Noname manuscript No.
(will be inserted by the editor)

SCL-Epred: A Generalised *De novo* Eukaryotic Protein Subcellular Localisation Predictor

Eukaryotic Protein Subcellular Localisation Prediction

Catherine Mooney · Amélie Cessieux · Denis C. Shields · Gianluca Pollastri

Received: date / Accepted: date

Abstract Knowledge of the subcellular location of a protein provides valuable information about its function, possible interaction with other proteins and drug targetability, among other things. The experimental determination of a protein's location in the cell is expensive, time consuming and open to human error. Fast and accurate predictors of subcellular location have an important role to play if the abundance of sequence data which is now available is to be fully exploited.

Catherine Mooney
Conway Institute of Biomolecular and Biomedical Science,
School of Medicine and Medical Science,
Complex and Adaptive Systems Laboratory,
University College Dublin, Belfield,
Dublin 4, Ireland
E-mail: catherine.mooney@ucd.ie

Amélie Cessieux
Département Génie Biologique,
Polytech'Nice-Sophia,
1645 Route des Lucioles,
06410 Biot, France
E-mail: cessieux@polytech.unice.fr

Denis C. Shields
Conway Institute of Biomolecular and Biomedical Science,
School of Medicine and Medical Science and
Complex and Adaptive Systems Laboratory,
University College Dublin, Belfield,
Dublin 4, Ireland
E-mail: denis.shields@ucd.ie

Gianluca Pollastri
Complex and Adaptive Systems Laboratory and
School of Computer Science and Informatics,
University College Dublin, Belfield,
Dublin 4, Ireland
E-mail: gianluca.pollastri@ucd.ie

In the post-genomic era, genomes in many diverse organisms are available. Many of these organisms are important in human and veterinary disease and fall outside of the well studied plant, animal and fungi groups. We have developed a general eukaryotic subcellular localisation predictor (SCL-Epred) which predicts the location of eukaryotic proteins into three classes which are important, in particular, for determining the drug targetability of a protein – secreted proteins, membrane proteins and proteins that are neither secreted nor membrane.

The algorithm powering SCL-Epred is a N-to-1 Neural Network, and is trained on very large non-redundant sets of protein sequences. SCL-Epred performs well on training data achieving a Q of 86% and a GC of 0.75 when tested in 10-fold cross-validation on a set of 15,202 redundancy reduced protein sequences. The three class accuracy of SCL-Epred and LocTree2, and in particular a consensus predictor comprising both methods, surpasses that of other widely used predictors when benchmarked using a large redundancy reduced independent test set of 562 proteins. SCL-Epred is publicly available at http://distillf.ucd.ie/distill/.

Keywords Subcellular Localisation Prediction · Eukaryotes · N-to-1 Neural Network · SCL-Epred

1 Introduction

As the gap between sequence-known and sequence-annotated proteins continues to grow so does the demand for quick and accurate automated annotation systems. Subcellular localisation prediction web servers have been widely and very successfully used in this context for more than a decade (Nielsen *et al.*, 1997; Nakai and Horton, 1999; Emanuelsson *et al.*, 2000). Knowledge of the subcellular location of a protein provides valuable information about its function, possible interaction with other proteins and drug targetability, among other things. Experimental approaches to determine subcellular localisation are time-consuming and expensive, whereas computational methods can provide fast and increasingly more accurate localisation predictions. Subcellular localisation predictors have an important role to play if the abundance of sequence data which is now available is to be fully exploited.

In this post-genomic era, the genomes of many diverse organisms are becoming available. Many of these organisms are important in human and veterinary disease and fall outside of the well studied plant, animal and fungi groups. In particular many important diseases are caused by parasites found in the "SAR" group of Stramenopiles, Alveolates and Rhizaria (Burki *et al.*, 2007). For example, Malaria caused by *Plasmodium spp.* parasites kills over a million people a year (Murray *et al.*, 2012); Toxoplasmosis, Cryptosporidiosis and Isosporiasis, caused by *Toxoplasma gondii*, *Cryptosporidium spp.*, and *Isospora belli* respectively, are especially important in people with HIV-AIDS or other immunocompromised individuals (Gellin and Soave, 1992); Babesiosis (caused by *Babesia bovis*) and Theileriosis (caused by *Theileria annulata*),

diseases in cattle cause economic loss especially to farmers in Africa, limiting economic development (Gardner *et al.*, 2005; Brayton *et al.*, 2007). Other neglected tropical diseases such as Chagas disease , Human African Trypanosomiasis and Leishmaniasis are caused by parasites in the Excavate supergroup (*Trypanosoma cruzi*, *Trypanosoma brucei* and *Leishmania spp.*).

New effective drugs are needed to fight these parasites and determining which proteins may be targatable is one of the first steps. Other than a few species specific predictors such as PATS (Zuegge *et al.*, 2001), PlasmoAP (Foth *et al.*, 2003) and PlasMit (Bender *et al.*, 2003), which are specialised for *Plasmodium falciparum* proteins, very few predictors are suitable for use within these groups. These species specific predictors do not generalise well to other species and also have the disadvantage of being trained on very small datasets. It is important for drug discovery to be able to identify if a protein is intracellular, extracellular or a membrane protein. Even this simple classification when performed on a genomic scale can greatly help in directing drug discovery efforts since currently more than half of all drug targets are found in membrane proteins (Bakheet and Doig, 2009). Drug targets are also more likely to be extracellular than intracellular, with few targets found in organelles (Bakheet and Doig, 2009).

SCL-Epred has been developed to address the need for an accurate general predictor of eukaryotic protein subcellular location into three classes: extracellular/secreted, membrane or "other". SCL-Epred is based on a N-to-1 Neural Network which we have previously described (Mooney *et al.*, 2011; Volpato *et al.*, 2013) and is trained in 10-fold cross-validation on very large redundancy reduced training sets generated from Swiss-Prot Release 2011_02 (Boeckmann *et al.*, 2003). At present there is no other predictor available which can preform this simple classification. The authors of the PSORT suite of web servers (Nancy *et al.*, 2010) provide a comprehensive list of subcellular localisation predictors on their web-page (http://www.psort.org/). Out of the approximately 50 web servers for the prediction of subcellular localisation in eukaryotes about half are specialised for the prediction of plant, animal and fungi proteins, or are even more specialised for a particular species e.g. human (Garg *et al.*, 2005) or rice (Kaundal and Raghava, 2009), or location e.g. Golgi membrane proteins (Yuan and Teasdale, 2002). About a quarter of the servers are no longer working, or are not currently available. We found only a small number of web servers that are available for the prediction of eukaryotes in general and we have benchmarked SCL-Epred against these predictors on an independent test set of 715 proteins generated from Swiss-Prot Release 2012_07.

SCL-Epred will be useful for the prediction of the subcellular location of proteins from complete proteomes of species which fall outside of the well provided for plant, animal and fungi groups and may also be used for predicting the subcellular location of proteins for species within those groups for example, as a first step towards screening a whole proteome for drug targets. Proteins classified as secreted can be further examined using SignalP to determine the location of the signal peptide and membrane proteins can be

further classified using MemLoci (Pierleoni *et al.*, 2011) which discriminates between three membrane protein localisations: plasma, internal and organelle membrane. Finally, intracellular locations could be further refined using LocTree2 (Goldberg *et al.*, 2012), for all eukaryotes, or BaCelLo (Pierleoni *et al.*, 2006), WoLF PSORT (Horton *et al.*, 2007), SherLoc (Shatkay *et al.*, 2007) or SCLpred (Mooney *et al.*, 2011) for plant, animal and fungi proteins, for example.

2 Materials and Methods

2.1 Datasets

We generated our training and test set from Swiss-Prot Release 2011_02. We found 129,516 eukaryotic sequences with a "SUBCELLULAR LOCATION" annotation. We removed sequences that had non-experimental qualifiers (Potential, Probable, By similarity) or were less than 30 or greater than 1500 residues in length, leaving 47,521 sequences. We internally redundancy reduced 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 (Table 1). Proteins were then classified as secreted (having the keyword "Secreted" within their "SUBCELLULAR LOCATION" annotation, and not "membrane" or "intermembrane"), membrane (having the keyword "membrane" or "intermembrane" within their "SUBCELLULAR LOCATION" annotation and not "Secreted"), or "other" (not "Secreted", "membrane" or "intermembrane").

We created an independent test set from Swiss-Prot Release 2012_07 using a similar method. We redundancy reduced this set with respect to the training set to less than 30% sequence similarity and then internally redundancy reduced the remaining sequences to less than 30% sequence similarity, leaving 715 protein sequences (Table 1). We refer to this dataset as the independent test set.

We further reduced the independent test set for use only when benchmarking against other predictors. We only included cell membrane proteins in this dataset to be fair to the other predictors, some of which were trained only to identify cell membrane, and not other membrane, proteins. We further redundancy reduced this dataset with respect to the LocTree2 training dataset in order to make a fair comparison between SCL-Epred and LocTree2. As the performance of SCL-Epred and LocTree2, on this redundancy reduced set, is significantly better than that of all the other predictors which we tested we did not consider it was necessary to redundancy reduce the dataset further with respect to any other predictor's training sets. This may give slightly inflated results for these predictors but does not affect our overall conclusions.

MSA

Multiple sequence alignments are extracted from UniRef90 from January 2011 containing 9,616,951 sequences (Suzek *et al.*, 2007). 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: N-to-1 Neural Network

The N-to-1 Neural Network (N1-NN) has been previously described in detail in Mooney *et al.* (2011). This implementation of the model maps a protein sequence of variable length N into three subcellular locations i.e. secreted, membrane or "other" (non-secreted, non-membrane). These features are stored in a vector $f = (f_1, f_2, f_3)$, if the i-th residue in the sequence is represented 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 is implemented as 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 and k is a normalisation constant. The feature vector f is obtained by combining information coming from all windows of $2c + 1$ residues in the sequence. In this work $c = 20$, therefore the motifs have a length of 41 residues. The feature vector f is mapped into the subcellular locations o, as follows:

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

where $\mathcal{N}^{(o)}$ is implemented as another two-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 and is 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. See Mooney *et al.* (2011) (Mooney *et al.*, 2011) for more details.

Training

Training was conducted in 10-fold cross-validation, i.e. 10 different sets of training were performed in which a different tenth of the overall set was 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 were not reserved for testing were further split into a validation set (1/10 of the overall set) and a training set. The training set was used to learn the free parameters of the network by gradient descent, while the validation set was used to monitor the training process. In order to mitigate the effect of the imbalance between the three classes every membrane sequence was presented twice as often during training as a sequence in the "other" class, and secreted sequence were presented four times as often. Three models were trained independently for each fold and ensemble averaged to build the final predictor. Differences among models were introduced by varying the architectural parameters of the network.

The weights in the networks were updated every 152 examples (protein sequences) and training continued until the walltime on the server was reached (10 days, corresponding to between 1200 and 1700 epochs of training per network). We saved the networks that performed best on the validation set, ensemble averaged them and evaluate them on the corresponding test set. The final result for the 10-fold cross-validation is the average of the results on each test set. When testing on the independent test sets we ensemble-combined all the models from all the cross-validation folds.

2.3 Evaluating performance

To evaluate the performance of SCL-Epred against other predictors we measure specificity (Spec), sensitivity (Sens), the false positive rate (FRP), the percentage of correctly predicted sequences (Q), Matthews correlation coefficient (MCC) and the generalised correlation (GC) (Baldi *et al.*, 2000) as follows:

$$
\begin{aligned}
Spec &= 100\frac{TP}{TP+FP} \\
Sens &= 100\frac{TP}{TP+FN} \\
FPR &= 100\frac{FP}{FP+TN} \\
MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \\
Q &= 100\frac{\sum_i z_{ii}}{N}
\end{aligned}
$$

$$GC = \sqrt{\frac{\sum_{ij} \frac{(z_{ij}-e_{ij})^2}{e_{ij}}}{N(K-1)}} \tag{3}$$

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 emphasise performances based on MCC or GC, as this index minimises the effect of class sizes (see Baldi *et al.* (2000) for more details).

3 Results and Discussion

SCL-Epred performs well when tested in 10-fold cross-validation on the training set of 15,202 protein sequences and on the 715 protein sequences of the independent test set (Table 2), predicting sequences into three classes: secreted, membrane and "other". The number of correctly predicted sequences (Q) is 86% and 83.2% respectively for the training and independent test sets. However, if the class size is unbalanced, as it is here (secreted sequences make up only 13% of the dataset), Q can be misleading, as if a predictor predicts all sequences into the largest class a reasonable Q can still be achieved. For example, a Q of 60% could be achieved on our independent test set by predicting all sequences into the "other" class, at the expense of predicting every secreted and membrane protein incorrectly. Therefore, GC and MCC provide a more accurate assessment of the performance of the predictor across all classes. SCL-Epred achieves a GC of 0.75 for the training set and 0.7 for the test set. The balanced performance of the predictor can be observed in Table 2, especially for the training set where MCC only varies from 0.72 to 0.78 across the three classes.

Although the models were trained on sequences of length greater than 30 residues and less than 1500 residues, SCL-Epred can predict shorter and longer sequences (up to 8192 residues). We tested the ability of the predictor

to accurately predict some of these short and long sequences using an internally redundancy reduced (< 30% sequence similarity) set of sequences that were less than 30 residues in length and had been excluded from the original training set due to their short length. We found that 359 of these 387 short sequences were labelled as secreted, and 97.5% of these were correctly predicted. However the 11 short membrane and 17 "other" short sequences were incorrectly predicted as secreted. We see the opposite pattern for sequences greater than 1500 in length. Only 13 of the 179 internally redundancy reduced sequences in the set of long proteins which were excluded from our training set were labelled as secreted. 36 were labelled as membrane proteins and 130 as "other". 95% of the "other" and 64% of the membrane sequences were predicted correctly, but only 8% of the secreted. It is clear that the relative length distribution of the classes needs to be considered when interpreting results for these very short or very long sequences.

SCL-Epred does not predict multiple locations, however we tested the behaviour of the web server when predicting sequences which are annotated with multiple locations in Swiss-Prot. We generated three test sets of sequences which were excluded from our dataset due to having multiple subcellular location annotations (membrane and cytoplasm; membrane and secreted; and secreted and cytoplasm). We internally redundancy reduced this dataset as previously described, and redundancy reduced it to less than 30% sequence similarity with the training set. When submitted to the SCL-Epred web server 31 of the 32 of membrane/cytoplasm sequences were correctly predicted as either membrane or "other" and all 5 of the secreted/cytoplasm sequences were correctly predicted as secreted or "other". There was only one membrane/secreted sequence which was predicted as "other".

We benchmarked SCL-Epred against a number of publicly available web servers using the benchmarking subset of the independent test set of 562 protein sequences. First we assessed the ability of SCL-Epred to correctly predict secreted proteins. A recent study which evaluated thirteen predictors of secreted proteins found that SignalP (Petersen *et al.*, 2011) achieved the best overall accuracy (Choo *et al.*, 2009). As well as SignalP we compared our performance to Signal-BLAST (Frank and Sippl, 2008), another predictor of signal peptides, and three other predictors not specifically trained to predict secreted proteins or signal peptides in isolation, LocTree2 (Goldberg *et al.*, 2012), SLPFA (Tamura and Akutsu, 2007) and ESLpred2 (Garg and Raghava, 2008). It is important to note that we do not specifically predict if a protein has a signal peptide as some proteins may use other secretary mechanisms (non-classical or leaderless protein secretion) (Bendtsen *et al.*, 2004). For predictors which predict into more than two classes we classified predictions as secreted (secreted–LocTree2/Extracellular–ESLpred2/Secretory(SP)–SLPFA)) or "other".

ESLpred2 with a Q of 81.7% only predicted four of the secreted proteins correctly. However, a GC of 0.13 more accurately reflects the poor performance of ESLpred2 on this dataset (Table 3) which demonstrated the importance of not relying on Q as an overall indicator of performance. SCL-Epred has the

highest Q and GC, however LocTree2 and SignalP also perform very well (all $Q > 90\%$ and GC > 0.7). We tested the performance of a two class consensus predictor made up of SCL-Epred, LocTree2 and SignalP, where a sequence was predicted as secreted if more that one of the three predictors predicted it as secreted, and predicted as non-secreted otherwise. Interestingly, the consensus improved over all individual results giving a Q of 95.2% and a GC of 0.84. This emphasises the importance of not relying on the results of a single predictor and the power of consensus.

For three class results we benchmarked SCL-Epred against LocTree2, Euk-mPLoc (Chou and Shen, 2010), SLPS (Jia *et al.*, 2007) and Cello (Yu *et al.*, 2006) using the benchmarking subset of the independent test set (Table 4). We chose these predictors for comparison with SCL-Epred as they all predict secreted and membrane proteins as well as other protein classes. As before, SCL-Epred and LocTree2 perform best in both Q and GC, with a Q of 90% and GC of 0.7 for both. Q for Euk-mPLoc, SLPS and Cello is good, 84.2%, 83.5% and 79.9% respectively, but all three exhibit a much lower GC (0.54%, 0.48% and 0.48% respectively).

As only two predictors performed well for three class predictions (SCL-Epred and LocTree2) we looked at ways of creating a consensus between the two. LocTree2 is more accurate at predicting membrane sequences than SCL-Epred (MCC of 0.65 and 0.57 respectively), however the MCC for SCL-Epred is higher than LocTree2 for secreted and "other" sequences. Therefore, we created a consensus predictor which took the SCL-Epred prediction if the sequence was predicted as secreted or "other" and we took the LocTree2 prediction if the sequences was predicted as membrane. This resulted in 90.7% of sequences being predicted correctly in the benchmarking test set (Q) and a GC of 0.72. Although this might be considered only a small improvement over the independent results of either predictor it does illustrate the potential advantage of having a number of highly accurate predictors which can be ensembled to give more accurate results, similar to meta-predictors which have been used successfully to improve protein structure predictions for many years (Mariani *et al.*, 2011).

To estimate the statistical significance of these results, we measured the standard deviation of the error distribution by sampling with replacement N sequences from the set of 562 sequences M times. In our case M = 1000 and N = 562. If we consider two standard deviations apart as safe (i.e. <5% probability that the order is random) we can observe that there is little to distinguish SCL-Epred, LocTree2, SignalP and the consensus predictor from each other. However, it is clear that these predictors are substantially outperforming the other predictors tested. It should be kept in mind that the test sets are small (562 sequences) and although we do not see a statistically significant separation in all cases, our results, and in particular the results of the consensus predictor, are generally better.

We wished to investigate if SCL-Epred could accurately predict the location of proteins sequences of species outside of the large plant, animal and fungi groups for which there are many specialised predictors. As the number

of sequences annotated with a subcellular location outside of these groups is limited, species specific predictors for non plant, animal or fungi groups have the disadvantage of being trained on very small datasets. Therefore, having trained SCL-Epred using all available eukaryotic sequences we examined the training set results (in 10-fold cross-validation) categorising each sequence into one of five supergroups (Keeling *et al.*, 2005), as follows, to assess the predictive power for each supergroup. We found 723 unique species in the training set. First we grouped these sequences by their UniProt taxonomy (Table 5), and then organised them into one of the five supergroups: Unikonts (Amoebozoa, Fungi and Metazoa), Plantae (Viridiplantae, Rhodophyta and Glaucocystophyceae), Chromalveolates (Stramenopiles, Alveolata, Cryptophyta and Haptophyceae), Excavates (Diplomonadida, Heterolobosea, Jakobida, Parabasalia and Euglenozoa) and Rhizaria. As the number of non-Plantae/Unikont sequences was still quite small we grouped the SAR group (Rhizaria and Chromalveolates) and Excavates together. Table 6 shows that 85% of the dataset sequences are Unikonts, with only 2% coming from the Chromalveolates/Rhizaria /Excavate supergroups. However, even with only such a small number of sequences SCL-Epred has some predictive power with a Q of 75.1% and a GC of 0.56. Importantly, the predictive power is balanced across all three classes with MCC only varying from 0.53 to 0.57. The real strength of the predictor, nevertheless, still lies within the Unikonts which have a very strong performance with a Q of 86.4% and a GC of 0.77. We then compared the performance of SCL-Epred on the SAR_Excavate dataset to that of LocTree2, Euk-mPLoc and Cello (Table 7). SCL-Epred is more accurate than the other predictors across each of the three classes except for the "other" class where the MCC for LocTree2 is better. GC and Q are better in all cases. A consensus predictor which selects the SCL-Epred prediction if the LocTree2 prediction is either secreted or membrane, and keep the LocTree2 prediction if it is "other" is more accurate than either SCL-Epred or LocTree2 across all classes (Table 7). It should be kept in mind that the SAR_Excavates dataset is a subset of SCL-Epred training data and that the method performs better on the training data than on independent test data (Table 2). Second, it may overlap with the training data of other methods. This should be kept in mind when evaluating these results.

4 Conclusion

We have developed a general *de novo* subcellular localisation predictor for eukaryotic protein sequences which predicts into three classes (secreted, membrane and "other") based on a N-to-1 Neural Network architecture (N1-NN). We have trained SCL-Epred in 10-fold cross-validation on large non-redundant subsets of annotated proteins from Swiss-Prot 2011_02 and benchmarked it against other subcellular localisation prediction servers on an independent test set. SCL-Epred performs favourably on these benchmarks. We have explored the possibility of using a subcellular localisation predictor trained on a

diverse set of eukaryotic sequences to predict the localisation of proteins from the often overlooked Chromalveolates, Rhizaria and Excavate supergroups. We have shown that SCL-Epred has some predictive power here and expect that as larger datasets become available these will to be especially beneficial towards improving prediction accuracy for these sequences. In this work, we have used the primary sequence and multiple sequence alignments as inputs to the network. Additional residue-level information could be included, such as predicted secondary structure, solvent accessibility, location of binding sites, or the inclusion of putative homology to proteins of known localisation, this is the subject of future work.

As the amount of sequence information generated by experimental methods keeps expanding at an ever-increasing pace, it is crucial to develop and make available faster and more accurate computational methods if this abundance of sequence data is to be fully exploited. Subcellular localisation prediction is a step toward 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.

Consensus or meta-predictors have been used successfully for many years to improve protein structure predictions accuracy (Mariani *et al.*, 2011) and we have shown that consensus predictions may be used to increase the accuracy of subcellular localisation prediction also. Alternative predictors have implemented different training methods (e.g. neural network-based methods such as SignalP or SCL-Epred, or methods trained using support vector machines (SVM) such as LocTree2), different training datasets, prediction into different locations or number of locations. Accordingly, some predictors are more accurate than others at prediction into any one class. This variability among methods can be exploited to lead to more accurate overall consensus predictions. We have shown that a simple consensus predictor built using predictions from SCL-Epred, LocTree2 and SignalP leads to improved prediction of secreted versus non-secreted proteins over any of the individual methods with a increase in GC from 0.79 to 0.84. Similarly, a consensus predictor of SCL-Epred and LocTree2 showed improvement in three class prediction accuracy from a GC of 0.7 to 0.72.

There is very little difference in terms of accuracy, ease of use and speed between SCL-Epred and LocTree2. As SCL-Epred and LocTree2 have both been implemented as web servers there is no need to install software locally, predictions are fast and are returned to the user via email. However, as we have shown a consensus of both predictors is better than either individually, and the end user should keep this in mind. Both methods are significantly more accurate than the other predictors tested (with the exception of two class predictions for SignalP). Importantly, we have shown that SCL-Epred performs better than any other predictor in the SAR-Excavates group and if this is the area of interest for the end user then SCL-Epred should be the predictor of choice, however, there are many other predictors available which are specialised for animal, plant and fungi sequences.

Table 1 Number of sequences per class for the training set and the independent test set.

	Training set	Independent test set	Independent test set (subset)
Secreted	1944	115	105
Membrane	4247	170	37
Other	9011	430	420
Total	15202	715	562

Table 2 Results for SCL-Epred trained and tested in 10-fold cross-validation on the training set and independent test set. SD - standard deviation.

	Spec	Sens	FPR	MCC	Q	GC
Training set						
Secreted	75.11	86.73	4.24	0.78		
Membrane	89.21	69.91	3.29	0.72		
Other	87.49	93.44	19.46	0.75		
					86.0 (0.28)	0.75 (0.005)
Independent test set						
Secreted	84.91	78.26	2.67	0.78		
Membrane	80.49	58.24	4.40	0.61		
Other	83.54	94.42	28.07	0.70		
					83.2 (1.38)	0.70 (0.024)

Table 3 Two class results (secreted/non-secreted) for SCL-Epred benchmarked against LocTree2, SignalP, SLPFA, Signal-BLAST and ESLpred2 on the benchmarking subset of the independent test set. The concensus predictor is made up of SCL-Epred, LocTree2 and SignalP. SD - standard deviation.

	Q	SD	GC	SD
Consensus	95.2	0.88	0.84	0.030
SCL-Epred	93.7	1.07	0.79	0.036
LocTree2	92.3	1.11	0.74	0.038
SignalP	91.8	1.13	0.72	0.038
SLPFA	87.2	1.39	0.55	0.047
Signal-BLAST	82.6	1.60	0.51	0.042
ESLpred2	81.7	1.63	0.13	0.055

SCL-Epred is available as part of our web server for protein sequence annotation. Our server is designed to allow fast and reliable annotation of protein sequences on a genomic-scale. The servers are freely available for academic users at http://distillf.ucd.ie/distill/. Linux binaries and the benchmarking sets are freely available for academic users upon request.

Acknowledgements The work was funded through a Science Foundation Ireland principal investigator grant (08/IN.1/B1864) to D. C. Shields and a Science Foundation Ireland

Table 4 Three class results (secreted, membrane and other) for SCL-Epred benchmarked against LocTree2, Euk-mPLoc, SLPS (results calculated on 170 sequences, "no determinant" for other 392 sequences) and Cello on the benchmarking subset of the independent test set. The concensus predictor is made up of SCL-Epred and LocTree2. Standard deviations are shown in brackets for GC and Q

	Consensus	SCL-Epred	LocTree2	Euk-mPLoc	SLPS	Cello
	MCC	MCC	MCC	MCC	MCC	MCC
Secreted	0.82	0.79	0.74	0.66	0.22	0.50
Membrane	0.59	0.57	0.65	0.38	0.61	0.45
Other	0.79	0.79	0.73	0.62	0.56	0.48
GC	0.72 (0.034)	0.70 (0.032)	0.70 (0.038)	0.54 (0.034)	0.48 (0.085)	0.48 (0.042)
Q	90.7 (1.25)	89.9 (1.25)	89.3 (1.34)	84.2 (1.46)	83.5 (1.88)	79.9 (1.69)

Table 5 UniProt taxonomic classification of the species in the training set.

	Number
Alveolata	104
Amoebozoa	237
Cryptophyta	20
Diplomonadida	10
Euglenozoa	123
Fungi	4528
Glaucocystophyceae	23
Haptophyceae	6
Heterolobosea	4
Jakobida	5
Metazoa	8175
Parabasalia	3
Rhizaria	12
Rhodophyta	127
Stramenopiles	26
Viridiplantae	1799

Table 6 Results for SCL-Epred tested in 10-fold cross-validation on the training set with sequences classified into the five supergroups. The results for Chromalveolates, Rhizaria and Excavates have been combined due to their small number. SD - standard deviation.

	SAR_Excavates		Plantae		Unikonts	
	Num	MCC	Num	MCC	Num	MCC
Secreted	19	0.53	107	0.57	1818	0.80
Membrane	88	0.56	467	0.70	3692	0.73
Other	206	0.57	1375	0.69	7430	0.77
GC		0.56 (0.039)		0.64 (0.019)		0.77 (0.005)
Q		75.1 (2.33)		85.0 (0.82)		86.4 (0.30)

Table 7 Results for SCL-Epred, LocTree2, Euk-mPLoc and Cello tested on the SAR_Excavates dataset. SD - standard deviation.

	Consensus	SCL-Epred	LocTree2	Euk-mPLoc	Cello
	MCC	MCC	MCC	MCC	MCC
Secreted	0.55	0.53	0.32	0.41	0.34
Membrane	0.66	0.56	0.40	0.40	0.19
Other	0.63	0.57	0.69	0.52	0.33
GC	0.61 (0.046)	0.56 (0.039)	0.51 (0.031)	0.45 (0.042)	0.33 (0.047)
Q	81.5 (2.14)	75.1 (2.33)	65.6 (2.54)	64.5 (2.56)	54.5 (2.69)

research frontiers grant (10/RFP/GEN2749) to G. Pollastri. The authors wish to acknowledge UCD IT Services, and in particular the Phaeton administrators, for the provision of computational facilities and support. We thank Tatyana Goldberg from the Rost Lab at TU Munich for providing LocTree2 predictions.

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.

Bakheet, T. and Doig, A. (2009). Properties and identification of human protein drug targets. *Bioinformatics*, **25**(4), 451–457.

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.

Bender, A., van Dooren, G., Ralph, S., McFadden, G., and Schneider, G. (2003). Properties and prediction of mitochondrial transit peptides from Plasmodium falciparum. *Mol and Biochem Parasit*, **132**, 59–66.

Bendtsen, J., Jensen, L., Blom, N., Von Heijne, G., and Brunak, S. (2004). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel*, **17**(4), 349–356.

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.

Brayton, K., Lau, A., Herndon, D., Hannick, L., Kappmeyer, L., Berens, S., Bidwell, S., Brown, W., Crabtree, J., Fadrosh, D., *et al.* (2007). Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoprotozoa. *PLoS Pathog*, **3**(10), e148.

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjæveland, Å., Nikolaev, S., Jakobsen, K., and Pawlowski, J. (2007). Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One*, **2**(8), e790.

Choo, K., Tan, T., and Ranganathan, S. (2009). A comprehensive assessment of N-terminal signal peptides prediction methods. *BMC Bioinformatics*, **10**(Suppl 15), S2.

Chou, K. and Shen, H. (2010). A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS One*, **5**(4), e9931.

Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., *et al.* (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*, **300**(4), 1005–1016.

Foth, B., Ralph, S., Tonkin, C., Struck, N., Fraunholz, M., Roos, D.S.and Cowman, A., and McFadden, G. (2003). Dissecting apicoplast targeting in the malaria parasite Plasmodium falciparum. *Science*, **299**, 705.

Frank, K. and Sippl, M. (2008). High-performance signal peptide prediction based on sequence alignment techniques. *Bioinformatics*, **24**(19), 2172–2176.

Gardner, M., Bishop, R., Shah, T., de Villiers, E., Carlton, J., Hall, N., Ren, Q., Paulsen, I., Pain, A., Berriman, M., *et al.* (2005). Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. *Science*, **309**(5731), 134.

Garg, A. and Raghava, G. (2008). ESLpred2: improved method for predicting subcellular localization of eukaryotic proteins. *BMC Bioinformatics*, **9**(1), 503.

Garg, A., Bhasin, M., and Raghava, G. (2005). Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J Biol Chem*, **280**(15), 14427–14432.

Gellin, B. and Soave, R. (1992). Coccidian infections in AIDS. toxoplasmosis, cryptosporidiosis, and isosporiasis. *Med Clin N Am*, **76**(1), 205.

Goldberg, T., Hamp, T., and Rost, B. (2012). LocTree2 predicts localization for all domains of life. *Bioinformatics*, **28**(18), i458–i465.

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**, W585–W5857.

Jia, P., Qian, Z., Zeng, Z., Cai, Y., and Li, Y. (2007). Prediction of subcellular protein localization based on functional domain composition. *Biochem Bioph Res Co*, **357**(2), 366–370.

Kaundal, R. and Raghava, G. (2009). RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. *Proteomics*, **9**(9), 2324–2342.

Keeling, P., Burger, G., Durnford, D., Lang, B., Lee, R., Pearlman, R., Roger, A., and Gray, M. (2005). The tree of eukaryotes. *Trends Ecol Evol*, **20**(12), 670–676.

Mariani, V., Kiefer, F., Schmidt, T., Haas, J., and Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9. *Proteins*, **79**(S10), 37–58.

Mooney, C., Pollastri, G., *et al.* (2011). SCLpred: protein subcellular localization prediction by N-to-1 Neural Networks. *Bioinformatics*, **27**(20), 2812–2819.

Murray, C., Rosenfeld, L., Lim, S., Andrews, K., Foreman, K., Haring, D., Fullman, N., Naghavi, M., Lozano, R., and Lopez, A. (2012). Global malaria mortality between 1980 and 2010: a systematic analysis. *Lancet*, **379**(9814), 413–431.

Nakai, K. and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, **24**(1), 34–35.

Nancy, Y., Wagner, J., Laird, M., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S., Ester, M., Foster, L., *et al.* (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**(13), 1608–1615.

Nielsen, H., Engelbrecht, J., Brunak, S., and Von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*, **10**(1), 1–6.

Petersen, T., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*, **8**(10), 785–786.

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

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

Pollastri, G. and McLysaght, A. (2005). Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, **21**(8), 1719–1720.

Shatkay, H., Höglund, A., Brady, S., Blum, T., Dönnes, 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.

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

Tamura, T. and Akutsu, T. (2007). Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition. *BMC Bioinformatics*, **8**(1), 466.

Volpato, V., Adelfio, A., and Pollastri, G. (2013). Accurate prediction of protein enzymatic class by N-to-1 Neural Networks. *BMC Bioinformatics*, **14**(Suppl 1), S11.

Yu, C., Chen, Y., Lu, C., and Hwang, J. (2006). Prediction of protein subcellular localization. *Proteins*, **64**(3), 643–651.

Yuan, Z. and Teasdale, R. (2002). Prediction of Golgi Type II membrane proteins based on their transmembrane domains. *Bioinformatics*, **18**(8), 1109–1115.

Zuegge, J., Ralph, S., Schmuker, M., McFadden, G., and Schneider, G. (2001). Deciphering apicoplast targeting signals – feature extraction from nuclear-encoded precursors of Plasmodium falciparum apicoplast proteins. *Gene*, **280**, 19–26.