Genomic language model predicts protein co-regulation and function

Yunha Hwang [1] **Andre Cornman** **Elizabeth H. Kellogg** [2] **Sergey Ovchinnikov** [3] **Peter R. Girguis** [1]

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, taxonomy). 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 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. We demonstrate gLM's potential for predicting gene function and regulation, and propose future research directions, including transfer learning capabilities of gLM.

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA [2]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA [3]John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA, USA. Correspondence to: Yunha Hwang <yhwang@g.harvard.edu>.

Preliminary work. Under review by the 2023 ICML Workshop on Computational Biology. Do not distribute.

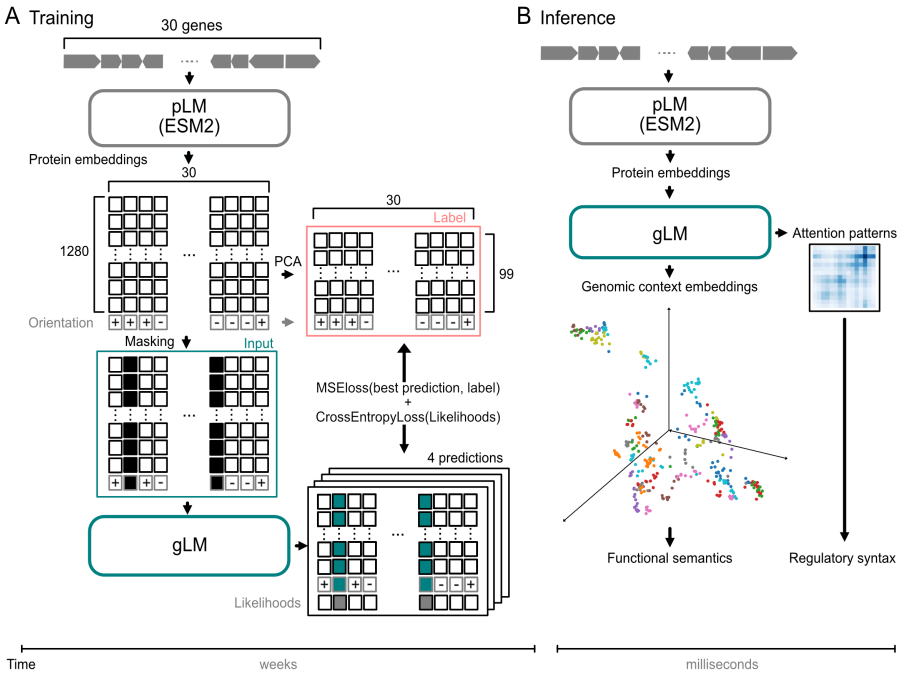

Figure 1. **gLM training and inference schematics**. **A)** Training begins with converting 15-30 gene metagenomic subcontigs to protein embeddings using ESM2. Orientation feature is concatenated for each protein and 15% of the proteins are masked randomly to generate training inputs. Labels are generated by applying PCA dimensionality reduction on the ESM2 protein embeddings, and concatenating the orientation feature. gLM is trained to make four possible predictions for the masked tokens, and their associated likelihoods. Training loss is calculated on both the prediction and likelihoods. The training stage takes several weeks on four NVIDIA A100 GPUs. **B)** At inference time, inputs are generated from a metagenomic subcontig using ESM2 output concatenated with an orientation feature. Hidden states and attention patterns of the trained gLM can be used for various downstream tasks.

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.

2.2. Results

2.3. Contextualized gene embeddings capture gene semantics

The mapping from gene to gene-function in organisms is not one-to-one. Similar to words in natural language, a gene can confer many different functions (Jeffery, 2018) depending on its context (Miskei, 2017), and many genes can confer similar functions (i.e. convergent evolution (Gherardini et al., 2007), remote homology (Ben-Hur & Brutlag, 2003)).

We explored an ecologically important example of genomic "polysemy" (multiple meanings conferred by the same word) of methyl-coenzyme M reductase (MCR) complex (Fig. 2ABC). The MCR complex is able to carry out a reversible reaction (Reaction 1 in Fig. 2D), whereby the forward reaction results in the production of methane (methanogenesis) while the reverse results in methane oxidation (methanotrophy). We first examine the McrA (methyl-coenzyme M reductase subunit alpha) protein in diverse lineages of ANME (ANaerobic MEthane oxidizing) and methanogenic archaeal genomes. These archaea are polyphyletic and occupy specific ecological niches. Notably, similar to how a semantic meaning of a word exists on a spectrum and

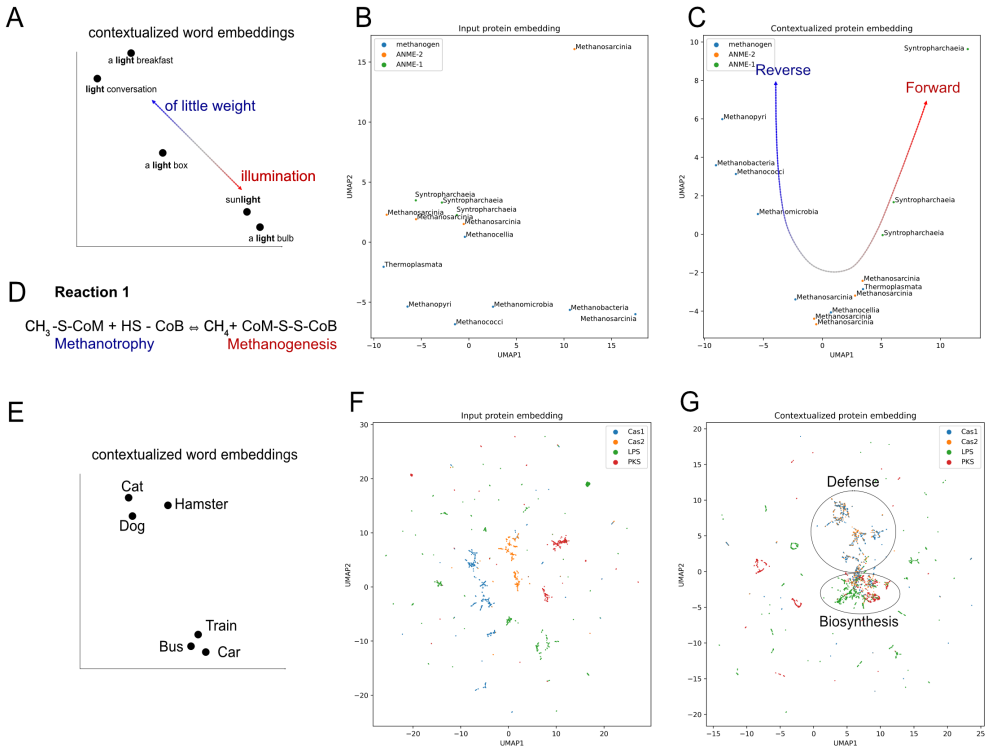

Figure 2. **Contextualized protein embedding analysis and comparison with concepts in natural language modeling. A**) A word's meaning upon contextualization varies across a continuous spectrum and can be ambiguous even with contextualization (e.g. double entendre). **B**) Input protein embeddings of McrA sequences in genomes, colored by metabolic classification of the organism (ANME, methanogen) based on previous studies and labeled by class-level taxonomy. **C**) Clustering of McrA sequences upon contextualization, with the likelihoods in the direction of Reaction 1 that the MCR complex carries out. **D**) Reaction 1, carried out by the MCR complex, either backward (Methanotrophy) or forward (Methanogenesis). **E**) Geometric relationship between contextualized protein embeddings based on the semantic closeness of words. **F**) Input (context-free) protein embeddings of Cas1, Cas2, lipopolysaccharide synthases (LPS) and polyketide synthases (PKS) showing clustering based on structural and sequence similarity. **G**) Clustering of contextualized protein embeddings where phage defense proteins cluster (Cas1 and Cas2) and biosynthetic gene products cluster (LPS and PKS).

a word can have multiple semantically appropriate meanings in a context (Fig. 2B), the MCR complex can confer different functions depending on the context. Previous reports demonstrate capacities of ANME (ANME-2 in particular) carrying out methanogenesis (Bertram, 2013) and methanogens conducting methane oxidation in specific growth conditions (Moran et al., 2007). The context-free ESM2 embedding of these proteins (Fig. 2E) shows little organization, with little separation between ANME-1 and ANME-2 McrA proteins. However, contextualized gLM embeddings Fig. 2C) of the McrA proteins show distinct organization where ANME-1 McrA proteins form a tight cluster, while ANME-2 McrA proteins form a cluster closer to methanogens. This organization reflects the phylogenetic relationships between the organisms that McrAs are found in, and reflect distinct operonic and structural divergence of MCR complexes in ANME-1 compared to those found in ANME-2 and methanogens (Shao, 2022). As proposed by Shao et al., the preferred directionality in Reaction 1 (Fig. 2G) in ANME-2 and some methanogens may be more dependent on thermodynamics.

We also demonstrate that contextualized gLM embeddings are more suitable for determining the functional relationship between gene classes. Analogous to how the words "dog" and "cat" are closer in meaning relative to "dog" and "train" (Fig. 2E), we see a pattern where Cas1 and Cas2 that appear diffuse in multiple subclusters in context-free protein embedding space (Fig. 2F) cluster in contextualized embedding space (Fig. 2G). This reflects their similarity in function (e.g. phage defense). This is also demonstrated in biosynthetic genes, lipopolysaccharide synthase (LPS) and polyketide synthase (PKS) genes clustering closer together in contextualized embedding space distinct from the Cas proteins (Fig. 2G). Contextualized protein embeddings are therefore able to capture relational properties semantic information (Reif, 2019), where proteins that are more similar in their function appear in more similar genomic contexts.

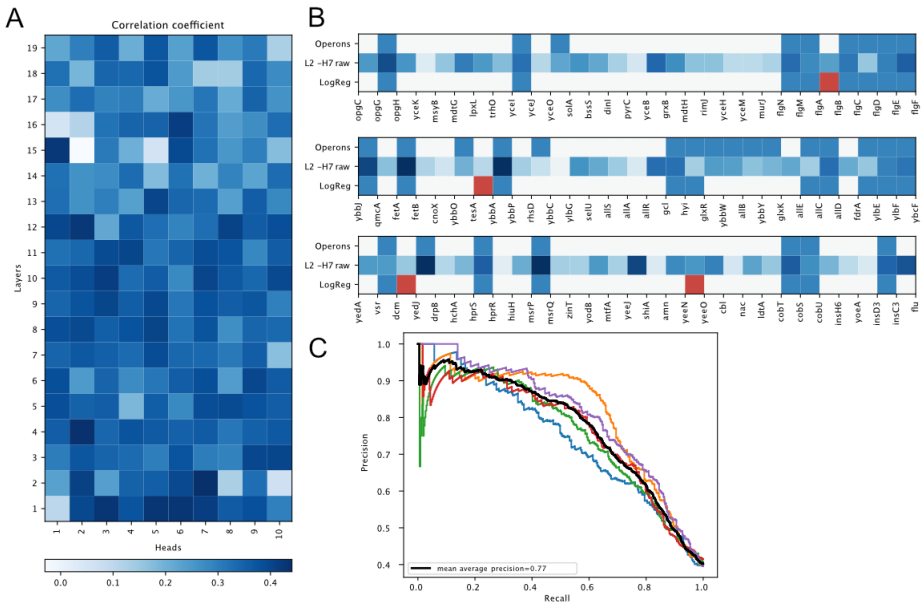

Figure 3. **Attention analysis**. **A**) Correlation coefficients (Pearson's rho) between attention heads across layers and operons. **B**) Three random examples of ground truth operons (top row), raw attention scores (middle row) between neighboring proteins in the highest correlating attention head and logistic regression prediction using all attention heads (last row) where false positive predictions are marked in red. **C**) Cross-validation precision-recall curve of logistic regression trained using all operons and attention heads.

2.4. Transformer's attention captures operons

The transformer attention mechanism16 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 in 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).

3. 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.

References

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

Ben-Hur, A. and Brutlag, D. Remote homology detection: a motif based approach. *Bioinformatics*, 19(Suppl 1): i26–i33, 2003.

Bertram, S. e. a. Methanogenic capabilities of anmearchaea deduced from13c-labelling approaches. *Environmental Microbiology*, 15:2384–2393, 2013. doi: 10.1111/1462-2920.12112.

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.

Gherardini, P. F., Wass, M. N., Helmer-Citterich, M., and Sternberg, M. J. E. Convergent evolution of enzyme active sites is not a rare phenomenon. *J. Mol. Biol.*, 372: 817–845, 2007.

Jeffery, C. J. Protein moonlighting: what is it, and why is it important? *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 373, 2018.

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.

Miskei, M. e. a. Fuzziness enables context dependence of protein interactions. *FEBS Lett.*, 591:2682–2695, 2017.

Moran, J. J., House, C. H., Thomas, B., and Freeman, K. H. Products of trace methane oxidation during nonmethyltrophic growth bymethanosarcina. *Journal of Geophysical Research*, 112, 2007. doi: 10.1029/2006jg000268.

Reif, E. e. a. Visualizing and measuring the geometry of bert. In *Adv. Neural Inf. Process. Syst.*, volume 32, 2019.

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.

Shao, N. e. a. Expression of divergent methyl/alkyl coenzyme m reductases from uncultured archaea. *Commun Biol*, 5:1113, 2022.

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.