Applying traditional machine learning and deep learning methods to the study of protein evolution

Essay 1

Neural networks to learn protein sequence-function relationships from deep mutational scanning data.

Background: Unsupervised learning methods such as EVmutation and DeepSequence train on large comparisons of evolutionarily relevant protein sequences. These methods can model the natural functions of protein families, but they cannot predict specific protein properties that are not constrained by long-term evolutionary selection. We newly propose a supervised deep learning framework that trains supervised neural networks to learn sequence-to-function mapping functions.

Dataset: consists of thousands to millions of protein sequence variants, each with a score used to quantify its activity or fitness in high-throughput assays. The dataset was divided into 81% training set, 9% tuning set (for optimization of hyper-references), and 10% validation set

Feature Engineering: We encode the protein sequence and the encoded information contains the physical and chemical properties of each amino acid at each position.One hot encoding is used to label the amino acid at a specific position and also AAindex encoding is used to capture the physical and chemical properties of the amino acid and PCA is applied to reduce it to 19 dimensions and finally the two encodings are connected. In Graph Convolutional Neural Network we get the residue distance matrix by calculating the 3D structure of the protein, thresholding the distances,and transforming the obtained matrix into an undirected graph to get the structure map.

The team tested a total of three neural networks: linear regression and fully connected, array convolution, and graph convolution. In a convolutional neural network, the convolutional layer learns to recognize common polar and nonpolar amino acids in the beta chain, allowing the network to assess the beta chain tendencies of the entire input sequence and correlate this information with protein function.

Applications:

1) Using visualization, visualize the connecting layers in the network as a 3D structure of the protein and color the protein according to its score, so as to understand which sequence positions have the greatest impact on the protein and to understand which regions are mutation-intolerant.

2 ) Designing new proteins based on wild-type proteins. The proteins with only localized changes relative to the wild type were evaluated using the trained model, and scored using a combination of four models, linear regression and three neural networks, where a random restart hill-climbing algorithm was used to maximize the minimum predictive function score.

Discussion:

Since mutations are often combined in an additive form, the linear model performs moderately well, but is unable to represent the interactions between mutations, whereas the convolutional network can generalize the effects of mutations between different locations to improve performance, and the graphical neural network does not outperform the sequence-based neural network, possibly due to a lack of data diversity, among other reasons.

Essay 2.

Learning the local landscape of protein structures with convolutional neural networks.

Adaptive mapping describes protein potentials as a function of conformational coordinates and is used to study evolution, identify proteins with new useful properties, or quantify variability. On the other hand, protein crystals are stable structures of proteins that are usually the lowest point in the energy terrain and dominate the population.

The basic AlphaFold implementation places protein sequences in the correct energy minima, but lacks the ability to apply deep learning to the opposite problem: how the structure restricts the proteins allowed for a given locus, which would limit the sampling of deleterious mutations and accelerate target mutagenesis and protein engineering efforts. We investigated this using 3DCNN, examining whether CNN models that predict hidden residues from local chemical environments (microenvironments) can predict residues well in wild-type residues and evolutionarily diverged homologs, and focusing on those amino acid distributions that are predicted with high accuracy.

Specific methods:

The network architecture has nine layers: six-layer feature extraction block: two pairs of 3D convolutional layers, each pair is followed by a reduced dimensionality maximum pooling layer, respectively, using 3\*3\*3 and 2\*2\*2 filters, using the Relu function applied to the output of each layer, the final dimensionality of the generated feature maps is 400\*3\*3\*3, and the spreading into the three-layer classification block. The classification block consists of three fully connected layers, the first two connected layers apply the relu function, and the third layer uses the softmax function to obtain a vector of probability scores for 20 amino acids.

Processing of the dataset:

We randomly sampled residues in the protein dataset to create a microenvironmental dataset that reflects the natural abundance of each amino acid. For each protein chain, we selected up to 50 residues and selected no more than 50% of the total number of residues to avoid biasing the model toward proteins with high molecular weights. The final microenvironment dataset consists of 1455978 microenvironments, split into 90% training set and 10% validation set.

Using these metadata, we generated a voxelized representation (4D tensor) of the microenvironment centered on the carbon of the α-specified residue and oriented with respect to the main chain such that the side chains are oriented along the +z axis. The voxelized representation has a resolution of 1 Å and consists of 3D space (x, y, z) and seven auxiliary channels. The auxiliary channels encode information about the nature of the atoms present in the voxel (C, H, O, N, S) as well as the partial charge and solvent accessible surface area.

A total of 5 epochs were trained in the CNN with an initial learning rate of 0.05,and the learning rate was halved if the validation accuracy did not increase by 0.1% for every 2000 batches.

Results:

Evaluated on the PSICOV dataset, an independent dataset containing 130 structures, the prediction accuracy of wild-type sequences was generally high, averaging 60%. After grouping biochemically similar amino acids, our ability to predict amino acid groupings was higher than our ability to predict specific amino acids, at 71%.

Next, we asked about the network's ability to predict the shared amino acids of a locus in a multiple sequence alignment (MSA). The degree of likelihood of this prediction depends on how well the microenvironment surrounding the given locus is conserved in the homologous structure. There is a 40% probability of predicting the correct shared amino acid, and a 55% probability of predicting the correct class ...

The percentage of correct amino acid predictions is highly dependent on sequence differences between homologs, and is less than 30% for proteins with large divergence from the wild type (less than 20% similarity).

Discussion:

Our results reinforce the idea that amino acids are closely linked to their local biochemical environment and that restrictions acting on amino acids in proteins tend to change over time as proteins evolve. The model's mispredictions may provide an opportunity for protein engineering, as it will point to sites that are ready for mutation. Specifically, we expect that whenever the network confidently predicts an amino acid that is different from the current wild type, that site will be ready for mutation, and that the specific incorrectly predicted amino acid is a good candidate for a stabilizing or function-acquiring mutation.

Essay 3.

Efficiently predicting protein stability changes upon single-point mutation with large language models.

This article describes in detail a new approach that utilizes Large Language Models and in particular ESM (Evolutionary Scale Modeling) models to efficiently predict changes in the stability of proteins after a single point mutation.

Background: Predicting changes in protein stability induced by single point mutations has been a challenge that has attracted the interest of many researchers for many years. Accurate prediction of protein thermal stability is crucial for applications in biochemistry such as drug development, protein evolution analysis and enzyme synthesis.

Assessment of protein thermal stability: assessed by ΔΔG = ΔG\_M - ΔG\_W (M: mutated; W: wild), higher ΔΔG means higher thermal stability

Machine learning-based predictions of protein thermal stability can be divided into two categories: (1) sequence-based methods, which extract effective features from protein sequences, and (2) structure-based methods, which extract structural variance to predict ΔΔG

Methods:

Dataset: Sequences and structures were carefully aligned, with the latter representing the coordinates of the backbone atoms within each amino acid. Our initial dataset was then derived from PROSTATA, containing wild-type and mutant sequences, and corresponding ΔΔG values, with a TM-score threshold of 0.5 (to prevent the appearance of similar proteins), for a total of 10,544 samples.

Model:

The model receives wild-type and mutant sequence-structure pairs as inputs. an ESM-assisted continuous graph convolutional network is applied independently to each input (i.e., wild-type and mutant proteins), which integrates the sequence and structural features into a concatenated low-dimensional feature vector. The two are then subtracted to capture feature changes caused by the presence of mutations, which are then converted to a one-dimensional output, i.e., ΔΔG values, by a nonlinear classifier to predict changes in the stability of the protein

Discussion:

The method has a higher accuracy while processing 15 times more than existing methods

The authors point to the need for further refinement of the sequence selection algorithm as well as more biochemical studies to improve the robustness and interoperability of the model.

Essay 4.

Rapid protein stability prediction using deep learning representations.

Presentation:

The Protein Language Model (PLM) learns the function of protein diversity by training to accomplish masked amino acids in a comprehensive protein sequence database, thereby generating salient biological features. However, due to evolutionary constraints and limited data limitations, PLM is unable to significantly improve protein activity in zero-sample learning environments and cannot be generalized to new environments.

Here we propose a new protein evolution model, EVOLVEpro, to evolve highly active protein variants through small-sample learning and minimal experimental testing, thus predicting highly active mutants from sequences only.

About EVOLVEpro's model development:

This is a deep learning based framework for directed evolution that involves:

1. A protein language model that encodes protein sequences into a potentially continuous space to facilitate optimization of their activity

2) A top-level regression model for studying the mapping between several data points from the latent space to protein activity.

We employ active learning using a regression model to rank protein sequences based on predicted compliance, from which the top-ranked sequences are selected for experimental validation . The loop is executed iteratively to evolve the identified protein activities.

Taking the ESM-2 protein language model as an example, our process is as follows:

A random set of first round variants were selected, and a random forest regression discriminant model was used to predict protein activity, using residue merge average embedding to select the top N selection strategies in each round of evolution. Finally, by capturing the residue site preferences of EVOLVEpro during evolution, it is possible to understand which positions are favorable for increasing activity

Application of the model:

Antibody optimization, Evolution of micro-DNA-guided CRISPR nuclease, Improved design of prime editor for Bxb1 integrase evolution, Evolution of T7RNA polymerase

Summary:

EVOLVEpro learns the general rules of protein activity and can generate highly active protein mutants in just a few evolutionary cycles.There is a powerful selection of functions in the potential space and top-level modules generated by PLM that represent proteins better than traditional coding methods such as OneHot coding and integer coding.

PLM learns masked sequence reconstruction tasks across evolutionary diversity, and the fitness terrain of its learning does not typically correlate with the fitness terrain of protein activity because it precludes typical directed evolution strategies, allowing EVOLVEpro to select highly active single mutants from tens of thousands of possible sequences and highly active proteins from hundreds of billions of multi-mutants.

EVOLVEpro has the following advantages:

1) High success rate

2) No special knowledge of proteins is required

3) Can be used for multi-objective optimization

4) Highly modular, allowing any quantifiable property to be used as input without fine-tuning.

Essay 5.

Structure-based self-supervised learning enables ultrafast protein stability prediction upon mutation.

Protein stability is highly susceptible to perturbations from the external environment and genetic mutations, and even small perturbations, such as single point mutations, can move active proteins towards non-functional, misfolded, or aggregated forms. We trained a self-supervised model, Pythia, for zero-sample prediction of free energy changes after protein mutations.

Model building:

Under the assumption that the energy of the protein in its unfolded state is largely unaffected by mutation, the derivation of the formula allows us to obtain a positive relationship between the free energy change and the sum of the probabilities that the amino acids of the wild-type and mutant proteins are in all Rotamer conformations:

\Delta\Delta G \propto -\ln \frac{P\_{\text{MUT}}}}{P\_{\text{WT}}

Therefore, we use a graph neural network architecture that takes as input the local structure of a protein, which is represented as a k-nearest neighbor graph (k-NN graph), where nodes represent amino acid residues and edges are defined based on Euclidean distances between Cα atoms. The features of each node include the amino acid type as well as three dihedral angles (φ, ψ, ω), and the edge features contain distances between main chain atoms, sequence position information, and chain information.

Pythia uses a message-passing neural network (MPNN) that is customized using Attention-based message passing and readout functions. In each layer of the Attention Message Passing Layer (AMPL), vertex representations are updated using Attention blocks and connected to edge connection representations to obtain message representations.

The training task of the model is to predict the correct amino acid type for the center node. In this way, Pythia is able to decode the intrinsic patterns between residues in a given protein and thus accurately predict the effects of mutations.

Interpretability of Pythia models:

Since Pythia employs an attention mechanism, we can utilize this model for investigating whether it is indeed successful in capturing complex interactions within proteins. The results show that Pythia has a high attention score for mutant pairs, indicating that Pythia is sensitive to the structure of the mutant and can effectively capture important relationships between mutant residues and their surroundings.

Model Evaluation:

Pythia outperforms existing deep learning methods and other traditional methods in terms of predictive relevance and computational speed, and achieves a 105-fold speedup over traditional force-field based methods.

Pythia was successfully applied to predict effective heat-stable mutations in limonene epoxide hydrolase (LEH) and demonstrated its potential for exploring the structure of 26 million high-quality proteins.