BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 20 2014, pages 2973–2974
doi:10.1093/bioinformatics/btu411

Sequence analysis Advance Access publication June 27, 2014

TPpred2: improving the prediction of mitochondrial targeting peptide cleavage sites by exploiting sequence motifs

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

[1]Biocomputing Group, University of Bologna, CIRI-Health Science and Technology/Department of Biology, 40126 Bologna and [2]Department of Computer Science and Engineering, University of Bologna, 40127 Bologna, Italy

Associate Editor: John Hancock

ABSTRACT

Summary: Targeting peptides are N-terminal sorting signals in proteins that promote their translocation to mitochondria through the interaction with different protein machineries. We recently developed TPpred, a machine learning-based method scoring among the best ones available to predict the presence of a targeting peptide into a protein sequence and its cleavage site. Here we introduce TPpred2 that improves TPpred performances in the task of identifying the cleavage site of the targeting peptides. TPpred2 is now available as a web interface and as a stand-alone version for users who can freely download and adopt it for processing large volumes of sequences.

Availability and implementaion: TPpred2 is available both as web server and stand-alone version at http://tppred2.biocomp.unibo.it.

Contact: gigi@biocomp.unibo.it

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

Received on April 22, 2014; revised on June 19, 2014; accepted on June 23, 2014

1 INTRODUCTION

The vast majority of mitochondrial proteins are encoded by the nuclear genome, synthesized on the cytosolic ribosomes and therefrom translocated to the different sub-mitochondrial localizations. The best characterized mechanism that controls the import into the mitochondrion involves the detection of an N-terminal pre-sequence, the so-called targeting peptide, which directs the protein across different translocating protein complexes embedded in the outer and inner mitochondrial membranes. The sorting peptide is proteolytically cleaved when the protein reaches the final destination (Schmidt *et al*., 2010). The detection of targeting peptides starting from protein sequences is then an important step to fully characterize the sequence of the mature protein and to annotate its function and localization. Different computational methods can help the annotation of targeting peptides in proteins. The best performing methods implement machine-learning approaches, such as artificial neural network, support vector machines (SVMs), hidden Markov models and Grammatical-Restrained Hidden Conditional Random Fields (GRHCRFs) (Emanuelsson *et al*., 2007; Indio *et al*., 2013; Petsalaki *et al*., 2006; Small *et al*., 2004).

Here we describe TPpred2, our new implementation for the prediction of mitochondrial targeting peptides and their cleavage sites. The new predictor combines TPpred, our previous GRHCRF-based method for the prediction of mitochondrial and plastidic targeting peptides (Fariselli *et al*., 2009; Indio *et al*., 2013), with an additional refinement step—SVM-based—for cleavage site identification. In particular, the SVM classifier is used to identify the correct cleavage site by exploiting TPpred prediction signals (extracted in terms of GRHCRF-based positional posterior probabilities) and a newly defined feature that captures the occurrence of sequence motifs previously recognized at the cleavage sites (Mossmann *et al*., 2012).

When tested on the non-redundant dataset of proteins of Indio *et al*. (2013; 202 mitochondrial and 8010 non-mitochondrial proteins; see Supplementary Section 1.1), TPpred2 significantly outperforms our previous TPpred predictor on mitochondrial cleavage site detection. TPPred2, being optimized for the prediction of cleavage sites of mitochondrial proteins, is not adequate for the identification of plant targeting peptides.

2 METHOD DESCRIPTION AND USAGE

2.1 Mitochondrial cleavage site sequence motifs

Cleavage regions in mitochondrial proteins exhibit characteristic sequence features for recognition by the cleavage machineries (Mossmann *et al*., 2012). Five different sequence motifs have been identified as typical of mitochondrial cleavage sites: (i) R2 = RX|X, (ii) R3a = RX [FLY]||[SA], (iii) R3b = RX[FLY]||X, (iv) R10 = RX[FLI]XX[TGS] XXXX|X and (v) Rnone = X|XS. The character '|' indicates the position of the cleavage; the 'X' character is a wildcard used to represent any residue type (the residues enclosed in square brackets indicate alternative choices). The different motifs are not mutually exclusive, and a single cleavage site can match with more than one of them. Here we rank the motifs according to their positive predictive value (PPV), defined as the ratio between the matches corresponding to an experimental cleavage site (true-positive matches) and the total number of occurrences in the non-redundant dataset including 202 proteins from eukaryotes with experimentally verified targeting peptides (Indio *et al*., 2013; see Supplementary Section 1.1 for more details). Adopting the PPV criterion, the top-ranking motif is R3a, followed by R10, R3b, R2 and Rnone. Not all cleavage sites exhibit these features: 139 of 202 mitochondrial proteins match at least one of the motifs (see Supplementary Section 1.1 for a complete statistics). Here we use these observations to derive a motif-based feature classifier and improve cleavage site identification.

2.2 TPpred2 overview

TPpred2 predicts mitochondrial targeting peptides and cleavage sites, adopting a two-step procedure. In the first step, the protein is filtered with TPpred (Indio *et al*., 2013). If no targeting peptide is detected, the

*To whom correspondence should be addressed.

input sequence is predicted as a non-mitochondrial protein (TPpred shows the lowest false-positive rate among the state-of-the-art methods; Indio *et al.*, 2013). If TPpred identifies a targeting peptide, the cleavage site prediction is refined using an additional SVM classifier (second step). In the refinement step, the classifier operates analyzing an optimal 29-residue-long segment centered on the TPpred-predicted cleavage site. The refinement SVM is fed using three input features: (i) two GRHCRF-derived positional posterior probabilities that score the likelihood for the cleavage site to be located at a given sequence position and (ii) a feature defined on the basis of the occurrence of the sequence motifs described in the previous section. The refined cleavage site is the position scoring with the highest SVM output probabilities. When the SVM probability is <0.5 for any position in the segment, the cleavage site predicted by TPpred is retained (for details, see Supplementary Section 1.2).

2.3 Usage and program requirements

TPpred2 is available both as web server and command-line tool. The TPpred2 web server (http://tppred2.biocomp.unibo.it) requires a single protein sequence in FASTA format. Once the input sequence is processed, the output page contains all the information about the identified targeting peptide (when detected), including the cleavage site location and a reliability score attached to the prediction. The command-line version of the program is well suited for large-scale genomic analyses (whereas the web server can handle a single protein sequence at a time). The program source code and example data are available at http://tppred2.biocomp.unibo.it. After unpacking, the TPpred2 command-line program is ready to use (see Supplementary Section 2.2 for further instructions about the initial configuration), and the user can operate by simply providing a multi-FASTA file as input to the program. For each protein sequence contained in the input file, TPpred2 lists the putative location of the cleavage site (when detected, the symbol '-' otherwise) and the corresponding reliability score (for both positive and negative predictions).

2.4 TPPRED2 performance

The overall performance of TPpred2 in discriminating mitochondrial from non-mitochondrial proteins is similar to that of TPpred (Indio *et al.*, 2013). However, TPpred2 is superior to TPpred in detecting the mitochondrial cleavage site. In Table 1, we compare the performance of both TPpred2 and the legacy TPpred obtained using a 5-fold cross-validation procedure. Two different scoring measures are computed: (i) the fraction of proteins whose cleavage sites have been perfectly identified (Cleave* columns) and (ii) the mean absolute error on cleavage site identification (ME* columns) that measures the average distance between predicted and observed cleavage sites. Scoring indexes are computed over the entire dataset of 202 proteins ($\text{Cleave}_{\text{all}}$, ME_{all}) and over the subset of 139 proteins whose cleavage site matches at least one sequence motif ($\text{Cleave}_{\text{motif}}$, ME_{motif}). Evidently, the new TPpred2 predictor significantly outperforms the legacy TPpred in the task of identifying the cleavage site. The improvement is particularly evident when the predictors are compared on the subset of proteins whose cleavage sites match with the sequence motifs. Furthermore, even when the site is not correctly located, TPpred2 provides better predictions, as highlighted by the reduced absolute error in cleavage identification.

Table 1. TPpred and TPpred2 performances on the prediction of cleavage sites of mitochondrial targeting peptides

Method	ME_{all}	$\text{Cleave}_{\text{all}}$ (%)	ME_{motif}	$\text{Cleave}_{\text{motif}}$ (%)
TPpred	7	17	5	24
TPpred2	6	32	4	46

Note: See Supplementary Section 1.2.5 for details about scoring measures.

3 CONCLUSION

In this article, we present TPpred2, a software tool for the prediction of mitochondrial targeting peptides and their cleavage sites. The improved performance of TPpred2 is achieved by jointly exploiting, for the first time to our knowledge, peculiar sequence motifs characterizing the cleavage site and GRHCRF prediction signals. The new TPpred2 predictor is able to better identify cleavage sites when compared with our previous implementation. The software is available as a web server and, most importantly, as a stand-alone program, making it well suited for large-scale genomic analysis.

Funding: PRIN 2010-2011 project 20108XYHJS (to P.L.M.) (Italian MIUR); COST BMBS Action TD1101 (European Union RTD Framework Program to R.C.); PON projects PON01_02249 and PAN Lab PONa3_00166 (Italian Miur to R.C. and P.L.M.); FARB-UNIBO 2012 (to R.C.).

Conflicts of interest: none declared.

REFERENCES

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.* (2009) Grammatical-restrained hidden conditional random fields for bioinformatics applications. *Algorithms Mol. Biol.*, **22**, 4–13.

Indio,V. *et al.* (2013) The prediction of organelle-targeting peptides in eukaryotic proteins with Grammatical-Restrained Hidden Conditional Random Fields. *Bioinformatics*, **29**, 981–988.

Mossmann,D. *et al.* (2012) Processing of mitochondrial presequences. *Biochim. Biophys. Acta*, **1819**, 1098–1106.

Petsalaki,E.I. *et al.* (2006) PredSL: a tool for theN-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinformatics*, **4**, 48–55.

Schmidt,O. *et al.* (2010) Mitochondrial protein import: from proteomics to functional mechanisms. *Nat. Rev. Mol. Cell Biol.*, **11**, 655–667.

Small,I. *et al.* (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.