Kaggle chanllenge: Proteins Prediction

Team: Ruan (Sicheng Mao, Yang Zhang, Yunhao Chen)

2022-2023 ALTEGRAD

January 29, 2023

Overview

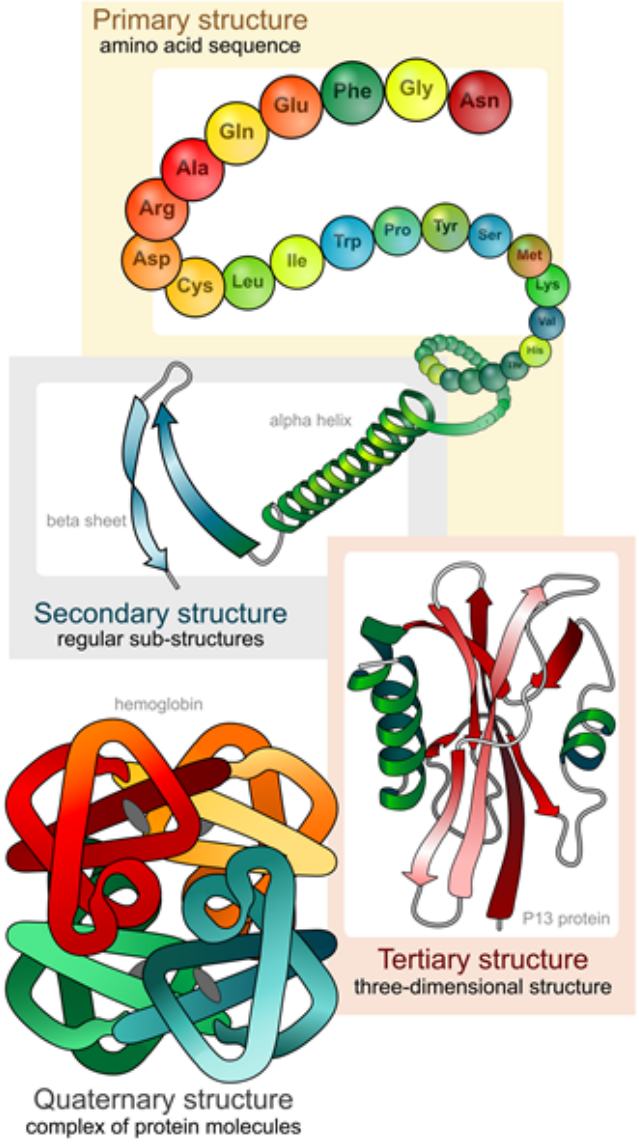
1. Introduction
2. Pipeline
3. Sequential features
4. Structural features
5. Classification
6. Evaluation
7. Conclusion

Introduction

Data:

- sequential information: amino acids sequence (primary structure)
- structural information: graphs (higher order structure)

Task: classify protein into 18 functionality classes.
Evaluation: Cross-Entropy Loss



Pipeline

1. Feature extraction
   - sequential features: apply language models
   - structural features: apply graph models
2. classification, hypertuning, model selection: autoML tool

Sequential features: TF-IDF

- protein sequence $\longleftrightarrow$ plain text
- TF(Term Frequency) * IDF(Inverse Document Frequency)
- Statistical approach
- Shortcomings:
  1. Only frequency information
  2. Lack of generalisation ability
  3. without any domain specific knowledge (biology)

Sequential features: Protvec

- **Protvec**: pretrained model with domain specific knowledge [1]
- Skip-gram model
- Trained on 546,790 protein sequences of Swiss-Prot database [2]
- Lack of diversity: only the **disordered proteins** [3] are considered

[1] Asgari E, Mofrad MRK (2015) Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics

[2] Swiss-Prot is the expertly curated component of UniProtKB (produced by the UniProt consortium)

[3] An intrinsically disordered protein (IDP) is a protein that lacks a fixed or ordered three-dimensional structure, typically in the absence of its macromolecular interaction partners, such as other proteins or RNA

Sequential features: ProteinBERT

- **ProteinBERT**: a self-supervised language model specifically designed for proteins [4]
- Inspired by BERT architecture, aiming at capturing **local and global** representations of proteins
- Gene Ontology **(GO) annotation prediction** as a novel pretraining task scheme

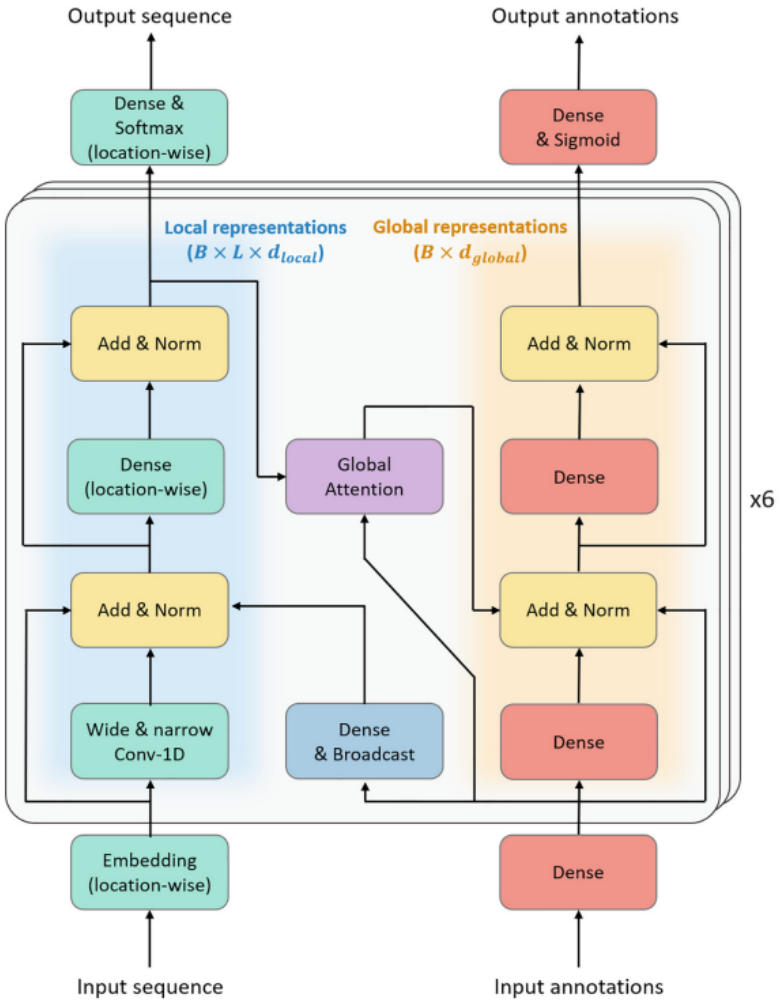
[4] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, Michal Linial, ProteinBERT: a universal deep-learning model of protein sequence and function

Sequential features: ProteinBERT

Information flow through:

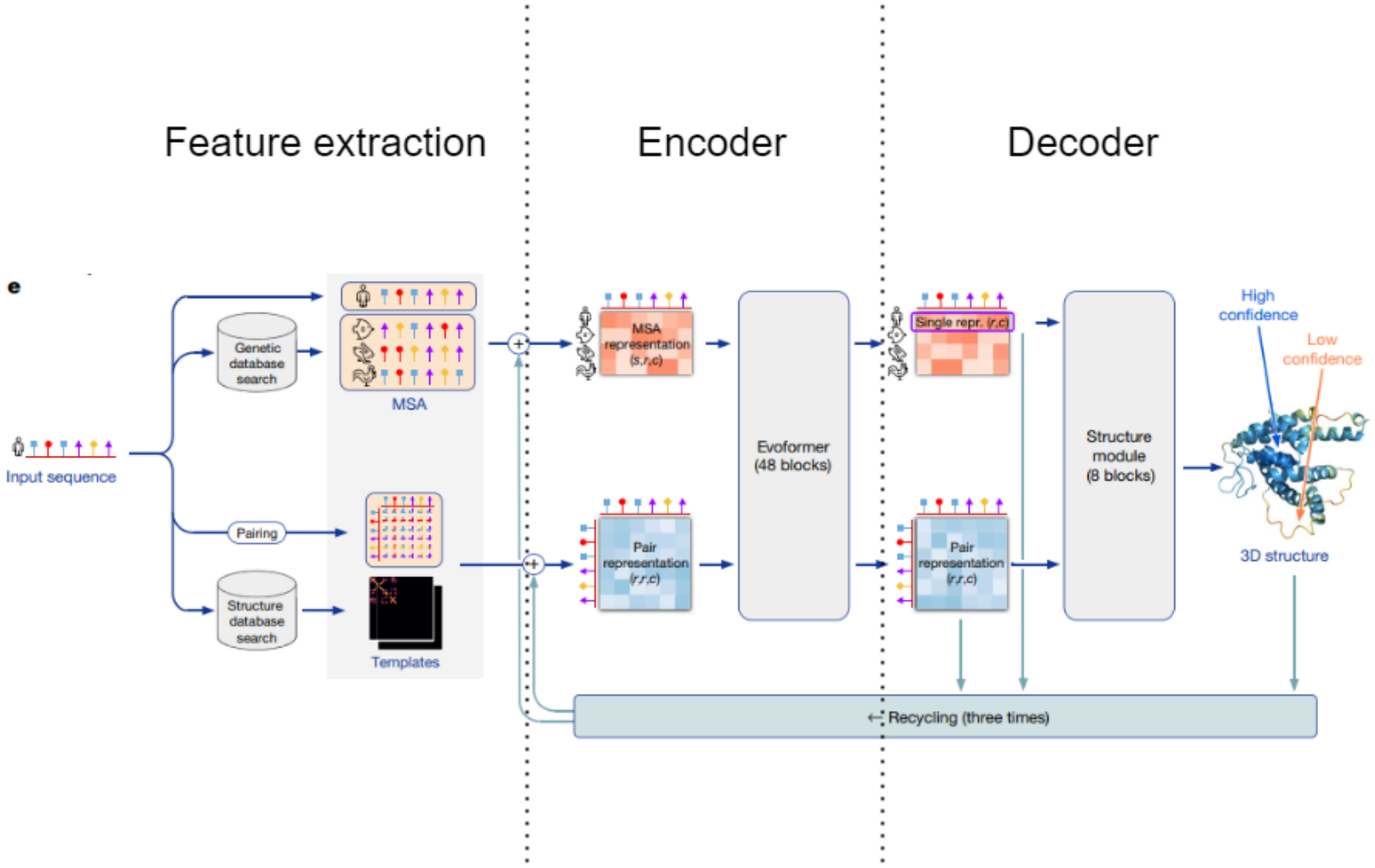
- broadcast dense layers (from global to local)
- attention layers (from local to global)

We make use of the **global** representations as our extracted features



Sequential features: AlphaFold2

- **AlphaFold2**: predicting a proteins 3D structure from its amino acid sequence [5]



[5] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold

Sequential features: ESMFold

- **ESMFold**: another competitive transformer-based model to predict protein fold problem[6]
- We make use of the minimum pretrained ESMfold encoder from Hugging Face
- Extract features for every amino acids in a protein sequence
- Aggregation strategy: Sum up

[6] Evolutionary-scale prediction of atomic level protein structure with a language model

GCN

- GCN: Graph convolutional network[7]
  - A first order approximation of a localized spectral filter on graph.
- Architecture: We use 2 GCN layers and mean readout function.

[7] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks

GCN

- Score: 1.88668
- Shortcomings: Simple GCNs can't have large scale message passing and deep GCNs will have smooth problem.
  Hard to design the architecture.
  mean readout function might be too simple for complex structure information.

GAT

- GAT: Grpah attention networks.[8]
  - A combination of a graph neural network and an attention layer
  - attention weights:

$$\alpha_{i,j} = \frac{exp(LeakyReLU(a^T[\theta x_i || \theta x_j]))}{\sum_{k \in N(i) \cup (i)} exp(LeakyReLU(a^T[\theta x_i || \theta x_k]))}$$

- Architecture: 2 GAT layers with mean readout function.

[8] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks

GAT

- Advantage: The attention weights can be used to evaluate the importance of the features.
- Score: 1.81989
- Problems: Architecture design, pooling method is too simple, no edge features are used in network.

GIN

- GIN: Graph Isomorphism Networks.[9]
  - generalizes the WL test and hence achieves maximum discriminative power among GNNs.
  - feature update:

$$h_v^{(k)} = MLP^{(k)}((1+\epsilon^{(k)}h_v^{(k-1)} + \sum_{u \in N(v)} h_u^{(k-1)})$$

  - GIN readout:

$$h_G = CONCAT(SUM(\{h_v^{(k)}|v \in G)|k = 0, 1, ..., K)$$

- Architecture: 2 GIN layers with mean readout function for pooling.

[9] Xu K, Hu W, Leskovec J, et al. How powerful are graph neural networks?

GIN

- Score: 1.84116
- Problems: Architecture design, pooling method is too simple, no edge features are used in network.

Classification

Feed the extracted features into an auto machine learning tool: AutoGluon.

- automatically perform classification on a collection of models
- automatically perform hyperparameter sweeping
- automatically ensemble models

	model	score_val	pred_time_val	fit_time	pred_time_val_marginal	fit_time_marginal	stack_level	can_infer	fit_order
0	WeightedEnsemble_L3	-0.891835	8.574015	6110.735480	0.001576	4.565410	3	True	26
1	CatBoost_BAG_L2	-0.919700	4.754232	5664.494449	0.105590	1903.196111	2	True	20
2	LightGBMXT_BAG_L2	-0.927140	4.811049	3913.547895	0.162406	152.249557	2	True	16
3	WeightedEnsemble_L2	-0.930269	2.540211	3183.077281	0.001617	5.116302	2	True	14
4	XGBoost_BAG_L2	-0.939213	5.009479	4027.771781	0.360837	266.473444	2	True	23
5	LightGBM_BAG_L2	-0.969090	4.797922	3995.934522	0.149280	234.636185	2	True	17
6	CatBoost_BAG_L1	-0.999276	0.116058	3022.558098	0.116058	3022.558098	1	True	8
7	NeuralNetFastAI_BAG_L2	-1.023223	6.562722	3771.529728	1.914080	10.231391	2	True	15
8	LightGBMLarge_BAG_L2	-1.035384	5.049778	4464.105117	0.401136	702.806779	2	True	25
9	NeuralNetTorch_BAG_L2	-1.058477	6.358984	3772.120343	1.710342	10.822006	2	True	24
10	ExtraTreesGini_BAG_L2	-1.078011	5.019598	3761.968069	0.370956	0.669732	2	True	21
11	RandomForestEntr_BAG_L2	-1.085409	4.936126	3770.630255	0.287484	9.331918	2	True	19
12	RandomForestGini_BAG_L2	-1.091064	5.006528	3763.367219	0.357886	2.068882	2	True	18
13	ExtraTreesEntr_BAG_L2	-1.091783	5.013200	3761.949035	0.364558	0.650698	2	True	22
14	XGBoost_BAG_L1	-1.102782	0.451378	128.769110	0.451378	128.769110	1	True	11
15	LightGBMXT_BAG_L1	-1.137872	0.284004	97.370624	0.284004	97.370624	1	True	4
16	NeuralNetFastAI_BAG_L1	-1.184091	0.891164	7.364564	0.891164	7.364564	1	True	3
17	LightGBM_BAG_L1	-1.209016	0.241845	134.024144	0.241845	134.024144	1	True	5
18	LightGBMLarge_BAG_L1	-1.260780	0.541614	350.762837	0.541614	350.762837	1	True	13
19	NeuralNetTorch_BAG_L1	-1.287691	0.169458	8.474578	0.169458	8.474578	1	True	12
20	RandomForestEntr_BAG_L1	-1.323998	0.223407	9.300573	0.223407	9.300573	1	True	7
21	RandomForestGini_BAG_L1	-1.326605	0.228892	1.465503	0.228892	1.465503	1	True	6
22	ExtraTreesGini_BAG_L1	-1.400114	0.245759	0.556181	0.245759	0.556181	1	True	9
23	ExtraTreesEntr_BAG_L1	-1.409887	0.251137	0.594346	0.251137	0.594346	1	True	10
24	KNeighborsDist_BAG_L1	-2.470747	0.458238	0.028554	0.458238	0.028554	1	True	2
25	KNeighborsUnif_BAG_L1	-2.527403	0.545687	0.029226	0.545687	0.029226	1	True	1

Evaluation

	Feature extraction	CE Loss
Sequence	TFIDF + SVD 100	1.19022
	Protvec + SVD 100	1.48465
	ESM + SVD 200	0.99529
	ProteinBert Finetuing	1.10325
	ProteinBert + SVD 100	0.94205
	ProteinBert + SVD 200	**0.89582**
	ProteinBert + SVD 300	0.90897
Structure	2 layer GCN + mean agg	1.88668
	2 layer GAT + mean agg	1.81989
	2 layer GIN + mean agg	1.84116
Both	ProteinBert + GCN + SVD 300	0.91830

Conclusion

- The best score we reach is 0.89(public)/0.87(private), with only sequential feature extracted with ProteinBert Encoder and dimension reduction to 200 dim with SVD.
- Sequential features are more useful than the structural features:
  - primary structure contains more information than higher order structure information.
  - graph models are too naive to extract necessary features.(the global aggregation is too noisy)

possible improvements:

- try more powerful protein models.
- try smarter way of global aggregation strategy (e.g. HGP-SL)