

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

Catherine Mooney<sup>1,3,4</sup>, Yong-Hong Wang<sup>2</sup> and Gianluca Pollastri<sup>1</sup> \*

<sup>1</sup> Complex and Adaptive Systems Laboratory and School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4

<sup>2</sup> Biophysics Institute, Hebei University of Technology, Tianjin, China

<sup>3</sup> Conway Institute of Biomolecular and Biomedical Science, University College Dublin, Belfield, Dublin 4

<sup>4</sup> School of Medicine and Medical Science, University College Dublin, Belfield, Dublin 4

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXXX

## 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 (secretory pathway, 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 2010.06. 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/distill/>.

**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. Most secretory pathway, mitochondrial and chloroplastic proteins have N-terminal peptides (SP, mTP and cTP respectively) but many proteins have no known motif (Emanuelsson, 2002; Nair and Rost, 2005), albeit such motif may exist. However, it would appear that for most proteins the sequence of the protein alone has sufficient information to predict the protein's location in the cell.

There are many methods for the prediction of SCL which can be roughly divided into two groups: homology-based, that rely on similarity to another sequence of known location; 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. The method we describe in this article falls into this latter category.

We predict SCL for eukaryotes only, which we divide into animals, plants and fungi. We consider four subcellular localizations for animals and fungi and five for plants: nucleus, cytoplasm, mitochondrion, chloroplast and the secretory pathway. In a first series of tests we adopt essentially the same experimental setting as in Casadio *et al.* (2008); 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,

\*to whom correspondence should be addressed

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 three eukaryotic kingdoms into four classes for animals and fungi and five classes for plants: secreted, cytoplasm, nucleus, mitochondrion and chloroplast. The predictor is trained on a non-redundant set of experimentally annotated sequences from release 48 of Swiss-Prot. Predictions are made from the full protein sequence, from the N- and C-terminal regions and evolutionary information in the form of a MSA. In Casadio *et al.* (2008) the performance of BaCelLo is benchmarked against LOctree, Protein Prowler, TARGETp and WoLF PSORT with a test set of protein sequences derived from a subset of Swiss-Prot 54. BaCelLo is available at <http://gpcr.biocomp.unibo.it/bacello/>.

**LOctree** Similarly to BaCelLo, LOctree (Nair and Rost, 2005) uses binary SVMs to predict SCL. Three versions of the predictor are available, trained specifically for plants, non-plants and prokaryotes. For prokaryotes predictions are into three classes: secreted, periplasm and cytoplasm. In the case of eukaryotes predictions are into six classes: extracellular space, nucleus, cytoplasm, chloroplast, mitochondrion and other organelles. LOctree is trained on a redundancy reduced subset of release 40 of Swiss-Prot. Predictions are made from the full sequence of the protein, a 50-residue N-terminal region, predicted secondary structure and the output of SIGNALp (for eukaryotes). LOctree is available at <http://cubic.bioc.columbia.edu/services/loctree/>.

**SherLoc** SherLoc (Shatkay *et al.*, 2007) uses SVM which integrate sequence and text-based features. There are three predictors, animal, fungal and 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 release 42.0. <http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc/>.

**TargetP** TargetP (Emanuelsson *et al.*, 2000) uses a feed-forward neural network specialized for the prediction of plant and non-plant SCL into three and four classes respectively, based on the N-terminal amino acid 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/>.

**Protein Prowler** Protein Prowler (Bodén and Hawkins, 2005; Hawkins and Bodén, 2006) is based on the ideas behind TargetP and trained on the same datasets, a redundancy reduced subset of Swiss-Prot releases 37 and 38. The predictor uses a series of neural networks and SVMs specialized for the prediction of plants or non-plants and predicts into the following classes: secretory pathway (presence of a signal peptide), mitochondrion (presence of a mitochondrial targeting peptide), chloroplast (presence of a

**Table 1.** Number of sequences per class for each of the three kingdoms in the BaCelLo training set and the BaCelLo\_2008 test set.

|                  | Animals | Fungi | Plants |
|------------------|---------|-------|--------|
| BaCelLo set      |         |       |        |
| Cytoplasm        | 439     | 211   | 58     |
| Mitochondrion    | 188     | 188   | 67     |
| Nucleus          | 1166    | 711   | 121    |
| Secreted         | 804     | 88    | 41     |
| Chloroplast      |         |       | 204    |
| Total            | 2597    | 1198  | 491    |
| BaCelLo_2008 set |         |       |        |
| Cytoplasm        | 846     | 331   | 102    |
| Mitochondrion    | 241     | 104   | 38     |
| Nucleus          | 979     | 256   | 99     |
| Secreted         | 722     | 26    | 18     |
| Chloroplast      |         |       | 1345   |
| Total            | 2788    | 717   | 1602   |

chloroplast transit peptide) and other. Protein Prowler is available at <http://pprowler.itee.uq.edu.au/>.

**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 amino acid sequence. Based on a number of features (amino acid composition, the presence of known sorting signal and target peptides etc, with different features for animals, fungi and plants) 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://wolffpsort.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 set. We also test this same version of SCLpred on the test dataset used in Casadio *et al.* (2008) (BaCelLo\_2008 set), which is based on Swiss-Prot 54 (Table 1). The BaCelLo\_2008 set is redundancy reduced excluding all sequences with a BLAST hit ( $e = 10^{-3}$ ) to the BaCelLo set. Next we create a new training and test set starting from Swiss-Prot Release 2010.06. From 97,939 Metazoa sequences we find 74,724 sequences with a "SUBCELLULAR LOCATION". We remove membrane proteins and sequences that have non-experimental qualifiers (Potential, Probable, By similarity), leaving 16,406 sequences. We internally redundancy reduce this set using an all-against-all BLAST (Altschul *et al.*, 1997) search (with  $e = 10^{-3}$ ) removing any sequence with a hit with more than 30% sequence identity to any other sequence in the set. All the sequences added to Swiss-Prot earlier than 2009 in the set are used as a training set (2010.06 training set). Sequences added to Swiss-Prot in 2009 or later are used for testing, as these sequences should

**Table 2.** Number of sequences per class for each of the three kingdoms in the 2010\_06 training set and the 2009+ test set.

|                      | Animals | Fungi | Plants |
|----------------------|---------|-------|--------|
| 2010_06 training set |         |       |        |
| Cytoplasm            | 1364    | 890   | 133    |
| Mitochondrion        | 315     | 413   | 81     |
| Nucleus              | 1830    | 1150  | 259    |
| Secreted             | 1584    | 111   | 98     |
| Chloroplast          |         |       | 523    |
| Total                | 5095    | 2564  | 1094   |
| 2009+ test set       |         |       |        |
| Cytoplasm            | 20      | 34    | 8      |
| Mitochondrion        | 5       | 19    | 7      |
| Nucleus              | 25      | 36    | 54     |
| Secreted             | 68      | 15    | 8      |
| Chloroplast          |         |       | 29     |
| Total                | 118     | 104   | 106    |

have less than 30% sequence similarity to any sequences used to train any of the predictors tested in this paper. We refer to as the 2009+ set. no new paragraph? We repeat this process for plants and fungi. Table 2 shows the number of sequences per class for each of the three kingdoms in these new training (2010\_06 set) and test sets (2009+ set).

The BaCelLo datasets are available on the BaCelLo website: <http://gpcr.biocomp.unibo.it/bacello/dataset.htm> and the SCLpred datasets are available upon request from the authors.

**MSA** Multiple sequence alignments are extracted from uniref90 (Suzek *et al.*, 2007) from February 2010 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).

**Input coding** Similarly to Pollastri and McLysaght (2005) the input at each residue is coded as a letter out of an alphabet of 25. Beside the 20 standard amino acids, B (aspartic acid or asparagine), U (selenocysteine), X (unknown), Z (glutamic acid or glutamine) and . (gap) are considered. The input presented to the networks is the frequency of each of the 24 non-gap symbols, plus the total frequency of gaps in each column of the alignment.

## 2.2 Predictive architecture: N1-NN

We will operationally call the model we describe in this work N-to-1 Neural Network, or N1-NN. The model is loosely based on previous models we have developed (e.g. Pollastri *et al.* (2006)) and on our framework to design Neural Networks 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 tackle this problem at the source, that is, they transform/compress the sequence into a fixed number of descriptors (or into descriptors of pairwise relations between sequences) beforehand, and they then 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  $f = (f_1, \dots, f_h)$ , and if we represent the  $i$ -th residue in the sequence as  $r_i$ , then  $f$  is obtained as:

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

where  $\mathcal{N}^{(h)}$  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).  $\mathcal{N}^{(h)}$  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:

$$o = \mathcal{N}^{(o)}(f) \quad (2)$$

where  $\mathcal{N}^{(o)}$  is a non linear function which we implement by a second 2-layered feed-forward neural network (the feature-to-output network). The whole, compound 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  $\mathcal{N}^{(h)}$  for a sequence of length  $N$ , there will be  $N$  contributions to the gradient for this network, which are simply 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 k-mers as soon as  $k > 2 - 3$ . If training is successful, only (soft) motifs relevant to the task at hand will be 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 will generally induce different representations of a sequence.

The number of free parameters in the overall N1-NN can be controlled by: the number of units in the hidden layer of the sequence-to-feature network  $\mathcal{N}^{(h)}()$ ,  $N_f^H$ ; the number of hidden units in the feature-to-output network  $\mathcal{N}^{(o)}()$ ,  $N_o^H$ ; the number of hidden states in the feature vector  $f$ , which is also the number of output units in the sequence-to-feature network,  $N_f$ . Given that only one instance of the sequence-to-feature network (i.e. only one set of free parameters) is replicated for all positions in the sequence, and there is only one feature-to-output network, the overall number of free parameters  $N_p$  of the N1-NN is:

$$N_p = (N_i + 1)N_f^H + (N_f^H + 1)N_f + (N_f + 1)N_o^H + (N_o^H + 1)N_o \quad (3)$$

where  $N_i$  is the size of the input vector representing one residue (including its context) and  $N_o$  is the number of output classes. The number of free parameters can be controlled by  $N_f^H$ ,  $N_f$  and  $N_o^H$ , while  $N_o$  is governed by the property being predicted, and  $N_i$  depends on the input representation and, importantly, by the size of the motifs being considered ( $2c + 1$  in equation 1).

**Training** For each training experiment (i.e. training on the BaCelLo set and training on the 2010\_06 set) we implement three predictors, one for each of the three kingdoms of animals, fungi and plants. Each training is conducted by 10-fold cross-validation, i.e. 10 different sets of training runs are performed in which a different tenth of the overall set is reserved for testing. The 10 tenths are roughly equally sized, disjoint, and their union covers the whole set. For each training the 9/10 of the set that are not

reserved for testing are further split into a validation set (1/10 of the overall set) and a proper training set. Given that some classes are far less numerous than others, in order to rebalance the training set we repeat the number of instances in the various classes until we have roughly the same number of examples in each of them. Examples in the testing and validation sets are not replicated. The training set is used to learn the free parameters of the network by gradient descent, while the validation set is used to monitor the training process. For each different architecture we run three trainings, which differ only in the training vs. validation split. Excluding different validation sets ensures that the resulting models are different, which yields larger gains when ensembling them.

For all sets and all kingdoms we use exactly the same architecture, in which  $N_f^H = N_f = N_o^H = 10$  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, which brings the training error to near zero in all cases. In all cases we save networks at epochs 1800, 1900 and 2000, ensemble average them and evaluate them on the corresponding test set. Saving the models that perform best on validation yields very similar results. The final results for each 10-fold cross-validation (different kingdoms, BaCelLo and 2010.06 set sets) are the average of the results on each test set. When testing on an entirely different set from the one used during training (BaCelLo for training and BaCelLo\_2008 for testing, 2010.06 set for training and 2009+ set for testing) we ensemble-combine *all* the models from all cross-validation folds of the best architecture.

Training is performed by gradient descent on the error, which we model as the relative entropy between the target class and the output of the network. The overall output of the network (output layer of  $\mathcal{N}^{(o)}()$ ) is implemented as a softmax function, while all internal squashing functions are implemented as hyperbolic tangents. The examples are shuffled between epochs. We use a momentum term of 0.9. Although this does not significantly affect the final results, it speeds up overall training times by a factor 10. The learning rate kept fixed at 0.2 throughout training.

### 2.3 Evaluating performance

To evaluate the performance of SCLpred against other state-of-the-art predictors we measure specificity (Spec), sensitivity (Sens), the false positive rate (FPR), Matthews correlation coefficient (MCC) and the generalized correlation (GC) (Baldi et al., 2000) as follows:

$$Spec = 100 \frac{TP}{TP + FP}$$

$$Sens = 100 \frac{TP}{TP + FN}$$

$$FPR = 100 * \frac{FP}{FP + TN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

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

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

(5)

where:

- True positives (TP): the number of sequences predicted in a class that are observed in that class.
- False positives (FP): the number of sequences predicted in a class that are not observed in that class.
- True negatives (TN): the number of sequences predicted not to be in a class that are not observed in that class.

- False negatives (FN): the number of sequences predicted not to be in a class that are observed in that class.
- $z_{ij}$ : the number of times a sequence of class  $i$  is predicted to be in class  $j$
- $e_{ij}$ : the number of sequences of class  $i$  expected to be predicted in class  $j$  by chance
- $N$ : the number of sequences
- $K$ : the number of classes

We emphasize performances based on GC (see Baldi et al. (2000) for more details), as this index minimizes the effect of class sizes, that are not only different between different classes, but in some cases vary considerably even for the same class between different datasets (see below). 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), ESLpred (Bhasin and Raghava, 2004), LOCSVMpsi (Xie et al., 2005), SLP-local (Matsuda et al., 2005), Protein Prowler (Bodén 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 sets (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 0.72 versus 0.67) and fungi (Q 75% versus 70% and GC 0.67 versus 0.66) while the accuracy for plants (Q) is the same (68%) but GC is still considerably higher for SCLpred (0.63 vs. 0.59).

Table 4 shows the accuracy of the same version of SCLpred tested on the full test dataset from Casadio et al. (2008) compared with 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. 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 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).

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 animals and fungi, and in the region of 25% larger for plants. The accuracy of this new version of SCLpred is shown in Table 5, and is broadly similar to that we obtain on the BaCelLo set. We then retest this version of SCLpred 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 (Shatkay et al., 2007), WoLF PSORT, Protein Prowler and TARGETp (Table

**Table 3.** Results for BaCelLo (taken from Pierleoni *et al.* (2006)) and SCLpred, trained and tested in 10-fold cross-validation on the BaCelLo set (Pierleoni *et al.*, 2006), extracted from Swiss-Prot 48.

|           | SCLpred     |      | BaCelLo |      | SCLpred     |      | BaCelLo |      | SCLpred     |      | BaCelLo     |      |
|-----------|-------------|------|---------|------|-------------|------|---------|------|-------------|------|-------------|------|
|           | Spec        | Sens | Spec    | Sens | Spec        | Sens | Spec    | Sens | Spec        | Sens | Spec        | Sens |
|           | Animals     |      |         |      | Fungi       |      |         |      | Plants      |      |             |      |
| Chlo      |             |      |         |      |             |      |         |      | 0.68        | 0.79 | 0.76        | 0.73 |
| Cyto      | 0.58        | 0.54 | 0.41    | 0.65 | 0.46        | 0.39 | 0.39    | 0.60 | 0.39        | 0.36 | 0.47        | 0.52 |
| Mito      | 0.77        | 0.74 | 0.66    | 0.76 | 0.72        | 0.78 | 0.72    | 0.81 | 0.49        | 0.34 | 0.54        | 0.51 |
| Nucl      | 0.83        | 0.85 | 0.85    | 0.65 | 0.83        | 0.82 | 0.85    | 0.67 | 0.83        | 0.76 | 0.76        | 0.72 |
| Secr      | 0.93        | 0.93 | 0.91    | 0.91 | 0.86        | 0.85 | 0.85    | 0.94 | 0.89        | 0.85 | 0.65        | 0.85 |
| <b>GC</b> | <b>0.72</b> |      | 0.67    |      | <b>0.67</b> |      | 0.66    |      | <b>0.63</b> |      | 0.59        |      |
| <b>Q</b>  | <b>0.82</b> |      | 0.74    |      | <b>0.75</b> |      | 0.70    |      | <b>0.68</b> |      | <b>0.68</b> |      |

**Table 4.** Results for SCLpred, trained on the BaCelLo set from Swiss-Prot 48 (Pierleoni *et al.*, 2006), compared to BaCelLo (Pierleoni *et al.*, 2006), LOctree (Nair and Rost, 2005), WoLF PSORT (Horton *et al.*, 2007), Protein Prowler (Hawkins and Bóden, 2006) and TARGETp (Emanuelsson *et al.*, 2000). Tested on the full BaCelLo\_2008 dataset from Swiss-Prot 54 (Casadio *et al.*, 2008). Results for the predictors other than SCLpred taken from Casadio *et al.* (2008).

|           | SCLpred     |      |      | BaCelLo     |      |      | LOctree |      |      | Protein Prowler |      |      | TARGETp |      |      | Wolf PSORT  |      |      |
|-----------|-------------|------|------|-------------|------|------|---------|------|------|-----------------|------|------|---------|------|------|-------------|------|------|
|           | Sens        | MCC  | FPR  | Sens        | MCC  | FPR  | Sens    | MCC  | FPR  | Sens            | MCC  | FPR  | Sens    | MCC  | FPR  | Sens        | MCC  | FPR  |
|           | Animals     |      |      |             |      |      |         |      |      |                 |      |      |         |      |      |             |      |      |
| Cyto      | 0.68        | 0.68 | 0.05 | 0.72        | 0.55 | 0.15 | 0.67    | 0.60 | 0.09 |                 |      |      |         |      |      | 0.65        | 0.59 | 0.09 |
| Mito      | 0.86        | 0.83 | 0.02 | 0.90        | 0.68 | 0.07 | 0.79    | 0.73 | 0.03 | 0.47            | 0.62 | 0.01 | 0.69    | 0.52 | 0.07 | 0.79        | 0.71 | 0.04 |
| Nucl      | 0.92        | 0.78 | 0.12 | 0.62        | 0.58 | 0.08 | 0.80    | 0.65 | 0.14 |                 |      |      |         |      |      | 0.84        | 0.70 | 0.13 |
| Secr      | 0.96        | 0.92 | 0.03 | 0.93        | 0.89 | 0.04 | 0.90    | 0.85 | 0.05 | 0.82            | 0.86 | 0.01 | 0.83    | 0.86 | 0.02 | 0.92        | 0.89 | 0.02 |
| Other     |             |      |      |             |      |      |         |      |      | 0.98            | 0.77 | 0.26 | 0.89    | 0.69 | 0.19 |             |      |      |
| <b>GC</b> | <b>0.81</b> |      |      | 0.70        |      |      | 0.72    |      |      | 0.74            |      |      | 0.7     |      |      | 0.75        |      |      |
| <b>Q</b>  | <b>0.85</b> |      |      | 0.75        |      |      | 0.78    |      |      | 0.89            |      |      | 0.86    |      |      | 0.81        |      |      |
|           | Fungi       |      |      |             |      |      |         |      |      |                 |      |      |         |      |      |             |      |      |
| Cyto      | 0.39        | 0.37 | 0.04 | 0.45        | 0.33 | 0.15 | 0.49    | 0.34 | 0.17 |                 |      |      |         |      |      | 0.27        | 0.35 | 0.03 |
| Mito      | 0.72        | 0.48 | 0.09 | 0.80        | 0.53 | 0.15 | 0.49    | 0.47 | 0.06 | 0.35            | 0.53 | 0.00 | 0.50    | 0.34 | 0.10 | 0.66        | 0.51 | 0.09 |
| Nucl      | 0.85        | 0.49 | 0.43 | 0.66        | 0.39 | 0.26 | 0.68    | 0.34 | 0.32 |                 |      |      |         |      |      | 0.91        | 0.43 | 0.48 |
| Secr      | 0.85        | 0.74 | 0.02 | 1.00        | 0.75 | 0.03 | 0.93    | 0.51 | 0.08 | 0.89            | 0.78 | 0.01 | 0.96    | 0.66 | 0.03 | 0.96        | 0.81 | 0.02 |
| Other     |             |      |      |             |      |      |         |      |      | 0.98            | 0.59 | 0.52 | 0.87    | 0.43 | 0.35 |             |      |      |
| <b>GC</b> | 0.57        |      |      | 0.56        |      |      | 0.44    |      |      | 0.67            |      |      | 0.53    |      |      | <b>0.59</b> |      |      |
| <b>Q</b>  | <b>0.60</b> |      |      | 0.59        |      |      | 0.57    |      |      | 0.89            |      |      | 0.84    |      |      | 0.58        |      |      |
|           | Plants      |      |      |             |      |      |         |      |      |                 |      |      |         |      |      |             |      |      |
| Chlo      | 0.78        | 0.42 | 0.25 | 0.79        | 0.48 | 0.19 | 0.48    | 0.26 | 0.13 | 0.07            | 0.03 | 0.05 | 0.14    | 0.06 | 0.09 | 0.15        | 0.01 | 0.16 |
| Cyto      | 0.63        | 0.63 | 0.02 | 0.43        | 0.32 | 0.06 | 0.78    | 0.52 | 0.08 |                 |      |      |         |      |      | 0.78        | 0.17 | 0.42 |
| Mito      | 0.37        | 0.09 | 0.15 | 0.29        | 0.53 | 0.00 | 0.55    | 0.11 | 0.24 | 0.71            | 0.13 | 0.32 | 0.66    | 0.14 | 0.26 | 0.50        | 0.30 | 0.05 |
| Nucl      | 0.79        | 0.82 | 0.01 | 0.84        | 0.48 | 0.11 | 0.80    | 0.41 | 0.14 |                 |      |      |         |      |      | 0.77        | 0.26 | 0.27 |
| Secr      | 0.83        | 0.48 | 0.02 | 0.94        | 0.42 | 0.04 | 0.83    | 0.56 | 0.02 | 0.79            | 0.32 | 0.06 | 0.83    | 0.35 | 0.05 | 0.33        | 0.15 | 0.04 |
| Other     |             |      |      |             |      |      |         |      |      | 0.83            | 0.23 | 0.49 | 0.83    | 0.22 | 0.50 |             |      |      |
| <b>GC</b> | <b>0.58</b> |      |      | 0.46        |      |      | 0.44    |      |      | 0.24            |      |      | 0.25    |      |      | 0.25        |      |      |
| <b>Q</b>  | <b>0.76</b> |      |      | <b>0.76</b> |      |      | 0.52    |      |      | 0.19            |      |      | 0.24    |      |      | 0.24        |      |      |

6). We obtained results for all the predictors through their respective web servers. We could not obtain results for LOctree in this case.

It is important to note that the 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 than the four animal and fungal and five plant classes which we use for SCLpred. For the purpose of these predictions for WoLF PSORT we count any proteins predicted as “vacu”, “lyso”, “E.R.”, “golgi” or “plas” as secreted, and any cytoplasmic proteins predicted as “cyto”, “cysk”, “cyto\_nucl” as correctly predicted, likewise any nuclear proteins predicted as either “nucl” or “cyto\_nucl” are counted as correct.

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 fewer class. GC, at 79%, is also 10% higher than BaCelLo, and higher than that of the two predictors with one fewer 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

**Table 5.** SCLpred, trained and tested in 10-fold cross-validation on the 2010.06 training set

|      | Animals |      |      | Fungi |      |      | Plants |      |      |
|------|---------|------|------|-------|------|------|--------|------|------|
|      | Spec    | Sens | MCC  | Spec  | Sens | MCC  | Spec   | Sens | MCC  |
| Chlo |         |      |      |       |      |      | 0.74   | 0.83 | 0.56 |
| Cyto | 0.62    | 0.65 | 0.50 | 0.58  | 0.57 | 0.35 | 0.42   | 0.29 | 0.27 |
| Mito | 0.73    | 0.59 | 0.64 | 0.69  | 0.62 | 0.59 | 0.23   | 0.14 | 0.13 |
| Nucl | 0.76    | 0.76 | 0.62 | 0.72  | 0.75 | 0.51 | 0.84   | 0.83 | 0.78 |
| Secr | 0.91    | 0.91 | 0.87 | 0.83  | 0.84 | 0.82 | 0.72   | 0.76 | 0.70 |
| GC   | 0.68    |      |      | 0.63  |      |      | 0.58   |      |      |
| Q    | 0.77    |      |      | 0.67  |      |      | 0.71   |      |      |

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.

#### 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 more SCL predictors become available predictions may be combined through the development of meta servers or consensus prediction methods similar to those developed for protein structure prediction and which have been shown to be successful at CASP. 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 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 (nucleus, cytoplasm, mitochondrion and the secretory pathway) 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 we have developed in Pollastri *et al.* (2007) and Mooney and Pollastri (2009)). 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 webservers for protein sequence annotation. Our server is designed to allow fast and reliable annotation of protein sequences on a genomic-scale: 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/>. 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/B1844. GP’s work is partly funded through SFI RFP grant 10/RFP/GEN2749.

#### ACKNOWLEDGEMENTS

We thank the authors of BaCellLo for making their datasets publicly available and Dr Andrea Pierleoni for assistance with the BaCellLo predictions. The authors wish to acknowledge UCD IT Services, and in particular the Phaeton administrators, for the provision of computational facilities and support.

#### REFERENCES

- Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17), 3389–3402.
- Baldi, P. and Pollastri, G. (2003). The principled design of large-scale recursive neural network architectures – DAG-RNNs and the protein structure prediction problem. *Journal of Machine Learning Research*, **4**, 575–602.
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**(5), 412–424.
- Bhasin, M. and Raghava, G. (2004). ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res*, **32**, W414 – 419.
- Bodén, M. and Hawkins, J. (2005). Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics*, **21**(10), 2279 – 2286.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M., Michoud, K., O’Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, **31**, 365–370.
- Casadio, R., Martelli, P., and Pierleoni, A. (2008). The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief Funct Genomic Proteomic*, **7**(1), 63 – 73.
- Emanuelsson, O. (2002). Predicting protein subcellular localisation from amino acid sequence information. *Brief Bioinform*, **3**(4), 361–376.

**Table 6.** Results for SCLpred, trained on the 2010.06 set, compared to BaCelLo (Pierleoni *et al.*, 2006), SherLoc (Shatkay *et al.*, 2007), WoLF PSORT (Horton *et al.*, 2007), Protein Prowler (Hawkins and Bóden, 2006) and TARGETp (Emanuelsson *et al.*, 2000). Tested on the 2009+ set.

|        | SCLpred     |      |      |      | BaCelLo |      |       |      | SherLoc |      |      |      | Protein Prowler |      |      |      | TARGETp |      |      |      | Wolf PSORT |      |      |      |  |
|--------|-------------|------|------|------|---------|------|-------|------|---------|------|------|------|-----------------|------|------|------|---------|------|------|------|------------|------|------|------|--|
|        | Spec        | Sens | MCC  | FPR  | Spec    | Sens | MCC   | FPR  | Spec    | Sens | MCC  | FPR  | Spec            | Sens | MCC  | FPR  | Spec    | Sens | MCC  | FPR  | Spec       | Sens | MCC  | FPR  |  |
| Animal |             |      |      |      |         |      |       |      |         |      |      |      |                 |      |      |      |         |      |      |      |            |      |      |      |  |
| Cyto   | 0.76        | 0.65 | 0.65 | 0.01 | 0.40    | 0.30 | 0.23  | 0.09 | 0.20    | 0.35 | 0.05 | 0.29 |                 |      |      |      |         |      |      |      |            |      |      |      |  |
| Mito   | 1.00        | 0.60 | 0.77 | 0.01 | 1.00    | 0.60 | 0.77  | 0.00 | 0.75    | 0.60 | 0.66 | 0.01 | 1.00            | 0.80 | 0.89 | 0.00 | 0.36    | 0.80 | 0.51 | 0.06 | 0.50       | 0.40 | 0.43 | 0.02 |  |
| Nucl   | 0.72        | 0.92 | 0.76 | 0.09 | 0.62    | 0.84 | 0.63  | 0.14 | 0.63    | 0.76 | 0.60 | 0.12 |                 |      |      |      |         |      |      |      | 0.69       | 0.80 | 0.67 | 0.10 |  |
| Secr   | 1.00        | 0.97 | 0.97 | 0.02 | 0.95    | 0.91 | 0.85  | 0.06 | 0.90    | 0.65 | 0.55 | 0.10 | 0.98            | 0.60 | 0.60 | 0.02 | 1.00    | 0.65 | 0.66 | 0.00 | 0.94       | 0.88 | 0.80 | 0.08 |  |
| Other  |             |      |      |      |         |      |       |      |         |      |      |      | 0.61            | 0.96 | 0.57 | 0.38 | 0.60    | 0.84 | 0.49 | 0.34 |            |      |      |      |  |
| GC     | <b>0.79</b> |      |      |      | 0.69    |      |       |      | 0.59    |      |      |      | 0.75            |      |      |      | 0.58    |      |      |      | 0.60       |      |      |      |  |
| Q      | <b>0.89</b> |      |      |      | 0.66    |      |       |      | 0.59    |      |      |      | 0.77            |      |      |      | 0.76    |      |      |      | 0.65       |      |      |      |  |
| Fungi  |             |      |      |      |         |      |       |      |         |      |      |      |                 |      |      |      |         |      |      |      |            |      |      |      |  |
| Cyto   | 0.58        | 0.65 | 0.41 | 0.11 | 0.57    | 0.50 | 0.33  | 0.19 | 0.50    | 0.32 | 0.19 | 0.16 |                 |      |      |      |         |      |      |      | 0.60       | 0.08 | 0.13 | 0.03 |  |
| Mito   | 0.79        | 0.79 | 0.74 | 0.04 | 0.71    | 0.89 | 0.75  | 0.08 | 0.71    | 0.53 | 0.54 | 0.05 | 0.89            | 0.42 | 0.56 | 0.01 | 0.61    | 0.58 | 0.51 | 0.08 | 0.71       | 0.79 | 0.69 | 0.7  |  |
| Nucl   | 0.77        | 0.75 | 0.64 | 0.06 | 0.64    | 0.64 | 0.45  | 0.19 | 0.66    | 0.86 | 0.60 | 0.24 |                 |      |      |      |         |      |      |      | 0.54       | 0.94 | 0.50 | 0.43 |  |
| Secr   | 0.92        | 0.73 | 0.79 | 0.01 | 0.86    | 0.80 | 0.80  | 0.02 | 0.52    | 0.73 | 0.54 | 0.11 | 0.80            | 0.80 | 0.77 | 0.03 | 0.86    | 0.80 | 0.80 | 0.02 | 0.73       | 0.73 | 0.69 | 0.04 |  |
| Other  |             |      |      |      |         |      |       |      |         |      |      |      | 0.82            | 0.93 | 0.57 | 0.41 | 0.86    | 0.89 | 0.60 | 0.29 |            |      |      |      |  |
| GC     | <b>0.69</b> |      |      |      | 0.66    |      |       |      | 0.52    |      |      |      | 0.67            |      |      |      | 0.63    |      |      |      | 0.58       |      |      |      |  |
| Q      | <b>0.72</b> |      |      |      | 0.71    |      |       |      | 0.61    |      |      |      | 0.72            |      |      |      | 0.73    |      |      |      | 0.64       |      |      |      |  |
| Plants |             |      |      |      |         |      |       |      |         |      |      |      |                 |      |      |      |         |      |      |      |            |      |      |      |  |
| Chlo   | 0.82        | 0.93 | 0.82 | 0.08 | 0.63    | 0.69 | 0.52  | 0.16 | 0.82    | 0.31 | 0.42 | 0.03 | 1.00            | 0.41 | 0.58 | 0.00 | 0.65    | 0.52 | 0.45 | 0.10 | 0.48       | 0.48 | 0.29 | 0.19 |  |
| Cyto   | 0.30        | 0.38 | 0.27 | 0.02 | 0.75    | 0.38 | 0.51  | 0.01 | 0.20    | 0.50 | 0.23 | 0.16 |                 |      |      |      |         |      |      |      | 0.35       | 0.75 | 0.46 | 0.11 |  |
| Mito   | 0.50        | 0.29 | 0.35 | 0.00 | 0.00    | 0.00 | -0.03 | 0.01 | 0.43    | 0.86 | 0.57 | 0.08 | 0.23            | 0.86 | 0.38 | 0.20 | 0.29    | 0.57 | 0.35 | 0.10 | 0.50       | 0.29 | 0.35 | 0.02 |  |
| Nucl   | 0.89        | 0.87 | 0.75 | 0.13 | 0.80    | 0.91 | 0.68  | 0.23 | 0.95    | 0.74 | 0.72 | 0.04 |                 |      |      |      |         |      |      |      | 0.88       | 0.83 | 0.72 | 0.12 |  |
| Secr   | 1.00        | 0.75 | 0.86 | 0.00 | 0.63    | 0.63 | 0.59  | 0.03 | 0.42    | 1.00 | 0.61 | 0.11 | 1.00            | 0.88 | 0.93 | 0.00 | 0.88    | 0.88 | 0.86 | 0.01 | 0.60       | 0.38 | 0.44 | 0.02 |  |
| Other  |             |      |      |      |         |      |       |      |         |      |      |      | 0.87            | 0.84 | 0.65 | 0.18 | 0.85    | 0.84 | 0.63 | 0.20 |            |      |      |      |  |
| GC     | <b>0.66</b> |      |      |      | 0.54    |      |       |      | 0.57    |      |      |      | 0.69            |      |      |      | 0.62    |      |      |      | 0.50       |      |      |      |  |
| Q      | <b>0.80</b> |      |      |      | 0.52    |      |       |      | 0.68    |      |      |      | 0.75            |      |      |      | 0.70    |      |      |      | 0.55       |      |      |      |  |

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

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

Hawkins, J. and Bóden, M. (2006). Detecting and sorting targeting peptides with recurrent networks and support vector machines. *Journal of Bioinformatics and Computational Biology*, **4**(1), 1 – 18.

Horton, P., Park, K., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C., and Naka, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res*, **35**, W5857.

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

Matsuda, S., Vert, J., 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 Sci*, **14**, 2804 – 2813.

Mooney, C. and Pollastri (2009). Beyond the twilight zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information. *Proteins*, **77**(1), 181–90.

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.

Pierleoni, A., Martelli, P. L., Fariselli, P., and Casadio, R. (2006). BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **422**(14), 408 – 416.

Pollastri, G. and McLysaght, A. (2005). Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, **21**(8), 1719–20.

Pollastri, G., Vullo, A., Frasconi, P., and Baldi, P. (2006). Modular DAG-RNN architectures for assembling coarse protein structures. *Journal of Computational Biology*, **13**(3), 631–650.

Pollastri, G., Martin, A., Mooney, C., and Vullo, A. (2007). Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics*, **8**(201), 12.

Shatkay, H., Höglund, A., Brady, S., Blum, T., Dönnies, P., and Kohlbacher, O. (2007). SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics*, **23**(11), 1410–1417.

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.

Suzek, B., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**(10), 1282.

Walsh, I., Vullo, A., and Pollastri, G. (2009). Recursive neural networks for undirected graphs for learning molecular endpoints. In *Pattern Recognition in Bioinformatics*, volume 5780 of *Lecture Notes in Bioinformatics*. Springer Verlag.

Xie, D., Li, A., Wang, M., Fan, Z., and Feng, H. (2005). LOCSVMP: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res*, **33**, W105 – W110.