Improving identification of animal secretory proteins

Jiratchaya Nuanpirom [1] and Unitsa Sangket [2,*]

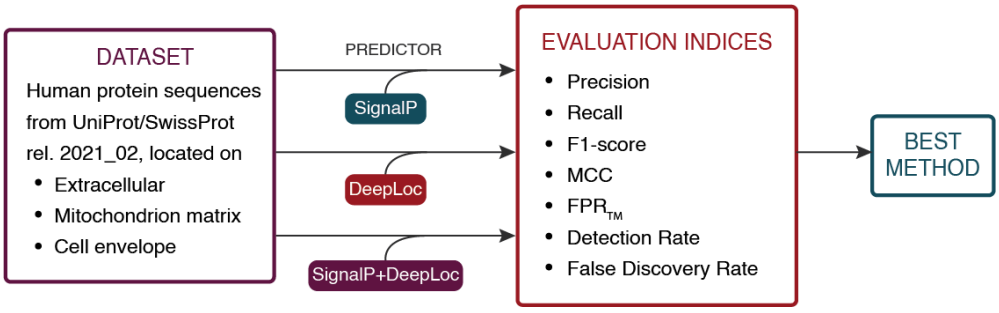
[1] Division of Biological Science, Faculty of Science, Prince of Songkla University, Hat Yai, Thailand. 90110; jirath.nuan@gmail.com

[2] Center for Genomics and Bioinformatics Research, Division of Biological Science, Faculty of Science, Prince of Songkla University, Hat Yai, Thailand. 90110; unitsa.s@psu.ac.th

* Correspondence: unitsa.s@psu.ac.th

Abstract: Nascent protein translated inside cells is controlled by a signal peptide to various cellular compartments, including the secretory protein. A preproprotein sequence has a signal peptide flanking it at the N-terminus. To distinguish the signal peptide from proteins that migrate to other compartments, several state-of-the-art tools have been developed. Because the signal peptide and the transmembrane protein are too close to each other, problems have arisen. In addition, several proteins can be found in multiple locations. Therefore, we proposed to use the integrated approach to bootstrap the performance of the traditional prediction method. The combination of SignalP and DeepLoc can provide a better result than any single predictor alone. This study was conducted using the protein sequences from the recently reviewed database and applied the scoring indices derived from the confusion matrix. In terms of recall and F1 score, the results show that the integrated method outperforms the individual predictors. Some indices are slightly different from those of the single predictor. Moreover, the integrated method increases the detection rates while decreasing the false discovery rates. It can be shown that the combination of multiple predictor algorithms outperforms the conventional predictor method.

Graphical abstract:



Keywords: SignalP; DeepLoc; signal peptide; extracellular

1. Introduction

Every protein in the cell must be directed to a specific cellular compartment to perform its functions. This includes secretory proteins that are located outside the cell. A signal peptide is the most important element in a protein sequence to indicate its pathway. A preprotein is the nascent protein from a translation that contains the signal peptide. In secretory proteins, the enzyme signal peptidase in the lumen of the endoplasmic reticulum (ER) cleaves the signal peptide from the nascent preprotein as it travels to its destination. The ER divides its membrane to form a transport vesicle toward the Golgi network for packaging and delivery outside the cell once the protein is properly folded [1].

In eukaryotes, the signal peptide is flanked at the N terminus of a preprotein sequence. It consists of 1-5 positively charged amino acids followed by 7-15 hydrophobic acids and 3-7 polar amino acids. Notably, the proteolytic cleavage site is located at the end of the polar region, between positions -3 and -1 [2]. Problems have arisen because the signal peptide is very similar to the integral membrane protein that spans the lipid bilayer of the cell and organelle and is called transmembrane protein. The hydrophobic portion of the transmembrane protein sequence is entangled with the more abundant positively charged region. The transmembrane protein differs in that it lacks a proteolytic cleavage site [3].

Once the properties of secretory proteins were introduced, many tools were developed to distinguish the signal peptide from other types of proteins. Secretory proteins can be predicted in 2 ways, by the presence of a signal peptide or by extracellular localization. SignalP, the latest version 5.0, combines a deep recurrent neural network with an optimized transfer learning algorithm. The use of two deep learning models led to improved results compared to the previous computational model [4]. DeepLoc 1.0 is a predictor of subcellular localization that uses a deep recurrent neural network. DeepLoc 1.0 also uses a feed-forward neural network with an attention mechanism to decode features and classify them into 11 different localization types, including extracellular. DeepLoc 1.0 outperforms previous homology-based predictors by using a neural network as a computational model. [5]. A single predictor, on the other hand, may incorrectly predict an outcome. Moreover, multiple proteins can be found in multiple locations. Proteins secreted to outside the cell may be incorporated inside the cell and enter the cytosol and nucleus. For this reason, we recommend using the integrated method to bootstrap predictive performance. The combination of SignalP 5.0 and DeepLoc 1.0 could be more effective than either predictor alone. This study was performed using protein sequences from a recently reviewed database and scoring indices derived from binary classification.

2. Materials and Methods

2.1 Evaluation indices

The study is evaluated using the confusion matrix, a 2x2 contingency table of actual and predicted classes consisting of true positive (TP), false positive (FP), true negative (TN), and false negative (FN). TP is a sample that was correctly predicted and labeled positive. FP is a negative sample that was predicted to be positive. TN is a negative sample that was correctly predicted and labeled as negative. FN is a positive sample that was predicted to be negative. The calculation derived from the confusion matrix used in this study is precision, recall, false positive rate of transmembrane for prediction as signal peptide (FPRTM), F1 score and Matthews Correlation Coefficient (MCC) [6,7]. Precision measures how many selected elements are relevant and is calculated by the proportion of a sample set that is correctly predicted as signal peptide (TP) and the sum of positive samples including TP and FP as shown in equation (1).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall, or sensitivity or true-positive rate (TPR), refers to how many relevant elements are selected, or the accuracy of the positive class, and is calculated by the proportion of samples that are correctly predicted to be a signal peptide (TP) and the sum of the ground truth, which includes TP and FN, as shown in equation (2).

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

The properties of signal peptide and transmembrane proteins are very similar. The false positive rate of transmembrane proteins for prediction as signal peptide (FPR_{TM}) is calculated by equation (3), where the false positive rate of transmembrane proteins is denoted as FP_{TM}, and a total number of transmembrane proteins, which are denoted as N_{TM}.

$$FPR_{TM} = \frac{FP_{TM}}{N_{TM}} \tag{3}$$

F1-Score refers to the harmonic mean of Precision and Recall and is usually optimized to balance Precision and Recall. The F1-Score is calculated according to the following equation (4).

$$F1\ score = \frac{2 \times TP}{2 \times (TP + FN + FP)} \quad (4)$$

MCC measures the quality of binary classification originated by Matthews (1975) [8]. MCC is calculated by equation (5), where the worst value leads to the best value from 0 to 1, respectively. [7,9].

$$MCC = \frac{TP \times (TN - FP) \times FN}{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)} \quad (5)$$

The measurement of MCC and F1 score may sometimes disagree, especially in a highly skewed dataset or in an outcome set that does not have a good ratio of positive and negative classes. Therefore, the measurement of F1-Score alone may indicate exaggerated results [9]. Moreover, the effectiveness of the computational methods for determining ground truth on different types of datasets is described by the detection rate in Equation (6). Moreover, the inverse of precision gives the false discovery rate (FDR).

$$Detection\ Rate = \frac{TP}{TP + FP + FN + TN} \quad (5)$$

All evaluation indices were calculated using R in R studio software. R scripts for data analysis were deposited in GitHub repository https://github.com/JirathNuan/BMB2021.

2.2 Experimental setup

All experiments in this study were conducted on a server with 64-bit architecture, 47 GB RAM and 1 TB SSD, running Ubuntu 19.10 (Eoan Ermine) LTS and Python version 3.8. Miniconda was used as a server-side version package check and prevents unbiased results. Miniconda can be downloaded and installed using the instructions in the Conda documentation (https://docs.conda.io/en/latest/miniconda.html).

2.3 Dataset

The protein sequences used to evaluate the tools were extracted from the Uniprot/SwissProt database, release 2021_02 [10]. The protein sequence set consisted of two subsets, a positive subset and a negative subset. Both subsets were searched based on the names of the subcellular localizations. The positive subset was searched for the keyword "Secreted" and the negative subset was searched for the keywords "Cell envelope" and "Mitochondrion matrix" to find protein sequences targeting the cell membrane and mitochondrion, respectively. For each subset, only protein sequences from humans were selected. The protein sequences must be longer than 100 aa. The protein sequences with keywords "uncharacterized", "probable" or "similar" were excluded. Also, a keyword "frag-ment" indicating the mature peptides is allowed, but partial, N- or C-terminal "fragment" are not allowed. A dataset that does not contain the evidence code was also excluded. Further details of the records to be excluded in each subgroup were described in Table 1.

Table 1. Description of dataset processing.

Dataset (and total sequences)	**Subcellular location term**	**Records to exclude**
Positive subset (805)	Secreted	Records containing keywords "Membrane", "Endoplasmic reticulum" and "Nucleus"
Negative subset (1,327)	Cell envelope	Records containing more than one different location such as "Secreted", "Cytoplasm", "Endoplasmic reticulum", "Nucleus", "Peroxisome", "Virion", "Mitochondrion", "lysosome", and "Golgi"
	Mitochondrion matrix	Records that are located at more than one different location, but the exception is a record that locates in the same compartment, e.g., mitochondrion matrix along with mitochondrion.

2.4 Prediction methods

Method SignalP (SignalP_alone) is designated from the use of SignalP 5.0 software and is the latest update version of SignalP. The standalone version of SignalP 5.0 was downloaded from https://services.healthtech.dtu.dk/software.php for research under license from Technical University of Denmark (DTU).

Method DeepLoc (DeepLoc_alone) is designated from the use of DeepLoc 1.0 software. The standalone DeepLoc 1.0 is available to download at https://services.healthtech.dtu.dk/software.php for research under license from the Technical University of Denmark (DTU).

Integrated method (SignalP_DeepLoc) is designated from combining the use of SignalP 5.0 and DeepLoc 1.0 software. First, input protein sequences were identified the presence of a signal peptide by SignalP 5.0. The signal peptide-flanking proteins were then used as input to DeepLoc 1.0 to identify its subcellular localization. The final results were the proteins that containing the signal peptide and secrete to the extracellular region.

3. Results

3.1 Overall performance on human dataset

Figure 1 illustrates the overall performance comparison between three methods, including the integrated prediction result of SignalP and DeepLoc. The prediction using the integrated method provides high value for recall and F1. While DeepLoc alone shows the best performance in terms of MCC, but the precision is slightly lower than the integrated method. Prediction with SignalP alone gives the lowest performance in all indices.

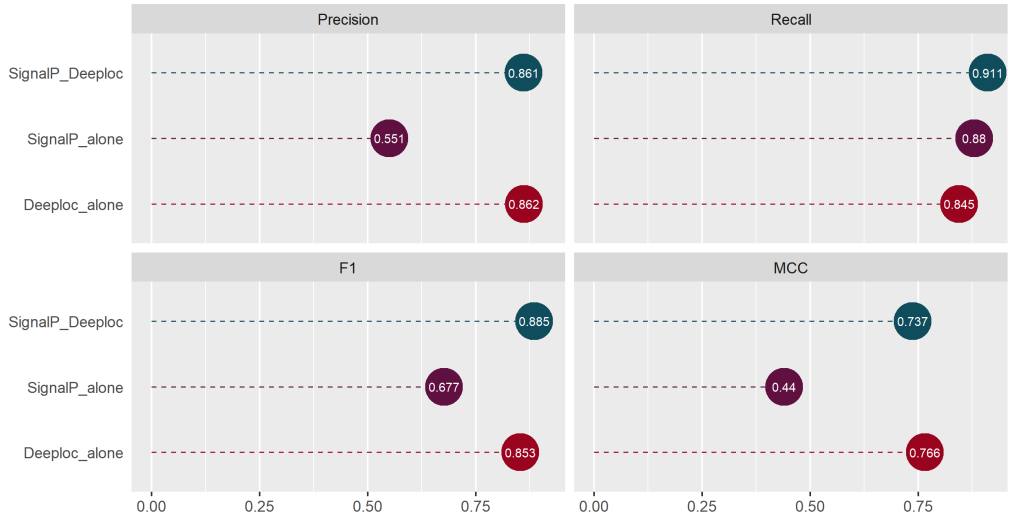


Figure 1. Evaluation of precision, recall, F1-score, and Matthews Correlation Coefficient (MCC).

3.2 Performance on discriminating between the signal peptide and transmembrane proteins

Prediction with DeepLoc alone is 2.25 times better than the integrated method in terms of FPR$_{TM}$ (Figure 2). In contrast, prediction with SignalP alone increases the FPR$_{TM}$ by 5.8-fold compared to the best method. This means that prediction with DeepLoc performs better than other methods in the human dataset.

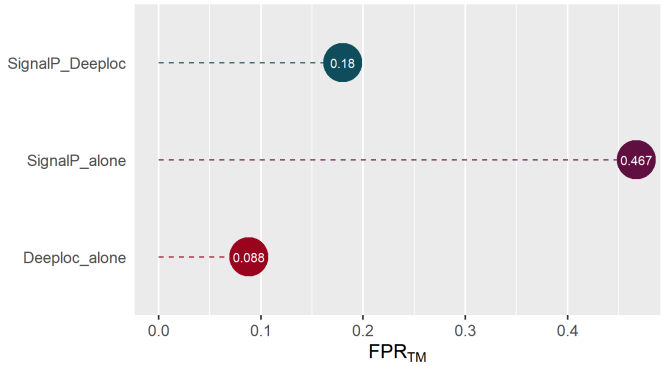


Figure 2. Evaluation of precision, recall, F1-score, and Matthews Correlation Coefficient (MCC).

3.3 Integrated method improves detection rate and false discovery rate

The detection rate of the integrated method is outstanding and is about 1.5 times higher than that predicted by SignalP or DeepLoc alone (Figure 3). This indicates that the integrated method outperforms the individual prediction methods. Moreover, the FDR of the integrated method and DeepLoc alone is slightly different, while the SignalP alone method has a higher FDR. This indicates that the combination of prediction method is not only effective in terms of detection rate, but also improves the FDR.

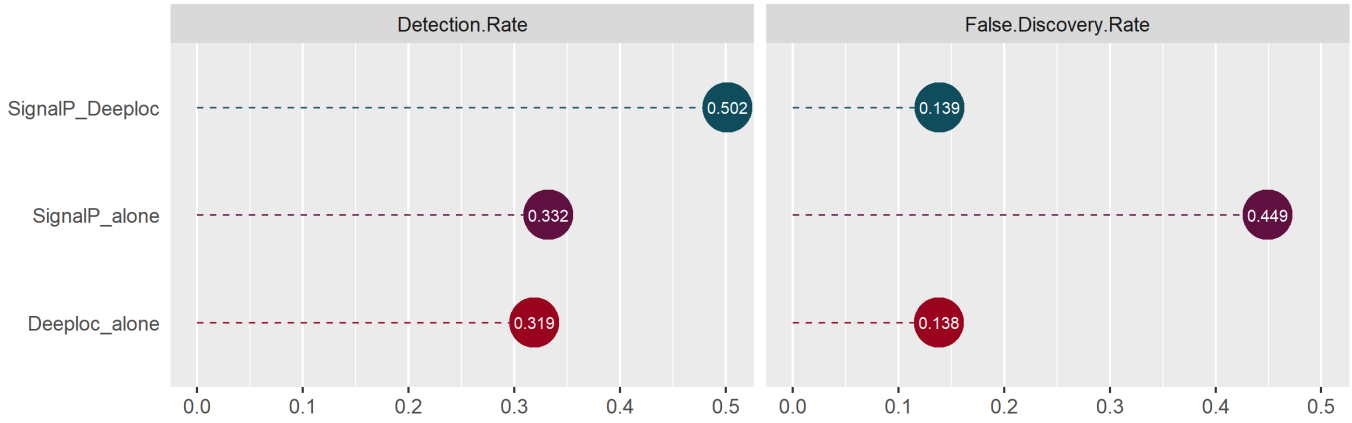


Figure 3. Detection rate and false discovery rate.

4. Discussion

Previously, secretory protein prediction was based on the results of a single prediction method. Here, we generated the dataset from the recently reviewed protein database based on subcellular location terms. Each location term was retrieved from the pro-tein sequences and the data were filtered to ensure that each protein was in a location that had been shown to exist. In particular, to ensure that the transmembrane proteins used for the analysis were not located in the organelle membrane, the cell envelope localization term was then selected. Proteins from the mitochondrial matrix encoded by mitochondrial DNA itself are also selected as the negative subset. In the positive subset, the secretory proteins are selected from the location term secreted. This means that these proteins are encoded by the genomic DNA and pass through the secretory pathway to reach outside the cells. Although, secretory proteins are diverse, some secretory proteins such as interleukins, neuropeptides and growth factor proteins often travel through the cell as they can also act as transcriptional regulators. Therefore, these may affect

subcellular localization as DeepLoc. However, some secreted proteins predicted by SignalP to contain a signal peptide are indifferent with respect to extracellular localization. This means that the predictor is confused, indicating either type I (FP) or type II (FN) errors in prediction. There is no single evaluation index that indicates the best prediction method. Therefore, the best prediction method should be selected based on its ability to reduce both types of errors. Moreover, it is better to consider the effectiveness of prediction and reduce the false discovery rate. However, to prove that the integrated method is worthy, it must be applied with other animal datasets.

5. Conclusions

In this study, a new strategy for identifying secretory proteins was proposed. Each predictor is performed with a different algorithm. Previously, a single predictor alone could be sufficient to identify secretory proteins. As the biological knowledge of protein subcellular lo-calization has been explored more deeply, we can thus learn more ground truth. The integration of the prediction algorithm is better as it improves the prediction of the positive classes. This affects many indices that indicate the effectiveness of prediction and reduces the false discovery rate.

Author Contributions: Conceptualization, J.N.; Methodology, J.N.; Software, J.N.; Validation, J.N.; Formal analysis, J.N.; Investigation, J.N. and U.S.; Resources, J.N. and U.S.; Data Curation, J.N. and U.S.; Writing - Original Draft, J.N.; Writing - Review & Editing, U.S.; Visualization, J.N.; Supervision, U.S.; Project administration, U.S.; Funding acquisition, U.S. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding

Data Availability Statement: Dataset, Commands, R script, and additional files are deposited in GitHub repository https://github.com/JirathNuan/BMB2021.

Acknowledgments: We would like to thank the Center for Genomics and Bioinformatics Research, Faculty of Science, Prince of Songkla University, Thailand for supporting the computer server in this research. Although the pandemic is still underway, we thank the BMB2021 committee for organizing this international conference.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dalbey, R.E.; von Heijne, G. 1 - Introduction/Overview. In Protein Targeting, Transport, and Translocation; Dalbey, R.E., von Heijne, G., Eds.; Academic Press: London, 2002; pp. 1–4 ISBN 978-0-12-200731-6.
2. von Heijne, G. The Signal Peptide. J. Membrain Biol. 1990, 115, 195–201, doi:10.1007/BF01868635.
3. von Heijne, G. Membrane Protein Structure Prediction: Hydrophobicity Analysis and the Positive-inside Rule. Journal of Molecular Biology 1992, 225, 487–494, doi:10.1016/0022-2836(92)90934-C.
4. Armenteros, J.J.A.; Tsirigos, K.D.; Sønderby, C.K.; Petersen, T.N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks. Nature Biotechnology 2019, 37, 420–423, doi:10.1038/s41587-019-0036-z.
5. Armenteros, J.J.A.; Sønderby, C.K.; Sønderby, S.K.; Nielsen, H.; Winther, O. DeepLoc: Prediction of Protein Subcellular Localization Using Deep Learning. Bioinformatics 2017, 33, 3387–3395, doi:10.1093/bioinformatics/btx431.
6. Savojardo, C.; Martelli, P.L.; Fariselli, P.; Casadio, R. DeepSig: Deep Learning Improves Signal Peptide Detection in Proteins. Bioinformatics 2018, 34, 1690–1696, doi:10.1093/bioinformatics/btx818.
7. Wu, J.-M.; Liu, Y.-C.; Chang, D.T.-H. SigUNet: Signal Peptide Recognition Based on Semantic Segmentation. BMC Bioinformatics 2019, 20, 677, doi:10.1186/s12859-019-3245-z.
8. Matthews, B.W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 1975, 405, 442–451, doi:10.1016/0005-2795(75)90109-9.
9. Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. BMC Genomics 2020, 21, doi:10.1186/s12864-019-6413-7.

10. The UniProt Consortium UniProt: The Universal Protein Knowledgebase in 2021. Nucleic Acids Research 2021, 49, D480–D489, doi:10.1093/nar/gkaa1100.