Title	SCLpred-EMS: Subcellular localization prediction of endomembrane system and secretory pathway proteins by Deep N-to-1 Convolutional Neural Networks
Authors(s)	Kaleel, Manaz; Zheng, Yandan; Chen, Jialiang; Feng, Xuanming; Simpson, Jeremy C.; Pollastri, Gianluca; Mooney, Catherine
Publication date	2020-06
Publication information	Bioinformatics, 36 (11): 3343-3349
Publisher	Oxford University Press
Item record/more information	http://hdl.handle.net/10197/12182
Publisher's statement	This is a pre-copyedited, author-produced PDF of an article accepted for publication in Bioinformatics following peer review. The definitive publisher-authenticated version Manaz Kaleel, Yandan Zheng, Jialiang Chen, Xuanming Feng, Jeremy C Simpson, Gianluca Pollastri, Catherine Mooney, SCLpred-EMS: subcellular localization prediction of endomembrane system and secretory pathway proteins by Deep N-to-1 Convolutional þÿ N e u r a l N e t w o r k s , B i o i n f o r m a t i c s , V o l u m e 3 6 , I s s u e 1 1 , J u n e available online at: https://doi.org/10.1093/bioinformatics/btaa156.
Publisher's version (DOI)	10.1093/bioinformatics/btaa156

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

SCLpred-EMS: subcellular localization prediction of endomembrane system and secretory pathway proteins by Deep N-to-1 Convolutional Neural Networks

Manaz Kaleel [1,2], Zheng Yandan [3], Chen Jialiang [3], Feng Xuanming [3], Jeremy C. Simpson [4,5], Gianluca Pollastri [1,2] and Catherine Mooney [1,3]*

[1]School of Computer Science, University College Dublin, Dublin, Ireland
[2]UCD Institute for Discovery, University College Dublin, Dublin, Ireland
[3]Beijing-Dublin International College, Beijing University of Technology, Chaoyang, China
[4]Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland
[5]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The subcellular location of a protein can provide useful information for protein function prediction and drug design. Experimentally determining the subcellular location of a protein is an expensive and time-consuming task. Therefore, various computer-based tools have been developed, mostly using machine learning algorithms, to predict the subcellular location of proteins.
Results: Here, we present a neural network based algorithm for protein subcellular location prediction. We introduce SCLpred-EMS a subcellular localization predictor powered by an ensemble of Deep N-to-1 Convolutional Neural Networks. SCLpred-EMS predicts the subcellular location of a protein into two classes, the endomembrane system and secretory pathway versus all others, with an MCC of 0.75-0.86 outperforming the other state-of-the-art web servers we tested.
Availability: SCLpred-EMS is freely available for academic users at http://distilldeep.ucd.ie/SCLpred2/
Contact: catherine.mooney@ucd.ie
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The subcellular location of proteins is considered crucial information for biological and medical research (Schwikowski *et al.*, 2000; Rajendran *et al.*, 2010; Kawai *et al.*, 1992). This biological and medical significance of protein subcellular location is due to the fact that protein location affects or determines protein interactions, accessibility to drug molecules, the potential for vaccines and countless other applications. Even though experimental protein subcellular location determination has significantly progressed in the last two decades, experimental methods are still expensive and time-consuming. As an alternative to experimental methods, in recent decades, computational prediction of protein subcellular location has been particularly active (Pierleoni *et al.*, 2006, 2011a,b; Yu *et al.*, 2014; Savojardo *et al.*, 2015, 2018).

The availability of accurately annotated large protein subcellular location databases such as UniProt (UniProt Consortium, 2019) makes this an attractive problem for machine learning algorithms. In the past fifteen years, most of the predictors of protein subcellular localization have been based on machine learning algorithms, from support vector machines (SVM) (Yu *et al.*, 2006) to Deep Neural Networks (Armenteros *et al.*, 2017). Even though many protein subcellular location prediction systems were introduced in the recent past, we found that the performance of these systems deteriorates considerably when an increasing level of homology reduction is introduced into their testing sets. This suggests that

the machine learning models predicting subcellular localization require stricter homology reduction to genuinely learn how to generalise.

2 Approach

Here we introduce SCLpred-EMS, an *ab-inito* protein subcellular location predictor powered by an ensemble of Deep N-to-1 Convolutional Neural Networks. The system was trained on a recent UniProt Knowledgebase (UniProtKB/Swiss-Prot) release 2019_06 (UniProt Consortium, 2019) and the results were measured in five-fold cross-validation. SCLpred-EMS predicts eukaryotic proteins into two classes, the endomembrane system and secretory pathway (EMS) versus all others, from the amino acid sequence. We employed a novel homology reduction protocol for stricter homology reduction and used an in lab encoding scheme (Kaleel *et al.*, 2019) that has led to significant performance improvements in similar predictive tasks. We comprehensively benchmarked SCLpred-EMS against other freely available state-of-the-art web servers published in the last five years and compatible for recasting the predicted classes into EMS versus others. We used a strict independent test set of 216 sequences and a less strict independent test set of 593 proteins for benchmarking. We show that SCLpred-EMS compares favourably with other state-of-the-art predictors on these sets. All of the predictors benchmarked use some form of machine learning in the core of the prediction algorithm. The benchmarked predictors and their prediction algorithms are: DeepLoc (Armenteros *et al.*, 2017) which uses a combination of convolutional neural networks (CNN), bidirectional long short term memory (LSTM), attention mechanism and hierarchical tree sorting pathways; LocTree3 (Goldberg *et al.*, 2014) which employs SVM and homology; and SCL-Epred (Mooney *et al.*, 2013) which employs an N-to-1 neural network.

3 Methods

3.1 Datasets

All eukaryotic entries were downloaded from the UniProt Knowledgebase (UniProtKB/Swiss-Prot) release 2019_06 (UniProt Consortium, 2019) – 189,818 proteins from 7,859 species. Approximately half of these proteins came from three species: Homo sapiens (Human), Mus musculus (Mouse) and Arabidopsis thaliana (Mouse-ear cress). 154,743 of these proteins were labelled with a subcellular location; we removed any proteins that were not supported by published experimental evidence (ECO code ECO:0000269), leaving 28,430 proteins. This set was split into two groups: the endomembrane system and secretory pathway; and "other" as shown in Figure 1. 3,531 proteins that had annotations which could not unambiguously be assigned to either EMS or "other" were removed. For example, proteins labelled as "Nuclear Envelope" could arguably be either EMS or "other", proteins that did not have subcellular location information other than "single-pass" or "multi-pass membrane", or proteins with subcellular location information for both EMS and "other" locations. Finally, 10,223 sequences remained in the EMS class and 14,676 in the "other" class.

This set of proteins was further reduced by removing any protein that was less than 30 amino acids or greater than 10,000 amino acids in length. The resulting set was then internally redundancy reduced, removing sequences so that no two sequences were more than 80% similar to each other leaving 20,722 sequences. A subset of sequences that were added to UniprotKB after 2016 was set aside in the 80% Independent Test Set (ITS-80). The remaining 19,579 sequences after removal of the ITS-80 became the Train-80 and were divided into five parts. These five parts were used to form five folds that are used for five-fold cross-validation.

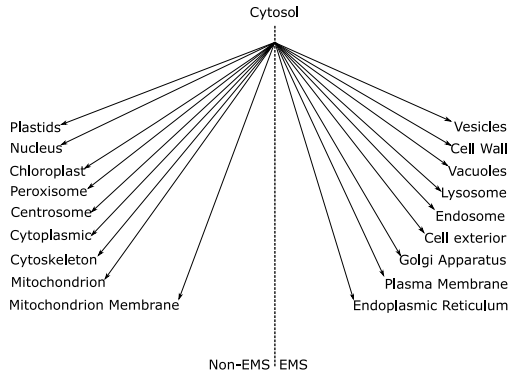

Fig. 1: Mapping of localization classes into EMS and other.

For each of the five-fold subdivisions, one of the five parts was assigned as the test set, a second part was assigned as validation set and the remaining three parts were combined to create the training set. The rest of the folds were formed by repeating the same procedure (See Figure 2). Table 1 shows the number of proteins in each fold. In the resulting folds, the validation set was redundancy reduced with respect to the test set using BLAST (Altschul *et al.*, 1997) with an e-value of 0.001. Then the test and validation sets were internally redundancy reduced using BLAST with an e-value of 0.001. Finally, the test set and validation sets were redundancy reduced with respect to their corresponding training set using BLAST with an e-value of 0.001.

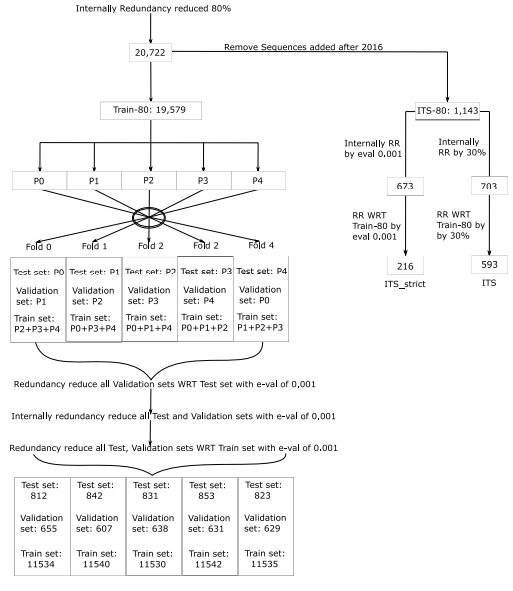

Fig. 2: Figure showing the redundancy reduction process. (RR: Redundancy Reduction; WRT: With Respect To.)

Two subsets of ITS-80 were used to benchmark other available web servers. The ITS-80 set was internally redundancy reduced and redundancy reduced to less than 30% sequence identity with respect to Train-80 leaving 593 sequences. We call the resulting set the ITS. We then created a second "strict" set by internally redundancy reducing the ITS-80 set using BLAST with an e-value of 0.001, and then redundancy reducing it with respect to Train-80 using BLAST with an e-value of 0.001, leaving 216 sequences.

	Train	Test	Validation	Total	EMS	EMS % of total
Fold 0	11,534	812	655	13,001	4,661	36%
Fold 1	11,540	842	607	12,989	4,662	36%
Fold 2	11,530	831	638	12,999	4,643	36%
Fold 3	11,542	853	631	13,026	4,674	36%
Fold 4	11,535	823	629	12,987	4,660	36%

Table 1. Table showing the number of protein sequences in each fold, the number and the percentage of sequences in the EMS (positive) class.

	Sequences	EMS	EMS % of total
ITS	593	236	40%
ITS_strict	216	110	51%

Table 2. Table showing the number of protein sequences in the ITS and ITS_strict datasets the number and the percentage of sequences in the EMS (positive) class.

We call this set ITS_strict. Table 2 shows the number of proteins belonging to each class in ITS and ITS_strict.

This rather complex strategy of redundancy reduction ensures that there is minimal similarity between the sequences in the test and validation set with respect to each other and to the training set (e-value ≤ 0.001). However, it allows for a larger training set with sequences with up to 80% sequence identity within the training dataset. Furthermore, the two independent test set have been redundancy reduced with respect to the training set to $\leq$ 30% sequence identity or e-value ≤ 0.001 ensuring that these two set are indeed independent of the training set. Although, reducing redundancy on the entire set of proteins at the beginning of the procedure would be more straight forward this would have resulted in a much smaller training set.

3.2 Alignments and data encoding

We generated alignments of multiple homologous sequences (MSA) for all datasets used in SCLpred-EMS by iterating PSI-BLAST (Altschul *et al.*, 1997) for three rounds with an e-value of 0.001 against the June 2016 version of UniRef90 (UniProt Consortium, 2019). These alignments are encoded into MSA profiles by calculating frequencies of residues and gaps. The frequencies of the amino acid present in the original sequence were then "clipped" to 1, similarly to (Kaleel *et al.*, 2019; Torrisi *et al.*, 2019) where each amino acid is represented by a vector of 22 number frequencies of an amino acid type from the list of homologous sequences.

3.3 Predictive architecture

In this work, we tested configurations of Deep Convolutional N-to-1 neural networks of various depths and width during preliminary experiments. Deep Convolutional N-to-1 Neural Network are composed of an input kernel mapping a window of amino acids into a feature vector followed by a stack of hidden convolutional kernels followed by an average pooling unit over the whole sequence, and a final fully connected network (Figure 3). The input kernel learns a non-linear function I from a window of amino acids $\hat{ic}_i$ at position i and predicts an intermediate state vector $\hat{is}_i$ at position i.

$$\begin{aligned} \hat{is}_i &= I(\hat{ic}_i) \\ \hat{ic}_i &= (i-c, ..., i, ..., i+c) \end{aligned}$$

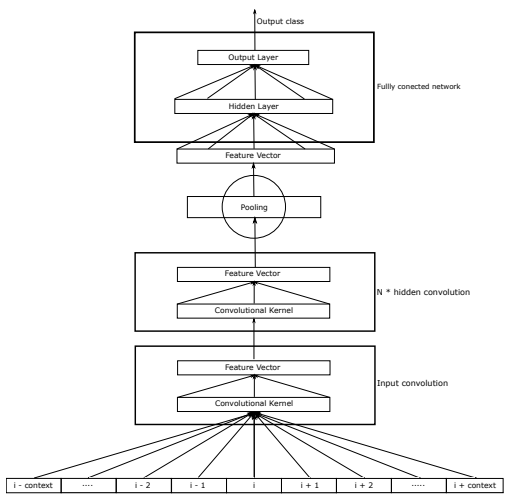

Fig. 3: Deep N-to-1 Convolutional architecture.

Each hidden convolutional kernel learns a non-linear function H^k at hidden layer k from a window of intermediate states $\hat{hc}_j^k$ at position j and predicts an intermediate state vector hs_i^k at position i.

$$hs_i^k = H^k(\hat{hc}_j^k)$$
$$\hat{hc}_j^k = (j-\gamma, ..., j, ..., j+\gamma)$$

The output vectors hs_p^l of the last hidden kernel at each position p are averaged element wise into a single vector $\hat{v}$. A fully connected network predicts the final subcellular location cls of each protein from the final vector $\hat{v}$. The fully connected network learns a non-linear function O.

$$cls = O(\hat{v})$$

Predictor	EMS	Other
DeepLoc-1.0	Extracellular	Nuclear
(Armenteros *et al.*, 2017)	Golgi apparatus	Cytoplasm
	Cell membrane	Mitochondrial
	Lysosome/Vacuole	Plastids
	Endoplasmic reticulum	Peroxisome
LocTree3	Secreted	Mitochondrion
(Goldberg *et al.*, 2014)	Plasma membrane	Nucleus
	Endoplasmic reticulum	Chloroplast
	Golgi apparatus	Cytoplasm
SCL-Epred	Secretory	Other
(Mooney *et al.*, 2013)	Membrane	

Table 3. Classes predicted by other available web servers re-classified as EMS/Other.

It should be noted that in our implementation, unlike standard Convolutional Neural Networks, all convolutional kernels and the final fully connected network are implemented by feed-forward neural networks with one hidden layer. That is, each convolutional kernel has two layers, and can be seen as a proper (1-layered) kernel followed by a non-linearity, followed by a further kernel of size 1 and a further non-linearity. In all cases we use sigmoidal non-linearities on the model's internal units, rather than rectified linear units. We found both the use of deeper convolutional stages and sigmoidal units to be beneficial in preliminary tests. The resulting architecture has a minimum of three internal (hidden) layers when no hidden-to-hidden convolutional kernel is present, while an architecture with k hidden-to-hidden kernels contains $3 + k$ hidden layers in total.

3.4 Training and ensembling

The models trained are stacks of two 2-layered convolutional layers followed by average pooling and a 2-layered fully connected network, with five inner layers in total containing roughly 2,000 weights and taking in all motifs of 21 residues. In preliminary testing, we observed a marked increase in performances for an increase in motif size up to 21, followed by a gentle degradation thereafter. We also observed modest performance improvements when increasing the depth of the stack up to five-seven inner layers, provided the total number of weights was kept approximately constant, and a slow degradation for deeper stacks. The models are trained in five-fold cross-validation. For each fold, the five models with the highest Matthews correlation coefficient (MCC) for the validation set were used in the final system. The five-fold cross-validation performance was assessed using the MCC for the test set for each fold. The final system tested on the ITS and ITS_strict sets is the ensemble of all 25 models selected from the five cross-validation folds.

3.4.1 Comparison to other predictors

We used the ITS and ITS_strict sets to benchmark SCLpred-EMS against other available subcellular localization predictors. Servers that were published in the last five years and are compatible with SCLpred-EMS class classification were benchmarked against SCLpred-EMS (see Figure 1). The predicted classes of the benchmarking servers were re-cast into EMS versus other as necessary (see Table 3). All the benchmarked predictors were run through their web-server interfaces.

3.4.2 Evaluating performance

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

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

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

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

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

where:

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

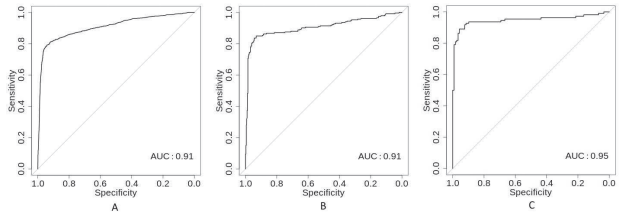

Fig. 4: ROC curve of SCLpred-EMS predictor performance (a) in five-fold cross-validation; (b) on the ITS; and (c) on the ITS_strict.

4 Results

For every protein, SCLpred-EMS predicts the probability (between 0 and 1) of that protein localising in the endomembrane system and the secretory pathway. The closer the predicted probability is to 1, the more confident SCLpred-EMS is in that prediction.

Initially, an extensive hyper-parameter search was run using the first fold of the dataset to find the optimal model (model selection). The hyper parameters were tuned based on the performance on the validation dataset. The final model contains an input kernel, a hidden kernel, an average pooling unit and, a fully connected network. The input kernel reads one amino acid at a time and compresses it into a 10 element vector. The hidden kernel reads a window of 31 of these compressed vectors (31 amino acids) and maps it into 10 element vectors. The average pooling unit compresses those 10 element vectors into a single vector with 10 elements for the whole protein sequence. Finally, the fully connected network reads the vector generated by the pooling unit and predicts the final results.

The five-fold cross-validation results for the training dataset, are shown in Table 4 and in Figure 4 as a Receiver Operating Characteristic (ROC) curve. The ROC curve is built by increasing between 0 and 1 the cut-off above which a protein is considered to be predicted as localising into the endomembrane system and the secretory pathway. The area under the curve (AUC) for the ITS is 0.91 while the AUC for the ITS_strict is 0.95.

4.1 Benchmarking results

We benchmarked SCLpred-EMS against DeepLoc-1.0 (Armenteros *et al.*, 2017), LocTree3 (Goldberg *et al.*, 2014) and SCL-Epred (Mooney

	MCC	Spec	Sen	FPR
5F-CV	0.75	88.96	77.55	4.84%
ITS	0.82	96.00	81.36	2.24%
ITS_strict	0.86	97.03	89.09	2.83%

Table 4. Performance of SCLpred-EMS in Five-fold cross-validation (5F-CV) and on the independent test set (ITS) and strict independent test set (ITS_strict).

ITS				
	MCC	Spec	Sen	FPR
SCLpred-EMS	0.82	96.00	81.36	2.24
DeepLoc	0.80	93.66	81.36	3.64
SCL-Epred	0.75	91.05	77.54	5.04
LocTree3	0.67	83.49	75.00	9.80
ITS_strict				
	MCC	Spec	Sen	FPR
SCLpred-EMS	0.86	97.03	89.09	2.83
DeepLoc	0.84	96.00	87.27	3.77
SCL-Epred	0.78	93.81	82.73	5.66
LocTree3	0.65	88.17	74.55	10.38

Table 5. Performance comparison of predictors on the ITS and ITS_strict. Performance was evaluated with Matthews correlation coefficient (MCC), specificity (Spec), sensitivity (Sens) and the false positive rate (FRP). Note: We benchmarked against the accurate version of DeepLoc as opposed to the faster version. LocTree3 did not predict subcellular locations for 11 proteins in the ITS dataset.

et al., 2013). Ideally we would have benchmarked SCLpred-EMS against DeepSig (Savojardo *et al.*, 2017) and SignalP 5.0 (Armenteros *et al.*, 2019) but, unfortunately, this is not possible. As not all transmembrane proteins actually have signal peptides they will potentially be predicted as "Other" by SignalP 5, but labelled as EMS in our dataset. Although DeepSig has a "transmembrane" output DeepSig only checks for transmembrane segments in the N-termini of the proteins. Again, proteins that lack signal peptides but have transmembrane segments in other parts of the sequence will be predicted as "Other" by DeepSig.

DeepLoc is trained on 13,858 protein sequences from the UniProt database, release 2016_04. LocTree3 is an improved version of LocTree2 (Goldberg *et al.*, 2012) that was developed on 2,240 sequences extracted from SWISS-PROT release 2011_04. LocTree3 uses a redundancy reduced dataset from LocTree2 and tested with an additional three UniprotKB datasets releases between 2011_04 and 2014. SCL-Epred used Swiss-Prot Release 2011_02 to train and test the system.

The predicted locations of these servers are recast into EMS versus all others for benchmarking purposes. Table 3 shows the class division for recasting multi-location web servers into two classes. In these tests SCLpred-EMS outperforms all other predictors tested on both the ITS and ITS_strict sets with an MCC of 0.82 and 0.86, respectively (Table 5). While we observe generally higher MCC on the stricter dataset, this is likely due to the strict dataset being more balanced.

5 Web-Server

SCLpred-EMS has been implemented as a publicly available web server. The user can submit a list of protein sequences in FASTA format, and SCLpred-EMS predicts the probability that each of these proteins will localise in the endomembrane system and secretory pathway versus all other locations. "Confidence" shows the level of confidence in the prediction. SCLpred-EMS accepts multiple queries at the same time, processes the queries in the background and sends the results to the user via an optional email address if provided. A submission query of up to 64 kbytes can be sent per submission, which is approximately 200 average sized proteins. Larger queries can be broken down into 64 kbyte chunks, or a request can be sent to lift the limit on a one-off basis. A submission query may contain spaces, newlines and tabs as they will be ignored by the system. Only 1 letter amino acid code format is understood and characters not corresponding to any amino acid will be treated as "X". Unlike some other web servers, the results are sent in the order that they were submitted via an optional email rather than a web-link which expires after some time. SCLpred-EMS responses are sent in text format and the results of multiple sequences are sent in a single email/web page. The web server version of the results is updated incrementally (every 60 seconds) until the query is complete.

6 Discussion

Recent advancements in protein subcellular localization prediction research have helped to shed light on protein interactions and subcellular location. Even though many protein subcellular localization predictors were introduced in the past fifteen years, protein subcellular location prediction still remains an unsolved problem due to the expensive and time consuming nature of experimental methods and the lack of universally reliable computational predictors.

In this work, we have introduced a new comprehensive dataset for training, testing and validation and two independent test sets. We have used a combination of pure sequence identity and BLAST e-values for stricter homology reduction. We have built a new predictor, SCLpred-EMS, based on deep Convolutional N-to-1 neural networks. This architecture has the ability to represent relatively lengthy sequence motifs while keeping the overall number of internal parameters small thereby minimising the risk of overfitting the data.

We developed and benchmarked SCLpred-EMS against other predictors published in the last five years. SCLpred-EMS outperforms the benchmarked predictors on both the ITS and the ITS_strict sets with an MCC of 0.82 and 0.86 respectively. SCLpred-EMS is freely available for academic users with a user-friendly web-interface (http://distilldeep.ucd.ie/SCLpred2). Multiple queries can be submitted to the predictor in fasta format. The datasets used in training and testing are available here: http://distilldeep.ucd.ie/SCLpred2/data/.

Even though Deep N-to-1 Convolutional neural networks perform well, convolutional neural networks do not learn the full sequential information from a protein, which is critical in the prediction of many biological sequence properties. In future work, as datasets grow ever larger and this becomes feasible, we anticipate that improvements in subcellular localisation prediction may arise from using combinations of recurrent neural network stages alongside convolutional layers similarly to Kaleel *et al.* (2019).

Acknowledgements

The authors acknowledge contributions of Amina Khalid and Tejaswini Kumar to the collection of the preliminary benchmarking results and the Research IT Service at University College Dublin for providing HPC resources that have contributed to the research results reported within this paper.

Funding

This work was supported by the Irish Research Council [GOIPG/2014/603 to M.K.] and a UCD School of Computer Science Bursary.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Armenteros, J. J. A., Sønderby, C. K., Sønderby, S. K., Nielsen, H., and Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, **33**(21), 3387–3395.

Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2019). Signalp 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology*, **37**(4), 420–423.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**(5), 412–424.

Goldberg, T., Hamp, T., and Rost, B. (2012). Loctree2 predicts localization for all domains of life. *Bioinformatics*, **28**(18), i458–i465.

Goldberg, T., Hecht, M., Hamp, T., Karl, T., Yachdav, G., Ahmed, N., Altermann, U., Angerer, P., Ansorge, S., Balasz, K., *et al.* (2014). Loctree3 prediction of localization. *Nucleic acids research*, **42**(W1), W350–W355.

Kaleel, M., Torrisi, M., Mooney, C., and Pollastri, G. (2019). Paleale 5.0: prediction of protein relative solvent accessibility by deep learning. *Amino acids*, **51**(9), 1289–1296.

Kawai, M., Cras, P., Richey, P., Tabaton, M., Lowery, D. E., Gonzalez-DeWhitt, P. A., Greenberg, B. D., Gambetti, P., and Perry, G. (1992). Subcellular localization of amyloid precursor protein in senile plaques of Alzheimer's disease. *The American Journal of Pathology*, **140**(4), 947–958.

Mooney, C., Cessieux, A., Shields, D. C., and Pollastri, G. (2013). SCL-Epred: a generalised de novo eukaryotic protein subcellular localisation predictor. *Amino acids*, **45**(2), 291–299.

Pierleoni, A., Martelli, P. L., Fariselli, P., and Casadio, R. (2006). BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**(14), e408–e416.

Pierleoni, A., Martelli, P. L., and Casadio, R. (2011a). MemLoci: predicting subcellular localization of membrane proteins in eukaryotes. *Bioinformatics*, **27**(9), 1224–1230.

Pierleoni, A., Indio, V., Savojardo, C., Fariselli, P., Martelli, P. L., and Casadio, R. (2011b). MemPype: a pipeline for the annotation of eukaryotic membrane proteins. *Nucleic Acids Research*, **39**(suppl_2), W375–W380.

Rajendran, L., Knölker, H.-J., and Simons, K. (2010). Subcellular targeting strategies for drug design and delivery. *Nature Reviews Drug Discovery*, **9**(1), 29–42.

Savojardo, C., Martelli, P. L., Fariselli, P., and Casadio, R. (2015). TPpred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. *Bioinformatics*, **31**(20), 3269–3275.

Savojardo, C., Martelli, P. L., Fariselli, P., and Casadio, R. (2017). DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics*, **34**(10), 1690–1696.

Savojardo, C., Martelli, P. L., Fariselli, P., Profiti, G., and Casadio, R. (2018). BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Research*, **46**(W1), W459–W466.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature Biotechnology*, **18**(12), 1257–1261.

Torrisi, M., Kaleel, M., and Pollastri, G. (2019). Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. *Scientific reports*, **9**(1), 1–12.

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**(D1), D506–D515.

Yu, C.-S., Chen, Y.-C., Lu, C.-H., and Hwang, J.-K. (2006). Prediction of protein subcellular localization. *Proteins: Structure, Function, and Bioinformatics*, **64**(3), 643–651.

Yu, C.-S., Cheng, C.-W., Su, W.-C., Chang, K.-C., Huang, S.-W., Hwang, J.-K., and Lu, C.-H. (2014). CELLO2go: A Web Server for Protein subCELlular LOcalization Prediction with Functional Gene Ontology Annotation. *PLOS ONE*, **9**(6), e99368.