurrences within the positive set and that within the negative set. This objective function is designed to find motifs with the largest number of occurrences in the positive set and smallest number of occurrences in the negative set, best suited for situations with multiple occurrences per sequence, also called two-component mixture (TCM) in MEME terminology. Optimization is achieved via heuristic hill-climbing.

Recently PWM is applied to protein motifs as well as DNA motifs in the DEME algorithm [48]. DEME uses the same criterion, the conditional likelihood, as Segal-Sharan does [43, 47]. It also employ a combination of global string search and conjugate gradient, but the string search is designed to be more sophisticated. Both OOPS and NOOPS are

supported. To be more effective on protein sequences DEME implemented the PAM120 substitution matrix which reflects our prior knowledge of amino acid similarities. However PWM does not allow insertion and deletion, making it less than optimal for protein sequence analysis. For a more detailed review of discriminative motif finding methods, see [48].

2.2 Identifying Targeting Motifs

Traditional motif finding algorithms are generative, only utilizing the positive set but not the negative set. For example, profile HMMs are widely used to model protein domains or motifs. The match, insert, and delete states of a profile HMM correspond to local alignment of a protein region to the motif profile. The match states represent conserved sites of a motif; the insert states represent insertions between two conserved sites in a protein; the delete states represent removal of a motif site in a protein. Another generative motif finding method, PWM, only models the conserved sites and does not model gaps in the local alignment.

These motif models are trained to optimize the maximum likelihood estimation (MLE) criterion. For this task, HMMs can be trained generatively with the Baum-Welch algorithm [37], and similarly PWMs are trained with the expectation-maximization (EM) algorithm, for example using MEME [36]. Note that for our purpose of finding motifs in proteins located in the same compartment, the models must be learned from unaligned sequences. These proteins do not belong to the same protein family and are too divergent for current multiple sequence alignment programs.

2.2.1 Discriminative training of HMM

Motif finding methods using generative training search for short sequences that are overrepresented in a given set compared to a background distribution. In contrast, discriminative motif finding methods use two or more sets and in each set they search for motifs that are overrepresented compared to the other sets. This allows for better motif models, especially for similar compartments or subcompartments. For simplicity we only use single-compartment proteins for discriminative motif finding.

Here we present a novel discriminative motif finder based on hidden Markov models. To train this model we use a discriminative criteria, maximum conditional likelihood (that is, the conditional likelihood of the correct class given the sequence). The conditional likelihood

is an established criteria in discriminative training methods, e.g. logistic regression. It has been shown in the speech recognition literature that the maximal mutual information estimate (MMIE) technique can train a set of HMMs to optimize this criteria. We use a MMIE algorithm termed extended Baum-Welch which iterates between aligning the motif sites and updating parameters based on the aligned sites. The update not only favors occurrence in the positive examples as in regular Baum-Welch, but also avoids occurrences in the negative examples. For simplicity we use the terms generative HMM and discriminative HMM for these two types of approaches below.

The MMIE algorithm is an extended version of the Baum-Welch algorithm [52]. Note that we do not know where the motif instances (with substitution, insertion and deletion) are without the motif parameters, but the motif parameters cannot be estimated without knowledge of where the motif instances are. As in the E-step of the Baum-Welch algorithm, we first infer the expected state of each position on each sequence based on current motif parameters (i.e. probabilities of each position being background, match, insert or delete state). This is equivalent to having a probabilistic alignment of the motif sites. Then in the M-step, we update the parameters to maximize the discriminative objective function based on the expected states above, or the probabilistic alignment. The E-step and M-step are repeated until the improvement upon objective function is too small.

The E-step in discriminative training is similar to that in Baum-Welch, using the forward and backward algorithm. The difference between generative and discriminative training is in the M-step, because the objective function to maximize is different. The update in M-step needs to increase occurrences of the motif in the positive examples and decrease occurrences in the negative examples. This is achieved by the following sequence weighting scheme based on the agreement between predictions and labels. Positive examples are weighted as the posterior probability of incorrect classification, $1 - p(\lambda^{(m)}|O_n)$, and negative examples are weighted as the negative of the probability of incorrect classification, $-p(\lambda^{(m)}|O_n)$. That is, a positive example is given a lower weight if its probability is high which is already correct, or given a higher weight otherwise. A negative example is given a smaller negative weight if its probability is low which is already correct, or a higher weight if it is incorrectly believed to be one of this class. In contrast, generative training weights positive examples as 1 and negative examples as 0 thus only focusing on occurrences in positive examples. Note that this interpretation is different from standard MMIE in speech recognition.

The Extended Baum-Welch Algorithm Using Sequence Weighting

We use the following notations. Let the training sequences be $\{O_1, O_2, \cdots O_N\}$, where N is the number of training examples. The sequences belong to M classes (for example, different branches in the tree of Figure 2.1) and the class labels of the sequences are given as $c_n \in \{1, 2, \cdots M\}, 1 \leq n \leq N$. The HMM for the m-th class is denoted as $\lambda^{(m)}$. The parameters for each HMM are denoted by $\lambda^{(m)} = (a_{ij}^{(m)}, b_{jk}^{(m)})$, where $a_{ij}^{(m)}$ and $b_{jk}^{(m)}$ are the transition and emission probabilities, respectively. The MMIE objective function, conditional likelihood of the correct class given the observed values, can be written as

$$\mathcal{F}_{\text{MMIE}} = \sum_n \log p(c_n|O_n) = \sum_n \log \frac{p(O_n|\lambda^{(c_n)})p(\lambda^{(c_n)})}{\sum_{m'} p(O_n|\lambda^{(m')})p(\lambda^{(m')})}$$

In the E-step, the expected count of state j at position t of sequence n according to model $\lambda^{(m)}$ is denoted as $\gamma_{nt}^{(m)}(j) = p(q_{nt} = j|O_n, \lambda^{(m)})$. The expected count of transition from state i to state j at position t of sequence n according to model $\lambda^{(m)}$ is denoted as $\xi_{nt}^{(m)}(i,j) = p(q_{nt} = i, q_{n,t+1} = j|O_n, \lambda^{(m)})$. These expected counts are calculated by the forward-backward algorithm. For simplicity we denote the expected count of transition and emission of the entire sequence n as $\xi_n^{(m)}(i,j)$ and $\phi_n^{(m)}(j,k)$, defined as

$$\begin{aligned} \xi_n^{(m)}(i,j) &= \sum_t \xi_{nt}^{(m)}(i,j) \\ \phi_n^{(m)}(j,k) &= \sum_t \gamma_{nt}^{(m)}(j) 1_{y_{nt}=k} \end{aligned}$$

We will show the update formulas for the M-step first and then the derivation. Positive examples are weighted as the posterior probability of incorrect classification, $1 - p(\lambda^{(m)}|O_n)$, and negative examples are weighted as the negative of the probability of incorrect classification, $-p(\lambda^{(m)}|O_n)$. After sequence weighting the probabilities are estimated similar to Baum-Welch, but a *smoothing constant* needs to be added to the probabilities of the same state before normalizing [52, 53]. The smoothing constants prevent negative probabilities due to negative sequence weights. The reestimation formulas in the M-step of MMIE are,

$$\hat{a}_{ij}^{(m)} \leftarrow \frac{\xi^{(m)}(i,j) - \xi^{(-m)}(i,j) + D_T a_{ij}^{(m)}}{\sum_{j'} \xi^{(m)}(i,j') - \xi^{(-m)}(i,j') + D_T a_{ij'}^{(m)}} \tag{2.1}$$

$$\hat{b}_{jk}^{(m)} \leftarrow \frac{\phi^{(m)}(j,k) - \phi^{(-m)}(j,k) + D_E b_{jk}^{(m)}}{\sum_{k'} \phi^{(m)}(j,k') - \phi^{(-m)}(j,k') + D_E b_{jk'}^{(m)}} \tag{2.2}$$

where $\xi^{(m)}(i,j)$, $\xi^{(-m)}(i,j)$, $\phi^{(m)}(j,k)$, $\phi^{(-m)}(j,k)$ are defined as follows for simplicity.

$$\xi^{(m)}(i,j) = \sum_{n|c_n=m} [1 - p(\lambda^{(m)}|O_n)]\xi_n^{(m)}(i,j), \quad \xi^{(-m)}(i,j) = \sum_{n|c_n \neq m} p(\lambda^{(m)}|O_n)\xi_n^{(m)}(i,j)$$

$$\phi^{(m)}(j,k) = \sum_{n|c_n=m} [1 - p(\lambda^{(m)}|O_n)]\varphi_n^{(m)}(j,k), \quad \phi^{(-m)}(j,k) = \sum_{n|c_n \neq m} p(\lambda^{(m)}|O_n)\varphi_n^{(m)}(j,k)$$

Following [54] we set the smoothing constants to twice the smallest value that ensures nonnegative transition and emission probabilities. This was found to lead to fast convergence empirically [54].

The MMIE literature does not use the sequence weighting perspective for the update formula, due to the large number of classes. Here we will show that the update formula originally developed for MMIE can be expressed as our sequence weighting forms. We will only derive the equation for transition probability since the derivations for emission probability is the same. The original update formula for MMIE is [52],

$$\hat{a}_{ij} \leftarrow \frac{a_{ij}^{(m)} \frac{\partial}{\partial a_{ij}^{(m)}} \mathcal{F}(\mathbf{\Lambda}) + D_T a_{ij}^{(m)}}{\sum_{j'} a_{ij'}^{(m)} \frac{\partial}{\partial a_{ij}^{(m)}} \mathcal{F}(\mathbf{\Lambda}) + D_T a_{ij'}^{(m)}} \tag{2.3}$$

The partial derivative of the objective function with respect to the transition probability a_{ij} can be calculated as follows.

$$\begin{aligned} & \frac{\partial}{\partial a_{ij}^{(m)}} \mathcal{F}(\mathbf{\Lambda}) \\ = \quad & \frac{\partial}{\partial a_{ij}^{(m)}} \sum_n \log \frac{p(O_n|\lambda^{(c_n)})p(\lambda^{(c_n)})}{\sum_{m'} p(O_n|\lambda^{(m')})p(\lambda^{(m')})} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial a_{ij}^{(m)}} \sum_{n|c_n=m} \log p(O_n|\lambda^{(m)})p(\lambda^{(m)}) - \frac{\partial}{\partial a_{ij}^{(m)}} \sum_n \log \sum_{m'} p(O_n|\lambda^{(m')})p(\lambda^{(m')}) \\
&= \sum_{n|c_n=m} \frac{\partial \log p(O_n|\lambda^{(m)})}{\partial a_{ij}^{(m)}} - \sum_n \frac{1}{\sum_{m'} p(O_n|\lambda^{(m')})p(\lambda^{(m')})} \frac{\partial p(O_n|\lambda^{(m)})p(\lambda^{(m)})}{\partial a_{ij}^{(m)}} \\
&= \sum_{n|c_n=m} \frac{\xi_n^{(m)}(i,j)}{a_{ij}^{(m)}} - \sum_n \frac{p(O_n|\lambda^{(m)})p(\lambda^{(m)})}{\sum_{m'} p(O_n|\lambda^{(m')})p(\lambda^{(m')})} \frac{\xi_n^{(m)}(i,j)}{a_{ij}^{(m)}} \\
&= \frac{1}{a_{ij}^{(m)}} \sum_{n|c_n=m} [1 - p(\lambda^{(m)}|O_n)]\xi_n^{(m)}(i,j) - \frac{1}{a_{ij}^{(m)}} \sum_{n|c_n \neq m} p(\lambda^{(m)}|O_n)\xi_n^{(m)}(i,j)
\end{aligned}
$$

Plugging in the above partial derivative to Equation 2.3 results in the sequence weighting update formula, Equation 2.1. Equation 2.2 can be derived in a similar way.

2.2.2 One Occurrences Per Sequence (OOPS) Model

For learning discriminative HMM, generative HMM, and MEME, we assume there is one occurrence of the motif in all sequences in the same compartment. Such a distribution is called one occurrence per sequence (OOPS) in MEME. Although a targeting motif may not appear on every sequence in a compartment, our analysis shows that distributions other than OOPS do not generate relevant motifs. We tried a method assuming zero or one occurrence per sequence (ZOOPS) but the results of these runs looked much poorer than the OOPS model when using MEME. For these runs ZOOPS usually found long subsequences shared among very few homologs (for example 3 proteins) which did not generalize well to the test data. Generative models do not have a constraint on the absence of motifs in the negative set, so without the requirement of OOPS (or, covering as many sequences as possible) there will be no mechanism to associate motifs extracted with localization. Another way to explain the improvement seen when using OOPS is that our method involves both the motif discovery and the use of SVM to classify proteins using the discovered motifs. Motifs that are too weak to provide any discriminative power will be of little use for the classifier and would thus not be used in the final outcome. However, the advantage of using weak motifs (which are usually found using OOPS) with SVM is that, while each one on their own may not be very informative a combination of weak motifs may still be very powerful. If such a combination exists the SVM method would identify it and use it to correctly classify proteins. In contrast, if these motifs were discarded at an earlier stage (in the motif discovery procedure) that would not have been available for the classifier to use.

2.2.3 Motif Finding Based on Predefined Pathways

An advantage of our discriminative training is that it can fully utilize the hierarchical structure that is used to sort proteins in the cell (see Figure 2.1). We now describe how this structure can be used for training and classification. For discriminative motif finding at a specific split, we find motifs for each branch discriminating only against proteins in the other branches of this split. Only proteins in compartments under the split are included as training examples. For example, starting from the root, we find 10 motifs for the secretory pathway and 10 motifs for all other (intra-cellular) compartments, discriminating between the two sets. Then for the splits under inter-cellular compartments, we find 10 motifs for nucleus and 10 motifs for the cytoplasm internal node, a union of proteins in cytosol, peroxisome and mitochondria. To compare these results to generative motif finding methods (MEME and generative HMM), we implemented a similar procedure for these methods as well. Training examples for the leaf node in the tree (the 9 compartments) are the same as in the flat structure. Training sets for the internal nodes are the union of descendant nodes, e.g. we search for 10 motifs for cytoplasm which is the union of proteins in cytosol, peroxisome, and mitochondria.

For the flat structure, all methods generate a total of 90 features (9 compartments each having 10 motifs, see Methods). For the hierarchical structure, all methods generate a total of 130 features (9 compartments, root and 3 internal nodes each having 10 motifs).

2.2.4 Selecting Motif Instances

After a PWM or HMM is learned, we would like to scan the sequences and only select the strongest matches as motif instances. That is, some sequences will have no instance of a motif while other sequences may have more than one instance. Hence for each candidate motif, we need to rank each possible position on all sequences.

For MEME, positions are ranked by the likelihood of the subsequence given the PWM. For HMM, the posterior probability given by posterior decoding is used. We consider two silent states of profile HMM, the begin state and the end state of the motif, instead of the first match state which may be skipped. For each position, we use the product of the posterior probabilities of its begin state and the nearest end state for ranking. All positions in all sequences are ranked by this product. According to the ranking we can retrieve the top positions when the number of instances is given (e.g. to retrieve the top 30 positions).

2.3 Predicting Localization

For evaluation of these three motif finders, we trained a support vector machine classifier (SVM). The feature set for the SVM are the motif scores. For MEME, the likelihood of the motif instance given the model is used as a feature. For generative and discriminative HMM, the log likelihood ratio of the entire sequence over the background model is used as a feature. The background model is the default one in HMMER. We carried out a 10-fold cross-validation procedure, so these three methods are trained on part of the dataset and tested on proteins not used to learn the motifs.

We use the SVM classifier in two different ways. The first is with a flat structure (one vs. all) and the second is with the hierarchical structure. For the hierarchical structure we train a separate SVM for each node in the tree and follow the classification results until we reach a leaf which denotes the final prediction for a compartment. For example, we first use our SVM to determine whether a protein is localized to the secretory pathway or to intra-cellular compartments. Based on that prediction, we use another SVM at each descending split (e.g. distinguishing between nucleus and cytoplasm). Some of the internal nodes have more than two descending splits (e.g. three compartments under cytoplasm), so each split is treated as an one vs. all classification. As in motif finding, only proteins in compartments under an internal node are included in SVM training. Accuracy calculation in hierarchical structure is the same as in flat structure. A prediction is considered correct only if it chooses the correct leaf node out of the 9 compartments; internal nodes are not counted toward accuracy.

The classifier enables us to rank the motifs found by MEME, generative and discriminative HMM. The ranking is based on the contribution for predicting locations. We rank the motifs by 1-step backward selection. For our SVM, the accuracy after removing each feature (corresponding to a motif) is recorded. The feature or motif that leads to the largest decrease in accuracy is selected as the top motif and the process is repeated until the desired number of motifs are selected.

The confidence of the prediction of each protein-compartment pair is also evaluated based on the classifier. For the flat structure or for each split on the tree structure, we convert SVM margins to conditional probability of observing the protein given a branch. For the tree structure, the confidence of a compartment given a protein sequence is the product of conditional probability on each split along the path from the root to the leaf

corresponding to the compartment.

2.4 Implementation and Details

We compare the results of our discriminative HMM model to generative training of HMMs and PWMs. Implementation of the generative and discriminative HMM are based on the HMMER 2.3.2 source code [37], and compared to motif finding using MEME 3.2.1 [36]. Unless explicitly mentioned below or in the main text, the default settings of HMMER 2.3.2 and MEME 3.2 are used. To make the comparison fair, we make similar assumptions and use the same options for discriminative HMM, generative HMM, and PWM learned by MEME: motif length is set to 4 (for HMM the number of match states is 4) and the one-occurrence-per-sequence (OOPS) model is assumed. For generative HMM, the Baum-Welch algorithm is executed with 100 random initialization and 100 iterations at most. HMMs are initialized by randomly selecting a 4-mer from the training sequences and setting all transition probabilities to the default values. The transition probabilities of background states before and after motif are defined according to the median protein length of our dataset, 530 amino acids. The emission probabilities of background states are the average amino acid composition of SwissProt 34 [55] as default. The background model is also the default one in HMMER, a HMM with a single state whose emission and transition probabilities the same as the background state in the profile HMM described above. For discriminative HMM, we first run the generative HMM algorithm using the same setting described above, then run the extended Baum-Welch algorithm on the one with highest likelihood with at most 100 iterations. For both generative and discriminative HMM, after a motif is found we mask the amino acids assigned to match states by Viterbi algorithm with random amino acids, and repeat the process until 10 motifs are found.

We classify the protein location by SVM using the motif scores as features. Training and testing of SVM are performed by the software SVMlight [56] with the linear kernel to avoid overfitting. The default options are used except that we tested three values for the trade-off between error and margin ($0.001, 0.01$ and the default value of SVMlight) and report the best cross-validation result.

	Output of classifier								
	Cyt (13.5)	ER (9.7)	Gol (0.6)	Vac (2.5)	Mit (30.2)	Nuc (32.2)	Per (0.0)	Mem (11.3)	Sec (0.0)
Cytosol (15.7)	25.4	2.0	0.0	0.7	26.8	43.8	0.0	1.1	0.2
ER (7.0)	14.2	30.7	1.3	0.6	22.5	26.3	0.0	4.4	0.0
Golgi (2.1)	8.3	16.5	3.2	2.4	23.7	38.9	0.0	6.8	0.0
Vacuole (2.5)	12.7	16.1	0.0	15.3	11.4	33.9	0.0	10.6	0.0
Mitochondria (25.8)	9.3	2.3	0.0	0.6	71.7	14.7	0.0	1.5	0.0
Nuclear (37.6)	12.9	0.6	0.0	0.6	19.5	64.6	0.0	1.5	0.2
Peroxisome (1.4)	5.0	0.0	0.0	0.0	70.0	25.0	0.0	0.0	0.0
Membrane (7.1)	8.5	7.0	0.8	2.3	7.0	29.0	0.0	45.5	0.0
Secreted (0.8)	25.0	12.0	0.0	0.0	19.0	13.3	0.0	30.7	0.0

Table 2.1: Confusion matrix of discriminative HMM using the tree compartment structure. Parenthesis after the columns are percentage of predictions (output) while parenthesis after the rows are percentage of labels (only single-compartment proteins counted as these are the training data).

2.5 Results

2.5.1 Prediction Accuracy

We applied our discriminative motif finding method to a yeast protein localization dataset [25]. This dataset consists of 1,521 *S. cerevisae* proteins with curated localization annotation in SwissProt [55]. Proteins were annotated with nine labels: nucleus, cytosol, peroxisome, mitochondria, endoplasmic reticulum (ER), Golgi apparatus, vacuole, plasma membrane, and secreted. We tested two different ways to search for motifs in discriminative training. The first uses a one vs. all approach by searching for motifs in each compartment while discriminating against motifs in all other compartments. The second uses a tree structure (Figure 2.1) to search for these motifs. The hierarchy of compartments utilizes the prior knowledge of cellular sorting by identifying refined sets of motifs that can discriminate compartments along the same targeting pathway. It has been shown previously that prediction accuracy can be improved by incorporating a hierarchical structure on subcellular compartments according to the protein sorting mechanism [24].

In addition to the two sets of motifs we find for discriminative HMMs, we find 10 motifs for each compartment using MEME and generative HMMs. For all methods the number of amino acid positions is set to four, although since HMMs allow for insertions and deletions the instances of motifs represented could be longer or shorter.

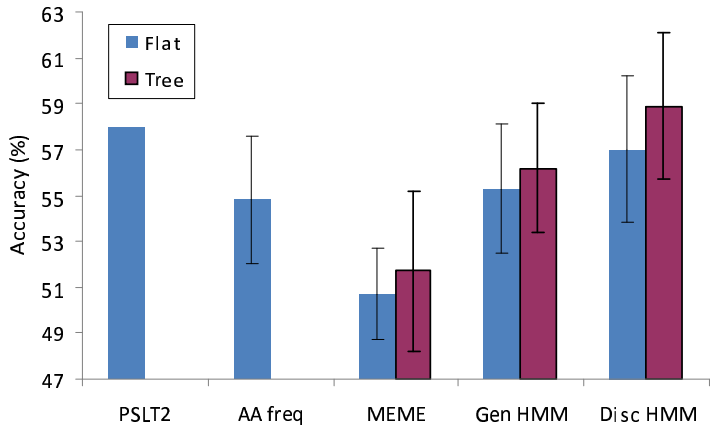


Figure 2.2: Accuracy of predictions based on known motifs for PSLT2, SVM using amino acid frequencies as features, and SVM using motifs discovered by MEME, generative and discriminative HMM. Results for the PSLT2 methods are taken from [25].

Because our goal is to identify novel targeting motifs and current understanding of targeting signals is still limited, we evaluate motif finding results by using them to predict localization as we describe above. We also compare the prediction accuracy of our method with that of a Bayesian network classifier that used curated motifs in InterPro [35]. The results for this prediction comparison are presented in Figure 2.2. As expected, the hierarchical structure, which provides another layer of biological information that is not available for the flat classification task, generally leads to improvement in classification results for all methods. When focusing only on generative training methods that do not utilize negative examples, profile HMMs outperformed MEME. This can be explained by the greater expressive power of the former model which allows for insertion and deletion events that cannot be modeled in MEME. Discriminative training that utilizes both this expressive set of options and positive and negative examples outperforms both other methods and its performance in the flat training setting is close to prediction based on known motifs. When using the hierarchical setting we can further improve the discriminative HMM results since internal nodes lead to more similar sets of motifs and discriminative training is most beneficial when the two groups are more similar to each other. For this setting discriminative HMMs achieve the most accurate classification results compared to all other methods we tested. Specifically, even though it does not use previous knowledge of motifs, discriminative HMMs improve upon results that were obtained using a list that included experimentally

validated motifs. The confusion matrix of the discriminative HMM is shown in Table 2.1. The coverage of compartments with fewer training sequences is low, e.g. proteins predicted as peroxisome and secreted are too few. This is most likely due to choosing the overall accuracy as the objective function to optimize.

We have applied the best classifier, discriminative HMM utilizing a hierarchical structure, to predict localization of all 6,782 proteins from SwissProt. The curated annotation of 1,521 proteins in the above dataset is used as training data. The predictions and the confidence are on the support website
(http://murphylab.web.cmu.edu/software/2009_TCBB_motif/).

Prediction Based on Amino Acid Composition

It is informative to compare classifications based on motif with those based on amino acid composition. We only utilize the amino acid composition of the whole sequence and not the N-terminal, C-terminal, or other more sophisticated compositions as in LOCtree [24]. We compared a number of SVM kernels for this data and concluded that a radial basis function (RBF) kernel works best. We set the gamma parameter of the RBF kernel to the default value of SVMlight. As shown in Figure 2.2, amino acid composition is as good as generative HMM and better than MEME, but accuracy is lower than discriminative HMM.

We have also used the classification result based on amino acid composition to evaluate whether our discriminative HMM method actually identified motifs, or was just utilizing the different AA decomposition of the proteins in each compartment. The predictions made by a SVM classifier based on discriminative HMM (using a tree structure) are compared to the predictions based on amino acid composition. 10-fold cross validation is used in both cases. Overall 27.1% of the proteins are only predicted correctly by our method and are assigned to wrong compartments by the amino acid composition classifier. A breakdown for each compartment is listed in Table 2.2. For peroxisome, vacuole, golgi, cytosol, and ER, most of the predictions require motifs and amino acid composition is not enough. For some compartments including nucleus, membrane and mitochondrion, there is a significant overlap between the two methods. This shows that the motifs identified (e.g. those in Figure 2.6 and 2.7 discussed below) are not just a different representation of amino acid frequencies but rather represent real sequence signature.

	Disc HMM recall	Disc HMM only not AA freq
Cytosol	53.0	43.7
ER	40.4	30.8
Golgi	13.0	11.7
Vacuole	23.3	20.9
Mitochondria	67.8	35.7
Nuclear	56.1	00.6
Peroxisome	04.2	04.2
Membrane	40.5	15.9
Secreted	00.0	00.0

Table 2.2: The first column is the percentage of proteins correctly predicted by our method in each compartment. The second column is the percentage of proteins correctly predicted by our method but not by a classifier based on amino acid composition. Discriminative HMM using the tree structure and amino acid composition using the flat structure are evaluated by 10 fold cross validation as described above as in Figure 2.2.

Prediction After Homology Reduction

It is important to examine how many homologous proteins are contained in this dataset, and how such redundancy affects the results. For this we have created a subset of proteins which contains no redundancy, and compare the classification performance of our method on this subset. This subset is filtered so that no pairs have more than 40% sequence identity, measured by BLASTALL 2.2.20. 98 proteins are filtered out, corresponding to only 6% of the original dataset. We performed the same procedures and parameters, and the cross validation accuracies are shown in Figure 2.3. The performance of the classifiers are robust against homology reduction compared to the results for the full dataset: amino acid composition and MEME have similar accuracy, generative HMM have slightly higher accuracy and discriminative HMM have slightly lower accuracy.

Precision-Recall Curves

We can obtain the precision and recall values of predicting one compartment at various threshold of confidence. Figure 2.4 shows the precision-recall curves of classification using SVM and three different motif finders, MEME, generative and discriminative HMM. Different methods performed better at different regions. For example, generative and discriminative HMM work well for mitochondria and ER, the later better on high precision

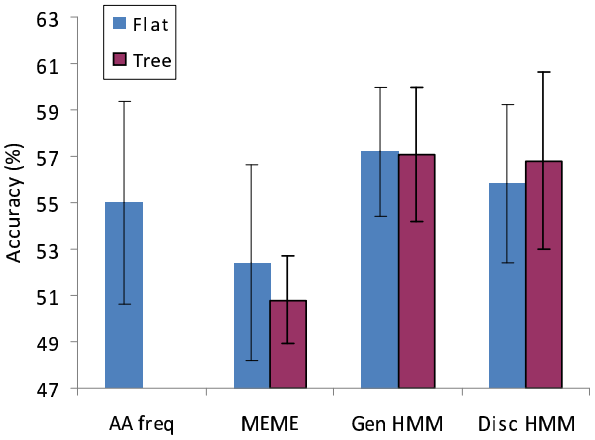


Figure 2.3: Cross validation accuracies of classification based on different methods using the redundancy-removed subset.

area. For peroxisome discriminative HMM is better than generative HMM which is better than MEME. In some areas, like high precision for membrane and secreted, MEME and generative HMM are better than discriminative HMM. Considering compartment sizes, overall discriminative HMM still outperforms other methods.

2.5.2 Recovering Known Motifs

After establishing the usefulness of our motif discovery algorithm for localization prediction we looked at the set of motifs discovered to determine how many of them were previously known.

Defining Known Targeting Motifs

There are a number of challenges we face when trying to compare the list of motifs identified by our methods with known motifs. Foremost is that evaluation of large sets of potential targeting motifs is hard when only a few targeting motifs are currently known. In addition, many of the motifs identified by our method are not directly involved in targeting proteins even if they are useful for subcellular classification. For example, DNA binding domains suggest that a protein would be localized to the nucleus though they are probably not the ones targeting it to that compartment. Thus restricting our comparison to classic motifs like ER retention signals may be misleading.

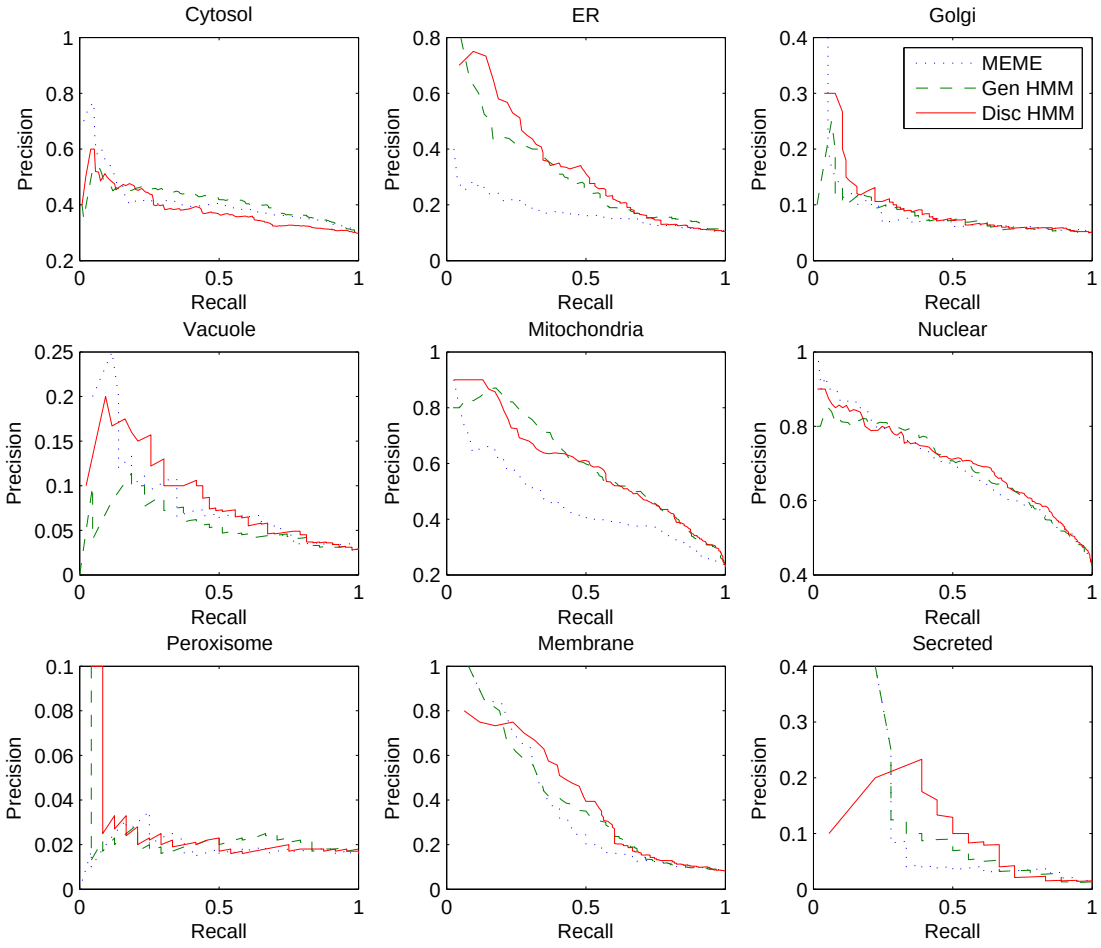


Figure 2.4: Comparing classifications using precision-recall curves of SVM whose features are motifs discovered by MEME, generative and discriminative HMM. Different thresholds are put on confidence derived from SVM margin.

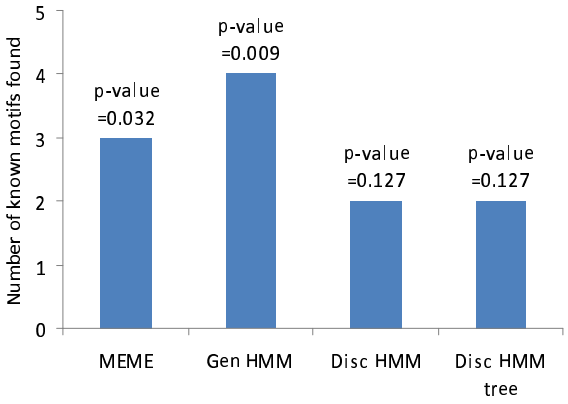


Figure 2.5: The number of known targeting motifs found by different methods and their significance. The p-values are calculated by generating random motifs.

To overcome these issues we collected a list of known targeting motifs from two databases, Minimotif Miner [57] and InterPro [35]. Minimotif Miner includes motifs that were experimentally validated to be involved in protein targeting. These motifs are represented as regular expressions. We also selected InterPro motifs that are associated with localization. To determine such association we perform a simple filtering step using the software InterProScan [58]. Any InterPro motif that occurs more than 4 times in one compartment and occurs in at most 3 compartments is considered associated with localization. Together we have a list of 56 known targeting motifs, 23 of them from MiniMotif Miner and 33 from InterPro.

Recovery Made by Different Methods

We ran MEME, generative and discriminative HMM on all sequences in our dataset to find 10 candidate motifs for each of the 9 compartments. The parameters of these methods are determined by cross-validation as described in the previous section. The candidate motif instances are matched against the known list derived from the Minimotif and InterPro scans. A known motif is considered to be recovered if one-third of its instances are correctly identified (overlapping at least half the motif length) when the number of predictions is 4 times the number of instances. For example, if a known motif has 12 instances, we retrieve the top 48 positions of each motif as described above and check if there are more than 4 overlaps.

Although directly comparing candidate motif models with known motifs has its advantages (e.g. not relying on a set of annotated sequence), it is difficult because each method outputs a different motif model. For example, MEME outputs a PWM while a HMM also allows for variable length insertions and deletions that cannot be accounted for in PWMs. We have thus decided to compare the different outputs by mapping their predictions back onto the proteins and comparing the proteins segments predicted to contain the motif with known motifs. This type of comparison has been used in the past [45,59]. Once the predictions are mapped to the proteins, determining whether the identified segment is a "hit" for a known motif also requires the determination of several parameters which we selected as above. We believe that these strike a good balance between specificity (overlap for at least half the motif) and sensitivity (a third of instances recovered). Note that the same criteria was applied to all methods so even if the criteria is not optimal the comparison is still valid and can be used to discuss the ability of each of the method to retrieve known instances.

The numbers of known motifs found are presented in Figure 2.5. Generative HMM was able to identify the most motifs followed by MEME. Although discriminative HMM works best for the classification task, it recovers less known motifs when compared to generative HMM and MEME. We provide possible explanations in the Discussion.

Significance of Known Motifs Recovered

To estimate statistical significance of recovering known motifs by MEME and HMMs, we generate 1000 sets each containing 90 random motifs as follows. Each motif is a randomly generated profile HMM. First a random 4-mer is generated assuming uniform distribution among the 20 amino acids. Then we construct a HMM and estimate the emission probabilities of the match states assuming this 4-mer is observed 10 times with a pseudocount of 1. Other emission and transition probabilities are set to default values of HMMER. After 90 such random HMMs are created, the same criteria for MEME and HMM motifs is used to count how many known motifs are recovered by these random HMMs. The p-value of recovering x known motifs is estimated as the number of motif sets that recovered x or more known motifs divided by 1000. For example generative HMM recovered 4 known motifs, and 9 motif sets out of 1000 recovered 4 or more known motifs, so the p-value is estimated as 0.009.

	Recovered?	Correct compartment (hits / total)	Other compartments (hits / total)
Microbodies targeting signal or PTS1 (SKL)	Yes	6 / 24	9 / 1497
Nuclear localization signal	Yes	191 / 647	119 / 874
Membrane C-terminal geranylgeranylation site	No	13 / 126	12 / 1395
ER retention signal (HDEL)	No	8 / 156	1 / 1365

Table 2.3: Distribution of known signals from MiniMotif Miner.

Distribution of Known Targeting Motifs

In order to understand why a certain known targeting motif is recovered while another is not, we analyzed the distribution of motifs in MiniMotif Miner [57] which include classical localization signals in the literature. Note that some well known targeting signals, like the signal peptide and the mitochondrial targeting sequence, are not in MiniMotif Miner due to lack of a clear consensus sequence. To our knowledge such signals rely one special programs like SignalP [28] and have not been represented as regular expression, PWM or HMM in previous knowledge-based localization prediction methods [31, 60]. Based on regular expression in MiniMotif Miner, there are four motifs that are significantly associated with localization on our yeast dataset, as listed in Table 2.3. Two of them are recovered by our method. We notice that not all well known localization signals are as discriminative as one would hope. Some signals like the ER retention signal are well conserved across species but can only explain a small portion of protein targeting in yeast.

Logos for Identified Motifs

The 20 most discriminative motifs and the known motifs found by discriminative HMM using flat and hierarchical compartment structure are shown in Figure 2.6 and 2.7 respectively. The most discriminative motifs are defined by backward feature selection as described in previous section. Motifs are visualized using HMM logos [61]. The nuclear localization signal motif is discovered by both methods. Discriminative HMM using flat structure finds the microbodies targeting signal, a motif known to be involved in peroxisome import [62]. Discriminative HMM using hierarchical structure finds the stress-induced protein motif (SRP1/TIP1), also known to be associated with the membrane in yeast [63]. Known motifs are sometimes ranked very highly, as SRP1/TIP1 above, but not always. This observation

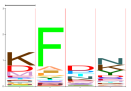
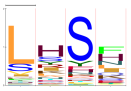
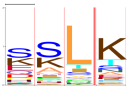
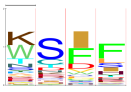
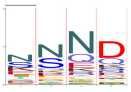
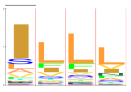
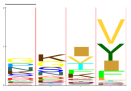
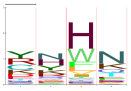
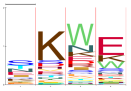
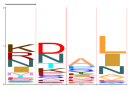
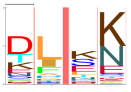
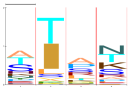
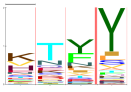
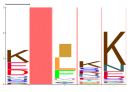
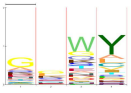
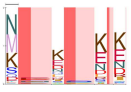
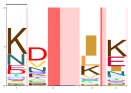
Rank	Compartment	HMM logo	Rank	Compartment	Known motif	HMM logo
1	Golgi		12	ER		
2	Cytosol		13	ER		
3	Vacuole		14	Golgi		
4	Golgi		15	Mitochondria		
5	Vacuole		16	Peroxisome		
6	Mitochondria		17	Cytosol		
7	ER		18	Golgi		
8	Mitochondria		19	Golgi		
9	Secreted		20	Mitochondria		
10	Secreted			Nuclear	Nuclear localization signal [KR]{4}	
11	Cytosol			Peroxisome	Microbodies targeting signal [STAGCN] [KRH] [LIVMAFY]$	

Figure 2.6: Top 20 motif candidates that are most predictive of localization, discovered by discriminative HMM using the flat compartment structure. Known motifs recovered by our methods are also shown with InterPro ID and regular expressions, which partially matches the HMM logo [61]. Pink columns are insert states of profile HMM; widths of dark and light pink columns correspond to the hitting probability and the expected length respectively (shortened when necessary to make the letters clear).

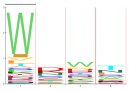
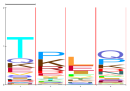
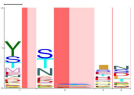
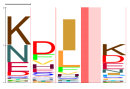
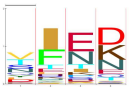
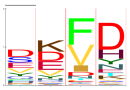
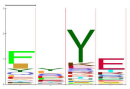
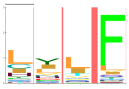
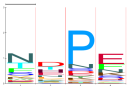
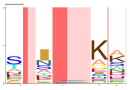
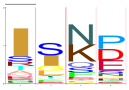
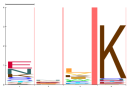
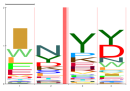
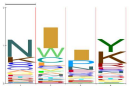
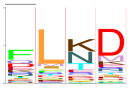
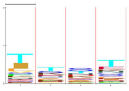
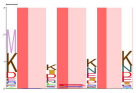
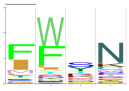
Rank	Compartment	Known motif	HMM logo	Rank	Compartment	Known motif	HMM logo
1	Secreted			12	Mitochondria		
2	Membrane	IPR000992 Stress-induced protein PWY[ST]{2}RL		13	Cytosol		
3	Cytosol			14	Cytosol		
4	Cytosol			15	Golgi		
5	ER			16	Peroxisome		
6	Cytosol			17	Membrane		
7	Cytosol			18	Cytosol		
8	Mitochondria			19	ER		
9	Peroxisome			20	Golgi		
10	Secreted				Nuclear	Nuclear localization signal [KR]{4}	
11	Golgi						

Figure 2.7: Top 20 motif candidates that are most predictive of localization and known motifs, discovered by discriminative HMM using the hierarchical compartment structure; detailed description in Figure 2.6.

suggests that there may be previously uncharacterized motifs that are highly associated with localization.

It is important to note that not all found motifs are necessarily involved in localization. Many may be involved in other functions that proteins in a given compartment need to carry out, or may reflect differences in amino acid composition between proteins localizing to different compartments. For example, the tryptophan motif for secreted proteins shown in Figure 2.7 presumably reflects a statistically higher frequency of that amino acid in secreted proteins than in other proteins but does not imply (or rule out) that that amino acid is important for the sorting process leading to secretion. Similarly, the "cytosolic retention signal" motif might not have any retention role but could simply be a motif associated with binding of cytosolic proteins to structures such as the cytoskeleton.

The motif found that matches to known NLS is presumably that of a single basic cluster corresponding to one half of a bipartite NLS. As such, non-basic amino acids in the conserved basic positions is perhaps surprising. However, it is possible that NLS still functions with the presence of non-basic amino acids either to the left or right of two or more basic amino acids. Since the HMM logos cannot capture correlation between positions (and the HMM only capture first order dependence), these motifs might match with some sequences that are unlikely to function as an NLS. It should however match well with many valid NLS. In other words, we might expect the motif in the form shown in Figure 2.6 and 2.7 to have some false positives but high recall of valid NLS.

2.5.3 Motif Conservation

Since at least some of the discovered motifs may play an as yet unidentified role in localization, we sought other ways of validating them as potential sorting signals. One approach was based on analysis of motif conservation: we expect motifs targeting proteins to their subcellular location to be more conserved among evolutionarily close species [64].

Protein Homolog Alignment

To evaluate the conservation of the motifs identified by each of the methods we used Saccharomyces Genome Database (SGD) fungal alignments for 7 yeast species [16]. The default alignment result is used. Sequence and homology information were derived from integration of two previous comparative genomics studies [65, 66]. For these species amino acid sequence alignment was performed by ClustalW, and four conservation states were defined

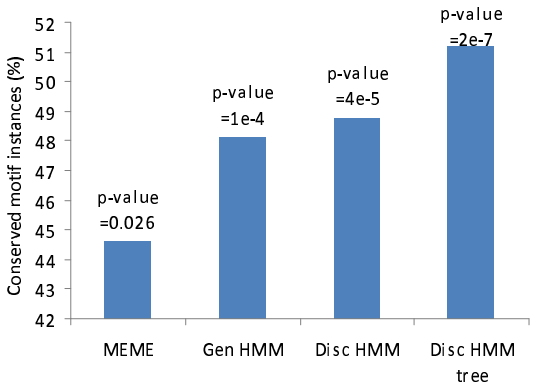


Figure 2.8: Percentage of conserved motif instances of the top 20 candidate motifs found by different methods. Conservation is based on SGD fungal alignment. A motif instance is considered conserved if all sites are strongly conserved. The p-values are denoted for each method (see Methods for the statistical test).

for each amino acid: no conservation versus weak, strong and identical conservation (across 7 species).

Measure of Conservation

The analysis below is based on the 20 most discriminative motif candidates, defined by backward feature selection as described previously. For each of the 20 motifs, we retrieve the top 30 positions based on likelihood or posterior probability. Then for each motif instance, it is considered conserved if all sites are labeled as having strong or identical conservation by ClustalW.

Significance of Motif Conservation

The statistical significance of motif conservation is calculated as follows. We scan through all proteins in our dataset using a sliding window of 4 amino acids (the motif length we used) to obtain the number of conserved 4-mer and total possible 4-mers. For each motif finding method, we have the number of conserved motif instances and the total number of top motif instances. With these counts we use a hypergeometric test to calculate a p-value for each method.

Conservation of Motifs Found by Different Methods

The percentage of conserved motif instances for MEME, generative and discriminative HMM (flat or hierarchical structure) as well as the significance for each of these methods are presented in Figure 2.8. The conservation analysis clearly indicates that motif instances discovered by all methods are significantly conserved when compared to random protein regions. Using a sliding window of the same length as the motifs, we find that only 41% of 4-mers are conserved. In contrast, for motifs identified by discriminative HMM using flat or hierarchical structure, 49% and 51% of motif instances are conserved respectively. For generative HMM 48% of motif instances are conserved and for MEME 45% instances are conserved. The conservation achieved by discriminative HMM using hierarchical structure is the highest among the methods we looked at.

Conservation After Randomizing Annotations

To further evaluate the significance of the motif conservation, we tested the conservation analysis on motifs extracted from training data with randomized compartment annotations. The fraction of each compartment (estimated from single-compartment proteins) is kept. We then perform motif discovery and conservation analysis on the randomized annotations. In Figure 2.9 we can see that the conservation after randomizing annotations is much lower than that using the correct annotations, except conservation of motifs found by MEME is similar to the original one which is not significant. Note that although the annotation is random, the motif finders may still extract overrepresented motifs related to other functions (not random 4-mers in hypergeometric test) and display some conservation that is stronger than background.

2.5.4 Reannotating Protein Localization

The motifs discovered by our method successfully predict the subcellular localization of close to 60% of all proteins. Still, we were interested in looking more closely at the other 40% for which we do not obtain the expected result. Several other factors can effect localization and our method clearly does not discover all targeting motifs. Still, we hypothesized that at least some of these mistakes can be explained by incorrect annotation in the SwissProt database.

To test this we have used the entire dataset as training set for both motif finding

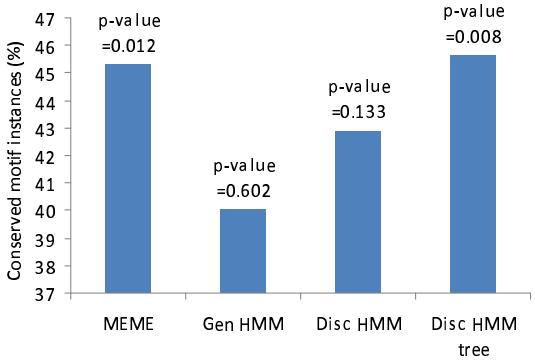


Figure 2.9: Motif conservation after randomizing annotation of the training data. The setting is the same as the conservation analysis in the main text using the correct annotations; percentage of conserved motif instances of the top 20 candidate motifs found by each method is shown. See methods in the main text for p-value calculation.

and the SVM classifier. Next, we examined more closely those proteins for which none of the motif-based methods (PSLT2, MEME, generative and discriminative HMM using hierarchical structure) agrees with the annotation in the SwissProt database. There are 42 such proteins out of 1,521 entries in the dataset we worked with. We have found at least 8 proteins for which there is strong reason to believe that the annotations in SwissProt are incomplete, discussed below.

Ski3/YPR189W

The protein superkiller 3 (Ski3), which is involved in mRNA degradation, was annotated as nuclear in the previous version of SwissProt used to create our annotated protein set. However, all motif-based classifiers (including MEME and HMM) predicted cytosol. The latest version of SwissProt, as well as SGD, lists it as localizing to both the nucleus and the Ski complex (in the cytoplasm). This illustrates that the motif-based classifiers can potentially complement protein databases and image-based annotations.

Frq1/YDR373W

The N-myristoylated calcium-binding protein, Frq1, is annotated as bud neck in SwissProt but manually curated as Golgi membrane on SGD, in agreement with the MEME prediction. The GFP image in the UCSF database is consistent with Golgi localization (Figure 2.10A).

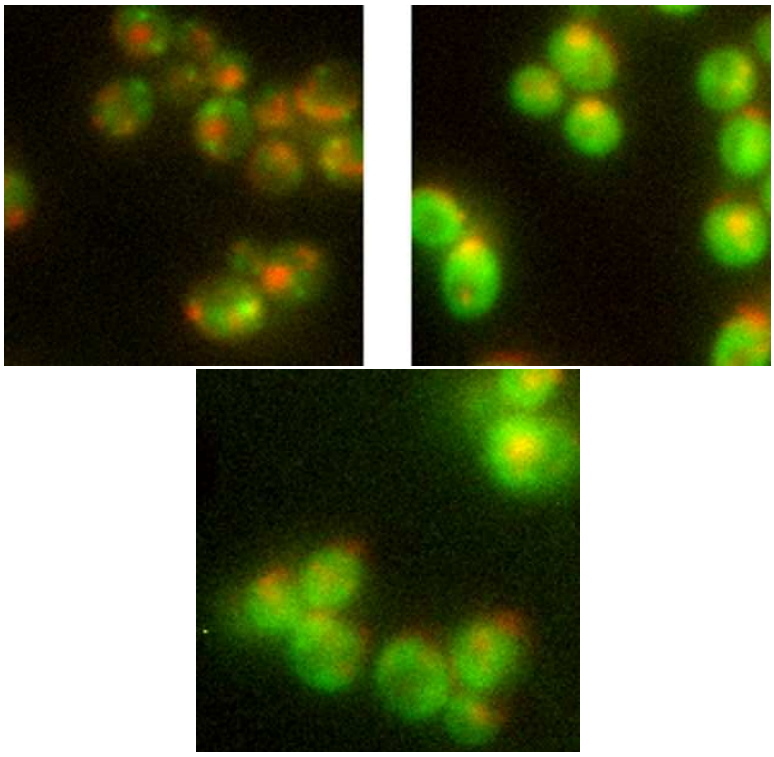


Figure 2.10: Fluorescence microscope images for some of the proteins whose subcellular location predicted from sequence differs from annotations in SwissProt. Each image shows the DNA-binding dye DAPI (red) and the GFP-tagged proteins (green). The proteins are Frq1/YDR373W (upper left), Ppt1/YGR123C (upper right), and Gsg1/YDR108W (lower). Images were obtained from the UCSF GFP-localization database (http://yeastgfp.ucsf.edu/).

Ppt1/YGR123C

Ppt1, or protein phosphatase T, is curated as present in both the cytoplasm and the nucleus on SGD. Cytoplasm is predicted by PSLT2 and MEME even though SwissProt only lists nucleus. The GFP tagged protein shows cytoplasmic localization (Figure 2.10B).

Vac8/YEL013W

Vac8 is labeled as vacuole by human experts, but is also involved in nucleus-vacuole (NV) junctions [67]. This could be the reason that an image-based automated classifier [15] and all motif finders agree on Vac8 being localized to the nucleus.

Pom152/YMR129W

According to SwissProt, the protein Pom152 is a component of the nuclear pore complex and is localized to the nucleus, but according to SGD it is localized to both nucleus (curated) and mitochondria (highthroughput). The GFP image on the UCSF database actually shows a cytoplasmic (non-nuclear) pattern, and an automated image-based classifier [15] predicted vacuole. The GFP evidences agrees with all motif based classifiers that predict Pom152 to be localized to the cytosol, although it is quite possible that the protein is mis-localized due to the GFP tagging. The results suggest that motif-based methods are helpful in identifying proteins at the boundary between two compartments.

Axl1/YPR122W

Axl1, a protein involved in axial budding, is labeled as bud neck or membrane on both SwissProt and SGD. GFP tagging also suggests cytosol, as predicted by the HMM motif finder.

Gsg1/YDR108W

Gsg1 is labeled as Golgi on SwissProt and SGD but the GFP image (Figure 2.10C) can also be interpreted as cytosol, agreeing with MEME and generative HMM.

Frt1/YOR324C

Frt1 is labeled as ER on SwissProt and SGD but predictions based on PSLT2 and generative HMM are mitochondria and cytosol respectively, which are possible based on the GFP

image.

2.6 Discussion

We have developed and used a new method that relies on discriminative HMMs to search for protein targeting motifs. We used our method to identify new motifs that control subcellular localization of proteins. Our method led to improvement over other methods when predicting localization using these motifs. While many of the motifs identified by our method were not known before, they are more conserved than average amino acids in protein coding regions indicating their importance for proper functioning of the proteins. We have also used our method to identify proteins that we believe are missannotated in public datasets. Some of the predicted annotations are supported by imaging data as well.

Our discriminative HMM can be considered as an extension over the maximum discrimination training of HMM suggested by Eddy et al [68]. The criterion used by both methods, conditional likelihood of the class given the data, is the same. However the maximum discrimination method proposed by Eddy et al only uses positive examples discriminating against background data. Thus, it cannot utilize negative examples as our method does.

When compared to known motifs, the set of motifs identified by discriminative HMM contains less known motifs than generative HMM and MEME, even though they lead to the highest prediction accuracy. One way to explain this result is the relatively small number of known targeting motifs. Thus, it could be that there are still many strong targeting motifs that are unknown and discriminative HMM was able to identify some of these. In addition, most known motifs are represented as consecutive peptides without insertion or deletion, hence they follow more closely the MEME model. It is worth noting that the results in this chapter are achieved without incorporating information on position relative to sequence landmarks like the N- or C-terminus or cleavage sites (like most motif finders). Thus it does not find elements, such as the signal peptide, that can be found using such alignments [69]. We will propose a solution in the next chapter.

Chapter 3

Inferring Targeting Pathways

In Chapter 2 we showed how to learn motifs and predict locations based on a tree representing targeting pathways. However this approach is too simplified to model the actual protein sorting mechanism. The tree we used is only a selected subset of the known pathways, and there may be more targeting pathways unknown to us. Hence we would like to model protein targeting using a more general structure and to discover new targeting pathways.

To perform their function(s), protein usually need to be localized to the specific compartment(s) in which they operate. Subcellular localization of proteins is typically achieved by sorting pathways involving carrier proteins. Disruption of these pathways leading to inaccurate localization plays an important role in several diseases, including cancer [3, 4, 8], Alzheimer's disease [5], hyperoxaluria [6] and cystic fibrosis [7]. Thus, an important problem in systems biology is to determine how proteins are localized to their target compartments, the carriers and motifs that govern this localization and the pathways that are being used.

While the above experimental methods provide some information on sorting pathways, no method exists to try and infer global sorting pathways from current localization information. In this chapter, we show that by integrating sequence, motif and protein interaction data we can develop global models for the process in which proteins are localized to subcellular compartments. We use a hidden Markov model (HMM) to represent sorting pathways. Carrier proteins and motifs are used to define internal states in this model and the compartments serve as the final (goal) state. Using this model we identified several sorting pathways, the carrier proteins that govern them and the proteins that are being sorted according to these pathways. Simulation data indicates that the models learned are accurate

(leading to 81% prediction accuracy with a noise level of 5%, see Figure 3.4). Using data from yeast we show that our model leads to accurate classification of protein compartments while at the same time enabling us to recover many known pathways and the proteins that govern these pathways. Several new predictions are provided by the model representing new putative sorting pathways.

3.1 Related Work

Recent advances in fluorescent microscopy coupled with automated image-based analysis methods provide rich information about the compartments to which proteins are localized in yeast [1, 15] and human [13, 14, 21]. Several computational methods have been developed to predict subcellular localization by integrating sequence data with other types of high throughput data [22–25, 33, 34, 70, 71]. These methods either treat the problem as a one vs. all classification problem [22, 23, 70, 71] or utilize a tree that corresponds to the current knowledge regarding intermediate compartments, for example LOCtree [24], BaCelLo [72] and discriminative HMMs [73]. The tree based methods were shown to be superior to the one vs. all methods; however, these methods do not attempt to learn the sorting pathways, relying instead on current (partial) knowledge of protein sorting mechanism.

A number of methods have learned decision trees for predicting subcellular localization. These include PSLT2 [25] which refines the location into sub-compartments using a decision tree learned from data and YimLOC [27] which learns a decision tree for the mitochondrion compartment only using features that include predictions from SherLoc [74], an abstract-based localization classifier. While the decision trees generated by these methods are often quite accurate, they are not intended to reflect sorting pathways, and they utilize features that, while useful for classification, are not related to the biochemical process of protein sorting.

In contrast to the global localization prediction methods, several experimental researchers have focused on trying to assign a specific sorting pathway to a small number of proteins. For example, proteins containing a signal peptide are exported through the secretory pathway [20], while some proteins without a classical N-terminal signal peptide are found to be exported via the non-classical secretory pathway [75]. A number of computational methods were developed to use this information to predict, for a given pathway, whether a protein goes through that pathway or not based on its sequence (for example, SignalP [28] and

SecretomeP [29]). However, these methods rely on the pathway as an input and cannot be used to infer new pathways.

There are many methods developed for reconstruction of pathways of other types, for example for signaling pathways [76–78] and metabolic pathways [79–81]. These pathways are used to describe information flow: one protein senses the environments and by activating a signaling or regulatory pathway passes that information along so that the cells can mount a response. We focused on a completely different meaning of pathway: physical movement of a specific protein. When referring to sorting pathways we mean that a single protein is being carried from one location to another. Unlike information flow pathways, which involve different molecules along the way, physical sorting pathways always involve the same proteins interacting with a set of different proteins. This makes it much more complicated to infer the order in which this is performed (since it is always the same protein). In addition, the outcome of an information flow pathway is often a change in genes expression which can be readily measured using microarrays. In contrast, the outcome of a sorting pathway is the localization of a single (or a few) proteins to a compartment. Again, this requires different methods for inference. We are not aware of any prior paper discussing computational methods for large scale inference of pathways describing physical movement of a protein.

3.2 Input Data

Our input data is composed of the localization of all proteins, their interactions and their sequences. Each protein is labeled with one or more locations. Generative HMM search for motifs present in one compartment and discriminative HMM search for motifs present in one compartment but absent in other compartments. We also collected all interacting partners of the protein and the occurrences of a set of known motifs from public databases (denoted as deterministic motifs to distinguish from novel motifs extracted from sequence described below), specifically InterPro [35] domains and and three signal sequence feature from UniProt [18]: signal peptides, transmembrane region, and GPI anchor (more detail in section 3.5.2). We perform feature selection by a hypergeometric test to identify features with a significant association with a location before learning our model.

We extract novel motifs associated with a location using the generative and discriminative HMM motif finder we have previously described [73]. We will compare two approaches

to convert each sequence to motif features: sequence likelihood and binary occurrence. The first approach use the sequence likelihood given the motif as feature, $\Pr(S|\lambda_k)$ where λ_k is the profile HMM of the motif (see next section). It represents how strong the instance matches the motif. Note that what really matters is the likelihood ratio of motif versus background, as described below. The second approach use a binary value to represent whether a motif occurs in a sequence instead of a real value. Binary motif occurrence are determined by posterior decoding as described in the previous chapter (also in the paper [73]).

3.3 Modeling Sorting Pathway by Hidden Markov Models

We used a HMM to model the process of sorting proteins to their compartments, determined by the interactions and sequence motifs. HMM is a generative model and thus provides the set of events that lead to the observed localization of the proteins (see Figure 3.1). An allowed pathway through the HMM state space structure represents a possible protein sorting pathway. All proteins start at the same start state, representing their translation in the cytoplasm. (While those few proteins that are translated in mitochondria would not begin in the cytoplasm, there were no mitochondrially-encoded proteins in our datasets and we can ignore this possibility.) The assigned (final) compartment of a protein is represented by a state in the model that does not have any outgoing transitions. Intermediate states correspond to intermediate compartments or to sorting events (for example, interaction with a carrier protein). These internal states emit observed features that are related to the sorting events, namely motifs (implying that the targeted protein uses that motif to direct it to that state) and carrier proteins that target proteins to the state. The emitted features of a protein are observed and determine its path in the state space. Emission is probabilistic and so certain proteins can pass through states even if they do not contain any of the motifs and do not interact with any of the carriers for that state. Note that while the compartment information is available during training, we do not know how many intermediate states should be included in the model (some sorting pathways may be short and others long, and several compartments can share parts of the pathways). Thus, unlike traditional HMM learning tasks that focus on learning the transition and emission probabilities, for our model we also need to learn the set of states that are used in the sorting HMM.

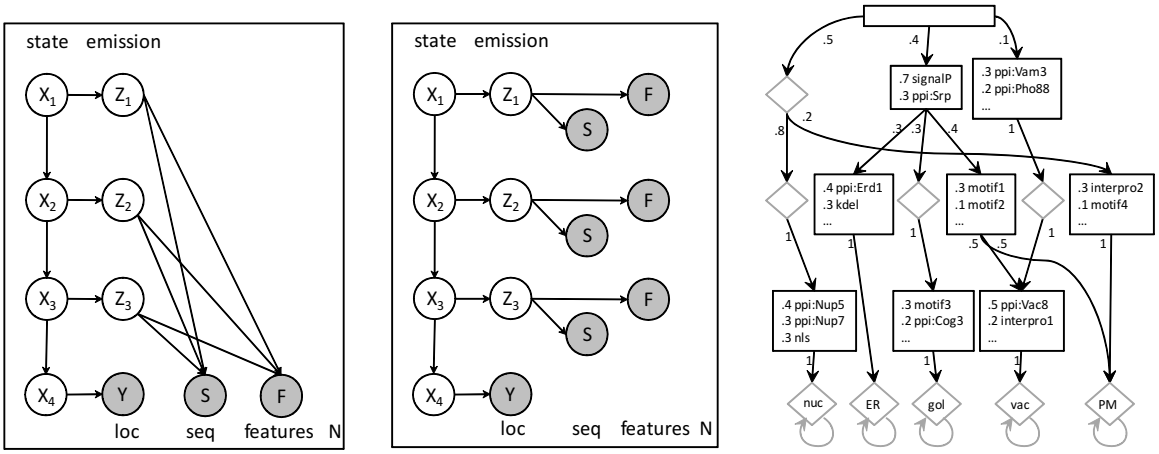


Figure 3.1: (**A**) The graphical model representation of a sample HMM for sorting pathways. Variables $X_1 \cdots X_4$ are unobserved intermediate sorting states at each level or each step. $Z_1 \cdots Z_3$ are the emission responsible for protein sorting at each step. S is the sequence and F corresponds to the binary feature observations. (**B**) The simplified HMM that maintains conditional independence between steps. (**C**) A sample state space: The top block is the root and its outgoing arrows correspond to initial probabilities. Bottom nodes are compartment states. The blocks are states and the arrows are transitions, with transition probabilities labeled. The items listed inside a blocks are top features emitted by the states, and emission probabilities are given on the left. Diamond-shaped blocks are silent states that emit the background feature only.

3.3.1 A HMM for the Sorting Pathways Problem

We will discuss the likelihood of our HMM in detail here (see Figure 3.1). The following description applies to using likelihood for motif features, but can be easily adapted to the case of binary motif features by removing the sequence variable S and include motif occurrences in the binary feature variables F (see below). As discussed above, in our HMM model all proteins move from a single start state to their final compartment. For reasons that will become clear when talking about learning the parameters of the model, we associate each state in our model with a specific level. The root state is level 0, all compartment states are associated with the final level (T) and each intermediate state is associated with a specific level t ($0 < t < T$). The number of levels T is inferred from the data during structure initialization as described in section 3.4. We require that a state at level t can be reached from the root after exactly t transitions; connections that are more than one level apart move through several "silent" states so that transitions are only between adjacent levels (diamond-shaped states in Figure 3.1). Silent states only emit a "background" feature (probabilities of the background feature are discussed later). Let X_t denote a hidden state at level t, $t = 1, 2, \cdots, T$ in a T-level model. The value of X_t can be one of J possible states, $X_t \in \{1, 2, \cdots, J\}$.

In addition to transition probabilities states are associated with emission probabilities. State X_t emits a feature index Z_t. Z_t can either be one of M motifs (represented as a likelihood score for each protein), or one of K binary features which include interactions with selected carriers, selected deterministic motif occurrences based on UniProt, or the background feature emitted by silent states. Hence $Z_t \in \{1, 2, \cdots M + K + 1\}$, where the motifs are indexed from 1 to M and the features are indexed from $M + 1$ to $M + K$.

Let S denote the sequence observed for each protein, F be the binary features from interaction databases and UniProt, and Y be the compartment assignments for a protein. The data likelihood of our HMM model (Figure 3.1), is defined as:

$$\Pr(S, F, Y|\Theta) = \sum_{X_1} \cdots \sum_{X_T} \sum_{Z_1} \cdots \sum_{Z_{T-1}} \Pr(S, F, Y, X_1, \cdots X_T, Z_1, \cdots Z_{T-1}|\Theta)$$

These joint probabilities can be decomposed based on the HMM independence assumptions as follows:

$$\Pr(S, F, Y, X_1, \cdots X_T, Z_1, \cdots Z_{T-1}|\Theta)$$

$$= \Pr(X_1)\prod_{t=1}^{T-1}\Pr(X_{t+1}|X_t)\Pr(Z_t|X_t)\Pr(S|Z_1,\cdots Z_{T-1})\Pr(F|Z_1,\cdots Z_{T-1})\Pr(Y|Z_t) \quad (3.1)$$

The parameters of our HMM are the initial, transition and emission probabilities, $\Theta = (\pi, A, B)$, defined as

$$\pi_i = \Pr(X_1 = i),\ A_{ij} = \Pr(X_{t+1} = j|X_t = i),\ B_{ik} = \Pr(Z_t = k|X_t = i).$$

where π_i is the initial probability of transition from the root to state i, A_{ij} is the transition probability between state i and state j, and B_{ik} is the emission probabilities from state i to emission k. Since each state only transits to a small number of states and emits a small number of features, these matrices are sparse.

3.3.2 Defining the Emission and Transition Probabilities for Our Model

As indicated above the feature observation includes the sequences and interactions selected carriers inferred by feature selection described above. Note that these observations are static and so may depend on all levels in the HMM. The emission probability for the sequence S is thus $\Pr(S|Z_1,\cdots Z_{T-1})$. Since probability depends on several motif models (one per level), which may be dependent (for example for overlapping motifs) and is thus computationally intractable given many combinations of motifs. As is commonly done [51] we approximate this term by the product of the conditional probabilities of the sequence given an individual emission at each level: $\prod_{t=1}^{T-1}\Pr(S|Z_t)$. Similarly we calculate the conditional probability of the binary features $\Pr(F|Z_1,\cdots Z_{T-1})$ using the product of the conditional probabilities of individual emissions (unlike for the sequence data this computation is exact since they are provided as independent events): $\prod_{t=1}^{T-1}\Pr(F|Z_t)$. This leads to the more typical HMM model shown in Figure 3.1B.

To translate the sequence information to a probability we use the likelihood of the sequence given the motif, $\Pr(S|\lambda_k)$, where λ_k is the motif mode. We use a profile HMM model but any other probabilistic models would also work, for example a position weight matrix (PWM) which specifies a weight for each amino acid at each motif position, assuming independence between positions. This likelihood is termed the motif score, and indicates how well the sequence agrees with the motif model. For states emitting one of the binary features or the background feature, the likelihood of the sequence is $\Pr(S|\lambda_0)$, where λ_0 is the background model for which we use a 0th-order Markov model, which assumes that each

position in the sequence are generated independently according to amino acid frequencies. Combined, the sequence likelihood is given by

$$\Pr(S|Z_t = k) = \begin{cases} \Pr(S|\lambda_k) & \text{if } 1 \leq k \leq M \\ \Pr(S|\lambda_0) & \text{if } M+1 \leq k \leq M+K+1 \end{cases} \tag{3.2}$$

The binary features observations, $F = (F_1, F_2, \cdots, F_K), F_k \in \{0, 1\}$ correspond to observed protein interactions and deterministic motifs as discussed above. As mentioned above we assume independence in noisy observation of these features, which is a necessary simplification. This lead to

$$\Pr(F|Z_t = k) = \prod_{j=1}^{K} \Pr(F_j|Z_t = k)$$

The conditional probability of observing a feature F_j given an emission Z_t is

$$\Pr(F_j = 1|Z_t = k) = \begin{cases} \nu_j & \text{if } k \neq M+j \\ \nu_0 & \text{if } k = M+j \end{cases}, \ 1 \leq j \leq K \tag{3.3}$$

where ν_j is probability of observing this interaction across all proteins in our dataset (background distribution) and $1-\nu_0$ is the probability of false negatives, .i.e. proteins that should go through this state but do not have this interaction / motif. Note that we need to use ν_j since an interaction or a motif may be observed even if the corresponding feature is not emitted by one of the states since many interactions are not related to protein sorting but rather to another pathway in which this protein is a member.

The conditional probability of the compartment given the final state is denoted by: $\Pr(Y|X_T)$. If a single compartment is given for a protein, the bottom state X_T is known for that protein and so this probability is 1 for that compartment and 0 for others. If the training data contains multiple compartments for a protein, it is reflected by the given compartment likelihood $\Pr(Y = y|X_T = c)$, which is assumed to be uniform for all compartments listed for that protein. In other words we consider multiple localization as uncertainty. For example, a protein might be considered to be 50% certain as one compartment and 50% certain as another compartment.

3.3.3 Approximation and Feature Levels

Unlike a typical HMM learning problem, the emission data we observe (sequence and interaction data) is static and so cannot be directly associated with any sequence of events. In addition, since our features are static, they can be emitted multiple times along the *same* path. However, if this happens the independence assumptions of HMMs are violated. Specifically, if a feature is emitted by a state in level t and then again by a state in level $t+1$ then it is not true anymore that the probability of emitting the feature given the state is independent of any emission events in previous states (since, if it was emitted before the protein can still emit it again). We thus constrain all features in our model so that each is only associated with a specific level and can only be emitted by states on that level. The level is determined in the initial structure estimation step discussed in the next section. Since no transitions are allowed between states on the same level no feature can thus be emitted more than once along the path and so the independence assumption holds. This requirement guarantees that the likelihood function obtained from the model presented in Figure 3.1B is a constant factor approximation of the likelihood function of our original model (Figure 3.1A).

Here we will describe how to approximate the full model in Figure 3.1A by the simplified model in Figure 3.1B, given that each feature has a fixed level. Recall that the joint probabilities of the original model in Figure 3.1A is given in Equation (3.1). First we focus on the emission probabilities of the feature observations, and show that the likelihood ratio of the emission versus the background equals the product of this likelihood ratio on all levels.

$$\frac{\Pr(F_j = 1|Z_1, \cdots Z_{T-1})}{\nu_j} = \prod_{t=1}^{T-1} \frac{\Pr(F_j = 1|Z_t)}{\nu_j} \tag{3.4}$$

where ν_j is the likelihood given the background feature. From Equation (3.4) we can naturally obtain

$$\Pr(F_j = 1|Z_1, \cdots Z_{T-1}) = \nu_j^{2-T} \prod_{t=1}^{T-1} \Pr(F_j = 1|Z_t)$$

for each feature, and it is combined as

$$\Pr(F|Z_1, Z_2, \cdots Z_{T-1}) = (\prod_j \nu_j^{2-T}) \prod_{t=1}^{T-1} \Pr(F|Z_t) \tag{3.5}$$

The full emission probability for each feature, $\Pr(F_j|Z_1, Z_2, \cdots Z_{T-1})$, is defined as a noisy observation (with false positive and false negative) of the OR function over Z_t,

$$\Pr(F_j = 1|Z_1 = k_1, Z_2 = k_2, \cdots Z_{T-1} = k_{T-1}) = \begin{cases} \nu_j & \text{if } \forall t \ k_t \neq M + j \\ \nu_0 & \text{if } \exists t \ k_t = M + j \end{cases}$$

However the OR function is unnecessary because we require feature F_j to have a fixed level, so only one level can emit the corresponding emission such that $Z_t = k_t = M + j$. Now to prove Equation (3.4), when one of the levels indeed emit the corresponding emission, we start from the right hand side of Equation (3.4) and apply Equation (3.3),

$$\prod_{t=1}^{T-1} \frac{\Pr(F_j = 1|Z_t)}{\nu_j} = \frac{\nu_0 \nu_j^{T-2}}{\nu_j^{T-1}} = \frac{\nu_0}{\nu_j} = \frac{\Pr(F_j = 1|Z_1, \cdots Z_{T-1})}{\nu_j}$$

and reach the left hand side of Equation (3.4). Similarly when none of the levels emit the corresponding emission,

$$\prod_{t=1}^{T-1} \frac{\Pr(F_j = 1|Z_t)}{\nu_j} = \frac{\nu_j^{T-1}}{\nu_j^{T-1}} = \frac{\nu_j}{\nu_j} = \frac{\Pr(F_j = 1|Z_1, Z_2, \cdots Z_{T-1})}{\nu_j}$$

Hence we have derived Equation (3.4) given the requirement that each feature must have a fixed level.

The above derivation for feature likelihood term is exact, but approximation is necessary for the sequence likelihood term. Similar to feature observations, we approximate the likelihood ratio of emission probabilities for sequence by a set of motifs over the background likelihood as the product of this likelihood at each level,

$$\frac{\Pr(S|Z_1, Z_2, \cdots Z_{T-1})}{\Pr(S|\lambda_0)} \approx \prod_{t=1}^{T-1} \frac{\Pr(S|Z_t)}{\Pr(S|\lambda_0)} \tag{3.6}$$

where λ_0 is the null model as in Equation (3.2). We assume that motifs are independent to each other since motif length is set to be short (either set to 4 peptides or 3 to 7 peptides) comparing to the sequence length, as is the case in most known targeting motifs. This is a common assumption (e.g. [51]) and necessary for avoiding overfitting. However as we discussed in section 2.4 this assumption requires that no motif is emitted twice in different levels, which is achieved by fixing the level of each feature. Similar to Equation (3.5) we

```
1. Estimate the associations between features and compartments
   using a hypergeometric test.
2. Select features significantly associated with at least one compartment.
2. Start with an initial structure estimated from associations
   between features and compartments.
3. While BIC score improves do
   a. For each level, create a candidate structure as follows.
      i. Add a node (state) at this level.
      ii. Link from all upper nodes and link to all lower nodes.
      iii. Run EM to optimize parameters.
      iv. Prune edges (transitions) rarely visited based on the parameters.
      v. Prune emissions rarely used based on the parameters.
      vi. Run EM again to adjust parameters.
   b. Create candidate structures by randomly splitting the state
      with largest number of out-transitions.
      i. Create a new state at the same level.
      ii. Each out-transition has 1/2 probability to be moved
          to the new state.
      iii. Copy the in-transitions to the new state.
      iv. Run EM to optimize parameters.
      v. Prune transitions and emissions rarely visited.
      vi. Repeat for a fixed number of times, e.g. the number of levels.
   c. Choose the candidate structure with highest BIC score.
   d. If improving, update to that structure; otherwise stop.
```

Figure 3.2: Algorithm for structure search.

also write the sequence likelihood term as

$$\Pr(S|Z_1, Z_2, \cdots Z_{T-1}) = \Pr(S|\lambda_0)^{2-T} \prod_{t=1}^{T-1} \Pr(S|Z_t). \tag{3.7}$$

By combining Equation (3.5) and (3.7), we show that the likelihood of the full model in Figure 3.1A and the likelihood of the simplified model in Figure 3.1B is approximately up to a constant factor, so that optimizing the simplified model also optimizes the original model.

3.4 Structure Learning

In addition to learning the parameters (emission and transition probabilities) we also need to learn the set of states that should be included in our model. The learning algorithm is formally presented in Figure 3.2. We start by associating potential features (protein interactions and known motifs) with compartments. For a potential feature, we use the hypergeometric distribution to determine the significance of this association (by looking at the overlap between proteins assigned to each compartment and proteins that are associated with each of the features). We next identify a set of significantly associated compartments (p-value < 0.01 with Bonferroni correction) for each potential feature. Features that are significantly associated with at least one compartment are selected and the remaining features are removed.

After feature selection, we estimate an initial structure by using the association between features and compartments. All features that correspond to the same set of associated compartments are grouped and assigned to a single state, such that this state emits these features with uniform probability. These features are fixed to the level corresponding to the number of compartments they are significantly associated with and can only be emitted by states on that level (we tried optimizing these feature levels as part of the iterative learning process but this did not improve performance while drastically increasing run time). Initial transition between states is determined from the inclusion relationship of the set of compartments (states for which features are associated with more compartments are assigned to higher levels). We initially only allow transitions between two states where the second state contains features that are associated with a subset of the compartments of the first state. That is, the initial structure resembles a partially ordered set when the states are ordered by inclusion. The transition probability out of a state is also set to the uniform distribution. The number of levels of this structure, T, will be fixed throughout the structure search process.

Starting with this initial model, we use a greedy search algorithm which attempts to optimize the Bayesian information criterion (BIC), which is the negative data log likelihood plus a penalty term for model selection.

$$\mathrm{BIC} = -2 \log \Pr(\mathbf{S}, \mathbf{F}, \mathbf{Y} | \Theta) + |\Theta| \log N$$

where $\mathbf{S}, \mathbf{F}, \mathbf{Y}$ are the collection of sequences, feature observations, and compartments of

the proteins in the training data. $\Theta = \pi, A, B)$ denote the parameters of the HMM. $|\Theta|$ is the number of parameters according to the structure, which is a function of the number of states and the number of transitions and emissions of each state. Complicated structures will have large $|\Theta|$ while simple structures will have small ones. N is the number of proteins in our training data. BIC is asymptotically consistent while Akaike information criterion (AIC) is not, and BIC is chosen particularly because we prefer sparser structures [82]. Since use of BIC can sometimes lead to overfitting, we compared the use of BIC to 4-fold internal cross-validation for model selection. BIC is faster than internal cross validation and performed better on simulated data (see section 3.5.1).

To improve the initial structure described above we perform two types of local moves at each search iteration: adding a new state and splitting the largest state. For each level, we try adding a state which is fully connected to all states in levels above and below it and emits all features on that level. We run standard EM algorithm [83] to optimize the parameters of the model for all states (transition and emission probabilities). Transitions and emissions with probabilities lower than a specific threshold are pruned. Features not emitted by any states are also pruned, so the feature set becomes smaller and smaller. Then we run EM algorithm again because the parameters are changed. A candidate model and structure is created by this process for each level. We also try splitting the largest state, defined as the state with the largest number of out-transitions. A randomly chosen half of the out-transitions will be moved to a newly created state which shares the same in-transitions and emissions. As above we run EM algorithm, prune transitions and emission, and run EM algorithm again to obtain a candidate structure. We try this for a fixed number of times, usually the number of levels so that half of the local moves are adding and half are splitting. Among all candidate structures obtained by adding and splitting, the one with the highest BIC score is chosen. This procedure is repeated until the BIC score no longer improves.

3.5 Results

3.5.1 Simulated Data

We first tested our method using simulated data in order to determine how well it can recover a known underlying structure given only information on destinations, carriers and motifs. We manually created structures with 7, 14, 23, 25, and 31 states with multiple

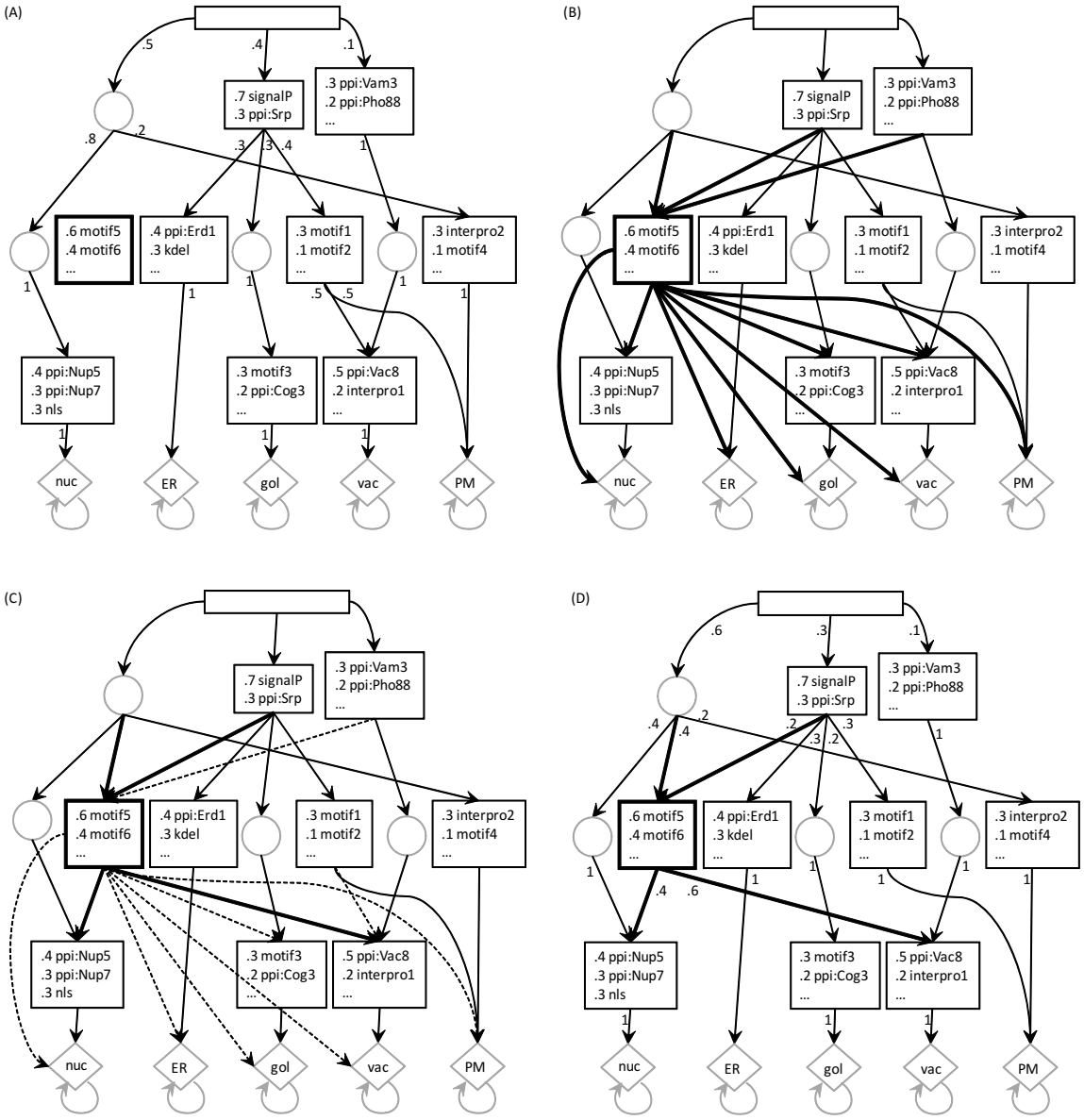


Figure 3.3: An example of a HMM state space that represents protein sorting pathways. Motifs or carriers are denoted as m_i. The top block is the initial state, and the compartments in a dataset (blocks with names) correspond to the bottom blocks. The shaded blocks and arrows are supplementary structures that make the state space compatible with a HMM of fixed length.

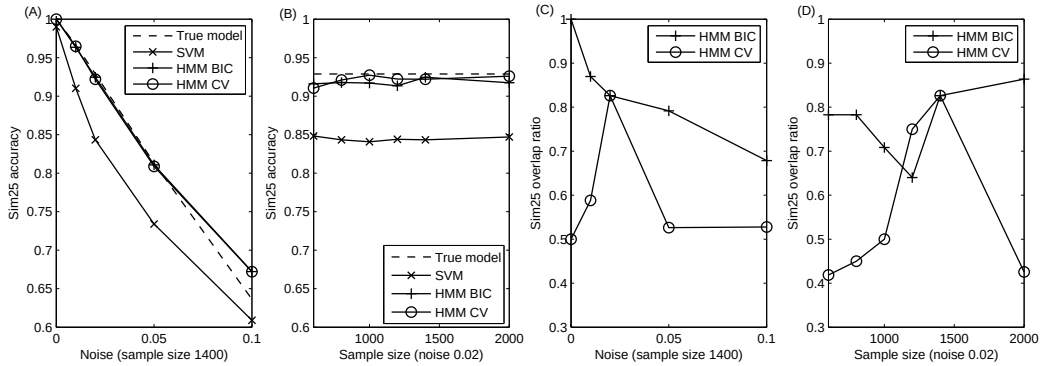


Figure 3.4: (**A**) Testing accuracy of simulated dataset generated from a structure with 25 states with varying levels of noise (false positive and false negative in features). The training sample size was fixed at 1400. (**B**) Testing accuracy versus different training sample sizes. The noise level was fixed at 2%. (**C**) The ratio of overlapping nodes and edges between the learned model and the true model with varying levels of noise. The training sample size was fixed at 1400. (**D**) The ratio of overlapping nodes and edges with varying training sample sizes. The noise level was fixed at 2%.

emitted features per state (see Supporting Website for the structure of these models). For each structure we simulate the probabilistic generative procedure and record the emitted features. 1,200 proteins are generated from the model, with varying levels of noise (leading to false positive and false negative features for proteins). We also tested various sizes of input sets with a fixed noise level.

Predicting Protein Locations

While it is not its primary goal, our method can provide predictions regarding the final localization of each protein. For each training dataset, we therefore generated a test dataset with 4,000 proteins from the same model and evaluated the accuracy of predicting protein localization for the test data using the structure and model learned by our method. Our method is compared to predictions made by the true model (note that due to noise, the true model can make mistakes as well) and by a linear support vector machine (SVM) learned from the training data using the features associated with each protein. Prediction accuracy on the 25-states dataset is shown in Figure 3.4 and the accuracy of other simulated datasets are available on the Supporting Website. As can be seen, when noise levels are low our model performs well and its accuracy is similar to that obtained by the true model for both simple and more complicated models. Both the learned model and the true model

outperform SVM which does not try to model the generative process in which proteins are sorted in cells relying instead on a one vs. all classification strategy. We compare model selection based on BIC versus 4-fold internal cross validation. BIC achieved similar accuracy with less computation, and matched the true structure better.

Recovering the True Structure

To quantitatively evaluate how well a learned structure resembles the true structure, we use the graph edit distance to measure their topological similarity [84]. First we need to match the nodes in a learned structure to a node in the true structure. We run the Viterbi algorithm on proteins in the testing data, and count the state co-occurrence matrix W whose elements W_{ij} is the co-occurrence of state i in the learned model and state j in the true model, i.e. the number of proteins in which the two states i and j occur in the Viterbi path inferred by the two models. The optimal one-to-one matching M, denoted as a set containing pairs of matched state indexes, can be found by running the Hungarian algorithm on the co-occurrence matrix W optimizing the objective function $\sum_{(i,j)\in M} W_{ij}$.

With the optimal matching we use the maximum common subgraph (MCS) and minimum common supergraph in the graph edit distance methodology to quantify similarity between two structures. Given two graphs G_1 and G_2, let $\hat{G}$ and $\check{G}$ be the MCS and minimum common supergraph of G_1 and G_2. Denote $|G|$ as the size, or the number of edges and nodes of a graph, we define the overlap rate as $|\hat{G}|/|\check{G}|$, i.e. the percentage of overlapping edges and nodes. The overlap rate comparing to the true model on the 25-states dataset is shown in Figure 3.4C. Structural comparison on other datasets is available on the supporting website. As can be seen, our algorithm successfully recovers the correct structure in all cases with 0% noise. As the noise increases the accuracy decreases. However, even for very high levels of noise the two models share a substantial overlap (around 40% of states and trnasitions could be matched).

3.5.2 Yeast Data

We next evaluated our method using subcellular locations of yeast proteins derived from fluorescence microscopy (the UCSF yeast GFP dataset [1]). This dataset contains 3,914 proteins that were manually annotated, based on imaging data, to 22 compartments. We collected the features from the following sources. Protein-protein interaction (PPI) data was downloaded from BioGRID (BiG) [85]. For deterministic motifs we use the annotated

occurrences of InterPro [35] domains and the following three signal sequences listed on UniProt [18]:

1. Signal peptides: UniProt defines this sequence feature based on the literature or consensus vote of four programs, SignalP, TargetP, Phobius and Predotar.

2. Transmembrane region: UniProt annotates a sequence with this feature either based on literature or consensus vote of four programs, TMHMM, Memsat, Phobius and Eisenberg.

3. GPI anchor: UniProt annotation for this feature either relies on literature or prediction by the program big-PI.

The above features are filtered by a hypergeometric test to identify features with a significant association with a final destination (p-value < 0.01 with Bonferroni correction) before learning the model.

To extract novel motifs associated with localization, we downloaded protein sequences from UniProt [18] and run generative and discriminative HMM motif finder [73]. We extract 20 motifs for each compartment, and compared setting all to length 4 versus setting the length to range from 3 to 7. The performance in all following evaluations are similar and we show results based on motif length as 4. We will compare using likelihood and binary occurrence for motif features. For binary motif occurrence, a motif is considered present if posterior probabilities of the begin state and the end state of the motif are both greater than 0.9 (detail in [73]).

Predicting Protein Locations

As with the simulated data, we first evaluated the accuracy of predicting the final subcellular location for each protein. This provides a useful benchmark for comparison to all other computational methods for which this is the end result. The performance is evaluated by 10-fold cross-validation. In each fold both feature selection and motif finding are restricted to the training data without accessing the testing data. We use three conventional measure in information retrieval: the accuracy, micro-averaging F1 and macro-averaging F1 [86]. For the accuracy, a prediction is considered correct if it matches any of the true locations. The F1 score is the harmonic mean of precision and recall [87]. Micro-averaging takes the average of the F1 score over all proteins, giving each protein an equal weight; in other

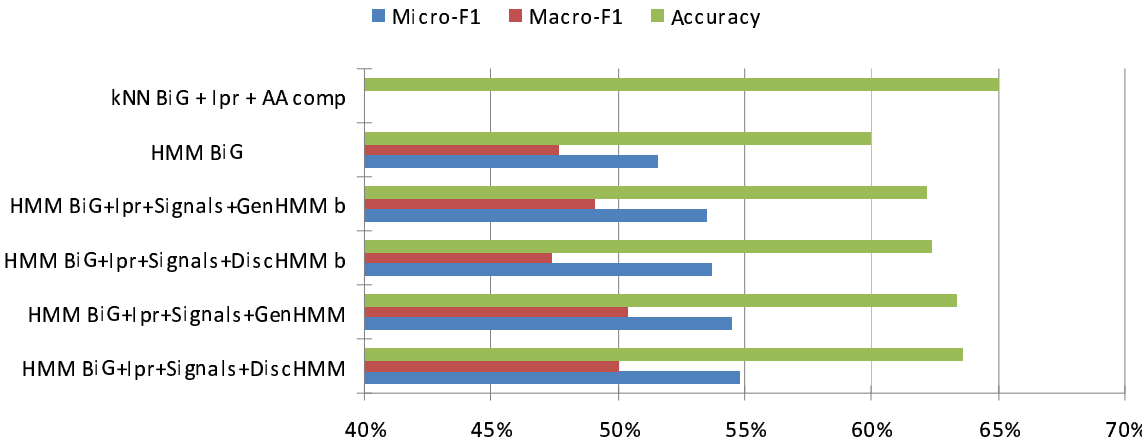


Figure 3.5: The accuracy of predicting the final subcellular location. For kNN we use the reported accuracy based on PPI information from BiG, deterministic InterPro motif annotation from UniProt, and amino acid composition of different length, gaps, and chemical properties using leave one out cross validation [26]. For HMM we also show micro-averaging and macro-averaging F1 score in 10-fold cross validation. The features for HMM include InterPro and BiG, and three signal sequences from UniProt. The novel motifs are learned using generative or discriminative HMM of length 4, represented by likelihood and binary features (GenHMM/DiscHMM b)

words, the classes are weighted by their sizes. Macro-averaging takes the average of the score over classes, giving each class an equal weight. Including macro-averaging F1 ensures smaller classes are not ignored since other measures are dominated by large classes. The result is shown in Figure 3.5. We compared our method with the k-Nearest Neighbors (kNN) from Lee *et al* [26] which was shown by the authors to outperform other methods. As can be seen in Figure 3.5 PPI information (BiG) provides the major contribution for accurate predictions while InterPro motifs do not contribute as much. This agrees with previous studies [25, 26]. When adding more features the performance improves and the best result is achieved using all features. Note that the accuracy of our method is very close to that of the kNN method. However, it is important to note that our method performs the much harder task of simultaneously learning the sorting pathways as well as predicting locations. Unlike these prior methods our method correctly determines pathways and not just end points. This is an important contribution of the method which is achieved while not compromising prediction accuracy.

Evaluation of the Learned Structure

To evaluate the accuracy of the learned structure, we collected information about known sorting pathways from the literature. We were able to find information regarding 13 classical and non-classical sorting pathways. For each of these pathways we identified a set of carriers or motifs that govern the pathway and, when available, the set of proteins that are predicted to use this pathway. Figure 3.6 presents the pathways we collected from the literature. For example the classical HDEL pathway into ER has two steps. In the first, proteins with signal peptide (SP) are introduced into this pathway by the SRP complex. In the second, proteins with the HDEL motif are retained in ER by interaction with proteins Erd1 and Erd2. The full list of carriers and motifs for these pathways is provided on the supporting website.

We first wanted to check if the databases we used for obtaining features contain the carrier information for the literature pathway. We filtered pathways for which carrier information in the BIG database did not contain enough proteins (and thus no method can identify this pathway based in this input data). This leaves 10 pathways that could, in principal, be recovered by computational models. Sorting steps that were filtered out in this way are represented as shaded links in Figure 3.6.

To determine whether we accurately recovered a pathway in our model we looked at the carriers and motifs that are associated with that pathway in the literature. A step in a literature pathway can be matched to a state if the state emits any carrier or motif in that step. A known pathway is considered recovered in a learned structure if its steps can be matched to the states along a path from the root to the compartment to which it leads. A pathway is partially recovered if only some of its steps can be matched. For example, the MVB pathway (Figure 3.6) is only partially recovered (66.7%) because the third step does not have a well-represented carrier in the data sources. The numbers of recovered pathways for different sets of features are listed in Table 3.1. The ranges correspond to the different folds in our cross validation analysis. Fractions represent partial matches as discussed above. When using the full set of input features our algorithm is able to recover roughly 80% of known pathways. Most of these pathways are recovered in all 10 folds (Table 3.1). Note that because some carriers do not appear in our database not all steps in all pathways can be matched and the best possible recovery is 8.7. Thus, the 7.7 recovery obtained is very close to optimal.

We rely on the hypergeometric test for feature selection. If a feature (e.g. a specific

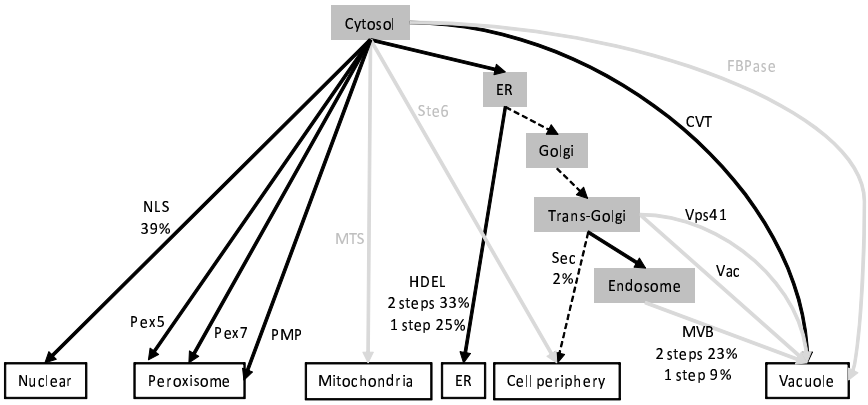


Figure 3.6: Protein sorting pathways collected from the literature. Each pathway is a path from cytosol to a compartment at the bottom, consisting of one or more steps (the links) that transport proteins between intermediate locations. Each step has a list of carriers and motifs responsible for the transportation by which we can verify whether the pathway is recovered. Shaded links denote steps whose carriers are underrepresented on BiG (covering less than 5% of proteins transported to the corresponding compartment in the GFP dataset). Dashed lines denote steps taken by default without specific carriers. The percentage under pathway name is the protein sorting precision when the pathway is recovered, as described in Table 3.2.

carrier) is not selected, it could never occur in the model. For carriers, feature selection depend on the data in BiG, but the interaction of a carrier and its cargos may not be present. For example, because of lack of evidence (the motif and carrier detection steps did not find the Vam3, Vam7, or the Vps41 features), the classical vacuole import pathway (Vac in Figure 3.6) and the alternative Vps41 pathway can only be 50% recovered (each missing a step). For both, the step of signal peptide (SP) is accurately found, but alternative motifs/carriers are selected to route proteins to the vacuole or cell periphery. We believe that Vam3 and Vam7 interact with more vacuolar proteins, but the interaction is missing in BiG so they are filtered out by the feature selection process.

We further collected lists of proteins indicated as following specific pathways in the literature for 4 of the pathways, NLS, HDEL, Sec and MVB, and tested whether the recovered pathways indeed sort proteins on the correct path to the correct destination (allowing close compartments as above). For each protein, we use the Viterbi algorithm to infer the highest probability path of states the protein is expected to follow according to our learned model, and compare the Viterbi path to the known pathways. Again counting partial match of a multi-step pathway as above, on average using all features results in correctly assigning

Table 3.1: Pathway recovery results of structure learned from different feature sets. The precision of inferred protein path is also listed here. Mean, minimum and maximum among the 10 folds are shown.

Features	Pathway recovery	Inferred protein path
HMM BiG	5.9 (4.7 - 8.0)	7% (4% - 10%)
HMM BiG + Ipr + Signals	7.2 (5.7 - 8.7)	8% (6% - 11%)
HMM BiG + Ipr + Signals + GenHMM b	6.2 (4.3 - 7.7)	8.4% (6% - 11%)
HMM BiG + Ipr + Signals + DiscHMM b	6.2 (5.3 - 7.3)	8.4% (6% - 11%)
HMM BiG + Ipr + Signals + GenHMM	7.7 (6.7 - 8.7)	17.9% (13% - 23%)
HMM BiG + Ipr + Signals + DiscHMM	7.7 (6.7 - 8.7)	19% (15% - 23%)

21% of 63 proteins. Focusing on a representative feature set, detailed protein path results for each pathway are also given in Table 3.2. The recovered NLS pathway sorted 39% of proteins correctly, and the recovered HDEL pathway sorted 33% correctly but sorted the other 25% via SP. Similarly the recovered MVB pathway sorted 23% to go through two of the three steps (SP and MVB) and other 9% to one of the three steps. The recovered Sec pathway only sorted 2% of the proteins to go through SP and end at cell periphery. However, this was due to the fact that while 17 of the 28 proteins collected from literature as being secreted were included in the GFP dataset, the majority are labeled as ER and vacule and none are labeled as cell periphery. Overall the GFP dataset include 40 out of the 63 proteins whose pathway is known, of which only 28% are labeled in agreement with our lierature survey.

It is important to note that our analysis of the learned structure may underestimate its accuracy, since it may have recovered correct pathways that could not be verified due to insufficient detection of relevant motifs or carriers in the input data.

Figure 3.7 shows one of the learned structures obtained using all features. Besides carriers and motifs included in our literature pathway collection (marked as boldface), many other features were found that are also known to participate in protein trafficking as curated in SGD [16] (marked with an asterisk). For those compartments not covered by our collection of known pathways, the general topology of this structure agrees with our basic understanding of cell biology. For example microtubule share a step with spindle pole, which in turn share a step with nuclear periphery, and cell periphery share steps with bud neck, which in turn share steps with bud and actin.

Table 3.2: Recovery and protein sorting results of each pathway using the features BiG + InterPro + Signals + DiscHMM 4.

Compartment	Pathway (#proteins)	Recovery (folds)	Steps	Sorting
Nucleus	NLS(15)	10/10	all	39%
Peroxisome	Pex5	1/10	all	
	Pex7	10/10	all	
	PMP	9/10	all	
ER	HDEL(11)	10/10	SP+HDEL	33%
			SP	25%
Cell periphery	Sec(28)	10/10	SP	2%
Vacuole	Vac	10/10	SP	
	MVB(9)	10/10	SP+MVB	23%
			SP	9%
	Vps41	10/10	SP	
	CVT	10/10	all	

Prioritizing Pathway Predictions for Possible Experiments

Given that the sorting routes taken by many proteins are currently unknown, the most important part of our work is the potential to identify novel pathways. In this regard, we note that, just like hand-constructed pathways, any novel putative pathways contained in our learned model can be readily tested experimentally by perturbing motifs and/or carriers. Our pathway HMM is composed of hidden states that correspond to intermediate locations, and the emissions correspond to carriers or motifs that are responsible for transportation into a location. Sometimes we do not have the same confidence over an entire path from root to destination. To perform validation experiments more efficiently, it would be better to focus on the more confident part of the learned structure. Hence we developed the following criterion to prioritize the hidden states, which also serve as basic units for possible experiments.

Our goal is to assign higher confidence to a state that leads to correct inference of the destination. We measure the association between occurrence of each state and the correctness of inferred destination. Occurrence of states is based on the optimal path inferred by the Viterbi algorithm. We use the testing data (held-out data not utilized during training) to calculate confidence. The hypergeometric test is used to rank whether a state is significantly associated to correct destination. This way a top state must have high precision (correct destination if proteins pass through this state) and high coverage

Table 3.3: Prioritized biological predictions on protein sorting mechanism. Each row denotes a HMM state with high confidence that corresponds to an intermediate location, and the carriers or motifs responsible for import into that location. Such states are the confident part of the learned pathways. Confidence of a state is based on whether it lead to correct inference of final destination. States significantly associated with correct inference of destination are listed, ranked by the p-value. Selected states all have high precision (accuracy given the occurrence of this state). Transportation mechanism into a state can be validated experimentally by perturbing one of the top 3 carriers or motifs. All possible destination compartments from a state are also listed. The upper part contains pathway prediction of the fold with highest accuracy in cross validation, and the lower part is from another fold with good performance.

State	p-value	Prec	Carrier / motif	Possible destination
111	$< .01$	86%	Dnf1, Trs120, Gga2	late Golgi
77	.01	100%	Trs33, Tvp15, Cka1	late Golgi
75	.04	100%	Sed5, Sec21, Cog1	Golgi, early Golgi
98	.04	100%	Cog3, Cho2, Sla1	punctate composite, early/late Golgi, vacuole, lipid particle
112	.04	100%	Svp26, Tos1, Tip20	Golgi
25	.05	80%	Arf1, Sys1, Erv14	Golgi, early/late Golgi
107	$< .01$	81%	Ypp1, Sec14, Mup1	punctate composite
80	$< .01$	75%	Vps51, Arl1, Chc1	punctate composite
106	.01	86%	Fth1, Vma10, Atg27	vacuole membrane
89	.01	78%	Cog3, signal peptides, Kex1	late Golgi, vacuole, vacuole membrane, nuclear periphery, peroxisome
27	.04	100%	Fah1, Get1, Drs2	cytosol, ER, ER-Golgi, late Golgi, actin, bud neck, spindle pole

(many proteins pass through this state). Note that confidence calculation does not involve any established knowledge in the literature, because our aim is to infer novel pathways.

The prioritized pathway predictions are listed in Table 3.3. The predicted transportation mechanisms are mostly based on carriers, but in one case also based on the motif of signal peptides. Many carriers listed in Table 3.3 are annotated as trafficking-related on SGD [16], but there could still be novel discovery. Interestingly, the highly confident states may center around one compartment in one learned structure (one fold in cross validation), while another structure is more confident around other compartments. The structure in the upper part of Table 3.3 focus on Golgi and the structure in the lower part is more confident around punctate composite and vacuole.

3.6 Discussion

The goal of this research is to propose hypotheses about protein sorting mechanisms, not just to make predictions. We propose, for what we believe is the first time, a method to learn sorting pathways from protein localization annotation, based on co-occurrence of interacting partner and sequence motif. Our method is able to recover a significant part of known pathways collected from the literature, and to infer the correct path of proteins known to follow these pathways.

Using a HMM naturally simulate the transportation path of a protein among unobserved intermediate states. Although the path is unobserved, the most likely one can be inferred by the Viterbi algorithm of the HMM based on observed features. The model is probabilistic and returns a distribution of possible compartments, instead of a single predicted compartment. Proteins that are targeted to more than one compartment in the training data can be handled by treating multiple localization as uncertainty.

An additional advantage of building comprehensive sorting models is that potential inconsistencies in canonical models can be identified and experiments performed to resolve them. We have derived a list of biological prediction of protein transportation mechanism based on carriers (receptors) and motifs. This list is ranked by confidence calculated on the learned structure, allowing biologists to focus on the more confident part of the inferred pathways and reduce the experimental efforts.

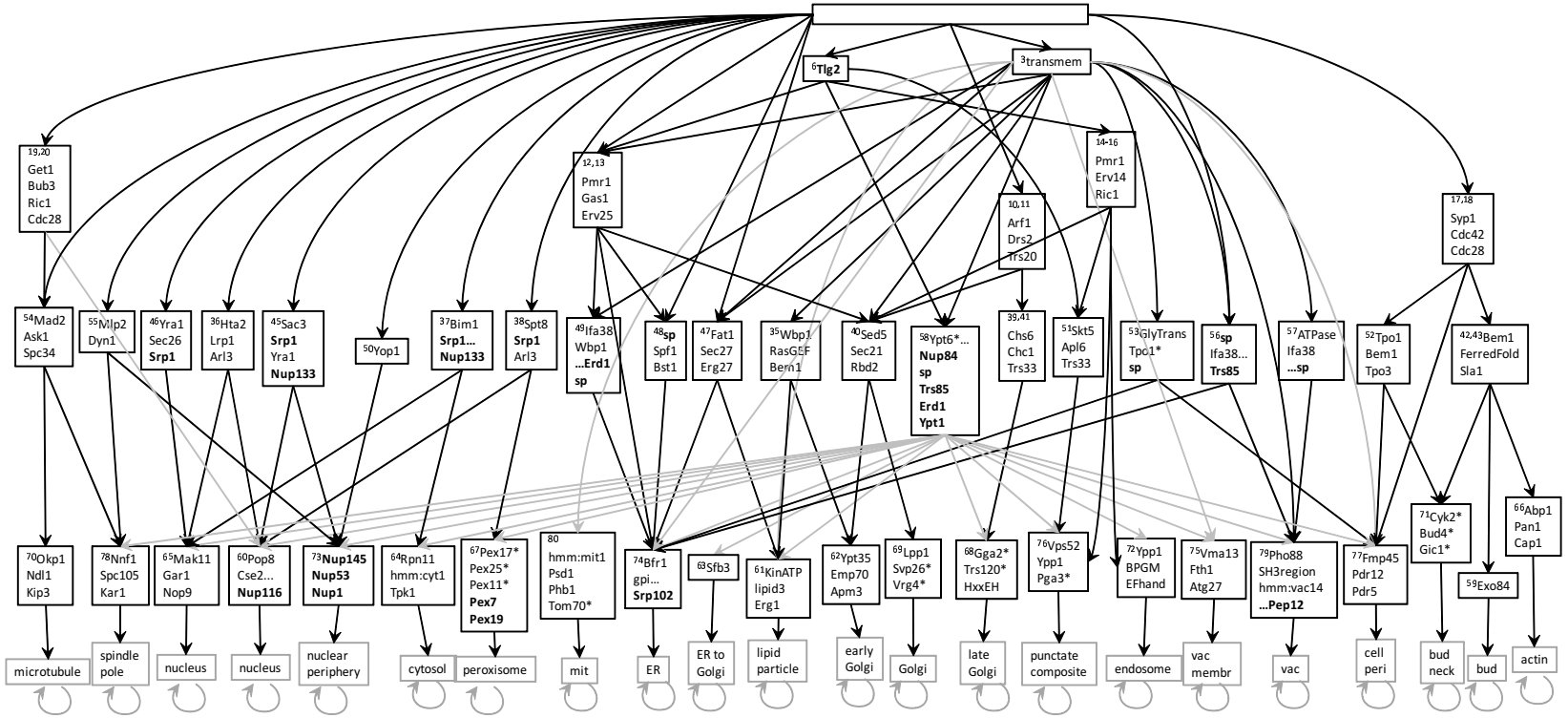


Figure 3.7: The HMM state space structure learned by our method that corresponds to potential protein sorting pathways. A state is represented by a block; its transitions are shown as arrows and its top 3 emitting features are listed inside the block. The sparse transition and emission probabilities are omitted here. The initial state probabilities are denoted as arrows from the root block at the top. The bottom states are the final destination compartments. Some transitions are shaded only because of visual clarity, including transitions across levels or from and to the highly connected state (state 58). While silent states are not explicitly displayed (to remove clutter) they are actually implicitly present. Any time an edge jumps more than one level it is going through silent state(s). For example, the right most edge coming out of the root goes through a silent state in the first level. Carriers and motifs that matches our literature pathway collection are shown in boldface; other features potentially related to protein trafficking according to SGD are marked with an asterisk.

Chapter 4

Extending to Higher Organisms

We have demonstrated the utility of discriminative motif finding using known targeting pathways in Chapter 2, and proposed to model and discover targeting pathways in budding yeast without using prior knowledge in Chapter 3. Although proteome information is more abundant in yeast, it is of more importance to understand targeting pathways in higher organisms, especially human. There are many potential biomedical applications, e.g. the study of cancer and other diseases [3–6]. Yet the mechanism of subcellular localization in human cells is not well understood as in yeast cells. Recently the Human Protein Atlas (HPA) has collected a large amount of location proteomic data in human [88]. About 5000 confocal microscopy images using antibodies are added to the Atlas to provide more detailed protein localization, and images of more proteins are expected to be generated [14]. It has been shown that automated determination of location based on the Atlas images is highly accurate [21]. This resource provides reliable training data for our model. The cellular transport machinery is more complex in human than in yeast. First, alternative splicing is much more common in human. Second, unlike in yeast which is unicellular, in human we need to consider many cell types combined with different conditions. For the HPA dataset, there are three cell lines and more than half of the proteins change one of the locations between cell lines. Most proteins are expected to remain in the same compartment across conditions and cell types, but some will have altered compartment under specific condition. We have extended our model in Chapter 3 to support alternative splicing and to incorporate condition into localization path prediction and inferring condition-specific targeting pathways. The extended model is applied to human localization data manually annotated based on HPA confocal microscopy images.

4.1 Related Work

Since most of the protein sorting mechanisms are conserved across a wide range of species, many localization classifiers support human or mammalian proteins. The programs TargetP and LOCtree are both tested in human and the results are compared with that in other species [23, 24]; PSLT is trained and tested for human proteins [89]. DC-kNN, a classifier that utilizes not just sequence but also Gene Ontology (GO) annotation, protein interaction, and known motifs, has been extensively tested in human as well as fruit fly and yeast [26]. However none of these sequence-based systems considers the unique challenges described above (either the cell conditions or alternative splicing).

When microscopy images under different conditions are available, automated image analysis systems can determine the localization under a large number of conditions (with combinations). This approach has been successfully applied to identify proteins whose location changes between human cancer and normal tissues, using immunohistochemistry (IHC) images provided by HPA [90]. With the immunofluorescence (IF) confocal microscopy images which provides higher resolution, more accurate automated analysis has been performed on three different cell lines [21]. We believe that there will be more such studies in the near future. However image-based analysis does not provide insight into the mechanism of location changes due to conditions, which is what we want to address in the next section.

4.2 Alternative Splicing

Most databases containing subcellular localization information (including the HPA dataset we use) associate locations to a gene, not an isoform. Although alternative splicing sometimes affects protein sorting [91, 92], there is little resource of isoform-specific localization information. Similarly, most of the relevant protein features are available on the gene level (sometimes based on the most representative isoform) and not the isoform level in databases. Such is the case for PPI and known motifs (sequence annotation on UniProt, see Results section for details). For simplicity we use the term protein for an entry in the localization and feature dataset (typically a gene) which may have many splicing variants (or isoforms). However we need to take special care for alternative splicing when utilizing novel motifs extracted from sequences. To support the large amount of alternative splicing in human we modify the two steps of generating motif features, motif discovery and feature vector calculation, as follows.

For motif discovery, all valid splicing variants of a protein in sequence databases are included. In generative motif finding, we search for motifs present in all splicing variants of all sequences in the positive set. In discriminative motif finding, we search for motifs present in all splicing variants of all proteins in the positive set and absent in all splicing variants of all proteins in the negative set. Note that the presence and absence of a motif are not strict but probabilistic. Sequences of proteins with only one splicing variant are duplicated three times (three being the median number of splicing variants), in order to avoid bias towards proteins with more variants.

As in the previous chapter, there are two approaches to convert each sequence to motif features: binary occurrence (of a motif instance), and sequence likelihood (representing how strong a motif instance is). With alternative splicing, we combine the feature vectors generated from a protein's isoform sequences into a feature vector of this protein. Our goal is that a motif is considered present in a protein if it is present in any of the splicing variants. For binary motif feature we combine the feature vector of the isoforms as follows. For a protein with V isoforms, let $F_k^{(v)}$ denote the occurrence of motif k on the isoform sequence v, $1 \leq v \leq V$. We define the combined binary motif feature F_k of this protein to be true if it is true in any $F_k^{(v)}$,

$$F_k \equiv \bigcup_{v=1}^{V} F_k^{(v)}.$$

For sequence likelihood feature, the feature vectors are combined as follows. For simplicity we use the sequence log likelihood instead of likelihood in Equation 3.2,

$$\log \Pr(S|Z_t = k) = \begin{cases} \ell(S|\lambda_k) & \text{if } 1 \leq k \leq M \\ \ell(S|\lambda_0) & \text{if } M+1 \leq k \leq M+K+1 \end{cases} \tag{4.1}$$

where $\ell(S|\lambda_0)$ is the combined background log likelihood and $\ell(S|\lambda_k)$ is the combined log likelihood of the protein sequences given motif k. The combined background log likelihood is the average over all isoforms,

$$\ell(S|\lambda_0) \equiv \frac{1}{V} \sum_{v=1}^{V} \log \Pr(S^{(v)}|\lambda_0).$$

The combined log likelihood given the motif k is set to the highest log likelihood ratio (LLR)

among all isoforms plus the combined background log likelihood,

$$\ell(S|\lambda_k) \equiv \ell(S|\lambda_0) + \max_v \log \Pr(S^{(v)}|\lambda_k) - \log \Pr(S^{(v)}|\lambda_0)$$

It is defined this way to make the combined LLR of a motif model versus background as the highest LLR among all isoforms.

4.3 Cell Line Specific Localization

In higher organisms there are much more variables (and their combinations) related to localization, including cell lines, tissue types, perturbations, diseases versus normal samples, etc. In the scope of this thesis we consider any such variable a condition, and focus on one variable, the cell lines, because of the data available. Note that the problem formulated below is not tied to cell lines and can apply to any simple set of condition (no structure among the conditions is considered). Our aim is to find out not only where the proteins are transported in different cell lines, but also how they are transported, i.e. motifs and interacting partners (carriers) that are activated or deactivated in certain cell lines (e.g. by post-translational regularization). As in the previous chapter our method consists of feature selection (motifs and carriers) and structure search for targeting pathways HMM. The extensions on these two parts are discussed below.

We use a simple extension to handle multiple cell lines in feature selection: treating locations in all cell lines as multiple locations. This directly applies to both the hypergeometric test for binary features (PPI and known motifs) and motif discovery. The underlying assumption is that such a motif (or carrier) is required even if a protein is transported to that location in only one cell line. This treatment tends to find motifs and carriers that are required in all cell lines, but a motif or carrier activated or deactivated in one cell line can also be found since the occurrences are probabilistic. After the features are selected, the activation or deactivation in the cell lines will be learned in the next phase, the HMM of targeting pathways.

An overview of the extension to the structure search algorithm is in Figure 4.2. We first collect a subset of proteins in the training data that do not change location among different cell lines, called the "common subset." Using this common subset only we learn a pathway HMM model, called the core model, by the standard structure search algorithm. For the core model the rarely visited transitions are pruned but the emissions are not. Then using

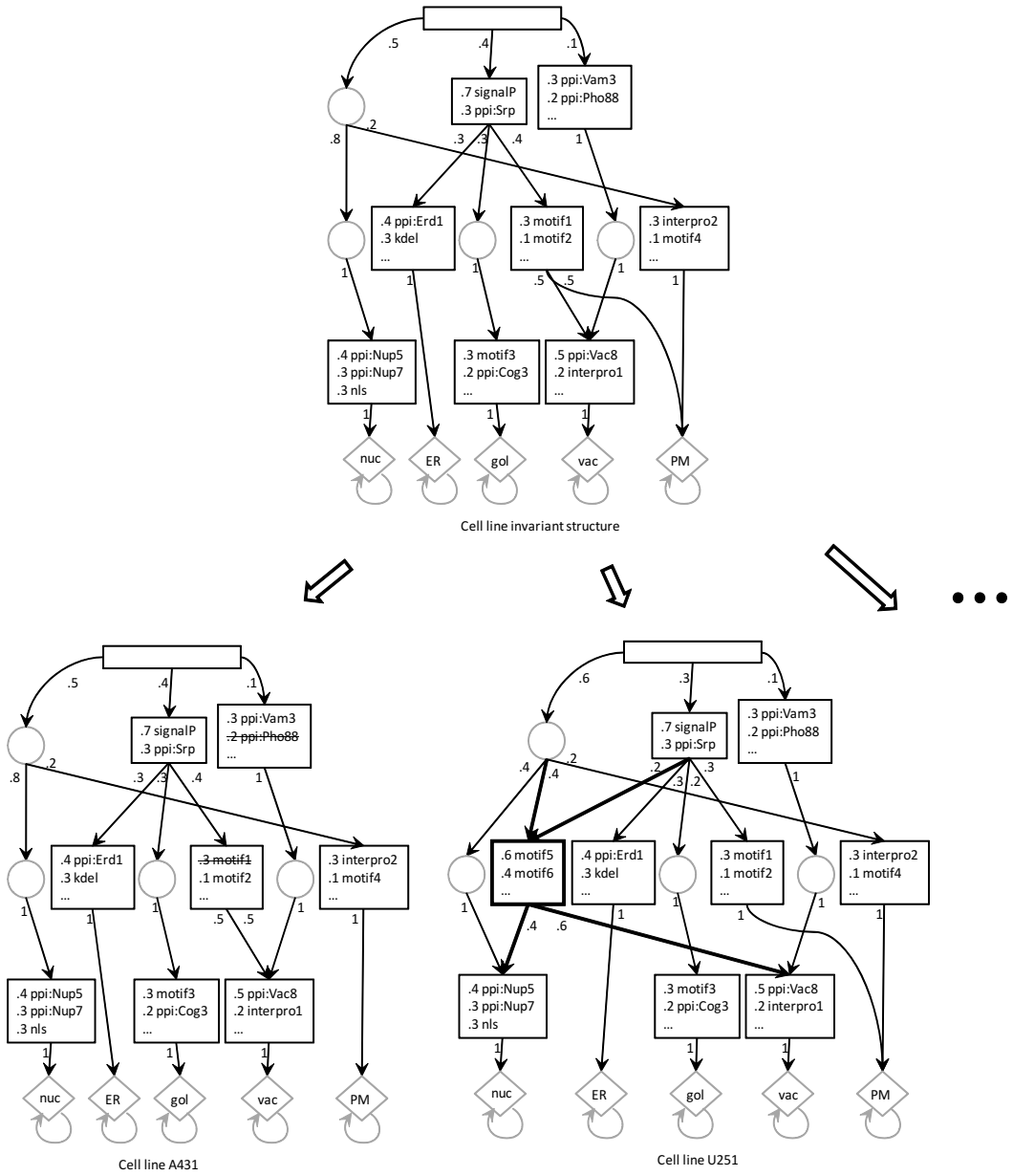


Figure 4.1: Overview of the two-phase structure search algorithm for multiple cell lines (this is a sample pathways structure). First we learn a structure from the subset of proteins whose localization is the same across all cell lines. Then for each cell line we run structure search again to fit cell line specific localizations, keeping track of the addition and removal of states, emissions and transitions. Cell line specific states represent pathways activated in an individual cell line.

```
1. Consider the locations in all cell lines as multiple locations
   and apply the feature selection procedure, including motif discovery.
2. Collect the common subset of proteins whose localization is the same
   across all cell lines.
3. Learn the core model by structure search with the common subset,
   pruning rarely visited transitions but not emissions.
4. For each cell line do
   a. Starting from the common structure above, run structure search
      with cell line specific localization annotations, removing
      rarely visited transitions and emissions.
   b. Record the transitions and emissions removed and states added
      in this cell line.
   c. Examine cell line specific states, emissions and transitions.
```

Figure 4.2: The two-phase structure search algorithm that supports multiple cell lines.

localization data for each cell line (regardless of whether the location is the same or different in other cell lines) we run structure search again. As in the standard structure search, the first step is to run EM algorithm to optimize the parameters and to prune transitions and emissions based on these parameters. The pruned transitions and emissions will be different in each cell line. At each search iteration after the first step, we try adding a new state or splitting the largest state, to see if it fits the training localization data in this cell line. Thus we obtain a modified structure for each cell line. The added states and pruned emissions and transitions correspond to pathways and carriers activated or deactivated in a specific cell line. See Figure 4.2 for the formal algorithm.

4.4 Results

We evaluate the extended algorithm on human protein localization data obtained from confocal microscopy images in the HPA database. Localization is annotated manually by experts in the HPA team based on fluorescent microscopy images of release 5.0, with further corrections after the public release [14]. 2,889 proteins, the majority of this dataset, are used except a few invalid ones that lack entries on Ensembl [93] or UniProt [18]. Localization is annotated in three different cell lines, A-431, U-251MG, and U-2 OS. Only 1,123 proteins are in the same location across three cell lines. The locations are grouped to ten classes: centrosome, cytoskeleton, cytosol, ER, Golgi, mitochondria, nuclei, nucleoli, plasma

Table 4.1: Features and data sources for HPA dataset

Feature type	Data source
Novel motifs	Extracted by generative and discriminative HMM [73]
Protein interactions	Downloaded from BiG [85]
Known short motifs	Represented by regular expression in Minimotif Miner [57]
Sequence annotations	Presence of sequence annotations defined on UniProt [18]
Active site	Amino acid(s) directly involved in the activity of an enzyme
Binding site	Binding site for any chemical group (co-enzyme, etc)
Calcium binding	Position(s) of calcium binding region(s) within the protein
Compositional bias	Region of compositional bias in the protein
Cross-link	Residues participating in covalent linkage(s) between proteins
Disulfide bond	Cysteine residues participating in disulfide bonds
DNA binding	Position and type of a DNA-binding domain
Domain	Position and type of each modular protein domain (InterPro)
Glycosylation	Covalently attached glycan group(s)
Initiator methionine	Cleavage of the initiator methionine
Lipidation	Covalently attached lipid group(s)
Metal binding	Binding site for a metal ion
Modified residue	Modified residues excluding lipids, glycans and protein cross-links
Motif	Short (up to 20 amino acids) sequence motif of biological interest
Nucleotide binding	Nucleotide phosphate binding region
Peptide	Extent of an active peptide in the mature protein
Propeptide	Part of a protein that is cleaved during maturation or activation
Signal	Sequence targeting proteins to the secretory pathway
Transit peptide	Extent of a transit peptide for organelle targeting
Transmembrane	Extent of a membrane-spanning region
Zinc finger	Position(s) and type(s) of zinc fingers within the protein

membrane (PM), and vesicles.

The features and corresponding data sources are described in Table 4.1. Novel motifs are extracted from amino acid sequences downloaded from Ensembl [93]. Again we use the generative and discriminative HMM motif finder described in Chapter 2 [73]. PPI data is downloaded from BiG [85]. Since human cells are more complicated than yeast, we use two more informative feature types for known motifs. One is short motifs represented as regular expression in the database Minimotif Miner [57]. We include motifs marked as traffick related. The other is presence of sequence annotations on UniProt [18]. Three such annotations have been utilized in yeast, but we extend to all sequence annotations except one that relies on localization information (resulting in circular reasoning) and those too

general (e.g. secondary structure and coiled coil). The list of annotation subtype is listed in Table 4.1. As in yeast, we apply the hypergeometric test to select features having a significant association with any destinate compartment (here using p-value < 0.05) before learning the model.

4.4.1 Predicting Protein Locations

Similar to the evaluation we performed in the previous chapter, although our goal is learning the pathways predicting the final subcellular location remains an objective way to evaluate the performance of our method. The performance is evaluated by 10-fold cross-validation, in which the testing data is kept away from both feature selection and model training. The result is shown in Figure 4.3. We compare our method to SVM using the same feature set as a classifier that do not utilize any pathway structure (the linear kernel and default setting of SVMlight are used [56]). As in section 3.5.2, evaluation is based on three conventional measures in information retrieval: the accuracy, micro-averaging F1 and macro-averaging F1 [86]. HMM performs better than SVM on the three measures in most of the feature sets, indicating the importance of learning the pathway structure. Yet when using likelihood scores as novel motif features HMM is less accurate than SVM or HMM using binary occurrence for novel motifs. We do not see significant improvement by adding novel motifs extracted from sequence, either generative or discriminative motifs. Most likely this is because the known motif information provided by sequence annotation in UniProt and regular expression in Minimotif Miner is already very comprehensive. The confusion matrix using the feature set of sequence annotation and Minimotif Miner is shown in Table 4.2. Proteins belonging to compartments with fewer training examples are often incorrectly predicted to be in larger compartments, especially nuclei in cell line U-251MG, cytosol in cell line U-2 OS and A-431. The most likely reason is that the optimization objective function, BIC score, correlates with overall likelihood which is dominated by larger compartments.

4.4.2 Evaluation of the Learned Structure

As in yeast, we also collected a list of known sorting pathways from the literature in human (see Figure 4.4). We identified 9 sorting pathways in human, most being well known and some less common. Each step in these pathways corresponds to a list of carriers or motifs responsible for that transport according to the literature. In yeast we rely on PPI for the validation of almost every known pathway. In human the classical sorting pathways are

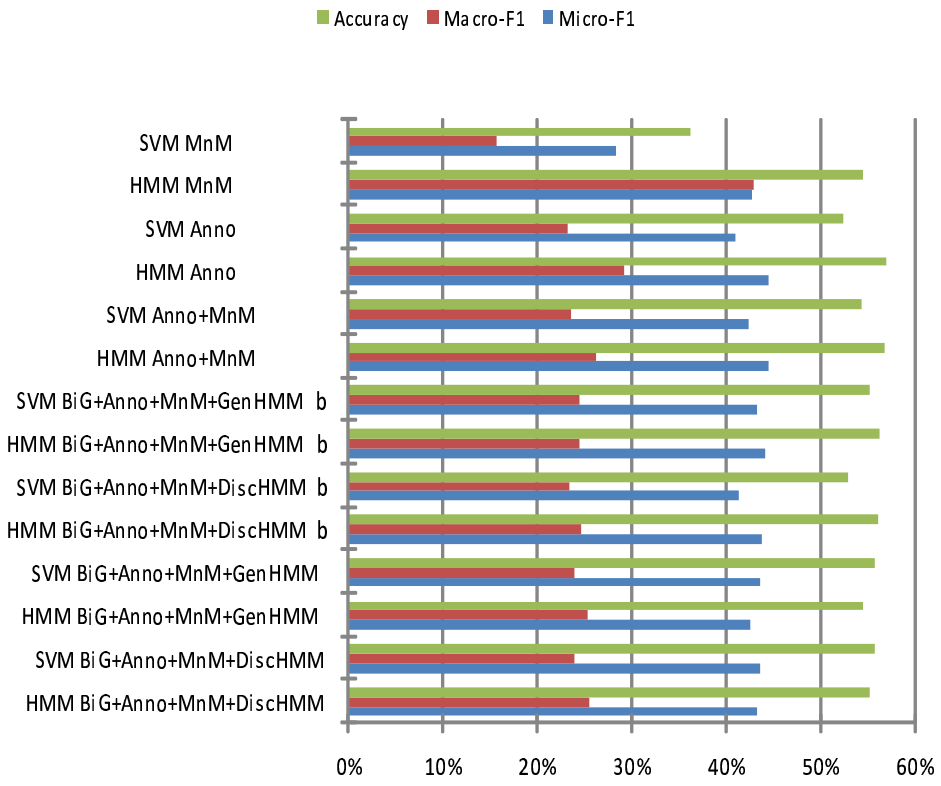


Figure 4.3: The performance of predicting the final subcellular location. Predication is evaluated by accuracy, micro-averaging F1, and macro-averaging F1 in 10-fold cross validation. We compared the result of different combinations of several feature types, including PPI in the BiG database, sequence annotation in UniProt, regular expression in Minimotif Miner, and novel motifs of length 4 extracted by generative and discriminative HMM.

	Prediction for cell line A-431									
	Cent	Cyto	ER	Golgi	Mito	Nuclei	Nucleoli	PM	Cytoskel	Vesicles
Cent	0.0	83.7	0.0	0.0	0.0	15.2	0.0	0.0	0.0	1.1
Cyto	0.1	78.3	0.3	0.6	1.9	16.9	0.1	0.9	0.0	0.9
ER	0.0	69.7	6.1	1.0	2.0	18.7	0.0	0.0	0.0	2.5
Golgi	0.0	71.1	0.0	0.4	0.9	22.9	0.9	2.0	0.0	1.9
Mito	0.0	68.2	0.7	0.0	20.8	9.0	0.0	0.3	0.4	0.7
Nuclei	0.1	71.7	0.0	0.5	1.7	24.1	0.1	0.9	0.0	0.9
Nucleoli	0.0	74.7	0.0	0.4	1.5	22.9	0.5	0.0	0.0	0.0
PM	0.0	75.8	0.5	0.3	1.9	17.0	0.5	2.5	0.0	1.5
Cytoskel	0.0	83.5	0.0	0.6	1.9	12.5	0.0	0.6	0.0	1.0
Vesicles	0.0	80.9	0.0	0.4	1.3	17.4	0.0	0.0	0.0	0.0
	Prediction for cell line U-251MG									
	Cent	Cyto	ER	Golgi	Mito	Nuclei	Nucleoli	PM	Cytoskel	Vesicles
Cent	0.0	10.5	0.0	0.0	2.0	87.5	0.0	0.0	0.0	0.0
Cyto	0.3	15.3	0.6	0.3	1.2	80.7	0.2	0.5	0.2	0.8
ER	0.0	10.4	4.8	0.0	1.1	81.6	0.0	0.0	0.0	2.0
Golgi	0.0	10.8	0.6	0.7	1.7	84.8	0.5	0.9	0.0	0.0
Mito	0.0	9.7	0.7	0.0	21.0	67.8	0.0	0.9	0.0	0.0
Nuclei	0.2	12.4	0.4	0.1	1.3	84.1	0.1	0.6	0.1	0.8
Nucleoli	0.0	11.2	0.0	0.0	2.6	84.5	0.5	0.9	0.0	0.3
PM	0.5	16.0	0.6	0.5	0.6	77.2	0.5	3.6	0.6	0.0
Cytoskel	0.0	14.2	0.0	0.0	0.7	85.1	0.0	0.0	0.0	0.0
Vesicles	0.0	15.0	0.0	0.4	1.5	82.5	0.0	0.5	0.0	0.0
	Prediction for cell line U-2 OS									
	Cent	Cyto	ER	Golgi	Mito	Nuclei	Nucleoli	PM	Cytoskel	Vesicles
Cent	0.0	90.8	0.0	0.0	0.0	7.5	0.0	0.0	0.0	1.7
Cyto	0.0	81.4	0.3	0.3	1.6	15.1	0.2	0.2	0.2	0.8
ER	0.0	80.8	3.5	0.0	0.0	13.7	0.0	0.0	1.0	1.0
Golgi	0.0	73.3	0.0	0.0	1.6	19.7	0.6	3.9	0.0	0.9
Mito	0.0	70.8	0.7	0.0	20.0	6.8	0.0	0.9	0.0	0.7
Nuclei	0.0	72.8	0.1	0.0	1.6	24.3	0.1	0.5	0.1	0.5
Nucleoli	0.0	81.1	0.0	0.5	1.5	16.0	0.5	0.0	0.0	0.4
PM	0.0	86.0	0.4	0.0	1.0	10.4	0.0	0.9	0.4	0.9
Cytoskel	0.0	87.3	0.0	1.0	1.2	9.9	0.0	0.7	0.0	0.0
Vesicles	0.0	80.2	0.0	0.5	2.0	14.4	0.0	1.0	0.0	1.9

Table 4.2: Confusion matrix of our pathway model in three cell lines A-431, U-251MG, and U-2 OS. Prediction is based on two feature types of known motifs: sequence annotation in UniProt and regular expression in Minimotif Miner.

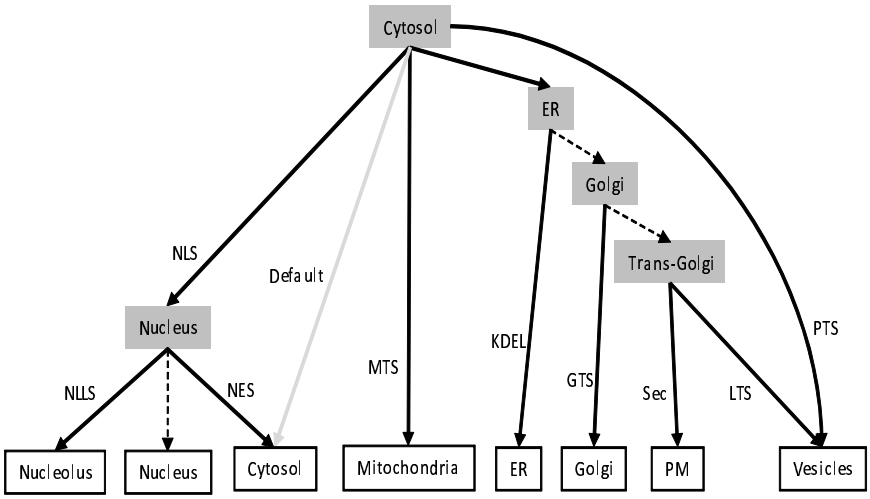


Figure 4.4: Protein sorting pathways collected from the literature. Each pathway is a path from cytosol to a compartment at the bottom, consisting of one or more steps (the links) that transport proteins between intermediate locations. Each step has a list of carriers and motifs responsible for the transportation by which we can verify whether the pathway is recovered. Dashed lines denote steps taken by default without specific carriers.

conserved and the few classical carriers (e.g. Importin proteins) are known to perform the same function. However none of these carriers have enough interactions present on BiG to be selected due to insufficient PPI information in human. Unable to validate the learned structure based on PPI, we rely on motif information in the literature to validate the learned structure instead. Fortunately the sequence annotation features provide more classical protein sorting motifs which are the key components of these pathways. Minimotif Miner also provides reference of involvement in sorting pathways for several motif features (i.e. the regular expressions). For example, using the Importin proteins we can validate the nuclear import pathway being recovered in yeast. Validation of the recovery of this pathway in human must rely on either the NLS motif in UniProt sequence annotation, or the corresponding regular expressions in Minimotif Miner, but not interaction with Importin.

The same method described in the previous chapter is also applied to examine the recovery of a specific pathway. A step in a literature pathway is considered recovered if there is a state on a path from the root to the destination that emits any motif in that step; a pathway is partially recovered if only some of its steps are recovered.

The pathway recovery results for different feature sets of features are listed in Table 4.3. This validation is based on known motif we collected from the literature. When the feature

Table 4.3: The number of pathways recovered out of 10 pathways based on different feature sets. The results are averaged over three cell lines and 10 folds. Minimum and maximum are also shown (best possible result would be 10). Fractions represent partial matches.

Features	Pathway recovery
HMM MnM	4.0 (2.5 - 6.0)
HMM Anno	6.3 (5.5 - 7.5)
HMM Anno + MnM	7.9 (7.0 - 9.0)
HMM BiG + Anno + MnM + GenHMM b	7.9 (6.5 - 9.5)
HMM BiG + Anno + MnM + DiscHMM b	7.9 (6.5 - 9.5)
HMM BiG + Anno + MnM + GenHMM	8.4 (7.5 - 8.5)
HMM BiG + Anno + MnM + DiscHMM	8.5 (8.0 - 8.5)

set only contains protein sorting motifs in MiniMotif Miner, our method is able to recover on average 40% of the pathways are recovered. When the feature set only contains sequence annotation, 63% are recovered, and using both 79% can be recovered. Using all features our method can recover about 85% of the known pathways. Again, a pathway might be recovered that we are not aware of, because the carrier or motif used is not in our collection.

4.4.3 Visualizing Differences in Sorting Pathways Learned from Localization in Three Cell Lines

We show a representative set of learned structures in Figure 4.5 (A-431), 4.6 (U-251MG) and 4.7 (U-2 OS). The relationship between compartments basically agrees with the established knowledge of protein sorting. The nuclei and cytosol share a path; compartments on the secretory pathway share several states as well, especially the state emitting the GPI signal sequence; within the secretory pathway Golgi is closer to PM. Because of our cell line specific structure search algorithm, we can match the common states in different cell lines to those in the common structure. Using this matching the differences in transitions and emissions in each cell line can be compared and displayed in the figure (marked by the thickness of lines). By this representation one can easily spot states, transitions and emissions common to all three cell lines, as well as cell line specific ones. Interestingly, in three cell lines our method added a state not previously learned the common subset (the thin block), but we can see that it correspond to a shared pathway unchanged between cell lines. On the other hand there are several transitions unique to one cell line or absent in one cell line. For example only in cell line U-2 OS there is a transition from the the secretory pathway to vesicles, and the transition from the secretory pathway to cytoskeleton is

absent in cell line U-251MG. It would be interesting to investigate whether such differences correspond to novel or known differential regularization in a specific cell line, but this is beyond the scope of this thesis.

4.5 Discussion

We have extended our targeting pathway model from yeast to human. The method supports alternative splicing which is common in higher organisms. The two phase structure search algorithm can utilize localization data spanning multiple cell lines, or potentially different cell types and conditions. It enables us to examine common and condition-specific carriers, motifs, and pathways. Using the extended model, we performed the first systematic discovery of targeting pathways in the human proteome based on confocal microscopy images on HPA. By comparing to a classifier without using a structure we show that incorporating the targeting pathways leads to more accurate prediction of the destinate compartment. The learned structure recovered about 85% of classical pathways we collected from the literature. The learned structure resembles our knowledge of protein sorting in the cell. Our cell line specific structure search algorithm enables visualization of the differences in sorting pathways between three cell lines, highlighting transitions unique to a cell line or absent in a cell line. For future work it would be interesting to examine whether such differences are related to unique properties of these cell lines. We would also like to investigate why all three structures learned from different cell lines added a similar state that should have been created in the common structure, for example try more random initialization or run more iterations (since the common structure is the basis for further structure search). Another possibility is that the common subset only has about half of the proteins, resulting in BIC choosing a simpler structure. We could try including the proteins that change locations between cell lines in the common subset but adjust the uncertainty of multiple localization. Our method can be applied to any conditions (e.g. diseases, drug effect, or different tissues). The inferred pathways, motifs and carriers can be tested experimentally as described in the previous chapter. We aim to further examine if we have discovered novel pathways in human.

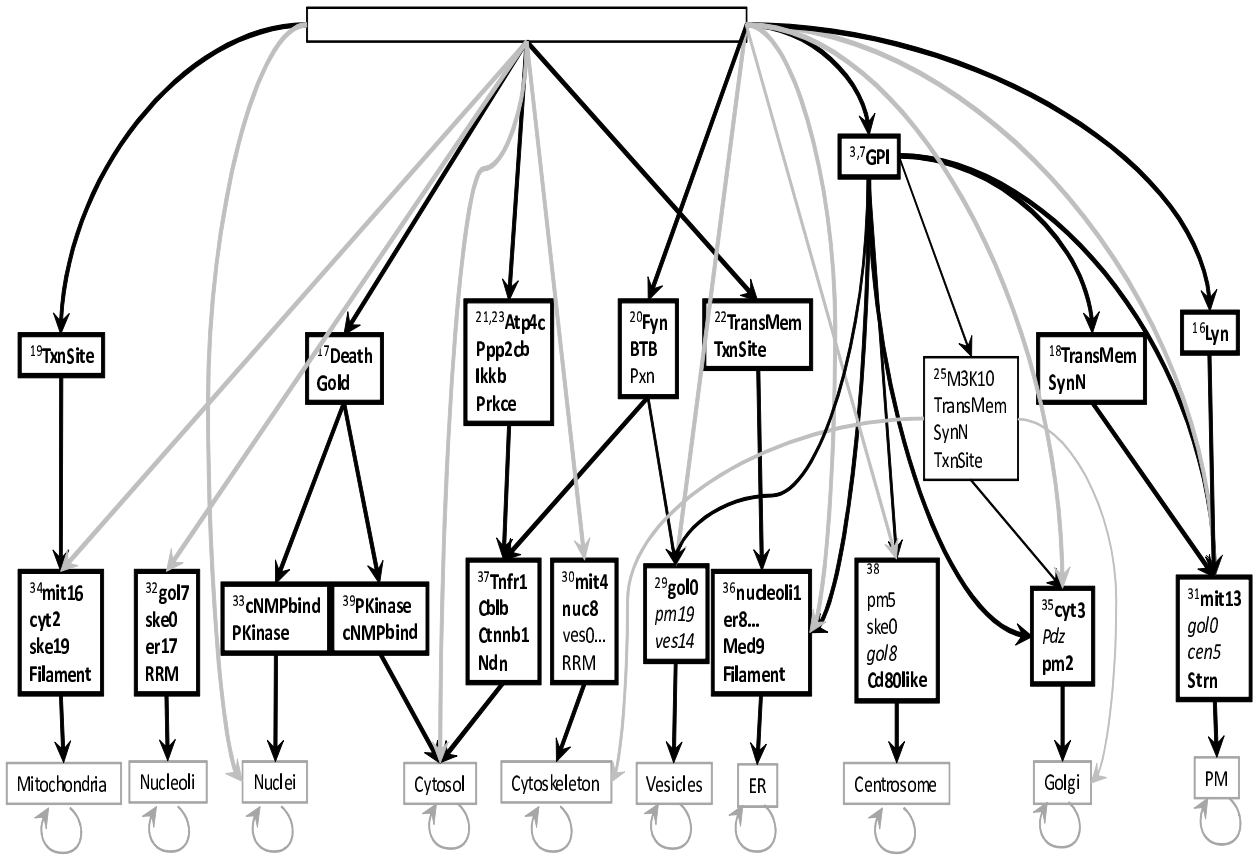


Figure 4.5: A representative HMM state space structure learned by our method that corresponds to potential protein targeting pathways in human cell line A-431. A state is represented by a block; its transitions are shown as arrows and its top 3 or 4 emitting features are listed inside the block. The bottom states (in gray) are the final destination compartments. Transitions across more than one level are shaded. Thin lines and blocks are specific to this cell line, and the thickest ones are shared among all three cell lines. Emissions in bold are shared among three cell lines, those in italic are shared in two cell lines, and others are specific to one cell line. Transitions across more than one levels are colored in gray for clarity.

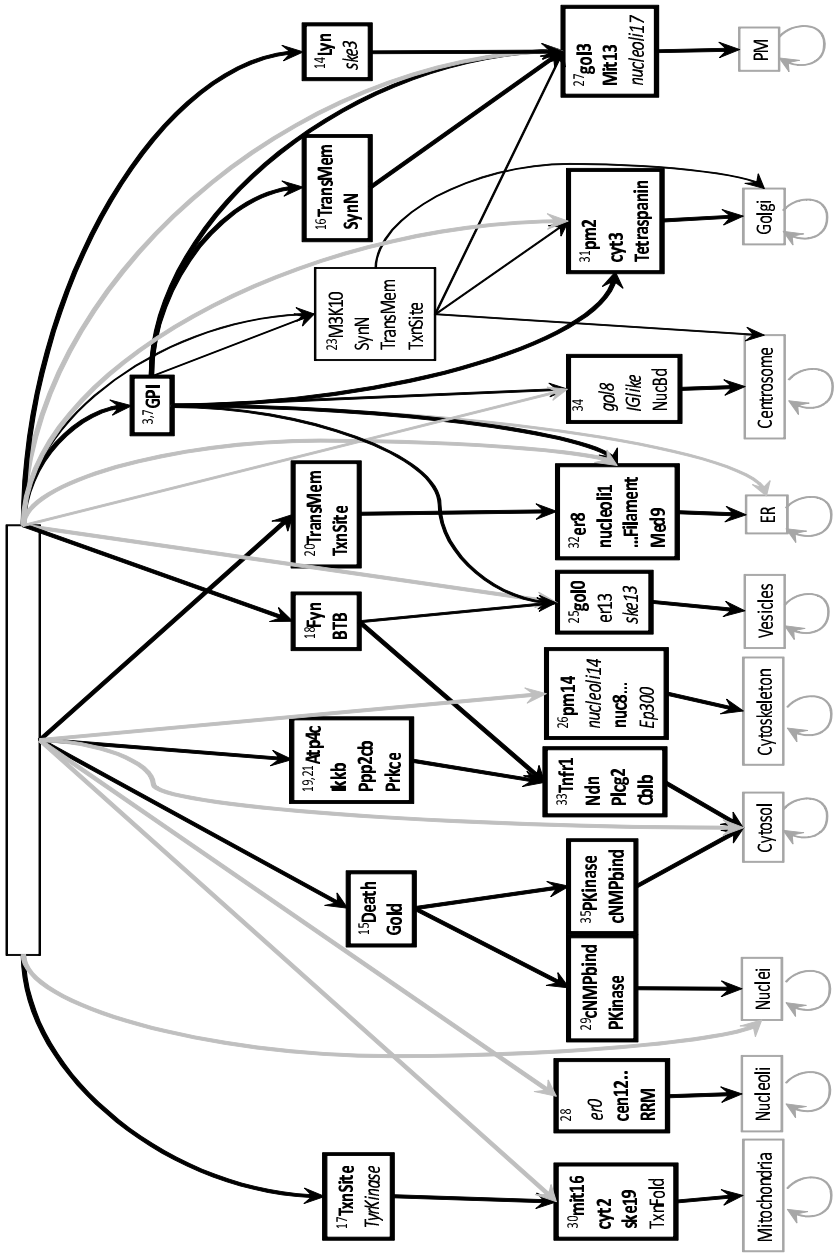


Figure 4.6: The HMM state space structure learned from localization data in cell line U-251MG. See Figure 4.5 for legend.

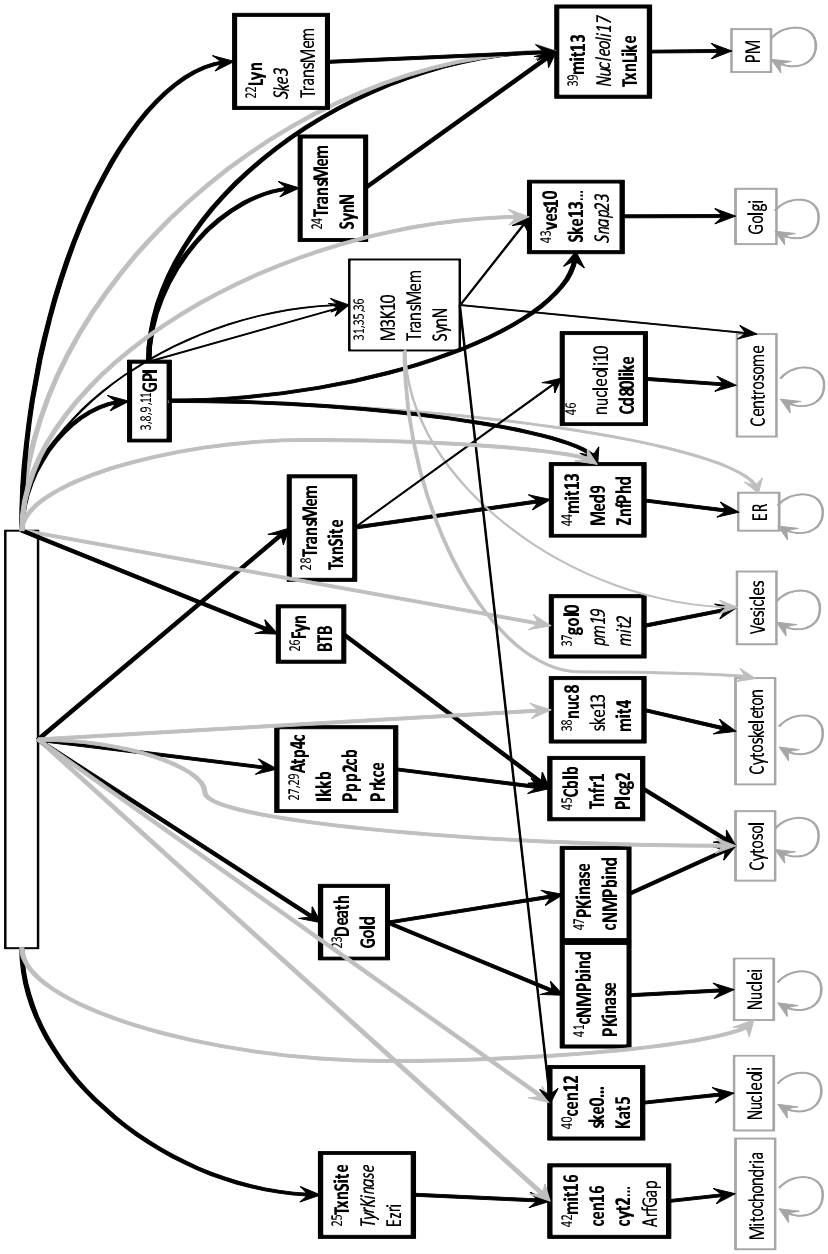


Figure 4.7: The HMM state space structure learned from localization data in cell line U-2 OS. See Figure 4.5 for legend.

Chapter 5

Conclusions and Future Work

In the previous chapters, we developed computational tools to study the protein sorting mechanism in different aspects. The proposed methods and the learned pathway structure provided better understanding of the protein sorting pathways in yeast and human.

This chapter summarizes the contributions and conclusions of this thesis (section 5.1) and points out potential future work (section 5.2).

5.1 Conclusions

Our goal is to learn novel cellular sorting pathways from proteomic data. For this purpose we first focused on extracting novel sorting motifs using the established sorting pathways (Chapter 2). These novel motifs served as features for discovering novel sorting pathways (in this case, not utilizing established pathways but using a flat structure instead). Combining the novel motifs with known motifs and PPI, we proposed a method to systematically learn novel sorting pathways in yeast (Chapter 3). We extended the method to higher organisms and applied to human data (Chapter 4). The highlights of each chapter are described as follows.

In Chapter 2 we developed the generative and discriminative HMM motif finding methods, and applied them to extract sorting motifs. Discriminative motif finding searches for motifs that are present in a compartment but absent in other, nearby, compartments by utilizing a tree structure that mimics the protein sorting pathways. We showed that both discriminative motif finding and the tree structure improve localization prediction on a benchmark dataset of yeast proteins. The motifs identified can be mapped to known

sorting motifs and the motif instances were found to be more conserved than the average protein sequence. Using our motif-based predictions we have identified potential annotation errors in public databases for the location of some of the proteins.

Instead of relying on the established sorting pathways, we developed a method to learn the pathway structure without prior knowledge on sorting pathways in Chapter 3. We use a HMM that naturally simulates the transportation of a protein among unobserved intermediate states; the path was determined by protein carriers and sequence motifs, based on the actual protein sorting mechanism. Our method relied on the co-occurrence of interacting partner and sequence motif to infer the structure. We believe that this is the first method to systematically learn sorting pathways from proteomic data. In simulation studies, the method has accurately recovered the underlying sorting models. Using the yeast GFP dataset, we showed that our model leads to accurate prediction of subcellular localization. We also showed that the pathways learned by our model recover many known sorting pathways and assign proteins following both classical and alternative pathways according to the literature to the correct path. The learned model identified new pathways and their putative carriers and motifs and these may represent novel protein sorting mechanisms. To experimentally validate potentially novel pathways efficiently, we have derived a list of highly confident prediction of protein sorting carriers and motifs.

In Chapter 4 we further extended our method to support higher organisms. Our method is applied to localization data from HPA based on confocal microscopy images on HPA. Two new issues common in higher organisms are addressed: alternative splicing and multiple cell lines. Our method learned cell line specific pathways which can be easily spot when visualized on each pathway structure. The learned structure led to accurate prediction of protein localization, and recovered about 85% of the classical pathways collected from the literature.

While most of the previous works only predict the final destination, we have addressed the question of where and why a protein goes to a specific location, and by what route. One key component of the protein sorting mechanism is sequence motif. When utilizing the classical sorting pathways, discriminative HMM has indeed extracted conserved sorting motifs that are informative enough to give good prediction accuracy. When we assume no prior pathway knowledge, these novel motifs are not enough and other features including PPI and known motifs are necessary. In yeast, the discriminative motifs still help improve prediction accuracy and pathway recovery, and are more informative than generative motifs

based on our evaluation. On the other hand, generative and discriminative HMMs are not as useful in human data comparing to UniProt sequence annotation that covers a wide range of protein properties. It is possible that the human protein sequences are more heterogeneous, making motif finding difficult. In the future we would try enhancing our motif finder to tackle the heterogeneity in large sequence dataset.

Another key component of the protein sorting mechanism is the interaction with a protein carrier. PPI information is critical for learning sorting pathways in yeast, but not so in human. This is certainly due to the lack of PPI information in human, but also because the budding yeast is the species that has the richest resource in PPI. The known motifs, especially UniProt sequence annotation, are the most informative features for human data. Some of the sequence annotations are based on experimental results and have better specificity comparing to motif scanning using regular expression or profile HMM. These annotation features are distinguished from those based on computational prediction which are marked as "potential". The later has lower specificity but wider coverage, and our model could utilize such differences because these are different features. These advantages provided by UniProt may explain the importance of sequence annotation in human data. As discussed above, the novel motifs extracted from sequence is not as useful in human as well in terms of prediction accuracy, because the sequence annotation already has a good coverage and human protein sequences are more heterogeneous.

By combining sequence motifs and the carrier proteins that recognize them (inferred from PPI), our results demonstrated that we can learn sorting pathways successfully in yeast and human. The sorting pathways HMM predicted the destination accurately and recovered the majority of known sorting pathways. Comparing to other systems that focus on prediction, we showed that a generative model that simulates the unobserved path a protein takes within the cell better explains the dynamic process of protein sorting.

5.2 Future Work

While the method has been successful, an HMM-based approach also suffers from a number of limitations. The input data used by our method is static while HMM expects sequential data. This requires us to rely on a number of assumptions including limiting each of the features to a unique level, and assuming independence between the features. The structure search algorithm requires substantial computation since the EM algorithm must be run

every time a candidate structure is being tried. Improving the search strategy is a direction for future work.

5.2.1 Physical Location of Intermediate States

Another issue we wish to address in the future is the inference of the actual location of the intermediate states. For example, we might associate an internal state with the ER or Golgi. To determine such locations, we will start with matching the learned structure to known pathways in the literature, which is part of our evaluation procedure. Because physical locations of the entire pathways have been described in the literature, we can assign the locations to the states that correspond to known pathways accordingly. Ideally the locations of these states should be uniquely determined, but there may be conflicts when a state is assigned to more than one location. How many conflicts exist would be the first issue to investigate. The assignment described above is limited to states that correspond to known pathways, not the more interesting novel pathways, but this assignment is highly confident when there is no conflict.

We also want to assign locations to other states even though there are uncertainties. When a state is matched to a transportation step on a known pathway, there may be other states on the path between this state and the destination. There are less uncertainty when there is only one such path. We can assign these states to the intermediate locations of the known pathway as well. For example, consider the case where a known pathway indicates transportation from ER to Golgi, and we already matched one state to the step of transportation into ER. If there is only one path from this state to Golgi, then the states in between are either in ER or Golgi. Again, this is limited to states on the path that correspond to a known pathway, only that the state (i.e. the emission of this state) does not match the known pathway directly.

Assigning locations to other states not described above is more difficult. We could list all final destinations of proteins in the training data that pass through a state. For example, 70% of the proteins passing through a state go to the nuclei and 30% go to the cytosol. This will be an informative reference for investigation, but there is no guarantee of the physical location even if 100% of the proteins go to a specific compartment.

5.2.2 Alternative Splicing

Although our method supports alternative splicing in higher organisms, there is no attempt to study whether alternative splicing could affect protein sorting. With our model it is possible to predict whether different isoforms will result in different localization. That is, if a protein sorting motif occurs in a cassette exon on one protein, the predicted location may be different between the isoforms. Specifically, both the novel motifs extracted from sequences and known motifs based on sequence annotation in UniProt [18] and regular expression in Minimotif Miner [57] already indicate the position of the motif instances. Combining with isoform information on Ensembl [93], it is straightforward to derive a list of predicted locations of each isoform. Such a list of isoform-dependent localization will be a unique contribution to understanding protein sorting. However, the relation between isoform and location is not included in HPA or in any database that we are aware of, making it difficult to validate this result. Since we may not have isoform-dependent location as training data, we could verify whether the locations of different isoform match the multiple locations. For example, if the model predicts that isoform 1 is targeted to cytosol but isoform 2 is targeted to PM, we could check whether the multiple locations are cytosol and PM. For proteins that are predicted to change location due to alternative splicing, if the prediction matches the multiple locations, they can be further investigated.

5.2.3 Model Identifiability Issue

Unfortunately in simulation studies we find that a structure with some difference from the true underlying structure can still achieve 100% accuracy. By using BIC score, we are making the assumption that a simpler structure is more preferable. To further resolve the identifiability issue, a specific post-processing procedure that eliminates duplicated states, transitions and emissions would be helpful. We would like to further investigate the design and application of such post-processing procedure in the future.

In Figure 3.4 (C) and (D), internal cross-validation achieved similar prediction accuracy as BIC does, but the learned structures sometimes do not resemble the true structure. A closer examination revealed that internal cross-validation added too many states. This is also related to the model identifiability issue. The reason BIC results in structures close to the true structures in the simulation study might be that the true structures are simple (at most 31 states). We tested internal cross-validation on human data using some of the

feature sets, and prediction accuracy is similar (result not shown). Because of the heavy computational cost we always end the search in 20 iterations for both BIC and internal cross-validation, which may limit the difference between the two as well. In the future if we can enhance the speed, we would like to perform further tests on internal cross-validation.

5.2.4 Combining with Unsupervised Learning of Locations from Images

With the availability of unsupervised learning of unmixing subcellular patterns of different locations [94], it would be even possible to perform such analysis without relying on human categorization of the patterns. This would enable learning new knowledge from the growing subcellular image collection, without relying on subjective manual annotation.

Bibliography

[1] W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea, "Global analysis of protein localization in budding yeast." *Nature*, vol. 425, no. 6959, pp. 686–691, Oct 2003.

[2] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction," *Proteins*, vol. 63, no. 3, pp. 490–500, 2006.

[3] T. R. Kau, J. C. Way, and P. A. Silver, "Nuclear transport and cancer: from mechanism to intervention." *Nat Rev Cancer*, vol. 4, no. 2, pp. 106–117, Feb 2004.

[4] A. B. Gladden and A. A. Diehl, "Location, location, location: the role of cyclin D1 nuclear localization in cancer." *Journal of cellular biochemistry*, vol. 96, no. 5, pp. 906–913, December 2005.

[5] B. De Strooper, M. Beullens, B. Contreras, L. Levesque, K. Craessaerts, B. Cordell, D. Moechars, M. Bollen, P. Fraser, P. St. George-Hyslop, and F. Van Leuven, "Phosphorylation, subcellular localization, and membrane orientation of the Alzheimer's disease-associated presenilins," *Journal of Biological Chemistry*, vol. 272, no. 6, pp. 3590–3598, February 1997.

[6] P. E. Purdue, Y. Takada, and C. J. Danpure, "Identification of mutations associated with peroxisome-to-mitochondrion mistargeting of alanine/glyoxylate aminotransferase in primary hyperoxaluria type 1," *J. Cell Biol.*, vol. 111, no. 6, pp. 2341–2351, December 1990.

[7] W. R. Skach, "Defects in processing and trafficking of the cystic fibrosis transmembrane conductance regulator," *Kidney International*, vol. 57, no. 3, pp. 825–831, March 2000.

[8] A. A. Cohen, N. Geva-Zatorsky, E. Eden, M. Frenkel-Morgenstern, I. Issaeva, A. Sigal, R. Milo, C. Cohen-Saidon, Y. Liron, Z. Kam, L. Cohen, T. Danon, N. Perzov, and U. Alon, "Dynamic Proteomics of Individual Cancer Cells in Response to a Drug," *Science*, vol. 322, no. 5907, pp. 1511–1516, December 2008.

[9] W. Yan, R. Aebersold, and E. W. Raines, "Evolution of organelle-associated protein profiling." *J Proteomics*, Dec 2008.

[10] T. Kislinger, B. Cox, A. Kannan, C. Chung, P. Hu, A. Ignatchenko, M. S. Scott, A. O. Gramolini, Q. Morris, M. T. Hallett, J. Rossant, T. R. Hughes, B. Frey, and A. Emili, "Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling." *Cell*, vol. 125, no. 1, pp. 173–186, April 2006.

[11] V. Starkuviene, U. Liebel, J. C. Simpson, H. Erfle, A. Poustka, S. Wiemann, and R. Pepperkok, "High-content screening microscopy identifies novel proteins with a putative role in secretory membrane traffic." *Genome Res*, vol. 14, no. 10A, pp. 1948–1956, Oct 2004.

[12] R. N. Aturaliya, J. L. Fink, M. J. Davis, M. S. Teasdale, K. A. Hanson, K. C. Miranda, A. R. R. Forrest, S. M. Grimmond, H. Suzuki, M. Kanamori, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and R. D. Teasdale, "Subcellular localization of mammalian type II membrane proteins." *Traffic*, vol. 7, no. 5, pp. 613–625, May 2006.

[13] E. G. Osuna, J. Hua, N. W. Bateman, T. Zhao, P. B. Berget, and R. F. Murphy, "Large-scale automated analysis of location patterns in randomly tagged 3T3 cells." *Ann Biomed Eng*, vol. 35, no. 6, pp. 1081–1087, Jun 2007.

[14] L. Barbe, E. Lundberg, P. Oksvold, A. Stenius, E. Lewin, E. Björling, A. Asplund, F. Pontén, H. Brismar, M. Uhlén, and H. A. Svahn, "Toward a confocal subcellular atlas of the human proteome." *Mol Cell Proteomics*, vol. 7, no. 3, pp. 499–508, Mar 2008.

[15] S. C. Chen, T. Zhao, G. J. Gordon, and R. F. Murphy, "Automated image analysis of protein localization in budding yeast." *Bioinformatics*, vol. 23, no. 13, pp. i66–i71, Jul 2007.

[16] J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, and D. Botstein, "SGD: Saccharomyces genome database." *Nucleic acids research*, vol. 26, no. 1, pp. 73–79, January 1998.

[17] G. Grumbling, V. Strelets, and T. F. Consortium, "FlyBase: anatomical data, images and queries," *Nucl. Acids Res.*, vol. 34, no. suppl_1, pp. D484–488, January 2006.

[18] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. Su, "The Universal Protein Resource (UniProt)." *Nucleic Acids Res*, vol. 33, no. Database issue, pp. D154–D159, Jan 2005.

[19] I. K. H. Poon and D. A. Jans, "Regulation of nuclear transport: central role in development and transformation?" *Traffic*, vol. 6, no. 3, pp. 173–186, Mar 2005.

[20] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, L. Zipursky, and J. Darnell, *Molecular Cell Biology*, fifth edition ed. W. H. Freeman, August 2003.

[21] J. Y. Newberg, J. Li, A. Rao, F. Pontén, M. Uhlén, E. Lundberg, and R. F. Murphy, "Automated Analysis Of Human Protein Atlas Immunofluorescence Images," *Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging*, pp. 1023–1026, 2009.

[22] P. Horton, K. J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. A. Collier, and K. Nakai, "WoLF PSORT: protein localization predictor." *Nucleic Acids Res*, vol. 35, no. Web Server issue, pp. W585–W587, Jul 2007.

[23] 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, Jul 2000.

[24] R. Nair and B. Rost, "Mimicking cellular sorting improves prediction of subcellular localization." *J Mol Biol*, vol. 348, no. 1, pp. 85–100, Apr 2005.

[25] M. S. Scott, S. J. Calafell, D. Y. Thomas, and M. T. Hallett, "Refining protein subcellular localization." *PLoS Comput Biol*, vol. 1, no. 6, Nov 2005.

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

[27] Y.-Q. Shen and G. Burger, "'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools," *BMC Bioinformatics*, vol. 8, pp. 420+, October 2007.

[28] 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, Jul 2004.

[29] J. D. Bendtsen, L. J. Jensen, N. Blom, G. Von Heijne, and S. Brunak, "Feature-based prediction of non-classical and leaderless protein secretion." *Protein Eng Des Sel*, vol. 17, no. 4, pp. 349–356, April 2004.

[30] 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. to appear, 2009.

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

[32] P. Horton and K. Nakai, "A probabilistic classification system for predicting the cellular localization sites of proteins." *Proc Int Conf Intell Syst Mol Biol*, vol. 4, pp. 109–115, 1996.

[33] M. Rashid, S. Saha, and G. P. Raghava, "Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs." *BMC Bioinformatics*, vol. 8, p. 337, 2007.

[34] H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano, "Extensive feature detection of N-terminal protein sorting signals." *Bioinformatics*, vol. 18, no. 2, pp. 298–305, Feb 2002.

[35] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle,

U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. G. Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. E. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. A. Sigrist, R. Vaughan, and E. M. Zdobnov, "The InterPro database, 2003 brings increased coverage and new features." *Nucleic Acids Res*, vol. 31, no. 1, pp. 315–318, Jan 2003.

[36] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, "MEME: discovering and analyzing DNA and protein sequence motifs." *Nucleic Acids Res*, vol. 34, no. Web Server issue, pp. W369–W373, Jul 2006.

[37] S. R. Eddy, "Profile hidden Markov models." *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.

[38] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman, "The Pfam protein families database." *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D281–D288, Jan 2008.

[39] C. Dingwall, J. Robbins, S. M. Dilworth, B. Roberts, and W. D. Richardson, "The nucleoplasmin nuclear location sequence is larger and more complex than that of SV-40 large T antigen." *J Cell Biol*, vol. 107, no. 3, pp. 841–849, Sep 1988.

[40] S. Subramanian, P. S. Sijwali, and P. J. Rosenthal, "Falcipain cysteine proteases require bipartite motifs for trafficking to the Plasmodium falciparum food vacuole." *J Biol Chem*, vol. 282, no. 34, pp. 24 961–24 969, Aug 2007.

[41] M. Doğruel, T. A. Down, and T. Jp, "NestedMICA as an ab initio protein motif discovery tool." *BMC bioinformatics*, vol. 9, pp. 19+, January 2008.

[42] P. Sumazin, G. Chen, N. Hata, A. D. Smith, T. Zhang, and M. Q. Zhang, "DWE: discriminating word enumerator." *Bioinformatics*, vol. 21, no. 1, pp. 31–38, January 2005.

[43] E. Segal, R. Yelensky, and D. Koller, "Genome-wide discovery of transcriptional modules from DNA sequence and gene expression." *Bioinformatics*, vol. 19 Suppl 1, pp. i273–i282, 2003.

[44] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, October 1992.

[45] T. L. Bailey and C. Elkan, "The value of prior knowledge in discovering motifs with MEME." *Proc Int Conf Intell Syst Mol Biol*, vol. 3, pp. 21–29, 1995.

[46] Y. Barash, G. Bejerano, and N. Friedman, "A Simple Hyper-Geometric Approach for Discovering Putative Transcription Factor Binding Sites," 2001, pp. 278–293.

[47] R. Sharan and E. W. Myers, "A motif-based framework for recognizing sequence families." *Bioinformatics*, vol. 21 Suppl 1, pp. i387–i393, Jun 2005.

[48] E. Redhead and T. L. Bailey, "Discriminative motif discovery in DNA and protein sequences using the DEME algorithm." *BMC Bioinformatics*, vol. 8, p. 385, 2007.

[49] A. D. Smith, P. Sumazin, and M. Q. Zhang, "Identifying tissue-selective transcription factor binding sites in vertebrate promoters." *Proc Natl Acad Sci U S A*, vol. 102, no. 5, pp. 1560–1565, Feb 2005.

[50] A. D. Smith, P. Sumazin, D. Das, and M. Q. Zhang, "Mining ChIP-chip data for transcription factor and cofactor binding sites." *Bioinformatics*, vol. 21 Suppl 1, pp. i403–i412, Jun 2005.

[51] S. Sinha, "On counting position weight matrix matches in a sequence, with application to discriminative motif finding." *Bioinformatics*, vol. 22, no. 14, pp. e454–e463, Jul 2006.

[52] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 107–113, Jan 1991.

[53] Y. Normandin, R. Cardin, and R. De Mori, "High-performance connected digit recognition using maximum mutual information estimation," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 299–311, April 1994.

[54] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, pp. 25–47, 2002.

[55] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." *Nucleic Acids Res*, vol. 31, no. 1, pp. 365–370, Jan 2003.

[56] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, Eds. MIT-Press, 1999.

[57] S. Balla, V. Thapar, S. Verma, T. Luong, T. Faghri, C. H. Huang, S. Rajasekaran, J. J. del Campo, J. H. Shinn, W. A. Mohler, M. W. Maciejewski, M. R. Gryk, B. Piccirillo, S. R. Schiller, and M. R. Schiller, "Minimotif Miner: a tool for investigating protein function." *Nat Methods*, vol. 3, no. 3, pp. 175–177, Mar 2006.

[58] E. M. Zdobnov and R. Apweiler, "InterProScan–an integration platform for the signature-recognition methods in InterPro." *Bioinformatics*, vol. 17, no. 9, pp. 847–848, Sep 2001.

[59] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu, "Assessing computational tools for the discovery of transcription factor binding sites." *Nat Biotechnol*, vol. 23, no. 1, pp. 137–144, Jan 2005.

[60] A. Drawid and M. Gerstein, "A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome." *J Mol Biol*, vol. 301, no. 4, pp. 1059–1075, Aug 2000.

[61] B. S. Böckler, J. Schultz, and S. Rahmann, "HMM logos for visualization of protein families." *BMC Bioinformatics*, vol. 5, p. 7, Jan 2004.

[62] S. J. Gould, G. A. Keller, and S. Subramani, "Identification of peroxisomal targeting signals located at the carboxy terminus of four peroxisomal proteins." *J Cell Biol*, vol. 107, no. 3, pp. 897–905, Sep 1988.

[63] L. R. Kowalski, K. Kondo, and M. Inouye, "Cold-shock induction of a family of TIP1-related proteins associated with the membrane in Saccharomyces cerevisiae." *Mol Microbiol*, vol. 15, no. 2, pp. 341–353, January 1995.

[64] R. Nair and B. Rost, "Sequence conserved for subcellular localization." *Protein Sci*, vol. 11, no. 12, pp. 2836–2847, Dec 2002.

[65] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander, "Sequencing and comparison of yeast species to identify genes and regulatory elements." *Nature*, vol. 423, no. 6937, pp. 241–254, May 2003.

[66] P. Cliften, P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. Cohen, and M. Johnston, "Finding Functional Features in Saccharomyces Genomes by Phylogenetic Footprinting," *Science*, vol. 301, pp. 71–76, 2003.

[67] X. Pan, P. Roberts, Y. Chen, E. Kvam, N. Shulga, K. Huang, S. Lemmon, and D. S. Goldfarb, "Nucleus-vacuole junctions in saccharomyces cerevisiae are formed through the direct interaction of Vac8p with Nvj1p." *Mol Biol Cell*, vol. 11, no. 7, pp. 2445–2457, 2000.

[68] S. R. Eddy, G. Mitchison, and R. Durbin, "Maximum discrimination hidden Markov models of sequence consensus." *J Comput Biol*, vol. 2, no. 1, pp. 9–23, 1995.

[69] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites." *Protein Eng*, vol. 10, no. 1, pp. 1–6, Jan 1997.

[70] K.-C. C. Chou and H.-B. B. Shen, "Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms." *Nature protocols*, vol. 3, no. 2, pp. 153–162, January 2008.

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

[72] A. Pierleoni, P. L. Martelli, P. Fariselli, and R. Casadio, "BaCelLo: a balanced subcellular localization predictor." *Bioinformatics*, vol. 22, 2006.

[73] T. H. Lin, R. F. Murphy, and Z. B. Joseph, "Discriminative Motif Finding for Predicting Protein Subcellular Localization," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 2, pp. 441–451, 2011.

[74] H. Shatkay, A. Höglund, S. Brady, T. Blum, P. Dönnes, and O. Kohlbacher, "SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data." *Bioinformatics*, vol. 23, no. 11, pp. 1410–1417, June 2007.

[75] A. Rubartelli and R. Sitia, "Secretion of mammalian proteins that lack a signal sequence," *Unusual Secretory Pathways: From Bacteria to Man. Austin, TX: RG Landes*, pp. 87–104, 1997.

[76] D. Ruths, L. Nakhleh, and P. T. Ram, "Rapidly exploring structural and dynamic properties of signaling networks using PathwayOracle." *BMC systems biology*, vol. 2, 2008.

[77] G. Bebek and J. Yang, "PathFinder: mining signal transduction pathway segments from protein-protein interaction networks." *BMC bioinformatics*, vol. 8, pp. 335+, September 2007.

[78] J. Scott, T. Ideker, R. M. Karp, and R. Sharan, "Efficient algorithms for detecting signaling pathways in protein interaction networks." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 13, no. 2, pp. 133–144, March 2006.

[79] J. Dale, L. Popescu, and P. Karp, "Machine learning methods for metabolic pathway prediction," *BMC Bioinformatics*, vol. 11, no. 1, pp. 15+, January 2010.

[80] E. Fischer and U. Sauer, "Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism," *Nature Genetics*, vol. 37, no. 6, pp. 636–640, May 2005.

[81] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson, "Integrating high-throughput and computational data elucidates bacterial networks," *Nature*, vol. 429, no. 6987, pp. 92–96, May 2004.

[82] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, corrected ed. Springer, July 2003.

[83] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[84] X. Gao, B. Xiao, D. Tao, and X. Li, "A survey of graph edit distance," *Pattern Analysis & Applications*, vol. 13, no. 1, pp. 113–129, February 2010.

[85] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets." *Nucleic acids research*, vol. 34, no. Database issue, pp. D535–D539, January 2006.

[86] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 1999, pp. 42–49.

[87] C. J. Van Rijsbergen, *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979.

[88] L. Berglund, E. Björling, P. Oksvold, L. Fagerberg, A. Asplund, C. A. K. Szigyarto, A. Persson, J. Ottosson, H. Wernérus, P. Nilsson, E. Lundberg, A. Sivertsson, S. Navani, K. Wester, C. Kampf, S. Hober, F. Pontén, and M. Uhlén, "A gene-centric Human Protein Atlas for expression profiles based on antibodies." *Mol Cell Proteomics*, vol. 7, no. 10, pp. 2019–2027, Oct 2008.

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

[90] E. Glory, J. Newberg, and R. F. Murphy, "Automated comparison of protein subcellular location patterns between images of normal and cancerous tissues," *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pp. 304–307, May 2008.

[91] K. N. Lazaridis, P. Tietz, T. Wu, S. Kip, P. A. Dawson, and N. F. LaRusso, "Alternative splicing of the rat sodium/bile acid transporter changes its cellular localization and transport properties," *Proceedings of the National Academy of*

Sciences of the United States of America, vol. 97, no. 20, pp. 11 092–11 097, September 2000.

[92] M. Nakao, R. A. Barrero, Y. Mukai, C. Motono, M. Suwa, and K. Nakai, "Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals," *Nucleic Acids Research*, vol. 33, no. 8, pp. 2355–2363, 2005.

[93] T. J. P. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek, "Ensembl 2009," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D690–D697, January 2009.

[94] T. Peng, G. M. C. Bonamy, E. Glory-Afshar, D. R. Rines, S. K. Chanda, and R. F. Murphy, "Determining the distribution of probes between different subcellular locations through automated unmixing of subcellular patterns," *Proceedings of the National Academy of Sciences*, vol. 107, no. 7, pp. 2944–2949, February 2010.