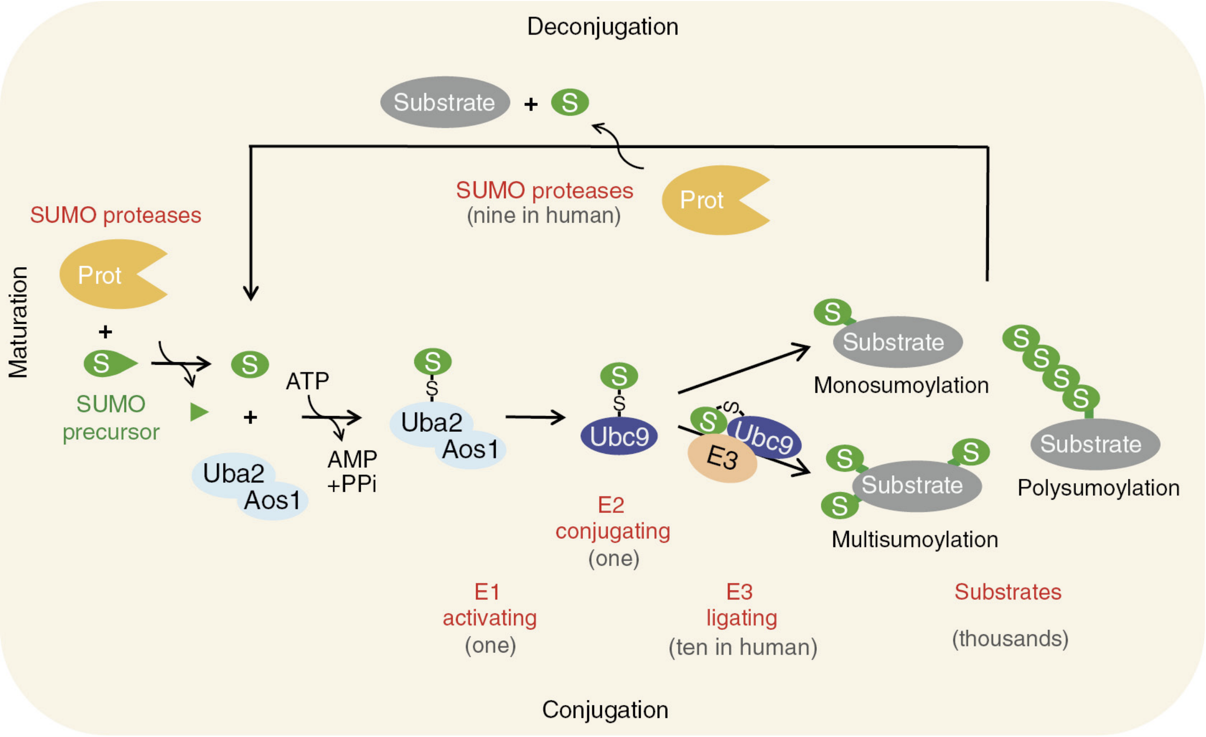
Intro

One of the major challenges in synthetic biology is to design better (more activite, stable, expressible) proteins with known desired function. Improved sequences could be used in industry to alleviate constraints in manufacturing processes. However, designing better sequences is not a trivial task. It is hard or even impossible to estimate the impact of a single amino acid mutation with currently available tools. Our goal is to develop a tool that is able to approximate these changes in property values (activity, fitness, stability) using data collected from the experimental work. The challenge is to predict protein property for a given sequence while using only 96 data points.

Ube2I

Ube2I, also known as small ubiquitin-related modifier (SUMO) E2 conjugase, plays an important role in the SUMO pathway modulating the amount of protein in eukaryotic cells. Mutations in Ube2I protein can result in abnormal enzyme function and disease-causing phenotype. With the help of E3 ligase, Ube2I transfers a SUMO signal molecule onto the target protein. The type of SUMO-ylation on the target protein decides its fate inside the cell - the intracellular location, protein stability and enzymatic activity. The fitness of Ube2I variants in the supplied dataset is described as the ability for the protein to partake in its natural functions, including the ability to form interactions with upstream functional proteins (Uba2, Aos1), downstream functional proteins (E3 ligase) and SUMO signal molecule and performing catalytic functions to transfer the SUMO signal molecule to the target protein.



In this figure Ube2I protein is named Ubc9. Source: https://www.degruyter.com/document/doi/10.1515/bmc-2016-0030/html

Files

- ube2i_single_point_mutants.fasta - contains 2854 sequences with experimental fitness values. Header contains a mutation and fitness value separated by ':'. Note that not all positions of protein have experimental data.
- features.csv - file that contains all features for the same 2854 sequences:
    - id - matches the first part of fasta header (mutation).
    - Volume - volume difference between wild type (original) amino acid and mutant amino acid.
    - MolWt - difference of molecular weight between wild type (original) amino acid and mutant amino acid.
    - NumHeavyAtoms - difference of number of heavy atoms (atomic number >1) between wild type (original) amino acid and mutant amino acid.
    - Hydropathy - difference of hydropathy between wild type (original) amino acid and mutant amino acid.
    - NumRotatableBonds - difference of number of rotatable bonds between wild type (original) amino acid and mutant amino acid.
    - FractionCSP3 - difference of fraction of C atoms that are SP3 hybridized between wild type (original) amino acid and mutant amino acid.
    - Acceptor - difference of number of hydrogen bond acceptors between wild type (original) amino acid and mutant amino acid.
    - Donor - difference of number of hydrogen bond donors between wild type (original) amino acid and mutant amino acid.
    - Aromatic - difference of aromaticity between wild type (original) amino acid and mutant amino acid.
    - Hydrophobe - difference of hydrophobicity value between wild type (original) amino acid and mutant amino acid.
    - LumpedHydrophobe - difference of lumped hydrophobe value between wild type (original) amino acid and mutant amino acid.
    - NetCharge - difference of net charge between wild type (original) amino acid and mutant amino acid.
    - Charge - total charge of residues within 10 angstroms of mutation.
    - NumPolarRes - number of polar residues within 10 angstroms of mutation.
    - NumApolarRes - number of apolar residues within 10 angstroms of mutation.
    - NumChargedRes - number of charged residues within 10 angstroms of mutation .
    - ProtRDF_all_2.0_3.0 - number of atom pairs that are 2-3 angstroms apart from each other. Only atoms within 10 angstroms of mutation are considered.
    - ProtRDF_all_3.0_4.0 - number of atom pairs that are 3-4 angstroms apart from each other. Only atoms within 10 angstroms of mutation are considered.
    - ProtRDF_all_4.0_5.0 - number of atom pairs that are 4-5 angstroms apart from each other. Only atoms within 10 angstroms of mutation are considered.
    - RosettaEnergy - total energy of protein (Rosetta energy units) calculated by Rosetta.
    - loss - loss value from trained protein language model. The bigger the loss, the worse sequence according to the model.

- ube2i_single_point_mutants_25_13_38_22_full_rep.npz - numeric representation from trained protein language model (internally built). Each sequence has a sequence_length x 512 matrix of float numbers.
- structures - this a directory of all modelled mutant structures. File name matches the first part of the fasta header (mutation).
- ube2i_msa.a3m - multi sequence alignment of ube2i homologous sequences in a3m format.
- ube2i_homologous.fasta - homologous sequences of ube2i sequences.

Objective

Using all information available, create a tool that is fitted/trained on 96 experimental values and is able to predict experimental values for unseen sequences. The performance of the tool will be measured using spearman correlation.

Papers

- Machine Learning-Assisted Directed Evolution Navigates a Combinatorial Epistatic Fitness Landscape with Minimal Screening Burden
- Low-N protein engineering with data-efficient deep learning
- Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences