II. LITERATURE REVIEW

Accurate Prediction of Protein-Nucleic Acid Complexes Using RoseTTAFoldNA

Introduction

Recent advances in computational biology have focused on the accurate prediction of protein-nucleic acid complexes, a crucial aspect of understanding biological functions and designing therapeutic interventions. The paper in question introduces RoseTTAFoldNA, a novel machine-learning approach that significantly enhances the capability to predict the structures of protein-nucleic acid complexes. This method builds upon the original RoseTTAFold technique, extending its predictive power to encompass both nucleic acids and their complexes with proteins. The significance of this advancement lies in its ability to model naturally occurring complexes and design sequence-specific RNA and DNA-binding proteins with a higher degree of accuracy than previously possible.

Methodology

RoseTTAFoldNA is trained on a diverse dataset from the Protein Data Bank (PDB), encompassing protein monomers, complexes, RNA monomers, dimers, and protein-RNA/DNA complexes. It employs a sophisticated architecture to achieve its predictive accuracy and incorporates several innovative features including:

Additional Tokens for Nucleic Acids (NA) in the 1D Track: Integration of 10 extra tokens for the four DNA and RNA nucleotides, alongside tokens for unknown DNA and RNA, enriches the model's capability to recognize and process nucleic acids alongside the 22 tokens used for amino acids.

Interactions Modelling in the 2D Track: Generalization of the 2D track to include interactions between nucleic acid bases and between bases and amino acids allows the model to capture the complex interplay within protein-nucleic acid complexes.

Detailed Representation of Nucleotides in the 3D Track: Extending the 3D track to accurately represent nucleotides facilitates the model's ability to construct all atoms in the nucleotide, mirroring the detailed approach used for amino acids.

The architecture is underpinned by a network of 36 three-track layers, supplemented by four structure refinement layers, totalling 67 million parameters.

Results

RoseTTAFoldNA showcases superior performance in predicting the structure of protein-nucleic acid complexes. Through the application of a comprehensive loss function, the model optimizes network parameters to accurately reflect the intricate interactions between proteins and nucleic acids. The paper delineates how RoseTTAFoldNA surpasses existing state-of-the-art methods, including its precursors, AlphaFold and RoseTTAFold, particularly in the domain of protein-nucleic acid complex prediction.

Comparison with Existing Methods

The literature review elucidates a detailed comparison of RoseTTAFoldNA with other methods:

DeepFoldRNA: RoseTTAFoldNA demonstrates comparable performance with average Local Distance Difference Test (lDDT) scores but exhibits superior accuracy in high-confidence predictions.

FARFAR2: RoseTTAFoldNA significantly outperforms FARFAR2, indicating its enhanced predictive accuracy.

CASP15 RNA Targets: While RoseTTAFoldNA's performance on CASP15 RNA targets does not surpass all contemporary machine-learning methods, its unique capability to predict protein-nucleic acid complexes underscores its importance.

Conclusion

RoseTTAFoldNA represents a significant advancement in the computational prediction of protein-nucleic acid complexes. By leveraging a deep learning architecture and a comprehensive training dataset, it not only improves upon the capabilities of previous models but also introduces a novel approach to understanding the complex interactions within these vital biological structures. The paper underscores the potential of RoseTTAFoldNA to revolutionize the modelling of naturally occurring complexes and the design of sequence-specific binding proteins, marking a pivotal step forward in the field of computational biology.

Highly accurate protein structure prediction with AlphaFold

Introduction

The groundbreaking work presented in "Highly accurate protein structure prediction with AlphaFold" introduces a major advancement in the field of computational biology. This study, led by the team at DeepMind, showcases the development of AlphaFold, a neural network-based model capable of predicting protein structures with unprecedented atomic accuracy. By leveraging a novel machine learning approach that integrates physical and biological insights into protein structures, along with multi-sequence alignments, AlphaFold significantly surpasses existing methods in protein structure prediction. This achievement not only marks a pivotal moment in solving the decades-old protein folding problem but also opens new avenues for large-scale structural bioinformatics, promising to accelerate scientific discoveries in biology and medicine.

Methodology

The AlphaFold model was trained on a comprehensive dataset that includes multiple sequence alignments (MSAs) and structural data from protein databases.

The training involved several key datasets and strategies:

Uniclust Dataset: An initial model was trained using structure predictions for a Uniclust dataset consisting of 355,993 sequences with full MSAs.

Final Model Training: The final model was trained with identical hyperparameters, sampling examples 75% of the time from the Uniclust prediction set (with sub-sampled MSAs) and 25% of the time from a clustered Protein Data Bank (PDB) set.

Big Fantastic Database (BFD): One of the sequence databases used was the Big Fantastic Database (BFD), a custom-made and publicly released collection used by several CASP teams. BFD is one of the largest publicly available collections of protein families, comprising 65,983,866 families represented as MSAs and hidden Markov models (HMMs), covering 2,204,359,010 protein sequences from reference databases, metagenomes, and metatranscriptomes.

MSA Depth Analysis: The model also utilized a per-residue normalized number of effective sequences (Neff) for each position of a query sequence, obtained by counting the number of non-gap residues in the MSA for this position and weighting the sequences using the Neff scheme.

The model signifies a huge leap in protein structure prediction, achieving near-experimental accuracy. This advancement is attributed to several key features and innovative methodologies:

Deep Learning Neural Network: Unlike its predecessors, AlphaFold's architecture is specifically designed for high accuracy, leveraging an advanced deep learning framework that doesn’t require ensembling, thus enhancing prediction speed by eight times without sacrificing accuracy.

Attention-Based Mechanism: Central to its design is an attention-based neural network, enabling the model to effectively capture long-range interactions between amino acids by focusing on specific sequence elements during the prediction process.

Integration of Evolutionary Information: The incorporation of multiple sequence alignments provides essential evolutionary insights, aiding in accurate residue contact prediction and overall structure determination.

Spatial Graph Convolutional Networks: To accurately model the geometric relationships between amino acids, AlphaFold employs spatial graph convolutional networks, crucial for understanding the protein's three-dimensional conformation.

End-to-End Prediction with Refinement Process: AlphaFold uniquely predicts protein structures directly from amino acid sequences, applying a refinement process post-prediction to improve structure accuracy based on physical and biological constraints.

Validation and Impact: Frame-aligned point error (FAPE), in the context of AlphaFold, might refer to a metric or methodology employed specifically by the model for assessing the accuracy of predicted protein structures by measuring the alignment error of predicted points (atoms in the protein structure) against their true positions in the experimental or reference structure.

Results

In the study led by John Jumper and colleagues, the AlphaFold model demonstrated exceptional prowess in predicting protein structures with near-experimental precision, establishing a new benchmark in computational biology. The model's performance in the CASP14 competition as 'AlphaFold2' highlighted its unprecedented accuracy over prior methods. Training on a comprehensive dataset, including the Uniclust dataset, showcased its efficient learning from a vast array of protein sequences and structures. This achievement not only signifies a leap forward in structural bioinformatics but also suggests AlphaFold's utility in addressing broader biophysical challenges, potentially revolutionizing biological research and drug discovery avenues.

Conclusion

The culmination of "Highly accurate protein structure prediction with AlphaFold" by John Jumper and his team at DeepMind signifies a landmark achievement in computational biology. AlphaFold's methodology, employing deep learning and an innovative neural network architecture, has demonstrated the ability to predict protein structures with near-experimental accuracy. This research not only overcomes a long-standing challenge in the protein folding problem but also heralds a new era for structural bioinformatics. With its exceptional performance in CASP14 and efficient training on extensive datasets like Uniclust and BFD, AlphaFold paves the way for significant advancements in understanding biological processes, disease mechanisms, and drug discovery. The model's impact extends beyond protein structure prediction, suggesting its potential to solve a wide range of biophysical problems and enhancing our capability to explore the complexities of life at a molecular level.

Identification of human lineage-specific transcriptional coregulators enabled by a glossary of binding modules and tunable genomic backgrounds

- Created a glossary of 108 non-redundant TF-8mer "modules" for 671 metazoan TFs from protein binding microarray data .

- Compared the precision of 8mer modules with position weight matrices using ENCODE TF chromatin immunoprecipitation sequencing datasets and RNA sequencing profiles .

- Developed GENRE (genomically equivalent negative regions) for constructing matched genomic background sequences for regulatory region analysis .

- Identified indirect binding motifs and associated tethering TFs more precisely than position weight matrices using the 8mer modules .

- GENRE outperformed four state-of-the-art approaches for background sequence construction .

- Suggested novel TF-TF interactions related to indirect binding motifs by analyzing the co-occurrence of tethering factors .

- Used 239 ENCODE TF chromatin immunoprecipitation sequencing datasets and associated RNA sequencing profiles for analysis .

- Anticipated that the tools developed will aid in elucidating tissue-specific gene-regulatory programs .

GENRE (genomically equivalent negative regions) is a tool developed in the paper "Identification of Human Lineage-Specific Transcriptional Coregulators Enabled by a Glossary of Binding Modules and Tunable Genomic Backgrounds" to construct matched genomic background sequences for the analysis of regulatory regions. It aims to provide a tunable tool for creating negative control regions that are equivalent to the genomic regions of interest. GENRE outperformed four state-of-the-art approaches to background sequence construction, making it a valuable resource for studying regulatory regions.

- The paper created a glossary of TF-8mer "modules" that have shared specificity for TFs, allowing for the identification of direct and indirect TF binding sites .

- The analysis of ENCODE TF chromatin immunoprecipitation sequencing datasets and associated RNA sequencing profiles using these 8mer modules showed their effectiveness in identifying indirect binding motifs and their associated tethering TFs .

- The paper also used these tools to analyze the co-occurrence of tethering factors, suggesting novel TF-TF interactions .

- By developing GENRE, a tunable tool for constructing matched genomic background sequences, the paper provided a means to analyze regulatory regions and study tissue-specific gene-regulatory programs .

Overall, the paper's tools and analyses contribute to predicting TF binding with DNA and understanding gene regulation.