3-1-2011

Discriminative motif finding for predicting protein subcellular localization.

Tien-ho Lin
Carnegie Mellon University

Robert F. Murphy
Carnegie Mellon University, murphy@cmu.edu

Ziv Bar-Joseph
Carnegie Mellon University, zivbj@andrew.cmu.edu

Published In
IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8, 2, 441-451.

Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2011 ; 8(2): 441–451. doi:10.1109/TCBB.2009.82.

Discriminative Motif Finding for Predicting Protein Subcellular Localization

Tien-ho Lin,
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Robert F. Murphy[Senior Member, IEEE], and
Machine Learning Department and Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA., Departments of Biological Sciences and Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

Ziv Bar-Joseph
Machine Learning Department and Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

Many methods have been described to predict the subcellular location of proteins from sequence information. However, most of these methods either rely on global sequence properties or use a set of known protein targeting motifs to predict protein localization. Here we develop and test a novel method that identifies potential targeting motifs using a discriminative approach based on hidden Markov models (discriminative HMMs). These models search for motifs that are present in a compartment but absent in other, nearby, compartments by utilizing an hierarchical structure that mimics the protein sorting mechanism. We show that both discriminative motif finding and the hierarchical structure improves localization prediction on a benchmark dataset of yeast proteins. 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. A software implementation and the dataset described in this paper are available from http://murphylab.web.cmu.edu/software/2009_TCBB_motif/

Index Terms

hidden Markov models; maximal mutual information estimate; discriminative motif finding; protein localization

I. Introduction

An important challenge in systems biology is to build detailed models of cell organization that provide accurate predictions of cell behaviors. Since many (if not all) of the proteins expressed by a given cell are likely to require proper subcellular localization in order to make their contributions to those behaviors, systems for assigning locations to proteins are critically needed. The problem is further complicated by the fact that 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 oncogenesis and differentiation [1], [2]. This implies that assignment of proteins to a fixed location or set of locations will not be sufficient. Given that many proteins are found in only a specific region of an organelle and that these refined localizations are conserved across species, it is also not likely that assignments at the level of whole organelles will be sufficient either.

Extensive work has been done on proteome-scale determination of location at various levels of resolution both by fractionation [3] and by microscopy [4]–[7]. Since collection of information for tens of thousands of proteins for perhaps hundreds of cell or tissue types under many conditions may be infeasible, computational prediction of location is an important alternative. Furthermore, knowledge of the sequence elements within proteins that determine location will be useful for modeling how changes in location are regulated.

A number of methods have been proposed for using sequence information to predict localization. These include WoLF PSort [8], TargetP [9], LOCtree [10], PSLT2 [11], TBpred [12], and iPSORT [13]. 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 [14], but does not try to search for novel targeting motifs. Beside known motifs TargetP also uses the motif finder MEME [15] to characterize manually curated special cases of the mitochondrial targeting signal [9]. 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 [12], 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.

To address these problems we developed a new method for searching among proteins that are known to localize to a certain compartment for motifs that may govern their localization patterns. We represent protein sequence motifs by profile hidden Markov models (HMMs), which models local alignments using match, insert, and delete states [16]. Profile HMMs have been successfully utilized to model protein families and domains, and they are used to represent domains in the Pfam database [17]. Unlike position weight matrices (PWMs) (for example, those used by MEME [15]), profile HMMs allow for variable length insertions and deletions that are common in protein motifs, for example the nucleoplasmin nuclear location sequence [18] and the sequence targeting proteases to the food vacuole in *P. falciparum* [19]. 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 [15] and NestedMICA [20] using PWMs, and HMMER [16] using profile HMMs, among others. While 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 (e.g. profile HMMs). Discriminative methods search for motifs that are present in one class (positive set) but absent in other classes (negative set). Such methods have been applied to search for DNA binding motifs represented by PWM, including gradient ascent optimizing conditional likelihood [21], heuristic hill-climbing [22], and enumerative search on a discrete space of PWM [23]. Recently PWM is applied to protein motifs as well as DNA motifs in the DEME algorithm using a combination of substring search and conjugate gradient [24]. However PWM does not allow insertion and deletion, making it less than optimal for protein sequence analysis. 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.

Since subcellular localization is determined by a hierarchical protein sorting process, several methods have been developed that utilize a hierarchical structure to improve prediction of protein localization [10], [25], [26]. 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 1. This allows us to take full advantage of biological knowledge regarding the organization of compartments within a cell.

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.

II. Discriminative motif finding

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 [16], and similarly PWMs are trained with the expectation-maximization (EM) algorithm, for example using MEME [15]. 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.

A. 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.

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 1) and the class labels of the sequences are given as $c_n \in \{1, 2, \cdots M\}$, $1 \le n \le N$. The HMM for the m-th class is denoted as $\lambda^{(m)}$. The MMIE objective function can be written as

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

The MMIE algorithm is an extended version of the Baum-Welch algorithm [27]. 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; see the appendix for more details.

We compare the results of our discriminative HMM model to generative training of HMMs and PWMs. To make the comparison fair, we make similar assumptions and use the same options for discriminative HMM, generative HMM, and PWM learned by MEME (see below). Implementation of the generative and discriminative HMM are based on the HMMER 2.3.2 source code [16], and compared to motif finding using MEME 3.2.1 [15]. Additional details may be found in the supplement.

B. Motif occurrences per sequence

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.

C. Motif finding on flat and hierarchical structure of compartments

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 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).

D. Retrieving top 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).

III. Predicting protein localization using identified motifs

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.

A. Ranking motifs found by different methods

To further evaluate the contribution for predicting locations, the motifs found by MEME, generative and discriminative HMM are ranked. 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.

IV. Results

A. Discriminative motif finding and hierarchical classification improves prediction accuracy

We applied our discriminative motif finding method to a yeast protein localization dataset [11]. This dataset consists of 1,521 *S. cerevisae* proteins with curated localization annotation in SwissProt [28]. 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 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 [10].

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.

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 [14]. The results for this prediction comparison are presented in Figure 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 I. 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 also compared these results to classifiers based on amino acid (AA) composition and determined that the discriminative HMM motif finding method outperforms these AA composition methods. This dataset contains only a few homologous proteins, about 6% of the sequences have >40% sequence identity by BLASTALL. Both MEME and HMM performed similarly after homology reduction and the improvement of HMM over PWM is still significant (see supplement).

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. Calculation of the confidence of each protein-compartment pair based on SVM margins is described in the supplement. The predictions and the confidence are on the support website (http://murphylab.web.cmu.edu/software/2009_TCBB_motif/).

B. Recovering known targeting 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.

1) 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.

To overcome these issues we collected a list of known targeting motifs from two databases, Minimotif Miner [29] and InterPro [14]. 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 [30]. 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.

2) 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 [31], [32]. 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 3. 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.

3) 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.

4) Logos for identified motifs—The 3 most discriminative motifs and the known motifs found by discriminative HMM using flat and hierarchical compartment structure are shown in Figure 4. The most discriminative motifs are defined by backward feature selection as described in previous section. Motifs are visualized using HMM logos [33]. In supplementary results we present the HMM logos for the top 20 motifs of both methods. 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 [34]. Discriminative HMM using hierarchical structure finds the stress-induced protein motif (SRP1/TIP1), also known to be associated with the membrane in yeast [35]. Known motifs are sometimes ranked very highly, as SRP1/TIP1 above, but not always. This observation 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 4 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 4 to have some false positives but high recall of valid NLS.

C. Identified motifs are highly conserved

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 [36].

1) 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 [37]. The default alignment result is used. Sequence and homology information were derived from integration of two previous comparative genomics studies [38], [39]. For these species amino acid sequence alignment was performed by ClustalW, and four conservation states were defined for each amino acid: no conservation versus weak, strong and identical conservation (across 7 species).

2) 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.

3) 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.

4) 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 5. 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.

D. Reannotating proteins

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 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. Four are discussed below and 4 more are discussed in the supplement.

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 6A).

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 6B).

Vac8/YEL013W—Vac8 is labeled as vacuole by human experts, but is also involved in nucleus-vacuole (NV) junctions [40]. This could be the reason that an image-based automated classifier [41] and all motif finders agree on Vac8 being localized to the nucleus.

V. 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 [42]. 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 our current method (like most motif finders) does not incorporate information on position relative to sequence

landmarks like the N- or C-terminus or cleavage sites. Thus it does not find elements, such as the signal peptide, that can be found using such alignments [43].

For future work, we would like to apply our method to more than 9 compartments, especially detailed subcompartments which were not well studied previously. We also aim to study higher eukaryotes, e.g. mouse and human, and ideally consider tissue-specific protein targeting. Since the discriminative criteria we use assumes all sequences have the motif, it may be more sensitive to incorrect annotations in training sequences [42]. Hence we also wish to make our sequence weighting more robust. This is particularly important in order to utilize high-throughput data on subcellular localization [4], [41], [44]. High-throughput experiments cover more proteins than curated datasets, but also contain more false positives and false negatives. Motif finding in these datasets would require better handling of incorrect annotations and using different confidence levels for different assignment depending on how well they are supported.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by a bioinformatics supplement to NIH National Technology Center for Networks and Pathways grant U54 RR022241 (Alan Waggoner). The authors would like to thank Sean Eddy for sharing the HMMER source code.

Appendix: The extended Baum-Welch algorithm for discriminative training of HMM

The objective function for discriminative training of HMM is the conditional likelihood of the correct class given the observed values,

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

where $\{O_1, O_2, \cdots O_N\}$ are the training sequences, $c_n \in \{1, 2, \cdots M\}$, $1 \le n \le N$ are the class labels of the sequences, and $\lambda^{(m)}$ is the HMM for the m-th class. 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 maximal mutual information estimate (MMIE) algorithm is an extended version of the Baum-Welch algorithm [27]. The E-step in extended Baum-Welch is the same as the one in the original Baum-Welch. The expected count of state j at position t of sequence n according to model $\lambda^{(m)}$ is denoted as $\gamma_{nt}^{(m)}(j)$. 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 $\xi_{nt}^{(m)}(i, j)$. 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 $\varphi_n^{(m)}(j, k)$, defined as

$$\xi_n^{(m)}(i,\ j) = \sum_t \xi_{nt}^{(m)}(i,\ j)$$
$$\varphi_n^{(m)}(j,\ k) = \sum_t \gamma_{nt}^{(m)}(j) 1_{y_{nt}=k}$$

The difference between generative and discriminative training is in the M-step. We will show that MMIE weights positive examples as the posterior probability of incorrect classification, $1 - p(\lambda^{(m)}|O_n)$, and weights negative examples 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 [27], [45]. The smoothing constants also prevent negative probabilities due to negative sequence weights. The reestimation formulas in the M-step of MMIE are,

$$\widehat{a}_{ij}^{(m)} \leftarrow \frac{\xi^{(m)}(i,\ j) - \xi^{(-m)}(i,\ j) + D_T a_{ij}^{(m)}}{\sum_{j'} \xi^{(m)}(i,\ j') - \xi^{(-m)}(i,\ j') + D_T a_{ij'}^{(m)}} \tag{1}$$

$$\widehat{b}_{jk}^{(m)} \leftarrow \frac{\varphi^{(m)}(j,\ k) - \varphi^{(-m)}(j,\ k) + D_E b_{jk}^{(m)}}{\sum_{k'} \varphi^{(m)}(j,\ k') - \varphi^{(-m)}(j,\ k') + D_E b_{jk'}^{(m)}} \tag{2}$$

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

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

Following [46] 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 [46].

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 [27],

$$\widehat{a}_{ij} \leftarrow \frac{a_{ij}^{(m)} \frac{\partial}{\partial a_{ij}^{(m)}} \mathcal{F}(\Lambda) + D_T a_{ij}^{(m)}}{\sum_{j'} a_{ij'}^{(m)} \frac{\partial}{\partial a_{ij}^{(m)}} \mathcal{F}(\Lambda) + D_T a_{ij'}^{(m)}} \tag{3}$$

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

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

Plugging in the above partial derivative to Equation 3 results in the sequence weighting update formula, Equation 1. Equation 2 can be derived in a similar way.

References

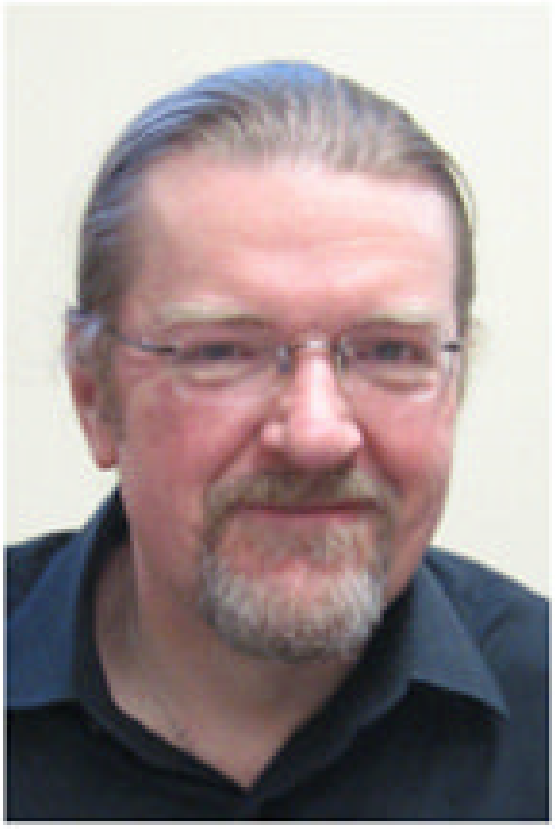
1. Kau TR, Way JC, Silver PA. Nuclear transport and cancer: from mechanism to intervention. Nat Rev Cancer Feb;2004 4(no. 2):106–117. [PubMed: 14732865]
2. Poon IKH, Jans DA. Regulation of nuclear transport: central role in development and transformation? Traffic Mar;2005 6(no. 3):173–186. [PubMed: 15702986]
3. Yan W, Aebersold R, Raines EW. Evolution of organelle-associated protein profiling. J Proteomics. Dec;2008
4. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. Global analysis of protein localization in budding yeast. Nature Oct;2003 425(no. 6959):686–691. [PubMed: 14562095]
5. Starkuviene V, Liebel U, Simpson JC, Erfle H, Poustka A, Wiemann S, Pepperkok R. High-content screening microscopy identifies novel proteins with a putative role in secretory membrane traffic. Genome Res Oct;2004 14(no. 10A):1948–1956. [PubMed: 15466293]
6. Aturaliya RN, Fink JL, Davis MJ, Teasdale MS, Hanson KA, Miranda KC, Forrest ARR, Grimmond SM, Suzuki H, Kanamori M, Kai C, Kawai J, Carninci P, Hayashizaki Y, Teasdale RD. Subcellular localization of mammalian type II membrane proteins. Traffic May;2006 7(no. 5):613–625. [PubMed: 16643283]
7. Osuna EG, Hua J, Bateman NW, Zhao T, Berget PB, Murphy RF. Large-scale automated analysis of location patterns in randomly tagged 3T3 cells. Ann Biomed Eng Jun;2007 35(no. 6):1081–1087. [PubMed: 17285363]
8. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. WoLF PSORT: protein localization predictor. Nucleic Acids Res Jul;2007 35(no. Web Server issue):W585–W587. [PubMed: 17517783]
9. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol Jul;2000 300(no. 4):1005–1016. [PubMed: 10891285]
10. Nair R, Rost B. Mimicking cellular sorting improves prediction of subcellular localization. J Mol Biol Apr;2005 348(no. 1):85–100. [PubMed: 15808855]
11. Scott MS, Calafell SJ, Thomas DY, Hallett MT. Refining protein subcellular localization. PLoS Comput Biol Nov;2005 1(no. 6):e66. [PubMed: 16322766]
12. Rashid M, Saha S, Raghava GP. Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics 2007;8:337. [PubMed: 17854501]
13. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S. Extensive feature detection of n-terminal protein sorting signals. Bioinformatics Feb;2002 18(no. 2):298–305. [PubMed: 11847077]

14. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJA, Vaughan R, Zdobnov EM. The InterPro database, 2003 brings increased coverage and new features. Nucleic Acids Res Jan;2003 31(no. 1):315–318. [PubMed: 12520011]
15. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res Jul;2006 34(no. Web Server issue):W369–W373. [PubMed: 16845028]
16. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14(no. 9):755–763. [PubMed: 9918945]
17. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, Bateman A. The Pfam protein families database. Nucleic Acids Res Jan;2008 36(no. Database issue):D281–D288. [PubMed: 18039703]
18. Dingwall C, Robbins J, Dilworth SM, Roberts B, Richardson WD. The nucleoplasmin nuclear location sequence is larger and more complex than that of sv-40 large t antigen. J Cell Biol Sep; 1988 107(no. 3):841–849. [PubMed: 3417784]
19. Subramanian S, Sijwali PS, Rosenthal PJ. Falcipain cysteine proteases require bipartite motifs for trafficking to the plasmodium falciparum food vacuole. J Biol Chem Aug;2007 282(no. 34):24 961–24 969.
20. Doruel M, Down TA, Hubbard TJ. NestedMICA as an ab initio protein motif discovery tool. BMC Bioinformatics 2008;9:19. [PubMed: 18194537]
21. Segal E, Yelensky R, Koller D. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics 2003;19 1:i273–i282. [PubMed: 12855470]
22. Sinha S. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. Bioinformatics Jul;2006 22(no. 14):e454–e463. [PubMed: 16873507]
23. Smith AD, Sumazin P, Das D, Zhang MQ. Mining ChIP-chip data for transcription factor and cofactor binding sites. Bioinformatics Jun;2005 21 1:i403–i412. [PubMed: 15961485]
24. Redhead E, Bailey TL. Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. BMC Bioinformatics 2007;8:385. [PubMed: 17937785]
25. Nakai K, Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics Dec;1992 14(no. 4):897–911. [PubMed: 1478671]
26. Horton P, Nakai K. A probabilistic classification system for predicting the cellular localization sites of proteins. Proc Int Conf Intell Syst Mol Biol 1996;4:109–115. [PubMed: 8877510]
27. Gopalakrishnan P, Kanevsky D, Nadas A, Nahamoo D. An inequality for rational functions with applications to some statistical estimation problems. Information Theory, IEEE Transactions on Jan;1991 37(no. 1):107–113.
28. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res Jan;2003 31(no. 1):365–370. [PubMed: 12520024]
29. Balla S, Thapar V, Verma S, Luong T, Faghri T, Huang CH, Rajasekaran S, del Campo JJ, Shinn JH, Mohler WA, Maciejewski MW, Gryk MR, Piccirillo B, Schiller SR, Schiller MR. Minimotif Miner: a tool for investigating protein function. Nat Methods Mar;2006 3(no. 3):175–177. [PubMed: 16489333]
30. Zdobnov EM, Apweiler R. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics Sep;2001 17(no. 9):847–848. [PubMed: 11590104]
31. Tompa M, Li N, Bailey TL, Church GM, Moor BD, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Rgnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol Jan; 2005 23(no. 1):137–144. [PubMed: 15637633]
32. Bailey TL, Elkan C. The value of prior knowledge in discovering motifs with meme. Proc Int Conf Intell Syst Mol Biol 1995;3:21–29. [PubMed: 7584439]

33. Schuster-Bckler B, Schultz J, Rahmann S. HMM logos for visualization of protein families. BMC Bioinformatics Jan;2004 5:7. [PubMed: 14736340]
34. Gould SJ, Keller GA, Subramani S. Identification of peroxisomal targeting signals located at the carboxy terminus of four peroxisomal proteins. J Cell Biol Sep;1988 107(no. 3):897–905. [PubMed: 2901422]
35. Kowalski LR, Kondo K, Inouye M. Cold-shock induction of a family of TIP1-related proteins associated with the membrane in *Saccharomyces cerevisiae*. Mol Microbiol Jan;1995 15(no. 2): 341–353. [PubMed: 7746155]
36. Nair R, Rost B. Sequence conserved for subcellular localization. Protein Sci Dec;2002 11(no. 12): 2836–2847. [PubMed: 12441382]
37. Saccharomyces Genome Database. 2008. http://db.yeastgenome.org/cgi-bin/FUNGI/showAlign
38. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature May;2003 423(no. 6937):241–254. [PubMed: 12748633]
39. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science Jul;2003 301(no. 5629):71–76. [PubMed: 12775844]
40. Pan X, Roberts P, Chen Y, Kvam E, Shulga N, Huang K, Lemmon S, Goldfarb DS. Nucleus-vacuole junctions in *Saccharomyces cerevisiae* are formed through the direct interaction of Vac8p with Nvj1p. Mol Biol Cell Jul;2000 11(no. 7):2445–2457. [PubMed: 10888680]
41. Chen SC, Zhao T, Gordon GJ, Murphy RF. Automated image analysis of protein localization in budding yeast. Bioinformatics Jul;2007 23(no. 13):i66–i71. [PubMed: 17646347]
42. Eddy SR, Mitchison G, Durbin R. Maximum discrimination hidden Markov models of sequence consensus. J Comput Biol 1995;2(no. 1):9–23. [PubMed: 7497123]
43. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng Jan;1997 10(no. 1):1–6. [PubMed: 9051728]
44. Ross-Macdonald P, Coelho PS, Roemer T, Agarwal S, Kumar A, Jansen R, Cheung KH, Sheehan A, Symoniatis D, Umansky L, Heidtman M, Nelson FK, Iwasaki H, Hager K, Gerstein M, Miller P, Roeder GS, Snyder M. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature Nov;1999 402(no. 6760):413–418. [PubMed: 10586881]
45. Normandin Y, Cardin R, De Mori R. High-performance connected digit recognition using maximum mutual information estimation. Speech and Audio Processing, IEEE Transactions on April;1994 2(no. 2):299–311.
46. Woodland P, Povey D. Large scale discriminative training of hidden Markov models for speech recognition. Computer Speech and Language 2002;16:25–47.

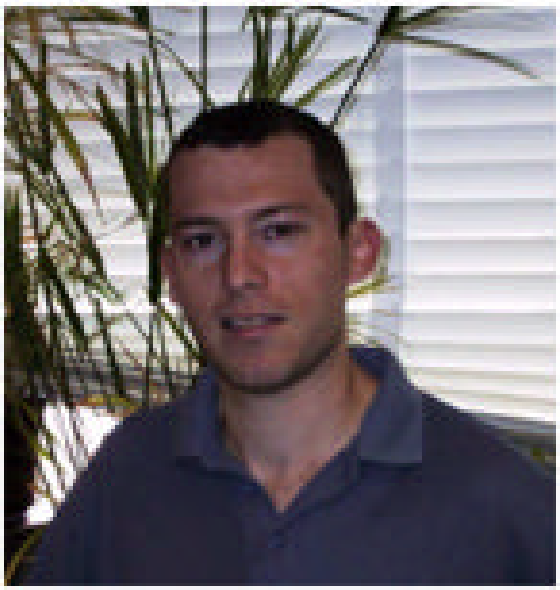
Biographies

Tien-ho Lin received the BSc degree in mathematics and MSc degree in computer science at the National Tsing Hua University, Hsinchu, Taiwan. Since 2004, he has been working toward the Ph.D. degree in language technology at Carnegie Mellon University, Pittsburgh, PA. His research interests are in the area of statistical learning and computational biology as well as applications in proteomics and gene expression time-courses.



Robert F. Murphy is the Ray and Stephanie Lane Professor of Computational Biology and Director of the Lane Center for Computational Biology at Carnegie Mellon University. He is also Professor of Biological Sciences, Biomedical Engineering, and Machine Learning and directs (with Ivet Bahar) the joint CMU-Pitt Ph.D. Program in Computational Biology. He served as the first full-term chair of NIH's Biodata Management and Analysis Study Section, was named a Fellow of the American Institute for Medical and Biological Engineering in 2006, received an Alexander von Humboldt Foundation Senior Research Award in 2008, and was appointed to the National Advisory General Medical Sciences Council in 2009. Dr. Murphy has co-edited two books and published over 160 research papers. He is President of the International Society for Advancement of Cytometry and is on the Editorial Boards of Cytometry and PLoS Computational Biology.

Dr. Murphy's group pioneered the application of machine learning methods to high-resolution fluorescence microscope images depicting subcellular location patterns in the mid 1990's. He currently leads NIH-funded projects for proteome-wide determination of subcellular location in 3T3 cells and continued development of the SLIF system for automated extraction of information from text and images in online journal articles.



Ziv Bar-Joseph received the BSc and MSc degrees in computer science from the Hebrew University of Jerusalem in 1997 and 1999, respectively, and the PhD degree in computer science from the Massachusetts Institute of Technology in 2003. He is currently on the faculty of the Machine Learning Department and the Computer Science Department, School of Computer Science, and the Lane Center for Computational Biology at Carnegie Mellon University. His areas of interest include time series gene expression analysis, systems biology, and machine learning. For more information, see http://www.sb.cs.cmu.edu.

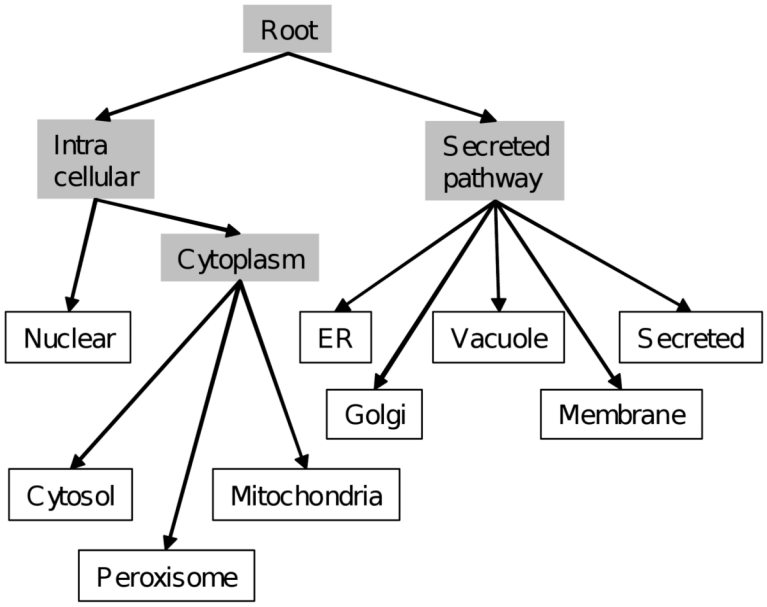


Fig. 1.
Hierarchical structure of compartments based on cellular sorting.

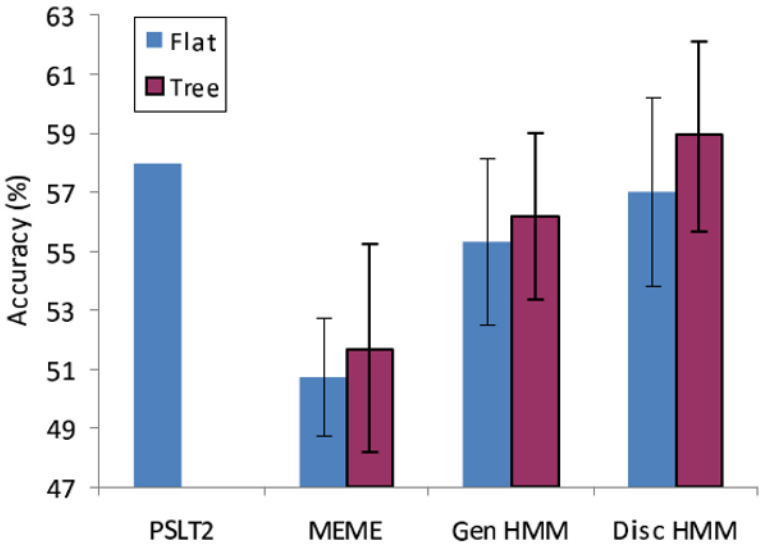


Fig. 2.

Accuracy of predictions based on motifs discovered by the different methods or a set of database motifs for PSLT2. Results for the PSLT2 methods are taken from [11]

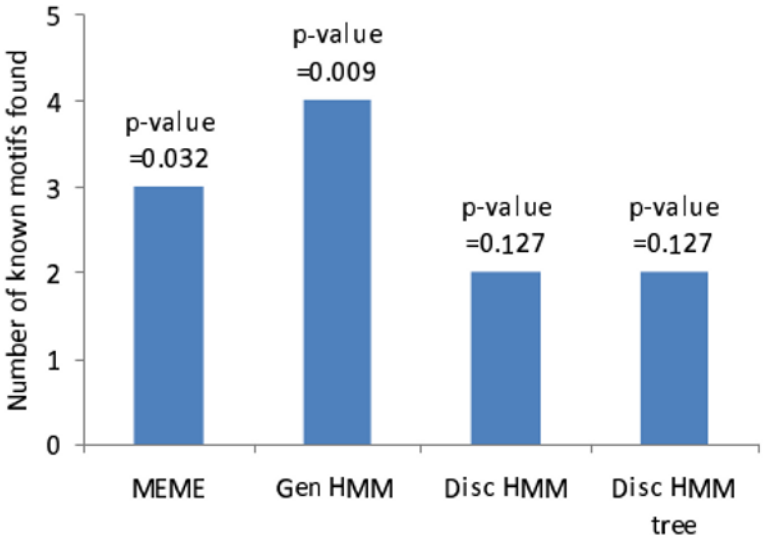


Fig. 3.

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

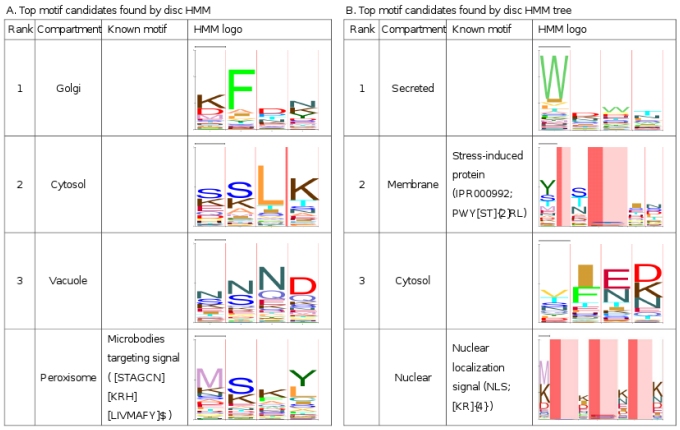


Fig. 4.

Top motif candidates that are most predictive of localization, discovered by discriminative HMM using flat (A) and hierarchical compartment structure (B). Known motifs recovered by our methods are also shown with InterPro ID and regular expressions, which partially matches the HMM logo [33]. NLS is also found when using the flat structure but not shown above. 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).

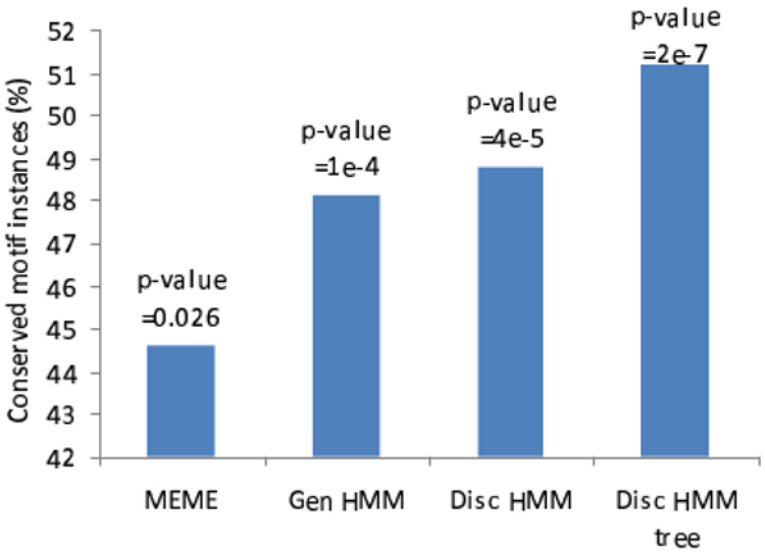


Fig. 5.

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).

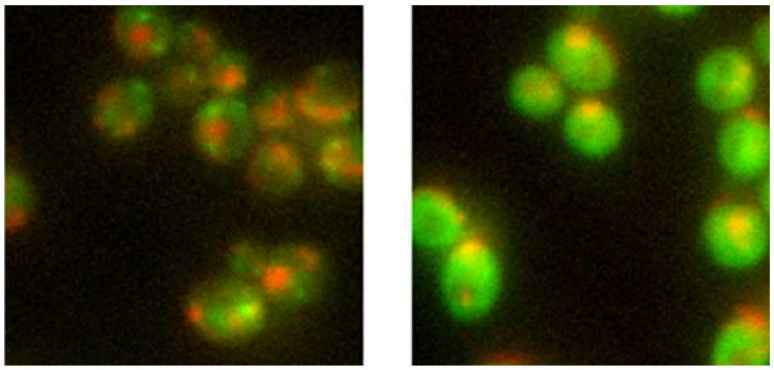


Fig. 6.

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 (left) and Ppt1/YGR123C (right). Images were obtained from the UCSF GFP-localization database (http://yeastgfp.ucsf.edu/)

TABLE I

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).

	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

IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2011 September 1.