QUANTORI

Beyond Words: The Expanding Role of Language Models in Biology

Omar Kantidze

Principal Scientist

27 November 2023

White Paper

TABLE OF CONTENTS

EXECUTIVE SUMMARY

It's a characteristic of human nature to find new ways to use recently discovered tools, especially when scientists adapt them to address practical problems that spark their curiosity.

Transformer-based pre-trained language models (PLMs) such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-Trained Transformers (GPT) have greatly changed the contemporary landscape of artificial intelligence (AI), and the real-world applications of such PLMs are continually expanding. Although originally designed for working with natural languages, these models have proven equally effective in handling languages of nature, such as nucleotide and amino acid sequences.

In this brief article, we aim to illustrate how these language models are reshaping the domains of biomedical text mining, functional genomics*, and protein engineering**, thus opening up fresh avenues for research and the development of innovative biologics-based therapies.

*Functional genomics investigates how genes and intergenic regions of the genome contribute to different biological processes.

**Protein engineering refers to the development of useful or valuable proteins through the design and production of unnatural polypeptides.

‘ATTENTION IS ALL YOU NEED' – TRANSFORMER, BERT AND GPT

The title of one of AI's holy texts, authored by Vaswani and colleagues in 2017, marks a significant milestone. In this publication, they unveiled the Transformer, a novel neural network architecture specifically tailored for processing sequences of data, particularly texts. A pivotal innovation within the Transformer was the introduction of "self-attention" mechanisms, facilitating a profound understanding of the intricate interplay between words within a sentence. Transformers revolutionized the field of natural language processing (NLP) and enabled the development of more advanced models like BERT and GPT (Devlin et al., 2018; Radford et al., 2018).

Both BERT and GPT are built on the Transformer architecture, need to be pre-trained on a massive amount of text data, and can be fine-tuned for specific NLP tasks. However, they have notable differences that dictate their respective practical applications. BERT excels in understanding language bidirectionally, taking into account the context of a word by analyzing the words that precede and follow it in a sentence. This affords it a more comprehensive grasp of word meanings in context. In contrast, GPT employs a unidirectional context, its primary role being to predict the subsequent word in a sentence based on the preceding words. This predictive capacity empowers GPT to generate coherent and contextually relevant texts.

TREASURES OF PUBMED – BIOMEDICAL TEXT LANGUAGE MODELS

Considering that the primary purpose of language models like BERT and GPT is text processing and insights extraction, it was evident that PLMs specifically trained on biomedical data corpora would quickly emerge. These datasets encompass various sources such as biomedical literature (abstracts and full-text articles, medical textbooks, clinical reports), electronic health records (EHRs), medical databases (e.g., DrugBank, ClinicalTrials.gov), websites, and ontologies (e.g., Unified Medical Language System, Gene Ontology) (Wang et al., 2023; Tian et al., 2023). This diverse data landscape allowed for fine-tuning these models to address a wide array of tasks within the biomedical field. Domain-specific PLMs can classify documents, perform named entity recognition, extract clinical information, answer medical questions, predict drug-drug and drug-target interactions, identify new targets, analyze genotype-phenotype relationships, summarize medical texts, offer clinical decision support, analyze sentiment in healthcare contexts, and even generate descriptions for medical images (Wang et al., 2023; Xie et al., 2022; Tian et al., 2023).

Since the introduction of BERT in 2018, over 40 biomedicine domain-specific text-mining PLMs have been developed. The majority of these models are based on BERT-like architectures (e.g., BioBERT, PubMedBERT, BioLinkBERT, DRAGON). Still, there are a few based on GPT (GPT-Neo, PubMed GPT, MedGPT) or PaLM architectures (Med-PaLM and Med-PaLM 2). The latter ones deserve special attention, as they are likely the most robust models, significantly outperforming others in accuracy when responding to US Medical Licensing Examination (USMLE)-style questions (MedQA dataset) (Singhal et al., 2023; Tian et al., 2023).

Looking ahead, it's likely that biomedical PLMs will become indispensable tools for researchers, clinicians, and drug discovery experts. In the healthcare field, their active integration into the analysis of EHRs and clinical decision support could greatly benefit from the establishment of consensus and regulations regarding the privacy of personal data in AI applications.

WHAT IS IN YOUR DNA? – GENOMIC LANGUAGE MODELS

The human genome is a continuous sequence consisting of four letters corresponding to nucleotides and is over three billion characters long. Genetic code deciphered by F. Crick provides an understandable way to translate such a code, broken into words-genes, into protein sequences. However, this decoding only works with protein-coding genome regions. A significant portion of the genome does not encode proteins but contains information about where and how genes should be expressed. These functional genome elements* (e.g., promoters, enhancers, insulators) often have distinct functions and activities in different biological contexts, can interact with each other over long distances, and form assemblies, thus providing a multilayer regulatory landscape. This suggests the presence of polysemy** and distant semantic relationships within nucleotide sequences, which are key characteristics of natural languages.

It came as no surprise that by mid-2020, DNABERT, a BERT-based model pre-trained on unlabeled human genome sequences, had already been established (Ji et al., 2021). This achievement served as a compelling demonstration of how Transformer-based architectures can be leveraged effectively to acquire comprehensive and adaptable insights into DNA, offering a 'one-model-does-it-all' approach. Subsequent fine-tuning with small task-specific labeled data allowed such genomic PLMs to tackle tasks like predicting promoters, splice sites, transcription factor binding sites, enhancer activity, chromatin profiles, and polyadenylation site strength.

Up to the present, only a handful of foundational genomic PLMs have emerged, including DNABERT-1 and -2 (in 2020 and 2023, respectively; (Ji et al., 2020; Zhou et al., 2023), BigBird (Zaheer et al, 2020), GENA-LM (Fishman et al., 2023), Nucleotide Transformer (Dalla-Torre et al., 2023), and HyenaDNA (Nguyen et al., 2023).

*Functional genome elements or cis-regulatory elements are non-coding DNA sequences that control gene expression by serving as binding sites for transcription factors, influencing when, where, and to what extent genes are transcribed.

**Polysemy in linguistics refers to a phenomenon in which a single word or phrase has multiple, related meanings or senses that are often connected by a common underlying concept.

While these models are primarily pre-trained on the human genome, it's worth noting that some, such as DNABERT-2 and Nucleotide Transformer, have also made use of multispecies datasets. Although functionally, these foundational genomic PLMs are similar, special attention is warranted for GENA-LM and HyenaDNA thanks to some of their key characteristics like the number of parameters, maximum input size, and context length. From an architectural standpoint, all of these models, except for HyenaDNA, can be regarded as variations of the BERT. The distinctive feature of HyenaDNA lies in its S4-related architecture (Toews 2023), enabling it to work with significantly longer contexts. Most genomic PLMs typically operate with context lengths ranging from 512 to 10k tokens*, whereas HyenaDNA uses an extended context length of up to 1 million tokens with single-nucleotide resolution. This has the potential to impact both the model's accuracy and its ability to tackle tasks that were previously beyond the reach of other genomic PLMs.

Conversations with some of the developers of genomic PLMs reveal that these models are in search of fresh and non-obvious challenges for their application. The advancement of this field would greatly benefit from the expertise of professionals engaged in translational biology and drug discovery. Moreover, it would be intriguing to witness the evolution of this domain towards generative AI, which holds promise from the perspective of synthetic biology**.

____

*In NLP, a "token" is a fundamental unit of text, which can be a word or a phrase, created through the process of tokenization to enable computational analysis of language. In genomic and protein PLMs, a token typically refers to individual nucleotides, amino acids, or their short sets, such as k-mers, which are used as the basic units for modeling and analysis.

**Synthetic biology is an interdisciplinary field that involves designing and engineering biological systems, often through genetic manipulation, for various practical applications.

DESIGNING TOOLS OF LIFE – PROTEIN LANGUAGE MODELS

When AI and proteins are mentioned together, AlphaFold and, for some, RosettaFold spring to mind. However, we won't delve into them here because these crucial tools for predicting the 3D structure of proteins* are not directly related to Transformer-based PLMs. Both AlphaFold and RosettaFold are based on the multiple sequence alignment** (MSA) approach, which essentially means making predictions based on the similarity of the primary structure (sequence) of proteins with known and unknown 3D structures. Thus, the quality of structure prediction hinges on the availability of homologs with known 3D structures for the protein of interest. This is where Transformer-based protein PLMs come into play, offering a promising application. Recent studies have already demonstrated their ability to outperform MSA-based approaches like AlphaFold2 in structure prediction for sequences that lack homologs*** (Lin et al., 2023).

Much like DNA, proteins possess their own unique code to determine their primary sequence, comprised of twenty letters that correspond to general amino acids. The availability of millions of protein sequences from various organisms has paved the way for the development and pre-training of foundational *protein PLMs*. Three primary applications of Transformer-based protein PLMs have emerged:

- protein structure determination,
- protein properties prediction,
- creation of novel protein molecules with tailored structures and properties.

*The three-dimensional (3D) structure of a protein refers to the specific spatial arrangement of its atoms, including the positions of individual amino acids, in a complex, folded, and functional form, which is crucial for its biological function.

**Multiple Sequence Alignment is a bioinformatics technique used to compare and align multiple biological sequences to identify similarities, differences, and conserved regions.

***A protein homolog is a protein that is evolutionarily related to another protein, sharing similar structural or sequence characteristics.

Protein Structure Prediction

Despite the prevailing use of MSA-based techniques as state-of-the-art tools in protein structure prediction, PLMs introduce novel prospects in this domain. For example, ESM-2, one of the most recent foundational protein PLMs, showcases the capability of language models to rapidly predict atomic-resolution protein structures directly from sequences (Lin et al., 2023). This breakthrough facilitates a departure from resource-intensive and costly structure prediction pipelines, eliminating the need for multiple sequence alignments, a requirement in methodologies like AlphaFold2 and RosettaFold. Furthermore, protein PLMs can be invaluable in tackling challenges associated with predicting the structures of orphan and rapidly evolving proteins, for which an MSA is unapplicable (Lin et al., 2023; Qiu & Wei, 2023). These models can also extend their utility to predicting the quaternary structure of proteins, i.e. the structures of protein oligomers (Avraham et al., 2023).

Protein Properties Prediction

Transformer-based protein PLMs, including models like ProteinBERT, Tranception, ESMs, ProGen, and others, have the capability to either predict global properties of proteins, such as their type, function, or cellular localization or infer local properties of specific protein residues, such as a corresponding 2D/3D structure or post-translation modifications (phosphorylation, cleavage sites, etc.) (Chandra et al., 2023). Furthermore, they are adept at addressing vital tasks in the field of drug discovery – identifying disease-causing mutations and predicting protein-protein and drug-protein interactions (Liu et al., 2022; Chandra et al., 2023). There have also been attempts to use protein PLMs in predicting the properties of clinically important biologics such as antibodies and peptides (Wang et al., 2023; Guntuboina et al., 2023).

Creation of Novel Proteins

Finally, protein PLMs offer a robust framework for generating a wide range of authentic proteins (Romero-Romero et al., 2023). Typically, PLMs generate sequences that are anticipated to conform to well-defined structures, even when they exhibit significant primary sequence divergence (Ferruz et al., 2022). The fine-tuning of these protein PLMs on specific protein families has showcased their capability to create protein sequences with precisely tailored properties. For example, when the ProGen model was employed to generate artificial lysozymes through this method, these synthetic lysozymes possessed enzymatic activities close to natural lysozymes, even when their sequence similarity was as low as 40-50% (Madani et al., 2023). Another noteworthy PLM, ZymCTRL, which was trained on enzyme sequences and their accompanying annotations, has effectively facilitated the design of protein sequences capable of executing user-defined enzymatic reactions (Munsamy et al., 2022). There are examples of protein PLMs tailored for even more specialized tasks – the antibody-specific IgLM model could produce diverse sets of antibodies mirroring those naturally present in immune repertoires (Shuai et al., 2022).

Owing to the undeniable importance of AI-driven protein research and protein engineering for drug discovery, accompanied by a wide array of diverse translational challenges, this field has gained significant progress. While the initial integration of AI into protein-related problem-solving didn't commence with language models, it has become evident that they have a meaningful role to play within this complex landscape. Highlighting the emergence of a novel frontier, the de novo design of protein molecules, it is worth noting that this rapidly evolving area is employing entirely distinct AI methodologies, including diffusion probabilistic models (Ingraham et al., 2022; Watson et al., 2023; Ni et al., 2023) and protein language models (Romero-Romero et al., 2023). The substantial investor confidence in startups focused on AI-based methods for de novo protein generation (Grinstein 2023) has the potential to trigger a rapid expansion of these technologies in the near future, facilitating the active integration of their products into clinical applications.

We develop cutting-edge technology systems, applications, and infrastructures for biotech, pharmaceutical, and healthcare companies that accelerate drug discovery and improve patient outcomes. Our innovative approach harnesses the power of data engineering and informatics, machine learning, emerging technologies, and cloud expertise to advance research and development and ultimately bridge the gap between meaningful data and patient success.

Omar Kantidze

Principal Scientist, Quantori

27 November 2023