Predicting eukaryotic protein secretion without signals

Nielsen, Henrik; Petsalaki, Eirini I.; Zhao, Linlin; Stühler, Kai

Published in:
Biochimica et Biophysica Acta - Proteins and Proteomics

Link to article, DOI:
[10.1016/j.bbapap.2018.11.011](https://doi.org/10.1016/j.bbapap.2018.11.011)

Publication date:
2019

Document Version
Peer reviewed version

[Link back to DTU Orbit](#)

Citation (APA):
Nielsen, H., Petsalaki, E. I., Zhao, L., & Stühler, K. (2019). Predicting eukaryotic protein secretion without signals. *Biochimica et Biophysica Acta - Proteins and Proteomics*, [140174].
<https://doi.org/10.1016/j.bbapap.2018.11.011>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Predicting eukaryotic protein secretion without signals

Henrik Nielsen^{1,*}, Eirini I. Petsalaki¹, Linlin Zhao², and Kai Stühler²

¹Department of Bio and Health Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark

²Molecular Proteomics Laboratory, Institute of Molecular Medicine, Universitätsklinikum Düsseldorf, Düsseldorf, Germany

*to whom correspondence should be addressed, hnielsen@bioinformatics.dtu.dk

Keywords: protein secretion; prediction; machine learning; artificial neural network; support vector machine

Abstract

Predicting unconventional protein secretion is a much harder problem than predicting signal peptide-based protein secretion, both due to the small number of examples and due to the heterogeneity and the limited knowledge of the pathways involved, especially in eukaryotes. However, the idea that secreted proteins share certain properties regardless of the secretion pathway used made it possible to construct the prediction method SecretomeP in 2004. Here, we take a critical look at SecretomeP and its successors, and we also discuss whether multi-category subcellular location predictors can be used to predict unconventional protein secretion in eukaryotes. A new benchmark shows SecretomeP to perform much worse than initially estimated, casting doubt on the underlying hypothesis. On a more positive note, recent developments in machine learning may have the potential to construct new methods which can not only predict unconventional protein secretion but also point out which parts of a sequence are important for secretion.

1 Introduction

Prediction of classical signal peptide-based protein secretion has a long history in bioinformatics, with the earliest methods being published in the 1980's [1–3]. The secretory signal peptide is probably the best known and most well-described protein sorting signal, and the large interest in signal peptide prediction is reflected by the high number of citations to the papers describing the SignalP method [4–6], which has been available online since 1996 and is currently in version 4.1 [7].

SignalP is an example of a signal-based method for protein sorting prediction, where the computational model recognizes the actual sorting signal. The two other approaches are global property-based methods and homology-based methods [8]. Global property-based methods exploit the fact that proteins in different compartments have different physicochemical properties, which is reflected in e.g. different amino acid compositions, especially regarding the surfaces of the proteins [9]. The earliest method for distinguishing between intra- and extracellular proteins based on amino acid and amino acid pair compositions was published in 1994 [10]. Homology-based methods, on the other hand, exploit the fact that proteins tend to stay in the same compartment during the course of evolution, meaning that subcellular location can often be inferred by homology to proteins with known location [11].

However, not all secreted proteins follow the “classical” signal peptide-dependent pathway. An increasing number of eukaryotic proteins have been found to be released without passing the endomembrane system, including proteins with very important functions like cytokines [12]. Such proteins will go undetected by signal peptide-dependent prediction methods such as SignalP.

When attempting to predict which proteins are secreted by unconventional “non-classical” signal peptide-independent routes, especially in eukaryotes, one is faced with two obstacles. First, the signal-based approach is not available, since it is generally not known where in the sequence the signals for secretion occur. Second, the number of experimentally confirmed data from which to build a training set is extremely small.

In bacteria, the situation is different, since there are many more examples known of signal peptide-independent secretion (rarely termed “non-classical” in bacteria). In Gram-negative bacteria, the type I, III, IV, and VI secretion pathways function without signal peptides, and in some cases, there is evidence of N-terminal or C-terminal sorting signals [8,13]. In Gram-positive bacteria, there are also a few known pathways (Wss, holin, and SecA2) [13,14]. This paper will discuss prediction of non-classical secretion in eukaryotes only; prediction in bacteria has been described elsewhere [8,14].

2 The SecretomeP method

SecretomeP is a method from 2004 [15] for predicting non-classically secreted proteins from Mammalia. It was published by our former colleagues in the Center for Biological Sequence Analysis, which later was transformed into Department of Bio and Health Informatics. SecretomeP 2.0, published in 2005 [16], added the possibility for prediction in Gram-positive and Gram-negative bacteria; the mammalian part was not modified or retrained.

The authors chose a novel way to deal with the two obstacles mentioned in the introduction. The method is built upon the hypothesis that extracellular proteins share certain features regardless of the pathway used to secrete them. If this is true, it must be possible to use the large number of known classically secreted proteins to define these features and use them for prediction. Accordingly, the authors extracted a positive training set of extracellular proteins with annotated signal peptides and removed the signal peptide part of the sequence. A negative training set was extracted with subcellular location annotated as cytoplasm and/or nucleus. Both datasets included mammalian proteins only. In addition, a small additional test set of 13 human proteins known to be secreted without a signal peptide was used to evaluate the prediction.

The features were selected from a set of 16 features that were either directly calculated from the sequence (such as number of atoms, theoretical isoelectric point, or number of positively charged residues) or predicted from the sequence (such as secondary structure or phosphorylation sites). Some degree of position-specific information in features such as secondary structure or phosphorylation sites was preserved by dividing the sequence into a number of equal-sized subsequences (bins) and using the average predicted value within each bin as feature values.

The features were subsequently used as inputs to artificial neural networks, which were constructed in a “bottom-up” fashion, inspired by the ProtFun protein function prediction method [17]. First, one network was trained on each feature in isolation; then, the most promising features were combined in pairs; and

again, the most promising feature pairs were selected to build up progressively larger feature combinations, until performance did not improve further. During this process, performance was always measured using five-fold cross-validation on a data set that had been homology partitioned so that no sequence in the test set had more than 26% identity to any sequence in the training set.

The final network has six input features: (1) number of atoms, (2) number of positively charged residues, (3) low-complexity regions assigned by SEG [18] (in five bins), (4) transmembrane helices predicted by TMHMM [19] (in five bins), (5) subcellular localization predicted by PSORT [20], and (6) propeptide cleavage sites predicted by ProP [21] (in five bins).

That the number of atoms has predictive information is not surprising, since extracellular proteins are on average shorter than cytoplasmic and nuclear ones (figure 1 of the SecretomeP paper [15]). The number of positively charged residues is strongly correlated with the number of atoms; but it makes sense that it was precisely this and not the number of negatively charged residues that was selected by the network training procedure, if you consider the “positive-inside” rule of transmembrane proteins which states that positively charged residues are more frequent in the cytoplasmic loops than in the extracellular loops [22]. Accordingly, the SecretomeP authors report that the Arginine plus Lysine content is higher in intracellular than secreted proteins.

Concerning the third feature, low-complexity regions seem to be less prevalent in secreted proteins than in intracellular proteins. This was apparently a novel observation by the SecretomeP authors.

The last three input features are more surprising. Proteins with transmembrane helices predicted by TMHMM were explicitly removed from the negative set in order to keep the network from learning the trivial fact that transmembrane proteins are not extracellular, so there should be no positive predictions by TMHMM in the data. However, the network has apparently utilized the probabilities for “inside” and “outside” given by TMHMM to help classify extracellular proteins – even though the TMHMM authors write in the instructions on their website: “Do not use the program to predict whether a non-membrane protein is cytoplasmic or not” [23].

That PSORT should be selected is also surprising, since that method by itself is not able to classify any of the 13 known human examples of non-classical secretion correctly. There are two old signal peptide predictors built into PSORT [2,3], so it is designed to predict classical secretion. But apparently, cytoplasm probability is after all slightly lower for extracellular proteins without their signal peptides than it is for cytoplasmic and nuclear proteins. The PSORT feature showed high correlation to the TMHMM feature.

Finally, the propeptides predicted by ProP are of the type recognized by members of the subtilisin/kexin-like proprotein convertase family, which is active in the secretory pathway. The surprising aspect here is that the number of predicted propeptide cleavage sites is actually *lower* in secretory proteins than in intracellular proteins. This might reflect the fact that the majority of the recognized cleavage sites are dibasic, leading to a higher number of false positive predictions in intracellular proteins due to the higher Lys+Arg content described above.

The predictive performance of SecretomeP is summarized in a Receiver Operating Characteristic (ROC) curve (figure 3 of the SecretomeP paper [15] – note that the curve shows false positive rate as a function of sensitivity where the convention is the exact opposite). As is remarked in the text, at a false positive rate of

5%, 40% of the positive examples are predicted. However, it is not clear whether this point on the curve corresponds to the recommended cutoff of 0.6. The reason for choosing 0.6 is unknown, and the false positive rate at this cutoff is not given.

Among the 13 known human examples of non-classical secretion, ten were positively predicted using the recommended cutoff of 0.6. A smoothed curve of the score distribution for these 13 sequences overlaps nicely with the score distribution of the positive training set (figure 4 of the SecretomeP paper [15]). These two observations together are taken as a confirmation of the underlying hypothesis that secreted proteins share characteristics regardless of the pathway used to secrete them.

3 Other dedicated methods

Besides SecretomeP, we are aware of five other published methods specifically designed to predict secretion without signal peptides in eukaryotes. These predictive tools have been summarized in Table 1.

Interestingly, all these methods, like SecretomeP, focus on mammalian proteins alone; no method is available for non-mammal eukaryotes. However, none of the papers actually argue for that choice or cite any references showing that non-classical secretion in mammals differs from the process in, e.g., birds, insects, fungi, or plants.

3.1 SRTpred

SRTpred from 2008 [24] used the SecretomeP dataset. In contrast to SecretomeP, the goal was not explicitly to make a predictor for non-classical secretion, but an overall predictor for secretion that did not rely on signal peptides. As the authors correctly remark, large scale genome sequencing projects sometimes assign the 5'-end of coding regions incorrectly, which can easily lead to missed signal peptides. Accordingly, the authors used a set of features that should be independent of signal peptides: 33 physicochemical properties averaged per sequence, amino acid composition, dipeptide (ungapped amino acid pair) composition, and sequence similarity to known proteins from the data set, measured by BLAST or PSI-BLAST [25].

However, the SRTpred authors chose to use the entire sequences instead of cutting off the predicted signal peptides like the SecretomeP authors did. This means that especially for short proteins, the signal peptides are allowed to influence the composition and physicochemical properties, making a direct comparison to SecretomeP performance problematic.

The SRTpred authors first tried artificial neural networks (ANNs), but found support vector machines (SVMs) to perform better. The final SRTpred method is an SVM integrating amino acid composition, dipeptide composition, and PSI-BLAST, so it is partly a homology-based method. The sensitivity of this hybrid method at a 5% false positive rate is 60%. Without the PSI-BLAST input, the corresponding rate is reported to be 44% – only slightly better than SecretomeP.

Keep in mind that when the focus is on predicting non-classically secreted proteins, the PSI-BLAST module is expected to be of little value, since the database of known proteins does not contain such proteins.

SRTpred is available as a web server, but it has the drawback of only being able to process one sequence per submission (where SecretomeP can process up to 100).

3.2 SecretP

Parallel to the development in SecretomeP, SecretP version 1 [26] is for mammalian proteins, while version 2 [27] is for bacteria. SecretP version 1 from 2010 aims to distinguish between three groups of proteins: classically secreted, non-classically secreted and non-secreted. For the first and last groups, the SecretomeP datasets were used. Unfortunately, the description of how the dataset of non-classically secreted proteins were extracted is lacking in detail. Two approaches are mentioned, where the first one is simply described thus: “Firstly, 864 mammalian proteins confirmed to route in non-classical secretory pathways were collected from Swiss-Prot through data mining”. 149 human proteins were put aside as a test set. In the second approach, a “secreted” keyword plus the *absence* of a signal peptide annotation was used in the selection. Using an absence of an annotation as a criterion is always risky, since the absence might simply reflect an incomplete annotation instead of a real absence of the feature.

The two approaches together gave a data set of 1248 non-classically secreted proteins. After homology reduction to 25% identity, there were 230 proteins left in the cross-validation set, and 92 in the exclusively human test set. Unfortunately, it is not clear whether homology reduction was only done *within* the two sets, or also *between* the cross-validation and the test set.

The features used in SecretP are amino acid composition and auto-covariance of seven physicochemical properties, fused into what is known as pseudo-amino acid composition [28]. In addition, five more features are used: signal peptides (predicted by SignalP 3.0 [5]), secondary structure content (predicted by SSCP [29] from amino acid composition alone), number of positively charged residues, isoelectric point, and subcellular localization (predicted by WoLF PSORT [30]). No selection process is described; these five features are apparently chosen manually. All the features are then used as inputs to an SVM.

The cross-validated performance of SecretP is reported to be 88.79% correct in the three categories. In the “independent” human test set, 76 out of 92 were correctly predicted to be non-classically secreted (83%). SecretomeP only predicted 50 of these 92 correctly (54%). The reason for the scare quotes around “independent” is that we are not sure whether there were homologous sequences in the human test set with more than 25% identity to sequences in the cross-validation set.

Like SRTpred, the SecretP web server can process only one sequence per submission. In addition, SecretP is currently broken, reporting an “internal server error” when a sequence is submitted.

3.3 SPRED

SPRED [31] is a random forest classifier for predicting both conventional and unconventional protein secretion in mammals. The authors extracted 780 extracellular proteins and 1980 intracellular proteins from UniProt by keyword searching and similarity reduction. 180 extracellular and 1380 intracellular proteins were kept as testing data. Just like in SecretomeP, the signal peptides of the proteins in the extracellular group were removed.

In total, 119 features were constructed for representing proteins, which include frequencies of amino acids in 10 functional groups and 7 physicochemical properties (hydrophobic, hydrophilic, neutral, positively charged, negatively charged, polar and non-polar amino acids), frequencies of structural elements and frequencies of short peptides and dipeptides and finally 31 physicochemical features selected from AAIndex [32]. The frequencies of amino acids were calculated both on the whole sequence level and within

various structural elements. An information gain-based criterion was used to select top features for building SPRED. The top 10 features achieved 80.38% accuracy on the test data. Adding more features up until a set of 75 increased the accuracy to 82.31%. Due to the limited dataset size, keeping on adding features increased the model complexity and overfitting emerged.

The authors also constructed a set of 19 proteins which were experimentally confirmed to be non-classically secreted for comparing the predictive power of SPRED to that of SRTPred and SecretomeP. SPRED correctly predicted 15 of them to be unconventionally secreted; SecretomeP predicted 13 whereas SRTPred predicted 5 proteins. The 19 proteins are an extension of the 13 proteins used in the SecretomeP paper, however, one of the 19 proteins (CALR_HUMAN) was later found to carry a signal peptide [33]. The remaining 18 proteins and their predictions are listed in Table 2.

SPRED [31] is available as a downloadable program, but in our experience, it does not work “out of the box”. It took some guidance, kindly provided by the first author, before we could make it run.

3.4 Sec-GO

Sec-GO [34] followed a totally different approach for the prediction of unconventional protein secretion for both mammals and Gram-positive bacteria by using gene ontology (GO) annotations [35]. In order to train the GO-based SVM models and benchmark with other existing methods, the author made use of the SPRED dataset with more stringent similarity reduction of 25% identity as training and test data for mammals. Each protein was represented by 60020 GO terms, which were encoded as a large sparse vector. A dimension reduction of GO feature space was applied due to the small dataset size. The author used the frequency difference of the same GO term between positive and negative datasets as the score for this term. This score stood for the discriminative power of the corresponding term for the positive and negative datasets. Then all terms were ranked according to their scores and eventually 436 GO terms were used for the mammalian data set. The top scoring GO-terms were vectorized and fed directly to an SVM for optimizing the model. The author reported that by analyzing feature contributions, the GO term “extracellular” was the most important, which was straightforward and intuitive; however, the approach did not give indications of what factors of sequences might lead to the extracellular location, which is of central interest for predicting protein secretion.

Sec-GO achieved for all manually or automatically GO annotated proteins an accuracy of 96.7% of mammal testing data compared to that of 82.2% from SPRED. Furthermore, a benchmark among SecretomeP, SRTpred, SPRED and Sec-GO on the 19 unconventionally secreted proteins from SPRED showed that Sec-GO remained the top one (see Table 2). However, the requirement for already existing GO terms makes it hard to use the approach for novel proteins.

The Sec-GO web server is no longer found at the address given in the paper, which makes it hard to evaluate Sec-GO’s efficacy on other datasets.

3.5 Hung et al. 2009

In this study [36], the authors made use of SecretomeP 1.0’s dataset for training SVM models on 30 features which were selected from physicochemical properties summarized in AAIndex. The feature selection and model parameter tuning were encoded as binary genes in an inheritable bi-objective genetic algorithm, which made this work different from others. The prediction accuracies for non-secretory

proteins and secretory proteins were 90.16% and 76.17%, respectively. However, the authors imposed a more stringent sequence similarity reduction to less than 25% identity, which made a direct comparison of performance to SecretomeP difficult. This unnamed method has never been made available as a web server or a downloadable program.

4 Multi-location predictors

Besides methods that predict whether or not a protein is secreted, there are also several methods available which predict a larger number of subcellular locations, including “secreted” or “extracellular”. Such multi-location predictors could potentially also be used to predict secretion without signal peptides. However, since the majority of secreted proteins have signal peptides, some kind of signal peptide prediction will usually be built into such methods, either implicitly or explicitly. If signal peptide prediction is essential for predicting the “secreted” or “extracellular” category, chances are not very high that the method will be useful for predicting unconventional protein secretion.

On the other hand, several of the multi-location predictors include some homology-derived features. The simplest approach is taken by the LocTree3 method [37] which directly uses the annotated subcellular location of the best hit in a PSI-BLAST search [25], while other methods use derived features of retrieved database hits such as Gene Ontology (GO) terms [35]. Such approaches could be useful in identifying non-classically secreted proteins, if they have close homologues that are known to be secreted. However, homology-based methods offer no new insights into the secretion signals or specific properties of non-classically secreted proteins, and they will have very limited chance of being able to predict the consequences of mutations affecting sorting signals because the wild-type and the variant would probably pick up the same homologues in a database search.

The predictors mentioned here are summarized in Table 3. In contrast to Table 1, this list is not meant to be complete; we have selected the most important and most used methods.

WoLF PSORT [30] is a successor to PSORT / PSORT II [20] for eukaryotic proteins. It is based on a combination of sequence-derived features and amino acid composition, integrated via a weighted version of the k nearest neighbours classifier. There are three features explicitly referring to signal peptides: the two old signal peptide predictors built into PSORT [2,3] and the signal peptide probability from iPSORT [38].

MultiLoc2 [39] is an SVM-based method integrating various sequence-derived features with amino acid composition, GO terms of homologues, and phylogenetic profiles. It also has explicit signal peptide prediction built into the model via the SVMTarget feature. SherLoc2 [40] extends the MultiLoc2 model by integrating also text mining of PubMed abstracts linked to the Swiss-Prot entries of retrieved homologues.

YLoc [41] is based on feature selection from a very large set of initial sequence-derived features. The selected features are subsequently combined via a naïve Bayes model, which makes it possible to indicate for each prediction which features were important. Certain of the selected features are clearly correlated to signal peptides. YLoc can optionally include GO terms of homologues.

iLoc-Euk [42] is a k nearest neighbours-based method mainly using GO terms of homologues, with additional evolutionary profiles used only in those cases where no homologues are found, or when the found homologues do not have GO annotation.

CELLO [43] is an SVM-based method that neither uses homology information nor has a built-in signal peptide model. It uses amino acid composition, amino acid pair composition, and n -peptide composition with reduced alphabets. However, it does also use partitioned amino acid composition, where the sequence is divided into a number of subsequences of equal length (like the bins of SecretomeP) and amino acid composition is calculated separately in each partition; a measure which could be influenced by the presence of signal peptides.

LocTree3 [37], as already mentioned, uses a direct transfer of subcellular location annotation from the best PSI-BLAST hit, if the significance of the hit is better than a certain E-value threshold. If this is not the case, it reverts to LocTree2 [44], which is an SVM-based method using a so-called profile kernel, a kind of string kernel based on sequence profiles found in a PSI-BLAST search. It is not easy to say whether the profile kernel approach recognizes signal peptides.

SubCons [45] is a consensus method incorporating predictions from CELLO, LocTree2, MultiLoc2 and SherLoc2. Its performance has been optimised on a set of human proteins, but it can be used for other eukaryotes also.

Finally, DeepLoc [46] is a method based on deep learning (convolutional and recurrent neural networks) without using annotation of homologues. It does not have an explicit signal peptide model, but it is apparent from the attention score output that it does seem to look specifically at the signal peptide region when predicting extracellular proteins.

5 A critical re-evaluation of SecretomeP performance

In the years since SecretomeP was first developed a lot more data has become available for protein sequences that are secreted in a non-classical manner. In addition, one common question addressed to the curators of the SecretomeP web service is whether it performs equally well for all eukaryotic sequences as it does for mammalian sequences. As such, an opportunity has presented itself for a critical reevaluation of SecretomeP's performance.

We collected two data sets from UniProt, one with mammalian protein sequences and one with eukaryotic sequences excluding mammalian. For each data set, the positive sub-set consisted of manually reviewed secreted sequences that lacked signal peptide annotation, and the negative sub-set consisted of protein sequences experimentally verified to be located in the cytoplasm or the nucleus, also lacking signal peptide annotation. Additional filtering was performed by excluding protein sequences that appeared to be fragments (not starting with a methionine) and sequences that were predicted to have a signal peptide by SignalP-3.0. The final data sets consisted of 543 non-classically secreted and 5997 non-secreted mammalian protein sequences, and 236 non-classically secreted and 7262 non-secreted eukaryotic (excluding mammalian) protein sequences.

In Table 4 you can see the performance of SecretomeP on these two data sets when using the recommended threshold (0.6) for the NN-score. As expected, SecretomeP performs better for mammalian sequences than other eukaryotic sequences, although the differences are not very significant. The most surprising is the low sensitivity and high false positive rate (>20%) in both cases.

This is underlined by their Receiving Operating Characteristic (ROC) curves (Figure 1). A ROC curve is made by varying the threshold for regarding a prediction as positive and plotting the ensuing sensitivity as a function of the false positive rate. The area under the curve (AUC) can then be used as a threshold-independent measure of predictive performance. The AUC can be directly interpreted as the possibility that a randomly chosen positive example will score higher than a randomly chosen negative example. A perfect prediction will have $AUC = 1$, while a random guess will give $AUC = 0.5$. The AUC of the SecretomeP ROC curves in Figure 1 is for both around 0.6, indicating a low discriminatory ability, only slightly better than a random classifier.

It should be noted that the performance was significantly better when truncated proteins were included (i.e. proteins not starting with methionine), with the ROC AUC equal to 0.67 and 0.78 for mammalian and other eukaryotic sequences respectively. This suggests that these proteins are in fact secreted in a classical manner, but the signal peptide in the N-terminus has been removed. It also leads to the conclusion that there are faults in the initial hypothesis behind the design of the SecretomeP, and that proteins that are secreted in a non-classical manner do not share as many features with the classically secreted proteins as initially considered.

The performances of SRTpred and SPRED are also shown in Table 4. SRTpred predicts slightly fewer mammalian non-classically secreted proteins than SecretomeP, but at a lower false positive rate. SPRED seems to be a bit better than the other two, detecting around half of the positive examples, but it still has a false positive rate of 14% for mammalian and 20% for non-mammalian proteins. It was not possible to draw ROC curves and calculate AUC values for SRTpred and SPRED, since they don't provide numeric output.

In addition to the three dedicated servers, we also benchmarked three of the multi-location predictors discussed in the previous section. In this analysis, we did not divide the data into Mammalia and other eukaryotes, since the predictors are trained on all eukaryotes together. The results are shown in Table 5.

From the table, it is apparent that CELLO and DeepLoc, using their default output, identify fewer non-classically secreted proteins than SecretomeP at the default threshold, but at a much lower false positive rate. DeepLoc thus identifies 11% of the secreted proteins practically without false positives. If we instead of using the most probable class from the methods use the numerical score for the "Extracellular" category, we can calculate ROC curves (shown in Figure 2), which show that CELLO and DeepLoc are actually better than SecretomeP in predicting non-classically secreted proteins.

Judged from the table, iLoc-Euk is even better, identifying almost half of the non-classically secreted proteins at a false positive rate of less than 2%. This is surprising, since DeepLoc is reported to be better than iLoc-Euk in general [46]. However, it should be kept in mind that iLoc-Euk is a homology-based method, retrieving GO terms of homologues from a UniProt-derived database, and there may be a considerable overlap between that database and our test set. For the same reason, we did not benchmark

LocTree3, since many of the predictions from that tool would be simple database retrievals of the annotations contained in our test set.

We would have liked to benchmark more methods, but time constraints and technical difficulties did not allow it. As an example, all the University of Tübingen servers (MultiLoc2, SherLoc2 and YLoc) were temporarily down at the time of writing, and although the text on the website says “We'll be back in a few days”, this was still the case at the time of revision.

6 Discussion

SecretomeP version 1 was, for its time, a bold and innovative suggestion for how to construct a predictor for secretion without signal peptides. It has been cited more than 800 times according to Google Scholar, and it is still being used extensively. However, its performance, measured on new independent data, is not nearly as good as we thought it would be, and the underlying hypothesis that extracellular proteins share features independent of the secretion pathway must be called into question.

SRTpred and SPRED do not represent real alternatives for predicting non-classical secretion, as they are built on the same questionable hypothesis and only perform marginally better. The small performance gain shown by SRTpred may even be attributed to the fact that the signal peptides were not removed in the training. SPRED seems to represent a more genuine performance gain, but is still limited by the constraints of the “common feature” hypothesis. Sec-GO represents an interesting analysis, but is not applicable in practice to the situation where a predictor would be most important, namely newly sequenced genomes.

SecretP, on the other hand, might be significantly better than SecretomeP, SRTpred and SPRED, but it is difficult to say how much confidence should be put in their data set. Given more time, the set of 864 mammalian proteins, available from the SecretP website, should be critically examined. Unfortunately, it was not possible for us to benchmark SecretP due to the restrictions on the website (1 sequence per submission, 50 sequences per day) and the fact that the webserver reports an error when you attempt to run it.

The three multi-location predictors that we benchmarked performed better than SecretomeP, even though they were not made with non-classical secretion in mind, but the performance is still not high. The best of them, iLoc-Euk, may have an inflated performance due to overlap between its database and our benchmark dataset. All in all, it is fair to say that prediction of non-classical (signal peptide-independent) secretion in eukaryotes is an unsolved problem.

However, the novel deep learning techniques (convolutional and recurrent neural networks) used in DeepLoc [46] has shown promising results for predicting signal peptide-independent secretion in bacteria (E. I. Petsalaki, J. J. Almagro Armenteros, O. Winther and H. Nielsen, unpublished results). In the future, we will apply this approach to eukaryotic non-classical secretion as well. The networks in DeepLoc have, in addition to the convolutional and LSTM (long short-term memory) recurrent layers, a so-called attention layer which calculates a relative weight for each position in the sequence. These attention weights can, for each prediction, point out which parts of the sequence were important for reaching that particular prediction. In this way, a deep neural network trained on non-classically secreted proteins could not only

give a prediction of secretion but also help localizing possible signals for non-classical secretion in the sequence.

Acknowledgements

The corresponding author is paid by the Technical University of Denmark. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors wish to thank Krishna Kumar Kandaswamy from the SPRED team for assistance with running the program.

References

- [1] G. von Heijne, Patterns of Amino Acids near Signal-Sequence Cleavage Sites, *Eur. J. Biochem.* 133 (1983) 17–21. doi:10.1111/j.1432-1033.1983.tb07424.x.
- [2] D.J. McGeoch, On the predictive recognition of signal peptide sequences, *Virus Res.* 3 (1985) 271–286. doi:10.1016/0168-1702(85)90051-6.
- [3] G. von Heijne, A new method for predicting signal sequence cleavage sites, *Nucleic Acids Res.* 14 (1986) 4683–4690. doi:10.1093/nar/14.11.4683.
- [4] H. Nielsen, S. Brunak, J. Engelbrecht, G. von Heijne, Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, *Protein Eng.* 10 (1997) 1–6. doi:10.1093/protein/10.1.1.
- [5] J.D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, Improved prediction of signal peptides: SignalP 3.0, *J. Mol. Biol.* 340 (2004) 783–95. doi:10.1016/j.jmb.2004.05.028.
- [6] T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions, *Nat Meth.* 8 (2011) 785–786. doi:10.1038/nmeth.1701.
- [7] H. Nielsen, Predicting Secretory Proteins with SignalP, in: D. Kihara (Ed.), *Protein Funct. Predict.*, Springer New York, 2017: pp. 59–73. doi:10.1007/978-1-4939-7015-5_6.
- [8] H. Nielsen, Protein Sorting Prediction, in: L. Journet, E. Cascales (Eds.), *Bact. Protein Secret. Syst.*, Humana Press, New York, NY, 2017: pp. 23–57. doi:10.1007/978-1-4939-7033-9_2.
- [9] M.A. Andrade, S.I. O'Donoghue, B. Rost, Adaptation of protein surfaces to subcellular location, *J. Mol. Biol.* 276 (1998) 517–525. doi:10.1006/jmbi.1997.1498.
- [10] H. Nakashima, K. Nishikawa, Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies, *J Mol Biol.* 238 (1994) 54–61. doi:10.1006/jmbi.1994.1267.
- [11] R. Nair, B. Rost, Sequence conserved for subcellular localization, *Protein Sci.* 11 (2002) 2836–2847. doi:10.1110/ps.0207402.
- [12] E.H. Duitman, Z. Orinska, S. Bulfone-Paus, Mechanisms of cytokine secretion: A portfolio of distinct pathways allows flexibility in cytokine activity, *Eur. J. Cell Biol.* 90 (2011) 476–483. doi:10.1016/j.ejcb.2011.01.010.
- [13] M. Desvaux, M. Hébraud, R. Talon, I.R. Henderson, Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue, *Trends Microbiol.* 17 (2009) 139–145. doi:10.1016/j.tim.2009.01.004.
- [14] H. Nielsen, Predicting Subcellular Localization of Proteins by Bioinformatic Algorithms, in: F. Bagnoli, R. Rappuoli (Eds.), *Protein Export Gram-Posit. Bact.*, Springer, 2016. doi:10.1007/82_2015_5006.
- [15] J.D. Bendtsen, L.J. Jensen, N. Blom, G. von Heijne, S. Brunak, Feature-based prediction of non-classical and leaderless protein secretion, *Protein Eng. Des. Sel.* 17 (2004) 349–356. doi:10.1093/protein/gzh037.
- [16] J.D. Bendtsen, L. Kiemer, A. Fausbøll, S. Brunak, Non-classical protein secretion in bacteria, *BMC Microbiol.* 5 (2005) 58. doi:10.1186/1471-2180-5-58.

- [17] L.J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H.H. Staerfeldt, K. Rapacki, C. Workman, C.A.F. Andersen, S. Knudsen, A. Krogh, A. Valencia, S. Brunak, Prediction of human protein function from post-translational modifications and localization features, *J. Mol. Biol.* 319 (2002) 1257–65. doi:10.1016/S0022-2836(02)00379-0.
- [18] J.C. Wootton, S. Federhen, Statistics of local complexity in amino acid sequences and sequence databases, *Comput. Chem.* 17 (1993) 149–163. doi:10.1016/0097-8485(93)85006-X.
- [19] A. Krogh, B. Larsson, G. von Heijne, E.L.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J. Mol. Biol.* 305 (2001) 567–580. doi:10.1006/jmbi.2000.4315.
- [20] K. Nakai, P. Horton, PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, *Trends Biochem Sci.* 24 (1999) 34–35. doi:10.1016/S0968-0004(98)01336-X.
- [21] P. Duckert, S. Brunak, N. Blom, Prediction of proprotein convertase cleavage sites, *Protein Eng. Des. Sel.* 17 (2004) 107–112. doi:10.1093/protein/gzh013.
- [22] G. von Heijne, Membrane protein structure prediction: Hydrophobicity analysis and the positive-inside rule, *J. Mol. Biol.* 225 (1992) 487–494. doi:10.1006/jmbi.1992.2250487.
- [23] TMHMM 2.0 Guide, (n.d.). <http://www.cbs.dtu.dk/services/TMHMM/TMHMM2.0b.guide.php> (accessed June 10, 2018).
- [24] A. Garg, G.P.S. Raghava, A Machine Learning Based Method for the Prediction of Secretory Proteins Using Amino Acid Composition, Their Order and Similarity-Search, *In Silico Biol.* 8 (2008) 129–140.
- [25] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25 (1997) 3389. doi:10.1093/nar/25.17.3389.
- [26] L. Yu, Y. Guo, Z. Zhang, Y. Li, M. Li, G. Li, W. Xiong, Y. Zeng, SecretP: A new method for predicting mammalian secreted proteins, *Peptides.* 31 (2010) 574–578. doi:10.1016/j.peptides.2009.12.026.
- [27] L. Yu, Y. Guo, Y. Li, G. Li, M. Li, J. Luo, W. Xiong, W. Qin, SecretP: Identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition, *J. Theor. Biol.* 267 (2010) 1–6. doi:10.1016/j.jtbi.2010.08.001.
- [28] K.-C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, *Proteins.* 43 (2001) 246–255. doi:10.1002/prot.1035.
- [29] F. Eisenhaber, F. Imperiale, P. Argos, C. Frömmel, Prediction of secondary structural content of proteins from their amino acid composition alone. I. New analytic vector decomposition methods, *Proteins Struct. Funct. Bioinforma.* 25 (1996) 157–168. doi:10.1002/(SICI)1097-0134(199606)25:2<157::AID-PROT2>3.0.CO;2-F.
- [30] P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C.J. Adams-Collier, K. Nakai, WoLF PSORT: protein localization predictor, *Nucleic Acids Res.* 35 (2007) W585–W587. doi:10.1093/nar/gkm259.
- [31] K.K. Kandaswamy, G. Pugalenthi, E. Hartmann, K.-U. Kalies, S. Möller, P.N. Suganthan, T. Martinetz, SPRED: A machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes, *Biochem. Biophys. Res. Commun.* 391 (2010) 1306–1311. doi:10.1016/j.bbrc.2009.12.019.
- [32] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, *Nucleic Acids Res.* 36 (2008) D202–D205. doi:10.1093/nar/gkm998.
- [33] G. Houen, C. Koch, Human placental calreticulin: purification, characterization and association with other proteins., *Acta Chem. Scand. Cph. Den.* 48 (1994) 905–911.
- [34] W.-L. Huang, Ranking Gene Ontology terms for predicting non-classical secretory proteins in eukaryotes and prokaryotes, *J. Theor. Biol.* 312 (2012) 105–113. doi:10.1016/j.jtbi.2012.07.027.
- [35] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene Ontology: tool for the unification of biology, *Nat. Genet.* 25 (2000) 25–29. doi:10.1038/75556.

- [36] C.-H. Hung, H.-L. Huang, K.-T. Hsu, S.-J. Ho, S.-Y. Ho, Prediction of non-classical secreted proteins using informative physicochemical properties, *Interdiscip. Sci. Comput. Life Sci.* 2 (2010) 263–270. doi:10.1007/s12539-010-0023-z.
- [37] T. Goldberg, M. Hecht, T. Hamp, T. Karl, G. Yachdav, N. Ahmed, U. Altermann, P. Angerer, S. Ansorge, K. Balasz, M. Bernhofer, A. Betz, L. Cizmadija, K.T. Do, J. Gerke, R. Greil, V. Joerdens, M. Hastreiter, K. Hembach, M. Herzog, M. Kalemánov, M. Kluge, A. Meier, H. Nasir, U. Neumaier, V. Prade, J. Reeb, A. Sorokoumov, I. Troshani, S. Vorberg, S. Waldraff, J. Zierer, H. Nielsen, B. Rost, LocTree3 prediction of localization, *Nucleic Acids Res.* 42 (2014) W350–W355. doi:10.1093/nar/gku396.
- [38] H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, S. Miyano, Extensive feature detection of N-terminal protein sorting signals, *Bioinformatics.* 18 (2002) 298–305. doi:10.1093/bioinformatics/18.2.298.
- [39] T. Blum, S. Briesemeister, O. Kohlbacher, MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction, *BMC Bioinformatics.* 10 (2009) 274. doi:10.1186/1471-2105-10-274.
- [40] S. Briesemeister, T. Blum, S. Brady, Y. Lam, O. Kohlbacher, H. Shatkay, SherLoc2: A High-Accuracy Hybrid Method for Predicting Subcellular Localization of Proteins, *J. Proteome Res.* 8 (2009) 5363–5366. doi:10.1021/pr900665y.
- [41] S. Briesemeister, J. Rahnenführer, O. Kohlbacher, Going from where to why—interpretable prediction of protein subcellular localization, *Bioinformatics.* 26 (2010) 1232–1238. doi:10.1093/bioinformatics/btq115.
- [42] K.-C. Chou, Z.-C. Wu, X. Xiao, iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins, *PLoS ONE.* 6 (2011) e18258. doi:10.1371/journal.pone.0018258.
- [43] C.-S. Yu, Y.-C. Chen, C.-H. Lu, J.-K. Hwang, Prediction of protein subcellular localization, *Proteins.* 64 (2006) 643–651. doi:10.1002/prot.21018.
- [44] T. Goldberg, T. Hamp, B. Rost, LocTree2 predicts localization for all domains of life, *Bioinformatics.* 28 (2012) i458–i465. doi:10.1093/bioinformatics/bts390.
- [45] M. Salvatore, P. Warholm, N. Shu, W. Basile, A. Elofsson, SubCons: a new ensemble method for improved human subcellular localization predictions, *Bioinformatics.* 33 (2017) 2464–2470. doi:10.1093/bioinformatics/btx219.
- [46] J.J. Almagro Armenteros, C.K. Sønderby, H. Nielsen, O. Winther, DeepLoc: prediction of protein subcellular localization using deep learning, *Bioinformatics.* 33 (2017) 3387–3395. doi:10.1093/bioinformatics/btx431.

Tables

Table 1: Summary of predictive tools dedicated for predicting unconventional protein secretion in eukaryotes. Abbreviations used: ANN, Artificial Neural Network; SVM, Support Vector Machine.

Method	Year	Model	Availability	Link
SecretomeP	2004 [15]	ANN	Web and Standalone	http://www.cbs.dtu.dk/services/SecretomeP/
SRTpred	2008 [24]	SVM, homology	Web	http://crdd.osdd.net/raghava/srtpred/
SecretP (v1)	2010 [26]	SVM	Web (Error)	http://cic.scu.edu.cn/bioinformatics/secretP/ (Internal Server Error)
SPRED	2010 [31]	Random Forest	Standalone	http://www.inb.uni-luebeck.de/tools-demos/spred/spred
Hung et al.	2010 [36]	SVM	No	-
Sec-GO	2012 [34]	SVM, GO- annotations	Web (not accessible)	https://iclab.life.nctu.edu.tw/secgo (404)

Table 2: The prediction results on 18 experimentally confirmed human non-classically secreted proteins. There were originally 19 proteins in the list in the SPRED paper, but one (CALR_HUMAN) has a signal peptide with experimental evidence in UniProt and has therefore been removed. Numerical output scores are given for SecretomeP and SRTpred. “+” and “-” indicate true positive or false negative predictions of unconventional secretion, respectively.

UniProt ID	UniProt AC	SecretomeP	SRTpred	SPRED	Sec-GO
FGF1_HUMAN	P05230	0.847 (+)	-0.81 (-)	+	+
FGF2_HUMAN	P09038	0.239 (-)	0.80 (+)	+	+
IL1B_HUMAN	P01584	0.610 (+)	0.96 (+)	+	+
IL1A_HUMAN	P01583	0.551 (-)	-0.2 (-)	+	+
LEG3_HUMAN	P17931	0.770 (+)	-1.16 (-)	+	-
MIF_HUMAN	P14174	0.776 (+)	-0.91 (-)	+	+
S10A4_HUMAN	P26447	0.724 (+)	-0.55 (-)	+	+
GSTP1_HUMAN	P09211	0.545 (-)	-0.7 (-)	+	+
PRDX1_HUMAN	Q06830	0.528 (-)	-0.94 (-)	+	+
IL18_HUMAN	Q14116	0.634 (+)	-1 (-)	+	+
H4_HUMAN	P62805	0.408 (-)	-1.12 (-)	+	+
S10A2_HUMAN	P29034	0.324 (-)	-0.48 (-)	+	+
LEG1_HUMAN	P09382	0.345 (-)	-0.62 (-)	+	+
THIO_HUMAN	P10599	0.370 (-)	0.71 (+)	+	+
CNTF_HUMAN	P26441	0.653 (+)	0.02 (+)	-	+
HME2_HUMAN	P19622	0.727 (+)	-1.39 (-)	-	+
THTR_HUMAN	Q16762	0.616 (+)	-1.2 (-)	-	+
HMGB1_HUMAN	P09429	0.068 (-)	-1.2 (-)	-	+

Table 3: Summary of selected predictors for multi-category protein localization in eukaryotes. Abbreviations used: ANN, Artificial Neural Network; k-NN, *k* Nearest Neighbours; SVM, Support Vector Machine.

Method	Year	Model	Availability	Link
CELLO	2006 [43]	SVM	Web	http://cello.life.nctu.edu.tw/
WoLF PSORT	2007 [30]	k-NN	Web	https://wolfpsort.hgc.jp/
MultiLoc2	2009 [39]	SVM, homology	Web (not accessible)	https://abi.inf.uni-tuebingen.de/Services (temporarily disabled)
SherLoc2	2009 [40]	SVM, homology	Web (not accessible)	https://abi.inf.uni-tuebingen.de/Services (temporarily disabled)
YLoc	2010 [41]	SVM, homology (optional)	Web (not accessible)	https://abi.inf.uni-tuebingen.de/Services (temporarily disabled)
iLoc-Euk	2011 [42]	k-NN, homology	Web	http://www.jci-bioinfo.cn/iLoc-Euk
LocTree3	2014 [37]	Homology, profile kernel SVM	Web and Standalone	https://roslab.org/services/loctree3/
SubCons	2017 [45]	Consensus	Web and Standalone	http://subcons.bioinfo.se/
DeepLoc	2017 [46]	Deep ANN	Web and Standalone	http://www.cbs.dtu.dk/services/DeepLoc/

Table 4: Performance of SecretomeP, SRTpred, and SPRED on two data sets comprising mammalian proteins and other eukaryotic proteins, respectively. AUC could not be calculated for SRTpred and SPRED since they do not provide numeric output values. The value marked by “*” is an estimate based on a subset of the negative data, since we had technical problems running SRTpred on the whole dataset. For the same reason, the TNR could not be calculated for the other eukaryotic proteins (marked “N/A*”). Abbreviations used: TPR, True Positive Rate (Sensitivity); TNR, True Negative Rate (Specificity); AUC, Area Under the receiver operating characteristic Curve.

Data	Method	TPR	TNR	AUC
Mammalia	SecretomeP	42.9 %	78.7 %	0.61
	SRTpred	38.3 %	86.7 %*	-
	SPRED	48.6 %	86.0 %	-
Eukaryota (excl. Mammalia)	SecretomeP	35.6 %	75.0 %	0.60
	SRTpred	35.2 %	N/A *	-
	SPRED	52.5 %	80.0 %	-

Table 5: Performance of three multi-category localization predictors on our new set of non-classically secreted proteins. Data were eukaryotic sequences (mammalian and non-mammalian combined). AUC could not be calculated for iLoc-Euk since it does not provide numeric output values. The value marked by “*” is an estimate based on a subset of the negative data, since we had technical problems running iLoc-Euk on the whole dataset. Abbreviations used: TPR, True Positive Rate (Sensitivity); TNR, True Negative Rate (Specificity); AUC, Area Under the receiver operating characteristic Curve.

Method	TPR	TNR	AUC
CELLO	18.0 %	97.2 %	0.73
DeepLoc	10.9%	99.8 %	0.72
iLoc-Euk	48.4%	98.6 %*	-

Figure captions

A File: Figure1A_ROC_SecretomeP_mam.eps

B File: Figure1B_ROC_SecretomeP_euk.eps

Figure 1: ROC curves (true positive rate vs. false positive rate) for SecretomeP predictions on new datasets. Panel A: mammalian protein sequences; panel B: eukaryotic protein sequences, excluding Mammalia. The dashed lines represent the theoretical performance of a random guess.

A File: Figure2A_ROC_CELLO_euk.eps

B File: Figure2B_ROC_DeepLoc_euk.eps

Figure 2: ROC curves (true positive rate vs. false positive rate) for those multi-location predictors where it was possible to get numeric scores. Data were eukaryotic sequences (mammalian and non-mammalian combined). Panel A shows results for CELLO, while panel B shows results for DeepLoc. The dashed lines represent the theoretical performance of a random guess.

Figure 1A

Figure 1B

SecretomeP for eukaryotic protein sequences (excl. mammalian)

Figure 2A

Figure 2B