SCLpred: Protein Subcellular Localization Prediction by N-to-1 Neural Networks

Catherine Mooney, Yong-Hong Wang, and Gianluca Pollastri

1 School of Computer Science and Informatics, University College Dublin, Ireland
2 School of Medicine and Medical Science, University College Dublin, Ireland
3 Complex and Adaptive Systems Laboratory, University College Dublin, Ireland
4 Conway Institute of Biomolecular and Biomedical Science, University College Dublin, Ireland
5 Biophysics Institute, Hebei University of Technology, Tianjin, China

ABSTRACT

Summary: Knowledge of the subcellular location of a protein provides valuable information about its function and possible interaction with other proteins. In the post-genomic era, fast and accurate predictors of subcellular location are required if this abundance of sequence data is to be fully exploited. We have developed a subcellular localization predictor (SCLpred) which predicts the location of a protein into four classes for animals and fungi and five classes for plants (secreted, cytoplasm, nucleus, mitochondrion and chloroplast) using machine learning models trained on large non-redundant sets of protein sequences. The algorithm powering SCLpred is a novel Neural Network (N-to-1 Neural Network, or N1-NN) we have developed, which is capable of mapping whole sequences into single properties (a functional class, in this work) without resorting to predefined transformations, but rather by adaptively compressing the sequence into a hidden feature vector. We benchmark SCLpred against other publicly available predictors using two benchmarks including a new subset of Swiss-Prot Release 2010_08. We show that SCLpred surpasses the state of the art. The N1-NN algorithm is fully general and may be applied to a host of problems of similar shape, that is, in which a whole sequence needs to be mapped into a fixed-size array of properties, and the adaptive compression it operates may shed light on the space of protein sequences.

Availability: The predictive systems described in this paper are publicly available as a web server at http://distill.ucd.ie/

Contact: gianluca.pollastri@ucd.ie

1 INTRODUCTION

With the recent advances in high throughput sequencing technology there has been a rapid increase in the availability of sequence information. To fully exploit this information sequences need to be annotated quickly and accurately, which has led to the development of automated annotation systems. A major step towards determining the function of a protein is determining its subcellular localization (SCL). Knowledge of the location of the protein sheds light not only on where it might function but also what other proteins it might interact with, as, in order to interact, proteins must inhabit the same location or physically adjacent compartments, at least temporarily. There is a growing gap between the number of proteins that have reliable SCL annotations and the number of known protein sequences. Experimental approaches to SCL prediction are time-consuming and expensive, whereas computational methods can provide fast and increasingly accurate localization predictions.

There are various different mechanisms by which a protein is directed to a particular location in the cell and there are many possible compartments in which eukaryotic proteins may be located. Some nuclear proteins have a nuclear localization signal (NLS) which may occur anywhere in the sequence (Cokol et al., 2000). Most secreted, mitochondrial and chloroplastic proteins have N-terminal cleavable peptides (SP, mTP and cTP) but many proteins have no known motif (Emanuelsson, 2002; Nair and Rost, 2005), and many are known not to have N-terminal peptides (Bendtsen et al., 2004a). Even in these cases the information contained in a protein sequence may be sufficient to predict the protein’s location in the cell, given that residue and k-residue frequencies correlate with locations (Nakashima and Nishikawa, 1994; Emanuelsson, 2002; Nair and Rost, 2003, 2005).

There are many methods for the prediction of SCL which can be roughly divided into two groups: homology or knowledge-based, that rely on similarity to another sequence of known location, or other known information about the sequence or similar sequences, for example WoLF PSORT (Horton et al., 2007) or SherLoc (Shatkay et al., 2007); and de novo or ab initio, sequence-based methods, which may use evolutionary information in the form of multiple sequence alignments (MSA), but do not depend on similarity to sequences of known location, for example BaCelLo (Pierleoni et al., 2006).

We predict SCL for eukaryotes only, which we divide into animals, plants and fungi. There are many potential classes of subcellular localization, and different prediction systems sometimes use different class subdivisions, ranging from 3 (Emanuelsson et al., 2000; Bödén and Hawkins, 2005; Hawkins and Bödén, 2006) up to more than 10 classes (Horton et al., 2007). Here, similarly to BaCelLo (Pierleoni et al., 2006), to which we directly compare our results, we consider four subcellular localizations for animals and
fungi and five for plants: nucleus, cytoplasm, mitochondrion, chloroplast and secreted. In a first series of tests we adopt essentially the same experimental setting as in Casadio et al. (2008) and Pierleoni et al. (2006), to which we compare our predictor. We then take a further step by developing new, redundancy reduced training and testing sets starting from Swiss-Prot Release 2010_06 (Boeckmann et al., 2003), and benchmark SCLpred on these sets against six state-of-the-art, publicly available predictors of SCL: BaCelLo, LOCtree, SherLoc, Protein Prowler, TargetP and WoLF PSORT, which we briefly describe in the following sections.

**BaCelLo**  
BaCelLo (Pierleoni et al., 2006) uses a hierarchy of binary support vector machines (SVM) to predict SCL for eukaryotes into four classes for animals and fungi and five for plants: secreted, cytoplasm, nucleus, mitochondrion and chloroplast. BaCelLo is trained on a non-redundant set of sequences from Swiss-Prot 48. Predictions are made from the full sequence, from the N- and C-terminal regions and evolutionary information. BaCelLo is available at [http://gpcr.biocomp.unibo.it/bacello/](http://gpcr.biocomp.unibo.it/bacello/).

**LOCtree**  
LOCtree (Nair and Rost, 2005) uses binary SVMs to predict SCL. Three versions of the predictor are available, for plants, non-plants and prokaryotes. For prokaryotes predictions are into three classes: secreted, periplasm and cytoplasm. For eukaryotes predictions are into six classes: extracellular space, nucleus, cytoplasm, chloroplast, mitochondrion and other organelles. LOCtree is trained on a redundancy reduced subset of Swiss-Prot 40. Predictions are made from the full sequence, a 50-residue N-terminal region, predicted secondary structure and the output of SIGNALp (for eukaryotes). LOCtree is available at [http://www.predictprotein.org/](http://www.predictprotein.org/).

**SherLoc**  
SherLoc (Shatkay et al., 2007) uses SVM which integrate sequence and text-based features. There are three predictors (animal, fungi, plant) which predict into ten locations for animals and fungi: cytoplasm, endoplasmic reticulum, extracellular, Golgi, lysosome, mitochondrion, nucleus, peroxisome, plasma membrane, vacuole and an extra class, chloroplast, for plants. The predictors are trained on sequences extracted from Swiss-Prot 42. http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc/.

**TargetP**  
TargetP (Emanuelsson et al., 2000) uses a feed-forward neural network for the prediction of plant and non-plant SCL into three and four classes respectively, based on the N-terminal sequence. The prediction is based on the presence of a chloroplast transit peptide (cTP), a mitochondrial targeting peptide (mTP) or a secretory pathway signal peptide (SP). TargetP is available at [http://www.cbs.dtu.dk/services/TargetP/](http://www.cbs.dtu.dk/services/TargetP/).

**Protein Prowler**  
Protein Prowler (Böden and Hawkins, 2005; Hawkins and Böden, 2006) is based on the ideas behind TargetP and trained on a subset of Swiss-Prot 37 and 38. The predictor uses neural networks and SVMs specialized for the prediction of plants or non-plants and predicts into the following classes: secretory pathway, mitochondrion, chloroplast and other. Protein Prowler is available at [http://pprowler.itee.uq.edu.au/](http://pprowler.itee.uq.edu.au/).

### Table 1. Number of sequences per class for each of the three kingdoms in the BaCelLo training set and the BaCelLo 2008 test set.

<table>
<thead>
<tr>
<th></th>
<th>Animals</th>
<th>Fungi</th>
<th>Plants</th>
<th>Animals</th>
<th>Fungi</th>
<th>Plants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytoplasm</td>
<td>439</td>
<td>211</td>
<td>58</td>
<td>846</td>
<td>331</td>
<td>102</td>
</tr>
<tr>
<td>Mitochondrion</td>
<td>185</td>
<td>188</td>
<td>67</td>
<td>241</td>
<td>104</td>
<td>38</td>
</tr>
<tr>
<td>Nucleus</td>
<td>1166</td>
<td>711</td>
<td>121</td>
<td>979</td>
<td>256</td>
<td>99</td>
</tr>
<tr>
<td>Secreted</td>
<td>804</td>
<td>88</td>
<td>41</td>
<td>722</td>
<td>26</td>
<td>18</td>
</tr>
<tr>
<td>Chloroplast</td>
<td>294</td>
<td></td>
<td>1345</td>
<td>2788</td>
<td>717</td>
<td>1602</td>
</tr>
<tr>
<td>Total</td>
<td>2597</td>
<td>1198</td>
<td>491</td>
<td>2788</td>
<td>717</td>
<td>1602</td>
</tr>
</tbody>
</table>

**WoLF PSORT**  
WoLF PSORT (Horton et al., 2007) is a version of the PSORT family of SCL predictors for the prediction of eukaryotic proteins based on their sequence. Based on a number of features (residue composition, presence of known sorting signal and target peptides etc) WoLF PSORT uses a k-nearest neighbor classifier, comparing these features to other Swiss-Prot annotated proteins, resulting in a ranked list of up to 12 possible locations: chloroplast, cytosol, cytoskeleton, endoplasmic reticulum, extracellular, Golgi apparatus, lysosome, mitochondrion, nuclear, peroxisome, plasma membrane, vacuolar membrane. WoLF PSORT is available at [http://wolfpsort.org/](http://wolfpsort.org/).

## 2 MATERIALS AND METHODS

### 2.1 Datasets

The first dataset which we use to train and test SCLpred is the dataset used by Pierleoni et al. (2006) to train BaCelLo in 10-fold cross-validation, for a direct comparison with this predictor. We call this set the BaCelLo training set. We also test SCLpred on the test dataset used in Casadio et al. (2008) (BaCelLo 2008 test set), which is based on Swiss-Prot 54 (Table 1). The BaCelLo 2008 test set is redundancy reduced excluding all sequences with e = 10^{-30} to the BaCelLo training set. Next we create a new training and test set starting from Swiss-Prot Release 2010_06. We start from 97,939 Metazoa, 27,540 Fungi and 28,998 Viridiplantae sequences. Of these 74,724, 20,196 and 22,442, respectively, have a “SUBCELLULAR LOCATION”. We remove membrane proteins and sequences that have non-experimental qualifiers (Potential, Probable, By similarity), leaving 16,406, 3,339 and 7,116 sequences, respectively. We internally redundancy reduce each of these sets using an all-against-all BLAST search (with e = 10^{-5}) to the BaCelLo training set. We then take a further step by developing new, redundancy reduced training and test sets starting from Swiss-Prot Release 2010_06 containing 6,464,895 sequences. The alignments are generated by three runs of PSI-BLAST with parameters b = 3000 (maximum number of hits) and e = 10^{-3} (expectation of a random hit).

### 2.2 Predictive architecture: N1-NN

We call the model we describe in this work N-to-1 Neural Network, or N1-NN. The model is based on our framework to design Neural Networks
Table 2. Number of sequences per class for each of the three kingdoms in the 2010\textsubscript{06} training set and the 2009+ test set.

<table>
<thead>
<tr>
<th></th>
<th>2010\textsubscript{06} training set</th>
<th>2009+ test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Animals</td>
<td>Fungi</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>1364</td>
<td>890</td>
</tr>
<tr>
<td>Mitochondrion</td>
<td>315</td>
<td>413</td>
</tr>
<tr>
<td>Nucleus</td>
<td>1830</td>
<td>1150</td>
</tr>
<tr>
<td>Secreted</td>
<td>1584</td>
<td>111</td>
</tr>
<tr>
<td>Chloroplast</td>
<td>523</td>
<td>194</td>
</tr>
<tr>
<td>Total</td>
<td>5095</td>
<td>2564</td>
</tr>
</tbody>
</table>

for structured data (Baldi and Pollastri, 2003; Walsh et al., 2009). The aim of the model is to map a sequence of variable length $N$ into a single property or fixed-width array of properties. Other models transform/compress the sequence into a fixed number of descriptors (or into descriptors of pairwise relations between sequences) beforehand, and then they map these descriptors into the property of interest. These descriptors are typically frequencies of residues or k-mers, sometimes computed separately on different parts of the sequence (e.g. around the termini, as in Pierleoni et al. (2006)). In some cases whole sections of the sequence are directly taken into account (again, typically the termini, where some signals are to be found), but even in this case the size of this section needs to be fixed and decided beforehand.

In N1-NN, instead, we do not compress all the information of a sequence into a handful of predefined features (e.g. k-mer frequencies, sequence length, etc.). Rather, we decide beforehand only how many features we want to compress a sequence into. If these features are stored in a vector $f = (f_1, \ldots, f_o)$, and if we represent the $i$-th residue in the sequence as $r_i$, then $f$ is obtained as:

$$f = k \sum_{i=1}^{N} N^{(h)}(r_{i-1}, \ldots, r_{i+c})$$

(1)

where $N^{(h)}$ is a non-linear function, which we implement by a two-layered feed-forward Neural Network with $h$ non-linear output units (the sequence-to-feature network), $N^{(h)}$ is replicated $N$ times ($N$ being the sequence length), and $k$ is a normalization constant. The feature vector $f$ is obtained by combining information coming from all windows of $2c+1$ residues in the protein. If $c = 20$, as in all the tests in this article, the motifs have a length of 41 residues. The feature vector $f$ thus obtained is mapped into the property of interest $o$ (for instance, cellular component class), as follows:

$$o = N^{(o)}(f)$$

(2)

where $N^{(o)}$ is a non-linear function which we implement by a second two-layered feed-forward neural network (the feature-to-output network). The whole neural network (the cascade of $N$ replicas of the sequence-to-feature vector network and one feature-to-output network) is itself a feed-forward neural network, thus can be trained by gradient descent via the back-propagation algorithm. As there are $N$ copies of $N^{(h)}$ for a sequence of length $N$, there will be $N$ contributions to the gradient for this network, which are added together.

The feature vector $f$ is a compression of the sequence into $h$ real-valued descriptors. These descriptors are automatically determined/learned in order to minimize the output error, hence to be most informative to predict the property of interest. Although there is a daunting number of possible motifs of length $2c+1$, the model does not need to count them or represent them all. Only a relatively small number of free parameters is available to represent all the motifs in a sequence. This prevents overparametrisation and model fitting problems that arise when one counts frequencies of n-mers as soon as $n > 2 - 3$. If training is successful, only (soft) motifs relevant to the task at hand are represented in $f$. Thus $f$ is effectively a compressed version of the sequence into a fixed-size array. The compression is property-driven, meaning that different predictive targets generally induce different representations of a sequence.

![Fig. 1. An N-to-1 Neural Network. $N$ copies of the $N^{(h)}$ network (only 3 represented for simplicity) process all the (overlapping) motifs of a predefined length in a sequence. The vectorial outputs $f_k$ of these networks are added up and the resulting feature vector $f$ is input to the $N^{(o)}$ network to produce the localization prediction.](image-url)
During preliminary experiments (run on the BaCelLo plant training set split into 2/3 for training and 1/3 for testing) we tested \( N_f^H \) values of 6, 8 and 10, which all yielded similar performances. When choosing a motif size, we considered that the average size for known signal peptides in eukaryotes is approximately 20 residues (Bendtsen et al., 2004b), and 35-40 is an upper size bound for most known signals and NLS (Bendtsen et al., 2004b; Cokol et al., 2000). It should be noted that, since all (overlapping) motifs of length 2c+1 are considered by an N-to-1 NN, it is not strictly necessary for 2c+1 to cover all motif sizes, as a signal larger than 2c+1 is still input to an N-to-1 NN as all its overlapping substrings of length 2c+1, although this may lead to the loss of some positional information. During preliminary experiments we tested \( c \) values of 10 and 15, which performed marginally less well than \( c = 20 \). We kept \( N_f^H \) and \( N_f \) fixed at 10 in all experiments. During the final cross-validations we used exactly the same architecture for all sets and all kingdoms, in which \( N_f^H = N_f = N_f^M = 10 \) and \( c = 20 \).

All trainings are also identical in that the weights in the networks are updated every 10 examples (proteins) and 2000 epochs of training are performed by gradient descent on the error, which we model as the relative entropy between the target class and the output of the network.

During preliminary experiments (run on the BaCelLo plant training set split into 2/3 for training and 1/3 for testing) we tested \( N_f^H \) values of 6, 8 and 10, which all yielded similar performances. When choosing a motif size, we considered that the average size for known signal peptides in eukaryotes is approximately 20 residues (Bendtsen et al., 2004b), and 35-40 is an upper size bound for most known signals and NLS (Bendtsen et al., 2004b; Cokol et al., 2000). It should be noted that, since all (overlapping) motifs of length 2c+1 are considered by an N-to-1 NN, it is not strictly necessary for 2c+1 to cover all motif sizes, as a signal larger than 2c+1 is still input to an N-to-1 NN as all its overlapping substrings of length 2c+1, although this may lead to the loss of some positional information. During preliminary experiments we tested \( c \) values of 10 and 15, which performed marginally less well than \( c = 20 \). We kept \( N_f^H \) and \( N_f \) fixed at 10 in all experiments. During the final cross-validations we used exactly the same architecture for all sets and all kingdoms, in which \( N_f^H = N_f = N_f^M = 10 \) and \( c = 20 \).

All trainings are also identical in that the weights in the networks are updated every 10 examples (proteins) and 2000 epochs of training are performed by gradient descent on the error, which we model as the relative entropy between the target class and the output of the network.

To measure performances for a given class \( i \) we use:

\[
\text{Spec} = \frac{TP}{TP + FN} \\
\text{Sens} = \frac{TP}{TP + FP} \\
FPR = \frac{FP}{FP + TN} \\
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

(5)

where:

- True positives (TP): \( z_{ij} \).
- False positives (FP): \( \sum_{j \neq i} z_{ij} \).
- True negatives (TN): \( \sum_{i \neq j} \sum_{j \neq i} z_{ij} \).
- False negatives (FN): \( \sum_{j \neq i} z_{ij} \).

We emphasize performances based on GC (see Baldi et al. (2000) for more details), as this index minimizes the effect of class sizes. For some of the experiments we extract performances of other predictors from the literature, hence not all indices are reported at all times.

### 3 RESULTS AND DISCUSSION

In previous tests BaCelLo (Pierleoni et al., 2006) was shown to outperform the following publicly available methods for the prediction of the subcellular localization: LOCtree (Nair and Rost, 2005), Psort II (Nakai and Horton, 1999), SubLoc (Hua and Sun, 2001), ESPl pred (Bhasin and Raghava, 2004), LOCSVMpsi (Xie et al., 2005), SLP-local (Matsuda et al., 2005), Protein Prowler (Bödén and Hawkins, 2005), TARGETp (Emanuelsson et al., 2000), PredoTar (Small et al., 2004) and pTARGET (Guda and Subramaniam, 2005).

In Table 3 we show the performance of SCLpred compared to BaCelLo on the BaCelLo training set (Pierleoni et al., 2006). Both predictors are assessed by 10-fold cross-validation on the same set. Overall SCLpred is far more accurate for animals (Q 82% versus 74% and GC 72% versus 67%) and fungi (Q 75% versus 70% and GC 67% versus 66%) while the accuracy for plants (Q) is the same (68%) but GC is still considerably higher for SCLpred (63% vs. 59%).
Table 4 shows the accuracy of the same version of SCLpred tested on the BaCelLo2008 test dataset from Casadio et al. (2008) compared to the other five SCL predictors tested on the same dataset (results from Casadio et al. (2008)). Notice that two of the predictors (Protein Prowler and TargetP) use a different class assignment ("easier" as comprised by fewer classes) and are thus not directly comparable to SCLpred. The results report to versions of the various predictors that were trained on datasets extracted from Swiss-Prot release 48 or earlier. Since the BaCelLo2008 test set is extracted from Swiss-Prot release 54 and redundancy reduced against Swiss-Prot 48, there is no significant overlap between the training sets of any predictors in the table and the BaCelLo2008 test set. For animals we obtain a Q of 85% and GC of 81%, higher than the second best predictor that is directly comparable (Wolf PSORT, with 81% and 75%, respectively). SCLpred also performs better than the two predictors that are not directly comparable on the two classes that are common (mitochondrion and secreted) on fungi. On fungi SCLpred has the best Q (60% vs. BaCelLo’s 59%) and the second best GC (57% vs Wolf PSORT’s 59%). On plants SCLpred has by far the highest GC (58% vs BaCelLo’s 46%) and the joint highest Q (76%, again with BaCelLo).

It should be noted that BaCelLo was optimized for balanced class accuracies (Pierleoni et al., 2006), that is, to maximize average class sensitivity (nQ measure). Based on nQ, SCLpred still outperforms BaCelLo on both the BaCelLo and BaCelLo2008 set for animal proteins (by 2.3% and 6.2%, respectively). BaCelLo fares better on fungi (by 4.5%, and 2%), while on plants BaCelLo does better on the BaCelLo training set (by 4.6%) and SCLpred on the BaCelLo2008 test set (by 2.2%). Overall BaCelLo shows a more balanced sensitivity across classes than SCLpred, although in the case of animal proteins this is at a lower average level.

We repeat the experiments on a new training set extracted from the 2010_06 release of Swiss-Prot, which is approximately twice the size of the BaCelLo set for all three kingdoms. The accuracy of this new version of SCLpred is shown in Table 5. On animal and fungi overall performances are lower, in absolute value, to those obtained on the BaCelLo set. We attribute this to the more balanced nature of the 2010_06 training set, which is thus intrinsically "harder". Assigning proteins randomly to classes with a probability proportional to class frequencies yields a Q measure 3% higher on the BaCelLo animal training set than on the 2010_06 set (33.1% vs. 30.1%) and 6.3% higher on fungi (41.3% vs. 35.0%). Always predicting the most numerous class yields a 9% higher Q on the BaCelLo set compared to 2010_06 for animals (44.9% vs 35.9%) and 14.4% higher for fungi (59.3% vs. 44.9%). Moreover, in both kingdoms the class which is overrepresented in the 2010_06 set compared to BaCelLo is cytoplasm (26.8% vs. 16.9% of the examples for animal, 34.7% vs. 17.6% of the examples for fungi), which in all out tests is the hardest to predict. Hence not only is 2010_06 more challenging because of its distribution of examples, but also because it contains a higher proportion of difficult instances. On plants Q is higher on the 2010_06 training than on the BaCelLo training set (71% vs. 68%) while GC is lower (58% vs 63%). This is the result of a larger chloroplast class (which is well predicted) in 2010_06, and of the mitochondrion class being only 7% of the 2010_06 set (vs. 14% in BaCelLo), which results in infrequent predictions for this class. While the improvement on the much larger chloroplast class dominates in terms of Q measure, the reduction of performances on mitochondrion dominates with respect to GC, which weighs all classes equally. Overall it should be noted that, because of different class composition, it is hard to compare Q and GC measures across different datasets, and different predictors should always be ranked on the same dataset, as we do throughout this paper.
Table 5. SCLpred, trained and tested in 10-fold cross-validation on the 2010/06 training set

<table>
<thead>
<tr>
<th></th>
<th>Animals</th>
<th>Fungi</th>
<th>Plants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spec</td>
<td>Sens</td>
<td>MCC</td>
</tr>
<tr>
<td>Chlo</td>
<td>0.62</td>
<td>0.65</td>
<td>0.50</td>
</tr>
<tr>
<td>Mito</td>
<td>0.73</td>
<td>0.59</td>
<td>0.64</td>
</tr>
<tr>
<td>Nucl</td>
<td>0.76</td>
<td>0.76</td>
<td>0.62</td>
</tr>
<tr>
<td>Secr</td>
<td>0.91</td>
<td>0.91</td>
<td>0.87</td>
</tr>
<tr>
<td>GC</td>
<td>0.68</td>
<td>0.63</td>
<td>0.58</td>
</tr>
<tr>
<td>Q</td>
<td>0.77</td>
<td>0.67</td>
<td>0.71</td>
</tr>
</tbody>
</table>

We then test the version of SCLpred trained on the 2010/06 set on the 2009+ dataset (a subset of Swiss-Prot 2010/06 with less than 30% sequence similarity to the training set, described in the dataset section). We compare its accuracy with BaCelLo, SherLoc, WoLF PSORT, Protein Prowler and TARGETp (Table 6).

Results for TARGETp and Protein Prowler are based on three class predictions for animals and fungi, and four for plants, whereas for WoLF PSORT and SherLoc prediction is possible into more four/five classes. For WoLF PSORT we count any proteins predicted as “vacu”, “lyso”, “E.R.”, “golg” or “plas” as secreted, and any “cyto”, “cysk”, “cyto_nucl” as cytoplasmic and any “nucl” or “cyto_nucl” as nuclear. For SherLoc any sequences predicted as “extracellular”, “ER”, “vacuolar”, “peroxisomal”, “Golgi” or “plasma” are counted as secreted.

On 2009+ SCLpred again performs best of all predictors. On animals Q is 89%, more than 20% better than the second best directly comparable predictor (BaCelLo, with 66.3%), and over 10% better than predictors using one less class. GC, at 79%, is also 10% higher than BaCelLo, and higher than that of the two predictors with one less class. On fungi both Q and GC (72% and 69%) are the highest of all four class predictors, and similar to those obtained by the three class predictors. On plants again Q (at 80%) is by far the highest (SherLoc in this case being the second best five class predictor at 68%), and GC (66%) is at least 9% higher than all other five class predictors, and only lower than Protein Prowler’s (69%) which tackles the simpler four class problem. In this case SCLpred also outperforms BaCelLo by nQ on all three kingdoms.

4 CONCLUSION AND FUTURE WORK

As the amount of sequence information churned out by experimental methods keeps expanding at an ever-increasing pace, it is crucial to develop and make available fast and accurate computational methods to make sense of it. SCL prediction is a step towards bridging the gap between a protein sequence and the protein’s function and can provide information about potential protein-protein interactions and insight into possible drug targets and disease processes. As different SCL predictors are specialized for prediction into different classes and number of classes, and as some predictors are more accurate than others at prediction into any one class, this information can be exploited to lead to more accurate overall consensus predictions, especially if the predictors are diverse in their behavior...

In this article we have developed a new method for SCL prediction (SCLpred) based on a novel Neural Network architecture (N1-NN). The architecture can map a sequence of any length into a set of individual properties for the whole sequence. We have developed three kingdom specific predictors for animals, fungi and plants and predict into four classes for animals and fungi (nuclear, cytoplasm, mitochondrion and the secreted) and an additional fifth class for plants (chloroplast). We have trained SCLpred in 10-fold cross-validation on large non-redundant subsets of annotated proteins from Swiss-Prot 2010/06 and benchmarked it against five other state-of-the-art SCL prediction servers on an independent set of recently annotated proteins. SCLpred performs favorably on these benchmarks, often by consistent margins, and we expect that its prediction accuracy will continue to improve with frequent re-trainings to take advantage of larger, more diverse, datasets of annotated proteins as they become available, and as our understanding of the underlying biological mechanisms improves. We expect larger datasets to be especially beneficial to our models, as these incorporate information from the whole sequence and normally have a higher number of free parameters than the alternatives.

Although here we have only used as input to the network information about the primary sequence and multiple sequence alignments, other residue-level information may be input to the model, such as predicted secondary structure, solvent accessibility, location of predicted binding sites, etc. Incorporating diverse information into the input to SCLpred is one of our future directions of investigation, as is the inclusion of putative homology to “templates”, or proteins of known localization/structure (e.g. by techniques similar to those in Mooney and Pollastri (2009)). In this work we predict subcellular localizations into a small number of classes (4 for animal and fungi, 5 for plants), to allow the comparison of our novel algorithms against a a number of existing predictors, and direct comparison against BaCelLo in particular, which has been shown as one of the best performing ab initio systems to date. We are currently testing our methods on a wider set of localization classes, as well as different functional tasks. A further direction of research is studying the space of f vectors (i.e. compressed, property-driven representations of whole proteins as fixed-size arrays) induced by different output targets (functional classes, protein folds/families), to determine whether they are satisfactory representations towards protein comparison, and whether they yield insights into the structure of the protein space.

SCLpred is available as part of our web servers for protein sequence annotation. Up to 32,768 residues can be handled in a single submission. The servers are freely available for academic users at http://distill.ucd.ie/distill/. Predictions are obtained by an ensemble of all models trained on the 2010/06 training set (as in Table 6). Linux binaries and the benchmarking sets are freely available for academic users upon request.

**FUNDING**

CM is supported by Science Foundation Ireland grant 08/IN.1/B1864. GP’s work is partly funded through SFI RFP grant 10/RFP/GEN2749.

**ACKNOWLEDGEMENTS**

We thank the authors of BaCelLo for making their datasets publicly available, Dr Andrea Pierleoni for assistance with the BaCelLo predictions and Tatyana Goldberg for providing LOCtree predictions.
We wish to acknowledge UCD IT Services for the provision of computational facilities and support.

REFERENCES


