Enhancing Transcription Factor Prediction through Multi-Task Learning (Student Abstract)

Liyuan Gao[1], Matthew Zhang [2], Victor S. Sheng[1]

[1] Texas Tech University, 2500 Broadway, Lubbock, 79409, Texas, USA
[2] Baylor University, 1311 S 5th St, Waco, 76706, Texas, USA.
{liygao, victor.sheng}@ttu.edu, matthew_zhang1@baylor.edu

Abstract

Transcription factors (TFs) play a fundamental role in gene regulation by selectively binding to specific DNA sequences. Understanding the nature and behavior of these TFs is essential for insights into gene regulation dynamics. In this study, we introduce a robust multi-task learning framework specifically tailored to harness both TF-specific annotations and TF-related domain annotations, thereby enhancing the accuracy of TF predictions. Notably, we incorporate cutting-edge language models that have recently garnered attention for their outstanding performance across various fields, particularly in biological computations like protein sequence modeling. Comparative experimental analysis with existing models, DeepTFactor and TFpredict, reveals that our multi-task learning framework achieves an accuracy exceeding 92% across four evaluation metrics on the TF prediction task, surpassing both competitors. Our work marks a significant leap in the domain of TF prediction, enriching our comprehension of gene regulatory mechanisms and paving the way for the discovery of novel regulatory motifs.

Introduction

Transcription factors (TFs) are critical proteins that bind to specific DNA sequences. By attaching to particular DNA regions, TFs regulate the process of transcribing DNA into RNA. However, predicting and understanding the nuanced behavior of TFs remains a formidable challenge in molecular biology. As biological datasets burgeon in size and complexity, there's a pressing need for advanced computational models that can adeptly harness this data to provide more accurate insights into the world of TFs. Traditional TF prediction model like TFPredict utilizes Support Vector Machines (SVMs) paired with other simple machine learning algorithms, which is not well-suited for large datasets and offer constrained prediction accuracy (Eichner et al. 2013). Though tools such as DeepTFactor, a deep learning model designed for predicting TFs across all domains of life, have already been developed, these tools require an extensive dataset to function optimally (Kim et al. 2021). In this paper, we propose a multi-task learning framework with a pre-trained protein language model to improve the TF prediction task. By training models on multiple related tasks simultaneously, it is believed that these models can harness shared information between tasks and enhance performance. Integrating this approach with the rich world of TF-specific annotations and associated domain annotations may hold the key to significantly improved TF predictions. Furthermore, diverging from conventional methods like one-hot encoding or K-mers encoding for protein sequences, this research adopts the pre-trained protein language model ESM, to procure protein sequence representations. The experimental results indicate that our multi-task learning framework achieves an accuracy exceeding 92% across four distinct evaluation metrics.

Methodology

ESM-2 is a state-of-the-art protein language model that outperforms all tested single-sequence protein language models across a range of structure prediction tasks and enables atomic resolution structure prediction (Lin et al. 2023). In this paper, we first utilize a pre-trained ESM-2 model to generate highly representative encodings of protein sequences that are subsequently fed to our multi-task prediction model. In addition, we employed a sliding window technique to condense protein representations and enhance computational efficiency. Specifically, we divide the protein sequence into windows of fixed size and then average the representation of amino acids within each window to obtain a condensed representation of these amino acids. Upon obtaining the condensed representations, these are subsequently fed into the subsequent multi-task learning framework. The comprehensive procedure is depicted in Fig.1.

Multi-task learning framework jointly trains several related tasks to improve their generalization performance by leveraging shared knowledge among them (Standley et al. 2020; Gao, Zhan, and Sheng 2023). Supposing there are T tasks, multi-task learning frameworks aim to solve these tasks simultaneously. These frameworks typically contain two sets of parameters: shared parameters θ and task-specific parameters $\{\psi_t\}_{t=1}^{T}$. The basic multi-task architectures aim to extract some common features in shared lower layers. By sharing information between related tasks, multi-task learning can generalize the model more effectively for the tasks. Following the shared layers, the remaining layers are split into multiple specific tasks. The optimization objec-

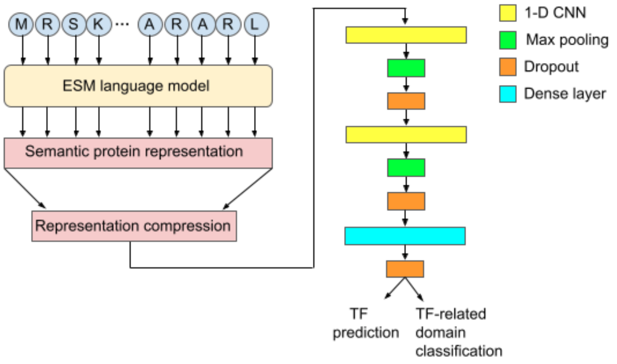
M R S K ··· A R A R L
ESM language model
Semantic protein representation
Representation compression
1-D CNN
Max pooling
Dropout
Dense layer
TF prediction
TF-related domain classification

Figure 1: Multi-task learning framework

Domains	ZF	Home	LZ	Other-TFs	Non-TFs	Total
TF	1710	1078	523	3058	0	6369
NTF	0	0	0	0	5000	5000

Table 1: Dataset distribution

tive of multi-task learning is as follows:

$$\text{Loss} = \sum_{t=1}^{T} \lambda_t(\theta, \{\psi_t\}_{t=1}^{T}) \tag{1}$$

where $\{\psi_t\}_{t=1}^{T}$ represent the task-specific loss weights, constraints $\lambda_t \geqslant 0$.

In this paper, the TF prediction model is structured to perform multi-task learning by concurrently training on TF-specific annotations and TF-related domain annotations. The shared layers consist of two 1-D CNN layers ((featuring 128 and 256 filters)), each followed by a max pooling and dropout layer (0.5). These 1-D CNN layers are then followed by a Dense layer and a dropout layer to further refine the extracted features and prevent overfitting. Finally, the output of TF prediction is generated by sigmoid activation and the output of TF-related domain classification is generated by softmax activation.

Dataset and Experimental Setup

The datasets were sourced from ”www.uniprot.org”, encompassing both TF-specific annotations and TF-related domain annotations. Within the scope of the TF prediction task, there are two categorizations: ’TF’ and ’Non-TF’ (NTF). Pertaining to the domain classification related to TFs, five distinct labels are identified: Zinc Finger domains (ZF), Homeodomain (Home), Leucine Zipper domains (LZ), other TF-associated domains (Other-TFs), and domains not related to TFs (Non-TFs). A detailed distribution of the dataset is presented in Table 1.

The deep learning models were performed for a total of 40 epochs with a batch size of 64. A learning rate of 1×10^{-4} was applied. Input protein sequences were standardized to 3000 amino acid residues, and the resulting sequences were further transformed into a predefined representation size of 300. Performance metrics were derived from five-fold cross-validation, averaging outcomes across all folds.

Model	F1 score	Accuracy
Single-task learning	0.8967	0.8842
Multi-task learning	**0.9158**	**0.9085**

Table 2: TF-related domain classification

Model	F1 score	Specificity	Sensitivity	Accuracy
TFpredict	0.7858	0.7754	0.7965	0.7860
DeepTFactor	0.9171	0.9056	0.9283	0.9169
Single-task	0.9225	0.8717	0.9466	0.9252
Multi-task	**0.9422**	**0.9253**	**0.9573**	**0.9432**

Table 3: TF prediction

Experimental Results and Analysis

The experimental results are delineated in 2 and 3, offering a comprehensive assessment of the proposed models' performance in TF related tasks. Table 2 distinctly demonstrates the superiority of the multi-task learning framework over its single-task counterpart in the domain of TF-related classification. In the context of TF prediction, as exhibited in Table 3, a detailed comparison amongst various models highlights the exceptional performance of the multi-task learning approach. This model distinctly surpasses all others in each evaluated metric, achieving remarkable F1 scores, specificity, sensitivity, and accuracy rates of 0.9422, 0.9253, 0.9573, and 0.9432, respectively. These results not only underscore the efficacy of the multi-task learning model in isolation but also suggest a synergistic enhancement when applied to the tasks of TF-related domain classification and TF prediction concurrently.

The observed data indicate a significant interplay between the TF-related domain classification and TF prediction tasks, where the shared features and insights gleaned from one task evidently contribute to the improved performance in the other. This interdependency underscores the potential of multi-task learning in extracting deeper, more nuanced patterns and relationships within the biological data, which might remain obscured under single-task learning paradigms.

Conclusion

In this paper, we proposed a multi-task learning framework for the TF prediction task. The experimental results demonstrated that the proposed multi-task learning framework consistently outperformed both the single-task learning model and the existing TF prediction model across diverse evaluation metrics. By harnessing the synergies between interrelated tasks, the multi-task learning model not only achieved superior accuracy but also advanced our understanding of gene regulatory mechanisms.

References

Eichner, J.; Topf, F.; Dräger, A.; Wrzodek, C.; Wanke, D.; and Zell, A. 2013. TFpredict and SABINE: sequence-based prediction of structural and functional characteristics of transcription factors. *PLoS One*, 8(12): e82238.

Gao, L.; Zhan, H.; and Sheng, V. S. 2023. Mitigate Gender Bias using Negative Multi-Task Learning. *Neural Processing Letters*, 1–16.

Kim, G. B.; Gao, Y.; Palsson, B. O.; and Lee, S. Y. 2021. DeepTFactor: A deep learning-based tool for the prediction of transcription factors. *Proceedings of the National Academy of Sciences*, 118(2): e2021171118.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130.

Standley, T.; Zamir, A.; Chen, D.; Guibas, L.; Malik, J.; and Savarese, S. 2020. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, 9120–9132. PMLR.