Prompt-Based Learning on Large Protein Language Models Improves Signal Peptide Prediction

Shuai Zeng, Duolin Wang, Lei Jiang, and Dong Xu(✉)

Department of Electrical Engineer and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
zengs@umsystem.edu, {wangdu,leijiang,xudong}@missouri.edu

Abstract. Signal peptides (SP) play a crucial role in protein localization in cells. The development of large protein language models (PLMs) provides a new opportunity for SP prediction. We applied a prompt-based learning framework, Parameter-Efficient Fine-Tuning (PEFT) for SP prediction, PEFT-SP, to effectively utilize pre-trained PLMs. We integrated low-rank adaptation (LoRA) into ESM-2 models to better leverage the protein sequence evolutionary knowledge of PLMs. Experiments show that PEFT-SP using LoRA enhances state-of-the-art results, leading to a maximum MCC gain of 0.372 for SPs with small training samples and an overall MCC gain of 0.048. Furthermore, we also employed two other prompt-based learning methods, i.e., Prompt Tuning and Adapter Tuning, into ESM-2 for SP prediction. More elaborate experiments show that PEFT-SP using Adapter Tuning can also improve the state-of-the-art results with up to 0.202 MCC gain for SPs with small training samples and an overall MCC gain of 0.030. LoRA requires fewer computing resources and less memory than the Adapter during the training stage, making it possible to adapt larger and more powerful protein models for SP prediction. The PEFT-SP framework is available at https://github.com/shuaizengMU/PEFT-SP. The web server for SP predic-tion leveraging the PEFT-SP framework is publicly available at https://www.mu-loc.org/peftsp/.

Keywords: Signal peptide · Large protein language model · ESM · Prompt-based learning · Parameter-Efficient Fine-Tuning · Low-rank adaptation

1 Introduction

Signal Peptides (SPs), short amino acid sequences typically located in the N-terminals of nascent polypeptides, play a crucial role in directing proteins through various translocation pathways. These pathways, such as the secretory (Sec) and the twin-arginine translocation (Tat) pathways, differ in their handling of protein conformation during translocation, with the Sec pathway transporting unfolded proteins and the Tat pathway translocating fully folded proteins. Upon

J. Ma (Ed.): RECOMB 2024, LNCS 14758, pp. 400–405, 2024.
https://doi.org/10.1007/978-1-0716-3989-4_40

successful translocation across the membrane, signal peptidase (SPase) precisely cleaves the SP, releasing the mature protein. SPases are categorized into three groups (SPase I, II, III), each dedicated to specific types of signal peptides. Precisely, SPase I (Sec/SPI) is responsible for cleaving general secretory signal peptides, while SPase II (Sec/SPII) and SPase III (Sec/SPIII) specialize in the cleavage of lipoprotein and prepilin signal peptides, respectively. Tat substrates are exclusively processed by SPase I (Tat/SPI) or SPase II (Tat/SPII).

Although these SP regions are recognizable, the absence of clearly defined consensus motifs presents a significant challenge to SP prediction. Advances in machine learning and deep learning have led to the development of various SP prediction tools, such as SignalP versions, SPEPlip, Deep-Sig, and SignalP 6.0. Large protein language models (PLMs), such as ProTrans and ESM-1 [4], have become foundational tools for various biological modeling tasks related to proteins. However, there is room for improvement. This paper presents a novel SP prediction framework, PEFT-SP, designed to harness the capabilities of PLM for signal peptide and cleavage site prediction. PEFT-SP consists of the ESM-2 model, a linear Conditional Random Fields (CRF) model, and PEFT modules, including Adapter Tuning [1], Prompt Tuning [3], and Low-Rank adaptation (LoRA) [2]. Our end-to-end solution performs better than SignalP 6.0, especially in SP types with limited training data. We evaluate different PEFT methods, including LoRA, and highlight the efficiency of our framework in utilizing PLMs for SP prediction. This study contributes to the exploration of PEFT on PLMs for SP prediction, emphasizing the importance of efficient PLM utilization in advancing prediction performance.

2 Methods

2.1 Pre-trained Large Protein Language Models

The recent surge in Protein Language Models (PLMs) has brought notable examples like ProtTrans, ESM-1, and the ESM-2 family. Among these models, the ESM-2 model family stands out, offering varying model sizes ranging from 8 million parameters to a substantial 15 billion parameters. The ESM-2 model family, encompassing ESM2-150M, ESM2-650M, and ESM2-3B, has showcased outstanding performance in structure prediction, surpassing many counterparts from ProtTrans and the ESM-1 model family in protein sequence-related tasks.

Unlike existing signal peptide prediction models that necessitate appending an organism identifier to the protein sequence, PEFT-SP with ESM-2 backbone streamlines the process by taking only the protein sequence as input. It encodes the sequence into token embeddings, which are then fed into a stack of multiple Transformer layers designed to capture contextual relationships between amino acids. These layers incorporate a self-attention mechanism and Position-wise Feed-Forward Networks (FFN) surrounded by separate residual connections.

The linear chain Conditional Random Field (CRF) is commonly employed in sequence labeling tasks, capturing relationships between labels and observed data. Viterbi decoding computes the most probable state sequence, including SP

regions (n, h, c, twin-arginine). Cleavage site (CS) prediction identifies CS based on the last SP class state. The forward-backward algorithm calculates marginal probabilities per sequence position. Predicting signal peptide type sums marginal probabilities of states and divides by sequence length.

2.2 Parameter-Efficient Fine-Tuning Methods for ESM-2

PEFT methods for ESM-2 can enhance the model's performance in various downstream tasks. PEFT introduces tunable parameters while freezing the original parameters in the backbone model, enabling the model to be tailored to new tasks with reduced computational overhead and fewer labeled examples. Unlike the original configuration of Adapter Tuning and LoRA, which integrates related modules into all Transformer layers, they are specifically inserted into the bottommost Transformer layers within the ESM-2 model, inspired from LLaMA-Adapter. Adapter Tuning involves incorporating adapter modules with a bottleneck architecture within the Transformer layer of the ESM-2 model. These modules compress the input data into a bottleneck layer with reduced dimensionality and reconstruct it to match the original input size. Prompt Tuning adds trainable embeddings, referred to as soft prompts, into the sequence embeddings, serving as inputs to the ESM-2 model. Soft prompts are continuously updated using gradients, while all parameters within the ESM-2 model remain fixed throughout the training process. LoRA enhances the fine-tuning of ESM-2 by introducing trainable rank decomposition matrices into the Transformer architecture. This reparameterization is applied to the projection matrices of the Query, Key, Value, and FFN modules within the Transformer.

2.3 Model Evaluation and Experiment Setting

We utilized the Matthews correlation coefficient (MCC), standard in SP prediction methods, for a fair assessment. Since most methods involve binary SP classification, we computed MCC1 using samples of transmembrane and soluble proteins. Additionally, MCC2 was calculated using a dataset where a specific SP type was the positive sample, and all other SPs and non-SPs as negatives. Our CS prediction depends on the last SP class region, outputting the cleavage site position rather than probabilities. Precision and recall evaluate CS prediction within a 3-position window. Precision is the correct CS ratio to predicted CSs, while recall is the correct CS ratio to true CSs. Accurate CS predictions must align with SP labels.

Our SP dataset is sourced from SignalP 6.0, comprising diverse protein sequences: 3,352 Sec/SPI, 2,261 Sec/SPII, 113 Sec/SPIII, 595 Tat/SPI, 36 Tat/SPII, 16,421 intracellular, and 2,615 transmembrane sequences. Sec/SPIII and Tat/SPII have limited samples. Each sequence is labeled with SP type and region details, with the final label indicating the CS. Initially obtained from Archaea, Eukarya, Gram-positive, and Gram-negative bacteria, the dataset is partitioned into three subsets for fairness and robustness. We used a nested three-fold cross-validation, resulting in six distinct test sets.

3 Results

3.1 Comparisons with State-of-the-Art

We employed PEFT-SP using LoRA for each model from the ESM-2 model family and trained them independently. We evaluated the MCC1 and MCC2 scores for each SP type within each organism group across test sets. Additionally, we calculated the mean MCC scores for MCC1 and MCC2 across all SP types and organisms.

PEFT-SP using LoRA with ESM2-3B backbone achieves the best performance (as shown in Fig. 1). It consistently outperforms SignalP 6.0 in the SP types (Sec/SPIII and Tat/SPII) with limited training samples, except for Tat/SPII in Gram-positive bacteria. It achieves a maximum MCC1 gain of 79.8% and an MCC2 gain of 87.3% in Sec/SPIII for Archaea. It attains a mean MCC1 improvement of 5.6% and a mean MCC2 improvement of 6.1%. It performs slightly worse than SignalP 6.0, with MCC1 differences ranging between 0.3%

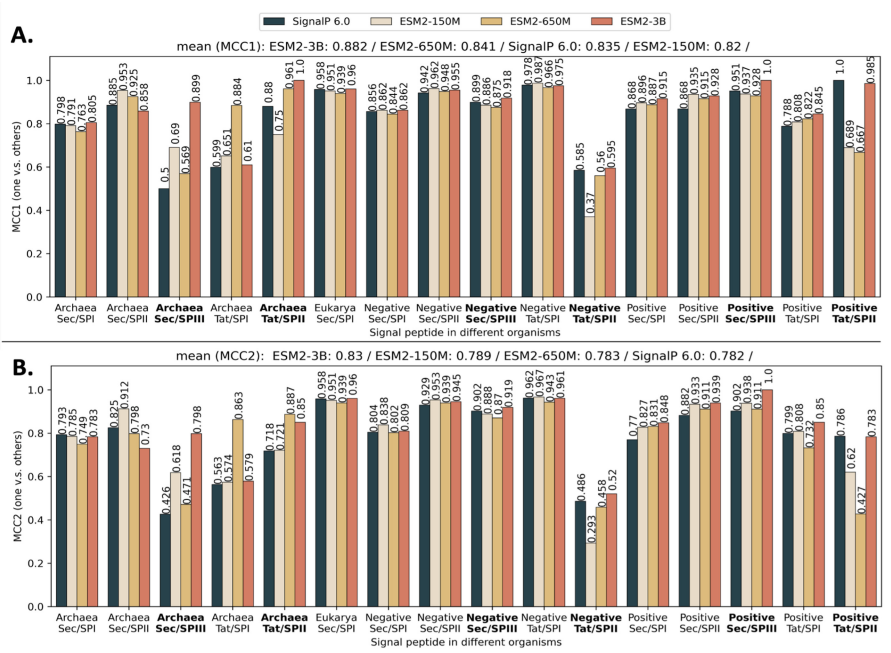

Fig. 1. PEFT-SP using LoRA and SignalP 6.0 performance in terms of MCC score for each SP type across different organisms. The bold text in the x-axis represents the SP type with small training samples. The MCC1 and MCC2 scores are shown along with the bars. The sorted mean for MCC1 and MCC2 are listed at the top. (A) MCC1 scores performance on the negative class composed of soluble and transmembrane proteins. (B) MCC2 scores performance on the negative class comprising soluble and transmembrane proteins and other SP types.

and 3.0% and MCC2 differences ranging between 0.4% and 11.5% in Sec/SPI and Sec/SPII for Archaea, and Tat/SPII for both Gram-negative and Gram-positive bacteria. For SP types (Sec/SPI, Sec/SPII, and Tat/SPII) with sufficient training data, PEFT-SP using LoRA with ESM2-3B demonstrates superior or closely comparable performance to SignalP 6.0.

3.2 Comparisons with Fine-Tuning and Other PEFT Methods

We compared PEFT-SP using different PEFT methods with ESM-3B, as well as SignalP 6.0 and finetuned ESM2-3B model. We trained all models independently with the same datasets generated from nest cross-validation. The performance of each model was measured using MCC2 across cross-validation.

Table 1 shows that the fine-tuning approach outperforms SignalP 6. This suggests that the ESM2-3B model holds promise as a potential candidate for other PEFT methods. The PEFT-SP using LoRA performs better than PEFT-SP using Prompt Tuning and Adapter Tuning regarding the mean MCC2. Moreover, the PEFT-SP using LoRA has fewer trainable parameters than fine-tuning and other PEFT methods during the training stage, dramatically reducing the computing resource and memory storage.

Table 1. Benchmark results of MCC2 for SignalP 6.0, Fine-tuning ESM2-3B, and PEFT-SP using different PEFT methods with ESM2-3B backbone. The SP type indicated with the symbol † represents SP types with limited training samples. The bold value indicates the highest value for each SP type among all methods.

Method/Backbone	-	Fine-tuning	Prompt Tuning	Adapter Tuning	LoRA
SP types	SignalP 6.0	ESM2-3B	ESM2-3B	ESM2-3B	ESM2-3B
Archaea Sec/SPI	0.793	0.771	0.777	**0.825**	0.783
Archaea Sec/SPII	0.825	**0.864**	0.509	0.783	0.730
Archaea Sec/SPIII †	0.426	0.724	0.500	0.351	**0.798**
Archaea Tat/SPI	0.563	0.564	**0.653**	0.538	0.579
Archaea Tat/SPII †	0.718	0.792	0.182	0.660	**0.850**
Eukarya Sec/SPI	0.958	0.948	0.954	0.954	**0.960**
Negative Sec/SPI	0.804	**0.813**	0.723	0.820	0.809
Negative Sec/SPII	0.929	0.946	0.886	0.950	**0.945**
Negative Sec/SPIII †	0.902	**0.982**	0.970	0.899	0.919
Negative Tat/SPI	**0.962**	0.902	0.853	0.899	0.961
Negative Tat/SPII †	0.486	0.358	0.325	0.405	**0.520**
Positive Sec/SPI	0.770	0.810	0.746	0.814	**0.848**
Positive Sec/SPII	0.882	0.908	0.833	0.911	**0.939**
Positive Sec/SPIII †	0.902	**1.000**	0.951	0.969	**1.000**
Positive Tat/SPI	0.799	0.746	0.590	0.752	**0.850**
Positive Tat/SPII †	**0.786**	0.603	0.148	0.669	0.783
Mean (MCC2)	0.781	0.796	0.663	0.762	**0.830**

4 Results

Our study introduced PEFT-SP, a new signal peptide prediction framework that operates without organism identifiers. Using LoRA with ESM2-3B, PEFT-SP effectively handles SP types with limited training data and matches or surpasses the baseline performance of the model across all SP types. The success of PEFT-SP with LoRA stems from two key factors: (1) leveraging the evolutionary insights of the ESM2-3B backbone model, and (2) implementing LoRA, a lightweight fine-tuning method, to adapt PLMs for SP prediction while maintaining their high quality. To our best knowledge, this is the first study to explore the effectiveness of PLM using the PEFT approach for SP prediction tasks.

Acknowledgments and Funding. We wish to thank Fei He and Yuexu Jiang for their useful discussions. This work was funded by the National Institutes of Health [R35-GM126985] and the National Science Foundation [DBI-2145226]. Funding for open access charge: National Science Foundation.

Availability of Data and Materials. The source code of PEFT-SP and trained models are publicly available at https://github.com/shuaizengMU/PEFT-SP. The web server is available at https://www.mu-loc.org/peftsp/.

Preprint. A detailed description of our methodology and results can be found in the preprint at https://www.biorxiv.org/content/10.1101/2023.11.04.565642v1.

References

1. Houlsby, N., et al.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning, pp. 2790–2799. PMLR (2019)
2. Hu, E.J., et al.: Lora: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
3. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)
4. Rives, A., et al.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. **118**(15), e2016239118 (2021)