BIOINFORMATICS

Vol. 22 no. 14 2006, pages e408–e416
doi:10.1093/bioinformatics/btl222

BaCelLo: a balanced subcellular localization predictor

Andrea Pierleoni, Pier Luigi Martelli, Piero Fariselli and Rita Casadio*

Biocomputing Group, Dept. of Biology University of Bologna, via Irnerio 42, 40126 Bologna, Italy

ABSTRACT

Motivation. The knowledge of the subcellular localization of a protein is fundamental for elucidating its function. It is difficult to determine the subcellular location for eukaryotic cells with experimental high-throughput procedures. Computational procedures are then needed for annotating the subcellular location of proteins in large scale genomic projects.

Results. BaCelLo is a predictor for five classes of subcellular localization (secretory pathway, cytoplasm, nucleus, mitochondrion and chloroplast) and it is based on different SVMs organized in a decision tree. The system exploits the information derived from the residue sequence and from the evolutionary information contained in alignment profiles. It analyzes the whole sequence composition and the compositions of both the N- and C-termini. The training set is curated in order to avoid redundancy. For the first time a balancing procedure is introduced in order to mitigate the effect of biased training sets. Three kingdom-specific predictors are implemented: for animals, plants and fungi, respectively. When distributing the proteins from animals and fungi into four classes, accuracy of BaCelLo reach 74% and 76%, respectively; a score of 67% is obtained when proteins from plants are distributed into five classes. BaCelLo outperforms the other presently available methods for the same task and gives more balanced accuracy and coverage values for each class. We also predict the subcellular localization of five whole proteomes, *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*, comparing the protein content in each different compartment.

Availability. BaCelLo can be accessed at http://www.biocomp.unibo.it/bacello/

Contact. casadio@alma.unibo.it, andrea@biocomp.unibo.it, gigi@biocomp.unibo.it, piero@biocomp.unibo.it

1 INTRODUCTION

The eukaryotic cell is a composite system internally subdivided into membrane-enveloped compartments that perform particular functions. Every subcellular compartment contains specific proteins, including enzymes, synthesized in the cytoplasm and translocated into the locations, where they carry out functional patterns. Therefore, knowing the localization of every protein is important for elucidating its interactions with other molecules and for understanding its biological function. Experimental high-throughput approaches have been applied to determine protein localization in *Saccharomyces cerevisiae* (Huh *et al*., 2003) and in *Arabidopsis thaliana* (Kleffmann *et al*., 2004). However these techniques cannot be generally applied to all the eukaryotic cells and computational predictive methods are needed in order to screen the huge amount of data derived from genomic projects and to guide the design of experiments.

Intracellular protein sorting involves several post-translational mechanisms that redirect a newly synthesized chain from the cytosol to its specific compartment on the basis of the information contained in its residue sequence. Pre-translational mechanisms, involving the sorting of the mRNAs inside the cytosol, seem to play a minor role in the translocation between different compartments (Gonsalvez *et al*., 2005). These considerations lead to the conclusion that the residue sequence of a protein is mostly responsible for its localization.

It is well known that many sequences contain cleavable peptides at the N-terminus that address the protein either to the secretory pathway, in which case they are called signal peptides, or to mitochondria and plastids and in this case they are called target or transit peptides. Some predictors have been implemented in order to recognize N-terminal signal peptides (Nielsen *et al*., 1997; Fariselli *et al*., 2003) or both the signal and the transit peptides (Emanuelsson *et al*., 2000). However some proteins get secreted by means of a non-classical way and do not require N-terminal signal peptides (Nickel, 2003; Bendtsen *et al*., 2004). Furthermore some proteins are translocated into mitochondria owing to a localization peptide at the C-terminal region (Lee *et al*., 1999; Izeta *et al*., 2003; Yamada *et al*., 2004). For these reasons in order to obtain an accurate prediction it is necessary to add information besides the N-terminal composition.

Concerning nuclear proteins, they have to span the nuclear membranes through the proteic Nuclear Pore Complexes. Several crossing mechanisms have been described, including free diffusion and mediated transport. Nuclear Localization Signals have been reported (Fried and Kutay, 2003) and approaches for finding them into a protein sequence were tested. However, the methods that incorporate only this type of information do not achieve satisfactory performance, probably due to the shortness and low specificity of the signals. Moreover only 30% of nuclear proteins is estimated to have a NLS (Cokol *et al*., 2000).

A number of different predictors for the subcellular localization have been released in the past years, based on different approaches. They can be divided into two major classes, following Nair and Rost (2005): predictors based on the knowledge extracted from the annotated databases and the so called *de novo* predictors. The former ones are based on the detection of similarity between the sequence to be predicted and sequences with known localization, by searching for homology (Marcotte *et al*., 2000) or for conserved domains or motifs (Scott *et al*., 2004). However these tools are able to predict

*To whom correspondence should be addressed.

the localization of half of the sequences in the data bases (Nair and Rost, 2005). The *de novo* methods, which we are interested in, are more general and rely only on the analysis of the residue sequence, without inferring the annotation from any known sequences. pTARGET (Guda and Subramaniam, 2005) and ESLpred (Bhasin and Raghava, 2004) are hybrid methods since they take into consideration both the results of PFAM or BLAST and the analysis of the residue composition. Among the *de novo* methods, some exploit only the information contained in the N-terminal portions of the sequences, some consider the overall residue composition, and others take advantage of the evolutionary information contained in sequence profiles. They make use of both standard statistical methods and machine learning approaches, such as Neural Networks and Support Vector Machines. Moreover they discriminate different number of classes, from 3 up to 12. All the available predictors infer their parameters from unbalanced data sets: the size of different classes merely reflects the presence in the data bases of annotated sequences and it is unlikely that it can represent an estimate of the actual proportions in a living cell, as previously estimated at genomic level (Nair *et al*., 2005). This lead to an overestimation of the prediction for the most represented classes, namely the nuclear proteins for animal or fungi and the chloroplastic ones for plants.

Here we describe a novel method for subcellular localization prediction that adopts a balancing procedure by assuming a uniform *a priori* probability for the classes. The predictor makes use of several Support Vector Machines (SVMs), draws information from both the protein sequence and its profile derived with a BLAST search in the database of eukaryotic proteins, and considers in an explicit way the compositions of the whole sequence and of both the N- and C- termini.

Particular care has been taken in selecting the training set. Indeed most of the available methods, except LocTree (Nair and Rost, 2005), take into consideration proteins sharing high level of identity, up to 95%, and do not adopt rigorous validation procedures for excluding homology between the training and the testing sets. The justification for this procedure is that the subcellular localization of a protein can change owing to the change of few residues in the sequence, as in the case of the short nuclear localization signals. However, as we prove in this paper, in most practical cases, the high level of similarity between two sequences determines the same localization for the two proteins; in these cases a simple assignment based on the transfer of annotation after a BLAST search achieves a very good performance, even better than those reported by more sophisticated methods. Redundant data sets can therefore lead to methods that have a poor generalization capability. For these reason we select non-redundant data sets comprising proteins sharing less than 30% identity.

We consider five subcellular compartments: the secretory pathway, the cytoplasm, the nucleus, the mitochondrion and the chloroplast, when present. We decided not to predict more classes since for other locations the number of annotated non homologous protein chians is very low, not enough to train a predictor with a good generalization capability. We did not consider membrane proteins, since efficient methods for the prediction of transmembranicity are available, with very low rate of false positives and false negatives (about 3%, Martelli *et al*., 2003).

Differences concerning the localization mechanisms among the different kingdoms have been reported and we trained three different systems for animals, fungi and plants, respectively. In particular BaCelLo is the first *de novo* predictor that distinguishes between animal and fungal organisms. BaCelLo also takes advantage of the evolutionary information contained in sequence profiles that are known to improve performances of predictors for protein structure and function.

2 MATERIALS AND METHODS

2.1 Data sets

Starting from release 48 of the SWISS-PROT data base (Bairoch *et al*., 2005), we generated three data sets for animals (*Metazoa*), fungi (*Fungi*) and plants (*Viridiplantae*), respectively. Proteins with an experimental annotation of the subcellular location were retained, excluding those in which the comments 'fragment', 'possible', 'probable' and 'by similarity' are reported. We also excluded proteins with multiple subcellular localization and proteins shorter than 50 residues. The entries annotated as 'membrane'or 'transmembrane' were discarded, since we are interested only in globular proteins. The three data sets were separately clustered with an identity level equal to 30% using the BLASTCLUST tool and checked with BLASTp (Altschul *et al*., 1990). One representative protein for each cluster was selected. This procedure led to 2597 proteins from animals, 1198 proteins from fungi and 491 proteins from plants, distributed in five locations: nucleus, cytoplasm, secretory pathway (comprehending proteins annotated as 'Secretory' and 'Extracellular'), mitochondrion and chloroplast (Table 1a). Other subcellular localizations have been excluded because too few (less than 20) non-redundant representatives have been annotated so far.

Available predictors have been trained with data sets up to the release 41 of SWISS-PROT. For sake of comparison, we reduced our training sets extracting only the non-redundant proteins contained in that release. The remaining proteins have been used as independent test sets (Table 1b). Since the number of proteins in the plant test set is small, we don't report the results. All the data sets are available at www.biocomp.unibo.it/bacello

For predicting the localization of proteins in whole genomes, we downloaded the protein sequences from the EnsEMBL web site (www.ensembl.org, Hubbard *et al*., 2005). We considered the releases NCBI 35 for *Homo sapiens*, NCBI m34 for *Mus musculus*, WS 140 for *Caenorhabditis elegans* and SGD 1 for *Saccharomyces cerevisiae*. The realease TAIR6 of the *Arabidopsis thaliana* has been downloaded from the TAIR web site (http://www.arabidopsis.org/, Rhee *et al*., 2003).

Two more data sets containing experimental annotated data are used: a data set of 2618 globular proteins deriving from the Yeast GFP fusion databases localized in cytoplasm, nucleus or mitochondria (http://yeastgfp.ucsf.edu, Huh *et al*., 2003) and a set of 499 globular proteins localized in the *Arabidopsis thaliana* chloroplast downloaded from the Plastid Protein Database (http://www.plprot.ethz.ch, Kleffmann *et al*., 2004)

2.2 BaCelLo architecture

Support Vector Machines (SVM) were first introduced by Cortes and Vapnik (1995) and are now broadly used in protein classification tasks. SVMs are able to discriminate two classes of examples by creating a hyperplane that optimally separates them with the best possible margin. Typically the hyperplane is built in a h-dimensional space H in which the examples are mapped by means of feature vectors, that result from the input vectors upon a transformation induced by a kernel function. We used the SVM-light package, version 6, freely available at http://svmlight.joachims.org. We adopted the Radial Basis Function (RBF) kernel since it gives the best performances (data not shown). All the parameters were set as default, except for Gamma and C, which were varied to get the best results.

Our predictor is composed of four support vector machines (SVM) arranged in a decision tree. Each node of the tree is a binary SVM. Different tree architectures were implemented and the most efficient were chosen. The architectures of the trees are shown in Figure 1 and are the same for animals

Table 1a. Number of proteins in the three kingdom specific datasets derived from SWISS-PROT 48

	Plants	Animals	Fungi
Nucleus	121	1166	711
Cytoplasm	58	439	211
Extracellular	41	804	88
Mitochondria	67	188	188
Chloroplast	204		
Total	491	2597	1198

Table 1b. Number of proteins in the training sets extracted from SWISS-PROT 41 and in the testing sets extracted from subsequent releases up to SWISS-PROT 48

	Animals		Fungi	
	Train	Test	Train	Test
Nucleus	803	363	589	122
Cytoplasm	302	137	181	30
Extracellular	632	172	72	16
Mitochondria	153	35	177	11
Total	1890	707	1039	179

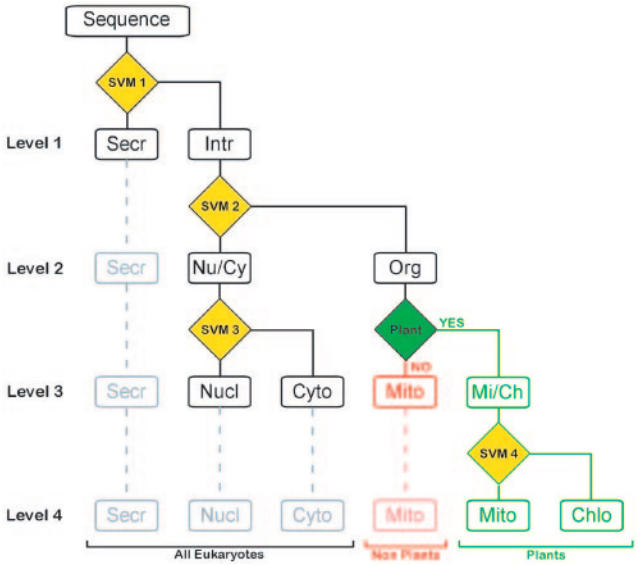

Fig. 1. Architecture of the BaCelLo's decision tree. Abbreviations: Secr: Secretory Pathway, Intr: Intracellular, Org: Organelles, Nu and Nucl: Nucleus, Cy and Cyto: Cytoplasm, Mi and Mito: Mitochondion, Ch and Chlo: Chloroplast.

and fungi, while the plant tree contains an additional node for separating chloroplastic and mitochondrial proteins. Different levels of prediction can be distinguished, each discriminating a different number of classes. For sake of comparison with the other methods we will always consider the level 2, discriminating three classes (Secretory pathway, Nuclear/Cytoplasmic and Organellar) and the level 4, discriminating four and five classes for non plants and plants, respectively.

Table 2. Input vector definition and RBF kernel parameters for the SVMs

SVM	Whole protein frequency	N-ter frequency	C-ter frequency	γ	C	Input vector length
1	+	+	–	3	6	160
2	+	+	+	3	6	280
3	+	–	–	150	2	40
4	+	–	–	150	2	40

SVMs are numbered as in Figure 1.

We used different information as input for SVMs at each node of the tree, depending on the type of discrimination to be performed. Depending on the SVM, we considered the compositions of the whole sequence, of the N-terminal and C-terminal portions. In all cases, both the sequence composition and the sequence profile composition were taken into account. Sequence profiles were obtained aligning with BLAST each sequence with the eukaryotic sequences released in the version 48 of SWISS-PROT. A threshold for the E-value equal to 10^{-4} was used. From the alignment a sequence profile is derived by counting the frequency of each residue in the aligned sequences in each position of the query sequence. The sequence profile composition for a given portion of the protein is obtained by summing up, over all the considered positions, the contributions of each one of the 20 residues. This procedure gives a 20-valued vector that is then normalized. Summing up, three different types of information were considered:

(1) the whole sequence composition, encoded with a 40-valued vector containing both the raw sequence composition (20 components) and the sequence profile composition (20 components);

(2) the N-terminus composition, encoded with a 120-valued vector, containing both the sequence and the profile compositions for three N-terminal portions, formed by the first 20, 40 and 60 residues, respectively;

(3) the C-terminus composition, encoded with a 120-valued vector, containing both the sequence and the profile compositions for three C-terminal portions, formed by the last 20, 50 and 100 residues, respectively;

Different input codes (including a thorough space search for the best input window lengths) have been tried for each node and Table 2 reports the best performing ones, together with the optimal SVM parameters Gamma and C that were selected.

2.3 Evaluation of the performances

We used different accuracy indexes that were computed starting from the confusion matrix $\boldsymbol{Z}$ in which any element z_{ij}, counts the number of examples belonging to the class i and predicted in the class j. First of all, for each class we computed the coverage (Cov) that is the percent of correctly predicted proteins over the total number of proteins belonging to the class. Defining the number of proteins of ith class:

$$x_i = \sum_j z_{ij} \tag{1}$$

coverage can be computed as:

$$Cov(i) = \frac{z_{ii}}{x_i} \cdot 100 \tag{2}$$

The other standard index for the evaluation is the accuracy (*Acc*) that measures the probability of correct prediction for a class:

$$Acc(i) = \frac{z_{ii}}{\sum_j z_{ji}} \cdot 100 \quad (3)$$

When evaluated on very unbalanced databases, the accuracy tends to be low for those classes containing a small number of sequences, since even a very low error rate in such a class can lead to a great number of false positive in the big complementary classes, increasing the denominator in Eq. 3. For these reasons we introduced the Normalized Accuracy (*nAcc*), in which any term of Eq. 3 is divided by the abundance of the respective class in the data set:

$$nAcc(i) = \frac{z_{ii}/x_i}{\sum_j z_{ji}/x_j} \cdot 100 \quad (4)$$

To define a global predictive performance for each class, we used the geometric average (*GAv*) between coverage and normalized accuracy:

$$GAv(i) = \sqrt{Cov(i) \cdot nAcc(i)} \quad (5)$$

In order to evaluate the global performance on all the classes, different parameters are adopted.

Routinely the overall accuracy is computed, defined as the number of correct predictions over the total number of proteins:

$$Q = \frac{\sum_i z_{ii}}{N} \cdot 100 \quad (6)$$

where N is the total number of proteins.

In unbalanced data sets, this parameter is biased towards the performance of the most abundant classes. We introduced the normalized overall accuracy, that counts the number of correct predictions assuming the equiprobability for each class:

$$nQ = \frac{\sum_i z_{ii}/x_i}{K} \cdot 100 \quad (7)$$

where K is number of the classes.

We also used the Generalized Correlation (GC) as proposed by Baldi *et al.* (2000) to analyze the multiclass accuracy. Defining the number of proteins predicted in the i^{th} class:

$$y_i = \sum_j z_{ji} \quad (8)$$

and the matrix:

$$e_{ij} = \frac{x_i y_j}{N} \quad (9)$$

the generalized correlation is computed as

$$GC = \sqrt{\frac{\sum_{ij} \frac{(z_{ij} - e_{ij})^2}{e_{ij}}}{N(K-1)}} \quad (10)$$

It is worth noticing that the generalized correlation does not make use of explicit normalization and can be considered independent of the normalized overall accuracy as defined in Eq. 7.

2.4 Balancing procedure

For all the three datasets the number of sequences for the different classes is highly uneven (Table 1) and the SVMs at each stage discriminate between two classes that are not equally represented in the training set. In the case of mitochondrial versus nuclear/cytoplasmic compartments, for example, the disproportion of the number of sequences is about 8. Under such condition, SVMs tend to over-predict the most abundant class and this can seriously affect the prediction of the under-represented classes (Wang *et al.*, 2004). To solve this problem we adopted the following procedure. For each kingdom, the data set was split in ten subsets. Eight sets are used to train the binary SVM classificators. Each one of the SVMs finds a (h-1)-plane in the h-dimensional hyperspace H of the features; for any point in the feature space a distance from the separating plane was defined. A conventional sign was computed in order to determine in which side the vector point lies with respect to the plane. Routinely a threshold of this 'signed distance' equal to zero is considered for separating the two classes, but a bias can be added to shift the hyperplane (Cortes and Vapnik, 1995). We adopted this possibility to overcome the problem of unbalanced data. The basic idea is to shift the plane in the direction that favors the classification for the less abundant class. Thus a threshold on the 'signed distance' is evaluated on a validation set. The goal is to minimize the unbalance of the predictive performances between the two classes and this can be obtained searching for the threshold on the signed distance that minimizes:

$$| GAv_{k(+)} - GAv_{k(-)} | \quad (11)$$

where *GAv* is the geometric average of the normalized accuracy and the coverage, as defined in Eq. 5. The threshold that minimizes Eq. 11 typically maximizes also the sum of the two geometric averages and leads to the optimal performance. The final performance is then evaluated on the remaining set, called the test set, that is not used to set the SVM parameter nor to pick the optimal threshold. The procedure is repeated ten times, in order to predict each one of the split sets with methods whose parameters have been computed using all the other sets.

Table 3. Performances of BLAST assignment in the RH dataset

Classes	no E-value threshold				E-value < 10^{-3}			
	Cov	*Acc*	*nAcc*	*GAv*	*Cov*	*Acc*	*nAcc*	*GAv*
Nucleus	93.3	98.0	82.4	87.7	98.5	91.9	95.4	96.9
Cytoplasm	86.1	90.1	74.7	80.2	93.2	82.5	81.6	87.2
Secretory Pathway	91.0	99.0	95.0	93.0	97.3	89.7	99.6	98.4
Mitochondria	69.2	90.6	91.4	79.5	81.5	82.2	96.0	88.5
Q	87.8				94.8			
nQ	84.9				92.6			
GC	0.81				0.92			
Not Assigned (%)	0.0				14.1			

Not assigned proteins are sequences for which no homologous (under the E-value threshold) can be found in the RH dataset. Values are normalized on the number of assigned sequences. Abbreviations: Cov: Coverage, Acc: Accuracy, nAcc: normalized Accuracy, GAv: Geometric Average, Q: number of proteines correctly predicted, nQ: normalized number of proteins correctly predicted, GC: Generalized Correlation: see the Materials and Methods section for their definition.

3 RESULTS AND DISCUSSION

3.1 Necessity of a non-redundant training set

Many available predictors with the exception of LocTree (Nair and Rost, 2005) were implemented using redundant training sets. Sequences sharing up to 95% identity were routinely selected. The rationale for this is the fact that little differences in the sequences can lead to different subcellular location. It is therefore important to quantify to which extent sequence identity affects subcellular location and to which extent redundancy in the training set leads to a tool with poor generalization capability when predicting sequences scarcely related to those considered for training. The most largely adopted of these data sets was firstly released by Reinhardt and Hubbard (1998) (RH Dataset) and contains 2427 eukaryotic proteins divided into four subcellular classes: extracellular (325 sequences), cytoplasm (684), nucleus (1097) and mitochondria (321).

Table 4. 10-fold crossvalidation performances of BaCelLo on three kingdom-specific datasets

Level	Classes	Plants							Animals							Fungi						
		Cov	*nAcc*	*Acc*	*GAv*	*nQ*	*Q*	*GC*	*Cov*	*nAcc*	*Acc*	*GAv*	*nQ*	*Q*	*GC*	*Cov*	*nAcc*	*Acc*	*GAv*	*nQ*	*Q*	*GC*
1	Secr	85.4	95.3	64.8	90.2	90.6	94.9	0.72	90.8	95.6	90.7	93.2	93.3	94.3	0.87	94.3	97.7	76.9	96.0	96.0	97.5	0.84
	Intr	95.8	86.7	98.6	91.1				95.8	91.2	95.8	93.5				97.7	94.5	99.5	96.1			
2	Nucl/Cyto	80.4	76.0	80.4	78.2	84.4	84.7	0.73	92.9	82.9	94.6	87.8	86.6	91.0	0.78	91.2	83.2	96.7	87.1	89.0	89.9	0.78
	Secr	85.4	89.8	64.8	87.6				90.8	85.0	90.7	87.9				94.3	92.8	76.9	93.5			
	Mito/Chlo	87.5	88.2	91.9	87.8				76.1	93.6	66.2	84.4				81.4	91.8	69.5	86.4			
3	Nucl	71.9	69.1	75.7	70.5	74.1	79.2	0.65	64.8	67.8	84.9	66.3	74.2	73.8	0.67	67.1	65.7	87.0	66.4	75.8	70.1	0.66
	Cyto	51.7	65.5	46.9	58.2				65.3	60.3	41.4	62.8				60.2	62.3	39.4	61.2			
	Secr	85.4	81.3	64.8	83.3				90.8	83.1	90.7	86.9				94.3	90.6	76.9	92.4			
	Mito/Chlo	87.5	78.1	91.9	82.7				76.1	87.4	66.2	81.6				81.4	83.8	69.5	82.6			
4	Nucl	71.9	66.8	75.7	69.3	66.6	68.2	0.59														
	Cyto	51.7	61.6	46.9	56.4																	
	Secr	85.4	80.0	64.8	82.7																	
	Mito	50.7	77.3	54.0	62.6																	
	Chlo	73.0	53.6	76.4	62.6																	

Abbreviations: See caption of Table 3 Secr: Secretory Pathway, Intr: Intracellular, Nucl: Nucleus, Cyto: Cytoplasm, Mito: Mitochondria, Chlo: Chloroplast.

We predicted the localization for every protein in the RH dataset with a BLAST search on the same dataset. The results of the assignment based on the closest non identical homologue are shown on Table 3. Setting a threshold for the E-value equal to 10^{-3} (corresponding approximately to a local sequence identity higher than 25%) 86% of the RH sequences are similar to at least another sequence of the set. A simple procedure based on transfer annotation is then able to correctly assign 94% of the proteins (1974 chains). When no E-value threshold is considered, and the annotation is transferred from the closest homologous regardless of the sequence identity level, the accuracy is still as high as 88%. Notably, this performance is similar to that achieved by methods trained on the same RH dataset: LOCSVMpsi (Xie *et al*., 2005), ESLpred (Bhasin and Raghava, 2005) and SubLoc (Hua and Sun, 2001) that reach overall accuracies (*Q*) as high as 90%, 88% and 79%, respectively.

The goal of a good predictor is to assign a subcellular localization especially when no homology is detectable and for this a non-redundant data set needs to be selected.

3.2 BaCelLo performances

Three non-redundant sequence sets were selected, for animals, fungi and plants, respectively. In each set the sequences share less than 30% identity. We generated three eukaryotic datasets in order to take into account the differences in the subcellular localization mechanism between evolutionary distant kingdoms.

BaCelLo is a system of SVMs organized in a tree structure, which exploits the information from the sequence composition and from the sequence profile composition. The input of the SVMs considers the whole sequence and different portion of the N- and C- termini, which are likely to contain localization signals. However it does not make explicit search for localization signal or implement annotation transfer by homology. Different tree architectures have been tried and the best performing one is adopted. The performance of BaCelLo, computed on the test sets with a rigorous 10-fold cross validation procedure, is shown in Table 4. The decision tree structure of our system allows to predict the subcellular location at different stages with different accuracy. In all the kingdoms extracellular/secretory proteins are well discriminated from intracellular proteins. When the normalized overall accuracy is considered (*nQ*), BaCelLo at the first level of the tree (Fig. 1) correctly discriminates 96%, 93% and 91% of the proteins from fungi, animals and plants, respectively. This level of discrimination is achieved using information from the whole sequence and from the N-terminal portions, where signal peptides are supposed to be. Adding one more level, BaCelLo discriminates among 3 classes: extracellular/secretory, nuclear/cytoplasmic and mitochondrial/chloroplastic. The normalized overall accuracy ranges from 89% in fungi to 84% in plants. At this stage the prediction exploits the information extracted from the whole sequence and from both the N- and the C-terminal portions. In the next step of the decisional tree nuclear and cytoplasmic proteins are discriminated. Since up to date no conserved nor general localization signal is known, this step is done using only information about the whole protein. The performance on the four classes ranges between 75% (fungi) and 74% (plants). For plant proteins an additional step is introduced to separate mitochondrial from chloroplastic proteins. Trying different input coding we verified that there is no advantage in using information from both termini portions (data not shown). For this reason the step exploits only the information from the whole protein sequence. The overall accuracy achieved for plant proteins is then 66% on 5 classes.

The data in Table 4 show that the best performance is reached in discriminating extracellular proteins. Prediction at level 2, where proteins of organelles are discriminated, is very good, while the distinction between nuclear and cytoplasmic proteins (level 3) and between chloroplastic and mitochondrial proteins (level 4) is more problematic, leading to a quite poor coverage for cytoplasmic and mitochondrial proteins and a quite poor accuracy for chloroplastic proteins.

It is evident that the accuracy value of each class is strongly influenced by the dimension of the class, being remarkably higher in most abundant classes, namely nuclear in non-plants or nuclear

Table 5. Comparative performances on the test set

3 classes		Fungi											Animals										
Method		**a**	b	c	d	e	f	g	h	i	j	k	**a**	b	c	d	e	f	g	h	i	j	k
Nucl/Cyto	*Cov*	**88.8**	87.5	88.1	86.8	97.4	98.0	94.1	96.1	87.5	90.8	*85.5*	**93.2**	88.8	92.6	84.0	92.6	95.2	96.6	97.6	89.4	90.6	*88.6*
	nAcc	**93.4**	75.9	56.6	54.0	48.1	63.8	55.3	59.9	65.6	76.5	*90.4*	**71.7**	67.9	60.7	47.5	44.4	54.1	60.7	62.9	70.8	66.6	*87.5*
	GAv	**91.1**	81.5	70.6	68.5	68.4	79.1	72.1	75.9	75.8	83.3	*87.9*	**81.7**	77.6	75.0	63.2	64.1	71.8	76.6	78.4	79.6	77.7	*88.0*
Secr	*Cov*	**93.8**	81.3	56.3	50.0	31.3	62.5	87.5	81.3	87.5	81.3	*43.8*	**85.5**	84.9	56.4	47.7	20.3	53.5	76.2	84.9	85.5	82.0	*62.8*
	nAcc	**99.3**	69.4	100	93.8	96.0	100	96.4	96.9	95.7	95.4	*100*	**90.7**	85.8	93.9	80.6	83.4	92.4	94.2	95.9	94.6	91.4	*100*
	GAv	**96.5**	75.1	75.0	68.5	54.8	79.1	91.8	88.8	91.5	88.1	*66.2*	**88.1**	85.3	72.8	62.0	41.1	70.3	84.7	90.2	89.9	86.6	*79.2*
Mito	*Cov*	**100**	63.6	63.6	63.6	63.6	81.8	36.4	54.5	66.7	90.9	*81.8*	**68.6**	60.0	54.3	48.6	54.3	58.3	57.1	54.5	74.3	65.7	*71.4*
	nAcc	**90.5**	94.2	94.1	74.0	98.0	97.6	93.3	97.6	88.6	94.5	*95.4*	**90.4**	85.1	84.7	75.8	81.2	88.1	95.4	96.8	89.0	88.5	*91.0*
	GAv	**95.1**	77.4	77.4	68.6	78.9	89.4	58.3	72.9	76.9	92.7	*88.3*	**78.7**	71.5	67.8	60.7	66.4	71.7	73.8	72.6	81.3	76.3	*80.6*
nQ		**94.2**	77.5	69.3	66.8	64.1	80.8	77.7	77.3	80.6	87.6	*70.4*	**82.4**	77.9	67.8	60.1	55.7	69.0	76.6	79.0	83.0	79.4	*74.3*
GC		**0.79**	0.57	0.53	0.45	0.58	0.77	0.62	0.70	0.62	0.70	*0.59*	**0.71**	0.64	0.48	0.38	0.36	0.57	0.70	0.75	0.70	0.67	*0.60*
Assigned to other locations		**-**	-	7.9	-		-	-	-	-	-	*14.5*	**-**	-	4.1	-	-	-	-	-	-	-	*14.0*
4 classes		Fungi											Animals										
Nucl	*Cov*	**66.4**	66.4	71.1	70.5	84.4	88.5					*62.3*	**66.1**	62.2	70.2	67.8	79.1	80.2					*73.3*
	nAcc	**71.3**	66.9	44.2	38.4	37.5	51.0					*63.5*	**56.4**	49.5	43.0	37.2	35.8	38.7					*64.2*
	GAv	**68.8**	66.6	56.1	52.0	56.3	67.2					*62.9*	**61.1**	55.5	54.9	50.2	53.2	55.7					*68.6*
Cyto	*Cov*	**56.7**	46.7	36.7	23.3	23.3	30.0					*56.7*	**54.0**	38.2	40.9	21.9	28.5	29.2					*46.0*
	nAcc	**65.4**	50.3	46.2	32.7	35.0	40.1					*70.4*	**50.7**	42.0	48.5	28.6	36.1	44.9					*63.1*
	GAv	**60.9**	48.5	41.2	27.6	28.6	34.7					*63.2*	**52.3**	40.1	44.5	25.0	32.1	36.2					*53.9*
Secr	*Cov*	**93.8**	81.3	56.3	50.0	31.3	62.5					*43.8*	**85.5**	84.9	56.4	47.7	20.3	53.5					*62.8*
	nAcc	**99.1**	61.7	100	92.4	82.4	100					*100*	**88.4**	80.0	93.4	67.1	76.3	90.1					*100*
	GAv	**96.4**	70.8	75.0	68.0	50.8	79.1					*66.2*	**86.9**	82.4	72.6	56.6	39.4	69.4					*79.2*
Mito	*Cov*	**100**	63.6	63.6	63.6	63.6	81.8					*81.8*	**68.6**	60.0	54.3	48.6	54.3	58.3					*71.4*
	nAcc	**79.6**	83.6	86.5	70.0	90.5	91.6					*89.2*	**86.2**	76.8	80.5	69.3	73.6	85.4					*86.7*
	GAv	**89.2**	72.9	74.2	66.7	75.9	86.6					*85.4*	**76.9**	67.9	66.1	58.0	63.2	70.6					*78.7*
nQ		**79.2**	64.3	56.9	51.9	50.7	65.7					*61.1*	**68.5**	61.3	55.5	46.5	45.5	55.3					*63.4*
GC		**0.68**	0.50	0.47	0.39	0.49	0.65					*0.54*	**0.60**	0.53	0.43	0.32	0.31	0.48					*0.54*
Assigned to other locations		**-**	-	7.9	-	-	-					*14.5*	**-**	-	4.1	-	-	-					*14.0*

a: BaCelLo, **b:** Loctree (Nair and Rost, 2005), **c**: Psort II (Nakai and Horton, 1999), **d**: SubLoc (Hua and Sun, 2001), **e**: ESLpred (Bhasin and Raghava, 2004), **f**: LOCSVMpsi (Xie *et al*., 2005), **g**: SLP-local (Matsuda *et al*., 2005), **h**: Protein Prowler (Boden and Hawkins, 2005), **i**: TARGETp (Emanuelsson *et al*., 2000), **j**: PredoTar (Small *et al*., 2004) , **k**: pTARGET (Guda and Subramaniam, 2005).
pTARGET results are highlighted in italic since it used for training proteins released after the release 41 of SWISS-PROT; Abbreviations: see caption for Table 3.

Table 6. Prediction of protein localization for whole genomes

	H.sapiens	*M.musculus*	*C.elegans*	*S.cerevisiae*	*A.thaliana*
Nucl	8725 *(26%)*	6811 *(24%)*	5878 *(23%)*	2078 *(32%)*	7050 *(25%)*
Cyto	10399 *(31%)*	10909 *(31%)*	6674 *(26%)*	1611 *(25%)*	6033 *(20%)*
Secr	4960 *(15%)*	5417 *(15%)*	4767 *(19%)*	227 *(3%)*	3001 *(10%)*
Mito	2452 *(7%)*	2793 *(8%)*	1516 *(6%)*	971 *(15%)*	963 *(3%)*
Chlo					4875 *(16%)*
Memb	7017 *(21%)*	7610 *(22%)*	6879 *(25%)*	1657 *(25%)*	8078 *(26%)*
Total	33553	35340	25714	6544	30600

For each species the number of proteins predicted in each localization and the Percentage with respect to the total are shown. Abbreviations: Secr: Secretory Pathway, Nucl: Nucleus, Cyto: Cytoplasm, Mito: Mitochondria, Chlo: Chloroplast, Memb: Membrane.

and chloroplastic in plants. This is due to the fact that also a great rate of false positives on the most abundant classes gives a low number of false positive on the other classes and then scarcely affects the accuracy. Since the proportion of the classes in the data set does not reflect any reliable a priori hypothesis, a more meaningful evaluation can be carried out considering the normalized parameters, defined in the Material and Methods section. It is worth noticing that the balancing procedure leads to performances in which the coverage and the normalized accuracy are similar for each class, except in the case of chloroplasts and mitochondria in plants, where the latter tend to be under-predicted. This may be due to the under representation of mitochondrial proteins from plants (67 examples of non-redundant mitochondrial proteins are known in plants while 188 are known in both animals and fungi).

The performances of the three kingdom-specific predictors are quite similar, but it is worth to noticing that merging the fungi and the animal proteins, similarly to what the other predictors do, leads to a poorer performance, in particular for fungi (data not shown).

3.3 Comparison with the other methods

The performance of BaCelLo has been compared to those of the best publicly available methods for the prediction of the subcellular localization. Some of them discriminate among three classes, namely TARGETp (Emanuelsson *et al*., 2000), ProteinProwler (Boden and Hawkins, 2005), SLP-local (Matsuda *et al*., 2005) and Predotar (Small *et al*., 2004); others discriminate among four classes in animals and five classes in plants and are Loctree, SubLoc, ESLpred and LOCSVMpsi. pTARGET and Psort II (Nakai and Horton, 1999) discriminate more classes than BaCelLo does, however we did not consider these classes since very few redundant examples are known. All the considered predictors, but pTARGET, have been trained on a dataset at most derived from SWISS-PROT, rel. 41. Then, in order to compare the performances, we retrained BacCelLo using only the subset of training sequences that were yet included in the release 41 of SWISS-PROT. The test and the comparison has been performed on the remaining sequences, up to release 48, that, by construction, are less than 30% identical to those of the training. It is important to note that some 30% proteins of this test set share identity with proteins included in SWISS-PROT 41, so that methods different from BaCelLo can still have some homology with the training set. Moreover, when comparing with pTARGET it has to be kept in mind that it was developed using proteins derived from release 46 of SWISS-PROT and then that it has been successively updated.

Predictions were run with default options, except for LOCSVMpsi, for which the four class classification option was selected. For the predictors that consider more than four classes, proteins predicted in classes not considered by BaCelLo are considered as badly assigned.

On the fungi dataset (Table 5), BaCelLo outperforms the other methods by at least 7% in terms of normalized overall accuracy, for the three classes predictions and 14% for the four classes ones. It performs remarkably better in discriminating secreted and mitochondrial proteins, even when compared with TARGETp, Proteins Prowler and Predotar that are explicitly designed to recognize signal and target peptides for these localizations. Furthermore we achieve the best average prediction (*GAv*) in each class and, outperforming other methods, a good balancing between coverage (*Cov*) and normalized accuracy (*nAcc*). As a general consideration, most of the methods achieve on this test a worse performance than that reported in the original papers, corroborating the notion that the redundancy of their data sets affects the generalization performances. Similar considerations are valid for the animal test set (Table 5). In this case the improvement with respect to other methods is about 5% on four classes. All the predictors perform worse on the animal set than on the fungi set. However the large improvement of the BaCelLo performances in predicting fungal proteins confirms the advantage in implementing a fungal-specific predictor.

3.4 Genome predictions

We adopted BaCelLo for high-throughput prediction of protein subcellular localization. Five proteomes have been tested in order to estimate the composition of protein localization. We predicted the localization of proteins from *H. sapiens*, *M. musculus* and *C. elegans* for animals, *S. cerevisiae* for fungi and *A. thaliana* for plants. Membrane proteins predicted with SpepLip (Fariselli *et al*., 2003) and ENSEMBLE (Martelli *et al*., 2003) were excluded from the set predicted with BaCelLo. The results of this large scale analysis are reported in Table 6, where both the number and the frequency of proteins predicted in each class are listed. In all the examined proteomes the sum of nuclear and cytoplasmic proteins accounts for about 50-60% of all proteins. Protein localization composition for Human and Mouse are very similar, as

expected, while *C. elegans* contains about the same number of secreted and membrane proteins as the other animals, but significantly fewer proteins in the nucleus and cytoplasm. *S. cerevisiae* is a unicellular organism, not endowed with an extracellular matrix nor communicating with other cells. Interestingly and accordingly, only 3% of its proteome is predicted as secreted. Concerning plants, up to 19% of the *A. thaliana* proteins are directed to organelles and more than 80% of them are directed to chloroplasts. The results are available at http://www.biocomp.unibo.it/bacello.

3.5 Comparison with high-throughput experimental data

As a proof of the reliability of the prediction of BaCelLo we compared predictions with data obtained with high-throughput methods in *A.thaliana* and in yeast. The PLPROT data base contains 499 annotated chloroplast proteins from *A.thaliana* and we correctly assign 53% of them, an amount similar to that reported for LocTree. From the Yeast GFP fusion database we extracted 2618 sequences experimentally annotated: 483 'mitochondrial', 496 'nuclear', 818 'cytoplasmic' and 821 'nuclear and cytoplasmic'. The performance on mitochondrial proteins is 87%. The rate of correct prediction for nuclear and cytoplasmic proteins reaches about 50%. Nevertheless considering 'nuclear', 'cytoplasmic' and 'nuclear and cytoplasmic' proteins together, level 2 of our predictor correctly assigns 88% of proteins.

4 CONCLUSIONS

BaCelLo is a new method for predicting the subcellular localization of a protein sequence from animals, fungi or plants. Three kingdom-specific sets of parameters have been inferred from non-redundant data sets of annotated proteins. This is at the basis of the implementation of the first *de novo* predictor specific for animals and fungi. The reduction of the redundancy of the training sets guarantees the generalization capability of BaCelLo. We prove that predictors trained on very redundant data sets don't perform better than a simple annotation transfer based on a BLAST search.

The key feature of BaCelLo is the procedure to balance the prediction scoring indexes, overcoming the biases in the dataset composition. BaCelLo outperforms other predictors on a test set of proteins that don't share identity with sequences used for training. Furthermore BaCelLo can be easily used for large scale analysis of whole genomes to produce an estimate and annotation of the protein content in each subcellular compartment.

BaCelLo is available at the site http://www.biocomp.unibo.it/bacello.

ACKNOWLEDGEMENTS

RC acknowledges the receipt of the following grants: PNR 2001-2003 (FIRB art.8) for a project on Bioinformatics for Genomics and Proteomics, a FIRB 2003 LIBI—International Laboratory of Bioinformatics and the support to the Bologna node of the Biosapiens Network of Excellence project within the European Union's VI Framework Programme (contract number LSHG-CT-2003-503265). PF acknowledges MIUR for a grant on Proteases. AP and PLM are supported by a FIRB 2003-LIBI grant.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M., Martin,M.J., Natale,D.A., O'Donovan,C., Redaschi,N. and Yeh,L.S. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A.F. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Bendtsen,J.D., Jensen,L.J., Blom,N., Von Heijne,G. and Brunak,S. (2004) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.*, **17**, 349–356.

Bhasin,M. and Raghava,G.P. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.*, **32**, W414–419.

Boden,M. and Hawkins,J. (2005) Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics*, **21**, 2279–2286.

Cortes,C. and Vapnik,V. (1995) Support vector networks. *Mach. Learn.*, **20**, 273–293.

Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal residue sequence. *J. Mol. Biol.*, **300**, 1005–1016.

Fariselli,P., Finocchiaro,G. and Casadio,R. (2003) SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics*, **19**, 2498–2499.

Fried,H. and Kutay,U. (2003) Nucleocytoplasmic transport: taking an inventory. *Cell. Mol. Life Sci.*, **60**, 1659–1688.

Gonsalvez,G.B., Urbinati,C.R. and Long,R.M. (2005) RNA localization in yeast: moving towards a mechanism. *Biol. Cell.*, **97**, 75–86.

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

Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T., Down,T., Durbin,R., Fernandez-Suarez,X.M., Gilbert,J., Hammond,M., Herrero,J., Hotz,H., Howe,K., Iyer,V., Jekosch,K., Kahari,A., Kasprzyk,A., Keefe,D., Keenan,S., Kokocinsci,F., London,D., Longden,I., McVicker,G., Melsopp,C., Meidl,P., Potter,S., Proctor,G., Rae,M., Rios,D., Schuster,M., Searle,S., Severin,J., Slater,G., Smedley,D., Smith,J., Spooner,W., Stabenau,A., Stalker,J., Storey,R., Trevanion,S., Ureta-Vidal,A., Vogel,J., White,S., Woodwark,C. and Birney,E. Ensembl (2005). *Nucleic Acids Res.*, **33**, D447–D453.

Huh,W.K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–91.

Izeta,A., Malcomber,S., O'Rourke,D., Hodgkin,J. and O'Hare,P. (2003) A C-terminal targeting signal controls differential compartmentalisation of Caenorhabditis elegans host cell factor (HCF) to the nucleus or mitochondria. *Eur. J. Cell. Biol.*, **82**, 495–504.

Kleffmann,T., Russenberger,D., von Zychlinski,A., Christopher,W., Sjolander,K., Gruissem,W. and Baginsky,S. (2004) The Arabidopsis thaliana chloroplast proteome reveals pathway abundance and novel protein functions. *Curr. Biol.*, **14**, 354–362.

Lee,C.M., Sedman,J., Neupert,W. and Stuart,R.A. (1999) The DNA helicase, Hmi1p, is transported into mitochondria by a C-terminal cleavable targeting signal. *J. Biol. Chem.*, **274**, 20937–20942.

Marcotte,E.M., Xenarios,I., van Der Bliek,A.M. and Eisenberg,D. (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, **97**, 12115–12120.

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

Matsuda,S., Vert,J.P., Saigo,H., Ueda,N., Toh,H. and Akutsu,T. (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Science*, **14**, 2804–2813.

Nair,R. and Rost,B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.*, **348**, 85–100.

Nakai,K. and Horton,P. (1999) PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–35.

Nickel,W. (2003) The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes. *Eur. J. Biochem.*, **270**, 2109–2119.

Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*., **10**, 1–6.

Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res*., **26**, 2230–2236.

Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M., Miller,N., Mueller,L.A., Mundodi,S., Reiser,L., Tacklind,J., Weems,D.C., Wu,Y., Xu,I., Yoo,D., Yoon,J. and Zhang,P. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*., **31**, 224–228.

Scott,M.S., Thomas,D.Y. and Hallett,M.T. (2004) Predicting subcellular localization via protein motif co-occurrence. *Genome Res*., **14**, 1957–1966.

Small,I., Peeters,N., Legeai,F. and Lurin,C. (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **6**, 1581–1590.

Wang,M., Yang,J., Liu,G.P., Xu,Z.J. and Chou,K.C. (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo-residue composition. *Protein Eng. Des. Sel*., **17**, 509–516.

Xie,D., Li,A., Wang,M., Fan,Z. and Feng,H. (2005) LOCSVMPSI: A web server for subcellular localization of eukaryotic proteins. *Nucleic Acids Res*., **33**, W105–W110.

Yamada,H., Chounan,R., Higashi,Y., Kurihara,N. and Kido,H. (2004) Mitochondrial targeting sequence of the influenza A virus PB1-F2 protein and its function in mitochondria. *EBS Lett*., **578**, 331–336.