Unified rational protein engineering with sequence-based deep representation learning

Ethan Alley
Grigory Khimulya
https://orcid.org/0000-0002-4184-9890
Surojit Biswas
Mohammed AlQuraishi
George M. Church
https://orcid.org/0000-0003-3535-2076

Method Article

Keywords: machine learning, protein engineering, deep learning

Posted Date: November 20th, 2019

DOI: https://doi.org/10.21203/rs.2.13774/v1

Version of Record: A version of this preprint was published at Nature Methods on October 21st, 2019. See the published version at https://doi.org/10.1038/s41592-019-0598-1.

Abstract

This protocol describes the computational steps necessary to reproduce the results described in the paper "**Unified rational protein engineering with sequence-only deep representation learning**" by Alley et al.

Introduction

Rational protein engineering requires a holistic understanding of protein function. Here, we apply deep learning to unlabelled amino acid sequences to distill the fundamental features of a protein into a statistical *representation* that is semantically rich and structurally, evolutionarily, and biophysically grounded. We show that the simplest models built on top of this unified representation (UniRep) are broadly applicable and generalize to unseen regions of sequence space. Our data-driven approach reaches near state-of-the-art or superior performance predicting stability of natural and *de novo* designed proteins as well as quantitative function of molecularly diverse mutants. UniRep further enables two orders of magnitude cost savings in a protein engineering task. Here we provide a protocol for reproducing these results.

Reagents

No reagents necessary

Equipment

Preferably, m5.12xlarge or m5.24xlarge AWS instance with Ubuntu Server 18.04 LTS AMI (for example, ami-0f65671a86f061fcd).

Code and dependencies described under "Requirements" in https://github.com/churchlab/UniRep-analysis

Procedure

1. Clone the repository containing the code with

$$gitclo \neq hps: / github. co\frac{m}{c}hurchla\frac{b}{U}niRep - analysis. git$$

2. Download and unzip the data using bash commands under "Getting the data" in the repository README

3. Reproduce figures and retrain top models by running ipython notebooks and python scripts as described under "Usage" in the repository README

Loading [MathJax]/jax/output/CommonHTML/jax.js

Troubleshooting

1. Check that the requirements are in place (see "Requirements" in the repository README)

2. Make sure the path to data folder is correct and accessible

3. Reach out for assistance

Time Taken

<1 hour for regenerating figures

~7 hours for retraining top models and recomputing metrics

Anticipated Results

Detailed description of results available here: https://www.biorxiv.org/content/10.1101/589333v1

References

Pre-print of Alley et al. "**Unified rational protein engineering with sequence-only deep representation learning" available at** https://www.biorxiv.org/content/10.1101/589333v1

Acknowledgements