CSE7850/CX4803 Machine Learning in Computational Biology

Lecture 18: Protein Function Prediction

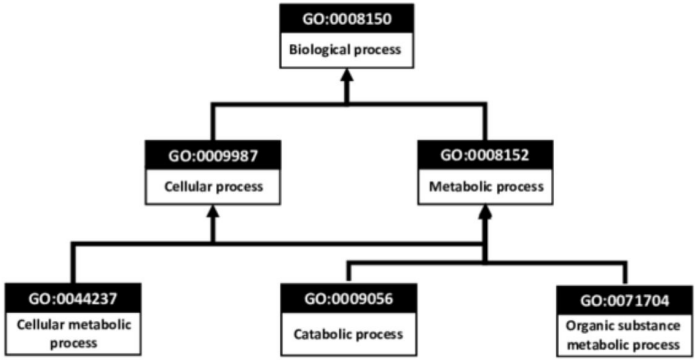
Yunan Luo

Protein function prediction

Two types of prediction problems for protein functions

- Categorical function annotation (classification) Paper #1
  - Scientists have defined a set of function labels
  - ML model predicts the functions for a given protein

Gene Ontology (GO)

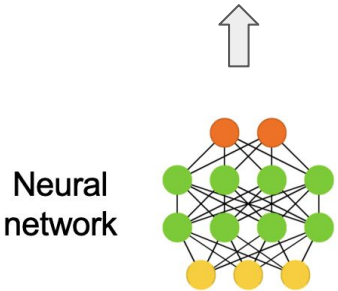


- Quantitative function activity (regression) Paper #2
  - For a given target property (fitness), predict the level of this property
    - Stability
    - Binding affinity

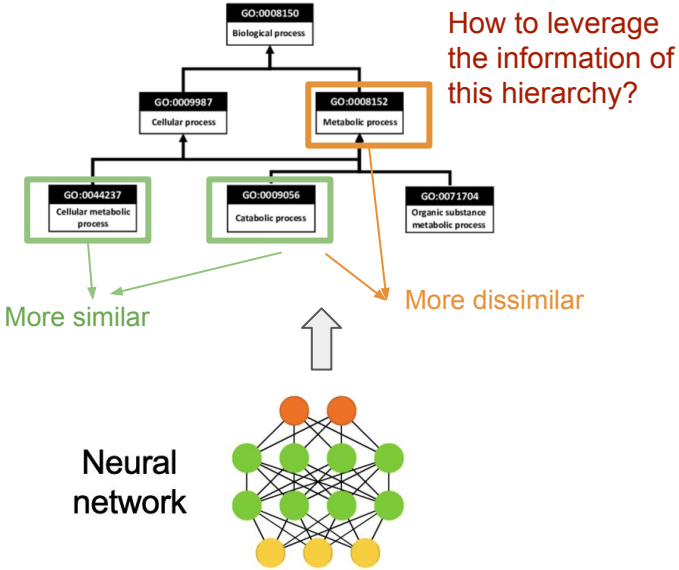
Sequence	Score
DNGVDGEWTYDDA...	1.0
DNGCDGEWTYDDA...	0.2
DNGVWGEWTYDDA...	5.4
DNGVSGEWTYDDA...	0.6
DNGVDGFWTYDDA...	1.1
DVGVDGEWTFGDA...	0.7
DNGVDGEWTYFDA...	2.5
YNGVDGEWTYDPA...	0.1

Structure of the label space

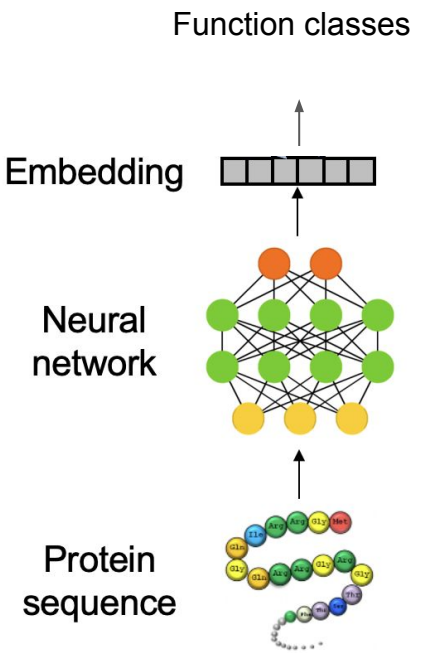
Previously, we considered the multi-class problems where the relationships of labels (classes) were not defined

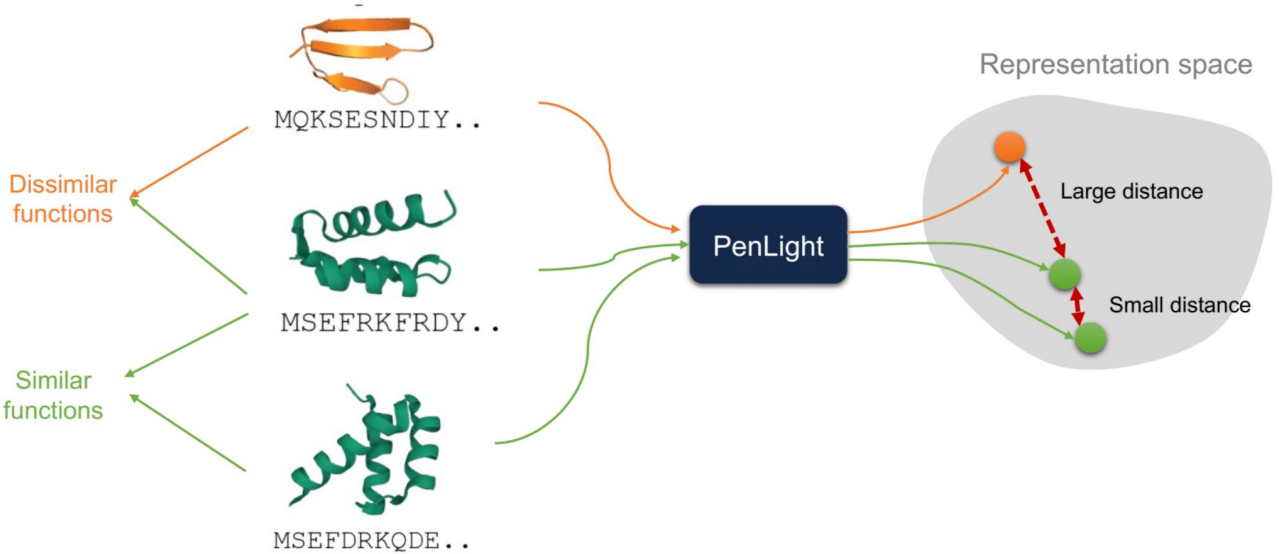


Now, we have a multi-class problems where the relationships of labels (classes) were arranged in a tree-like structure

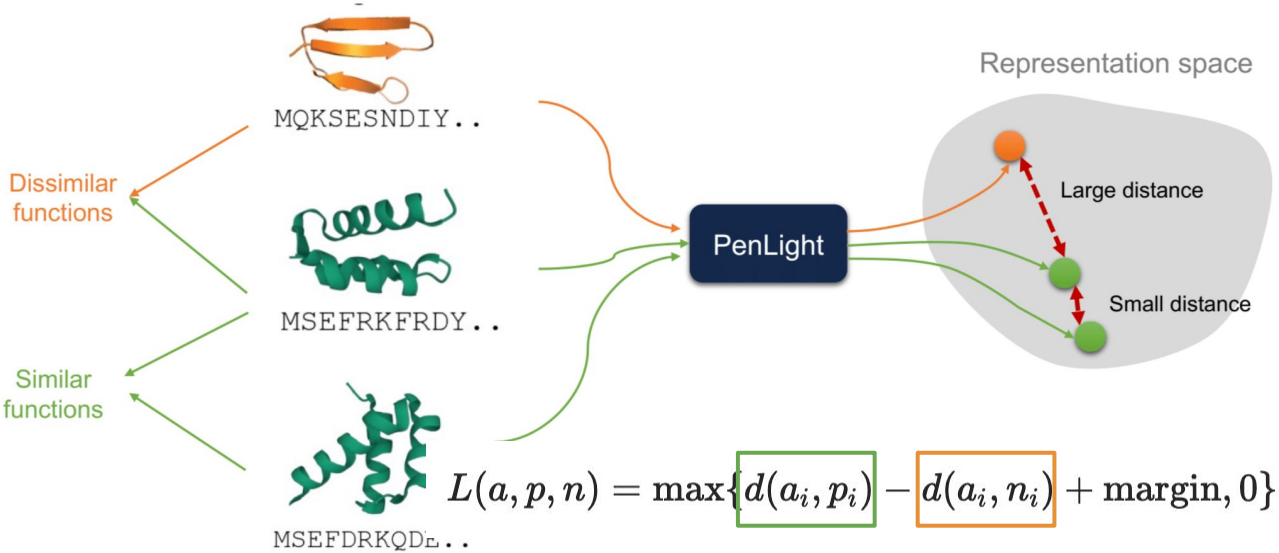


Contrastive learning





Contrastive learning: triplet margin loss



PyTorch code

```
def distance_loss2(self, output_seq1, output_seq2, output_seq3, margin):
    dist_seq1 = self.cos(output_seq1, output_seq2)
    dist_seq2 = self.cos(output_seq1, output_seq3)
    margin = margin.to(output_seq1)
    zeros = torch.zeros(dist_seq1.shape).to(output_seq1)
    loss = torch.mean(torch.max(dist_seq1 - dist_seq2 + margin, zeros))
    return loss
```

Science

RESEARCH

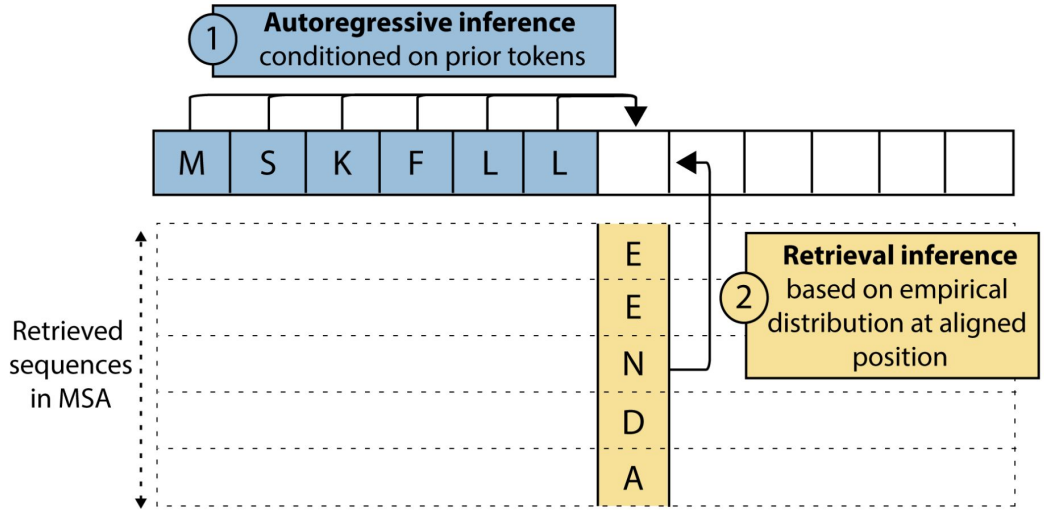
FUNCTION PREDICTION

Enzyme function prediction using contrastive learning

Tianhao Yu[1,2,3]†, Haiyang Cui[1,2,3]†, Jianan Canal Li[3,4], Yunan Luo[5], Guangde Jiang[1,2], Huimin Zhao[1,2,3,6]*

Paper #1 generalizes this idea to multiple hierarchy systems

Paper #2



Sequence	Score
DNGVDGEWTYDDA...	1.0
DNGCDGEWTYDDA...	0.2
DNGVWGEWTYDDA...	5.4
DNGVSGEWTYDDA...	0.6
DNGVDGFWTYDDA...	1.1
DVGVDGEWTFGDA...	0.7
DNGVDGEWTYFDA...	2.5
YNGVDGEWTYDPA...	0.1

An unsupervised approach based on protein language models (PLMs)

- Q: Why PLMs can predict the fitness even without training on labeled data?