Contents



# 1. Introduction

## 1.1. 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.

## 2. Biological Background

### 2.0.1. Proteins and their structure

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 standard building blocks: the 20 canonical amino acids arranged in different

assortments. These amino acids are linked though peptide bonds, which is why proteins are also known as polypeptides.

Amino acids share 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. [+LINKING SENTENCE HIGHLIGHTING THE FACT THAT DIFFERENT AAs PRODUCE DIFFERENT PURPOSE PROTEINS- atm unclear why amino acid chemical properties are important]. Generally the sequences are made up of between 50 and 2000 amino acids. The ordering of the amino acids dictates how a protein folds into its three-dimensional structure, known as its conformation. Although each conformation is unique, two common folding patterns occur in many proteins: the $\alpha$ helix and the $\beta$ sheet.

Protein structure is broken down into four levels of organization. The amino acid sequence is known as the primary structure. The regularly repeating local structures stabilized by chemical bonds, such as helices and sheets, are referred to as the secondary structure. The overall three-dimensional shape of a single protein molecule constitutes the tertiary structure. Finally, if a protein molecule is formed as a complex of more than one polypeptide chain, the complete structure is the quaternary structure [1].

Each amino acid is chemically distinct and can occur at any position in a protein chain, giving rise to $20^n$ possible polypeptide sequences of length $n$. For a typical protein of 300 amino acids, the number of possible sequences is astronomically large. However, only a small fraction of these sequences are capable of folding into a stable three-dimensional conformation. Natural selection has enabled living organisms to explore sequence space, favoring those sequences that reliably fold into stable structures.

### 2.0.2. Protein families and evolutionary information

Proteins do not evolve in isolation; instead, they belong to protein families, groups of proteins that share a common evolutionary origin and therefore exhibit related structures and functions [2].

Define: proteins with shared evolutionary origin, similar structure/function.

Point out that protein sequences vary across species but retain conserved features critical for function.

Explain why families are important: by comparing homologous sequences across organisms, we can extract information not obvious from a single sequence.

This sets the stage for Multiple Sequence Alignments.

### 2.0.3. 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.

Protein families [3]

Here define the LENGTH L, NUMBER M of sequences, — this will introduce the importance of picking the correct generative model in the next section

### 2.0.4. Protein Residues and Contacts

One of the earliest challenges for researchers in computational biology, encouraged by the CASP (Critical Assessment of Structure Prediction) competition, has been

understanding how a linear sequence of amino acids folds into a three-dimensional protein structure. Since a protein's 3D shape largely determines its biological function, predicting structure from sequence is a fundamental goal. Two important concepts in this context are residues and contacts:

- **Residues**: An individual amino acid within a protein sequence. During peptide bond formations, the chemically linked amino acids usually lose certain atoms, and what remains is referred to as the "residue" of the original amino acid. In structural biology, "residue" also refers to a specific position in a protein sequence.
- **Contacts**: A pair of residues that are spatially close in a protein's folded three-dimensional structure. These structures can be represented in Cartesian coordinates $(x, y, z)$, hence the contacts are defined using distance thresholds. Two residues are generally considered to be in contact if the distance between selected atoms is below a set threshold, usually 8 Ångströms (Å). Residues that are far apart in the linear sequence, may be close together in the 3D structure. may be far apart in [3]

important to add: there is short, mid and long range + most models evalute long range separately- it is the most important and hardest to predicting

This leads us to:

#### 2.0.5. The problem of inference

computational biology challenge: given MSA, oobserve correlations between residues (MI, co-variation) but correlations can be direct vs. indirect

lead into DCA

AlphaFold / Boltz-2 … will be used for structural integrity check.

## 3. Mathematical Foundations

### 3.1. Proteins as Statistical Systems

### 3.2. Direct Coupling Analysis

## 4. Evolution of DCA Methods

### 4.1. Mean-field DCA (mfDCA)

- Simplified inference method; first successful applications to contact prediction
- Strengths/weaknesses

### 4.2. Pseudo-likelihood Maximization DCA (plmDCA)

- More accurate & scalable inference
- Widely used in practice

### 4.3. Boltzmann Machine DCA (bmDCA)

- Boltzmann learning to directly fit the Potts model
- Shown to generate functional protein variants (e.g., chorismate mutase)

### 4.4. Deep Generative and Hybrid Models

- DeepSequence (variational autoencoder on MSAs)
- arDCA (autoregressive formulation, avoids MCMC)

ArDCA is an autoregressive network built on the basis of the Direct Coupling Analyis framework defined by Trinquier et. al. It will serve as the basis of this exploration. ArDCA was built to explore the ability of generative models for protein design coming from sequence-data. ArDCA emerges as a capable model for the extraction of structural and functional protein information which encoded in rapidbly growing protein databases. The paper compares the generative model's performance to existing solutions involv-

ing Boltzmann machines and deep generative models. It finds that this lightweight approach not only performs at a similar accuracy, but at a substantially lower computational cost (a factor between $10^2$ and $10^3$). It presents an important innovation also due to its mathematical advantages, which will be explored further, leading to improved applicability in sequence generation and evaluation.

- Attention-based DCA (bridging DCA and transformer attention)

### 4.5. Transformer-based models (could be combined above^)

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

## 5. Implementation and Extension

- Julia -> Python re-implementation (vectorization, frameworks, missing functions)
- Benchmarks

(MAYBE) Improvements:

- Allowing arbitrary sequence length (GPT-style transformers)
- Incorporating attention mechanism (Potts with attention)

Evaluation with Boltz-2 / AlphaFold3

## 6. Results and Discussion

## Bibliography

[1] "The Shape and Structure of Proteins," in *Molecular Biology of the Cell*, 4th ed., New York: Garland Science, 2002. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK26830/

[2] European Bioinformatics Institute (EMBL-EBI), "What are protein families?." [Online]. Available: https://www.ebi.ac.uk/training/online/courses/protein-classification-intro-ebi-resources/protein-classification/what-are-protein-families/

[3] B. Adhikari and J. Cheng, “Protein Residue Contacts and Prediction Methods,” *Methods in Molecular Biology*, vol. 1415, pp. 463–476, 2016, doi: 10.1007/978-1-4939-3572-7_27.