Genomic language model predicts protein co-regulation and function

Yunha Hwang
Dept of Organismic and Evolutionary Biology
Harvard University
MA, 02138
yhwang@g.harvard.edu

Andre Cornman
Independent contributor
ancornman1@gmail.com

Elizabeth Kellogg
Dept of Molecular Biology and Genetics
Cornell University
NY, 14853
ehk68@cornell.edu

Sergey Ovchinnikov
John Harvard Distinguished
Science Fellowship Program
Harvard University
MA, 02138
so@fas.harvard.edu

Peter Girguis
Dept of Organismic and Evolutionary Biology
Harvard University
MA, 02138
pgirguis@oeb.harvard.edu

Abstract

Deciphering the relationship between a gene and its genomic context is fundamental to understanding and engineering biological systems. Machine learning has shown promise in learning latent relationships underlying the sequence-structure-function paradigm from massive protein sequence datasets; However, to date, limited attempts have been made in extending this continuum to include higher order genomic context information. Here, we trained a genomic language model (gLM) on millions of metagenomic scaffolds to learn the latent functional and regulatory relationships between genes. gLM learns contextualized protein embeddings that capture the genomic context as well as the protein sequence itself, and appears to encode biologically meaningful and functionally relevant information (e.g. enzymatic function). Our analysis of the attention patterns demonstrates that gLM is learning co-regulated functional modules (i.e. operons). Our findings illustrate that gLM's unsupervised deep learning of the metagenomic corpus is an effective and promising approach to encode functional semantics and regulatory syntax of genes in their genomic contexts and uncover complex relationships between genes in a genomic region.

1 Introduction

Evolutionary processes result in the linkage between protein sequences, structure and function. The resulting sequence-structure-function paradigm has long provided the basis for interpreting vast amounts of genomic data. Recent advances in neural network (NN)-based protein structure prediction methods Jumper (2021); Baek (2021), and more recently protein language models (pLMs) Rives (2021); Elnaggar (2020); Madani (2023) suggest that data-centric approaches in unsupervised

NeurIPS 2021 AI for Science Workshop.

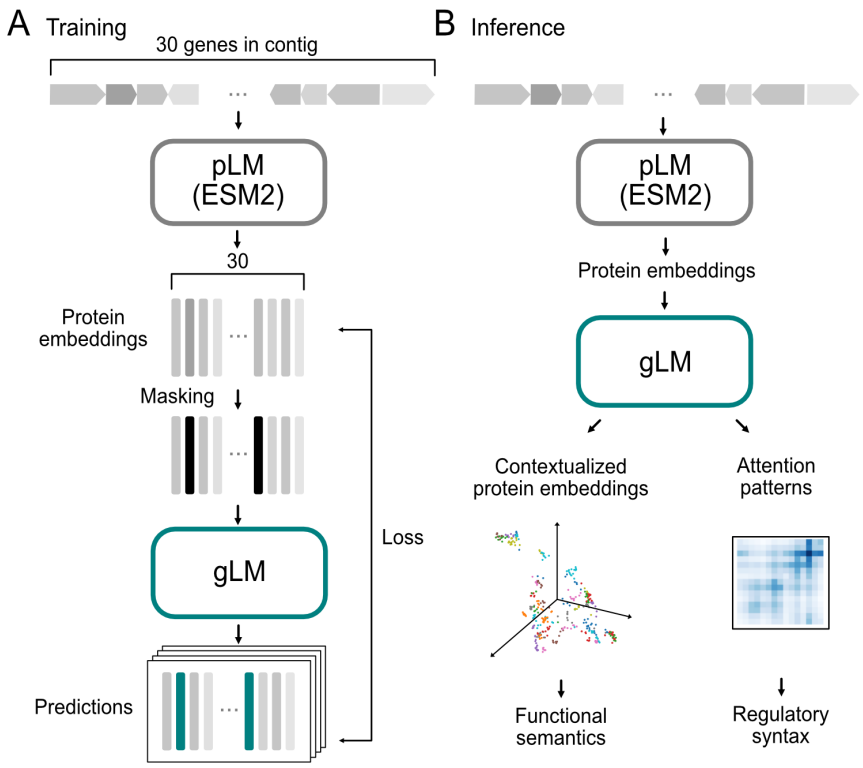

Figure 1: gLM training and inference schematics. A) For training, contigs (contiguous genomic sequences) containing up to 30 genes are first translated into proteins, which are subsequently embedded using a pLM encoder (ESM2). Masked inputs are generated by random masking at 15% probability and gLM (a transformer encoder) is trained to make four predictions for each masked protein, with associated likelihoods. Training loss is calculated on both the prediction and likelihoods. B) At inference time, inputs are generated from a contig using ESM2 output. Contextualized protein embeddings (last hidden layer of gLM) and attention patterns are used for various downstream tasks.

learning can represent these complex relationships shaped by evolution. To date, These models largely consider each protein as an independent and standalone entity. However, proteins are encoded in genomes, and the specific genomic context that a protein occurs in is also determined by evolutionary processes, where each gene gain, loss, duplication and transposition event is subject to selection and drift Wright (1948); Lynch & Conery (2003); Cordero & Polz (2014). These processes are particularly pronounced in prokaryotic genomes where frequent horizontal gene transfers (HGT) shape genomic organization and diversity Treangen & Rocha (2011); Shapiro (2012). Thus, there exists an inherent evolutionary linkage between genomic context and gene function Kountz & Balskus (2021), which can be explored by characterizing patterns that emerge from large metagenomic datasets.

Recent machine learning based approaches have shown predictive power of genomic context in gene function Miller et al. (2022) and metabolic trait evolution Konno & Iwasaki (2023) in prokaryotic genomes. However, both these models represent genes as categorical entities, despite genes existing in continuous space, where multidimensional properties such as phylogeny, structure, and function are abstracted in their sequences. In order to close the gap between genomic-context and gene sequence-structure-function, we developed the first, to our knowledge, genomic language model (gLM) that represents proteins using pLM embeddings that have been shown to encode relational properties Rives (2021) and structure information Lin (2023). Our model, based on the transformer architecture Vaswani et al. (2017), is trained using millions of unlabelled metagenomic sequences. We trained gLM with the masked language modeling Devlin et al. (2018) objective, with the hypothesis that its ability to attend to different parts of a multi-gene sequence will result in the learning of gene functional semantics and regulatory syntax (e.g. operons). Here, we report evidence of the learned contextualized protein embeddings and attention patterns capturing biologically relevant information.

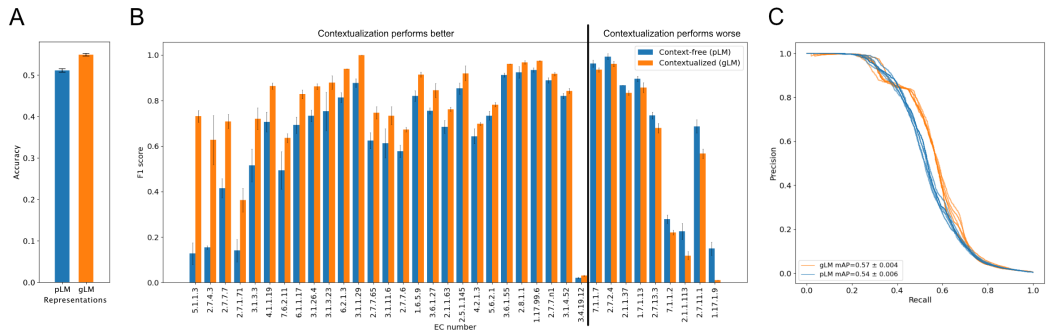

Figure 2: Contextualization of enzyme function. A) Linear probe EC classification accuracy for pLM (ESM2) representations and gLM (1st hidden layer) representations. B) F1-score comparisons of statistically significant (Benjamini/Hochberg corrected p-value < 0.05) differences in performance of pLM- and gLM-based EC number linear probes. EC classes are ordered with the largest gain with contextualization on the left to the largest loss with contextualization on the right. C) Precision-Recall curves of pLM- and gLM-based EC number linear probes.

2 Methods

2.1 Masked language modeling of genomic sequences

To model genomic sequences, we trained a 19-layer transformer model (Fig. 1A) on seven million metagenomic contig fragments consisting of 15 to 30 genes from the MGnify Richardson (2023) database. Each gene in a genomic sequence is represented by a 1280 feature vector (context-free protein embeddings) generated by using ESM2 pLM Rives (2021), concatenated with an orientation feature (forward or backward). For each sequence, 15% of genes are randomly masked, and the model learns to predict the masked label using the context. Based on the insight that more than one gene can legitimately be found in a particular genomic context, we allow the model to make four different predictions and also predict their associated probabilities. Thus, instead of predicting their mean value, the model can approximate the underlying distribution of multiple genes that can occupy a genomic niche We assess the model's performance using a pseudo-accuracy metric, where a prediction is considered correct if it is closest to the masked protein in euclidean distance compared to the other proteins encoded in the sequence. Training and inference code and analysis scripts are available at `https://github.com/y-hwang/gLM`.

3 Results

3.1 Model performance

We validate our model's performance on the Escherichia coli K-12 genome by excluding from training 5.1% of MGnify subcontigs in which more than half of the proteins are similar (>70% sequence identity) to E. coli K-12 proteins. The goal here is not to remove all E. coli K-12 homologs from the training, which would have removed a vast majority of training data as many essential genes are shared across organisms. Instead, our goal was to remove as many E.coli K-12-like genomic contexts (subcontigs) from training, which is more appropriate for the training objective. gLM achieves 71.9% in validation pseudo-accuracy and 59.2% in validation absolute accuracy. Notably, 53.0% of the predictions made during validation are with high confidence (with prediction likelihood > 0.75), and 75.8% of the high confidence predictions are correct, indicating gLM's ability to learn a confidence metric that corresponds to increased accuracy. We baseline our performance with a bidirectional LSTM model trained using the same language modeling task on the same training dataset, where validation performance plateaus at 28% pseudo-accuracy and 15% absolute accuracy.

3.2 Contextualization improves enzyme function prediction

To test the hypothesis that the genomic context of proteins can be used to aid function prediction, we evaluated how contextualization can improve the expressiveness of protein representations for

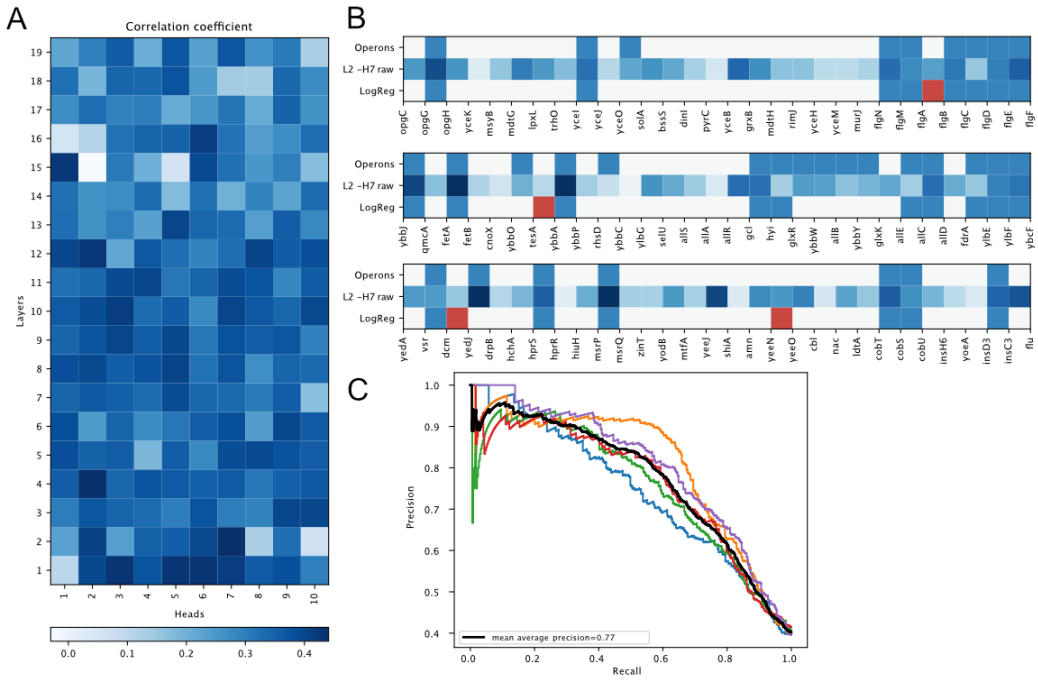

Figure 3: Attention analysis. A) Correlation coefficients (Pearson's rho) between attention heads across layers and operons. Darker color corresponds to stronger correlation with previously identified operons. Attention patterns of the second layer-seventh head [L2-H7] is most strongly correlated with the operons. B) Three random examples of contigs and predicted operonic relationship between neighboring proteins. Proteins are listed in the order they are encoded in the contig. Ground truth E. coli K-12 operons (top row), raw attention scores in the attention head [L2-H7] most correlated with operons (middle row) and logistic regression prediction using all attention heads (last row) where false positive predictions are marked in red. C) Five-fold cross-validation precision-recall curves of logistic regression trained using all operons and attention heads.

enzyme function prediction. First, we generated a custom MGYP-EC dataset where the train and test data were split at 30% sequence identity for each EC class Yu (2023). Second, we apply a linear probe (LP) to compare the expressiveness of representations at each gLM layer, with and without masking the queried protein (Extended Data 8). By masking the queried protein, we can assess gLM's ability to learn functional information of a given protein, only from its genomic context, without the propagation of information from the protein's pLM embeddings. We observed that a large fraction of contextual information pertaining to enzymatic function is learned in the first six layers of gLM. We also demonstrate that context information alone can be predictive of protein function, reaching up to 24.4 ± 0.8% accuracy. In contrast, without masking, gLM can incorporate information present in the context with the original pLM information for each queried protein. We observed an increase in expressivity of gLM embeddings also in the shallower layers, with accuracy reaching up to 51.6 ± 0.5% in the first hidden layer. This marks a 4.6 ± 0.5% increase from context-free pLM prediction accuracy (Fig. 2A) and mean average precision (Fig. 2C) Thus, we demonstrate that information that gLM learns from the context is orthogonal to information captured in pLM embedding. We also observed diminishing expressivity in enzyme function information with deeper layers of gLM; this reflects the masked pretraining objective that is independent of enzyme function prediction task and is consistent with previous examinations of LLMs, where specific layers perform better than others for downstream tasks. Finally, to further examine the expressiveness of these representations, we compared per-class F1 score gains (Fig. 2B). We observe statistically significant differences in F1 scores (t-test, Benjamini/Hochberg corrected p-value < 0.05) between the two models in 36 out of 73 EC classes with more than ten samples in the test set. Majority (27 out of 36) of the statistical differences resulted in improved F1 score in LP trained on gLM representations.

3.3 Transformer's attention captures operons

The transformer attention mechanism models pairwise interaction between different tokens in the input sequence. Previous examinations of the attention patterns of transformer models in natural language processing (NLP) Rogers et al. (2020) have suggested that different heads appear to specialize in syntactic functions. Subsequently, different attention heads in pLMs Vig (2020) have been shown to correlate to specific structural elements and functional sites in a protein. For our gLM, we hypothesized that specific attention heads focus on learning operons, a "syntactic" feature pronounced in in microbial genomes where multiple genes form regulatory modules. We used the E.coli K-12 operon database Salgado (2018) consisting of 817 operons for validation. gLM contains 190 attention heads across 19 layers. We found that heads in shallower layers correlated more with operons (Fig. 3A), with raw attention scores in the 7th head of the 2th layer [L2-H7] linearly correlating with operons with 0.44 correlation coefficient (Pearson's rho, Bonferroni adjusted p-value $<$ 1E-5) (Fig. 3B). We further trained a logistic regression classifier using all attention patterns across all heads. This classifier predicted the presence of an operonic relationship between a pair of proteins in a sequence with mean average precision of 0.77 (Fig. 3C).

4 Discussion

The unprecedented amount and diversity of metagenomic data, coupled with advances in deep learning presents an exciting opportunity for building a large computational model that can learn hidden patterns and structures of biological systems. Such a model builds upon the conceptual and statistical frameworks that evolutionary biologists have developed for the past century. The work presented here demonstrates the concept of genomic language modeling. Our implementation of the masked genomic language modeling illustrates the feasibility of training, evidence of biological information being captured in learned contextualized embeddings, and meaningful interpretability of the attention patterns.

One of the most powerful aspects of the transformer-based language models is their potential for transfer learning and fine-tuning. Promising future directions for applying gLM for advancing biological research include: 1) Fine-tuning gLM for the protein-protein-interactome prediction task, 2) Using gLM features to encode genomic contexts as additional input for improved and contextualized protein structure predictions. Genomic language modeling presents an avenue to bridge the gap between atomic structure and organismal function, and thereby bringing us closer to genomically engineering organisms.

5 Acknowledgements

We would like to thank the EBI MGnify team for generating and maintaining the metagenome database. We would also like to thank Meta AI's ESM developers who made both the folded MGnify proteins structures and source-code openly available

References

Baek, M. e. a. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373:871–876, 2021.

Cordero, O. X. and Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.*, 12:263–273, 2014.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:[cs.CL]*, 2018.

Elnaggar, A. e. a. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:[cs.LG]*, 2020.

Jumper, J. e. a. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021.

Konno, N. and Iwasaki, W. Machine learning enables prediction of metabolic system evolution in bacteria. *Sci Adv*, 9:eadc9130, 2023.

Kountz, D. J. and Balskus, E. P. Leveraging microbial genomes and genomic context for chemical discovery. *Acc. Chem. Res.*, 54:2788–2797, 2021.

Lin, Z. e. a. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023.

Lynch, M. and Conery, J. S. The origins of genome complexity. *Science*, 302:1401–1404, 2003.

Madani, A. e. a. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.*, pp. 1–8, 2023.

Miller, D., Stern, A., and Burstein, D. Deciphering microbial gene function using natural language processing. *Nat. Commun.*, 13:5731, 2022.

Richardson, L. e. a. Mgnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.*, 51:D753–D759, 2023.

Rives, A. e. a. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.*, 118, 2021.

Rogers, A., Kovaleva, O., and Rumshisky, A. A primer in bertology: What we know about how bert works. *Trans. Assoc. Comput. Linguist.*, 2020. doi: 10.1162/tacl_a_00349/96482.

Salgado, H. e. a. Using regulondb, the escherichia coli k-12 gene regulatory transcriptional network database. *Curr. Protoc. Bioinformatics*, 61:1.32.1–1.32.30, 2018.

Shapiro, B. J. e. a. Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336:48–51, 2012.

Treangen, T. J. and Rocha, E. P. C. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet*, 7:e1001284, 2011.

Vaswani, A., Shazeer, N., and Parmar, N. Attention is all you need. In *Adv. Neural Inf. Process. Syst.*, 2017.

Vig, J. e. a. Bertology meets biology: Interpreting attention in protein language models. *arXiv [cs.CL]*, 2020.

Wright, S. On the roles of directed and random changes in gene frequency in the genetics of populations. *Evolution*, 2:279–294, 1948.

Yu, T. e. a. Enzyme function prediction using contrastive learning. *Science*, 379:1358–1363, 2023.