Introduction

Proteins are biomolecules that are fundamental to nearly all biological processes. Their diverse roles include transporting nutrients, catalyzing chemical reactions, providing structural support, and more. In alignment-based analyses, the function of a protein is determined by its primary sequence, a chain composed of 20 amino acids and an additional symbol representing an alignment gap. A single protein sequence can vary greatly in length and order of its amino acids, leading to a very large number of possible configurations.

The size of the protein sequence space makes exhaustive experimental exploration infeasible. However, the rapid growth of biological databases, driven by advances in sequencing technologies has transformed biology into a data-rich discipline. Large repositories such as UniProt, Pfam, and the Protein Data Bank store millions of sequences and structures, providing crucial resources for computational approaches.

This wealth of data has enabled the development of advanced statistical and machine learning models capable of simulating protein sequence evolution, prediciting structural conformations, and generating novel seuqneces with desired properties. Breakthroughs in protein structure prediction, most notably the Nobel Prize winning AlphaFold, demonstrated that computational methods can rival experimental accuracy, in a much cheaper manner.

In this work, I explore the implementation of an autoregressive neural network for protein sequence generation leveraging the Direct Coupling Analysis (DCA) framework to model residue-residue dependencies as defined in [REFER PAPER]. I then evaluate the generated sequences for structural plausibility and functional relevance using AlphaFold3 and Boltz-2.

Biological Background

Proteins

Proteins are essential biological molecules responsible for a wide range of functions in living organisms. Despite their functional diversity, all proteins are polymers of the same set of 21 standard building blocks: the 20 canonical amino acid and one alignment gap symbol used in bioinformatic sequence analysis.

Protein Structure Levels

Each amino acid shares a common core structure consisting of a central carbon atom bonded to a hydrogen atom, an amino group, a carboxyl group, and a variable side chain. This side chain is the defining feature of each amino acid, giving rise to differences in size, shape, chemical reactivity, and polarity. The chemical composition of amino acids involves carbon, hydrogen, oxygen, and nitrogen atoms, with some also containing sulfur. Based on the properties of their side chains, amino acids can be broadly classified as hydrophobic (nonpolar), hydrophilic (polar), and charged. The specific sequence of amino acids dictates how a protein folds into its three-dimensional structure, which in turn dictates its function. Even small changes in sequence can dramatically affect stability, activity, and interaction patterns.

Multiple Sequence Alignments

Studying proteins at the sequence level can be challenging because of the immense diveristy and complexity of possible configurations. However, comparitive analyses across related proteins, grouped into protein families, allow researchers to identify conserved positions and co-evolving residue pairs, offering insights into structural and functional constraints. This is the basis of approaches such as Direct Coupling Analysis, which seeks to uncover the statistical dependencies between amino acid positions that reflect physical contacts in the folded structure.

Brief Review of the paper

- what it is used for
- why it is important
- what sets it apart

Literature Review

Deep Learning for Proteins - Historical Context & SOTA

Early DCA approaches

- mean-field DCA
- bmDCA
- plmDCA

Deep Generative models

- deep sequence
- ardca

Transformer-based models

- MSA Transformer evoformer
- Attention-Potts Model: factored self-attention -> potts model

Contextualization & Key advances

- arrange chronologically to trace evolution from statistical physics to deep interpretable neural approaches
- highlight SOTA for different tasks

Preliminary Methods

Direct Coupling Analysis

arDCA - technical review

Implementation

Results and Conclusions