SSProteinFitnessPrediction:
Data Documentation

Alicia Olivares-Gil, José A. Barbero-Aparicio, Juan J. Rodríguez,
José F. Díez-Pastor, César García-Osorio, Mehdi D. Davari

Contents

1 Introduction

This repository contains a compilation of 19 datasets of protein-fitness pairs containing single-substituted variants (17 datasets), double-substituted variants (1 dataset) and multiple-substituted variants (1 dataset). All the datasets and their characteristics are listed in Table 1. The generation and processing of the files contained in this repository has been carried out using the scripts contained in this GitHub repository.

2 Datasets generation

In addition to the files containing the sequence-fitness pairs and the wild type of the protein, this repository contains the files generated by applying the following pre-processing steps:

1. A search for homologous sequences in the Uniref100 database using the HMMER software.

```
jackhmmer --incT <0.5 * wild type length> --cpu 64 --nali -A <output .sto file path>
<input .fasta wild type file path> <Uniref100 path>
```

2. Inference of a statistical coupling model from the homologous sequences found using the PLMC software.

```
bin/plmc -o <output .params file path> -n 64 -le <0.2 * active sites> -m 3500 -g
-f <identifier> <input .a2m file path>
```

3. Encoding of the sequences in the sequence-fitness pairs and the homologous sequences using the Unirep [1], eUnirep [3], PAM250 [4] and DCA [8] encodings.

4. Serialisation of the encoded sequence sets using Python's Pickle package.

Dataset	Abbreviation	Sequence length	#variants	#homologous sequences	Substitutions
YAP1_HUMAN_Fields2021 [2]	yap1_human	34	313	248 330	single
UBE4B_MOUSE_Klevit2013 [19]	ube4b_mouse	102	518	97 159	single
GAL4_YEAST_Shendure2015 [10]	gal4_yeast	64	803	171 882	single
BLAT_ECOLX_Tenaillon2013 [9]	blat_ecolx_4	286	975	30 356	single
PABP_YEAST_Fields2013 [11]	pabp_yeast_1	75	1 142	948 246	single
RL401_YEAST_Bolon2013 [17]	rl401_yeast_1	75	1 154	96 590	single
BRCA1_HUMAN_Fields2015_y2h [20]	brca1_human_2	303	1 278	1 909	single
RL401_YEAST_Bolon2014 [16]	rl401_yeast_2	75	1 282	96 590	single
MTH3_HAESTABILIZED_Tawfik2015 [14]	mth3_haeaest.	329	1 611	75 260	single
POLG_HCVJF_Sun2014 [13]	polg_hcvif	86	1 613	17 184	single
BG_STRSQ_Abate2015 [15]	bg_strsq	478	2 598	124 487	single
BRCA1_HUMAN_Fields2015_e3 [20]	brca1_human_1	303	2 846	1 909	single
HSP82_YEAST_Bolon2016 [12]	hsp82_yeast	230	4 065	59 490	single
BLAT_ECOLX_Ostermeier2014 [7]	blat_ecolx_1	286	4 799	30 356	single
BLAT_ECOLX_Ranganathan2015 [21]	blat_ecolx_3	286	4 921	30 356	single
BLAT_ECOLX_Palzkill2012 [5]	blat_ecolx_2	286	4 922	30 356	single
HG_FLU_Bloom2016 [6]	hg_flu	564	10 337	71 029	single
PABP_YEAST_Fields2013 [11]	pabp_yeast_2	75	33 771	948 246	double
avGPF_Kondrashov2016 [18]	avgfp	235	32 610	697	multiple

Table 1: Columns represent (1) the name and reference of the dataset, (2) the abbreviation used in the file names, (3) the number of amino acids in the protein sequence, (4) the number of sequences present in the dataset, (5) the number of homologous sequences found using a jackhmmer search in UniRep100 database and (6) the number of substitutions present in each variant.

3 Files description

This repository contains two types of .zip files:

- Files starting with **raw_data_** contain:
  - `<dataset>.fasta`: The wild type sequence in fasta format.
  - `<dataset>.csv`: A csv file containing all the sequence-fitness pairs. The first column contains the sequences represented as the point mutation(s) made on the wild type (For example: D101E means that the 'D' in the 101-th position of the wild type has been replaced by an 'E'). The second column contains the numerical value of that sequence's fitness.
  - `<dataset>_jhmmer.sto`: The Stockholm alignment file generated when performing a jackhmmer search using the HMMER software for searching homologous sequences of the wild type in the UniRef100 database.
  - `<dataset>_jhmmer.a2m`: The same alignment file but in A2M format.
  - `<dataset>_plmc.params`: A binary file containing all inferred model parameters using the PLMC software taking the A2M file as an input.
- The rest of the .zip files contain Pickle files with the datasets as Numpy arrays in various encodings.
  - `<dataset>_wt.pk`: The wild type (list of amino-acids).
  - `<dataset>_variants.pk`: The list of substitutions as they appear in `<dataset>.csv`.
  - `<dataset>_sequences.pk`: The list of sequences (list of amino-acids). resulting from making the corresponding substitutions on the wild type.
  - `<dataset>_y.pk`: The fitness values corresponding to the sequences in `<dataset>_sequences.pk`.
  - `<dataset>_unambiguous_homologs.pk`: The list of homologous sequences (list of amino-acids) found performing a jackhmmer search in the UniRef100 database. Ambiguous symbols such as 'X', 'B', 'J' and 'Z' have been randomly replace by one of their appropriate amino-acids.
  - `<dataset>_Xl_unirep.pk`: The list of sequences in Unirep encoding.
  - `<dataset>_Xl_eunirep.pk`: The list of sequences in eUnirep encoding.
  - `<dataset>_Xl_pam250.pk`: The list of sequences in PAM250 encoding.
  - `<dataset>_Xu_pam250.pk`: The list of unambiguous homologous sequences in PAM250 encoding.
  - `<dataset>_Xl_dcae.pk`: The list of sequences in DCA encoding. Since the DCA encoding process purges some sequences, the number of sequences present in this file is lower than the others.
  - `<dataset>_indexes.pk`: The indexes needed to match the sequences in other encodings with the sequences in DCA encoding. For example, `<dataset>_Xl_unirep[indexes]` returns the corresponding sequences in the DCA file but in Unirep encoding.
  - `<dataset>_Xu_dcae.pk`: The list of unambiguous homologous sequences in DCA encoding.
  - `<dataset>_y_dcae.pk`: The fitness values corresponding to the sequences in DCA encoding.

4 Requirements

Depending on the steps to be reproduced, the following requirements need to be fulfilled:

- In order to perform the jackhmmer search, the HMMER software must be installed on your work environment and the Uniref100 database in fasta format needs to be downloaded in your file system.
- In order to infer the coupling statistical model with the PLMC software package you must first download and compile it in your computer. The sequences alignment file used as input must be in A2M format, so the Stockholm alignment file generated in the previous step must be converted.
- To read the `<dataset>_plmc.params` files you need to use the script `scripts/read_params.m` in the PLMC software package.
- The pickle files have been generated using the following package versions:
  - python 3.10.12
  - numpy 1.25.0
  - pickle 4.0

References

[1] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

[2] Carlos L Araya, Douglas M Fowler, Wentao Chen, Ike Muniez, Jeffery W Kelly, and Stanley Fields. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences*, 109(42):16858–16863, 2012.

[3] Surojit Biswas, Grigory Khimulya, Ethan C Alley, Kevin M Esvelt, and George M Church. Low-N protein engineering with data-efficient deep learning. *Nature methods*, 18(4):389–396, 2021.

[4] M Dayhoff, R Schwartz, and B Orcutt. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure*, volume 5, pages 345–352. National biomedical research foundation Silver Spring, MD, USA, 1978.

[5] Zhifeng Deng, Wanzhi Huang, Erol Bakkalbasi, Nicholas G Brown, Carolyn J Adamski, Kacie Rice, Donna Muzny, Richard A Gibbs, and Timothy Palzkill. Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *Journal of molecular biology*, 424(3-4):150–167, 2012.

[6] Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses*, 8(6):155, 2016.

[7] Elad Firnberg, Jason W Labonte, Jeffrey J Gray, and Marc Ostermeier. A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular biology and evolution*, 31(6):1581–1592, 2014.

[8] Alexander-Maurice Illig, Niklas E Siedhoff, Ulrich Schwaneberg, and Mehdi D Davari. A hybrid model combining evolutionary probability and machine learning leverages data-driven protein engineering, 2022. Preprint at https://www.biorxiv.org/content/early/2022/06/07/2022.06.07.495081.

[9] Hervé Jacquier, André Birgy, Hervé Le Nagard, Yves Mechulam, Emmanuelle Schmitt, Jérémy Glodt, Beatrice Bercot, Emmanuelle Petit, Julie Poulain, Guilène Barnaud, et al. Capturing the mutational landscape of the beta-lactamase tem-1. *Proceedings of the National Academy of Sciences*, 110(32):13067–13072, 2013.

[10] Jacob O Kitzman, Lea M Starita, Russell S Lo, Stanley Fields, and Jay Shendure. Massively parallel single-amino-acid mutagenesis. *Nature methods*, 12(3):203–206, 2015.

[11] Daniel Melamed, David L Young, Caitlin E Gamble, Christina R Miller, and Stanley Fields. Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly (a)-binding protein. *Rna*, 19(11):1537–1551, 2013.

[12] Parul Mishra, Julia M Flynn, Tyler N Starr, and Daniel NA Bolon. Systematic mutant analyses elucidate general and client-specific aspects of hsp90 function. *Cell reports*, 15(3):588–598, 2016.

[13] Hangfei Qi, C Anders Olson, Nicholas C Wu, Ruian Ke, Claude Loverdo, Virginia Chu, Shawna Truong, Roland Remenyi, Zugen Chen, Yushen Du, Sheng-Yao Su, Laith Q Al-Mawsawi, Ting-Ting Wu, Shu-Hua Chen, Chung-Yen Lin, Weidong Zhong, James O Lloyd-Smith, and Ren Sun. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis c viral fitness and drug sensitivity. *PLoS pathogens*, 10(4):e1004064, 2014.

[14] Liat Rockah-Shmuel, Ágnes Tóth-Petróczy, and Dan S Tawfik. Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLoS computational biology*, 11(8):e1004421, 2015.

[15] Philip A Romero, Tuan M Tran, and Adam R Abate. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*, 112(23):7159–7164, 2015.

[16] Benjamin P Roscoe and Daniel NA Bolon. Systematic exploration of ubiquitin sequence, e1 activation efficiency, and experimental fitness in yeast. *Journal of molecular biology*, 426(15):2854–2870, 2014.

[17] Benjamin P Roscoe, Kelly M Thayer, Konstantin B Zeldovich, David Fushman, and Daniel NA Bolon. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *Journal of molecular biology*, 425(8):1363–1377, 2013.

[18] Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

[19] Lea M Starita, Jonathan N Pruneda, Russell S Lo, Douglas M Fowler, Helen J Kim, Joseph B Hiatt, Jay Shendure, Peter S Brzovic, Stanley Fields, and Rachel E Klevit. Activity-enhancing mutations in an e3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*, 110(14):E1263–E1272, 2013.

[20] Lea M Starita, David L Young, Muhtadi Islam, Jacob O Kitzman, Justin Gullingsrud, Ronald J Hause, Douglas M Fowler, Jeffrey D Parvin, Jay Shendure, and Stanley Fields. Massively parallel functional analysis of brca1 ring domain variants. *Genetics*, 200(2):413–422, 2015.

[21] Michael A Stiffler, Doeke R Hekstra, and Rama Ranganathan. Evolvability as a function of purifying selection in tem-1 β-lactamase. *Cell*, 160(5):882–892, 2015.