De Novo Protein Subcellular Localization Prediction by N-to-1 Neural Networks

Catherine Mooney[1] and Yong-Hong Wang[2] and Gianluca Pollastri[1⋆]

[1] Complex and Adaptive Systems Laboratory and School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4
[2] Biophysics Institute, Hebei University of Technology, Tianjin, China

Abstract. 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 (SCL_pred) 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 high throughput machine learning techniques trained on large non-redundant sets of protein sequences. The algorithm powering SCL_pred is a novel Neural Network (N-to-1 Neural Network, or N1-NN) 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 SCL_pred against other publicly available predictors using two benchmarks including a new subset of Swiss-Prot release 57. We show that SCL_pred compares favourably to the other state-of-the-art predictors. Moreover, 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 even shed light on the space of protein sequences. The predictive systems described in this paper are publicly available at http://distill.ucd.ie/distill/.

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

⋆ To whom correspondence should be addressed

adjacent compartments, at least temporarily. At present 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 more 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 eukaryotics protein may be located. Here we consider four for animals and fungi and five for plants: nucleus, cytoplasm, mitochondria, chloroplast and the secretory pathway. Some nuclear proteins have a nuclear localisation 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 [8, 16]. 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 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. In a first series of tests we adopt essentially the same experimental setting and 4/5 location classes as in [6, 18], 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 57 [5], and benchmark SCL_pred on these sets against five state-of-the-art, publicly available predictors of SCL: BaCelLo, LOCtree, Protein Prowler, TARGETp and WoLF PSORT.

BaCelLo BaCelLo [18] uses a hierarchy of binary SVMs 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 [6] 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 [16] 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/`.

Protein Prowler Protein Prowler [4, 10] 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 specialised 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 chloroplast transit peptide) and other. Protein Prowler is available at `http://pprowler.itee.uq.edu.au/`.

TargetP TargetP [7] uses a feed-forward neural network specialised 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/`.

WoLF PSORT WoLF PSORT [11] 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 neighbour 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, mitochondria, nuclear, peroxisome, plasma membrane, vacuolar membrane. WoLF PSORT is available at `http://wolfpsort.org/`.

2 Materials and Methods

2.1 Datasets

The first dataset which we use to train and test SCL_pred is the dataset used by [18] to train BaCelLo in ten-fold cross validation, for a direct comparison with this predictor. We call this set the BaCelLo set. We also test this version of SCL_pred on the test datasets used in [6] (BaCelLo_2008 set), which is based on Swiss-Prot 54. Next we combine the BaCelLo and BaCelLo_2008 sets and redundancy reduce the union by an all-against-all PSI-BLAST [1] search (with $e = 10^{-3}$) removing any sequence with a hit with more than 30% sequence

identity to any other sequence in the set. Table 1 shows the number of sequences per class for each of the three kingdoms in this new set (BaCelLo_union set).

Using the BaCelLo_union set we re-train SCL_pred in ten-fold cross validation and test it on a independent set extracted from Swiss-Prot 57 (SP_57 set). To create this we first remove from Swiss-Prot 57 any protein present in Swiss-Prot 54 (from which BaCelLo_union is obtained), which leaves 203,860 sequences out of the original 462,764 entries. We then search for metazoa, fungi and viridiplante with an appropriate SCL, that is entries with the keywords "nucleus", "cytoplasm","mitochondrion", "Plastid, chloroplast" or "secreted" in the SUBCELLULAR LOCATION subfield. We exclude membrane proteins, entries with multiple keywords and non-experimental qualifiers (Potential, Probable, By similarity) and sequences with fewer than 30 residues. Then we PSI-BLAST the remaining sequences against Swiss-Prot 54 with $e = 10^{-3}$ and remove any sequences with a hit with more than 30% sequence identity to any sequence in Swiss-Prot 54. Finally we run an internal redundancy reduction on the remaining sequences, removing any entry with more than 90% sequence identity to another sequence in the set. Table 2 shows the number of sequences per class for each of the three kingdoms.

All the BaCelLo datasets are publicly available on the BaCelLo website: `http://gpcr.biocomp.unibo.it/bacello/dataset.htm`.

Table 1. Number of sequences per class for each of the three kingdoms in the BaCelLo_union set. See text for details.

	Animals	Fungi	Plants
Cytoplasm	689	466	89
Mitochondrion	263	271	72
Nucleus	1488	884	155
Secreted	881	881	48
Chloroplast			277
Total	3321	1717	641

MSA Multiple sequence alignments are extracted from a redundancy reduced, 2004 version of the NR dataset containing over 1 million sequences. The alignments are generated by three runs of PSI-BLAST with parameters $b = 3000$ (maximum number of hits), $e = 10^{-3}$ (expectation of a random hit) and $h = 10^{-10}$ (expectation of a random hit for sequences used to generate the PSSM).

Input coding Similarly to in [20] 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)

Table 2. Number of sequences per class for each of the three kingdoms in the SP_57 set. See text for details.

	Animals	Fungi	Plants
Cytoplasm	29	82	1
Mitochondrion	6	55	9
Nucleus	78	84	65
Secreted	107	2	3
Chloroplast			18
Total	220	223	96

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. [21]) and on our framework to design Neural Networks for structured data [2, 23]. In this case, instead of compressing all the information of a sequence into a handful of predefined features (e.g. k-mer frequencies, sequence length, etc.), we decide beforehand only *how many* features we want to compress a sequence into. If these features are stored in a vector $f = (f_1, \ldots, f_h)$, and if we represent the i-th residue in the sequence as r_i, then f is obtained as:

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

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. $\mathcal{N}^{(h)}$ is replicated N times (N being the sequence length), and k is a normalisation constant. Notice that the feature vector f is obtained by combining information coming directly from all windows of $2c + 1$ residues in the protein, and is based on motifs that may be fairly long (e.g. if $c = 10$, as in all the tests in this article, the motifs have a length of 21 residues). The feature vector f thus obtained, is mapped into the property of interest o (for instance, cellular component class), as follows:

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

where $\mathcal{N}^{(o)}$ is a non linear function which we implement by another 2-layered feed-forward neural network. The whole, compound neural network (the cascade of N sequence to feature vector networks and one feature vector to output network) is itself a feed-forward neural network, thus can 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. Notice that the feature vector f is not a predefined transformation/compression of the sequence, but instead is automatically learned in order to minimise 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 has only a relatively small number of free parameters to represent them, hence does not suffer from over-parametrisation 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, and the compression is property-driven.

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:

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

where N_i is the size of the input vector representing one residue and N_o is the number of output classes.

Training, Ensembling For each training experiment (i.e. training on the BaCelLo set and training on the BaCelLo_union 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. The training set is used to learn the free parameters of the network by gradient descent, while the validation set is used to choose model and hyperparameters (network size and architecture, i.e. N_f^H, N_f and N_o^H). For each different architecture we run three trainings, which differ only in the training vs. validation split. We choose the architecture which performs best on validation. For each fold the three networks for the best architecture are ensemble averaged and evaluated on the corresponding test set. The final results for each 10-cross validation (different kingdoms, BaCelLo and BaCelLo_union 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, BaCelLo_union for training and SP_57 for testing) we ensemble-combine

all the models from all cross-validation folds of the best architecture. Table 3 shows details of network size and training times for each of the three predictors.

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. Training terminates when either the walltime on the server is reached (6 days for fungi and plants, 10 days for animals) or the epoch limit is reached (40k for fungi and plants and 20k for animals). The gradient is updated 360 times for each epoch (or once every 2-6 examples, depending on the set), and the examples are shuffled between epochs. The learning rate is halved every time a reduction of the error is not observed for more than 50 epochs. Models are saved at regular intervals (every 100 epochs) during training. When training is complete the model with the best performance on the validation set is chosen to be part of the final ensemble for each predictor.

Table 3. Network size and training times for the three network architectures. N_f: size of the feature vector; N_o^H: number of hidden units in the feature-to-output network; N_f^H: number of hidden units in the sequence-to-feature network.

	Animals	Fungi	Plants
N_f	12	10	8
N_o^H	11	7	4
N_f^H	13	11	9
Epochs limit	20k	40k	40k
Walltime	10 days	6 days	6 days

2.3 Evaluating performance

We measure accuracy/specificity (Cov) and coverage/sensitivity (Acc) per class as in [16, 4, 7, 18] and the geometric average (GAv) as in [16, 18]:

$$\begin{aligned} Cov &= 100\frac{TP}{TP+FP} \\ Acc &= 100\frac{TP}{TP+FN} \\ GAv &= \sqrt{Acc.Cov} \end{aligned} \tag{4}$$

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.
– False negatives (FN): the number of sequences predicted not to be in a class that are observed in that class.

The overall accuracy of the predictors is measured by Q%:

$$Q\% = 100\frac{\text{number of proteins correctly predicted}}{\text{number of proteins in data set}} \tag{5}$$

3 Results and Discussion

Table 4. Q% for BaCelLo and SCL_pred, trained and tested in ten-fold cross validation on the BaCelLo set [18], extracted from Swiss-Prot 48.

	Animals		Fungi		Plants	
	BaCelLo	SCL_pred	BaCelLo	SCL_pred	BaCelLo	SCL_pred
Cytoplasm	41.4	44.4	39.4	36.0	46.9	46.6
Mitochondrion	66.2	58.5	69.5	67.6	54.0	16.4
Nucleus	84.9	84.8	87.0	88.8	75.7	75.2
Secreted	90.7	90.1	76.9	86.4	64.8	85.4
Chloroplast					76.4	83.4
Q	73.8	**77.7**	70.1	**76.0**	**68.2**	68.0

Table 5. Q% for SCL_pred compared to BaCelLo [18], LOCtree [16], WoLF PSORT [11], Protein Prowler [10] and TARGETp [7]. Tested on the full and reduced (in brackets) BaCelLo_2008 dataset (from Swiss-Prot 54) (see [6] for details). Results for the predictors other than SCL_pred taken from [6].

	Animals		Fungi		Plants	
Predictor	3 Class	4 Class	3 Class	4 Class	4 Class	5 Class
SCL_pred	**92** (89)	**82** (**74**)	78 (79)	55 (52)	**85** (69)	**84** (67)
BaCelLo	89 (**91**)	75 (64)	82 (84)	**59** (**57**)	77 (**76**)	76 (69)
LOCtree	90 (81)	78 (62)	81 (75)	57 (47)	53 (**76**)	52 (**70**)
WoLF PSORT	91 (88)	81 (71)	86 (82)	58 (51)	25 (69)	24 (57)
PProwler	89 (**91**)		**89** (**86**)		19 (63)	
TARGETp	86 (88)		84 (82)		24 (67)	

In previous tests BaCelLo [18] was shown to outperform the following publicly available methods for the prediction of the subcellular localization: Loctree [16], Psort II [17], SubLoc [12], ESLpred [3], LOCSVMpsi [24], SLP-local [13], Protein Prowler [4], TARGETp [7], PredoTar [22] and pTARGET [9]. In Table 4 we show

the performance of SCL_pred compared to BaCelLo on the BaCelLo sets [18]. Both predictors are assessed by ten-fold cross-validation. Overall SCL_pred is more accurate for animals (77.7% versus 73.8%) and fungi (76% versus 70.1%) while the accuracy for plants is similar (68% versus 68.2%). The accuracy per class differs somewhat, with BaCelLo being more accurate for mitochondrial proteins, especially in the case of plants where the SCL_pred prediction is poor (16.4% versus 54%). However SCL_pred is more accurate for secreted proteins in fungi and plants by 9.5% and 20.6% respectively. Prediction accuracy is similar for proteins in the cytoplasm and nucleus and SCL_pred is again more accurate for chloroplastic proteins by 7%.

Table 5 shows the accuracy of the same version of SCL_pred tested on the two test datasets from [6] compared with the other five SCL predictors tested on the same dataset (results from [6]). SCL_pred performs well for animals and plants (better than all the other servers in 5 out of 8 cases), however it performs less well on fungi. We interpret these mixed results as a consequence of the small size of the sets: given we take into account the whole, unprocessed sequence, rather than a handful of features extracted from it, the networks we use have at least a few thousand adjustable parameters and SCL_pred is more prone to overfitting the training set than the other systems.

To check whether larger datasets may alleviate the problem, we repeat the experiments on the BaCelLo_union set, which is approximately 30% larger than the BaCelLo set (3321 proteins for Animals, 1717 for fungi, 641 for plants). The accuracy of this new version of SCL_pred is shown in Table 6. We then retest this version of SCL_pred on the SP_57 (a subset of Swiss-Prot 57, described in the dataset section) and again compare its accuracy with BaCelLo, LOCtree, WoLF PSORT, Protein Prowler and TARGETp (Table 7). We obtained results for WoLF PSORT and Protein Prowler through their respective web servers, and results for TARGETp were obtained by downloading the stand alone version of TARGETp available from the TARGETp website, which we then ran locally, hence we have no control on the sequence identity cutoffs between the training sets of these predictors and SP_57. BaCelLo results were kindly provided by Dr Pierleoni. We could not obtain results for LOCtree in this case. In five out of the six cases SCL_pred is the most accurate predictor overall (Table 7).

In Table 8 we show a more detailed analysis of these results for SCL_pred, BaCelLo and WoLF PSORT. It is important to note that due to efforts to reduce redundancy between this new subset of Swiss-Prot 57 and Swiss-Prot 54 (used in training) the number of samples per class is very small in some instances (only one sequence for plant cytoplasm, two for secreted proteins in fungi and three in plants and only six and nine animal and plant mitochondrial proteins respectively).

The most accurately predicted classes for each predictor are the classes with the greatest number of examples: nucleus and secreted in animals; cytoplasm, mitochondrion and nucleus in fungi; and nucleus and chloroplast in plants. Overall SCL_pred continues to perform well, comfortably outperforming BaCelLo and WoLF PSORT in the most densely populated classes for plants and fungi

(nucleus and chloroplast, and nucleus and cytoplasm respectively) and also performing well for mitochondrial proteins in fungi. Performance of the animal predictor is more mixed, with none of the three predictors performing well in the less densely populated classes of cytoplasm and mitochondrion. In the other two classes of nuclear and secreted proteins the performance of the three predictors for coverage, accuracy and geometric average is the same for SCL_pred and BaCelLo (75%) when averaged across these three measures for the two classes and 79% for WoLF PSORT. The overall Q performance is slightly better for SCL_pred (68.6%) than for WoLF PSORT (68.2%) and BaCelLo (66.8%). Given the small size (96-223 proteins) of these sets, and their unbalanced nature, further testing on larger, more balanced sets would be desirable when such sets become available.

We also test the accuracy of a consensus prediction between SCL_pred, BaCelLo and WoLF PSORT. The combination of several prediction methods has been used successfully in many cases, for instance for structure predictions at CASP [15]. Here we take a majority vote between the three predictors, and where there is a tie (i.e. each of the three predictors predicts a different class) we trust SCL_pred. The consensus predictor is more accurate for the animal predictor but SCL_pred is more accurate than the consensus for fungi and plants. We do consider that this is an area worth further investigation and a SCL meta-server may be of use to the community of biologists.

Table 6. Coverage (Cov), accuracy (Acc) and geometric average (GAv) per class for SCL_pred, trained and tested in ten-fold cross validation on the combined and redundancy reduced datasets from [18] and [6] (Swiss-Prot 48 and 54).

	Animals			Fungi			Plants		
	Cov	Acc	GAv	Cov	Acc	GAv	Cov	Acc	GAv
Cytoplasm	53.1	50.5	51.8	60.5	47.0	53.3	46.9	42.7	44.8
Mitochondrion	65.6	64.6	65.1	69.2	66.4	67.8	46.7	19.4	30.1
Nucleus	79.2	81.5	80.3	72.8	82.4	77.4	77.6	78.1	77.8
Secreted	88.8	88.1	88.4	88.4	87.5	88.0	72.7	66.7	69.6
Chloroplast							66.7	79.4	72.8
Q	75.5			70.5			66.3		

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

Table 7. Q(%) for SCL_pred compared to BaCelLo [18], WoLF PSORT [11], Protein Prowler [10] and TARGETp [7] tested on the SP_57 set.

	Animals		Fungi		Plants	
Predictor	3 Class	4 Class	3 Class	4 Class	4 Class	5 Class
SCL_pred	84.5	**68.6**	**89.2**	**68.6**	**82.3**	**82.3**
BaCelLo	**90.0**	66.8	87.9	57.4	76.0	76.0
WoLF PSORT	83.6	68.2	75.8	52.9	71.9	67.7
PProwler	70.5		82.5		77.1	
TARGETp	65.0		80.7		71.9	

Table 8. Coverage (Cov), accuracy (Acc) and geometric average (GAv) per class for the three four-class animal and fungus predictors and the five-class plant predictors tested on the new subset of Swiss-Prot 57.

Animals

	Cytoplasm			Mitochondrion			Nucleus			Secreted			Q
Predictor	Cov	Acc	GAv	Cov	Acc	GAv	Cov	Acc	GAv	Cov	Acc	GAv	
SCL_pred	**25.0**	27.6	**26.3**	25.0	16.7	20.4	65.3	62.8	64.1	85.3	86.9	86.1	68.6
BaCelLo	17.7	**31.0**	23.4	30.0	**50.0**	38.7	67.3	47.4	56.5	94.2	**91.6**	**92.9**	66.8
WoLF PSORT	18.9	24.1	21.4	16.7	33.3	23.6	**72.2**	**66.7**	**69.4**	**95.7**	83.2	89.2	68.2
Consensus	21.2	24.1	22.6	**50.0**	**50.0**	**50.0**	68.5	64.1	66.3	89.8	90.7	90.2	**71.4**

Fungi

	Cytoplasm			Mitochondrion			Nucleus			Secreted			Q
Predictor	Cov	Acc	GAv	Cov	Acc	GAv	Cov	Acc	GAv	Cov	Acc	GAv	
SCL_pred	**75.6**	**41.5**	**56.0**	**84.0**	76.4	80.1	**61.1**	**91.7**	**74.8**	0	0	0	**68.6**
BaCelLo	48.2	32.9	39.8	78.9	81.8	80.4	52.9	64.3	58.3	**25.0**	**100**	**50.0**	57.4
WoLF PSORT	46.8	26.8	35.4	77.3	61.8	69.1	57.5	72.6	64.6	12.5	50.0	25.0	52.9
Consensus	73.0	32.9	49.0	81.4	**87.3**	**84.3**	60.7	88.1	73.1	20.0	50.0	31.6	67.3

Plants

	Cytoplasm			Mitochondrion			Nucleus			Secreted			Chloroplast			Q
Predictor	Cov	Acc	GAv	Cov	Acc	GAv	Cov	Acc	GAv	Cov	Acc	GAv	Cov	Acc	GAv	
SCL_pred	0.0	0.0	0.0	0.0	0.0	0.0	**96.9**	**95.4**	**96.1**	**100**	**66.7**	**81.6**	**53.6**	**83.3**	**66.8**	**82.3**
BaCelLo	0.0	0.0	0.0	0.0	0.0	0.0	88.1	90.8	89.4	50.0	**66.7**	57.7	50.0	66.7	57.7	76.0
WoLF PSORT	0.0	0.0	0.0	**50.0**	**22.2**	**33.3**	93.3	86.1	89.7	0.0	0.0	0.0	33.3	38.9	36.0	67.7
Consensus	0.0	0.0	0.0	0.0	0.0	0.0	96.8	92.3	94.5	100.0	**66.7**	**81.6**	50.0	**83.3**	64.5	80.2

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 specialised 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 behaviour.

In this article we have developed a new method for SCL prediction (SCL_pred) 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, mitochondria and the secretory pathway) and an additional fifth class for plants (chloroplast). We have trained SCL_pred in ten-fold cross-validation on two large non-redundant subsets of annotated proteins from Swiss-Prot releases 48 and 54 and benchmarked them against five other state-of-the-art SCL prediction servers on independent sets. SCL_pred performs favourably on these benchmarks 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 SCL_pred is one of our future directions of investigation, as it is the inclusion of putative homology to "templates", or proteins of known localisation/structure (e.g. by techniques similar to those we have developed in [19, 14]). 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.

SCL_pred 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

This work is supported by Science Foundation Ireland grants 10/RFP/GEN2749 and 05/RFP/CMS0029 and grant RP/2005/219 from the Health Research Board of Ireland.

Acknowledgements

We thank authors of BaCelLo for making their datasets publicly available and in particular we thank Dr Andrea Pierleoni for providing the BaCelLo predictions.

References

1. S Altschul, T Madden, A Schäffer, J Zhang, Z Zhang, W Miller, and D Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997.
2. P. Baldi and G. Pollastri. 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, 2003.
3. M. Bhasin and G.P. Raghava. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res*, 32:W414 – 419, 2004.
4. M. Bóden and J. Hawkins. Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics*, 21(10):2279 – 2286, 2005.
5. B Boeckmann, A Bairoch, R Apweiler, M Blatter, A Estreicher, E Gasteiger, M Martin, K Michoud, C O'Donovan, I Phan, S Pilbout, and M Schneider. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 31:365–370, 2003.
6. R Casadio, PL Martelli, and A Pierleoni. The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief Funct Genomic Proteomic*, 7(1):63 – 73, 2008.
7. O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*, 300:1005 – 1016, 2000.
8. Olof Emanuelsson. Predicting protein subcellular localisation from amino acid sequence information. *Brief Bioinform*, 3(4):361–376, 2002.
9. C. Guda and S. Subramaniam. pTARGET: a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics*, 21:3963 – 3969, 2005.
10. J. Hawkins and M. Bóden. Detecting and sorting targeting peptides with recurrent networks and support vector machines. *Journal of Bioinformatics and Computational Biology*, 4(1):1 – 18, 2006.
11. P Horton, KJ Park, T Obayashi, N Fujita, H Harada, C.J. Adams-Collier, and K Naka. WoLF PSORT:protein localization predictor. *Nucleic Acids Res*, 35:W5857, 2007.
12. S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17:721 – 728, 2001.

13. S. Matsuda, J.P. Vert, H. Saigo, N. Ueda, H. Toh, and T. Akutsu. A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci*, 14:2804 – 2813, 2005.
14. C. Mooney and Pollastri. Beyond the twilight zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information. *Proteins*, 77(1):181–90, 2009.
15. John Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol*, 15(3):285–289, 2005.
16. R. Nair and B Rost. Mimicking cellular sorting improves prediction of subcellular localization. *J Mol Biol*, 348:85 – 100, 2005.
17. K. Nakai and P Horton. PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem Sci*, 24:34 – 35, 1999.
18. Andrea Pierleoni, Pier Luigi Martelli, Piero Fariselli, and Rita Casadio. BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, 422(14):408 – 416, 2006.
19. G. Pollastri, A.J.M. Martin, C. Mooney, and A. Vullo. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics*, 8(201):12, 2007.
20. G Pollastri and A McLysaght. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8):1719–20, 2005.
21. G. Pollastri, A. Vullo, P. Frasconi, and P. Baldi. Modular DAG-RNN architectures for assembling coarse protein structures. *Journal of Computational Biology*, 13(3):631–650, 2006.
22. I. Small, N. Peeters, F. Legeai, and C. Lurin. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, 6:1581 – 1590, 2004.
23. I. Walsh, A. Vullo, and G. Pollastri. Recursive neural networks for undirected graphs for learning molecular endpoints. In *Pattern Recognition in Bioinformatics*, volume 5780 of *Lecture Notes in Bioinformatics*. Springer Verlag, 2009.
24. D. Xie, A. Li, M. Wang, Z. Fan, and H. Feng. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res*, 33:W105 – W110, 2005.