Sequence analysis

TARGET: a new method for predicting protein subcellular localization in eukaryotes

Chittibabu Guda\textsuperscript{1,2,*} and Shankar Subramaniam\textsuperscript{3,4,5}

\textsuperscript{1}Genomic Center for Excellence in Cancer Genomics, \textsuperscript{2}Department of Epidemiology and Biostatistics, University at Albany, State University of New York, One Discovery Drive, Rensselaer, NY 12144-3456, USA and \textsuperscript{3}San Diego Supercomputer Center, \textsuperscript{4}Department of Bioengineering and \textsuperscript{5}Department of Chemistry and Biochemistry, University of California, San Diego, San Diego, CA 92093, USA

Received on May 3, 2005; revised on and accepted on August 26, 2005

Advance Access publication September 6, 2005

ABSTRACT

\textbf{Motivation:} There is a scarcity of efficient computational methods for predicting protein subcellular localization in eukaryotes. Currently available methods are inadequate for genome-scale predictions with several limitations. Here, we present a new prediction method, pTARGET that can predict proteins targeted to nine different subcellular locations in the eukaryotic animal species.

\textbf{Results:} The nine subcellular locations predicted by pTARGET include cytoplasm, endoplasmic reticulum, extracellular/secretory, golgi, lysosomes, mitochondria, nucleus, plasma membrane and peroxisomes. Predictions are based on the location-specific protein functional domains and the amino acid compositional differences across different subcellular locations. Overall, this method can predict 68–87% of the true positives at accuracy rates of 96–99%. Comparison of the prediction performance against PSORT showed that pTARGET prediction rates are higher by 11–60% in 6 of the 8 locations tested. Besides, the pTARGET method is robust enough for genome-scale prediction of protein subcellular localizations since, it does not rely on the presence of signal or target peptides.

\textbf{Availability:} A public web server based on the pTARGET method is accessible at the URL http://bioinformatics.albany.edu/~ptarget. Datasets used for developing pTARGET can be downloaded from this web server. Source code will be available on request from the corresponding author.

\textbf{Contact:} cguda@albany.edu

\textbf{Supplementary data:} Accessible as online-only from the publisher.

INTRODUCTION

Protein subcellular localization, consequent to protein sorting or protein trafficking, is a key functional characteristic of proteins. The eukaryotic cell is a highly ordered structure where nucleus-encoded proteins are synthesized in the cytoplasm and all non-cytosolic proteins are transported to their destined subcellular locations. Subcellular localization of proteins in the intended compartments is vital for the structural and functional integrity of the cell. Therefore, comprehensive knowledge on the subcellular localization of proteins is essential for understanding their roles and interacting partners in cellular metabolism. Exhaustive experimental studies have been carried out in yeast to elicit the subcellular localization of the entire proteome (Kumar \textit{et al}, 2002; Huh \textit{et al}, 2003); however, such diligent feats are not practicable in all species. Therefore, experimental annotation of protein subcellular localization is not able to keep up with the large number of sequences that continue to emerge from the genome sequencing projects. To bridge this gap, there is a need to develop faster, accurate and genome-scale computational methods for predicting subcellular localization of proteins.

Several computational methods have been developed over the past decade for predicting subcellular localization of eukaryotic proteins. These methods are broadly classified into four groups. (1) Methods based on the sorting signals rely on the presence of protein targeting or signal peptides that are recognized by location-specific transport machinery to enable their entry (Nielsen \textit{et al}, 1997; Nakai and Horton, 1999; Emanuelsson \textit{et al}, 2000). Among these, PSORT is a popular method (Nakai and Horton, 1999) that could predict proteins targeted to 12 different subcellular locations. Nevertheless, these methods can predict only those proteins with known sorting signals. (2) Methods based on the differences in the amino acid composition or amino acid properties of proteins from different subcellular locations. These methods use hydrophobicity index of amino acids (Feng and Zhang, 2001), amino acid composition (Cedano \textit{et al}, 1997; Reinhardt and Hubbard, 1998; Feng, 2000; Hua and Sun, 2001; Cui \textit{et al}, 2004), etc.; however, the overall prediction accuracy of these methods is rather low. (3) Methods based on lexical analysis of keywords (LOCkey) from the functional annotation of proteins (Nair and Rost, 2002). The reliability of this method depends on the consistency and the accuracy of keyword assignments given to the proteins. (4) The fourth group of prediction methods uses phylogenetic profiles (Marcotte \textit{et al}, 2000), domain projection (Mott \textit{et al}, 2002) or a combination of evolutionary and structural information (Nair and Rost, 2003). But, these methods are useful for predicting only a limited number of locations.

Recently, we published a new prediction method, MITOPRED based on functional domain occurrence patterns and amino acid compositional differences between sequences belonging to different subcellular locations (Guda \textit{et al}, 2004). However, this method can predict only those proteins targeted to mitochondria. Here we present another method, referred to henceforth as pTARGET that can predict proteins targeted to nine different subcellular locations in eukaryotic species.
predicts up to nine subcellular locations in animal species (includes predicting only mitochondrial proteins (Guda et al., 2004). Recently, we published MITOPRED, a variant of this algorithm for ALGORITHM.

**METHODS**

**Data collection and filtering.** We used protein sequences from the SWISS-PROT database release 45.0 (http://www.ebi.ac.uk/swissprot), for training and testing of pTARGET. To obtain high-quality datasets, we filtered the data as follows. (1) Included sequences only from the animal species (includes Fungi and Metazoa) that have annotation for ’subcellular localization’. (2) Removed sequences with ambiguous and uncertain annotations such as ’by similarity’, ’potential’, ’probable’, ’possible’, ’equals’, etc. (3) Removed sequences known to exist in more than one subcellular location such as those that shuttle between cytoplasm and nucleus, etc. In each location, we clustered sequences at 95% identity using the cd-hit program (Li et al., 2001) to remove highly homologous sequences. (4) Finally, we selected only those subcellular locations with at least 100 annotated sequences. These locations include (the number of sequences is shown in parentheses), CYT-cytoplasm (2062), EXC-endoplasmic reticulum (693), NUC-nucleus (3446), PLA-plasma membrane (4162) and POX-peroxisomes (173).

**Calculation of amino acid composition.** For proteins from each location, we calculated the average relative amino acid compositions (AACs) separately for the N-terminal 25 residues (NTAAC) and for the rest of the sequence (CTAAC), as described in Guda et al. (2004).

**Determination of location-specific Pfam domains.** Pfam database (database of protein families, version 16.0) has a collection of 7677 unique protein functional domains built based on Hidden Markov Models (HMMs) (http://pfam.wustl.edu; Bateman et al., 2004). We searched all protein sequences in each location against the Pfam-A database at gathering thresholds using a faster ’hmmpfam’ program (Chukkapalli et al., 2004) modified from the HMMER software (Eddy, 1998). By comparing the occurrence patterns of Pfam domains across nine subcellular locations, we determined the location-specific Pfam domains (Fig. 1).

**Comparison of pTARGET with PSORT.** We downloaded and locally installed the PSORT stand-alone program from the URL http://psort.nibb.ac.jp. The datasets used for training and testing of PSORT are identical to those used for pTARGET.

**ALGORITHM**

Recently, we published MITOPRED, a variant of this algorithm for predicting only mitochondrial proteins (Guda et al., 2004), whereas the current algorithm implements an improved scoring system that predicts up to nine subcellular locations in animal species (includes Fungi and Metazoa). This prediction algorithm calculates two distinct scores, i.e. a score based on the presence or absence of location-specific Pfam domains in a given location (Pfam score) and a score based on the relative amino acid weights calculated from AAC (AAC score). The sum of these two scores is used in the final prediction.

**Score based on Pfam domain occurrence patterns**

Each location has a set of location-specific Pfam domains that are not known to exist in other locations. A query sequence is searched against the Pfam-A database and if any Pfam domains are found, a Pfam score is calculated for each location based on the matching location-specific domains. Pfam score is an arbitrary value (we chose +50 for rewards and −50 for penalties) assigned to locations based on the presence or absence of location-specific domains. For example, protein sequence ‘ABF1_HUMAN’ contains the ‘Homeobox’ domain that is nucleus-specific. If the query sequence contains the Homeobox domain, the Pfam score for nuclear locations is +50 and it is −50 each, for the rest of the locations. If the query sequence contains ‘shared’ domain(s), only those locations in which the domain is shared will get a Pfam score of 0, while the other locations will get −50 since it is a non-specific domain for them. Finally, if the query sequence does not have any known Pfam-A domain, the Pfam score is 0 for all locations, in which case prediction is based on the amino acid composition scores alone.

**Score based on the amino acid composition**

pTARGET program considers 9 subcellular locations and for each location, there are two distinct regions i.e. NT and CT (N- and C-terminal regions), making it 18 effective locations with distinct amino acid compositions (Table 1). We compared the AACs from each location against those of similar regions in the other locations, in all pairwise combinations. For each pairwise comparison, we calculated residue-specific weights using equation (1) and identified the residues whose compositions differ by at least 20% (Table 2).

\[ W_{Ai} = \begin{cases} \frac{|f_{Ai} - f_{Bi}| \times \min(f_{Ai}, f_{Bi})}{10} & i = 1, 2, 3, \ldots, 20 \end{cases} \]

where, \( W_{Ai} \) is the weight for amino acid \( i \) at location \( A \) in comparison with that at location \( B \). \( f_{Ai} \) and \( f_{Bi} \) are relative frequencies of residue \( i \) at location \( A \) and \( B \), respectively. The AAC of a location is represented in a 20-element vector. The total number of pairwise comparisons in all combinations equals to \( 2^{n} \times (n(n-1))/2 \) where, \( n \) is the number of locations \( n = 9 \) with two distinct regions i.e. NT and CT in each location.

AAC scores have been calculated separately for each of the nine locations where the location with highest score wins the prediction. For each current location, there are 16 ‘other’ locations including 8 NT and 8 CT locations, and the AAC score for the current location is the sum of 16 arbitrary scores (either zero or 10), one from each pairwise comparison against ‘other’ locations. In each pairwise comparison, two raw scores are calculated, one for the current location (Equation 2) and the second for the ‘other’ location (Equation 3). Every time the raw score of the current location is higher that of the ‘other’ location, an arbitrary score of 10 is added to the AAC score of the current location; if not, ‘zero’ is added and vice versa. For example, the AAC score for a cytoplasmic location is calculated by comparing the scoring residues in the query AAC against matching residue averages of cytoplasmic AACs or the...
Table 1. Location-specific relative amino acid composition for the N-terminal and C-terminal sequences

| Location | A   | C   | D   | E   | F   | G   | H   | I   | K   | L   | M   | N   | P   | Q   | R   | S   | T   | V   | W   | Y   |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| CYT_NT   | 8.39| 1.56| 4.90| 6.98| 3.59| 7.44| 1.98| 4.80| 6.45| 8.49| 4.56| 3.76| 5.37| 4.36| 5.00| 7.84| 4.81| 6.60| 0.94| 2.17|
| END_NT   | 9.06| 2.54| 5.26| 2.39| 4.84| 6.39| 1.31| 4.54| 3.39| 18.90| 4.90| 1.82| 4.47| 2.54| 3.81| 8.47| 4.46| 7.75| 2.21| 2.11|
| EXC_NT   | 10.01| 4.57| 2.55| 2.99| 4.24| 6.25| 1.63| 4.25| 3.92| 17.18| 4.58| 2.40| 4.85| 3.28| 3.49| 7.68| 4.96| 7.37| 1.58| 2.22|
| GOL_NT   | 8.04| 1.97| 2.81| 3.84| 5.21| 5.75| 1.46| 4.51| 5.03| 14.77| 5.64| 2.75| 5.09| 3.04| 6.13| 8.45| 4.93| 6.74| 1.64| 2.19|
| LYS_NT   | 12.13| 3.23| 1.96| 2.18| 2.70| 7.61| 1.50| 2.45| 2.30| 20.94| 4.47| 1.52| 7.24| 2.62| 5.54| 7.54| 4.08| 6.56| 2.34| 1.19|
| MIT_NT   | 11.17| 1.77| 1.90| 2.24| 4.04| 6.57| 2.09| 4.10| 4.43| 12.03| 5.27| 2.62| 5.18| 3.53| 9.13| 9.24| 5.50| 5.94| 1.47| 1.79|
| NUC_NT   | 8.15| 1.45| 5.11| 6.80| 3.02| 6.83| 2.07| 3.04| 6.09| 7.14| 5.34| 9.51| 69.1| 4.37| 6.68| 9.98| 5.06| 4.85| 0.64| 2.54|
| PLA_NT   | 8.16| 2.60| 3.30| 4.74| 4.56| 4.72| 1.70| 4.11| 2.93| 14.14| 5.38| 3.75| 6.20| 3.63| 4.32| 8.44| 5.82| 5.98| 1.86| 1.77|
| POX_NT   | 10.26| 1.13| 5.46| 4.88| 2.79| 6.15| 2.04| 3.95| 5.46| 9.74| 4.25| 3.70| 6.40| 4.36| 6.07| 8.39| 5.02| 7.04| 0.66| 2.23|

The upper diagonal shows differences in the N-terminal region and the lower diagonal shows differences in the C-terminal region.

Table 2. N-terminal and C-terminal scoring residues differing by at least 20% in their AAC from all-against-all comparison of subcellular locations

| Location | A   | C   | D   | E   | F   | G   | H   | I   | K   | L   | M   | N   | P   | Q   | R   | S   | T   | V   | W   | Y   |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| CYT      | C   | D   | E   | F   | H   | L   | N   | P   | Q   | R   | W   |       |     |     |     |     |     |     |     |     |     |     |
| END      | C   | F   | W   |   |     |     |     |     |     |     |     |     |       |     |     |     |     |     |     |     |     |     |
| EXC      | C   | I   | M   | N   | V   | W   | Y   |   |     |     |     |     |       |     |     |     |     |     |     |     |     |     |     |
| GOL      | G   | K   | W   | F   | Q   |     |     |     |     |     |     |     |       |     |     |     |     |     |     |     |     |     |     |
| LYS      | C   | E   | F   | K   | W   | Y   |   |     |     |     |     |     |       |     |     |     |     |     |     |     |     |     |     |
| MIT      | C   | D   | M   | W   | Y   | F   | I   | L   |   |     |     |     |       |     |     |     |     |     |     |     |     |     |     |
| NUC      | F   | I   | P   | R   | S   | V   | W   |   |     |     |     |     |       |     |     |     |     |     |     |     |     |     |     |
| PLA      | C   | D   | E   | F   | I   | K   | L   | Q   | S   | W   | Y   |       |     |     |     |     |     |     |     |     |     |     |
| POX      | C   | E   | W   | F   | M   |     |     |     |     |     |     |     |       |     |     |     |     |     |     |     |     |     |     |

Gray–NT sequences; white–CT sequences.
‘other’ 16 non-cytoplasmic AACs. In other words, this translates into a higher score for cytoplasmic locations and a lower score for the non-cytoplasmic locations, if the AAC of query sequence is closer to that of the cytoplasmic averages and vice versa. Note that for each comparison, the scoring residues differ depending on the ‘other’ location being compared, since we use only those residue weights that differ by at least 20% in any given vector pair comparison (Table 2). While calculating the cytoplasmic score, residues from the first row of Table 2 are used for N-terminal AAC comparisons; while, residues from the first column of Table 2 are used for C-terminal AAC comparisons. Cytoplasmic (C_c) and ‘other’ location (O_o) scores have been calculated using Equations (2) and (3), respectively.

\[ C_c = \sum_{i=1}^{N} (d_i^*W_{COi}) \quad \text{where, } d_i = \begin{cases} Q_i - O_i, & \text{if } W_{COi} \geq 2 \\ O_i - Q_i, & \text{if } W_{COi} \leq -2 \end{cases} \quad (2) \]

\[ O_o = \sum_{i=1}^{N} (d_i^*W_{COi}) \quad \text{where, } d_i = \begin{cases} C_i - Q_i, & \text{if } W_{COi} \geq 2 \\ Q_i - C_i, & \text{if } W_{COi} \leq -2 \end{cases} \quad (3) \]

where \( W_{COi} \) is the weight for residue \( i \) when the AACs from a cytoplasmic location and location \( O \) are compared, \( Q_i \), \( C_i \) and \( O_i \) are relative frequencies of residue \( i \) in the query sequence, cytoplasmic location and location \( O \), respectively. The final AAC score for the cytoplasmic location (S_c) is the sum of arbitrary scores determined using Equation (4).

\[ S_c = \sum_{a=0}^{R} S_a = \begin{cases} a, & \text{if } C_i > O_i \\ 0, & \text{if } C_i \leq O_i \end{cases} \quad (4) \]

where \( R \) is the number of non-cytoplasmic locations (total 16), \( S_a \) is the score for ‘other’ location \( O \) and \( a \) is an arbitrary value of 10. If the query sequence is cytoplasmic, \( C_i \) is expected to be higher than \( O_i \) at all locations, i.e. the total cytoplasmic score equals to \( R \) times \( a \) (maximum 160). For example, ADH_HUMAN protein is a cytoplasmic enzyme that functions as aldehyde oxidase and this protein gets the maximum score of 160 in the current scoring scheme. Likewise, the final AAC score for each location is calculated and adjusted to a maximum score of 50 in order to equalize it with the Pfam score.

Using Pfam and AAC scores in the prediction

The sum of Pfam and AAC scores is used in the prediction; however, their relative contribution in the final prediction vary depending on the presence, absence, shared or unknown nature of the Pfam domains in the query sequence. In a nutshell, (1) when a query sequence contains at least one location-specific domain, the Pfam score itself is enough to make a prediction; (2) when a query sequence has domain(s) shared across multiple locations, the combined score is necessary for prediction and (3) when a query sequence has no known domain(s), the prediction is entirely based on the AAC score. A detailed explanation of this process with actual scores and examples is provided in Supplementary Table 3.

Algorithm testing

We used various measures of quality including specificity, sensitivity and Mathew’s correlation coefficients (MCC) for testing the algorithm, as described in Guda et al. 2004. To characterize the prediction performance for individual locations, we used the ROC (Receiver Operating Characteristic) plots (Swets, 1988).

IMPLEMENTATION

Analysis of Pfam domain occurrence patterns

Eukaryotic cells are organized into a complex network of membranes and compartments where metabolic pathways are distributed across different subcellular locations. Since, enzymes or proteins involved in these pathways contain one or more functional domains (Pfam domains), by keeping track of the functional domains specific to a location, it is possible to predict the location of a protein that contains such domains. We analyzed about 23000 protein sequences from the SWISSPROT database containing subcellular location information (from empirical studies) and determined unique Pfam domains specific to each of the nine locations (Fig. 1). A query sequence is searched against the Pfam database to find if any Pfam domains are present in that sequence. The Pfam score is calculated for each location depending on the presence or absence of matching location-specific Pfam domains in the query sequence. For multidomain proteins, the total Pfam score is the sum of all domain scores; however, the presence of one location-specific Pfam domain is enough to assign a query protein to that location.

The Pfam-A database release 16.0 contains about 7677 functional domains (HMM models), yet we used only 2146 unique domains in this program because only the eukaryotic and non-plant sequences were used in the dataset. The limitation of predicting solely based on Pfam score is that for any given genome, ~30-40% of the proteins do not have reliable Pfam-A annotations at gathering thresholds, and some functional domains are shared across multiple subcellular locations. To predict such proteins, the current method uses AAC differences across different subcellular locations in the scoring system.

Analysis of AAC differences across different subcellular locations

It has been known that protein sorting usually relies on the presence of N-terminal targeting sequences that are recognized by location-specific translocation machinery (Rusch and Kendall, 1995). To take full advantage of such targeting signals, we analyzed the AAC of N-terminal 25 residues (NT) separately from the rest of the C-terminal (CT) sequence (Table 1). We determined the AAC differences across different locations in all pairwise combinations (36 pairs) for 9 subcellular locations and chose only those residues showing at least 20% difference, as the scoring residues (Table 2, also Fig. 1 in the Supplementary data). Inclusion of residues with fewer than 20% differences in the scoring system lowered the prediction performance of this method (data not shown).

Analysis of AAC from different subcellular locations revealed remarkable differences in the NT region compared to the CT region (Table 2) because the targeting signals are mostly found in the N-terminal region except for the endoplasmic reticulum and peroxisomal proteins where KDEL/HDEL and SKL signals, respectively, are found at the C-terminus (Stornaiuolo et al., 2003; Subramani et al., 2000). For the mitochondria or other organelles involved in the secretory pathway (Endoplasmic reticulum → Golgi → Lysosomes → Extracellular), N-terminal target peptides are identified based on the cleavage sites (Emanuelsson et al.,...
Prediction of protein subcellular localization

Evaluation of the prediction performance

For each location we used two test sets; the first one is all known positives and the second set is all known negatives for that location. We evaluated pTARGET’s performance in predicting nine subcellular locations based on specificity and sensitivity, MCC values (Table 3) and ROC plots (Fig. 2). We also determined the rates of false positives (FPs) and false negatives (FNs) using proteins from all-against-all locations (Table 1 in the Supplementary data). pTARGET can make predictions at different score thresholds resulting in different values for the evaluation parameters stated above. Score threshold of 50 is a cutoff where predictions could be either from the Pfam score alone or from the AAC score alone, while 1 is the lowest possible score.

Specificity and sensitivity test

Specificity and sensitivity are two competing but non-exclusive measures of quality useful for testing the performance of classification methods. An ideal classification method should have both values close to 1. As shown in Table 3, the maximum sensitivity of pTARGET ranges from 0.68 (GOL) to 0.87 (MIT) at the lowest score threshold of 1, while for all but the GOL location sensitivity rates peaked above 0.75. At the other end, specificity rates are almost perfect (~1) for all locations at the highest score threshold of 50, while at the highest sensitivity level (score threshold of 1) the specificity rates are still above 0.96. In other words, the worst case false positive rate expected for any location would not be >4%.

Figure 2 shows the relationship between specificity and sensitivity using ROC plots. For all but CYT locations, the ROC curves climb rapidly towards the upper left hand corner of the graph which is a good characteristic of ROC plots. This shows that the pTARGET program has high sensitivity as well as high specificity. The overall prediction performance of pTARGET is the lowest for cytoplasmic proteins. This is probably because CYT is the default location for protein synthesis as well as the hub of cellular core metabolism and, therefore, it is likely to have the most number of ‘shared’ functional domains thus negatively affecting the prediction performance.

Matthew’s correlation coefficient test

MCC provides a single measure of evaluating specificity and sensitivity together, where it equals one for perfect predictions and zero for random assignments (Matthew, 1975). At the highest specificity level (score ≥ 50), MCC values for different locations range from 0.53 to 0.76, while at the highest sensitivity level (Score ≥ 1) the range is between 0.50 to 0.84 (Table 3).
Our results suggest that the prediction performance of pTARGET is consistent and better than that of PSORT, for most of the locations tested. Unlike PSORT, the current method is sufficiently robust for genome-scale prediction of proteins in eukaryotic animal species and does not require species-specific training datasets. Previously, we used MITOPRED for genome-scale prediction of mitochondrial proteins in six eukaryotic proteomes (Guda et al., 2004). One of the limitations of pTARGET is its inability to accurately predict proteins localized in multiple locations such as those shuttling between cytoplasm and nucleus. Based on the number of ‘shared’ domains in our study (~500, data not shown), we estimate that in eukaryotic proteomes, at least 20% of the proteins are localized to multiple locations. In the future, we will focus on developing sophisticated scoring methods to accurately predict proteins targeted to multiple locations.

ACKNOWLEDGEMENTS

Authors are thankful to Dr. Giridhar Chukkapalli at the San Diego Supercomputer Center for assistance in running genome-scale HMM jobs. This project has been supported by the start-up funds to CG from the State University of New York at Albany and the University of California Life Sciences Informatics (LSI) Program/Mitokor grant (L99-10077) to SS.

Conflicts of Interest: none declared.

REFERENCES


