Direct-Coupling analysis of protein structure using Expectation Reflection

Evan Cresswell-Clay[1], Vipul Periwal[1]

***Abstract*— Sequence covariation in multiple sequence alignments of homologous proteins has been used extensively to provide insights into protein structure predictions (PSP). However, global statistical inference is required in order to acquire direct relationships between amino acid positions in these sequences that are not simply correlations which could arise as a result of third party influence. Methods for statistical inference of such sequence covariation have been developed to exploit the growing availability of sequence data to provide hints about the folded protein structure which provides critical *a priori* information for more advanced PSP such as AlphaFold2. We present a novel method for protein structure inference using an iterative parameter-free model estimator which uses the formalism of statistical physics. With no tunable learning rate, our method scales to large system sizes while providing improved performance in the regime of small sample sizes. We apply this method to a more than 6500 PDB structures and compare its performance to that of other methods.**

***Index Terms*— Protein Contact Prediction, Data-driven modeling, Proteomics**

I. Introduction

Amino acids that are spatially proximal in the tertiary protein structure may have co-evolved due to functional requirements. Therefore phylogenetic analysis of proteins may provide insights into the functionally important contacts in the protein's network of interacting residues. However, global statistical inference is required in order to infer direct relationships between amino acid positions in these sequences that are not simply correlations that could arise as a result of third party influence. Direct Coupling Analysis (DCA) and pseudo-likelihood variants (PLM) have been developed over the past decade to exploit the growing availability of sequence data to obtain insights into native protein structures, potentially to gain intuition for proteins without known crystal structures. Current neural network PSP systems rely on DCA output as input, either as sumarized pairwize potentials or raw direct-information objects [1]. The major problem in all such inference is the exponential growth of sequence state space which cannot be overcome with the present growth of sequencing data, even setting aside computational limitations.

[1]Laboratory of Biological Modeling, NIDDK

This work was supported by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases.

This work utilized the computational resources of the NIH HPC Biowulf cluster (`http://hpc.nih.gov`)

II. Expectation Reflection for Protein Structure Contact Prediciton

We present a novel method for statistical inference of such sequence covariation to infer the network of interacting residues in proteins. The co-evolution of sequence position pairs of homologous proteins has been used to obtain a measure of interaction strength to great effect in previous work [2], [3], [4].These methods defined a Direct Information (DI) between all position pairs using the computed interaction matrix. Using the formalism of statistical physics, we define a formal free energy of observations from a partition function with an energy function chosen precisely to enable an multiplicative model update that avoids the pitfalls of cost minimization. Our method has no tunable learning rate, scales to large system sizes, and is particularly suited to sparse high-dimensional datasets to infer networks of such interacting residues [5], [6].

A. Results

We analyze 6500 PDB structures and demonstrate that our method outperforms DCA and PLM methods across all ranges of protein family size and sequence length. When comparing the true-positive versus false-positive rates of contact prediction of our method and the best DCA variant, our method provided the most accurate contact prediction for a significant majority of structures. We also demonstrate the ability of our method to provide insight to metamorphic protein structures.

References

[1] M. AlQuraishi, "Machine learning in protein structure prediction," *Current opinion in chemical biology*, vol. 65, pp. 1–8, 2021.

[2] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, "Identification of direct residue contacts in protein–protein interaction by message passing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 1, pp. 67–72, 2009.

[3] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "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.

[4] M. Ekeberg, T. Hartonen, and E. Aurell, "Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences," *Journal of Computational Physics*, vol. 276, pp. 341–356, 2014.

[5] D.-T. Hoang, J. Jo, and V. Periwal, "Data-driven inference of hidden nodes in networks," *Physical Review E*, vol. 99, no. 4, p. 042114, 2019.

[6] D.-T. Hoang, J. Song, V. Periwal, and J. Jo, "Network inference in stochastic systems from neurons to currencies: Improved performance at small sample size," *Physical Review E*, vol. 99, no. 2, p. 023311, 2019.