An Extended Ensemble Approach for Protein Sequence Variation

Hoda Akl[1], Brooke Emison[1], Xiaochuan Zhao[1] and Purushottam Dixit[123]

***Abstract*— Proteins are subject to various biophysical and evolutionary constraints limiting the possible variability in their sequences. We explore the space of putatively functional proteins through a generative model for protein sequences based on extended ensembles in statistical physics. We infer a latent space with a tunable dimension allowing us to generate de-novo sequences while preserving higher order statistics of the protein family. Computational structural assessment shows that the generated sequences are likely to be thermodynamically stable. Notably, our approach enhances accuracy with faster computation time than the Potts model, and more robustness to training set size than variational auto-encoders.**

***Index Terms*— generative models for protein sequences, representation learning**

I. Introduction

Quantifying the constraints on amino acid variation in protein sequences within a protein family is crucial to our understanding of evolutionary and biophysical forces that dictate protein structure and function. Since many factors simultaneously constrain sequence variation (thermodynamic stability, enzymatic activity, interaction partners etc.), bottom-up mechanistic models for sequence variation are infeasible. Data-driven approaches such as maximum entropy (Max Ent) methods [1] and machine learning techniques [2] have leveraged the advancement in sequencing and the availability of multiple sequence alignments for modeling of protein sequences. However, such approaches face conceptual and practical limitations. Max Ent requires the modeler to identify appropriate constraints which are often chosen to be pairwise couplings. As a result, it is computationally prohibitive to fit Max Ent models for large proteins. Finally, Max Ent modeling assumes statistical independence among phylogenetically related sequences. Variational auto-encoders can potentially model large proteins, however the quality of their representation is heavily dependent on the multiple sequence alignment size, thus facing difficulty in accurately representing proteins that exist in a small number of organisms, for example, mammalian proteins.

To address these difficulties, we present a probabilistic generative model based on the extended ensemble approach. Here, we model protein sequences as arising from their own Gibbs-Boltzmann distribution with temperature-like intensive parameters that are unique to individual sequences and energies that are shared across all sequences. The temperatures and the energies are learnt from the multiple sequence alignment using likelihood maximization. The intensive parameters embed individual sequences in a latent space of tunable dimensionality. Our method can assign probabilities to arbitrary sequences thereby allowing us to generate ensembles of de novo sequences.

This work was funded by NIH grant R35GM142547.

[1]Department of Physics, University of Florida, Gainesville, FL, USA.
[2]Genetics Institute, University of Florida, Gainesville, FL, USA.
[3]Department of Chemical Engineering, University of Florida, Gainesville, FL, USA.

II. Results

Using several protein families, we show that the sequences generated using our approach are novel and the ensemble of generated sequences preserves several higher order correlations across positions on the protein sequence. Using the enzyme dihydrofolate reductase (DHFR) from E. coli and computational structural assessment, we show that the generated sequences are likely to be thermodynamically stable. Moreover, the probabilities assigned by our approach to point mutants correlates strongly with the experimentally estimated fitness of those mutants. Notably, our approach has several practical advantages over several state-of-the-art methods. While our model shows higher accuracy, it is faster than the Potts model [3] and requires less training data than variational auto-encoders [4]. Notably, our framework is applicable to all types of categorical data including nucleotide sequences and binary data [5] such as presence/absence of genes in genomes, neuronal spikes, etc.

References

[1] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, "Inverse statistical physics of protein sequences: a key issues review," vol. 81, no. 3, p. 032601. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6633/aa9965

[2] Z. Wu, K. E. Johnston, F. H. Arnold, and K. K. Yang, "Protein sequence design with deep generative models," vol. 65, pp. 18–27. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S136759312100051X

[3] A. P. Muntoni, A. Pagnani, M. Weigt, and F. Zamponi, "adabmDCA: adaptive boltzmann machine learning for biological sequences," vol. 22, no. 1, p. 528. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04441-9

[4] A. Hawkins-Hooker, F. Depardieu, S. Baur, G. Couairon, A. Chen, and D. Bikard, "Generating functional protein variants with variational autoencoders," vol. 17, no. 2, p. e1008736. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1008736

[5] X. Zhao, G. Plata, and P. D. Dixit, "SiGMoiD: A super-statistical generative model for binary data," vol. 17, no. 8, p. e1009275. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1009275