MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins

Chittibabu Guda¹, Eoin Fahy¹,4 and Shankar Subramaniam¹,2,3,*

¹San Diego Supercomputer Center and ²Department of Bioengineering, ³Department of Chemistry and Biochemistry, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA and ⁴MitoKor, 11494 Sorrento Valley Road, San Diego, CA 92121, USA

Received on July 31, 2003; revised on November 4, 2003; accepted on November 7, 2003
Advance Access publication March 22, 2004

ABSTRACT

Motivation: Currently available methods for the prediction of subcellular location of mitochondrial proteins rely largely on the presence of mitochondrial targeting signals in the protein sequences. However, a large fraction of mitochondrial proteins lack such signals, making those tools ineffective for genome-scale prediction of mitochondrial-targeted proteins. Here, we propose a method for genome-scale prediction of nucleus-encoded mitochondrial proteins. The new method, MITOPRED, is based on the Pfam domain occurrence patterns and the amino acid compositional differences between mitochondrial and non-mitochondrial proteins.

Results: MITOPRED could predict mitochondrial proteins with 100% specificity at a 44% sensitivity rate and with 67% specificity at 99% sensitivity. Additionally, it was sufficiently robust to predict mitochondrial proteins across different eukaryotic species with similar accuracy. Based on Matthews correlation coefficient measure, the prediction performance of MITOPRED is clearly superior (0.73) to those of the two popular methods TargetP (0.51) and PSORT (0.53). Using this method, we predicted the nucleus-encoded mitochondrial proteins from six complete genomes (three invertebrate, two vertebrate and one plant species) and estimated the total number in each genome. In human, our method estimated the existence of 1362 mitochondrial proteins corresponding to 4.8% of the total proteome.

Availability: MITOPRED program is freely accessible at http://mitopred.sdsc.edu. Source code is available on request from the authors.

Contact: shankar@ucsd.edu

Supplementary information: Training data sets are also available at http://mitopred.sdsc.edu

INTRODUCTION

Mitochondria are very interesting organelles, not only due to their unique role in a variety of cellular functions but also because they are under the control of two genomes, their own and that of the host cell nucleus. Functionally, mitochondria act as the powerhouse of the cell and hence are essential for the maintenance of basic processes such as energy metabolism. Mitochondria are also involved in complex biochemical processes such as programmed cell death (Gottlieb, 2000) and ionic homeostasis (Jassem et al., 2002). As a result, mitochondria are the functional centers for proteins associated with over 100 known human diseases (http://www.neuro.wustl.edu/neuromuscular/mitosyn.html), including Alzheimer’s disease (Hutchin and Cortopassi, 1995), Type II diabetes (Gerbitz et al., 1996) and Parkinson’s disease (Wooten et al., 1997). The human mitochondrial genome is a small, circular DNA encoding 37 genes, of which 24 are RNAs and only 13 are proteins (all 13 are subunits of the respiratory chain). Hence, virtually all proteins functioning in mitochondria are nucleus-encoded and are imported from the cytoplasm. Current estimates of the number of such proteins, based on visualization by two-dimensional (2D) gel electrophoresis, are predicted to be of the order of 1500 sequences (Lopez et al., 2000). Our database (http://www.mitoproteome.org) has already cataloged ~1000 human mitochondrial protein sequences obtained either from public sources or from experimental LC/MS/MS studies (Taylor et al., 2003).

Since several functional domains are shared across different cellular organelles, it is a challenge to predict the proteins destined to function in a specific subcellular location. Some mitochondrial proteins carry cleavable ‘matrix targeting signals’ (MTTs) predominantly at their N-terminal regions. In yeast, such proteins account for only 25% of all known or predicted mitochondrial proteins (Marcotte et al., 2000), suggesting that the N-terminal targeting
mitochondrial proteins across different eukaryotic species. Could be used for genome-scale prediction of nucleus-encoded proteins, and the method is robust, indicating that the prediction performance of this method is superior to the performance of currently available methods. Moreover, this method is restricted to proteins that contain MTS sequences.

It is evident that any prediction method based on a single criterion is not sufficient to address fully the complex mechanisms involved in the transport of proteins into mitochondria. We have developed a method, referred to henceforth as MITOPRED, based on multiple criteria such as location-specific Pfam (http://pfam.wustl.edu) domains and their occurrence patterns, amino acid composition and the PI value differences between mitochondrial and non-mitochondrial locations. Our results indicate that the prediction performance of this method is superior to the performance of currently available methods. Moreover, this method is robust and hence could be used for genome-scale prediction of nucleus-encoded mitochondrial proteins across different eukaryotic species.

METHODS

Data refinement and sorting
We extracted ∼35 000 eukaryotic protein sequences containing annotation for 'subcellular localization' from SWISSPROT release 40.41 (http://www.ebi.ac.uk/swissprot). This set was filtered to ∼19 000 by removing the entries with low confidence annotations such as 'by similarity', 'potential', 'probable' and 'possible'. Sequences were separated into plant and non-plant groups (including yeast and animal proteins), and within each group sequences were sorted into different subcellular locations based on their annotation. Under each location, sequences were clustered at 90% identity using the CD-HIT program (Li et al., 2001) to obtain relatively accurate average residue compositions by avoiding over- or underrepresentation of certain residues from highly homologous sequences. Clustering at sequence identities lower than 90% produces smaller and diverse data sets, resulting in the loss of identification of some unique Pfam domains (data not shown). Only those locations containing at least 100 sequences were included in the final data set. The number of sequences under each location in the final set is presented in Figure 1.

Algorithm implementation
This algorithm contains two distinct scoring methods, and the final prediction is based on the combined score, as shown in Figure 2. Scores are calculated based on the occurrence patterns of Pfam domains and the differences in the amino acid composition and the PI value of proteins from mitochondrial and non-mitochondrial locations.

Analysis of Pfam domains and their occurrence patterns
Pfam, a database of protein families based on HMMs, hosts the largest public collection of protein functional domains (Bateman et al., 2002). We used Pfam version 7.8, which contains 5049 HMM models, each representing one Pfam domain. All mitochondrial and non-mitochondrial protein sequences were searched against the Pfam database (Pfam-A) using the 'hmmsearch' program supplied in the HMMER package (Eddy, 1998), at an E-value threshold of 1 × 10^−5. Unique Pfam domains from the mitochondrial and non-mitochondrial sets were classified and separated into three categories, i.e. 'Mito only'—domains that occur exclusively in mitochondria (238 domains); 'Non-mito only'—domains that occur exclusively in non-mitochondrial locations (1024 domains) and 'Shared'—domains (217 domains) found both in mitochondria and non-mitochondrial locations.

Fig. 1. Number of sequences used in the training set in each subcellular location. CYT, cytoplasmic; NUC, nuclear; MIT, mitochondrial; EXC, extracellular or secretory; END, endoplasmic reticulum; PLA, plasma membrane; GOL, Golgi complex; and POX, peroxisomal.
Amino acid compositions (AACs) were calculated separately, for the N-terminal (NT) 25 residues and for the rest of the C-terminal (CT) sequence of each protein sequence using Equation (1), where

\[
f_i = 100 \frac{n_i}{N} \quad i = 1, 2, 3, \ldots, 20,
\]

where, \(n_i\) is the frequency of amino acid \(i\) and \(N\) is the number of amino acid residues in the protein sequence.

Amino acid compositions (AACs) were calculated separately, for the N-terminal (NT) 25 residues and for the rest of the C-terminal (CT) sequence of each protein sequence using Equation (1), where

\[
N = \begin{cases} 
25, & \text{if NT} \\
(L - 25), & \text{if CT}
\end{cases}
\]

and \(L\) is the length of the full protein. Average AACs for NT and CT sequences were calculated for each subcellular location (Table 1).

Calculation of location-specific amino acid weights This program considers eight different locations as shown in Table 1; however, there are 12 effective locations (eight NT and four CT) because separate AACs were calculated for all NT sequences and four CT sequences (mitochondrial, cytoplasmic, nuclear and extracellular). For each non-mitochondrial location (total of 10), location-specific amino acid weights were calculated by comparing its AAC with that of the corresponding mitochondrial location (Table 2). Amino acid weights are represented by the vector

\[
W = [w1, w2, w3, \ldots, w20].
\]

Location-specific weights were calculated using Equation (2).

\[
w_{Li} = \{(f_{Mi} - f_{Li})/f_{Li}\} \times 10, \quad i = 1, 2, 3, \ldots, 20,
\]

where, \(w_{Li}\) is the weight for amino acid \(i\) at location \(L, f_{Mi}\) is the mitochondrial amino acid frequency for residue \(i\) and \(f_{Li}\) is the amino acid frequency for residue \(i\) at location \(L\).

**Calculation of score based on location-specific amino acid weights** For each location, residues with (20% higher or lower) compositional values significantly different from those of mitochondria are determined. The composition of such residues in the query sequence is compared with that of the location average to see how it deviates from that location-specific AAC. For example, if a residue \(X\) is determined to be significantly higher in mitochondria than in the cytoplasm and the compositional value of \(X\) in the query is more than that of the cytoplasmic average, then the square value of the cytoplasmic \(X\) weight is added to the mitochondrial score. Similarly, the same score is added to the cytoplasmic score if the compositional value of \(X\) in the query is less than that of the mitochondrial average. Hence, for each non-mitochondrial location, using location-specific amino acid weights, two separate scores were calculated, one for the mitochondria and the other for the corresponding location. If the mitochondrial score is more than the location score, an arbitrary positive score \((a)\) is assigned, and if not, zero is assigned. The final score \((S)\) is the sum of all the scores from all the locations.

\[
S = \sum_{l=0}^{R} S_l = \begin{cases} 
\alpha, & \text{if } M_{ls} > L_{ls} \\
0, & \text{if } M_{ls} \leq L_{ls}
\end{cases}
\]

\(R\) is the total number of non-mitochondrial locations, \(S_l\) is the arbitrary score for non-mitochondrial location \(l, M_{ls}\) is the mitochondria score with weights from location \(l\) and \(L_{ls}\) is the location score for location \(l\). If the query sequence is mitochondrial, it is expected to have a higher score at each location \(l\), i.e. \(S_l = \alpha\), at all locations.

**Calculation of estimated pI value** pI values were calculated for NT regions as the pH value at which the protein has zero net charge. The net charge in a protein was estimated based on the free \(\alpha\)-amino and \(\alpha\)-carboxyl groups as well as the nature and the number of its ionizable R groups (Nelson and Cox, 2000). In Figure 3, we present the average pI values for each subcellular location.
Table 1. Location-specific amino acid composition for NT and CT sequences

| Location | A  | B  | C  | D  | E  | F  | G  | H  | I  | J  | K  | L  | M  | N  | P  | Q  | R  | S  | T  | V  | W  | Y  |
|----------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| NT       |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| CT       |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |

Fig. 3. Average pl values of NT protein sequences from different cellular locations. CYT, cytoplasmic; MIT, mitochondrial; NUC, nuclear; END, endoplasmic reticulum; EXC, extracellular/secretory; GOL, Golgi complex; PLA, plasma membrane; and POX, Peroxisomal.

Testing the MITOPRED program

Preparation of training and testing data

Non-plant sequences were used for training and testing the program. Total sequences were randomly divided into four equal sets, and within each set, training and testing sequences were separated in a 9:1 ratio, respectively. We created four distinct training sets by pooling training data from three sets at a time and leaving one out. Each training set was cross-validated by testing with the testing data from the leftover set.

Confusion matrix method

Using the confusion matrix (Kohavi and Provost, 1998), various measures of quality such as accuracy, precision, specificity, sensitivity and Matthews correlation coefficient (MCC) (Matthews, 1975) were determined. In the following Equations (4)–(10), TP refers to true positives (correctly predicted mitochondrial proteins), TN refers to true negatives (correctly predicted non-mitochondrial proteins), FN refers to false negatives (incorrectly predicted mitochondrial proteins) and FP refers to false positives (incorrectly predicted non-mitochondrial proteins).

\[
\text{accuracy (AC)} = \frac{\text{TP}}{\text{TP} + \text{TN}} \quad (4)
\]

\[
\text{specificity (SP)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (5)
\]

\[
\text{sensitivity (SN)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)
\]

\[
\text{positive precision (PP)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)
\]
negative precision ($P_N$): proportion of predicted negative cases that were correct

$$P_N = \frac{TN}{TN + FN}$$  \hspace{1cm} (8)

estimation of actual positives ($P_{est}$): actual positives in a genome were obtained by dividing TPs with sensitivity.

$$P_{est} = \frac{(FP + TP) \times P_P}{SN}$$  \hspace{1cm} (9)

MCC: This is regarded as a more rigorous measure of evaluating the performance of class prediction methods. MCC equals 1 for perfect predictions, while it is zero for completely random predictions (Matthews, 1975).

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$  \hspace{1cm} (10)

Comparison of MITOPRED performance

The performance of MITOPRED was compared with that of similar programs such as PSORT and TargetP to predict mitochondrial proteins. The PSORT program was downloaded from http://psort.nibb.ac.jp, locally installed and trained with the same data used for training MITOPRED. TargetP predictions were carried out from TargetP Server v1.01 (http://www.cbs.dtu.dk/services/TargetP). Two non-plant data sets were used for testing purposes, one containing known mitochondrial sequences and the other with known non-mitochondrial sequences.

Genome-scale prediction of mitochondrial proteins

We downloaded complete proteome sets of six eukaryotic genomes from the proteome analysis at the EBI resource (http://www.ebi.ac.uk/proteome), representing three invertebrates (yeast, Caenorhabditis elegans and Drosophila), two vertebrates (mouse and human) and one plant species (Arabidopsis). For all species except Arabidopsis, predictions were carried out on the program trained with non-plant data set, whereas for Arabidopsis the plant data set was used for training the program.

RESULTS AND DISCUSSION

Patterns of Pfam domain occurrence and their relevance in scoring

In eukaryotes, many biological pathways are localized to specific cellular locations in part or full, and hence it is logical that certain functional domains occur exclusively in certain compartments. As described in the Methods section, Pfam domains are divided into several groups, i.e. ‘Mito only’, ‘Non-mito only’ and ‘Shared’, based on their occurrence patterns across different cellular locations.

Biologically, exclusive mitochondrial domains participate in pathways exclusive to mitochondria (e.g. electron transport chain, β-oxidation, etc.), and hence it is valid to postulate that sequences possessing any one of the Mito-only domains are mitochondrial, and the same logic may be applied to non-mitochondrial domains. For instance, the electron transfer flavoprotein (ETF) domains ETF-α and ETF-β function as a heterodimer, serving as a specific electron acceptor for various mitochondrial dehydrogenases, and transfer electrons to the main respiratory chain via ETF-ubiquinone oxidoreductase. This pathway is seen only in mitochondria (a similar system also occurs in some bacteria), and hence the ETF domains are considered Mito-only. Such information on the location-specific occurrence patterns of Pfam domains is used in the scoring scheme.

Each query sequence is searched against the Pfam database to find the HMM models (Pfam domains), if any, occurring in that sequence. If one or more Pfam domains are found, an arbitrary positive or negative score is awarded based on the group each domain belongs to, and if not, a score of zero is awarded. In the case of multi-domain proteins, the total Pfam score is the sum of all domain scores. Domains belonging to the Mito only group are rewarded with a positive score (+50), and those that fall into Non-mito only are penalized with a negative score (−50).

Pfam database release 7.8 contains about 5000 functional Pfam domains (HMM models); however, only about 1500 domains are considered in this program because only eukaryotic non-plant sequences with reliable annotation for the subcellular localization have been used in the training set. The prediction rate of this program is expected to improve as more Pfam domains and more information on subcellular localization become available. The limitation of predicting solely based on the Pfam score is that for any given genome about 30–40% of the proteins do not have reliable (E-value ≤ 1 × 10−5) Pfam annotations. Hence, this method also uses other criteria for score calculation such as the AAC and pI values. If the query sequence contains a known Pfam domain, prediction is based on the total score from the Pfam, AAC and pI scores, and if not, the prediction is determined by the AAC and pI scores only. The Pfam and AAC scores complement each other in the final prediction. They are equally weighted to enable a prediction based on the AAC score only, in cases where a protein does not contain a defined Pfam domain. The pI score is considered only in certain occasions as described in the pI score section.

Score based on the amino acid composition

The AAC of mitochondrial sequences differs significantly for certain residues with respect to certain subcellular locations (Table 1). Also, within mitochondrial sequences the AAC varies between NT and CT regions especially at the 25 NT residues (Table 1). Hence, it is important to determine the location-specific and region-specific (NT, CT) differences in the AAC between mitochondrial and non-mitochondrial locations.
The AACs of NT sequences from all non-mitochondrial locations show striking differences compared with their mitochondrial counterparts; however, the same is true only for cytoplasmic, nuclear and extracellular locations in the case of CT sequences. Hence, we calculated AAC for all the eight NT locations and for only four CT locations (Table 1).

Analysis of N-terminal amino acid composition In most cases, signal peptides or target peptides that are crucial for protein transport are located in the NT region and hence we analyzed the AAC of NT 25 residues (NT-AAC) separately from the rest of the sequence (CT-AAC). As shown in Table 1, proteins from each location have a unique compositional pattern compared with mitochondrial proteins. Overall, Gly (G), Ile (I), Met (M), Ser (S) and Thr (T) show minor fluctuations across different locations. In general, hydrophobic residues such as Ala (A), Phe (F) and basic residues such as His (H) and Arg (R) are significantly higher (>20%) in mitochondrial proteins, while Tyr (Y) and acidic residues such as Asp (D) and Glu (E) are significantly lower (<20%). This explains why mitochondrial sequences generally have higher pI values compared with those of any other location (Taylor et al., 2003). Other residues such as Cys (C), Lys (K), Leu (L), Asn (N), Pro (P), Gln (Q), Val (V) and Trp (W) show mixed but significant differences across different locations in comparison with those of mitochondrial proteins.

Analysis of CT-AAC It is known that targeting signals, if any, are not always located at the NT region of mitochondrial proteins. In some cases, there are largely unidentified signals located within the sequence (Herrmann and Neupert, 2000) and in a few cases cleavable signal peptides are located at the CT regions (Lee et al., 1999). NT-AAC alone cannot identify these differences, and hence we also used the CT-AAC, from the 26th residue to the end of the sequence to calculate location- and region-specific weights. As shown in Table 1, C is significantly lower (<20%) in mitochondrial sequences compared with all non-mitochondrial CT locations, while, F, I, L, M, W and Y are significantly higher.

Calculation of AAC score The AAC score is calculated using location-specific residue weights (Table 2) for each of the 10 non-mitochondrial locations, and the total AAC score is the sum of all location scores. The scoring method is designed in such a way that each location contributes to the score equally and independently, i.e. each location contributes either a positive score of a (we used a = 10) or zero, depending on whether the mitochondrial score is more or less than the location score (see Methods section). Thus, a true mitochondrial sequence could achieve a maximum score of 100 and a true non-mitochondrial sequence could score zero. The total AAC score for a sequence (maximum 100) is scaled down to 50 in order to maintain equal weighting with the Pfam score.

Score based on pI values
Protein pI values are calculated from the NT-sequence only when the Pfam score is zero or more. This is because when the Pfam score is negative, it is obvious that the query sequence already possesses a non-mitochondrial Pfam domain and hence should be discouraged from further scoring to avoid FPs. Since we estimated pI values mainly from the charged residues in the NT region, i.e. Cys (C), Asp (D), Glu (E), His (H), Lys (K), Arg (R) and Tyr (Y) (Nelson and Cox, 2000), one may consider the pI score as redundant to the NT-AAC score. However, not all the seven residues show significant compositional differences at all NT locations, and hence addition of the pI score only helps to improve the prediction rate. As shown in Figure 3, the average pI value of mitochondrial proteins (pI = 9.68) is much higher than that from any other location. Hence, if the query sequence has a pI value of 10 or more in the NT region, its pI value is added to the total score.

Evaluation of MITOPRED performance
Unlike most other methods, MITOPRED predicts only two locations (two-class classifier method), i.e. mitochondrial and non-mitochondrial. Predictions can be performed at different levels of specificity and sensitivity by defining various thresholds for the final score, which is the sum of the AAC, Pfam and pI scores. The prediction performance of MITOPRED has been evaluated using various parameters as follows.

Jackknife test To evaluate the accuracy and consistency of predictions, MITOPRED has been trained separately with four distinct training sets as described in the Methods section. However, all predictions use the same Pfam domain database, and the differences are only seen in AAC. Test data from the leftover set is used to test each trained set.

A jackknife test (also called the test of leave-one-out) shows that the relative prediction rates for both positives and negatives are consistent across all sets at all thresholds (data not shown). This indicates that the training data is not biased and that the program may be used to predict mitochondrial proteins across different species.

Accuracy and precision test Since it is clear from the jackknife test that different training data sets are not biased (data not shown), we used the combined data set for the training and testing of the program. At each threshold, the numbers of TPs, TNs, FPs and FNs are calculated, and based on these values, parameters such as accuracy, precision, specificity and sensitivity are determined, using Equations (4–8).

As seen in Table 5, the prediction accuracy ranges from 0.71 to 0.92 at different thresholds. At lower thresholds, the Precision falls rapidly (from 1 to 0.32), while the Recall remains at higher levels (from 0.92 to 0.99), resulting in higher sensitivity. Sensitivity and specificity tests will further help understand the performance of this program.
Table 3. Confusion matrix values and dependent parameters at each threshold value

<table>
<thead>
<tr>
<th>Score threshold</th>
<th>Positives tested (1408)</th>
<th>Negatives tested (8729)</th>
<th>Accuracy</th>
<th>P&lt;sub&gt;P&lt;/sub&gt;</th>
<th>P&lt;sub&gt;N&lt;/sub&gt;</th>
<th>SN</th>
<th>SP</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥70</td>
<td>624</td>
<td>784</td>
<td>8729</td>
<td>0</td>
<td>0.923</td>
<td>1</td>
<td>0.918</td>
<td>0.443</td>
</tr>
<tr>
<td>≥60</td>
<td>732</td>
<td>676</td>
<td>8721</td>
<td>8</td>
<td>0.933</td>
<td>0.989</td>
<td>0.928</td>
<td>0.520</td>
</tr>
<tr>
<td>≥50</td>
<td>902</td>
<td>506</td>
<td>8633</td>
<td>96</td>
<td>0.941</td>
<td>0.904</td>
<td>0.945</td>
<td>0.641</td>
</tr>
<tr>
<td>≥40</td>
<td>1099</td>
<td>309</td>
<td>8383</td>
<td>346</td>
<td>0.935</td>
<td>0.761</td>
<td>0.964</td>
<td>0.781</td>
</tr>
<tr>
<td>≥30</td>
<td>1210</td>
<td>198</td>
<td>7860</td>
<td>869</td>
<td>0.895</td>
<td>0.582</td>
<td>0.975</td>
<td>0.860</td>
</tr>
<tr>
<td>≥20</td>
<td>1295</td>
<td>113</td>
<td>7051</td>
<td>1678</td>
<td>0.823</td>
<td>0.436</td>
<td>0.984</td>
<td>0.920</td>
</tr>
<tr>
<td>≥10</td>
<td>1374</td>
<td>34</td>
<td>6121</td>
<td>2608</td>
<td>0.739</td>
<td>0.345</td>
<td>0.994</td>
<td>0.976</td>
</tr>
<tr>
<td>&gt;0</td>
<td>1394</td>
<td>14</td>
<td>5827</td>
<td>2902</td>
<td>0.712</td>
<td>0.324</td>
<td>0.998</td>
<td>0.990</td>
</tr>
</tbody>
</table>

Values in parentheses indicate the number of proteins in each set.

Specificity and sensitivity test

Specificity and sensitivity are two competing measures of quality for any two-classifier method. While the goal is to maintain both at the highest possible value of 1, it is not possible practically to achieve this, and a reasonable trade-off point should be found, depending upon the goal of the search. Table 3 shows the specificity and sensitivity of MITOPRED at different threshold levels. At the highest specificity (1.0), the sensitivity is at 0.44, and at the highest sensitivity level (0.99), the specificity is reasonable at 0.67. At a mid-range sensitivity level of 0.78, 78% of all positives can be predicted with only 4% FPs. Figure 4 shows the relationship between specificity and sensitivity using a receiver operating characteristic (ROC) plot (Swets, 1988).

MCC test

MCC values have been calculated for predictions at various score thresholds (Table 3). At higher score thresholds (score threshold ≥60), MCC values are slightly lower because at these levels predictions are primarily based on the Pfam score, which is limited by the number of known Pfam domains. The best prediction performance (MCC = 0.733) is obtained at a score threshold of ≥40, where the AAC score alone could play a significant role in the final predictions. However, below this level, the MCC value decreases with the score threshold since the predictions become more sensitive and less specific.

Comparison of MITOPRED performance

The performance of MITOPRED has been compared against two variants of itself as well as against two other popular programs, PSORT and TargetP. Since MITOPRED predictions are primarily based on the AAC and Pfam scores, the two self-variants include predictions based on the AAC score.
Table 4. Comparison of prediction performance by MITOPRED, TargetP and PSORT

<table>
<thead>
<tr>
<th>Prediction method</th>
<th>Test seta (1408)</th>
<th>Test setb (8727)</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MITOPRED</td>
<td>1099 309</td>
<td>8381 346</td>
<td>0.733</td>
</tr>
<tr>
<td>TargetP</td>
<td>507 901</td>
<td>8628 99</td>
<td>0.509</td>
</tr>
<tr>
<td>PSORT</td>
<td>773 635</td>
<td>8296 431</td>
<td>0.534</td>
</tr>
</tbody>
</table>

Predictions were calculated at a specificity threshold of >0.95 in each method.

a Test set containing 1408 known mitochondrial non-plant sequences.
b Test set containing 8727 known non-mitochondrial non-plant sequences.

only or the Pfam score only. In all the cases except TargetP, we used the same training and testing data sets. Since TargetP predictions are based on the presence of a signal peptide, predictions have been done only for a subset of the test set known to possess cleavable signal peptides. The sensitivity and specificity are calculated for each method at different thresholds, and ROC plots are generated (Figure 4). Comparison of the ROC plots shows that MITOPRED has a balanced specificity and sensitivity at all thresholds, while predictions based on the other methods lack one or the other. For example, PSORT predictions have good specificity (0.95), but the highest sensitivity rate is only 0.51. Similarly, predictions based solely on the AAC score lack specificity at higher sensitivity levels, while those based only on the Pfam score are incapable of reaching the highest sensitivity level theoretically due to the lack of Pfam motifs for all sequences. TargetP shows a performance comparable with that of MITOPRED; however, this method could only be used to predict sequences with signal peptides (Table 4) and hence is not useful for genome-scale predictions. Comparison of ROC plots shows that MITOPRED performs better than TargetP even for the predictions using the signal peptide containing dataset (Fig. 4).

MITOPRED, TargetP and PSORT predictions on two non-plant datasets (see Methods section) have been compared at threshold levels equivalent to >95% specificity rate in each program. As seen in Table 4, the MCC for MITOPRED (0.733) is far higher compared with those of TargetP (0.509) and PSORT (0.534), indicating its superior performance in predicting mitochondrial proteins. This is not surprising because TargetP has inherent limitations in predicting sequences lacking targeting peptides, whereas the sensitivity of the PSORT method is not high (Fig. 4). Our method predicted a large number of human mitochondrial proteins that were not predicted by the other two methods. Examples of some human proteins and their association with certain diseases based on online Mendelian inheritance in man (OMIM) annotations (http://www.ncbi.nlm.nih.gov/omim) are presented in Table 5. As more and more functional domains are identified (in the future Pfam database versions), the performance of MITOPRED should improve significantly, whereas in the case of other methods the scope of improvement per se is limited.

Genome-scale prediction of mitochondrial proteins

Mitochondrial proteins are predicted from six complete eukaryotic genomes including three invertebrate (Saccharomyces cervisiae, C.elegans and Drosophila), two vertebrate (mouse and human) and one plant (Arabidopsis) species (Table 6). Genome-scale predictions are carried out for score thresholds of 50 or more because the precision declines rapidly below this point (Table 3). The total number of mitochondrial proteins in each genome has been estimated at a score threshold of 60 because this is the highest precision level where the predictions are based on one or more criteria. At any threshold higher than 60, predictions could be based on a combined score from the Pfam and AAC scores, resulting in reduced sensitivity, whereas at lower thresholds, predictions could be based purely on the AAC score, resulting in reduced specificity. Table 6 shows the total number of estimated mitochondrial proteins in each genome, using Equation (9).

The current method estimates the existence of 613 mitochondrial proteins in yeast, representing about 10% of its genome, and this is close to the previous estimation of about 679 proteins (11% of the genome) based on the phylogenetic profile method (Marcotte et al., 2000). However, Kumar et al. (2002) estimated ~800 proteins (13% of the genome) based on immunological studies. We believe that our estimations are very conservative since they are based on predictions made at a high score threshold (>60). For C.elegans and Drosophila genomes, MITOPRED estimates 887 and 1047 proteins, corresponding to about 4.3 and 6.3% of the genomes, respectively. Compared with C.elegans, the genome fraction of mitochondrial proteins in Drosophila is much higher both due to the reduction in the total genome size and due to the increase in the number of mitochondrial proteins. Our estimation in C.elegans (~4.3% of the genome) matches with that of Marcotte et al. (2000), who reported that only ~4% of its genome represents mitochondrial proteins. In human, MITOPRED estimates about 1362 mitochondrial proteins (4.8% of the genome), and this is supported by the previous estimation of about 1500 proteins in the human mitochondrial (Lopez et al., 2000; Taylor et al., 2003). Also, all the 13 mitochondrial-encoded proteins in human are correctly predicted as mitochondrial, at the highest score threshold of 70 or more. As expected, mouse genome estimations (1282) are close to that of the human (1362), supporting their close evolutionary relationship (Emes et al., 2003). In Arabidopsis, our method estimates about 1187 mitochondrial proteins, and substantial evidence is not available to corroborate this figure. However, Werhahn and Braun (2002) have resolved about 800 mitochondrial proteins in Arabidopsis based on 2D gel analysis. This number may well be a gross underestimation because many hydrophobic proteins
Table 5. Examples of some important human mitochondrial proteins predicted by MITOPRED but not by other methods

<table>
<thead>
<tr>
<th>Swissprot_Id</th>
<th>Description</th>
<th>Complex/pathway</th>
<th>Associated diseasea</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT92_HUMAN</td>
<td>ATPase subunit C</td>
<td>ATP synthase</td>
<td>Batten’s disease</td>
</tr>
<tr>
<td>COXX_HUMAN</td>
<td>Hemo O synthase (COX10)</td>
<td>Heme synthase</td>
<td>Encephalopathy, tubulopathy</td>
</tr>
<tr>
<td>CPNI_HUMAN</td>
<td>Steroid 11-β-hydroxylase</td>
<td>Steroidogenesis</td>
<td>Adrenal hyperplasia—IV</td>
</tr>
<tr>
<td>ETFB_HUMAN</td>
<td>Electron transfer flavoprotein β subunit</td>
<td>Electron transport</td>
<td>Glutaricaciduria-IIB</td>
</tr>
<tr>
<td>GCSP_HUMAN</td>
<td>Glycine cleavage system P-protein</td>
<td>Glycine degradation</td>
<td>Hyperglycinemia</td>
</tr>
<tr>
<td>GPDM_HUMAN</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase</td>
<td>Glyceral metabolism</td>
<td>Diabetes mellitus</td>
</tr>
<tr>
<td>HEM0_HUMAN</td>
<td>5-aminolevulinic acid synthase</td>
<td>Heme biosynthesis</td>
<td>Sideroblastic anemia</td>
</tr>
<tr>
<td>HXX1_UMAN</td>
<td>Hexokinase type I</td>
<td>Glycolysis</td>
<td>Hemolytic anemia</td>
</tr>
<tr>
<td>MUTA_HUMAN</td>
<td>Methylmalonyl-CoA mutase</td>
<td>Amino acid degradation</td>
<td>Methylmalonic aciduria</td>
</tr>
<tr>
<td>ODDB_BUMAN</td>
<td>2-oxoisovalerate dehydrogenase β subunit</td>
<td>Amino acid metabolism</td>
<td>Maple syrup urine disease</td>
</tr>
<tr>
<td>PCCA_HUMAN</td>
<td>Propionyl-CoA carboxylase α subunit</td>
<td>Fatty acid catabolism</td>
<td>Propionic acidemia-I</td>
</tr>
<tr>
<td>PGN_HUMAN</td>
<td>Paraplegin</td>
<td>Metalloprotease</td>
<td>Spastic paraplegia-7</td>
</tr>
<tr>
<td>PROD_HUMAN</td>
<td>Proline oxidase</td>
<td>Proline metabolism</td>
<td>Hyperprolinemia-I</td>
</tr>
<tr>
<td>THTR_HUMAN</td>
<td>Thiosulfate sulfurtransferase</td>
<td>Cyanide detoxification</td>
<td>Leber’s disease</td>
</tr>
</tbody>
</table>

aDisease information is obtained from OMIM database.

Table 6. Genome-scale prediction of mitochondrial proteins using MITOPRED

<table>
<thead>
<tr>
<th>Score threshold</th>
<th>Yeast (6203)</th>
<th>C.elegans (20482)</th>
<th>Drosophila (16674)</th>
<th>Mouse (27709)</th>
<th>Human (28424)</th>
<th>Arabidopsisa (25972)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥70</td>
<td>258</td>
<td>319</td>
<td>379</td>
<td>433</td>
<td>478</td>
<td>114</td>
</tr>
<tr>
<td>≥60</td>
<td>322</td>
<td>466</td>
<td>550</td>
<td>674</td>
<td>716</td>
<td>164</td>
</tr>
<tr>
<td>≥50</td>
<td>566</td>
<td>1073</td>
<td>1093</td>
<td>1604</td>
<td>1694</td>
<td>1131</td>
</tr>
<tr>
<td>Total estimatedb</td>
<td>613</td>
<td>887</td>
<td>1047</td>
<td>1282</td>
<td>1362</td>
<td>1187</td>
</tr>
<tr>
<td>Per cent of genome</td>
<td>9.9%</td>
<td>4.3%</td>
<td>6.3%</td>
<td>4.6%</td>
<td>4.8%</td>
<td>4.6%</td>
</tr>
</tbody>
</table>

Values in parentheses indicate the number of tested proteins in each genome.

aArabidopsis genome was predicted using training data only from plant species that also include chloroplast-localized sequences. In the case of plants, since the training data are limited, with fewer Pfam annotations, estimations were determined based on the predictions at a score threshold of ≥55, where most predictions are based on the AAC score.

bEstimation of total number of mitochondrial proteins is based on the predictions at a score threshold of ≥60, except in the case of Arabidopsis (see above).

are not separated in the isoelectric focusing (IEF) dimension and also because low-abundance proteins are either not visible on these gels or are masked by high abundance proteins (Taylor et al., 2003).

Genome databases such as Saccharomyces genome database (SGD, http://genome-www.stanford.edu/Saccharomyces) provide Gene Ontology (GO, http://www.geneontology.org) annotations for subcellular locations, and in the case of yeast about 450 mitochondrial proteins are annotated. Recent experimental investigations in other species such as rat (Fountoulakis et al., 2002), human (Taylor et al., 2003) and Arabidopsis (Werhahn and Braun, 2002) have been providing more information about the number of mitochondrial proteins. However, exhaustive experimental studies similar to those carried out in yeast (Kumar et al., 2002; Huh et al., 2003) are still awaited in these species. Since most prediction programs including ours rely on training data, the performance of these programs will be improved considerably as more and more information on mitochondrial proteins accumulates from experimental studies.

Additionally, our method does not require the presence of signal peptide sequences and hence could be used for genome-scale prediction and is sufficiently robust to use against all eukaryotic species without having been trained by species-specific data.

ACKNOWLEDGEMENTS

The authors are thankful to Dr Roderick Capaldi (University of Oregon) for technical discussions and Dr Giridhar Chukkapalli (San Diego Supercomputer Center) for assistance with running genome-scale HMM jobs. This project is supported by the University of California Life Sciences Informatics (LSI) Program/Mitokor grant (L.99-10077).

REFERENCES
