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

2.1. 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.

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. The distinct amino acids are bonded together through peptide bonds to form proteins, also known as polypeptides. Generally, protein 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. A protein's conformation helps it in completing its task, therefore defines its function.Although each conformation is unique, two common folding patterns occur in many proteins: the α helix and the β 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 the previously mentioned helices and sheets, make up 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 known as 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.

2.2. Protein families and evolutionary information

Proteins do not evolve in isolation; they often belong to protein families, groups of proteins that share a common evolutionary origin, therefore exhibiting related sequence features and functional properties [2]. Throughout years of evolution, mutations accumulate in these families: some are beneficial, altering the protein activity in ways that give rise to new functions, while many others are neutral and have no effect on stability or activity. Harmful changes, by contrast, disrupt folding or function and are eliminated by natural selection. The result is a collection of homologous proteins that retain overall structural and functional characteristics, but also display sequence variability that encodes the evolutionary history of the family [1].

The study of evolutionary history and relationships among biological entities is referred to as phylogenetics. Potein families provide the domain for phylogenetic analysis, as examining families provides insight that cannot be obtained from a single sequence. Analyzing homologous proteins across diverse organisms allows the detection of correlated mutations between amino acid positions, which in turn represent structural or functional constraints enforced by evolution. These statistical patterns are exploited by computational approaches to predict three-dimensional structure and understand protein function. [3]

2.3. Multiple sequence alignments

To extract important evolutionary clues from protein families, the homologous sequences need to be organized in a systematic way. This is done through a multiple sequence alignment (MSA), in which three or more sequences are arranged so that homologous sites are placed in the same column [4]. To maximise the positional

correspondence of sequences with varied length, alignment gaps are introduced when necessary.

Multiple sequence alignments reveal patterns of conservation and variation across the family. Conserved positions, those which are unchanged in multiple sequences, represent sites that are critical for mainting structure or function, while variable positions indicate sites that can tolerate mutations without disruption to the protein. Beyond conservation, MSAs also capture covariation: pairs of positions that mutate in a correlated way across sequences. These covariation signals reflect couplings, where a mutation at one site requires compensation at another to maintain protein integrity. [ADD SOURCE-wikipedia]

2.4. Protein residues and contacts

[[POSSIBLY REMOVE One of the earliest challenges for researchers in computational biology–encouraged also by the CASP (Critical Assessment of Structure Prediction) competition–has been to understand how the linear sequence of amino acids folds into its conformation.]]

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 [5].

Contacts are further broken down into short, medium, and long range predictions. Most computational approaches evaluate the long range contacts separately, as they are the most important for accurate predictions and unsurprisingly, the hardest to predict [5].

2.5. The problem of inference

The specific challenge in computational biology that this paper explores is the prediction, and generation, of protein sequences given a multiple sequence alignment. As previously mentioned the protein families intrinsicly contain correlations between residues which allows the model to build a predictive thing. The inherent challenge of these models is to distinguish between direct and indirect correlations.

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

3. Mathematical Foundations

3.1. Proteins as statistical systems

By analyzing statistical distributions of amino acids present in MSAs

- sequence variation = samples from prob dist over sequences
- statistical mechanics: Hamiltonian descibes high-dim dist

define the Hamiltonian used in arDCA with field and couplings

- statmech: define interactions -> derive properties
- protein inference: inverse! observe samples and reconstruct underlying interaction parameters

3.2. Direct Coupling Analysis

- how DCA solves this problem

4. Evolution of DCA Methods

4.1. Mean-field DCA

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

4.2. Pseudo-likelihood Maximization DCA

- More accurate & scalable inference
- Widely used in practice

4.3. Boltzmann Machine DCA

- 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 involving 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

5.1. Implementation details

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

5.2. Computational challenges

- Benchmarks

5.3. (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] F. Morcos *et al.*, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences*, vol. 108, no. 49, pp. E1293–E1301, 2011, doi: 10.1073/pnas.1111471108.

[4] M. Wiltgen, "Algorithms for Structure Comparison and Analysis: Homology Modelling of Proteins," *Encyclopedia of Bioinformatics and Computational Biology*. Academic Press, pp. 38–61, 2019. doi: https://doi.org/10.1016/B978-0-12-809633-8.20484-6.

[5] 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.