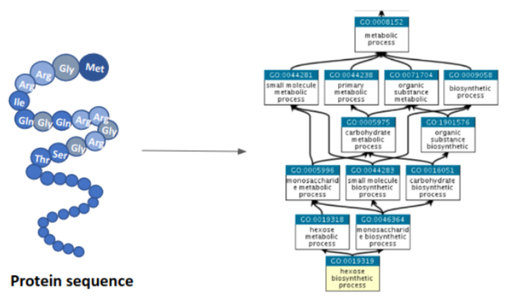
Automated Protein Function Description for Novel Class Discovery

Meet Barot[1], Vladimir Gligorijevic[2], Richard Bonneau[1,2,3], Kyunghyun Cho[1,2]

[1]New York University Center for Data Science, [2]Prescient Design (Genentech), [3]NYU Biology

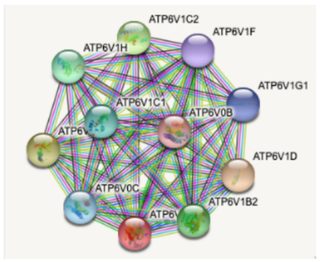
Protein function prediction as it is

Supervised multilabel problem, where sequences are mapped to labels organized into a hierarchy, e.g. the Gene Ontology



Protein function prediction as it **should be**

Given a set of proteins, describe their common function.



→ "Proton-transporting atpase activity, rotational mechanism."

Motivation

- Why use sets of proteins?
  - A function description is our abstraction of the common property of a group of proteins.
  - We discover functions by understanding that a group of proteins do something in common.
- Why use natural language?
  - We can avoid having our predictions be limited to a pre-defined set of functions
  - With language, the model can compose new functions out of the same pieces that we use to explain the world to each other.

To discover new categories of protein function with info to guide experimental design to test for them, we need a model that generates functional descriptions.

Current methods for function discovery

- Many methods exist for function prediction, but most do not consider the problem of discovering novel functions.
- Clustering-based methods are not able to give much information about the new functional categories that they predict
- DeepGOZero [1] predicts for terms not included in the training set, but this is limited to terms with ontological relations with known terms.

Proposed model

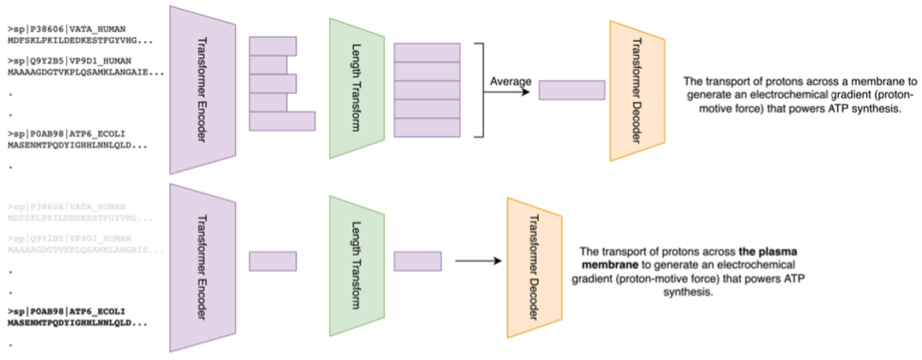


Figure: Transformer encoder-decoder model with length transform [2] to handle sets of sequences. Estimates $P(d|S)$ where S is a set of protein sequences and d is a given description.

Data

- Uniprot-KB Swiss-Prot (manually annotated and reviewed), 566,996 proteins total
  1. Maximum number of proteins per GO term: 1280
  2. Minmum number of proteins per GO term: 32

Table: Number of proteins and GO terms in training and test sets.

	Train P&F	Train P, Test F	Test P, Train F	Test P&F
Prots	316k	181k	20k	20k
Funcs	9k	2k	879	1.5k

Evaluation Metrics

1. Correctness $\in [0, 1]$:
   - Average number of times a correct GO term outranks an incorrect one for a given sequence set prediction.
2. Specificity $\in [0, 1]$:
   - Among correct GO terms, the average number of times a child outranks its parent.
3. Robustness $\in [-1, 1]$:
   - The average rank correlation between the predictions of a pair of different identically annotated sequence sets.

Results

Table: Model Performances.

Metric	Train P, Test F	Test P, Train F	Test P&F
Correctness	0.8844	0.8014	0.7157
Specificity	0.5765	0.5526	0.5701
Robustness	0.4020	0.1977	0.2362

Test set generation examples

- Prediction: the directed movement of proteins from endoplasmic reticulum to the nucleus .
- Actual description: the targeting and directed movement of proteins into a cell or organelle . not all import involves an initial targeting event .

- Prediction: the process whose specific outcome is the progression of the eye over time , from its formation to the mature structure .
- Actual description: the process in which the anatomical structures of appendages are generated and organized . an appendage is an organ or part that is attached to the trunk of an organism .

- Prediction: any process that modulates the frequency , rate or extent of cell differentiation .
- Actual description: any process that activates or increases the frequency , rate or extent of cell differentiation .

- Prediction: any process involved in forming the mature 3 ' end of a dna (mrna) molecule .
- Actual description: a protein complex that contains the gins complex , cdc45p , and the heterohexameric mcm complex , and that is involved in unwinding dna during replication .

[1] Maxat Kulmanov and Robert Hoehndorf. "DeepGOZero: Improving protein function prediction from sequence and zero-shot learning based on ontology axioms". In: bioRxiv (2022). doi: 10.1101/2022.01.14.476325.

[2] Raphael Shu et al. "Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior". In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. 05. 2020, pp. 8846–8853.

Contact me at meetbarot@nyu.edu