BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 8 2013, pages 981–988
doi:10.1093/bioinformatics/btt089

Sequence analysis

Advance Access publication February 21, 2013

The prediction of organelle-targeting peptides in eukaryotic proteins with Grammatical-Restrained Hidden Conditional Random Fields

Valentina Indio[1,2], Pier Luigi Martelli[1,3,*], Castrense Savojardo[1,4], Piero Fariselli[1,4] and Rita Casadio[1,2,3]

[1]Biocomputing Group, University of Bologna, 40126 Bologna, [2]Giorgio Prodi Interdepartmental Center for Cancer Research, University of Bologna, 40138 Bologna, [3]Department of Biology and [4]Department of Computer Science and Engineering, University of Bologna, 40126 Bologna, Italy

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Targeting peptides are the most important signal controlling the import of nuclear encoded proteins into mitochondria and plastids. In the lack of experimental information, their prediction is an essential step when proteomes are annotated for inferring both the localization and the sequence of mature proteins.

Results: We developed TPpred a new predictor of organelle-targeting peptides based on Grammatical-Restrained Hidden Conditional Random Fields. TPpred is trained on a non-redundant dataset of proteins where the presence of a target peptide was experimentally validated, comprising 297 sequences. When tested on the 297 positive and some other 8010 negative examples, TPpred outperformed available methods in both accuracy and Matthews correlation index (96% and 0.58, respectively). Given its very low–false-positive rate (3.0%), TPpred is, therefore, well suited for large-scale analyses at the proteome level. We predicted that from ~4 to 9% of the sequences of human, *Arabidopsis thaliana* and yeast proteomes contain targeting peptides and are, therefore, likely to be localized in mitochondria and plastids. TPpred predictions correlate to a good extent with the experimental annotation of the subcellular localization, when available. TPpred was also trained and tested to predict the cleavage site of the organelle-targeting peptide: on this task, the average error of TPpred on mitochondrial and plastidic proteins is 7 and 15 residues, respectively. This value is lower than the error reported by other methods currently available.

Availability: The TPpred datasets are available at http://biocomp.unibo.it/~valentina/TPpred/. TPpred is available on request from the authors.

Contact: gigi@biocomp.unibo.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 30, 2012; revised on January 15, 2013; accepted on February 18, 2013

1 INTRODUCTION

Mitochondria and plastids (in plants) are membrane-enclosed organelles contained in eukaryotic cells that take part into essential biological processes, including cell bioenergetics and metabolism. Following the endosymbiotic theory, the DNA-containing organelles derive from free-living bacteria that have been incorporated into the cytoplasm of early eukaryotic cells. The organelles retain a small portion of their original genome, encoding few tens and few hundreds proteins in mitochondria and plastids, respectively. However, both experimental and computational methods estimate that some thousands different proteins are located in organelles (Sickmann *et al.*, 2003; van Wijk, 2004). The last release of MitoMiner lists 2755 human proteins experimentally characterized as mitochondrial with mass spectrometry or with green-fluorescent protein tagging (Smith *et al.*, 2012), and the AT_CHLORO database lists 1323 *Arabidopsis thaliana* proteins annotated as chloroplastic on the basis of mass spectrometry experiments (Ferro *et al.*, 2010). Most organellar proteins are, therefore, encoded by the nuclear genome, synthesized by the cytoplasmic ribosomes and then targeted to the organelles.

The most common, although not unique, type of targeting signal consists of an N-terminal sequence (the targeting peptide, often referred as pre-sequence in mitochondria and transit peptide in plastids). The targeting peptide is cleaved when the native protein is translocated through the outer membrane of the organelles by means of the translocon protein complexes. In multicellular organisms, it has been estimated that ~10–25% of nuclear genes encode for proteins endowed with an N-terminal peptide targeting the protein to mitochondria and plastids (Emanuelsson *et al.*, 2000). Targeting peptides are highly heterogeneous in terms of length (ranging from 10 to 150 residues) and primary sequence (Bruce, 2001; Patron and Waller, 2007; Staiger *et al.*, 2009). Phylogenetic and structural studies recognized a modular architecture in targeting peptides, often consisting of two or three separate domains, potentially forming amphiphilic α-helices or β-strands when interacting with the organelle outer membrane (Bruce, 2001; Habib *et al.*, 2007). The modular organization is probably involved in the sub-organellar trafficking of proteins and gives origin to a great variety of primary sequences for targeting peptides (Texeira and Glazer, 2012). In a small but increasing number of proteins, the same targeting peptide can mediate the translocation to both the mitochondria and the chloroplasts (Carrie *et al.*, 2009).

*To whom correspondence should be addressed.

Presently, in UniProtKB, only 11% of ~9200 sequences that are endowed with a targeting peptide are supported by experimental annotations. Because of the biological importance of targeting mechanisms and the scarcity of experimentally validated knowledge, machine-learning tools have been introduced to predict the presence of targeting peptides and the position of the corresponding cleavage sites. Early methods are specific for chloroplasts, e.g. ChloroP (Emanuelsson *et al*., 1999) and PCLR (Schein *et al*., 2001), and mithochondria, e.g. MitoProt (Claros and Vincens, 1996). More recent tools integrate the prediction of targeting peptides with the prediction of secretory signal peptides, although they have different compositional and length features. Among these are TargetP (Emanuelsson *et al*., 2007), which incorporates ChloroP, iPSORT (Bannai *et al*., 2002), Predotar (Small *et al*., 2004) and PredSL (Petsalaki *et al*., 2006).

All methods predict the presence of the targeting peptide, but only MitoProt, TargetP and PredSL predict the position of the cleavage site. In general, all the predictors analyse an N-terminal portion of the native protein, ranging from 40 to 100 residues, depending on the method. ChloroP (Emanuelsson *et al*., 1999), TargetP (Emanuelsson *et al*., 2007), PCLR (Schein *et al*., 2001) and Predotar (Small *et al*., 2004) are based on neural networks (NNs) and their input includes residue composition, hydrophobicity and abundance of charged residues. iPSORT adopts a rule-based algorithm that considers 434 different propensity scales (Bannai *et al*., 2002). MitoProt defines a discriminant function on a pool of 47 physicochemical properties (Claros and Vincens, 1996). PredSL combines NNs, hidden Markov models (HMM) and scoring matrices (Petsalaki *et al*., 2006). Because of the paucity of data, all methods have been trained on datasets containing also non-experimentally validated targeting peptides, predicted with computational methods and/or inferred by similarity.

Prediction of targeting peptide can be considered as a labelling problem, where residues of the N-terminal region of the sequence are assigned either to 't' (targeting peptide) or 'n' (non-targeting peptide) labels. Grammatical-Restrained Hidden Conditional Random Fields (GRHCRF) is a recently introduced machine-learning tool well suited to solve labelling problems (Fariselli *et al*., 2009, Savojardo *et al*., 2011). GRHCRFs offer several advantages: (i) like HMMs, they can incorporate previous knowledge on the problem by introducing a grammar on the prediction labels; (ii) like the Hidden Conditional Random Fields (HCRF), they are discriminative models and do not require the strong independence assumptions made in HMMs (Fariselli *et al*., 2009; Lafferty *et al*., 2001); and (iii) similar to NNs, they can analyse complex and heterogeneous input encodings.

Here, we introduce TPpred, a new predictor for targeting peptides based on GRHCRFs. TPpred is trained on a non-redundant dataset containing only experimentally validated targeting peptides and efficiently predicts both the presence of targeting peptides and the localization of the cleavage sites.

2 METHODS

2.1 Dataset

We gathered the eukaryotic proteins longer than 45 residues from SwissProt (release November 2011) and annotated with existing evidence at the protein level, with the exclusion of fragments. Starting from this set, we collected both the positive and the negative datasets. The positive dataset (proteins endowed with an experimentally detected targeting peptide) was collected searching in the feature field for the keyword 'TRANSIT PEPTIDE', which identifies all the pre-sequences directing a protein to an organelle in UniProtKB (http://www.uniprot.org/keywords/KW-0809). We excluded annotations labelled as 'by similarity', 'probable' or 'potential', and we retained only proteins from mitochondria and plastids provided with a known cleavage site. Proteins lacking the keyword 'TRANSIT PEPTIDE' were collected in the negative set. By this, we obtained 757 positive and 47 363 negative examples. To obtain a non-redundant dataset, sequences were then compared with Basic Local Alignment Search Tool, and a graph was built linking the pairs of sequences (nodes) that share >30% identity on local alignments (HSP) with e-value $<10^{-3}$. The graph was clustered by extracting its connected components with a transitive closure algorithm. After this procedure, sequences in different clusters share <30% identity. We also checked that the 160-residue long N-terminal regions of sequences are <30% identical when extracted from different clusters. The non-redundant training dataset was built by randomly selecting one sequence per cluster. The final dataset consists of 297 sequences with targeting peptide (DB+) and 8010 without targeting peptide (DB−) (Table 1). To test whether the prediction is affected by the presence of transmembrane helices, we extracted a subset of proteins with an α-helix annotated in the 160 residue-long N-terminal segment by UniProtKB (values in parentheses of Table 1). The dataset is available at: http://biocomp.unibo.it/~valentina/TPpred/.

2.2 GRHCRF

Prediction of targeting peptides can be posed as a labelling problem with a strong grammatical constraint: the targeting peptide region ('t') precedes the non-targeting peptide region ('n'). Starting from HMMs that are the prototypical models addressing this type of problems, Conditional Random Fields (CRF) have been introduced: they are discriminative models that allow relaxing the strong independence assumptions of HMMs by means of a global normalization procedure (Lafferty *et al*., 2001). GRHCRFs have been developed to overcome the limitations of the coincidence between labels and states typical of CRFs (Fariselli *et al*., 2009; Savojardo *et al*., 2011). GRHCRFs decouple the set of labels from the set of states and allow defining a one-to-many mapping between them. Like HMMs, GRHCRFs can be represented through an automaton comprising a set of labelled states connected by transitions. The topology of the automaton casts the grammar to be modelled. The same label can be shared among different states. This ensures a great expressive power of the method and a large flexibility in the automaton design. A feature function is associated to each state and to each transition. The parameters of the feature functions are learned from the association between the input sequences included in the training set and their known labellings. Discriminative learning has been implemented for finding the parameters that maximize the probability of a label given the input (see Fariselli *et al*., 2009 for details). Given a trained model, the labelling of a

Table 1. The training dataset

Organism	Without TP (DB−)	Chloroplastic TP	Mitochondrial TP	With TP (DB+)
Plants	605 (86)	95 (12)	18 (0)	113 (12)
Non-plants	7405 (1081)	—	184 (12)	184 (12)
Total	8010 (1167)	95 (12)	202 (12)	297 (24)

TP, Targeting peptide. Values in parentheses refer to proteins where a transmembrane helix is annotated in the 160 residue-long N-terminal segment.

sequence is predicted with the posterior-Viterbi algorithm. This implements a decoding procedure that preserves the grammar, and it is based on the posterior probabilities for each label as computed by the model (Fariselli *et al.*, 2005).

2.3 Input features

For each sequence, 160 N-terminal residues were considered for building the input to GRHCRFs. Each position of the segment was encoded with a 25-valued vector describing the type and the physicochemical features of the corresponding residue. The 25-valued vector comprises four different modules: (i) a 20-valued binary vector, describing the residue type, whose elements are all null but the one corresponding to the residue to be encoded (seq); (ii) one value encoding the average Kyte–Doolittle hydrophobicity (Kyte and Doolittle, 1982) of a seven-residue long window centred on the residue to be encoded (kd); (iii) two values encoding the number of positively and negatively charged residues in a seven-residue window (ch); and (iv) two values describing the hydrophobic moments (hm) computed considering 100° and 160° angles for simulating ideal α-helices and β-sheets, respectively. The program *hmoment* included in EMBOSS (Rice *et al.*, 2000) was adopted to carry out the computation of the hydrophobic moments. For each position in the sequence, the feature function of each GRHCRF state considers an 11-residue window; therefore, it takes in input $11 \times 25 = 275$ different variables. When encoding the five N- and C-terminal residues of each sequence, we padded the empty positions of the window with '0' values.

2.4 Training procedure

We adopted a 5-fold cross-validation procedure for training and testing, by randomly splitting the non-redundant training set into five subsets. Three subsets were used for training the method (training set), one for validation (validation set) and the remaining for evaluating the performance (test set). The best model topology, the best parameters and the best input were selected on the basis of the results obtained on the validation set. Five training runs were performed, and performance was computed collecting all the results obtained for the five test sets.

2.5 Scoring the performance

Different scoring indexes were used to evaluate the prediction performances at the protein level. For the two protein classes, namely, 'with targeting peptide' (+) and 'without targeting peptide' (−), we indicated with TP and TN the number of true-positive and true-negative predictions, respectively, and with FP and FN the number of false-positive and false-negative predictions, respectively.

General prediction scores are the overall accuracy (Acc) and the Matthews correlation coefficient (MCC), defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FP)}} \quad (2)$$

Unlike Acc, MCC is only slightly affected by bias deriving from the unbalance between positive and negative examples in the dataset, and it is then the most useful global score. Scoring indexes evaluating the performance on the single class are sensitivity (Sn), specificity (Sp) and false-positive rate (FPR) computed as:

$$Sn(c) = \frac{TP}{TP + FN} \quad (3)$$

$$Sp(c) = \frac{TP}{TP + FP} \quad (4)$$

$$FPR(c) = \frac{FP}{TN + FP} \quad (5)$$

where c is the class at hand. A thorough explanation of the purposes of these indexes can be found in Baldi *et al.* (2000).

2.6 Prediction with available methods

For sake of comparison, we predicted the sequences included in our dataset with the following methods: (i) the TargetP server was accessed at http://www.cbs.dtu.dk/services/TargetP/; (ii) the executable version of iPSORT was downloaded from http://ipsort.hgc.jp/caml-iPSORT/; (iii) the executable version of PredSL was downloaded from http://hannibal.biol.uoa.gr/PredSL/source.html; (iv) the Predotar server was accessed at http://urgi.versailles.inra.fr/predotar/predotar.html; (v) the software of MitoProt was downloaded from ftp://ftp.biologie.ens.fr/pub/molbio; and (vi) PCLR was re-implemented in house using the parameters listed in http://www.andrewschein.com/cgi-bin/pclr/weights.html. When required, the proper prediction parameters were selected, dividing plant and non-plant proteins.

2.7 Whole-proteome analysis

The complete sets of proteins from *Homo sapiens* (GRHh37.p5), *A.thaliana* (TAIR10) and *Saccharomyces cerevisiae* (EF4) were downloaded from the EnsEMBL website (www.ensembl.org). These sets comprise 93 588, 35 386 and 6692 protein sequences (including splicing variants), respectively. They are encoded by 21 160, 27 416 and 6692 genes, respectively. We predicted all the proteins with our TPpred and checked the agreement between the prediction of targeting peptide and the Gene Ontology (GO) annotation of the subcellular localization as reported in EnsEMBL. We retained only experimental annotations labelled with the following evidence codes: EXP (experimental), IDA (inferred from direct assay), IPI (inferred from physical interaction), IMP (inferred from mutant phenotype), IGI (inferred from genetic interaction) and IEP (inferred from expression pattern) (http://www.geneontology.org/GO.evidence.shtml).

3 RESULTS AND DISCUSSION

3.2 Main features of targeting peptides

3.1.1 Length The length of the mitochondrial and plastidic targeting peptides ranges from 10 up to 150 residues (Fig. 1). Mitochondrial-targeting peptides are, on average, shorter than plastidic ones, being the average lengths 35 and 59 residues, respectively. The dispersions around the average lengths are, however, very high and of ∼16 and 22 residues, respectively. We chose not to separate the two datasets, because of (i) the scarcity of non-redundant proteins experimentally annotated for targeting peptides and (ii) the possibility that the same targeting peptide mediates the translocation to both mitochondria and plastids (Carrie *et al.*, 2009).

3.1.2 Residue composition The residue composition of targeting peptides is plotted in Figure 2 and compared with the whole-sequence composition of the proteins included in our dataset. Relative standard deviations of the samples are not shown for sake of clarity and have been evaluated to be ∼20% of the plotted data. Mitochondrial and plastidic proteins (represented in Fig. 2 with blue and cyan bars, respectively) do not show major compositional differences with proteins included in the negative set (red bars). On the contrary, targeting peptides are characterized by a peculiar composition. The differences in composition between the targeting peptides and the whole sequences have been assessed in terms of log-odds and *P*-values

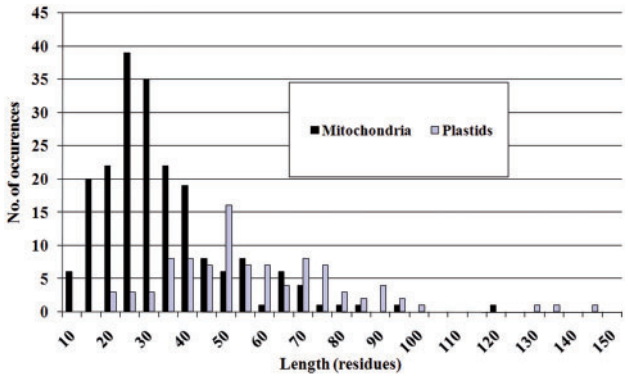

Fig. 1. Length distribution of the targeting peptides of proteins included in the DB+ dataset

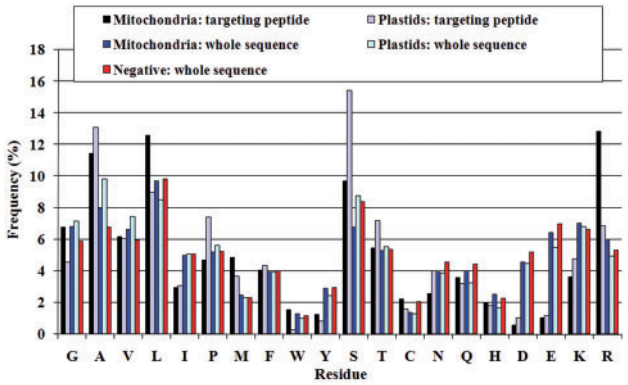

Fig. 2. Residue composition of targeting peptides and of sequences included in the training set. Average values are computed on 202 mitochondrial proteins, 95 plastidic proteins and 8010 proteins non-containing the targeting peptide (negative). Relative standard deviations of the samples, evaluating the dispersion around the reported average value, are ~20%, and they are not represented in the figure

(Supplementary Table S1). At a significance level equal to 10^{-10}, both mitochondrial- and plastidic-targeting peptides (black and grey bars, respectively) are enriched in alanine (A) and serine (S) and depleted in negatively charged residues (D, E), tyrosine (Y) and isoleucine (I). Other differences can be detected when mitochondrial- and plastidic-targeting peptides are separately analysed: the former are enriched in methionine (M), leucine (L) and arginine (R), whereas lysine (K) is underrepresented. In plastidic-targeting peptides glycine (G) is less frequent. The composition of targeting peptides of our dataset is similar to that of previous analyses (Texeira and Glaser, 2012), and it accounts for the interactions with proteins involved in protein import and peptide cleavage, whose structural details are still unknown (Jarvis and Robinson, 2004; Pfanner and Geissler, 2001).

3.1.3 Cleavage site The strongest compositional information is thought to reside in the region neighbouring the cleavage site, as it is recognized by the active site of peptidase complexes. Starting from the 297 proteins of the positive dataset, we aligned the eight residues downstream and upstream the cleavage site. The resulting profile is visualized in the sequence logo of Figure 3. The cleavage site is between positions 0 and 1. The logo

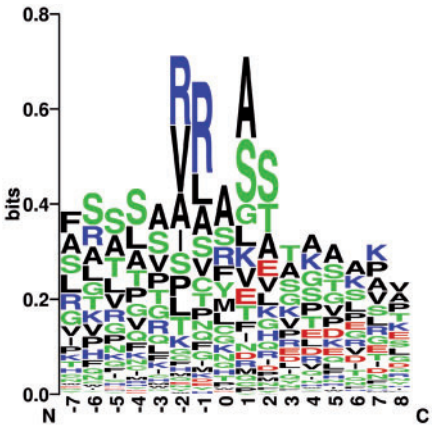

Fig. 3. Sequence logo of the positions neighbouring the cleavage site. Sequence logo (Schneider and Stephens, 1990) is computed by the WebLogo server (weblogo.berkeley.edu). Position '1' is the first residue of the mature protein. Height of letters is proportional to their information content in profile. Information is measured in bits and ranges between 0 and $\log_2(20) \approx 4.3$. Colour codes cluster residues in apolar (black), polar (green), positively charged (blue) and negatively charged (red)

represents the conservation of each residue present in a position with a letter whose height is proportional to its information content, ranging from 0 to $\log_2 20 = 4.3$. It is evident that the positions neighbouring the cleavage site are slightly more conserved than the others, although the overall information content is moderate. The most conserved residues are in position −1 and −2, downstream, and +1 and +2, upstream the cleavage site.

3.2 The automaton for predicting the targeting peptide

We cast the information on the features of targeting peptides in the automaton described in Figure 4. It comprises 45 states associated to the label 't' (targeting states represented with squares) and 31 states associated to the label 'n' (non-targeting states, represented with circles). The overall model mainly consists of two sub-models: the grey coloured states describe the 160 N-terminal residues of proteins with targeting peptides and white coloured states describe the same portion in proteins without targeting peptide.

The sub-model for proteins without targeting peptide consists of states connected with a dense topology that allows a general description of a broad range of different sequences.

The sub-model for proteins endowed with targeting peptides is more specific and consists of different groups of states aiming at capturing the modular organization of targeting peptides. As reported in Section 1, phylogenetic and structural studies recognized a modular architecture in targeting peptides that often contain two or three separate domains, potentially forming amphiphilic α-helices or β-strands when interacting with the organelle outer membrane (Bruce, 2001; Habib *et al.*, 2007). The targeting model consists of four modules: (i) the N-terminal region, probably unstructured and variable in length, is modelled with a set of densely connected states; (ii) two central domains, consisting of forward-connected states, aim at capturing the features of the modules described in literature; and (iii) the cleavage

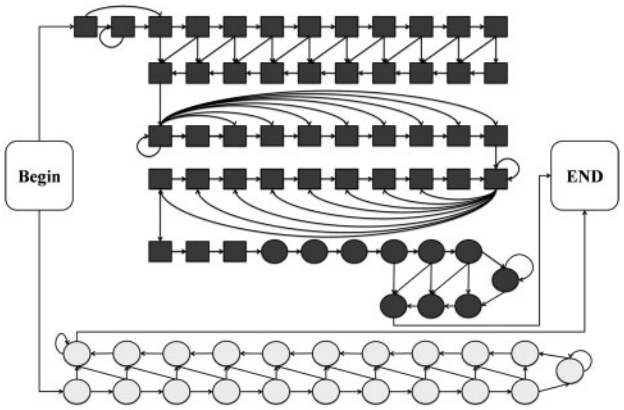

Fig. 4. Automaton for targeting peptide prediction. Squares represent states labelled as targeting peptide ('t'), whereas circles represent non-targeting peptide states ('n'). Grey states model the 160-residue long N-terminal region of proteins endowed with targeting peptide (target model). White states model proteins devoid of signal peptide (non-target model). See text for further details

site region, where three residues upstream and three residues downstream are explicitly modelled with six different states. To model the large variability of targeting peptide lengths, several states are self-connected. Owing to the overall topology, the length of predicted targeting peptides spans from 10 to 155 residues, in agreement with the lengths deduced from the analysis of our dataset (Fig. 1). As the modular organization cannot be easily recognized in all the known targeting peptides, the model topology has to maintain a high flexibility.

Different model topologies have been adopted, and we retained the model best performing on the validation sets. The performance scores are computed in cross-validation on test sets independent of both the training and the validation sets.

3.3 Targeting peptide prediction with different inputs

The GRHCRF was trained adopting the strategy described in Section 2.4 and adopting a 5-fold cross-validation procedure. Three sets are used for training, one (validation set) for choosing the best automaton parameters (including the window size and the input encoding) and the remaining (testing set) for computing the indexes scoring the performance. As the complete dataset was reduced for similarity (Section 2.1), this procedure ensures a reliable evaluation of the generalization capability of the method.

Table 2 lists the performance scores evaluated on the complete set (altogether comprising 297 positive and 8010 negative examples) when different inputs are fed to the GRHCRF. The single sequence leads to an overall accuracy value (Acc) as high as 95% and an MCC value equal to 0.50. By incrementally adding the different features described in Section 2.3, performance increases on testing sets and reaches the maximum value when information on sequence, hydrophobicity, charge and hydrophobic moment of the N-terminal regions are included. The final performance of TPpred is as high as 96% accuracy and 0.58 MCC. When proteins belonging to organisms from different kingdoms are evaluated separately, the MCCs are as high as 0.74 and 0.52 for plants and non-plants, respectively (for a more detailed evaluation see Supplementary Table S2).

Table 2. Performance of GRHCRF with different input encoding

Input	Acc (%)	MCC	Sn(+) (%)	Sp(+) (%)	Sn(−) (%)	Sp(−) (%)	FPR (%)
Seq	95	0.5	69	40	96	99	3.8
Seq + kd	95	0.54	73	42	96	99	3.7
Seq + kd + ch	96	0.56	75	46	97	99	3.3
Seq + kd + ch + hm (TPpred)	96	0.58	75	48	97	99	3.0

seq, sequence; kd, Kyte–Doolittle hydropathy; ch, charge; hm, hydrophobic moment. For a thorough description of the input, see Section 2.3. The evaluation dataset comprises 297 positive and 8010 negative examples. Scoring indexes are computed with a 5-fold cross-validation procedure, by collecting the results on the test sets. For index definition, see Section 2.4.

To assess the advantage in adopting GRHCRF for predicting the presence of targeting peptides, we also implemented a predictor based on an HMM with the same topology described in Figure 4 and trained on the same dataset. When sequence information is adopted as input, the HMM scores with an MCC equal to 0.39, significantly lower than that reached with GRHCRF on the same input ('seq' line in Table 2; see Supplementary Table S3 for a detailed comparison).

3.4 Benchmark with available methods

Table 3 lists the prediction performance of TPpred on the non-redundant dataset as compared with that of other available methods predicting both mitochondrial- and plastidic-targeting peptides. TPpred outperforms all of these methods. When evaluated with the Fisher r-to-z transformation (Fisher, 1921), the difference in MCC between TPpred and the best performing predictor (Predotar) is significant, with $P < 10^{-4}$. It is worth noticing that in Table 3, only TPpred is evaluated by adopting a cross-validation procedure. Indeed, the overlap between the training datasets of other methods and that adopted for training/validating TPpred can lead to overestimate the performances of the other tools.

The low specificity on the positive class [Sp(+)] is the major pitfall of all available predictors, probably because of the unbalance in the datasets adopted for training them. TPpred is by far the most specific predictor (48%). This is reflected in a lower sensitivity, that, however, reaches a high value [Sn(+) = 75%]. Moreover, TPpred scores with an FPR (3%) that is less than a half with respect to the best available tools. The same trend is confirmed, when comparing with predictors specific for mitochondria (MitoProt) and plastids (PCLR) (Table 4): also in this case, TPpred scores with the highest accuracy and MCC.

TPpred scores with a lower sensitivity than other methods and this corresponds to an increase of the false-negative rate [FNR = 1 − Sn(+)]. FNR is equal to ~25% for TPpred, higher than that reported by other methods (7–21%). However, when implementing a prediction method suitable for large-scale annotation of proteins, it is important to keep the error rate as low as possible in the most abundant class, to keep the number of wrong predictions low. At the whole-proteome level, proteins without targeting peptide (the negative set) are by far more

Table 3. Benchmark results on non–organelle-specific predictors

Method	Acc (%)	MCC	Sn(+) (%)	Sp(+) (%)	Sn(−) (%)	Sp(−) (%)	FPR (%)	FPR TM[a] (%)
TPpred[b]	96	0.58	75	48	97	99	3	2.5
TargetP[c]	89	0.44	93	24	89	100	10.9	11.1
Predotar[c]	92	0.49	90	30	92	100	7.7	8.8
iPSORT[c]	89	0.38	79	22	90	99	10.2	12.4
PredSL[c]	91	0.45	88	26	91	100	9.4	12.5

Scoring indexes are computes as described in Section 2.4. Tools other than TPpred were run as described in Section 2.5. The evaluation dataset comprises 297 positive and 8010 negative examples. [a]FPR computed on the negative examples endowed with a transmembrane helix within the 160 N-terminal residues (see Table 1, first column, values in parentheses). [b]Sequences are predicted in cross-validation on the test sets. [c]The benchmark dataset overlaps with the training set.

Table 4. Benchmark results on organelle-specific predictors

Method	Acc (%)	MCC	Sn(+) (%)	Sp(+) (%)	Sn(−) (%)	Sp(−) (%)	FPR (%)	FPR TM[a] (%)
Mitochondrial proteins								
202 positive (mitochondrial) and 8010 negative(1167 TM) examples from all eukaryotes								
TPpred[b]	96	0.52	74	39	97	100	3	2.5
MitoProt[c]	77	0.23	89	9	76	100	22	29.9
Plastidic proteins								
95 positive (plastidic) and 605 negative (86 TM) examples from plants								
TPpred[b]	94	0.73	73	81	97	96	2.6	0
PCLR[c]	86	0.6	93	48	84	99	15.5	12.8

Scoring indexes are computes as described in Section 2.4. Tools other than TPpred were run as described in Section 2.5. Only the values relative to TPpred are computed in cross-validation. [a]FPR computed on the negative examples endowed with a transmembrane helix within the 160 N-terminal residues (see Table 1, first column, values in parentheses). [b]Sequences are predicted in cross-validation on the test sets. [c]The benchmark dataset overlaps with the training set.

abundant than proteins with targeting peptide (the positive set), and this is why we consider that the low FPR (that is the rate of error on the negative set) reported by TPpred is an interesting feature for its adoption in large-scale analyses.

We also tested the FPRs of predictors on the subset of 1167 negative examples endowed with a transmembrane helix within the 160 N-terminal residues (last column of Tables 3 and 4). Less than 2.5% of this set is predicted by TPpred as proteins endowed with a targeting peptide, and this is the best result obtained with the currently available methods. This enables TPpred to be safely adopted for analysing membrane proteins. In particular, when used as a pre-filter to identify cleaved peptides, it lowers the risk of removing N-terminal transmembrane helices.

3.5 Prediction of the cleavage site

TPpred also predicts the position of the cleavage site along the sequence. This information is important, as it allows knowing the sequence of the mature and functional protein. Some predictors, however, do not report it (e.g. Predotar). In Table 5, we comparatively assessed the prediction of cleavage site performed with TPpred, TargetP, PredSL and MitoProt (with the last one, only on mitochondrial proteins). Mitochondrial and plastidic sequences are evaluated separately because of the different average length of the corresponding targeting peptides (35 and 59 residues, respectively, see also Section 3.1.1). For each prediction, we evaluated the error (E) as the difference between the positions of the real and the predicted cleavage sites. We then computed the mean error (ME) and the number of prediction for which the error is lower than the standard deviation (σ) of the length distribution of targeting peptides ($E < \sigma$ score). Standard deviations are equal to 16 and 22 residues for mitochondria and plastids, respectively, as discussed in Section 3.1.1. Our TPpred outperforms the other methods both in terms of ME and $E < \sigma$ score, particularly for mitochondria. This indicates that TPpred correctly predicts the correct length of the targeting peptide, even if the length distribution is spread.

3.6 Prediction of targeting peptides in whole proteomes

The whole proteomes of three species were downloaded from EnsEMBL and predicted with TPpred. Results of the prediction are reported in Table 6. We estimate that 4.0, 9.0 and 6.1% of proteins are endowed with targeting peptide in human, *Arabidopsis* and yeast, respectively. The estimates are somewhat lower than those previously reported with other methods (10–25%, Emanuelsson *et al.*, 2000). This result is possibly because of the low FPR of TPpred (3%) that limits the number of mispredictions in the negative set.

For proteins predicted with targeting peptide, we tested the compatibility with the GO annotations for cellular component reported in EnsEMBL and labelled with an experimental evidence code (see Section 2.6). GO terms were divided into three subsets: (i) terms directly related to mitochondrial or plastidic

Table 5. Benchmark on the cleavage site prediction

Method	Mitochondria		Plastids	
	ME (res)	E<σ score (%)	ME (res)	E<σ score (%)
TPpred[a]	7	89	15	74
TargetP[b]	12	71	16	71
PredSL[b]	12	75	17	73
MitoProt[b]	13	75	—	—

ME: mean prediction error on the position of the cleavage site. E<σ score: proportion of predictions with error lower than the standard deviation of the length distribution of targeting peptides. [a]Sequences are predicted in cross-validation on the test sets. [b]The benchmark dataset overlaps with the training set.

Table 6. Targeting peptides predicted at the whole-organism scale

	H.sapiens	*A.thaliana*	*S.cerevisiae*
Whole organism			
No. of proteins (no. of genes)	93 588 (21 160)	35 386 (27 416)	6692 (6692)
With predicted targeting peptide			
No. of proteins (no. of genes)	3744 (1685)	3194 (2521)	407 (407)

Predicted proteomes are available at http://biocomp.unibo.it/~valentina/TPpred/. Values reported in parentheses refer to the number of genes.

Table 7. Comparison between targeting peptide predictions and experimental GO annotations

GO annotation (EnsEMBL)	*H.sapiens*	*A.thaliana*	*S.cerevisiae*
Mitochondrion	288 (8%)	286 (9%)	228 (56%)
Plastid	—	1297 (41%)	—
Compatible	5 (0%)	10 (0%)	3 (1%)
Incompatible	158 (4%)	221 (7%)	40 (10%)
Not annotated	3293 (88%)	1370 (43%)	136 (33%)

Percentage values are computed with respect to the number of protein sequences predicted as endowed with targeting peptide (3744 in *Homo*, 3194 in *Arabidopsis* and 407 in *Saccharomyces*).

localizations; (ii) terms compatible with mitochondrial or plastidic localizations, as they include them as subsets (e.g. cell part, intracellular, cytoplasm and membrane); and (iii) terms incompatible with mitochondrial or plastidic localization.

The results for the three proteomes are reported in Table 7. In the case of the human proteome, only 12% of the proteins predicted with targeting peptide are endowed with experimental annotation of their localization: 8% of proteins are mitochondrial and are, therefore, correctly predicted; 4% are localized in other cellular components (mostly in the nucleus) and can, therefore, be considered as false predictions. The rate of experimental annotation in *A.thaliana* and yeast is much higher (57 and 67%, respectively) and mostly confirms the predictions of TPpred: 50 and 56% of proteins predicted with targeting peptide in *Arabidopsis* and yeast, respectively, are localized in mitochondria or plastids. When considering only the set of annotated proteins, the rates of success in the two well-annotated organisms are, therefore, 50/57 = 87% and 56/67 = 84%, respectively. Proteins with incompatible localization are, in both organisms, mostly annotated as nuclear or, in the case of *Arabidopsis*, as plasma membrane.

The good agreement between the prediction and the experimental annotations confirms the suitability of TPpred for performing prediction of whole proteomes. In the three analysed organisms, we also suggest a new annotation for a large amount of proteins: 3293 in human, 1370 in *Arabidopsis* and 136 in yeast.

4 CONCLUSIONS

We implemented TPpred, a new predictor for targeting peptides in mitochondrial and plastidic proteins. TPpred is based on GRHCRFs, a recently introduced machine-learning approach. Differently from available methods, it is trained only on experimentally validated targeting peptides. TPpred outperforms other available methods, both in predicting the presence and the length of targeting peptides. TPpred is significantly more specific than the available predictors and scores with a very low-FPR. This feature makes TPpred useful for predicting the targeting peptides in proteomes of whole organisms. In particular, when tested on the proteomes of *H.sapiens*, *A.thaliana* and *S.cerevisiae*, the estimate of the amount of proteins endowed with targeting peptide is ~4–9%. The good agreement between the predictions of TPpred and the experimental annotations suggests that this method can be combined with subcellular localization predictors for improving their performance in genome-wide annotation procedures.

Funding: This work has been supported by the following grants: PRIN 2009 project 009WXT45Y (Italian Ministry for University and Research: MIUR), COST BMBS Action TD1101 (European Union RTD Framework Program) and PON project PON01_02249 (Italian Ministry for University and Research: MIUR). PhD fellowship of the Italian Ministry for University and Research: MIUR (to C.S.).

Conflict of Interest: none declared.

REFERENCES

Baldi,P. *et al.* (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Bannai,H. *et al.* (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305.

Bruce,B.D. (2001) The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. *Biochim. Biophys. Acta*, **1541**, 2–21.

Carrie,C. *et al.* (2009) Protein transport in organelles: dual targeting of proteins to mitochondria and chloroplasts. *FEBS J.*, **276**, 1187–1195.

Claros,M.G. and Vincens,P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.*, **241**, 779–786.

Emanuelsson,O. *et al.* (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.*, **8**, 978–984.
Emanuelsson,O. *et al.* (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
Emanuelsson,O. *et al.* (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. *Nat. Protoc.*, **2**, 953–971.
Fariselli,P. *et al.* (2005) A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics*, **6**, S12.
Fariselli,P. *et al.* (2009) Grammatical-restrained hidden conditional random fields for bioinformatics applications. *Algorithms Mol. Biol.*, **4**, 13.
Ferro,M. *et al.* (2010) AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. *Mol. Cell. Proteomics*, **9**, 1063–1084.
Fisher,R.A. (1921) On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, **1**, 3–32.
Jarvis,P. and Robinson,C. (2004) Mechanisms of protein import and routing in chloroplasts. *Curr. Biol.*, **14**, R1064–R1077.
Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
Lafferty,J. *et al.* (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proc. ICML01*, 282–289.
Habib,S.J. *et al.* (2007) Analysis and prediction of mitochondrial targeting signals. *Methods Cell Biol.*, **80**, 761–781.
Patron,N.J. and Waller,R.F. (2007) Transit peptide diversity and divergence: a global analysis of plastid targeting signals. *BioEssays*, **29**, 1048–1058.
Petsalaki,E.I. *et al.* (2006) PredSL: a tool for the N-terminal sequence-based prediction of subcellular localization. *Genomics Proteomics Bioinformatics*, **4**, 48–55.
Pfanner,N. and Geissler,A. (2001) Versatility of the mitochondrial protein import machinery. *Nat. Rev. Mol. Cell Biol.*, **2**, 339–349.
Rice,P. *et al.* (2000) EMBOSS: European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.
Savojardo,C. *et al.* (2011) Improving the prediction of disulfide bonds in eukaryotes with machine learning methods and protein subcellular localization. *Bioinformatics*, **27**, 2224–2230.
Schein,A.I. *et al.* (2001) Chloroplast transit peptide prediction: a peek inside the black box. *Nucleic Acids Res.*, **29**, e82.
Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
Sickmann,A. *et al.* (2003) The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc. Natl. Acad. Sci. USA*, **103**, 13207–13212.
Small,I. *et al.* (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.
Smith,A.C. *et al.* (2012) MitoMiner: a data warehouse for mitochondrial proteomics data. *Nucleic Acids Res.*, **40**, 1060–1067.
Staiger,C. *et al.* (2009) Diversity in degrees of freedom of mitochondrial transit peptides. *Mol. Biol. Evol.*, **26**, 1773–1780.
Texeira,P.F. and Glaser,E. (2012) Processing peptidases in mitochondria and chloroplasts. *Biochim. Biophys. Acta.*, **1833**, 360–370.
van Wijk,K.J. (2004) Plastid proteomics. *Plant Physiol. Biochem.*, **42**, 963–977.