A Bayesian System Integrating Expression Data with Sequence Patterns for Localizing Proteins: Comprehensive Application to the Yeast Genome

Amar Drawid [1]

&

Mark Gerstein [1,2] *

Departments of (1) Molecular Biophysics & Biochemistry
and (2) Computer Science
266 Whitney Avenue, Yale University
PO Box 208114, New Haven, CT 06520
(203) 432-6105, FAX (203) 432-5175
Mark.Gerstein@yale.edu

Version - Final

ABSTRACT

We develop a probabilistic system for predicting the subcellular localization of proteins and estimating the relative population of the various compartments in yeast. Our system employs a Bayesian approach, updating a protein's probability of being in a compartment based on a diverse range of 30 features. These range from specific motifs (e.g. signal sequences or HDEL) to overall properties of a sequence (e.g. surface composition or isoelectric point) to whole-genome data (e.g. absolute mRNA expression levels or their fluctuations). The strength of our approach is the easy integration of many features, particularly the whole-genome expression data. We construct a training and testing set of ~1300 yeast proteins with an experimentally known localization from merging, filtering, and standardizing the annotation in the MIPS, Swiss-Prot and YPD databases, and we achieve 75% accuracy on individual protein predictions using this dataset. Moreover, we are able to estimate the relative protein population of the various compartments without requiring a definite localization for every protein. This approach, which is based on an analogy to formalism in quantum mechanics, gives greater accuracy in determining relative compartment populations than that obtained by simply tallying the localization predictions for individual proteins (on the yeast proteins with known localization, 92% vs. 74%). Our training and testing also highlights which of the 30 features are informative and which are redundant (19 being particularly useful). After developing our system, we apply it to the 4700 yeast proteins with currently unknown localization and estimate the relative population of the various compartments in the entire yeast genome. An unbiased prior is essential to this extrapolated estimate; for this, we use the MIPS localization catalogue, and adapt recent results on the localization of yeast proteins obtained by Snyder and colleagues using a minitransposon system. Our final localizations for all ~6000 proteins in the yeast genome are available over the web at http://bioinfo.mbb.yale.edu/genome/localize.

INTRODUCTION

The subcellular localization of a protein – the location or compartment it occupies within the cell – is one of its most basic features, and there is an involved machinery within the cell for sorting newly synthesized proteins and sending them to their final locations. However, with the advent of whole-genome sequencing, we are now in the position of knowing the sequences of many proteins without knowing their localization.

Various methods have been employed in the past to predict the subcellular localization of proteins.

Nakai and colleagues developed an integrated expert system to sort proteins into different compartments using sequentially applied "if-then" rules (Nakai & Kanehisa, 1991, 1992, 1999). This eventually culminated in the PSORT system available over the web (psort.nibb.ac.jp). The rules were based on different signal sequences, cleavage sites, and the amino acid composition of individual proteins. At every node of the "if-then" tree, a protein was classified into a category (left or right descendent of the node) based on whether it satisfied a certain condition. One advantage of this process was that it could potentially mimic the actual physical decisions in the real sorting process. In further work, Nakai & Horton (1996) developed a more probabilistic approach, and they used a "k nearest neighbors" method to classify proteins according to the localization of their closest relatives (Nakai & Horton, 1997).

Other integrated approaches to predicting subcellular localization focussing on sequence composition have been developed recently. Reinhardt & Hubbard (1998) used overall composition in conjunction with neural networks to classify proteins directly into different compartments. Andrade *et al.* (1998) concentrated on using the composition of surface residues to predict subcellular localization.

There also has been much activity in predicting individual sorting signals -- e.g. signal sequences targeting proteins to the secretory pathway or mitochondrial targeting peptides. In particular, von Heijne and colleagues have worked extensively on identifying these, using neural networks and weight matrices (Claros *et al.*, 1997; Nielsen *et al.*, 1997, 1999; Sipos & von Heijne, 1993; von Heijne, 1986, 1992; von Heijne *et al.*, 1997). Their individual predictions for the various sorting sequences collectively form an impressive system for protein localization. However, it is not always clear how to combine the individual predictions into a unified framework. Related work on the identification of sorting sequences has been carried out in other laboratories (e.g. Claros & Vincens, 1996; Ladunga *et al.*, 1991; Milanesi *et al.*, 1996).

In this paper, we describe an integrated system for localizing yeast proteins using Bayesian formalism. Initially, we assume that each protein has certain default probabilities of being in the various compartments. We sequentially update these "prior" expectations using Bayes' rules and a variety of features (clues) to obtain the final

probabilities that the protein has of being in the different compartments. By analogy to formalism in quantum mechanics, we also develop a way of estimating the overall compartment population (i.e. the total number of proteins in a compartment) without rigidly localizing proteins to a single compartment. We carefully construct various sets of yeast proteins of known localization based on merging, filtering, and standardizing the annotation in MIPS, Swiss-Prot and YPD, and we test and train our system against these sets. Finally, we apply our system in an "extrapolative" fashion to predict the subcellular location of the yeast proteins with currently unknown localization. This allows us to estimate tentatively the overall relative population of the various yeast compartments. Our work follows upon a recent structural and functional characterization we have done on the yeast genome (Gerstein, 1998a; Hegyi & Gerstein, 1999; Jansen & Gerstein, 2000).

OUR FORMALISM

Compartments as States and the State Vector

Our overall formalism is schematized in Figure 1a. In our actual results (next section), we assume that a protein exists in one of five "generalized" compartments. However, in this section we will use only three compartments to explain our formalism: cytoplasm (C), nucleus (N), and the extracellular environment and secretory pathway (E). We will discuss the localization L of protein *m* in terms of its *probability state vector*:

$$\vec{\mathbf{P}}_{\mathrm{m}}(\mathrm{L}) = (p_m(C), p_m(N), p_m(E)) \qquad (1)$$

In this vector, each component gives the probability that a protein can be found in the corresponding subcellular compartment. This formalism is directly analogous to the state vector of an individual particle used in statistical quantum mechanics.

Feature Vector

A feature (or clue) is an observation made about a protein. For example, it could be a protein's absolute mRNA expression value, or its isoelectric point, or the fact that it does not have a signal sequence. We encapsulate our knowledge about the association between a feature and the compartments in a *feature vector*. We count the number of proteins in each compartment that possess the feature. Each component of the feature vector equals the fraction of the total number of proteins in that compartment possessing that feature:

$$\vec{\mathbf{P}}(\text{feature} \mid \mathrm{L}) = (p(feature \mid C), p(feature \mid N), p(feature \mid E)) \qquad (2)$$

For example, for the feature "NLS = true," we obtain $p(NLS=true \mid N)$ by counting the fraction of the total number of nuclear proteins that contain the nuclear localization signal

(NLS). Note that unlike the components of the state vector, the components of each feature do not sum to 1.

Updating the State Vector Using Bayes' Rule with Feature Vectors

Following along with the schematic in figure 1a, we start our analysis with a *prior* - a state vector that contains the assumed default probabilities of a protein being in the different compartments. We update the prior using a feature vector that corresponds to a feature that the protein possesses, and then obtain an *a posteriori* state vector. Thus, if we update the state vector of protein *m* with the feature vector corresponding to the feature "nuclear localization signal present (NLS=true)," we obtain

$$\vec{\mathbf{P}}_{\mathrm{m}}(\mathrm{L} \mid \mathrm{NLS} > \mathrm{true}) = (p_m(C \mid NLS{=}true),\ p_m(N \mid NLS{=}true),\ p_m(E \mid NLS{=}true)) \tag{3}$$

which could, for instance, look like (0.1, 0.6, 0.3). Specifically, we use Bayes' Rule of conditional probability for the updates (Pitman, 1997):

$$p_m(L \mid feature) = p_m(L) \cdot p(feature \mid L) \,/\, Z \tag{4}$$

where Z is a normalization factor. Z equals the product of the fraction of the total number of proteins in each location having that feature and the prior probability of the protein *m* being in that location, summed over all locations,

$$Z = \sum_{L} p(feature \mid L) \,.\, p_m(L) \tag{5}$$

For instance, *L* could be cytoplasm and *feature* could be "NLS=true." Then $p(NLS{=}true \mid C)$ is the fraction of all cytoplasmic proteins with an NLS (that is, the cytoplasmic component of the feature vector $\vec{\mathbf{P}}_{\mathrm{m}}(\mathrm{L} \mid \mathrm{NLS} > \mathrm{true})$), and $p_m(C \mid NLS{=}true)$ is the chance that the given protein *m* with an NLS is cytoplasmic.

After an update, we make the *a posteriori* state vector our new prior, and repeat the procedure using a different feature vector. We then get a new *a posteriori* state vector, which serves as the prior for another feature. We sequentially apply all available feature vectors, updating the state vector every time.

Thresholding a State Vector to Localize a Protein in a Specific Compartment

After we apply all the features and arrive at a "final" state vector for each protein, we feel justified in localizing the protein to a single compartment if the probability density is strongly concentrated to that compartment. We call this procedure "thresholding," and make this determination in two ways:

(i) **Top-2 difference**. We choose the two compartments that have the greatest probability values in the state vector. If the difference between their probability values is greater than a particular threshold, we localize the protein to the compartment corresponding to the larger value. Otherwise, we leave the protein unlocalized.

(ii) **Entropy.** We calculate the entropy of the state vector using the standard formula, viz:

$$S(\vec{\mathbf{P}}_m) = . \sum_L p_m(L) . \ln p_m(L) , \quad (6)$$

where the sum is over all locations *L*.

A protein with low entropy has a high probability of being in a particular compartment and a low probability of being in the others. Hence, it is localized well. In this paper, we have used an entropy threshold to differentiate between the localized and unlocalized proteins, although the top-2 difference threshold performs almost as well.

Estimating Relative Compartment Populations with an Overall Compartment Population Vector

To estimate the relative population of the various compartments (i.e. the ratio of the total number of proteins present in those compartments), we could simply tally the specific localizations found via the entropy threshold. However, some proteins are not strongly localized by our procedure. Moreover, we feel it is quite reasonable for some proteins not to have a definite localization. For instance, several proteins have been found experimentally in more than one compartment (Hodges *et al.*, 1999; Ross-Macdonald *et al.*, 1999). In particular, the transcription factor complex NFλB is known to shuttle between an inactive form in the cytoplasm and an active form in the nucleus (Kopp & Ghosh, 1994), and various structural proteins also appear both in the nucleus and the cytoplasm (see TUB4 example below).

Hence, we use a different procedure to estimate the relative population of each compartment. As schematized in Figure 1b, we build an *overall compartment population vector* $\vec{\mathbf{N}}(\mathrm{L})$, in which each component represents the overall population of a certain compartment:

$$\vec{\mathbf{N}}(\mathrm{L}) = (v(C), v(N), v(E)) \quad (7)$$

We obtain each component *v(L)* by summing over the state vectors of all the proteins the probability density that a protein would be in that compartment. For instance, for the cytoplasmic component of the vector, *v(C)*, we have

$$v(C) = \text{int}(\sum_m p_m) C) \ ). \quad (8)$$

$\tilde{\mathbf{N}}(\mathrm{L})$ provides an estimate of the overall populations of the different compartments without requiring individual predictions.

While this summation of probabilities may appear to be intuitively obvious, its formal justification is not trivial. We present the analogy of our problem of estimating the compartment populations to the density matrix formalism in quantum mechanics at our website. While not strictly necessary, we think this analogy is stimulating and useful in connecting our analysis with a number of powerful mathematical tools.

IMPLEMENTATION

To successfully implement the formalism, we need (i) high-quality training and testing data, (ii) a good prior, (iii) relevant features, and (iv) a cross-validation protocol.

The Localized-1342, the Training and Testing Dataset

To train and test our system, we used the localizations from Swiss-Prot (Bairoch & Apweiler, 2000) and MIPS (Frishman *et al.*, 1998; Mewes *et al.*, 1998, 1999; Frishman & Mewes, 1997) -- and to a lesser extent from the Yeast Protein Database (YPD, version 9.08) (Hodges *et al.*, 1999). We prepared 4 different datasets of localized yeast proteins. We called them *Localized-465*, *Localized-704*, *Localized-1342* and *Localized-2013*, where the terminal number (e.g. "-465") represented the number of proteins in the dataset. The four datasets are described in detail in figure 2. They differ in their overall "quality."

Our quality factor for each protein describes the degree to which we were sure that its localization was based on real experimental evidence (rather than computational predictions), and that this localization was consistent amongst the various data sources (e.g. MIPS versus Swiss-Prot). In particular, a Swiss-Prot localization was characterized as high-quality only if it was *not* annotated as "predicted" or "possible," and if the protein could be easily assigned to a single collapsed location (e.g. excluding cytoskeletal proteins or proteins with multiple locations). Similar exhaustive characterizations were performed for proteins with MIPS localizations.

Consideration of the data quality was critical for training and testing, since we had to be careful to guard against "circular logic" -- that is, training our computational prediction algorithm on computationally predicted localizations in the training set. For example, if the training data contained proteins that were predicted to have membrane (T) localization according to transmembrane prediction programs, the results of our algorithm

could not be considered valid as it also makes use of a generic transmembrane prediction program.

Amongst our four datasets, the smallest one (Localized-465) contained only the proteins with the highest quality localizations, i.e. proteins which had consistent localizations in MIPS, Swiss-Prot and YPD, and which were not annotated to have predicted localization in any of these data sources. The largest one (Localized-2013) contained a number of additional proteins with more problematic localizations that could potentially be derived from computational predictions. Unfortunately, we were not sure of the degree to which localization was derived from computational predictions because of the incomplete annotations of many yeast proteins. Our third dataset (Localized-1342) included all proteins that had non-conflicting localizations in either MIPS or Swiss-Prot or both, and that were not annotated to have a predicted localization. We felt that this dataset gave the best balance between overall quality and the number of proteins and largely avoided the "circular validation" problem. The cross-validation and extrapolation results in this paper are based on this dataset.

Five Collapsed Compartments (C, E, N, M and T) and their Prior Population

The proteins in the Localized-1342 dataset are mainly associated with 12 subcellular compartments (see table 1, figure 3). However, many of these compartments contain only a very small number of proteins, greatly skewing the statistics. For instance, there are few proteins in vesicles and vacuoles (<30 in each), in contrast to the 494 nuclear proteins. Hence, we found it advantageous to *collapse* the 12 compartments into five new "generalized" compartments that lumped together a number of the related smaller compartments, allowing for a more even distribution of proteins. Our compartments are the nucleus (N), mitochondria (M), cytoplasm (C), membrane (T for Transmembrane), and secretory pathway (E for Endoplasmic reticulum or Extracellular). Our T compartment contains all the integral transmembrane (cell membrane, plasma membrane and membranes of various compartments such as mitochondria, nucleus, golgi) proteins, whereas our E compartment contains all the secreted proteins and proteins in the secretory pathway and small organelles (i.e., proteins in the endoplasmic reticulum, golgi, vacuoles, vesicles and peroxisome).

Our five compartments are mutually exclusive; a protein cannot logically be in two compartments simultaneously. We excluded all cytoskeletal proteins from our training data, because most of these proteins could not be easily localized to a single one of our five compartments. For example, cytoskeletal Gamma-Tubulin (TUB4), a protein localized to the spindle pole body (Sobel & Snyder, 1995), has the following MIPS subcellular localization annotation: "spindle pole body; cytoplasm; nucleus."

For our initial training and testing, we used a prior based on the relative proportions of the Localized-1342 proteins in the different compartments. This is shown in figure 4. We used a new composite prior for the extrapolation (discussed later).

A Diverse Set of 30 Features: motifs, overall-sequence, and whole-genome

The features that we used to implement the Bayesian formalism are described in detail in table 2. We used a total of 30 features. The features were first divided into three categories depending on the information they were derived from: (i) motifs (16 features), (ii) overall-sequence (4 features), and (iii) whole-genome (10 features). The features in the "motif" category were based on a small sequence pattern in a protein. For instance, the feature HDEL (the endoplasmic reticulum retention signal) denoted the presence or absence of the HDEL motif at the C-terminus of a protein. The features in the "overall-sequence" category were based on the entire sequence of a protein. For example, the feature PI was the isoelectric point pI of a protein, whereas the feature TMS1 denoted the number of predicted transmembrane segments in a protein. Finally, the "whole-genome" features were derived from considering whole-genome level data; the specific values of a protein's whole-genome features were only meaningful in the context of the values of all other proteins in the genome. For instance, the feature MAYOUNG contained the mRNA absolute expression data in the experiments of Young and colleagues (Holstege *et al.*, 1998), whereas the feature MRCYCSD contained the standard deviation in mRNA expression level over time (i.e. expression fluctuation) for proteins in the yeast cell cycle experiment (Spellman *et al.*, 1998).

We also subdivided the 30 features into three groups depending on how much they contributed to the overall predictive strength of our system: the 10 most important features, 9 other included features and 11 redundant features. These are described in detail later. For all the results reported in this paper, we excluded the 11 redundant features and based our system on the best 19 features.

For each feature, we divided proteins into a specific number of bins (see table 2) according to the different values. For instance, for the knockout feature (KNOCKOUT), we divided proteins into two bins: lethal versus viable knockouts. On the other hand, for the Young expression data (MAYOUNG feature), we divided proteins into 10 different bins of identical size, according to their absolute levels of mRNA expression. We then used each bin as a separate feature, and created a separate feature vector for each bin. We updated the state vector of a protein by using the feature vector of the bin it belonged to. For example, if a protein was associated with bin four of the Young dataset (MAYOUNG bin=4), we used the feature vector $\vec{\mathbf{P}}(\text{MAYOUNG bin} > 4 \mid \text{L})$ to update its state vector.

For identifying a number of the "motif" features, we used fairly simple rules that could be summarized in either regular expressions or weight matrices. We could, of course, have used more advanced signal recognition methods (such as the system of Claros & Vincens (1996) for identifying mitochondrial proteins). However, many of these are implemented as complex neural network programs accessible only via mail servers, and using them would have substantially increased the software engineering complexity of our system. We decided to use the simpler approach first, as our goal here was primarily to see how we could integrate and balance many diverse features, particularly those involving the

whole-genome information, with the traditional sorting signals. We did elect, nevertheless, to use the SignalP server to help identify proteins in the secretory pathway (feature SIGNALP in our scheme)(Nielsen *et al.*, 1997, 1999; von Heijne *et al.*, 1997). This identifies the most basic sorting signal, and we found that incorporating it did improve the overall performance slightly beyond that obtained from the simple weight matrix approach (feature SIG1).

Cross-validation and Correlated Features

Our Bayesian system is a "naive" or "simple" case of a more general Bayesian network in that it implicitly assumes that all features are independent and uncorrelated (Friedman *et al.*, 1997). This is, of course, not completely true for the features we are using. However, by partitioning our dataset into separate training and test sets and using cross-validation to measure the performance of our system, we can avoid misleading results due to over-parameterization (Efron & Tibshirani, 1986). Furthermore, we can identify the most redundant features -- those that contribute the least to the overall prediction accuracy or actually hurt the prediction -- and remove them. We can also highlight the features that contribute the most to the strength of the overall prediction.

Specifically, we trained and tested our system using a seven-fold jackknife on the proteins with known localizations. We divided the Localized-1342 set into 7 subsets, each containing ~190 proteins. The proteins in each subset were selected completely randomly. Each protein belonged to only a single subset, and there were no duplicated proteins in any subset. We then predicted the localization of the proteins in each subset based on training our system on the remaining ~1150 proteins that belonged to the other 6 subsets.

When we performed the cross-validation, we observed that the addition of some features with redundant information decreased the overall prediction accuracy, as is to be expected because of the implicit correlations. Table 2 summarizes the degree to which each feature raised or lowered the prediction accuracy when it was added to or subtracted from our system. This gives a rough measure of the "information content" of the feature. In particular, we found that using either the absolute expression levels from the SAGE or GeneChip experiments individually gave better results than when we used both these features simultaneously (features MAYOUNG vs. MASAGEG). In such cases, we included only one of these features in our implementation, and denoted the other feature as “redundant” (table 2). We would have obtained very similar results if we interchanged any of the “redundant” features with their partners that are included in our current implementation.

PERFORMANCE ON 1342 PROTEINS WITH KNOWN LOCALIZATION

Analysis of Individual Protein Predictions

When we applied our procedure to the testing set proteins, we found that they had a fairly even distribution of entropies, from 0 to 1.55, as shown in the entropy vs. coverage graph (figure 6a). Consequently, the error rate varies linearly with the coverage (figure 6b). We localized 2/3 of the proteins in the Localized-1342 set with an entropy threshold of 0.91. For each of these proteins, we compared our predicted compartment with the protein's observed location. We correctly predicted the compartments of 75% of these proteins (figure 5a). Nuclear proteins were predicted extremely well, whereas membrane and secretory pathway proteins were predicted relatively poorly (N 88% correct vs. T 39% correct and E 50% correct).

As shown in the table related to figure 2, the overall cross-validation prediction accuracy for the Localized-465 dataset (the dataset with highest-quality protein localizations) was very high at 88%. The Localized-704 dataset had the same accuracy (75%) as the Localized-1342 set. The low-quality dataset, Localized-2013, had a low prediction accuracy (72%). Thus, the cross-validation prediction accuracy decreased with the quality of localization in the dataset.

Relative Compartment Populations in the Localized-1342 Set

We estimated the populations of the different compartments in the Localized-1342 set by constructing an overall compartment population vector. The comparison between the known compartment populations of the Localized-1342 proteins and the compartment populations estimated by our method are shown in figure 5b. We estimated more nuclear (N) proteins and fewer cytoplasmic (C) proteins than those present in the Localized-1342 set.

Note how the overall compartment population estimate is considerably more accurate than the simple summation of the individual "well localized" protein predictions (92% vs. 74%), particularly for proteins in the less populated compartments (see caption for figure 5b). One can readily rationalize this: the state vectors of the proteins belonging to one of the small compartments (i.e. the golgi which is part of E) may contain appreciable probability values for that compartment. However, these values are usually not high enough to actually localize the proteins to this minor compartment when the state vectors are thresholded, and are thus ignored when one imposes a single localization on every protein. On the other hand, these probability values accumulate in the overall compartment population vector.

EXTRAPOLATION TO THE 4700 YEAST PROTEINS WITH UNKNOWN LOCALIZATION

After the testing and training, we used our system to localize the 4700 yeast proteins that did not have a known localization (we called this set the *Unknown-4700*). We were then in a position to estimate the overall populations of the various compartments in the entire yeast genome.

A Composite Prior Combining MIPS and Snyder-Lab Results

A fair and unbiased prior is essential to obtain accurate extrapolation results. As should be apparent from the above, the Localized-1342 prior is rather skewed towards nuclear proteins, perhaps reflecting the interests of investigators. We extensively examined a variety of other possible priors, based on our other possible training sets (i.e. the Localized-465, Localized-704, and Localized-2013), the overall composition of the MIPS database, and the experimental data from the Snyder lab (Ross-Macdonald *et al.*, 1999). These are shown in the various subpanels of figure 4. In the Snyder lab experiments, randomly selected ~400 genes were disrupted on a large scale using a minitransposon system, and the location of the subsequent epitope-tagged proteins was determined using immunofluorescence.

The MIPS prior represents a large number of localized yeast proteins (1935), but it is biased towards nuclear proteins. Due to the way the experiments are done, the Snyder prior, which represents a smaller sampling of yeast proteins (367), accurately estimates the relative fraction of the yeast genome devoted to nuclear proteins, but tends to shift, in an understandable way, proteins from the mitochondrial, membrane, and secretory pathway compartments to cytoplasmic proteins. Consequently, we combined the E, M, and T parts of MIPS prior with the N part of Snyder prior to construct a more representative prior. We call this the *composite* prior. Its construction is described in detail in the caption to figure 4.

Extrapolation Results

To determine the locations of the Unknown-4700 yeast proteins, we trained the feature vectors on the entire Localized-1342 set, and used the composite prior. We calculated the overall compartment population vector for the Unknown-4700 proteins, which gave us an estimate of the different compartment populations. We compared these relative compartment populations with those in the Localized-1342 set (figure 7a). We estimated many more membrane (T) proteins than those expected by the relative populations of the Localized-1342 set (33% vs. 13%). This reflects the fact that a high proportion of yeast proteins are membrane proteins and that the training data (the Localized-1342) was biased against membrane proteins. Nuclear (N) proteins also constituted a large part of our estimates (33%), probably due to the fact that the training data was so strongly biased

towards them. However, we did not obtain a higher proportion of cytoplasmic (C) proteins (18% vs. 32%).

We localized 2/3 of the Unknown-4700 proteins to individual compartments using an entropy threshold of 1.12, and left the remaining proteins unlocalized. We compared the relative compartment populations thus obtained with those obtained from the overall compartment population vector (figure 7b). Membrane and nuclear proteins (T 24%, N 23%) also dominated the predictions obtained from thresholding, and only 11% of the proteins were localized to the cytoplasm (C). The relative ratios of cytoplasmic and nuclear proteins obtained from the above two methods (C 11% vs. 18%, and N 23% vs. 33%) suggested that the overall compartment population vector gave a more reliable estimate of the relative populations of the different compartments than the thresholded predictions.

Combined Results for All ~6000 Proteins in Yeast Genome, on the Web

We determined the total populations of the different compartments in the entire yeast genome (~6000 proteins) by adding (i) the populations of the different compartments in the overall compartment population vector of the Unlocalized-4700 proteins, and (ii) the compartment populations of the Localized-1342 proteins. The results are shown in figure 7c. We present our feature vectors and state vectors for all yeast proteins on our website http://bioinfo.mbb.yale.edu/genome/localize.

DISCUSSION AND CONCLUSION

In this paper, we described a system for determining the subcellular localization of proteins using Bayesian formalism. Our approach has a number of key attributes: the incorporation of expression data, the identification of redundant features, the use of an entropy threshold to identify readily localized proteins, the estimation of overall compartment populations without localizing every protein, and the development of an unbiased composite prior. We successfully implemented our system and performed a comprehensive analysis of the yeast genome. Since our system is generally applicable, it can be used to predict the subcellular locations of proteins in other organisms, such as the worm. Below, we discuss a number of aspects of our system in relation to the other approaches for localization prediction.

Flexible but does not Model Physical Process

The main advantage of our system lies in its flexibility. In a rule-based system, the order of the application of rules is fixed. At every step, the program has to classify a protein irreversibly into certain compartments according to these rules. On the other hand, our simple Bayesian approach never rigidly classifies proteins into any compartments until the very end. The probabilities of a protein being in different compartments change gradually. In the end, depending upon the various thresholds applied, our system may localize a protein to a compartment, or leave it unlocalized. New and diverse features can be added to our system easily, because the order of the features is not important.

The "flip" side of the flexibility of our system is that it models the physical process of protein sorting less directly than an expert system. In an expert system, the order of the application of rules can potentially match the actual way the protein is sorted -- e.g. in the secretory pathway.

Integration of Whole-Genome Data, Particularly from Expression Studies

The flexibility of our system is particularly important with regard to the integration of the whole-genome data such as that from the expression studies. We believe that the incorporation of the whole-genome features is a particularly novel attribute of our system.

One further aspect of this addition is that it may enable our system to better localize proteins for which gene-prediction places the N-terminus incorrectly. As pointed out by Reinhardt & Hubbard (1998), many genes are automatically assigned in large genome analysis projects, and these assignments are often unreliable for the 6'-regions. This can result in missing or only partially included leader sequences, thereby causing problems for sequence-motif-based localization algorithms. Similar considerations apply to the localization of EST fragments.

Dependence on the Prior and Feature Training Data

One of the major disadvantages of our approach is its dependence on the prior and the data used to build the feature vectors. The dependence on the prior is, of course, a general issue in Bayesian analysis that cannot be avoided. We have tried to overcome this difficulty to some degree by using the data from the Snyder lab.

Similar data-dependency issues arise with either neural network or expert system approaches. In this context, one of the advantages of the Bayesian approach is that the effects of the various features and assumptions are a bit more transparent than they are for neural networks, allowing one to appreciate biases in the training data better. However, the addition of an incorrect prior or a badly constructed feature will globally affect the proportions of all the compartments in our system. This is in contrast to the expert system approach, where a bad rule applied late in the sorting process will only have a "local" effect, changing the balance between only two or three compartments.

Accommodation of Unlocalized Proteins

Our system can readily accommodate proteins that do not strongly localize to a single compartment. A protein can be in such an unlocalized state either because it is in reality present in more than one compartment (e.g. NFλB, see above), or because we currently do not have sufficient clues to determine its localization. In analogy with quantum mechanics methodologies, our system builds an overall compartment population vector to determine the total populations of the different compartments. This vector correctly integrates the information from these ambiguously localized proteins.

Future Improvements

In the future, we hope to build on the strengths of our system (its flexibility) and try to compensate for its problems (dependence on the prior and features). In particular, we hope to add more features. These will be based on more advanced methods for signal sequence recognition (i.e. connecting to neural network based servers), on other whole-genome data (e.g. the protein abundance data of Gygi *et al.* (1999)), on protein-protein interaction maps (Enright *et al.*, 1999; Uetz *et al.*, 2000), and on transfering annotation from sequence homology to other proteins of known localization (Wilson *et al*., 2000).

We would also like to use more data from the Snyder minitransposon experiments to construct our composite prior and training set. Eventually, these experiments should be extended to a large fraction of the genome. Finally, with more data and better features, we should be able to use more distinct compartments, and localize proteins more specifically to the minor compartments.

ACKNOWLEDGEMENTS

We thank W Krebs for help with the website, V Alexandrov and Y Kluger for help with the mathematical formalism, and M Snyder and A Kumar for help with the localization experiments. MG thanks the NIH and the Keck foundation for financial support.

TABLE 1 – COLLAPSED COMPARTMENTS

Collapsed Compartment	Compartments used for collapsing	Description
C	Cytoplasm	Cytosolic proteins (not in any organelles or membranes or cytoskeleton)
M	Mitochondria	Mitochondrial proteins
N	Nucleus	Nuclear proteins
T	Membrane Plasma membrane	Integral transmembrane proteins (in the cell membrane, the plasma membrane, or the membranes of various compartments such as mitochondria, nucleus, golgi)
E	Endoplasmic reticulum (ER) Golgi apparatus Peroxisome, Vacuole, Vesicle Extracellular	Proteins involved in the secretory pathway and those in small organelles

TABLE 2 – FEATURES

The table describes the 30 features used in our system. In the first table, each row contains the name of a feature, its general type and subtype, its contribution towards the overall prediction strength (in terms of a percentage change described below), its status regarding our implementation, and the number of bins used to model it. The second table provides more extended description of each feature.

The positive values in the "%Change" column denote the fall in the prediction accuracy if the cross-validation is performed *without* the corresponding feature -- i.e. if it is excluded from the 19 basic features used for the analysis. Note that the prediction accuracy for cross-validation for the Localized-1342 set is 75% (74.7% to be exact) when we use the 19 basic features. Thus, for example, when the feature MIT1 is excluded, prediction accuracy falls by 5.1% (to 74.7 - 5.1 = 69.6%). Negative values in the "%Change" column denote a fall in the prediction accuracy if the cross-validation is performed after *including* the corresponding feature in the system, beyond the 19 basic ones. Thus, when the feature COILDCO is included in our system, prediction accuracy falls by 0.1% (to 74.7 - 0.1 = 74.6%). A feature has "Important" status if the prediction accuracy falls by more than 0.5% after the exclusion of the feature. Such features are included in our final implementation. The status of the feature is "Included" if the feature is included in our final implementation along with the "Important" features. A feature has "Redundant" status if its inclusion decreases the prediction accuracy. Such features are not included in our final implementation. (We could also have computed the redundancy of each of our features by computing the mutual information between each of them.) Some further notes: (i) "from-MIPS" means "this information could be derived from MIPS or PEDANT" (Frishman *et al.*, 1998; Mewes *et al.*, 1998, 1999; Frishman & Mewes, 1997). (ii) "from-YPD" means "as given in the Yeast Protein Database, YPD" (Hodges *et al.*, 1999). We mostly used version 8.15. However, some features were taken from a newer version (9.08). (iii) "from-NK92" means "as described in Nakai & Kanehisa (1992)." (iv) The protein sequence patterns are written in the UNIX regular expression format.

A. Brief Description

Feature	Type	Subtype	%Change	Status	Bins
MIT1	Motif	Signal	5.1	Important	2
GLYC	Motif	Signal	1.2	Important	10
SIGNALP	Motif	Signal	1.0	Important	2
SIG1	Motif	Signal	0.7	Important	2
NUC1	Motif	Signal	0.6	Important	6
PI	Overall-sequence	Isoelectric Point	1.3	Important	10
TMS1	Overall-sequence	Transmembrane helix	0.9	Important	5
MAYOUNG	Whole-genome	Absolute expr. (GeneChip)	3.6	Important	10
KNOCKOUT	Whole-genome	Knockout mutation	1.8	Important	2
MRDIASD	Whole-genome	Expr. fluctuation (Diauxic Shift)	1.4	Important	10
PLMNEW1	Motif	Signal	0.3	Included	2
FARN	Motif	Signal	0.3	Included	2
GGSI	Motif	Signal	0.3	Included	2
MIT2	Motif	Signal	0.2	Included	2
HDEL	Motif	Signal	0.1	Included	2
NUC2	Motif	Signal	0.1	Included	3
POX1	Motif	Signal	0.1	Included	2
MRCYELU	Whole-genome	Expr. fluctuation (Cell Cycle)	0.4	Included	10
MRCYCSD	Whole-genome	Expr. fluctuation (Cell Cycle)	0.2	Included	10
COILDCO	Motif	Coiled coils	-0.1	Redundant	2
CKIISITE	Motif	Kinase target site	-0.1	Redundant	2
CDC28SITE	Motif	Kinase target site	-0.3	Redundant	4
PKASITE	Motif	Kinase target site	-0.5	Redundant	5
ROSTALL	Overall-sequence	Surface residue composition	-0.8	Redundant	9
LENGTH	Overall-sequence	Protein length	-1.6	Redundant	10
MASAGEL	Whole-genome	Absolute expr. (SAGE)	-0.3	Redundant	10
MRCYC15	Whole-genome	Expr. fluctuation (Cell Cycle)	-0.4	Redundant	10
MRCYC28	Whole-genome	Expr. fluctuation (Cell Cycle)	-0.6	Redundant	10
MASAGEG	Whole-genome	Absolute expr. (SAGE)	-0.7	Redundant	10
MASAGES	Whole-genome	Absolute expr. (SAGE)	-0.9	Redundant	10

B. Extended Description

Feature	Description
MIT1	More than one N-terminal residue is cut (good chance of being mitochondrial) (from-YPD).
GLYC	Glycosylation site (from-NK92).
SIGNALP	Secretory signal peptide according to the SignalP server (Nielsen *et al.*, 1997, 1999; von Heijne *et al.*, 1997).
SIG1	Results of a simple program that predicts if a protein has a signal sequence. The pattern consists of a charged residue within the first seven residues, followed by a stretch of 14 residues with an average GES hydrophobicity less than -1 kcal/mole.
NUC1	Four-residue patterns of 1. All basic amino acids (K or R) or 2. Three basic amino acids (K or R), and one H or P (from-NK92).
PI	pI (Isoelectric Point) values (from-MIPS)(from-YPD).
TMS1	Results of a program that predicts whether a protein has transmembrane (TM) segments. TM segments were identified using the GES hydrophobicity scale (Engelman *et al.*, 1986). The values from the scale for amino acids in a window of size 20 were averaged, and then compared against a cutoff value. We used the Boyd and Beckwith MaxH criteria to set the cutoffs as in previous analyses (Boyd *et al.*, 1998; Klein *et al.*, 1985; Gerstein *et al.*, 2000).
MAYOUNG	Absolute mRNA expression in a GeneChip experiment (Holstege *et al.*, 1998).
KNOCKOUT	Knockout mutation (lethal or viable). (from-MIPS)(from-YPD) (Baudin *et al.*, 1993; Shoemaker *et al.*, 1996; Wach *et al.*, 1994).
MRDIASD	Standard deviation in mRNA expression level over time (i.e. expression fluctuation) for a protein in the diauxic shift experiment (DeRisi *et al.*, 1997).
PLMNEW1	Plasma membrane signal (from-NK92). We checked for this signal in the entire sequence, rather than just at the C-terminal.
FARN	C-terminal farnesylation site: the sequence pattern consists of a Cysteine followed by two aliphatic residues and one more residue at the C-terminus (C[ALIVG][ALIVG].$) (Stryer, 1996).
GGSI	C-terminal geranylgeranylation site (CC$\|C.C$|CC..$) (Stryer, 1996).
MIT2	Mitochondrial matrix import sequence: The N-terminal of the protein has repeated alternating hydrophobic and hydrophilic patterns, and the protein contains at least 4 S or T residues in its 20 N-terminal residues.
HDEL	Endoplasmic reticulum retention signal (HDEL) (from-NK92). We checked for the presence of this signal in the 9 C-terminal residues.
NUC2	Pattern starting with a P and followed within 3 residues by a basic 4-residue segment containing K or R residues (P.{0,3}[KR]{4}) (from-NK92).
POX1	C-terminal Peroxisome import signal ([SA][KRH]L) (from-NK92).
MRCYELU	Standard deviation in mRNA expression level over time (i.e. expression fluctuation) for a protein in the elutriation time series experiment in Yeast Cell Cycle Analysis Project (Spellman *et al.*, 1998).
MRCYCSD	Standard deviation in mRNA expression level over time (i.e. expression fluctuation) for a protein in the Alpha-factor arrest time series experiment in Yeast Cell Cycle Analysis Project (Spellman *et al.*, 1998).

COILDCO	Results of the Multicoil program that predicts the presence of coiled coils (from-MIPS) (Wolf *et al.*, 1997).
CKIISITE	Potential casein kinase II protein kinase sites (from-YPD).
CDC28SITE	Potential cdc28 protein kinase sites (from-YPD).
PKASITE	Potential protein kinase A (cAMP-dependent protein kinase) target sites (from-MIPS)(from-YPD).
ROSTALL	As given in the Andrade *et al.* paper, we calculated the two eigenvectors for all proteins using their surface amino acid compositions. We then plotted the proteins in the plane of the eigenvectors and divided the plane into 9 compartments, each of which served as a bin (Andrade *et al.*, 1998; Rost & Sander, 1994).
LENGTH	Length of a mature protein after the removal of N- and C- terminal peptides (from-MIPS)(from-YPD).
MASAGEL	Absolute mRNA expression of l phase proteins in the SAGE experiment (Velculescu *et al.*, 1997).
MRCYC15	Standard deviation in mRNA expression level over time (i.e. expression fluctuation) for a protein in the cdc15 arrest time series experiment in Yeast Cell Cycle Analysis Project (Spellman *et al.*, 1998).
MRCYC28	Standard deviation in mRNA expression level over time (i.e. expression fluctuation) for a protein in the cdc28 time series experiment in Yeast Cell Cycle Analysis Project (Spellman *et al.*, 1998).
MASAGEG	Absolute mRNA expression of g/m phase proteins in the SAGE experiment (Velculescu *et al.*, 1997).
MASAGES	Absolute mRNA expression of s phase proteins in the SAGE experiment (Velculescu *et al.*, 1997).

FIGURE 1 - BAYESIAN FORMALISM AND OVERALL COMPARTMENT POPULATIONS

Part A: Bayesian Formalism. The pie charts in the figure show the state vector of protein *m* (the distribution of probabilities of protein *m* being in the different compartments) at various stages of the Bayesian analysis. The bar graphs show the feature vectors $\vec{\mathbf{P}}(\text{feature} \mid \text{L})$ for each feature. The patterns of the 3 compartments are shown in the schematic of the cell in the top left corner (N for nucleus—black, C for cytoplasm—white, E for extracellular environment and secretory pathway—gray). First, the state vector of the protein is updated from "Initial Prior" to "Posterior" vector using the feature vector for feature "NLS=true" and Bayes' rule. For example, $p_m(N \mid NLS=true) = p_m(N) \cdot p(NLS=true \mid N) / Z$, where Z is a normalization factor. See text for further explanation. This new state vector is then sequentially updated using various feature vectors. The feature *mRNA expr=high* can be regarded as the feature *MAYOUNG bin=10* in our actual analysis. Similarly, the feature *pI>9* can be regarded as the feature *PI bin=10* in our actual analysis. The final state vector shows that the protein *m* has a high probability of being a nuclear protein.

For reference Bayes' Rule is as follows: If $B_1,...,B_n$ are all possible mutually exclusive results of the first stage of a procedure, and A is an observation at a second stage, we can calculate the probabilities of events B_i given the occurrence of A as:

$$P(B_i|A) = P(A|B_i) \cdot P(B_i) / [P(A| B_1) \cdot P(B_1) + ... + P(A| B_n) \cdot P(B_n)]$$

Where $P(B_i)$ = unconditional probabilities (prior probabilities), $P(A|B_i)$ = conditional probabilities (likelihoods) (Pitman, 1997).

Notes: (1) The sum of the probabilities in each individual bar graph is not equal to 1, because each bar depends on the properties of its compartment and other related complementary features. For example, $p(NLS=true \mid N) + p(NLS=false \mid N) = 1$, but $p(NLS=true \mid N) + p(NLS=true \mid C) + p(NLS=true \mid E) \equiv 1$. (2) If the probability for a location in a feature vector is 0, or if a component of the state vector becomes 0 (after normalizing) while updating, the probability for the corresponding location in the resultant state vector will always be equal to 0. To avoid this, we add a pseudo-count of 0.001 to all feature vector components that are equal to 0. We also add a pseudo-count of 0.001 to all state vector components that are equal to 0 every time we update the state vectors.

<u>Part B</u>: Estimating Overall Compartment Populations. Two ways of estimating the relative populations of the various compartments (i.e. the ratio of the total number of proteins present in those compartments) in the entire yeast genome are shown. In the top part, the small pie charts on the right are the state vectors of individual proteins. To estimate the relative population of each compartment, we build an *overall compartment population vector* $\vec{\mathbf{N}}(\mathrm{L})$, in which each component represents the overall population of a certain compartment: $\vec{\mathbf{N}}(\mathrm{L})$ = (*v(C), v(N), v(E)*). We obtain $\vec{\mathbf{N}}(\mathrm{L})$ by adding the probability state vectors of all proteins. More specifically, we obtain each individual component *v(L)* by summing $p_m(L)$ of all proteins (and then rounding it to the nearest integer). In the bottom part, each protein probability state vector is "thresholded" to a specific compartment using an entropy value. If the protein lies below the entropy threshold, it is classified as "unlocalized" (shown as hashed). We can estimate the relative population of the various compartments by simply adding up the number of proteins belonging to each compartment (and those left unlocalized).

FIGURE 2 – CREATION OF FOUR TRAINING DATASETS

The Venn diagram shows how we analyzed the known protein localizations from different data sources to build our test and training sets. We were particularly concerned about making sure that our training data was of high quality -- that it was based on experimentally determined localizations and that these localizations were consistent among the various data sources. See text for more discussion. The Venn diagram consists of 4 circles. The bottom circle represents proteins in Swiss-Prot with high-quality localization (704). This is our core data. The right circle represents proteins in MIPS which have some localization annotation and which can be easily collapsed into a single compartment (e.g. excluding cytoskeletal proteins or proteins with multiple locations; see text; 1935). The left circle represents proteins in YPD which have some localization annotation and which can be easily collapsed into a single compartment (2143). The top circle represents proteins that have "predicted" localization annotation and thus are flagged as low-quality. (Note by definition this cannot intersect the Swiss-Prot circle.)

From these circles, we form four subsets (described as "sets") as follows. Set 1: Proteins that have the same collapsed localization in Swiss-Prot, MIPS and YPD, and have high-quality localization in Swiss-Prot and MIPS. Set 2: Proteins that have high-quality localization in Swiss-Prot, but do not have the same collapsed localization in all of Swiss-Prot, MIPS and YPD (including the proteins that do not have any localization annotation in either MIPS or YPD or both). Set 3: Proteins with high-quality localization in MIPS that have either low-quality or no localization in Swiss-Prot. Set 4: Proteins in MIPS that are annotated as predicted, and that have either low-quality or no localization in Swiss-Prot. From these four sets we simply derived our four training and testing datasets as follows:

Dataset	Formation	Number of Proteins	% Correct Predictions after Cross-validation
Localized-465	Set 1	465	88
Localized-704	Localized-465 + Set 2	704	75
Localized-1342	Localized-704 + Set 3	1342	75
Localized-2013	Localized-1342 + Set 4	2013	72

Our system was independently trained and tested using each of these 4 datasets. In each case, cross-validation was performed using a seven-fold jackknife test, a prior based on the relative proportions of the corresponding dataset (fig 4), entropy localization and the comparison of individual protein predictions with observed locations. The last column of the table denotes the percentage of the total proteins that were predicted to have correct localization after thresholding individual protein state vectors.

One issue with training on these 4 datasets is the degree to which circular logic enters into our analysis. We scrutinized the Swiss-Prot and MIPS localization annotations of all proteins to find if they were experimentally observed to lie in a compartment or if they were predicted or guessed to be present in a location. Our first 3 datasets (Localized-465, Localized-704 and Localized-1342) contain only those proteins that were experimentally

observed to belong to a compartment, and hence circular logic cannot apply to them. As one can see from the table, the results of the cross-validation using these datasets are in fact better than those obtained by using the dataset Localized-2013.

The Localized-1342 dataset has the largest number of proteins that are annotated to have high-quality localization information, and hence this dataset is independent of any circular logic. The cross-validation and extrapolation results in this paper are based on the Localized-1342 set.

FIGURE 3 - THE CELL AND ITS COMPARTMENTS

A schematic of a cell is shown with various compartments.

FIGURE 4 - PRIORS

Various priors are shown in the figure. Each prior (except the composite prior) represents the known compartment populations in the corresponding dataset. The Localized-1342 prior is used for the testing and training procedure to obtain the results shown in this paper. The composite prior is devised by combining the data obtained from MIPS and the Snyder lab. We assume that it fairly represents the relative compartment populations in the entire yeast genome. The composite prior is used for the extrapolation of the locations of the proteins with currently unknown localization.

Here we describe the construction of the composition prior in detail: Due to the particularities of the experimental technique used, the Snyder data accurately estimated the number of nuclear proteins (in a random sampling) but tended to over-assign proteins to the cytoplasm, shifting them from membrane, mitochondrial or secretory compartments (E, T, M). In contrast, the MIPS dataset was biased against cytoplasmic proteins (which are often not annotated with a localization) and towards nuclear proteins. Consequently, for our composite prior we used the relative populations of the integral membrane (T), mitochondrial (M) and secretory pathway (E) proteins in the overall MIPS data for the composite prior. However, to overcome the nuclear bias in the MIPS prior, we used the Snyder prior to estimate the relative population of nuclear proteins in yeast genome (23%). After doing this, the relative population of cytoplasmic proteins is fixed at 36% (by the requirement that the prior sums to 100%). We believed that this new composite prior was unbiased, and that it fairly represented the relative populations of the various compartments in yeast (figure 4).

There is, in addition, much corroboration for our decision to use 23% integral membrane (T) proteins in the composite prior. Comparable results have been found by many investigators in whole-genome surveys of yeast and other completely sequenced genomes (Arkin *et al.*, 1997; Boyd *et al.*, 1998; Gerstein, 1997, 1998b; Gerstein & Hegyi, 1998; Goffeau *et al.*, 1993; Jones, 1998; Rost, 1996; Rost *et al.*, 1995; Tomb *et al.*, 1997; Wallin & von Heijne, 1998). In particular, our membrane-prediction program, which predicted whether a protein had transmembrane helices, indicated that 22% of the proteins in the yeast genome were integral membrane (T) proteins (those with more than one transmembrane helix).

FIGURE 5 - CROSS-VALIDATED PERFORMANCE

Part A: Correct predictions of the individual proteins after thresholding their state vectors. The bars in the front show the percentage of correct predictions (predicted location is the same as the observed location of the protein) for low-entropy Localized-1342 proteins (2/3 of the total dataset). The bars in the back show the percentage correct predictions for all Localized-1342 proteins. Each bar shows the percentage of correct predictions for proteins in each compartment. Membrane (T) and secretory pathway (E) proteins are predicted worse than average, whereas nuclear (N) proteins are predicted extremely well.

Part B: Comparison between the actual known compartment populations of the Localized-1342 set (outer circle), and those obtained from the overall compartment population vector (inner circle, heavy line). The root mean square (RMS) of the difference in the population compartments is calculated using the standard formula

$$R = \frac{1}{\sqrt{Q}} \left\| \vec{\mathbf{N}}_{pred}(L) \,.\, \vec{\mathbf{N}}_{obs}(L) \right\| = \sqrt{\frac{\sum_{L}^{Q})v_{pred}(L) \,.\, v_{obs}(L)^{*2}}{Q}}$$

, where $v_{pred}(L)$ is the predicted population of the compartment L, $v_{obs}(L)$ is the observed population of the compartment L, and Q is the number of compartments ($Q = 5$). Since the total number of proteins in our dataset (U) is 1342, the average population of a compartment $\bar{v} = U / Q = 268.4$. The error rate in overall population prediction (Y) is the ratio of the RMS difference (D) to the average population of compartments ($Y = D/\bar{v}$). Then, the accuracy of prediction $A = 1 – Y = 1 - D/\bar{v}$. For the overall compartment population vector, the RMS is 22.4, the error rate is 8%, and the accuracy is 92%. If we threshold individual state vectors of all Localized-1342 proteins and perform similar calculations, we obtain 74% accuracy.

FIGURE 6 - ANALYSIS OF INDIVIDUAL PROTEIN PREDICTIONS

Part A: Variation of the entropy of proteins with coverage during the cross-validation of the Localized-1342 dataset. We consider proteins in an increasing order of their entropies. Hence, the coverage (the fraction G from 0 to 1 of the dataset) is from low-entropy to high-entropy proteins. The entropy for the state vector of protein *m* is given by the formula $S(\vec{\mathbf{P}}_m) = . \sum_{loc} p_m(L) . \ln p_m(L)$, where $p_m(L)$ is the probability that protein *m* lies in compartment *L.* The equation of the trend-line for the graph is S = 1.5G - 0.096, where S is the entropy and G is the coverage.

Part B: Variation of the error rate with coverage. The error rate is the ratio of the number of wrong individual localization predictions to the total number of proteins U in the dataset. The equation of the trend-line is Y = 30G + 6.6, where Y is the error rate (as a percentage) and G is the coverage. Variation of the entropy with the error rate (not shown here) can be described by the equation of the trend-line S = 0.044Y – 0.3, where S is the entropy and Y is the error rate.

FIGURE 7 - EXTRAPOLATION

Part A: Comparison between the relative compartment populations of the Unknown-4700 proteins as expected from the known populations of the Localized-1342 proteins (outer circle), and as obtained from the overall compartment population vector (inner circle, heavy line). The composite prior is used to calculate the overall compartment population vector.

Part B: Comparison between the relative compartment populations of the Unknown-4700 proteins as obtained by thresholding individual state vectors (outer circle), and as obtained from the overall compartment population vector (inner circle, heavy line – same as in part A). The composite prior is used to calculate both. In the thresholding procedure, 1/3 of the Unknown-4700 proteins (those with entropy values greater than 1.12) are left unlocalized and are shown with horizontal lines.

Part C: Estimate of the relative compartment populations in the entire yeast genome. The compartment populations were calculated by adding the observed compartment populations of the Localized-1342 proteins, and the overall compartment population vector of the Unknown-4700 proteins.

REFERENCES

Andrade, M., O'Donoghue, S. & Rost, B. (1998). Adaptation of Protein Surfaces to Subcellular Location. *Journal of Molecular Biology* **276**, 517-525.

Arkin, I., Brunger, A. & Engelman, D. (1997). Are there dominant membrane protein families with a given number of helices? *Proteins* **28**, 465-466.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45-8.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F. & Cullin, C. (1993). A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucl. Acids Res.* **21**, 3329-3330.

Boyd, D., Schierle, C. & Beckwith, J. (1998). How many membrane proteins are there? *Prot. Sci.* **7**, 201-205.

Claros, M. G., Brunak, S. & von Heijne, G. (1997). Prediction of N-terminal protein sorting signals. *Curr Opin Struct Biol* **7**, 394-8.

Claros, M. G. & Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem* **241**, 779-86.

DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-6.

Efron, B. & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science* **1**, 54-77.

Engelman, D. M., Steitz, T. A. & Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annual Review of Biophysics & Biophysical Chemistry* **15**, 321-53.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.

Friedman, N., Geiger, D. & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning* **29**, 131-163.

Frishman, D., Heumann, K., Lesk, A. & Mewes, H. W. (1998). Comprehensive, comprehensible, distributed and intelligent databases: current status. *Bioinformatics* **14**, 551-61.

Frishman, D. & Mewes, H.-W. (1997). PEDANTic genome analysis. *Trends in Genetics* **13**, 415-416.

Gerstein, M. (1997). A Structural Census of Genomes: Comparing Eukaryotic, Bacterial and Archaeal Genomes in terms of Protein Structure. *J. Mol. Biol.* **274**, 562-576.

Gerstein, M. (1998a). How Representative are the Known Structures of the Proteins in a Complete Genome? A Comprehensive Structural Census. *Folding & Design* **3**, 497-512.

Gerstein, M. (1998b). Patterns of Protein-Fold Usage in Eight Microbial Genomes: A Comprehensive Structural Census. *Proteins* **33**, 518-534.

Gerstein, M. & Hegyi, H. (1998). Comparing Microbial Genomes in terms of Protein Structure: Surveys of a Finite Parts List. *FEMS Microbiology Reviews* **22**, 277-304.

Gerstein, M., Lin, J. & Hegyi, H. (2000). Protein Folds in the Worm Genome. *Pac. Symp. Biocomp.* **5**, 30-42

Goffeau, A., Slonimski, P., Nakai, K. & Risler, J. L. (1993). How Many Yeast Genes Code for Membrane-Spanning Proteins? *Yeast* **9**, 691-702.

Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19**, 1720-30.

Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* **288**, 147-64.

Hodges, P. E., McKee, A. H., Davis, B. P., Payne, W. E. & Garrels, J. I. (1999). The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res* **27**, 69-73.

Holstege, F. C. P., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S. & Young, R. A. (1998). Dissecting the Regulatory Circuitry of a Eukaryotic Genome. *Cell* **95**, 717-728.

Jansen, R. & Gerstein, M. (2000). Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res* **28**, 1481-1488.

Jones, D. T. (1998). Do transmembrane protein superfolds exist? *FEBS Lett* **423**, 281-5.

Klein, P., Kanehisa, M. & DeLisi, C. (1985). The detection and classification of membrane-spanning proteins. *Biochim Biophys Acta* **815**, 468-76.

Ladunga, I., Czako, F., Csabai, I. & Geszti, T. (1991). Improving signal peptide prediction accuracy by simulated neural network. *Comput Appl Biosci* **7**, 485-7.

Mewes, H. W., Hani, J., Pfeiffer, F. & Frishman, D. (1998). MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Res* **26**, 33-7.

Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S. & Frishman, D. (1999). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **27**, 44-8.

Milanesi, L., Muselli, M. & Arrigo, P. (1996). Hamming-Clustering method for signals prediction in 5' and 3' regions of eukaryotic genes. *Comput Appl Biosci* **12**, 399-404.

Nakai, K. & Horton, P. (1996). A Probabilistic Classification System for Predicting the Cellular Localization Sites of Proteins. *Intelligent Systems for Molecular Biology* **4**, 109-115.

Nakai, K. & Horton, P. (1997). Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier. *Intelligent Systems for Molecular Biology* **5**, 147-152.

Nakai, K. & Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24**, 34-6.

Nakai, K. & Kanehisa, M. (1991). Expert System for Predicting Protein Localization Sites in Gram-Negative Bacteria. *PROTEINS: Structure, Function, and Genetics* **11**, 95-110.

Nakai, K. & Kanehisa, M. (1992). A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells. *Genomics* **14**, 897-911.

Nielsen, H., Brunak, S. & von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng* **12**, 3-9.

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Sys.* **8**, 581-599.

Pitman, J. (1997). *Probability.* Springer, New York.

Reinhardt, A. & Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Research* **26**, 2230-2236.

Ross-Macdonald, P., Coelho, P. S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K. H., Sheehan, A., Symoniatis, D., Umansky, L., Heidtman, M., Nelson, F. K., Iwasaki, H., Hager, K., Gerstein, M., Miller, P., Roeder, G. S. & Snyder, M. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption [see comments]. *Nature* **402**, 413-8.

Rost, B. (1996). PHD: Predicting One-dimensional Protein Secondary Structure by Profile-Based Neural Networks. *Meth. Enz.* **266**, 525-539.

Rost, B., Fariselli, P., Casadio, R. & Sander, C. (1995). Prediction of helical transmembrane segments at 95% accuracy. *Prot. Sci.* **4**, 521-533.

Rost, B. & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins* **20**, 216-226.

Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. (1996). Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy [see comments]. *Nat Genet* **14**, 450-6.

Sipos, L. & von Heijne, G. (1993). Predicting the topology of eukaryotic membrane proteins. *Eur J Biochem* **213**, 1333-40.

Sobel, S. G. & Snyder, M. (1995). A highly divergent gamma-tubulin gene is essential for cell growth and proper microtubule organization in Saccharomyces cerevisiae. *Journal of Cellular Biology* **131**, 1775-88.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998). Comprehensive Identification of Cell Cycle-

regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol Biol Cell* **9**, 3273-97.

Stryer, L. (1996). *Biochemistry*. 4. W. H. Freeman and Company, New York.

Tomb, J.-F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karpk, P. D., Smith, H. O., Fraser, C. M. & Venter, J. C. (1997). The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature* **388**, 539-547.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. & Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett , D. E. J., Hieter, P., Vogelstein, B. & Kinzler, K. W. (1997). Characterization of the Yeast Transcriptome. *Cell* **88**,

von Heijne, G. (1986). Net N-C charge imbalance may be important for signal sequence function in bacteria. *J Mol Biol* **192**, 287-90.

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225**, 487-94.

von Heijne, G., Nielson, H., Engelbrecht, J. & S., B. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* **10**, 1-6.

Wach, A., Brachat, A., Pohlmann, R. & Philippsen, P. (1994). New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. *Yeast* **10**, 1793-808.

Wallin, E. & von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* **7**, 1029-38.

Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing Annotation Transfer for Genomics: Quantifying the Relations between Protein Sequence, Structure and Function through Traditional and Probabilistic Scores. *J Mol Biol* **297**, 233-249.

Wolf, E., Kim, P. S. & Berger, B. (1997). MultiCoil: a program for predicting two- and three-stranded coiled coils [In Process Citation]. *Protein Sci* **6**, 1179-89.