The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation

Rita Casadio, Pier Luigi Martelli and Andrea Pierleoni

Advance Access publication date 18 February 2008

Abstract

Automated sequence annotation is a major goal of post-genomic era with hundreds of genomes in the databases, from both prokaryotes and eukaryotes. While the number of fully sequenced chromosomes from microbial organisms exponentially increased in the last decade above 600, presently we know the whole DNA content of only 25 eukaryotic organisms, including *Homo sapiens*. However, the process of genome annotation is far from being completed. This is particularly relevant in eukaryotes, whose cells contain several subcellular compartments, or organelles, enclosed by membranes, where different relevant functions are performed. Translocation across the membrane into the organelles is a highly regulated and complex cellular process. Indeed different proteins and/or protein isoforms, originated from genes by alternative splicing, may be conveyed to different cell compartments, depending on their specific role in the cell. During recent years the prediction of subcellular localization (SL) by computational means has been an active research area. Several methods are presently available based on different notions and addressing different aspects of SL. This review provides a short overview of the most well performing methods described in the literature, highlighting their predictive capabilities and different applications.

Keywords: subcellular localization; functional genomics; machine learning; proteome annotation; sequence-based predictions

INTRODUCTION

The so-called post-genomic era for a specific organism starts when its genome is released [1]. The operations that are routinely necessary are: (i) a more accurate annotation process; (ii) a better understanding of the proteome and its function; (iii) a search for both basic and regulation factors of the transcription machinery to detect gene expression co-regulation. The three different steps may take advantage of both a computational and an experimental phase, during which a close integration of experimental, computational and database search methods can refine the initial genome draft annotation. A major goal of recent bioinformatics developments is indeed to provide automatic methods to relate sequence, structure and functional spaces. Within this framework, tools specifically developed to predict structural and functional features starting from the gene/protein sequence are helping in inferring sequence to structure–function relationship. In eukaryotes, it is known that protein function may also depend on the protein final destination within the cell complex compartmentalization. This in turn may be related to another important still scarcely exploited mechanism, known as ‘alternative splicing’, namely the diverse generation of mature RNAs starting from what is recognized as a ‘gene’ [2–4] and their redirection to different subcellular compartments, depending on the structure/function relationship of the different isoforms. In humans, for example, both the number of proteins and performed functions are much higher than the number of putative genes that
may be deduced from the genome (about 25,000, Ensembl 47, [5–7]). Therefore, when a genome is annotated, also the likelihood of alternative splicing sites and the final destination of the different protein variants are necessary additional features. Presently, due to technical difficulties, experimental data are still limited [8], and both the presence of alternative splicing sites and the subcellular localization (SL) are addressed by means of computational approaches, such as the alignments of cDNAs and genome sequences and/or predictive methods [9]. Recently, the pilot project of the Encyclopaedia of DNA elements (ENCODE) has undertaken a comprehensive analysis of 44 selected regions of the human genome (about 1%, [10]) for identifying all the functional elements. The first reference set of manually annotated splice variants, made available by the GENCODE consortium [11], was also filtered at the protein level by means of different predictive algorithms developed within the European Biosapiens network of excellence. The results corroborated the view that by means of structure/function prediction methods the alternative protein isoforms can be also properly validated, characterized as functionally active proteins and that different isoforms can be addressed to different subcellular compartments, depending on their structural/functional properties [12].

From this and previous observations [12, 13], ‘subcellular localization’ stems, therefore, in eukaryotes as an additional important feature to help the overall process of protein annotation from sequence to function and to be considered/included when automated annotation platforms are implemented. As already reported in literature, prediction of protein SL may also help the annotation process of prokaryotic genomes [14], especially when the secretome needs to be recognized.

In the following, we outline the presently available methods for predicting SL by grouping them depending on their basic characteristics and by benchmarking them on a specific eukaryotic protein reference set. This experiment suggests that the present state of the art still deserves improvement and that more experimental work is necessary to fully exploit the computational capabilities of the different methods.

METHODS BASED ON SEQUENCE SIMILARITY SEARCH

As a putative consequence of molecular evolution, proteins sharing high sequence identity are usually targeted to the same subcellular compartment. This has been described in several computational studies [15–17]. Therefore, with few exceptions, the best way to annotate the SL of a non-annotated sequence is the annotation transfer upon homology search. Although similarity-based annotations are often the most accurate, in some cases they can miss short and/or sparse signals that are able to change the location of a protein without greatly affecting the overall similarity. This is the case when short Nuclear Localization Signals (NLS) and N- or C-terminal signal peptides are present in the sequence. Moreover, for many protein chains, similarity-based annotation cannot be carried out, particularly when poorly studied species are taken into consideration [18].

Most servers for similarity-based annotation of the SL exploit the information contained in well-annotated databases such as SwissProt. Among them LocKey [19] and Proteome Analyst [20] search for both SL annotations and terms reported in the SwissProt entry referring to functions that are strictly correlated to a specific subcellular compartment. Other methods, like SherLoc [21], analyse also the title and the abstract of papers referenced in the SwissProt entry. After the homology search routinely performed with BLAST, significant features are found in the retrieved entries and are then processed with different methods, including Support Vector Machines (SVM) and Bayesian statistics-based ones. Two databases are also available reporting pre-calculated SL annotations of whole proteomes derived by homology search: eSLDB [18] and PA-GOSUB [22]. Both rely on information extracted from the SwissProt database after a BLAST search. When homology search does not give profitable results, eSLDB reports results computed with Machine Learning (ML)-based tools (see below the section ‘Database of SL prediction for whole genome’).

METHODS BASED ON FUNCTIONAL MOTIFS/DOMAINS SEARCH

A less common approach for the annotation of SL relies on searching for sequence motifs and/or domains that are correlated with a determined localization. These methods could in theory be applied also to partial residue sequences; however, they suffer for the same limits of homology-based
searches since they are not able to classify every sequence. Furthermore, domains, families and motifs are difficult to detect when remote sequence homology is involved.

Different methods rely on different family, domain and motif databases. pTARGET [23] searches for PFAM domains related to particular SL, discriminating up to 10 localizations. The predictions are then complemented with ML-based methods when no suitable domain is found in PFAM. PSLT2 [24] is a fungi-specialized predictor that processes by means of a Bayesian network several features, deriving from the search InterPro motifs, signal peptides and transmembrane domains. On similar bases a method was developed by searching for the co-occurrence of SMART domains [25].

The only available evaluation of a motif-based method as compared to similarity-based ones was reported by Jia et al. [26]. Their motif-based method is 1% more accurate than that based on the BLAST search. This only discussion suggests that to date motif/domain-based methods can capture roughly the same information as similarity-based ones. However, the continuous growth of databases of known motifs associated with the different localizations will possibly help the prediction task.

SEQUENCE-BASED METHODS

Great efforts have been done to develop methods able to predict the SL of a protein starting simply from its residue sequence, without relying on the existence of homologous annotated sequences nor on the recognition of specific motifs and domains; they aim at capturing from the single sequence of the protein all those features, if present, to address the protein to a given SL. This is the most challenging task for the annotation in projects of functional genomics. An increasing number of papers describe advances in this field. Apparently, ML-based methods are particularly suited to this aim. ML is an active field in computational science that develops techniques by extrapolating information from known examples and generalizing rules for classifying never seen before examples. ML-based methods are always applicable and are the only ones that can be adopted to predict the SL for proteins without annotated homologues and/or without recognizable domains.

An impressive number of SL predictors have been developed (a non-exhaustive list can be found at www.psort.org) and a wide spectrum of ML techniques have been adopted, ranging from Hidden Markov Models (HMM) to Neural Networks (NN) and SVM. The review from Schneider and Fechner [27] offers an overview of the applications of ML techniques for the prediction of protein SL. In most cases the training examples are derived from the annotations reported in SwissProt or mapped on the Gene Ontology definitions. The predictors are generally designed to predict difficult sequences that do not share significant similarity with annotated proteins. For sake of generalization, the implementation of ML methods, during the so-called training phase, requires particular attention in avoiding redundancy among training and testing datasets. This is necessary to prevent the predictor from biases towards the most represented examples and from overestimating their performances obtained with cross-validation procedures. Unfortunately, most of the available predictors are trained on datasets of sequences sharing up to 90% sequence identity; for this reason, the reported performances hardly describe the real generalization capability of the predictor and their usefulness in addressing ‘real world’ problems. Indeed it was proven that methods trained on redundant datasets reach performances comparable (or even inferior) to those obtained with a simple similarity search [16] since two sequences >30% identical usually share the same subcellular compartment [15].

The most common way to encode a sequence is to use the residue composition by means of 20-valued vectors. Most methods make use of the composition of the whole sequence and/or of portions of the sequence that are likely to carry localization signals, such as the N- and/or the C-termi. Some methods adopt a more sophisticated input, including the dipeptide composition (or even the composition of higher order peptides). This encoding requires a larger input (e.g. a 400-valued vector in the case of dipeptide composition) that can be cut down by means of reduced residue alphabets [28]. However, despite the increased input information, these methods do not outperform methods based on the sole sequence composition, as previous rigorous comparative evaluations have proven [29]. Other methods take advantage of a large spectrum of physicochemical properties derived from the residue sequence and used as input along with the sequence composition. However, the most important improvement was obtained introducing the composition of sequence profiles derived from multiple
sequence alignments [16, 30]. It should be pointed-out that these methods do not make use of the annotations associated with the aligned sequences as described earlier so that they cannot be classified within the homology-based ones.

The encoded representation of the sequence is routinely fed into a ML tool. Among the ML tools, SVMs are the most widely used since they are proven to outperform other techniques also in SL discrimination task [31].

We consider as sequence-based SL predictors also tools that search for explicit localization signals such as signal peptides, transit peptides or NLS. Among them, TargetP [32] and Protein Prowler [33] predict the presence of a signal peptide for secretion or a transit peptide for targeting the protein into an organelle, while PredictNLS [34] recognizes short NLS patterns along the sequence by means of regular expressions. Even being specialized in discriminating a limited number of SL classes, these predictors generally do not outperform more general tools able to distinguish a greater number of SL classes. However, in some cases, the output from the former methods is used as input to the latter: this is the strategy adopted for example by the PSORT series of SL predictors. Most of methods for SL do not explicitly discriminate membrane proteins, since very efficient predictors are available to solve this problem, with very low false positive (FP) and false negative rates [36]. It should be considered that to date no method is able to predict the target membrane of the transmembrane proteins.

**THE TOP SCORING SEQUENCE-BASED METHODS FOR SL PREDICTION**

When different predictors are available, it is difficult to exploit their relative performance, especially considering that they have been implemented through the years with different datasets and that some of the training sets may have contained redundancy with the testing set. Independent and uniform evaluation of predictors is, therefore, needed and different papers were recently addressing this issue.

A general evaluation was performed in 2005 by Klee and Ellis [36], indicating that TARGETp was the best predictor for secreted proteins in eukaryotes out of six methods. A more general comparison of SL predictors was reported in 2006 by Sprenger et al. [29], who considered five methods and nine types of SL (including the membrane location). Wolf PSORT [37] outperformed the other sequence-based methods, particularly in discriminating localizations endowed with the greatest number of known sequences, including nucleus, cytosol, mitochondrion and extracellular regions. In the same year BaCelLo [16] was implemented and the prediction performances of 11 methods in discriminating five SL classes on kingdom specific datasets was compared. BaCelLo, Targetp [32] and LOCtree [30] emerged as the best predictors [16].

In 2007 Klee and Sosa [38] compared the performance of 10 publicly available methods in predicting secreted human proteins. Protein Prowler, BaCelLo and SIGNALp [32] where shown to outperform other methods when globular proteins were considered. On the other hand, Wolf PSORT was the best method for discriminating secreted from membrane proteins, a task not performed by the above mentioned predictors. Brameier et al. [39] compared five methods in the task of predicting nuclear proteins: BaCelLo and LOCtree neatly resulted as the most efficient. From these efforts it appears that five predictors can be considered top scoring when performing in different tasks and in different kingdoms. Therefore, we will focus on these ones, whose main features are listed in Table 1 and briefly outlined in the following.

(a) TARGETp [32] is a tool that analyses the N-terminal portion of a sequence with a combination of feed-forward NNs for discriminating four classes: secreted, mitochondrial, chloroplastic and other proteins. It is also able to predict the cleavage site of signal and transit peptides.

(b) Protein Prowler [33] implements a series of sequential cascaded and recurrent NNs, whose results are collected and combined with a SVM. It considers the same information and data as TARGETp, discriminating the same four localizations.

(c) LOCtree [30] is based on several binary SVMs arranged in three different decision trees specific for plants, non-plants and prokaryotes, respectively. In prokaryotes it classifies secreted, periplasmic and cytoplasmic proteins. In eukaryotes it discriminates up to six SLs: extracellular space, nucleus, cytoplasm,
chloroplast, mitochondrion and other organelles.

The input information consists of the compositions of sequence profiles of the whole sequence and of a 50-residues long N-terminal portion. This input is supplemented with the predicted secondary structure composition, and, for eukaryotes, with the output of SIGNALp, a predictor for signal peptides starting from the sequence. Training data were derived from the release 40 of SwissProt and a rigorous homology reduction procedure was applied.

(d) BaCelLo [16] implements three predictors specific for three eukaryotic kingdoms: animals, plants and fungi. As in the case of LOCtree, it is based on a decision tree of binary SVMs. It is able to discriminate secreted, nuclear, cytoplasmic, mitochondrial and chloroplastic proteins. BaCelLo takes into account information derived from the whole sequence and from both the N- and C-terminal regions, and it also considers the composition of sequence profiles deriving from multiple sequence alignments. The hierarchical structure of the decision tree also allows defining intermediate classes as nucleo/cytoplasmic and chloroplastic/mitochondrial, allowing a more efficient discrimination into fewer classes. The training set contains non-redundant experimentally annotated sequences from the release 48 of SwissProt. The peculiarity of BaCelLo is the procedure adopted in order to make the predictor independent of the abundances of the different classes in training datasets.

(e) Wolf PSORT [37] is a classifier that extends the previously released PSORT II and iPSORT and computes, starting from the sequence, a large number of features (comprising the composition, the presence of signal and target peptides, transmembrane domains, NLSs, post-translational modification sites, DNA binding domains). It predicts SL comparing these features with those of the annotated proteins in SwissProt and implementing a \( k \)-nearest-neighbor classifier. Wolf PSORT discriminates up to 12 SL classes and has different parameterizations for animals, fungi and plants. Interestingly, it defines some dual classes, containing proteins with multiple localizations, such as proteins that shuttle between the cytoplasm and the nucleus. The result of a prediction with Wolf PSORT is a ranked list of possible SL annotations.

**COMPARISON AMONG THE BEST PERFORMING SEQUENCE-BASED SL PREDICTORS**

In this article, we compute a new assessment of the performance of the five earlier described methods on a set of experimentally annotated proteins extracted from the release 54 of the SwissProt database. SL annotations were searched in the ‘SUBCELLULAR LOCATION’ subfield of the ‘COMMENT’ field in the SwissProt entries and proteins annotated as ‘membrane’ were excluded, since most of the considered methods are specialized in predicting

---

**Table 1: Main features of the best performing sequence-based predictors for SL**

<table>
<thead>
<tr>
<th>Name</th>
<th>Kingdom specificity</th>
<th>No classes*</th>
<th>Multiple localization*</th>
<th>Method</th>
<th>Input</th>
<th>Training dataset (SwissProt release)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BaCelLo [16]</td>
<td>Animals, fungi, plants</td>
<td>5</td>
<td>No</td>
<td>Decision tree of SVMs</td>
<td>Aminoacidic composition, profiles, N- and C-termini</td>
<td>48</td>
</tr>
<tr>
<td>LOCtree [30]</td>
<td>Non-plants, plants, prokaryotes</td>
<td>6</td>
<td>No</td>
<td>Decision tree of SVMs</td>
<td>Profiles, secondary structure, N-terminus</td>
<td>40</td>
</tr>
<tr>
<td>Protein Prowler [33]</td>
<td>Non-plants, plants, prokaryotes</td>
<td>4</td>
<td>No</td>
<td>NN</td>
<td>N-terminal sequence</td>
<td>38</td>
</tr>
<tr>
<td>TARGETp [32]</td>
<td>Non-plants, plants, prokaryotes</td>
<td>4</td>
<td>No</td>
<td>NN/HMM</td>
<td>N-terminal sequence</td>
<td>38</td>
</tr>
<tr>
<td>Wolf PSORT [37]</td>
<td>Animals, fungi, plants</td>
<td>12</td>
<td>Yes</td>
<td>Rule based</td>
<td>Aminoacidic composition, N-terminal sequence, empirical rules</td>
<td>45</td>
</tr>
</tbody>
</table>

*Number of discriminated classes: BaCelLo discriminates secreted, cytoplasmic, nuclear, mitochondrial and chloroplastic proteins; LOCtree discriminates secreted, organellar, cytoplasmic, nuclear, mithochondrial and chloroplastic proteins; Protein Prowler and TARGETp discriminates secreted, mitochondrial, chloroplastic and other proteins; Wolf PSORT discriminates chloroplastic, cytosolic, cytoskeletal, endoplasmic reticulum, extracellular, golgi, lysosomal, mitochondrial, nuclear, peroxisomal, plasma membrane and vacuolar membrane proteins.

*Multiple localization refers to the ability of discriminating dual classes (e.g. nucleus/cytoplasm).
the localization of globular proteins. Since the mentioned predictors have been trained on sequences derived at the last from the release 48, we considered only the 5107 entries of eukaryotic annotated sequences deposited starting from the release 49. This set contains 2788 proteins from animals, 717 from fungi and 1602 from plants and will be referred as the ‘whole dataset’ in the sequel of this article.

A second dataset, referred as ‘reduced dataset’, was extracted from ‘whole dataset’, considering only the sequences sharing low identity with proteins from the release 48. To this aim we adopted an $E$-value lower than $1E^{-3}$ upon a BLAST search corresponding to a maximum sequence identity of about 30%. This set ensures to test predictors only on really new examples and contains 1412 sequences, 575 coming from animals, 437 from fungi and 400 from plants. In order to avoid to bias the evaluation of the prediction performances towards classes of proteins that are overrepresented in the testing set, we clustered sequences sharing the same localization and that align with an $E$-value lower than $1E^{-3}$, as computed with BLAST. The procedure led to 987 groups: 432 in animals, 418 in fungi and 132 in plants. In order to evaluate the prediction performances, the rates of correct and wrong predictions were computed for each group and then averaged over all the groups belonging to the same SL class. Prediction scores were then evaluated following the usual definitions [40], using these averaged rates of true and false predictions. This procedure ensures the evaluation of the complete set without biasing towards the most abundant groups of sequences.

Only five major SLs have been taken into consideration: nucleus, cytoplasm, mitochondrion, chloroplast and secretory pathway, searching for ‘nucleus’, ‘cytoplasm’, mitochondrion’, ‘chloroplast’ and ‘secreted’ keywords, respectively. Entries endowed with multiple keywords or with non-experimental annotations were discharged. The amount of proteins in each class for each kingdom is reported in the first column of Table 2. A virtual class containing both nuclear and cytoplasmic proteins has been generated to deal with TARGETp and Protein Prowler that do not discriminate the two classes. On the other hand, the 12 classes considered by Wolf PSORT were reduced to five considering as ‘secretory pathway’ the predictions ‘extracellular’, ‘golgi’, ‘lysosome’, ‘vacuolar’ and ‘endoplasmic reticulum’. Moreover, transmembrane predictions were neglected and the subsequent prediction in the ranked list was considered. In the case of LOCtree, secreted and organelle classes (that do not include mitochondrial and chloroplastic) were considered as ‘secretory pathway’.

All predictions were performed using the publicly available web-servers, except in the case of LOCtree whose predictions were kindly performed by the authors.

All the datasets used in the comparison are publicly available at http://gpcr.biocomp.unibo.it/bacello/dataset.htm.

Results of the comparison on each SL class and for each kingdom are shown in Table 2. Recall values measure the rate of correctly predicted proteins in each class, FP rate reports the proportion of proteins not belonging to a given class that were erroneously predicted as being positive, while the Matthews Correlation Coefficient (MCC) measures how much the predictors outperform a random assignment in each class [40]. Performances obtained on the whole and the reduced sets are reported in roman and italic, respectively. For sake of completeness we also reported the performances of the annotation by similarity: for each protein in the dataset, similar proteins were searched in the release 48 of SwissProt, with an $E$-value lower than $1E^{-3}$, and the corresponding SL annotation was assigned to the target sequence. When multiple hits were found, only the best one was taken into consideration. This procedure allows estimating the number of really new examples in our dataset; the relative recalls estimate the rate of proteins in each class that can be correctly annotated by similarity. On the overall about 75% of proteins from plants and animals can be annotated by similarity, with relevant biases on some localization classes. For fungi this rate lowers to about 40%. Table 3 lists the overall performances using the global accuracy score and the generalized correlation index as previously defined [40]. We will briefly discuss the results obtained for each discriminated class.

**Secretory pathway**

It is generally the best predicted class, probably owing to the presence of a recognizable N-terminal signal. BaCelLo and Wolf PSORT outperform the other methods when proteins from animals and fungi are considered; in the case of plants the performance of Wolf PSORT drops to random and LOCtree and
Table 2: Comparison of the performance of the top scoring sequence-based predictors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No of proteinsa</td>
<td>Recall (%)</td>
<td>FP rate (%)</td>
<td>MCC</td>
<td>Recall (%)</td>
<td>FP rate (%)</td>
</tr>
<tr>
<td>Animals (78%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secretory Pathway</td>
<td>272</td>
<td>93</td>
<td>4</td>
<td>0.89</td>
<td>90</td>
<td>5</td>
</tr>
<tr>
<td>Mitochondrion</td>
<td>75</td>
<td>93</td>
<td>3</td>
<td>0.88</td>
<td>79</td>
<td>9</td>
</tr>
<tr>
<td>Nucleus</td>
<td>241</td>
<td>90</td>
<td>7</td>
<td>0.68</td>
<td>79</td>
<td>3</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>979</td>
<td>62</td>
<td>8</td>
<td>0.58</td>
<td>80</td>
<td>14</td>
</tr>
<tr>
<td>Nucleus or cytoplasm</td>
<td>224</td>
<td>57</td>
<td>17</td>
<td>0.41</td>
<td>66</td>
<td>27</td>
</tr>
<tr>
<td>Fungi (39%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secretory Pathway</td>
<td>26</td>
<td>100</td>
<td>3</td>
<td>0.75</td>
<td>93</td>
<td>8</td>
</tr>
<tr>
<td>Mitochondrion</td>
<td>104</td>
<td>80</td>
<td>15</td>
<td>0.53</td>
<td>49</td>
<td>6</td>
</tr>
<tr>
<td>Nucleus</td>
<td>256</td>
<td>77</td>
<td>13</td>
<td>0.58</td>
<td>42</td>
<td>8</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>152</td>
<td>72</td>
<td>33</td>
<td>0.38</td>
<td>63</td>
<td>32</td>
</tr>
<tr>
<td>Nucleus or cytoplasm</td>
<td>331</td>
<td>45</td>
<td>15</td>
<td>0.33</td>
<td>49</td>
<td>17</td>
</tr>
<tr>
<td>Plants (75%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secretory Pathway</td>
<td>18</td>
<td>94</td>
<td>4</td>
<td>0.42</td>
<td>83</td>
<td>2</td>
</tr>
<tr>
<td>Mitochondrion</td>
<td>38</td>
<td>29</td>
<td>0</td>
<td>0.53</td>
<td>55</td>
<td>24</td>
</tr>
<tr>
<td>Chloroplast</td>
<td>1345</td>
<td>72</td>
<td>17</td>
<td>0.54</td>
<td>77</td>
<td>12</td>
</tr>
<tr>
<td>Nucleus</td>
<td>99</td>
<td>84</td>
<td>11</td>
<td>0.48</td>
<td>80</td>
<td>14</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>36</td>
<td>88</td>
<td>22</td>
<td>0.60</td>
<td>72</td>
<td>11</td>
</tr>
<tr>
<td>Nucleus or cytoplasm</td>
<td>102</td>
<td>43</td>
<td>6</td>
<td>0.32</td>
<td>78</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>27</td>
<td>2</td>
<td>0.38</td>
<td>33</td>
<td>3</td>
</tr>
</tbody>
</table>

*The amount of sequences/groups in each considered class is listed in the second column.

bLast column lists the recall in each class, when a similarity based annotation is performed. Each protein in the whole database has been aligned with BLAST (E-value <1E-3) against the SwissProt 48 database. The bold values represent the percentage of proteins annotated in each kingdom.

Recall is computed as the percent rate of correctly classified sequences in each class. FP rate for a class is computed as the percent proportion of proteins not belonging to the class that were erroneously predicted as being positive. MCC is the Matthews Correlation Coefficient [40]. Figures in roman are computed on the whole dataset comprising 5106 annotated sequences whose entries have been deposited in the SwissProt database from release 49 up to release 54; proteins annotated as 'membrane' were removed. Figures in italic are computed on a reduced dataset, containing 1412 sequences extracted from the whole dataset, considering only the sequences sharing low identity (that is an E-value lower than 1E-3 upon a BLAST search) with proteins from release 48. These proteins were further grouped into 937 homology groups, adopting the same E-value threshold. The number of groups for each class is indicated in the second column and in italic. For each group the number of correct and wrong predictions was computed. These numbers were then averaged on all the groups belonging to the same class and the performance scores were computed with the usual definitions [40]. A bootstrapping procedure was run to assess the statistical variance of the dataset. After 1000 bootstrapping cycles the absolute standard deviation associated to recalls is <3% for sequences from animals and fungi, and it is <8% for sequences from plants. In the cases of FP rate and MCC, absolute standard deviation values are always lower than 1% and 0.05, respectively.
BaCelLo are the best performers. It has to be considered that, for this last case, only 18 examples are available and statistical fluctuations are likely to influence the results. However, a bootstrap procedure proved that the standard deviation on the MCC is lower than 0.05 (see Table 2 for details).

Mitochondrion
When animals and fungi are considered BaCelLo, Wolf PSORT and LOCtree (and Protein Prowler in fungi) score with similar MCC, but BaCelLo has a better recall when the whole dataset is considered. On the reduced dataset, containing the sequences that do not share any similarity with the training set, BaCelLo outperforms the other predictors. The MCC performance drops out in the case of plants for all the predictors. In the case of BaCelLo and LOCtree this is due to a poor separation of chloroplastic and mitochondrial proteins along the decision trees (data not shown).

Chloroplast
BaCelLo outperforms all the methods in the whole dataset and three predictors (TARGETp, Protein Prowler and Wolf PSORT) score similarly to a random guess, with a MCC near to zero. This is quite surprising, since all of them search explicitly for chloroplastic target peptides. In the reduced dataset a general improvement in the prediction performances is reported, and LOCtree outperforms the other methods. The dependence of the scores on the dataset can be due to the fact that the chloroplastic proteins of the whole dataset are highly redundant and clustered in few groups in the reduced set.

Nucleus or cytoplasm
This class contains all the proteins that are either in the nucleus or in the cytoplasm and that are not endowed with signal or transit peptides. For all the methods, performances are generally good in the case of animals and generally decrease in the case of fungi. For plants, BaCelLo and LOCtree clearly outperform the other predictors, though their performances are not as high as in the case of animals.

Nucleus
Wolf PSORT outperforms others methods in predicting nuclear proteins for animal and fungi, but reports very high FP rate in fungi. BaCelLo and LOCtree are the best nuclear predictors for plants, due also to the fact that their decision trees better discriminate the ‘nucleus or cytoplasm’ class, as observed in the preceding paragraph.

Cytoplasm
The class seems to be the most difficult to be predicted for all the predictors, particularly when the performance on the reduced dataset is considered. In this case performances on the whole and

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Q (%)</td>
<td>GC</td>
<td>Q (%)</td>
<td>GC</td>
<td>Q (%)</td>
</tr>
<tr>
<td>Animals</td>
<td>3</td>
<td>89</td>
<td>0.79</td>
<td>91</td>
<td>0.78</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>75</td>
<td>0.70</td>
<td>78</td>
<td>0.59</td>
<td>78</td>
</tr>
<tr>
<td>Fungi</td>
<td>3</td>
<td>82</td>
<td>0.65</td>
<td>81</td>
<td>0.49</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>59</td>
<td>0.56</td>
<td>57</td>
<td>0.44</td>
<td>86</td>
</tr>
<tr>
<td>Plants</td>
<td>4</td>
<td>77</td>
<td>0.48</td>
<td>73</td>
<td>0.42</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>76</td>
<td>0.46</td>
<td>70</td>
<td>0.59</td>
<td>63</td>
</tr>
</tbody>
</table>

Classes: number of discriminated classes. The overall accuracy (Q) is computed as the percent proportion of correctly classified sequences over the total of tested proteins. The generalized correlation (GC) is a generalization of the Matthews Correlation Coefficient and was computed as previously described [40]. Figures in roman are computed on the whole data set as described in text; figures in italic are computed on the reduced dataset (see Table 2 for details). A bootstrapping procedure was run to assess the statistical variance of the dataset. After 1000 bootstrapping cycles, the absolute standard deviation associated to the overall accuracy and generalized correlation are always <2% and 0.02, respectively.
reduced sets greatly differ, since many of the collected cytoplasmic sequences have homologues in SwissProt 48. The MCCs on the reduced sets are generally poor. In most cases cytoplasmic proteins are confused with nuclear ones. For this reason the ‘nucleus or cytoplasm’ class can give a better performance.

CONCLUSIONS
Secreted proteins are in all cases the easiest class to predict, due to the presence of signal peptides and compositional differences, as the relative abundance in cysteine residues. Methods such as TARGETp and Protein Prowler, searching only for the signal peptide, lack proteins secreted via non classical ways, while methods such as BaCelLo and LOCtree are able to recognize them since they take into consideration other features. The drawback is the incapability of determining the signal peptide cleavage site.

Good predicting performances are also achieved in the prediction of ‘nucleus or cytoplasm’ class; however, for all the tested predictors it is generally a very hard task to further discriminate between cytoplasmic and nuclear proteins. Similarly, in plants it is quite easy to predict mitochondrial or chloroplastic proteins, but it is very hard to discriminate the two classes. In particular, methods based on the explicitly search of an N-terminal target peptide fail in this task. This can be due to the fact that other signals, such as C-terminal target peptides, can determine the translocation to the chloroplasts and mitochondria.

On the overall BaCelLo and Wolf PSORT seem the best performing predictors for animals and fungi sequences when four localizations are discriminated, while BaCelLo and LOCtree are the only satisfactory predictors in the case of plant proteins. It has to be pointed out that Wolf PSORT aims to discriminate up to 12 classes and we tested it reducing the class definitions. However, it is impossible to reliably assess the performances in predicting more than four or five classes, due to the paucity of non-homologous examples.

Presently, it can be concluded that the golden rule for SL annotation of a novel protein is to search for annotated homologues and eventually to check the annotation using ‘state of the art’ predictors. When no homologue is found sequence-based predictors are necessary but reliable results can be obtained only when few major classes are discriminated.

DATABASE OF SL PREDICTION FOR WHOLE GENOME
For annotating the proteins out of genome sequencing projects of eukaryotes, two main databases containing pre-computed annotations are publicly available. The first one is PA-GOSUB that contains the annotation by homology of 11 genomes. Since no sequence-based prediction is used, only a portion of the genomes can be annotated, ranging from 50 to 90% [22].

The other database, eSLDB, on the contrary, contains three types of annotation: experimental, homology derived and predicted using a pipeline containing SPeP [41] (for the prediction of signal peptide), ENSEMBLE [42] (for the discrimination of membrane proteins) and BaCelLo. In this way the whole genome of five organisms has been annotated.

FUTURE DIRECTIONS
As a general conclusion, it seems, therefore that predictions should be improved, provided that the number and the quality of experimental data necessary for training are also improved. This will be possible when more mechanisms responsible for the localization of proteins will be taken into consideration and when the data will allow to discover and/or to better characterize the delivery processes to different cell compartments in the different kingdoms. For example, it is known that mitochondrial localization is influenced by the UTR regions of the mRNA transcripts; however, these data are not directly available at the protein level [43] and their inclusion in a SL predictor should be worked out. Another interesting problem is related to those proteins that may have multiple localizations, as in the case of proteins shuttling between nucleus and cytoplasm. Wolf PSORT is the only predictor that presently addresses this last problem on statistical bases. Finally, it is worth noticing that proteins, such as the so-called moonlight proteins may be or may be not endowed with different structures in different compartments, depending on the different function. All the above issues, however, need more experimental characterization in order to be properly taken into consideration when developing a SL predictor.

WEB SITES
BaCelLo http://gpcr.biocomp.unibo.it/bacello/
eSLDB http://gpcr.biocomp.unibo.it/esldb/
Key Points

- Prediction of SL is useful for the functional annotation of whole proteomes, especially for eukaryotes.
- Different tools have been implemented for predicting SL, based either on similarity search, motif detection and sequence analysis with ML methods.
- Similarity and motif-based predictors cannot annotate all the proteins in a proteome and sequence-based methods have to be adopted.
- When comparing the best performing sequence-based predictors, BaCelLo and Wolf PSORT are the best ones for animals and fungi, in discriminating four localizations. LocTree and BaCelLo are the only satisfactory predictors in the case of plant proteomes.
- At the moment the golden rule for SL annotation of a novel protein is to search for annotated homologues and eventually to check the annotation using ‘state of the art’ predictors. When no homologue is found sequence-based predictors are necessary but reliable results can be obtained only when few major classes are discriminated.

Acknowledgements

We thank Dr Rajesh Nair for kindly providing the prediction with LocTree. R.C. acknowledges the receipt of the following grants: 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). A.P. is supported by a FIRB 2003-LIBI grant.

References


