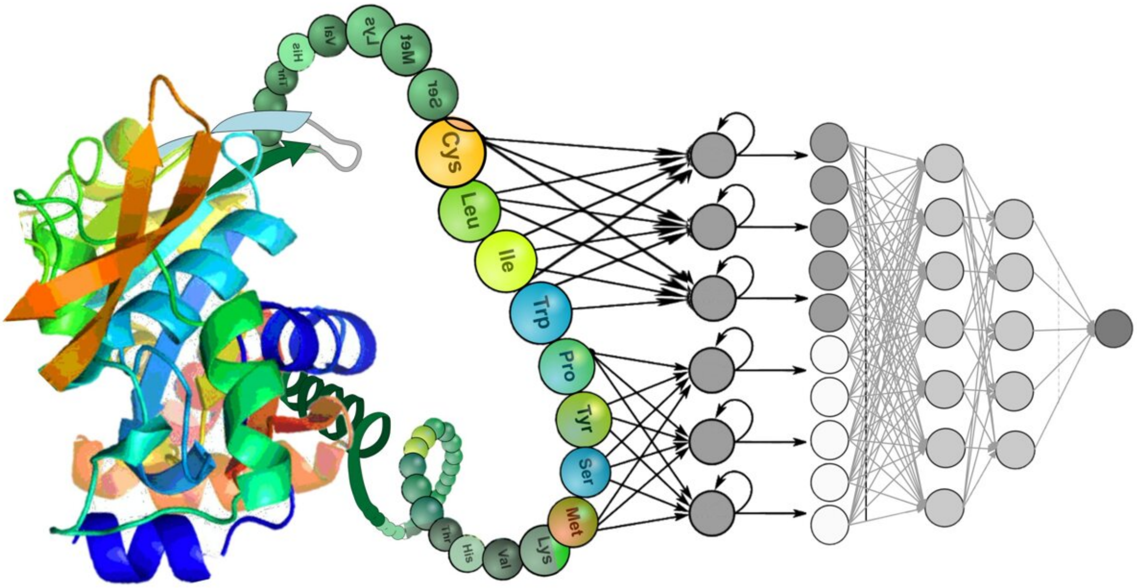
PHYS ORG

Study evaluates deep learning models that decode the functional properties of proteins

April 18 2022, by Ingrid Fadelli



Complex functional information hidden in amino acid sequence patterns are automatically learned by large-scale and self-supervised deep learning models Credit: Unsal et al

Deep learning–based language models, such as BERT, T5, XLNet and GPT, are promising for analyzing speech and texts. In recent years, however, they have also been applied in the fields of biomedicine and biotechnology to study genetic codes and proteins.

Bioinformaticians, genetics researchers and neuroscientists have been trying to infer the biological roles of genes and proteins for decades. To do this, however, they need to analyze extremely large and highly complex biological data.

Researchers at Hacettepe University, Middle East Technical University and Karadeniz Technical University, Turkey, have recently carried out a study evaluating the potential of deep learning–based language models for studying proteins and predicting their functional properties. Their paper, published in *Nature Machine Intelligence*, provides a valuable summary of the advantages and disadvantages of different state-of-the-art approaches.

"Molecular biology data can be modeled as a language (i.e., the language of genes/proteins), such that the sequence of a gene or protein can be thought of as a sentence with a specific meaning in natural language, and the semantics of this protein language is the specific biological, physical and chemical properties of these biomolecules," Tunca Doğan, one of the researchers who carried out the study, told Phys.org. "Based on this idea, our work tries to build machine learning models that take language model-derived high dimensional numerical embeddings of proteins as input and predict their functional properties with high accuracy."

In their paper, Doğan and his colleagues assessed the ability of different protein language modeling approaches to extract hidden patterns containing important clues about the functional properties of proteins. Their evaluations included all the most well-known natural language modeling architectures (i.e., BERT, T5, XLNet, ELMO, etc.), each of which can contain hundreds of millions or, in some cases, billions of parameters.

"Self-supervised pre-training of these models requires huge resources," Doğan explained. "Thanks to valuable previous work on this topic, which aimed to pre-train protein language models using these architectures, we mostly focused on our secondary supervised training for predicting functional properties."

In order to effectively assess the protein language models and compare their performances, the team first had to compile large and reliable testing datasets, each with a different difficulty level. Ultimately, they created four benchmark datasets that allowed them to investigate semantic similarities, ontology-based functional definitions, drug target protein families, and physical interactions between proteins. All of these are crucial biological mechanisms that are known to be closely linked to the occurrence and progression of genetically inherited diseases, such as different types of cancer.

"Perhaps our most notable finding was that these deep language models are able to successfully learn the functional properties of proteins using the amino acid sequences as the sole input, which is quite a difficult problem," Doğan said. "These results are also consistent with the findings of recent protein structure prediction studies (e.g., Deepmind's AlphaFold2 and Baker Lab's RoseTTAFold), which uses the sequence as its input and predicts the 3D monomer structure with extremely high performance."

In the future, the models evaluated by this team of researchers could help to enhance precision medicine interventions, for instance analyzing the molecular make-up of patients resulting from genomic variations to devise personalized treatments. While the results gathered by Doğan and his colleagues highlight the huge potential of deep learning–based protein modeling tools, existing methods will still need to be significantly improved before they can be integrated into real-life clinical decision-making systems.

"We are now working on a new system to better represent proteins," Doğan added. "In addition to amino acid sequences, this system utilizes network-based data (i.e., known protein-protein interactions) and knowledge hidden in the unstructured biomedical texts (e.g., scientific articles) at the input level, together with integrative deep learning approaches. Our ultimate aim is to obtain a universal protein representation that can successfully be used in any biomedical or biotechnological modeling task."

More information: Serbulent Unsal et al, Learning functional properties of proteins with language models, *Nature Machine Intelligence* (2022). DOI: 10.1038/s42256-022-00457-9

Citation: Study evaluates deep learning models that decode the functional properties of proteins (2022, April 18) retrieved 16 May 2024 from https://phys.org/news/2022-04-deep-decode-functional-properties-proteins.html