Combining rational protein engineering and deep representation learning

William Johnson
PhD in Molecular Biology, University of Alberta, Canada

Abstract (600 word limit):

In order to rationally engineer proteins, it is necessary to understand protein function holistically. We apply deep learning to unlabeled amino acid sequences to distill the fundamental properties of proteins into a structurally, evolutionarily, and biophysically grounded statistical representation. Based on this unified representation (UniRep), we show that the simplest models are broadly applicable and generalize to regions of sequence space previously unexplored. Our data-driven approach predicts the stability of natural and de novo designed proteins, as well as the quantitative function of molecularly diverse mutants, competitively with state-of-the-art methods. UniRep also increases the efficiency of protein engineering tasks by two orders of magnitude. UniRep provides a versatile summary of fundamental protein features that can be used across the protein engineering informatics domain. Protein engineering has the potential to transform synthetic biology, medicine, and nanotechnology. A traditional approach to protein engineering relies on random variation and screening/selection without modeling the relationship between sequence and function. Although proteins share a number of engineering-relevant properties, there is a smaller group of fundamental properties that make them work. An existing quantitative protein modeling approach aims to approximate one or a small subset of these. One example is biophysical modeling, which consists of structural approaches. Deep learning is a machine learning paradigm for learning data representations from raw input data. Recently, this flexibility has been demonstrated in the prediction of protein structure by replacing complicated informatics pipelines with models that predict structure directly from sequence. Raw sequence data for proteins are widely available. As the number of these sequences increases, so does their variety. There has been no evaluation of either of these approaches on a comprehensive collection of protein informatics problems, nor have they been applied to learning representations at scale. Using recurrent neural networks (RNN), we construct statistical representations of proteins based on 24 million UniRef50 sequences. Multiplication-based long-/short-term memory (mLSTM) RNNs develop rich representations of natural language to achieve state-of-the-art performance on critical tasks. As a measure of how semantically related proteins are represented, we evaluated UniRep's ability to partition structurally similar sequences that share limited sequence identity and enable unsupervised clustering of homologous sequences. Protein stability is a fundamental factor in their function, as well as an engineering endpoint affecting yield, reaction rate, and shelf life of protein catalysts, sensors, and therapeutics. Next, we evaluated UniRep's ability to predict the stability of a large collection of de novo designed mini proteins. This result was unexpected, since de novo designed proteins represent a small proportion of the UniRep training.

About Research Topic (200 word limit)

Building models that generalize from local data to distant regions of sequence space where more functional variants are present is a core challenge of rational protein engineering. In general, deep-learning models cannot generalize beyond their domain of training. The reason we hypothesized UniRep captures general features of protein fitness landscapes extends beyond task-specific training data is that UniRep is trained in an unsupervised manner on a wide variety of proteins and is relatively compact. Known characteristics of proteins are encapsulated in UniRep features . Since UniRep is derived from raw data, it is not constrained by existing mental models for understanding proteins, so it may also approximate yet unknown features, enabling protein engineering prediction

tasks beyond those examined here. In contrast to other methods, UniRep does not require experimentally determining or computationally folding a structural intermediate. UniRep may improve protein engineering workflows or, at best, help discover sequence variants not accessible by experiments or structural approaches while enabling rapid generalization to distant, unseen regions of the fitness landscape. Additionally, UniRep offers several natural extensions. Similarly to previous work with proteins and small molecules, it can already be used generatively for deep protein design. As well as in engineering, our results suggest UniRep distance may aid vector-parallelized semantic protein comparisons, even without any training data, regardless of evolutionary depth. Using sequence likelihood-based scoring, we also envision future work on data-free UniRep variant-effect prediction.

Biography (200 word limit)

William Johnson with a PhD in Molecular Biology. My experience includes managing projects, conducting research, and teaching. Molecular genetics has been a focus of my expertise. In addition, I have contributed to the development of biotechnology programs in both public and private companies. Currently, I am working at the University of Alberta, Canada. I am studying the role of the vagus nerve in the development and treatment of dietary obesity.

About Institution (200 word limit)

Washington University in St. Louis (WashU, or WUSTL) is a private research university in Greater St. Louis with its main campus (Danforth) mostly in unincorporated St. Louis County, Missouri, and Clayton, Missouri. It also has a West Campus in Clayton, North Campus in the West End neighborhood of St. Louis, and Medical Campus in the Central West End neighborhood of St. Louis. Founded in 1853 and named after George Washington, the university has students and faculty from all 50 U.S. states and more than 120 countries. Washington University is composed of seven graduate and undergraduate schools that encompass a broad range of academic fields. To prevent confusion over its location, the Board of Trustees added the phrase "in St. Louis" in 1976. Washington University is a member of the Association of American Universities and is classified among "R1: Doctoral Universities – Very high research activity".

References (15-20)

1. Quraishi M ProteinNet: a standardized data set for machine learning of protein structure. BMC Bioinform. 20, 311 (2019).
2. Robertson S Understanding inverse document frequency: on theoretical arguments for IDF. J. Documentation 60, 503–520 (2004).
3. Park H et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. J. Chem. Theory Comput 12, 6201–6212 (2016).
4. Alford RF et al. The rosetta all-atom energy function for macromolecular modeling and design. J. Chem. Theory Comput 13, 3031–3048 (2017).
5. Glorot X, Bordes A & Bengio Y Domain adaptation for large-scale sentiment classification: a deep learning approach. In Proc. 28th International Conference on International Conference on Machine Learning 513–520.

6. Håndstad T, Hestnes AJH & Sætrom P Motif kernel generated by genetic programming improves remote homology and fold detection. BMC Bioinform. 8, 23 (2007).
7. Li S, Chen J & Liu B Protein remote homology detection based on bidirectional long short-term memory. BMC Bioinform. 18, 443 (2017).
8. Lovato P, Cristani M & Bicego M Soft Ngram representation and modeling for protein remote homology detection. IEEE/ACM Trans. Comput. Biol. Bioinform 14, 1482–1488 (2017).
9. Pedregosa F et al. Scikit-learn: machine learning in python. J. Mach. Learn. Res 12, 2825–2830 (2011).
10. Jones E, Oliphant T & Peterson P SciPy: Open source scientific tools for Python (SciPy, 2001)