Multi-Dataset Integration and Residual Connections Improve Proteome Prediction from Transcriptomics Using Deep Learning

This manuscript (permalink) was automatically generated from xomicsdatascience/transcriptome-proteome-nas-manubot@5d989ce on July 8, 2024.

Authors

- **Caleb W. Cranney**
  0000-0001-8482-758X · CCranney · CalebCranney
  Department of Computational Biomedicine, Cedars Sinai Medical Center · Funded by Grant R35GM142502, National Institute of General Medical Sciences (NIGMS)

- **Jesse G. Meyer**
  0000-0003-2753-3926 · jessegmeyerlab · j_my_sci
  Department of Computational Biomedicine, Cedars Sinai Medical Center

✉ — Correspondence possible via GitHub Issues

Abstract

Proteomes are well known to poorly correlate with transcriptomes measured from the same sample. While connected, the complex processes that impact the relationships between transcript and protein quantities remains an open research topic. Many studies have attempted to predict proteomes from transcriptomes with limited success. Here we use publicly available data from the Clinical Proteomics Tumor Analysis Consortium to show that deep learning models designed by neural architecture search (NAS) achieve improved prediction accuracy of proteome quantities from transcriptomics. We find that this benefit is largely due to including a residual connection in the architecture that allows input information to be remembered near the end of the network. Finally, we explore which groups of transcripts are functionally important for protein prediction using model interpretation with SHAP.

Introduction

The central dogma of biology posits a linear flow of information from genetic encoding to mRNA transcripts to functional proteins. However, this seemingly straightforward relationship belies a more intricate reality, where the interactions between omic layers are multifaceted and complex. Elucidating these relationships is crucial for understanding biological systems in both healthy and diseased states. By disentangling these interactions, researchers can identify markers and patterns specific to complex diseases, ultimately enabling the development of targeted treatments.

Among the multi-omic data type relationships, the connection between transcripts and their corresponding proteins is particularly enigmatic. Despite being quantifiable and having a direct derivative relationship, their relative quantities often exhibit only weak correlations [1], [2]. Research has identified several possible causes for this discrepancy, including alternate rates of protein generation and decay [3], [4], [5], varying reactions to environmental stimuli [6], [7], or simply systematic experimentation error bias [8], [9]. Recent works have even found that the most predictive transcripts of a protein include those that are involved in protein-protein interactions [10]. While predicting protein quantity through direct transcript-to-protein correlation remains elusive, using contextual transcript information may reveal more complex proteomic-transcriptomic relationships that could be leveraged to predict one from the other. Moreover, predicting proteins from transcripts could reduce the time and financial costs associated with future studies, as transcript quantification is generally easier to accomplish than protein quantification [11].

The National Cancer Institute's (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC) provides a valuable resource for exploring this issue, offering multi-omic datasets that enable research into healthy and cancerous disease states [12]. Specifically, CPTAC datasets comprise transcriptomic and proteomic data (among others) from various cancer types and adjacent healthy tissue, which makes this data useful for detection of inter-omic relationships. Several research collaborations have occurred to use this data to study RNA-protein relationships. The 2017 NCI-CPTAC DREAM Proteogenomics challenge, a sub challenge of which aimed to predict protein quantities from transcript abundance, is a notable example [13]. Contestants used a variety of methods, including random forest regression, genetic models, spline regression, linear regression, and elastic net methods, as well as ensemble combinations [14]. Using a test set only from breast cancer and ovarian cancer, the winning model achieved a pearson correlation of 0.41 and 0.47 between the true and predicted protein quantities, respectively [15]. Notably, deep learning neural networks were not strongly represented in the challenge results.

Since the close of the challenge, research into deep learning has expanded significantly, with a focus on replicating human behaviors like computer vision and natural language processing (NLP). However, the underlying principles of deep learning are equally applicable to biological data [16], [17], [18]. The key challenge lies in developing a model architecture that best fits a given problem. One underutilized strategy for deep learning model design is the NAS [19]. NAS serves a function like hyperparameter tuning, in that a range of values for specific hyperparameters are evaluated to obtain the optimal configurations. In the case of NAS, the concept is expanded to include model architecture in addition to hyperparameters, designing optimal and unique model architectures in an automated fashion. While NAS has been applied to genomic data [20], its application to predicting proteomic data from transcriptomic data remains unexplored.

We previously showed how machine learning can accurately predict the metabolome from the proteome, and how model interpretation revealed important biological insights [21]. Here, we extend that work to transcript-to-protein deep learning prediction models and demonstrate that utilizing NAS improved the accuracy. Furthermore, we highlight the potential of model interpretation to identify

patterns in transcript-protein relations that underpin biological processes characteristic of a disease state.

References