BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 9 2011, pages 1224–1230
doi:10.1093/bioinformatics/btr108

Sequence analysis
Advance Access publication March 2, 2011

MemLoci: predicting subcellular localization of membrane proteins in eukaryotes

Andrea Pierleoni[1,2], Pier Luigi Martelli[1] and Rita Casadio[1,*]
[1]Biocomputing Group, Computational Biology Network, via San Giacomo 9/2, 40126 Bologna and [2]Externautics s.p.a., Via Fiorentina 1, 53100 Siena, Italy

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Subcellular localization is a key feature in the process of functional annotation of both globular and membrane proteins. In the absence of experimental data, protein localization is inferred on the basis of annotation transfer upon sequence similarity search. However, predictive tools are necessary when the localization of homologs is not known. This is so particularly for membrane proteins. Furthermore, most of the available predictors of subcellular localization are specifically trained on globular proteins and poorly perform on membrane proteins.

Results: Here we develop MemLoci, a new support vector machine-based tool that discriminates three membrane protein localizations: plasma, internal and organelle membrane. When tested on an independent set, MemLoci outperforms existing methods, reaching an overall accuracy of 70% on predicting the location in the three membrane types, with a generalized correlation coefficient as high as 0.50.

Availability: The MemLoci server is freely available on the web at: http://mu2py.biocomp.unibo.it/memloci. Datasets described in the article can be downloaded at the same site.

Contact: casadio@biocomp.unibo.it

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

Received on December 4, 2010; revised on February 22, 2011; accepted on February 23, 2011

1 INTRODUCTION

Knowledge of protein subcellular localization is crucial for unraveling its cellular function/s (Casadio *et al*., 2008). Different membrane types partition eukaryotic cells into functional compartments and mediate exchange of matter, energy and information. Biological activity within the different compartments strictly depends on proteins that are associated to their surrounding membranes and in turn knowing the subcellular localization of a given membrane protein helps in elucidating its functional role.

About 30% of the human proteins included in the SwissProt database are annotated as interacting with membranes (6609 out of 20 286 in release 08_2010). Nearly 20% of these are integral membrane proteins (also known as transmembrane proteins) that span the lipid bi-layer, either with α-helical or β-barrel domains. The remaining portion consists of peripheral membrane proteins associated to membranes with different mechanisms (Ghosh *et al*., 2008) that includes:

(i) posttranslational covalent attachment to lipid anchors such as glycophosphatidylinisitol (GPI) anchors (Chatterjee and Mayor, 2001), palmitoyl anchors (Greaves and Chamberlain, 2007) and myristoyl anchors (Maurer-Stroh *et al*., 2002);
(ii) interaction between the phospholipid bilayer and specific membrane-targeting domains exposed on the protein surface (Cho and Stahelin, 2005; Lemmon, 2008); and
(iii) association with proteins that directly interact with membranes.

The interaction of any single protein with a specific membrane can involve more than one of the above mechanisms.

Eukaryotic cell membranes can be divided into two main classes on the basis of both functional and phylogenetic considerations: the endomembrane system and the organelle membranes. The endomembrane system comprises the plasma membrane, the endoplasmic reticulum (ER), the Golgi complex, the nuclear envelope, the vesicles, the lysosomes and the vacuoles. These membranes are thought to have originated from invaginations of the ancestral prokaryotic plasma membrane (Jékely, 2007). Moreover, they form a single functional system, since they are either directly connected or communicate with each other through the vesicle transport system. The plasma membrane holds a special role in the endomembrane system, being the interface between the cell and its external environment.

The other class of membranes comprises the inner and outer membranes of mitochondria, chloroplasts and other plastids that do not exchange materials through the vesicle transport and that are likely to have an independent phylogenetic origin (de Duve, 2007).

The trafficking of membrane proteins from ribosomes to their final localization is a complex process that involves several molecular mechanisms, only partially unraveled. The best known is the recognition of specific targeting peptides that are present in the N- or C-terminal segments of the sequence. The most investigated among them are:

(i) N-terminal signal peptides targeting the proteins towards the ER, from where they are re-directed to the rest of the endomembrane system and to the secretory pathway (Hegde and Bernstein, 2006);
(ii) N-terminal transit peptides targeting the proteins towards the intracellular organelles, namely mitochondria and plastids (Balsera *et al*., 2009; Chacinska *et al*., 2009);

[*]To whom correspondence should be addressed.

(iii) C-terminal propeptides that are cleaved upon the attachment to a GPI-anchor that associates a protein to the plasma membrane (Pierleoni *et al*., 2008).

Other sequential and structural features controlling the trafficking of membrane proteins are known. Specific recognition of short sequence motifs exposed on the surface of proteins associated to the ER triggers their inclusion into vesicles and their redirection towards other membranes of the endomembrane system (Barlowe, 2003; Sato and Nakano, 2007), including the plasma membrane (Rodriguez-Boulan *et al*., 2005). Unfortunately, no consensus sequence is available for these motifs. Moreover, proteins can be associated to membranes by means of domains that are able to recognize the specific phospholipid composition and the shape of the target membranes (Cho and Stahelin, 2005; Lemmon, 2008).

The lack of a deep knowledge on the targeting mechanisms of membrane proteins is also due to the difficulty in carrying out the experimental determination of their localization: this is particularly true in the case of peripheral proteins. Several large-scale approaches have been attempted in order to evaluate the proteome of different membranes in different organisms (for a review, see Sadowski *et al*., 2008). In particular, modern proteomic and labeling techniques have been applied for determining the plasma membrane proteome in *Arabidopsis* and rice (Komatsu, 2008), and the inner and outer membrane proteomes of mitochondria in rat and mouse (Distler *et al*., 2008). However, owing to the intrinsic limitations of the experimental procedure, the complete characterization of the proteomes of the different eukaryotic membranes is still a challenging task (Sadowski *et al*., 2008), and computational methods can be adopted for rapid and inexpensive prescreening and additional analysis.

Several tools have been developed for recognizing membrane-interacting proteins starting from their residue sequences. They are based on the prediction of transmembrane domains, on the recognition of posttranslational modification sites, and on the identification of lipid-binding domains. However, none of these predictors is able to identify the membrane where the protein is localized (Eisenhaber and Eisenhaber, 2007; Punta *et al*., 2007). Currently, similarity-based methods or general purpose localization predictors, mainly trained on globular proteins, are adopted for accomplishing this task (Briesemeister *et al*., 2010; Guda and Subramaniam, 2005; Horton *et al*., 2007; Lin *et al*., 2009; see for recent reviews: Casadio *et al*., 2008; Imai and Nakai, 2010). The only method specifically suited to the localization of membrane proteins is described in a recent work from Sharpe *et al*. (2010). The method is based on the analysis of the differences in residue composition and length of membrane-spanning regions of transmembrane proteins from ER, Golgi and plasma membranes. A predictor is available only for single-span transmembrane proteins, not localized in mitochondria or plastids, and coming from vertebrates and fungi. In this article, we describe MemLoci, a tool based on support vector machines (SVMs) specifically suited to predict the localization of all the eukaryotic membrane proteins, both integral and peripheral. This is prompted by considering the limitations of tools based on sequence similarity for annotating subcellular localization of membrane proteins and by proving that the existing general purpose predictors for subcellular localization are largely unsatisfactory in the case of membrane proteins.

Scarcity of available annotated data prevents from implementing a predictor that includes a complete classification of all the membranes. Three localizations are then taken into account: the plasma membrane, the internal membranes (i.e. the endomembrane system with the exclusion of the plasma membrane) and the membranes of organelles (mitochondria and plastids). The input coding includes 123 different features, which cast the available knowledge about the molecular mechanism and the sequence signals controlling the sorting of membrane proteins.

The method is freely available on the web at: http://mu2py.biocomp.unibo.it/memloci.

2 METHODS

2.1 The dataset

We extracted from the SwissProt database (release 8_2010, July 2010) all the eukaryotic membrane proteins with known subcellular localization by parsing the 'SUBCELLULAR LOCATION' section of the COMMENT field. Sequences whose annotations were marked as 'PROBABLE', 'POSSIBLE', 'BY SIMILARITY' and 'FRAGMENT' were discarded. We ended up with 24 640 sequences of membrane proteins, including transmembrane and peripheral ones, annotated according to 10 different subcellular localizations. Localizations were then reduced to the three major classes (plasma membrane, organelle membranes and internal membranes) following the scheme detailed in Table 1. We discarded proteins endowed with multiple localizations in our three-class partition.

In order to reduce the redundancy of the dataset, proteins were grouped into similarity sets. All-against-all alignments were performed with BLAST. A graph was built by linking all the pairs of sequences sharing >80% of identical residues. The connected components of this graph, as determined with the transitive closure algorithm (Cormen *et al*., 2001), defined 10 634 non-overlapping subsets. Only one protein per subset was retained. As listed in Table 1, our complete dataset consists of 10 634 sequences: 4016 from plasma membrane, 2308 from organelle membranes and 4310 from the internal membranes. We adopted this reduced dataset for training and evaluating MemLoci.

In order to compile training/testing sets containing sequences with low homology, the reduced dataset (10 634 sequences) was then clustered into subsets containing proteins that align with an E-value $\leq 10^{-3}$ with a transitive

Table 1. The dataset

Membrane	Metazoa	Fungi	Viridiplantae	Total
Plasma	7458 (*3233*)	623 (*371*)	579 (*412*)	8660 (*4016*)
Organelle	5065 (*1047*)	1519 (*1089*)	440 (*172*)	7024 (*2308*)
Mitochondrion	5065 (*1047*)	1519 (*1089*)	339 (*147*)	6923 (*2283*)
Plastids	–	–	101 (*25*)	101 (*25*)
Internal	5390 (*1688*)	2904 (*2172*)	662 (*450*)	8956 (*4310*)
ER	2617 (*939*)	1575 (*1245*)	239 (*165*)	4431 (*2349*)
Golgi	1439 (*514*)	593 (*447*)	222 (*165*)	2254 (*1126*)
Nucleus	198 (*93*)	174 (*152*)	20 (*15*)	392 (*260*)
Vesicles	238 (*65*)	120 (*74*)	0 (*0*)	358 (*139*)
Vacuole	24 (*7*)	438 (*354*)	134 (*78*)	596 (*349*)
Peroxisome	142 (*52*)	87 (*80*)	23 (*18*)	87 (*80*)
Lysosome	305 (*110*)	18 (*13*)	4 (*4*)	327 (*127*)
Endosome	263 (*66*)	230 (*143*)	22 (*14*)	515 (*223*)
Microsome	596 (*264*)	51 (*46*)	26 (*19*)	663 (*329*)
TOTAL	17 913 (*5968*)	5046 (*3632*)	1681 (*1034*)	24 640 (*10 634*)

Figures in roman and italic types refer to the complete dataset and to the reduced dataset adopted for training and testing MemLoci, respectively.

closure procedure. We ended up with 1547 clusters and by construction proteins belonging to two different clusters share low similarity, since they align with an E-value $>10^{-3}$

In order to build an independent validation set, we retained all the proteins released after January 1, 2008 and included in clusters containing a single sequence. With this procedure we obtained 100 sequences that: (i) share low similarity with all the other proteins in the dataset; and (ii) were not used to train the other subcellular localization methods, except the tool by Sharpe *et al*. (2010). The validation set consists of 32 plasma membrane, 50 internal membranes and 18 organelle membranes proteins.

The remaining 10 534 sequences included in 1447 clusters were used to train MemLoci. To this aim, 10 cross-validation sets were then defined by randomly grouping these non-similar clusters.

2.2 The predictor architecture

The MemLoci prediction system consists of three binary SVM classifiers specifically suited at discriminating one of the three considered localization classes (plasma membrane, organelle membranes and internal membranes). During the training procedure of each one of the three SVMs, the sequences belonging to the class to be discriminated were presented as positive examples, while all the remaining sequences were presented as negative examples. The SVM-light package (version 6, freely available at http://svmlight.joachims.org) was used, adopting the radial basis function (RBF) kernel and setting the parameters Gamma and C to 3 and 6, respectively, upon a grid search in the parameter space. The three SVMs were trained with three different sets of features, each one optimized for a specific class (see Section 2.3). In this way, the training procedure determined three optimal separating hyperplanes in three different feature spaces. Given a protein sequence, each SVM assigns a binary positive/negative classification based on the discriminating hyperplane. The reliability of the classification improves at increasing distance from the hyperplane. A jury among the three binary predictions is built in the following way. For each example and for each class-specific SVM we compute a signed distance, whose absolute value is the distance from the discriminating hyperplane and whose sign indicates whether the example is classified in the positive or negative set. The final prediction is based on a simple winner-takes-all jury that assigns the example to the class corresponding to the highest value for the signed distance.

2.3 Input coding

The input vector encodes features that can be grouped into the following five classes:

(i) *The composition of the sequence profile*: sequence profiles are compiled aligning the target sequence with the release 8_2010 of the SwissProt database using BLAST and adopting a threshold E-value equal to 10^{-5}. The profile consists of a sequence of 20-valued vectors reporting the residue composition of the alignment for each position of the query sequence. The sequence profile composition for a given segment is obtained by averaging the 20-valued vectors over the segment, and consists of a 20-valued vector with elements ranging between 0 and 1.

(ii) *The composition of the filtered sequence profile*: in order to enhance the contribution of evolutionary conserved positions, the sequence profile is also filtered by retaining only the residues whose frequencies at a given position are higher than a given threshold. Two thresholds are adopted, 0.33 and 0.75. This procedure extracts highly conserved local signatures, some of which are likely to contain important localization signals. The filtered profile composition of each segment consists of a 20-valued vector with elements ranging between 0 and 1.

(iii) *The average hydrophobicity of N and C sequence termini*: the average hydrophobicity of N and C sequence termini as computed considering for each terminus two segments comprising either 40 or 60 residues. This feature aims at highlighting N and C sequence termini that carry signal and transit peptides or posttranslational modification signals, often consisting of highly hydrophobic regions. The Kyte–Doolittle scale is adopted (Kyte and Doolittle, 1982). In order to standardize the range of the input values, the Kyte–Doolittle parameters are linearly rescaled so that the highest value (4.5, for isoleucine) and the lowest value (−4.5, for arginine) are mapped to 1 and −1, respectively. The transformed values are then averaged over the selected segment. The average hydrophobicity of a segment is encoded with a single real value ranging between −1 and 1.

(iv) *The residue composition of highly hydrophobic stretches*: highly hydrophobic stretches are identified by averaging the (rescaled) Kyte–Doolittle hydrophobicity over a 13-residue sliding window and by selecting segments reporting an average hydrophobicity >0.33. The residue composition of highly hydrophobic stretches is then calculated. This input is then encoded in a 20-valued vector whose elements range between 0 and 1. The hydrophobic stretches are likely to contain localization signals included in membrane spanning regions and the surrounding residues (Sharpe *et al*., 2010).

(v) *The protein length*: the number of residues of the protein sequence is rescaled by suitable factor (1/2000). The scaling factor was chosen considering that 90% of the proteins in our dataset are less than 2000 residues long.

Each of the three class-specific SVM-based predictors is implemented independently of the others two with its specific input encoding. The input encoding was selected after a search in the encoding space, adopting a 10-fold cross-validation procedure. Eight sets were used to train the binary SVMs, one was used for evaluating the best input coding, and the remaining one was used to compute the final performances reported in the Section 3. By this, the evaluation of the predictor performance is unaffected by any possible overfitting resulting from the input optimization procedure.

Since the best performing input encoding is very similar across the 10 cross-validation groups, we adopted a majority consensus for selecting the feature included in the final predictor. Features used for each predictor are listed in Table S1 (Supplementary Material). The optimization of the RBF kernel parameters for each training set was performed with a grid search. The 10 sets gave very similar values, which were therefore averaged to obtain the parameters used for training the final predictor, namely $\gamma=3$ and C = 6.

2.4 Scoring indexes

The performance is evaluated by predicting all the membrane proteins in the dataset with the 10-fold cross-validation procedure described above. Since proteins belonging to the same cluster can share up to 80% identity, we adopted the following normalizing procedure in order to avoid biases of the predictive scores towards the most populated clusters.

For each one of the three classes (i) and for each one of the 1352 similarity clusters (j) we computed $\mathrm{TP}_j(i)$ and $\mathrm{TF}_j(i)$, the number of correct predictions in the class i and in the complementary class, respectively, normalized to the total number of sequences in the cluster. In a similar way we computed the ratio of false positive and false negative predictions, $\mathrm{FP}_j(i)$ and $\mathrm{FN}_j(i)$, respectively. These numbers were then summed over all the clusters, obtaining TP(i), TN(i), FP(i) and FN(i), that give an unbiased estimation of the number of correct and wrong predictions in the class i.

The recall (Rec), the false positive rate (FPR) and Matthews correlation coefficient (MCC) for the class i are then computed as follows:

$$\mathrm{Rec}(i)=\frac{\mathrm{TP}(i)}{\mathrm{TP}(i)+\mathrm{FN}(i)} \tag{1}$$

$$\mathrm{FPR}(i)=\frac{\mathrm{FP}(i)}{\mathrm{TN}(i)+\mathrm{FP}(i)} \tag{2}$$

$$\mathrm{MCC}(i)=\frac{\mathrm{TP}(i)\cdot\mathrm{TN}(i)-\mathrm{FP}(i)\cdot\mathrm{FN}(i)}{\sqrt{\left(\mathrm{TP}(i)+\mathrm{FP}(i)\right)\left(\mathrm{TP}(i)+\mathrm{FN}(i)\right)\left(\mathrm{TN}(i)+\mathrm{FP}(i)\right)\left(\mathrm{TN}(i)+\mathrm{FN}(i)\right)}} \tag{3}$$

The global performances were evaluated with the overall accuracy (Q)

$$Q=\frac{\sum_{i=1}^{3}\mathrm{TP}(i)}{N} \tag{4}$$

where N is the total number of clusters.

Moreover, we adopted the generalized correlation (GC) described by Baldi *et al*. (2000), which is a generalization of the MCC in the case of many classes.

3 RESULTS

3.1 MemLoci at work

We developed MemLoci, a predictor of subcellular localization specifically suited for integral and peripheral membrane proteins. Three major classes are discriminated: plasma, internal and organelle membrane. A complete partition is presently not feasible given the paucity of non-redundant experimental data. The predictor consists of three binary SVMs that analyze the composition, the conservation, the hydrophobicity of the whole sequence and for each sequence the same features for its N- and C-terminal portions. A winner-takes-all jury further processes the output of the three class-specific SVMs (see Section 2.2 for details).

A rigorous 10-fold cross-validation procedure was adopted for evaluating the performance of the method. By construction, pairs of sequences belonging to two different sets align with an E-value $> 10^{-3}$. In order to estimate to which extent the residual sequence identity could affect the scores, we also performed a cross-validated annotation transfer by similarity: for each sequence in each one of the 10 sets, we searched with BLAST the most similar sequence in the other sets, and we transferred the corresponding annotation.

Table 2 summarizes the prediction performance. In order not to over-represent the most populated similarity clusters, the scores have been weighted as detailed in Section 2. The low values of the correlation coefficients obtained with the annotation transfer by similarity indicate that the prediction is close to random. This means that the E-value threshold adopted for splitting the dataset fits with the need of avoiding any overestimation caused by the similarity between training and testing sets.

The reported performances corroborate the notion that MemLoci is suited for solving the annotation problem when no similar sequence endowed with an annotation for subcellular localization is available. The generalized correlation index is 0.51, while the Matthews correlation indexes are as high as 0.43, 0.42 and 0.60 for plasma, internal and organelle membranes, respectively. A detailed analysis of the prediction reveals that the most frequent error of MemLoci is to include plasma membrane proteins within the internal membrane class, and vice-versa. About 30% of plasma membrane proteins are predicted as internal membrane and 20% of internal membrane proteins are predicted in the plasma membrane class (Data not shown).

3.2 Comparison with general purpose subcellular localization predictors

Many methods are available for predicting the protein subcellular localization starting from sequence. Almost none of them is however specific for membrane proteins. Some include plasma membrane proteins in a specific class and do not make any distinction between globular and membrane proteins for the other localizations. Among these tools, the most popular are WoLF-PSORT (Horton *et al*., 2007), pTARGET (Guda and Subramaniam, 2005) and Yloc-HiRes (Briesemeister *et al*., 2010). Other predictors, such as BaCelLo (Pierleoni *et al*., 2006), explicitly exclude membrane proteins from their training/testing sets. BaCelLo and WoLF-PSORT resulted the two best performing tools for predicting subcellular localization in a recent comparative evaluation performed on a set of globular proteins (Casadio *et al*., 2008). YLoc and KnowPred were reported to outperform both WoLF-PSORT and BaCelLo on a recent test consisting of animal sequences (Briesemeister *et al*., 2010)

We tested MemLoci and the other state-of-the-art methods on an independent dataset comprising only sequences released after January 2008 and comprising 32 plasma membrane, 50 internal membranes and 18 organellar membranes proteins. None of the protein in the independent test aligns with an E-value lower than 10^{-3} to the MemLoci training dataset.

Only 52 proteins from animals or fungi have been submitted to pTARGET since this tool is not designed to predict proteins from plants. When using YLoc we adopted the HiRes version that defines 11 localizations. YLoc can take advantage of gene ontology (GO) term annotations extracted from proteins similar to the query sequences, as retrieved from UniProtKB. Therefore, for sake of a fair comparison, we disabled this option. This procedure, according to the authors, results only in small differences when predicting proteins devoid of annotation (Briesemeister *et al*., 2010). Yloc and BaCelLo define three different sets of parameters, for animal, plants and fungi, and each sequence was predicted adopting the appropriate parameter set.

The following schema is adopted for reducing the classes defined by different predictors (5 for BaCelLo, 11 for Yloc-HiRes, 12 for WoLF-PSORT, 9 for pTARGET and 10 for KnowPred) to the three major classes we take into consideration, namely ‘plasma membrane’, ‘internal membranes’ and ‘organelle membranes’

- Proteins predicted as secreted or localized in the plasma membrane are considered in the ‘plasma membrane’ class. Secreted proteins are indeed often associated to the external

Table 2. Performance of MemLoci and of the annotation upon similarity search (BLAST) using a 10-fold cross-validation procedure

Method	Plasma membrane			Internal membranes			Organelle membranes			Global indexes	
	Rec%	FPR%	MCC	Rec%	FPR%	MCC	Rec%	FPR%	MCC	*Q*%	GC
MemLoci	56	15	0.43	72	30	0.42	70	9	0.60	66	0.51
BLAST	53	37	0.14	49	40	0.09	26	10	0.19	45	0.16

For the definitions, see Section 2.4.

Table 3. Performance of general purpose localization predictors on the validation set

Method	Plasma membrane			Internal membranes			Organelle membranes			Global indexes	
	Rec%	FPR%	MCC	Rec%	FPR%	MCC	Rec%	FPR%	MCC	*Q*%	GC
MemLoci	56	9	0.52	86	34	0.53	50	9	0.43	70	0.50
BaCelLo	38	15	0.26	60	44	0.16	56	20	0.32	52	0.28
WolfPSort	25	10	0.17	82	60	0.24	44	7	0.41	57	0.32
pTARGET[a]	18	31	−0.13	60	69	−0.09	40	5	0.39	48	0.32
KnowPred	44	31	0.12	63	58	0.05	17	1	0.30	48	0.25
YLoc	56	34	0.21	56	40	0.16	44	4	0.50	54	0.37
BLAST	63	44	0.17	48	26	0.23	22	11	0.13	48	0.19

For the definitions, see Section 2.4. [a]Only fungi and animals sequences were submitted to pTARGET.

leaflet of the plasma membrane. It is worth stressing that BaCelLo does not define a plasma membrane class.

- Proteins predicted as mitochondrial or chloroplastic are considered in the ‘organelle membranes’ class.
- Proteins predicted in all the other localizations are considered in the ‘internal membranes’ class.

In the case of KnowPred, which computes a score value for each localization, only the best score is taken into consideration. Based on this schema, all the predictions were compared to the UniProtKB experimental annotations. We performed an annotation transfer by similarity, with the same procedure described in Section 3.2, in order to evaluate the performances of a baseline predictor that considers only similarity among the sequences included in the validation set and those that form the training set.

Scoring indexes on validation set are listed in Table 3. MemLoci, specifically trained on membrane proteins, outperforms the other tools in all the considered classes: MCC values of ‘plasma membrane’ and ‘internal membranes’ double with respect to the best prediction achieved with the other methods (from 0.26 to 0.52 and from 0.24 to 0.53, respectively). Only Yloc-HiRes outperforms MemLoci in the ‘organelle membranes’ class, but it scores much worse in the two other classes. The global scoring indexes of MemLoci ($\mathrm{GC}=0.50$ and $Q=70\%$) are therefore far higher than those obtained with all the other methods. The comparison with BLAST results confirm that the good performances are not due to residual similarity between the training and the validation sets. All the other general purpose methods performs poorly on the independent dataset and performances are in some case very close to those reported by the baseline annotation based on transfer by similarity. BaCelLo and WolfPSort hardly achieve half of the GC coefficients previously reported in the case of non-membrane proteins (Casadio *et al*., 2008). It is worth noticing that BaCelLo, which explicitly excludes membrane proteins from its training set, scores with a performance similar to those of WoLF-PSORT and pTARGET, which include them. The most accurate of the general purpose methods is YLoc that scores with GC coefficient and overall accuracy (Q) of 0.37 and 54%, respectively. However both index values are considerably lower than those obtained by MemLoci ($\mathrm{GC}=0.50$ and $Q=70\%$).

The GC value of MemLoci on the validation set well correlates with that reported on the training set (Table 3 and Supplementary Table S2, respectively), proving that the adopted cross-validation procedure is not affected by biases. Considering separately the three classes, the validation results are even higher than the training ones, but in the case of organelle membranes. For most of the other tools, the validation performances are lower than those achieved on the MemLoci training set (reported in Supplementary Table S2) since many sequences in this set were used to parameterize the different methods. As in the case of MemLoci, on the ‘organelle membranes’ class also the other methods, but Yloc, show lower results on the validation set than on the training set.

The only method devoted to the prediction of subcellular localization of membrane proteins has been recently developed by Sharpe *et al*. (2010).This neural network-based method is able to discriminate between single-spanning membrane protein localized in the plasma, post-Golgi, pre-Golgi and ER membrane. It explicitly excludes proteins from mitochondria and plastids. Moreover, it is suited only to fungal and animal single-span transmembrane proteins. These limitations restrict the comparison to only 18 proteins (8 plasma membrane and 10 internal membrane). All the internal membrane proteins are correctly predicted in the post-Golgi, pre-Golgi or ER membrane localizations. Only three out of eight plasma membrane sequences are correctly predicted. On the same 18 sequence-test, MemLoci correctly predicts all the 10 proteins from internal membranes and 7 out of 8 plasma membrane proteins.

3.3 Analysis of the most significant features

Predictions of MemLoci derive from non-linear combination of several features as performed by SVMs based on RBF kernel. In order to evaluate the most discriminating ones, for each binary classifier the means and the SDs of all the features were computed over each one of the 10 testing sets. The significance of the difference between the average values of the positive and the negative examples was assessed with the P-value, as estimated with a two-tailed Student’s t-test. Supplementary Tables S3–S5 list, for each localization type, features with $P\text{-values} \leq 10^{-10}$ in all the 10 testing sets. A summary of the findings is reported in Table 4. Several compositional differences are observed, mainly in the whole sequence and in N-terminal regions. Increasing length is relevant for plasma membrane proteins and decreasing length is characteristic of organelle membrane proteins. At the C-terminus, the composition in Ser residues is enriched in plasma membrane proteins and depleted in organelle membrane proteins. Differences in C- and N-termini are likely to cast specific features of the signals responsible for the sorting mechanisms, often mediated by protein–protein interaction.

Table 4. Most relevant features for each binary discrimination

	Whole sequence	N-terminal regions	C-terminal regions	Length
Plasma membrane	Cys↑, Ile↑, Lys↓	Cys↑,Gly↑, Lys↓	Ser↑	↑
Internal membranes	Asp↑,Met↓, Thr↓	Asp↑,Met↓, Thr↓		
Organelle membranes	Lys ↑, Met↑, Cys ↓, Ala↑	Met↑, Asp↓, Glu↓, Val↓	Ser↓	↓

Up arrow: features enriched in the class; down arrow: features enriched in the complementary class.

For example, it has been shown that transit peptides of mitochondrial proteins tend to be positively charged and rich in apolar residues (Pfanner and Geissler, 2001), and these observations agree with the increased composition in Lys and Ala and the decreased composition in Asp and Glu that we extract as significant features distinctive of organelle membrane proteins. Out of the 10 significant features distinctive of proteins from organelle membranes (Table 4), four are also significant for plasma membrane proteins, although with the opposite sign. The same holds for three features typical of proteins from organelle membranes that are also significant for internal membrane proteins. The better performance of MemLoci on organelle membrane proteins may be accounted for by considering that this class share significant features with both the others with opposite sign and it can therefore be better discriminated. This analysis highlights some features that are significantly different in the different protein classes, and are therefore relevant for accomplishing the prediction. Most of them, to our knowledge, were not previously identified. It is worth stressing that, although statistically significant, the reported features are not strong enough to perform an efficient discrimination, unless they are combined in a complex way with all the other considered features. Indeed SVMs take advantage of a non-linear combination that it is not simply accounted for by feature differences among classes.

4 CONCLUSIONS

Here we present MemLoci, an efficient method for predicting the subcellular localization of integral and peripheral membrane proteins. MemLoci outperforms currently state-of-the-art methods for the prediction of subcellular localization, which were not trained specifically to solve this task. MemLoci can then be adopted for characterizing the localization of both the experimental and predicted membrane proteomes of any eukaryotic organism. MemLoci can be included in a pipeline, in succession to predictive methods devoted to the identification of proteins associated to the membrane phase. Efficient predictors are available for discriminating the different types of integral membrane proteins, both transmembrane or lipid-anchored (Chou and Chen, 2007; Martelli *et al*., 2003; Nugent and Jones, 2009; Pierleoni *et al*., 2008). In particular, in the case of transmembrane proteins, predictive methods achieve the discrimination task with $<1\%$ false positive or false negative rates. Moreover, MemLoci can be adopted to define the localization of proteins detected with different experimental techniques specifically developed for determining the membrane proteome (Lu *et al*., 2008; Rabilloud, 2009). One possible application of MemLoci is then to discriminate the fraction of the membrane proteome interacting with the plasma membrane and by this significantly restricting the protein set where new antigens, biomarkers and drug targets can be discovered.

Funding: FIRB 2003 LIBI–International Laboratory of Bioinformatics (to R.C.)

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.

Balsera,M. *et al*. (2009) Protein import machineries in endosymbiotic organelles. *Cell. Mol. Life Sci.*, **66**, 1903–1923.

Barlowe,C. (2003) Signals for COPII-dependent export from the ER: what's the ticket out? *Trends Cell. Biol.*, **13**, 295–300.

Briesemeister,S. *et al*. (2010) Going from where to why–interpretable prediction of protein subcellular localization. *Bioinformatics*, **26**:1232–1238.

Casadio,R. *et al*. (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief. Funct. Genomic. Proteomic.*, **7**, 63–73.

Chacinska,A. *et al*. (2009) Importing mitochondrial proteins: machineries and mechanisms. *Cell*, **138**, 628–644.

Chatterjee,S. and Mayor,S. (2001) The GPI-anchor and protein sorting. *Cell. Mol. Life Sci.*, **58**, 1969–1987.

Cho,W. and Stahelin,R.V. (2005) Membrane-protein interactions in cell signaling and membrane trafficking. *Ann. Rev. Biophys. Biomol. Struct.*, **34**, 119–151.

Chou,K.C. and Shen,H.B. (2007) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.*, **360**, 339–345.

Cormen,T.H. *et al*. (2001) *Introduction to algorithms*. 2nd edn. MIT press, Cambridge, MA, USA.

De Duve,C. (2007) The origin of eukaryotes: a reappraisal. *Nat. Rev. Genet.*, **8**, 395–403.

Distler,A.M. *et al*. (2008) Proteomics of mitochondrial inner and outer membranes. *Proteomics*, **8**, 4066–4082.

Eisenhaber,B. and Eisenhaber,F. (2007) Posttranslational modifications and subcellular localization signals: indicators of sequence regions without inherent 3D structure? *Curr. Protein Pept. Sci.*, **8**, 197–203.

Ghosh,D. *et al*. (2008) The identification and characterization of membranome components. *J. Proteome Res.*, **7**, 1572–1583.

Greaves,J. and Chamberlain,L.H. (2007) Palmitoylation-dependent protein sorting. *J. Cell. Biol.*, **176**, 249–254.

Guda,C. and Subramaniam,S. (2005) pTARGET: a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics*, **21**, 3963–3969.

Hegde,R.S. and Bernstein,H.D. (2006) The surprising complexity of signal sequences. *Trends Biochem. Sci.*, **31**, 563–571.

Horton,P. *et al*. (2007) WoLF PSORT: protein localization predictor. *Nucl. Acid. Res.*, **35**, W585–W587.

Imai,K. and Nakai,K. (2010) Prediction of subcellular locations of proteins: where to proceed? *Proteomics*, **10**, 3970–3983.

Jékely,G. (2007) Origin of eukaryotic endomembranes: a critical evaluation of different model scenarios. *Adv. Exp. Med. Biol.*, **607**, 38–51.

Komatsu,S. (2008) Plasma membrane proteome in *Arabidopsis* and rice. *Proteomics*, **8**, 4137–4145.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Lemmon,M.A. (2008) Membrane recognition by phospholipid-binding domains. *Nat. Rev. Mol. Cell. Biol.*, **9**, 99–111.

Lin,H.N. *et al*. (2009) Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. *BMC Bioinformatics*, **10** (Suppl. 15), S8.

Lu,B. *et al.* (2008) Strategies for shotgun identification of integral membrane proteins by tandem mass spectrometry. *Proteomics*, **8**, 3947–3955.

Martelli,P.L. *et al*. (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics*, **19**, i205–i211.

Maurer-Stroh,S. *et al*. (2002) N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. *J. Mol. Biol.*, **317**, 541–557.

Nugent,T. and Jones,D.T. (2009) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**, 159.

Pfanner,N. and Geissler,A. (2001) Versatility of the mitochondrial protein import machinery. *Nat. Rev Mol. Cell Biol.* **2**, 339–349.

Pierleoni,A. *et al*. (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.

Pierleoni,A. *et al*. (2008) PredGPI: a GPI-anchor predictor. *BMC Bioinformatics*, **9**, 392.

Punta,M. *et al*. (2007) Membrane protein prediction methods. *Methods*, **41**, 460–474.

Rabilloud,T. (2009) Membrane proteins and proteomics: love is possible, but so difficult. *Electrophoresis*, **30**, S174–S180.

Rodriguez-Boulan,E. *et al*. (2005) Organization of vesicular trafficking in epithelia. *Nat. Rev. Mol. Cell. Biol.*, **6**, 233–247.

Sadowski,P.G. *et al*. (2008) Sub-cellular localization of membrane proteins. *Proteomics*, **8**, 3991–4011.

Sato,K. and Nakano,A. (2007) Mechanisms of COPII vesicle formation and protein sorting. *FEBS Lett.*, **581**, 2076–2082.

Sharpe,H.J. *et al*. (2010) A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell*, **142**, 158–69.