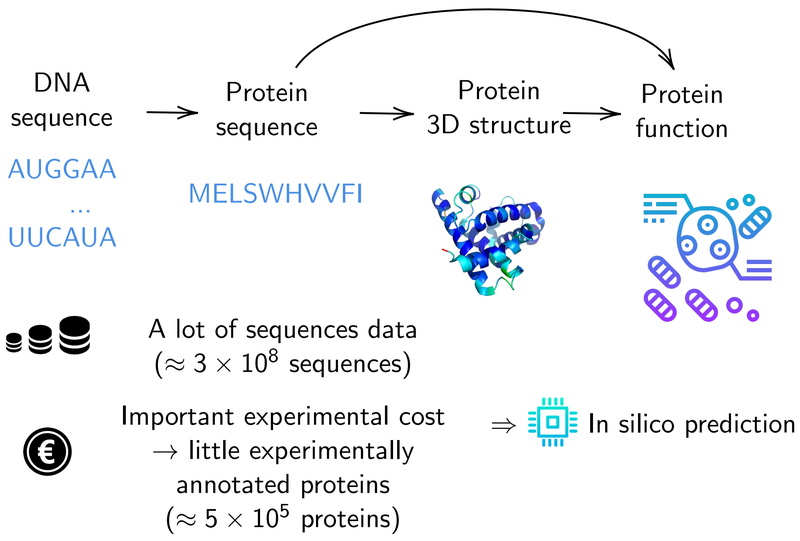
EnzBert: Deep attention network for enzyme class predictions

Nicolas Buton
Yann Le Cunff
François Coste
Univ. Rennes, Inria, CNRS, IRISA, Rennes, France

1. Task

1.1 Functional annotation of protein sequences



A lot of sequences data ($\approx 3 \times 10^8$ sequences)

Important experimental cost → little experimentally annotated proteins ($\approx 5 \times 10^5$ proteins)

⇒ In silico prediction

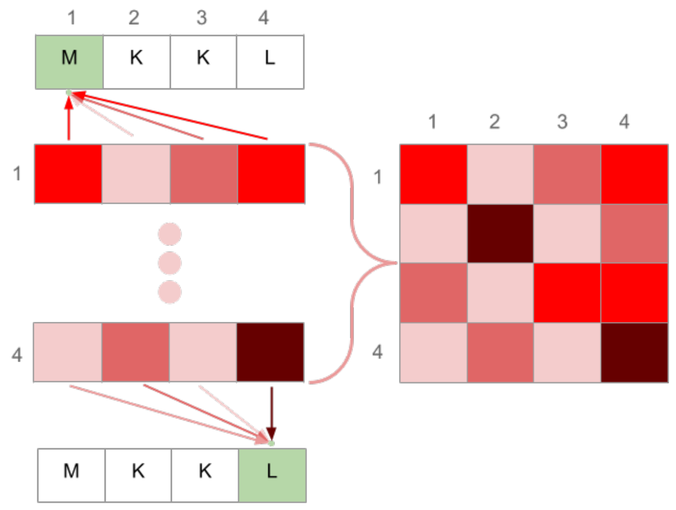
1.2 Challenges

- Sequence information (2.1)
- Underlying 3D structure (2.1)
- Few examples per category (2.2)
- Interpretability (2.3)

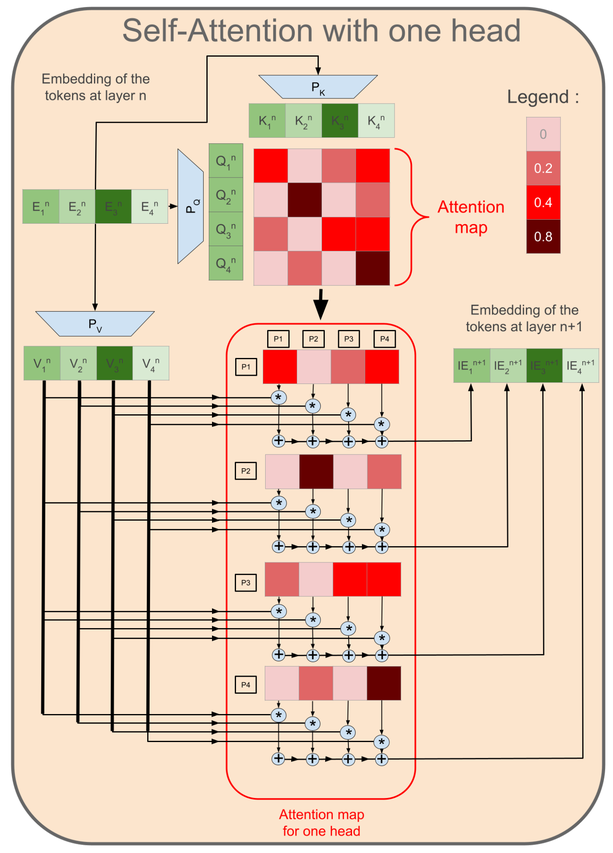
2. Methods

2.1 Attention mechanism and Transformer

Attention is the basis of Transformers, for each position it gives weights to other residues.



Can account for long distance relations between the elements of the sequence



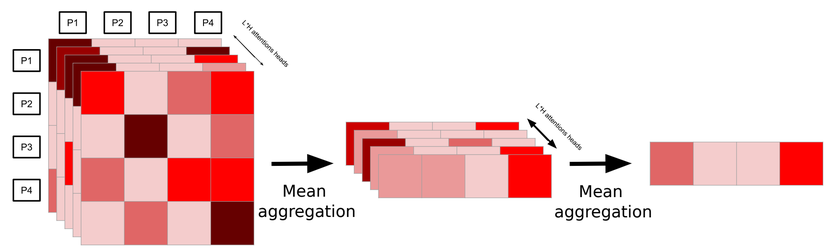
2.2 Two phases training

Unsupervised pre-training on all proteins and supervised fine-tuning step on enzyme classification

Produces a meaningful embedding space for proteins and exploits this during the fine-tuning for better generalization

2.3 Interpretability method

Attention maps from multiple layers and multiple heads. A new interpretability method to get residues' importance scores : attention aggregation.



3. Results

3.1 Enzyme classification

Comparison with state-of-the art method at level 1 and 2 of the enzyme commission (EC) number on EC40 dataset

Model	Accuracy at level 1	Accuracy at level 2
UDSMProt[1]	0.87	0.84
EnzBert	**0.97**	**0.95**

Macro metrics, comparison with state-of-the-art method at different levels of the enzyme commission (EC) number on ECPred40 dataset.

Model	Level	Macro-precision	Macro-recall	Macro-f1
ECPred[2]	0	0.78	0.78	0.77
EnzBert	0	**0.80**	**0.80**	**0.80**
ECPred	1	**0.69**	**0.84**	**0.73**
EnzBert	1	0.62	0.55	0.55
ECPred	2	0.48	**0.61**	0.51
EnzBert	2	**0.57**	0.60	**0.55**
ECPred	3	0.50	0.56	0.50
EnzBert	3	**0.53**	**0.57**	**0.52**
ECPred	4	0.43	0.41	0.40
EnzBert	4	**0.49**	**0.49**	**0.47**

3.2 Interpretability

Benchmark: Evaluated on the identification of catalytic residues on enzymes with respect to different interpretability methods.

Method type	PRG-AUC (x100)	max F-Gain (%)	Time (s)
Random	42.54 ± 4.37	69.85 ± 1.04	–
Grad	75.01	81.27	4.64
Grad X input	63.62	78.66	7.74
Integrated grad	76.41	81.70	2.48×10^2
Attn last layer	87.80	85.62	**2.87**
Attn aggregation	**98.01**	**96.05**	3.72
Rollout	66.08	76.77	2.95
TGLRP	90.92	88.56	4.05×10^1
TGradCam	81.00	76.77	4.35×10^1
LIME	93.46	91.44	1.73×10^4

Attention-based interpretability performs best

NH(3)-dependent NAD(+) synthetase enzyme

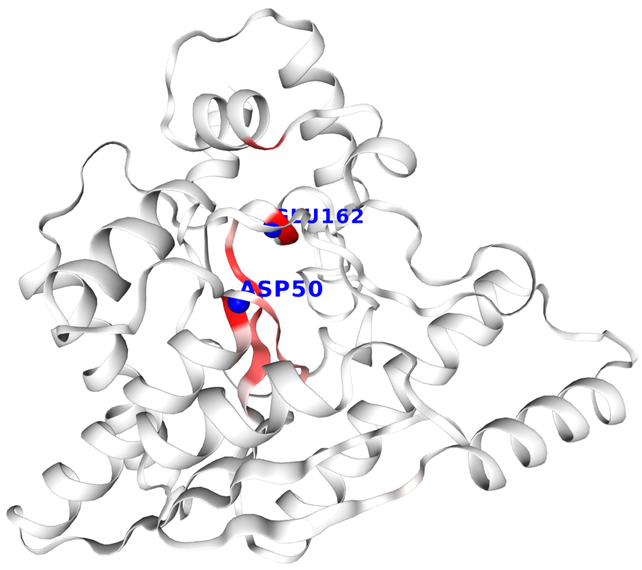


Fig. : Top 5% of residues' importance score of our interpretability method are highlighted in red and catalytic sites are represented by blue spheres.

The two highest scores on attention aggregation correspond to the two catalytic residues of the enzyme.

4. Conclusion

- State-of-the-art prediction on enzymes' classes from sequences only for our model EnzBert
- New interpretability method for Transformers that works very well on enzymes
- Prospects: Considering the hierarchy of Enzyme Commission (EC) number may improve our results

Contacts

email : nicolas.buton@irisa.fr

Bibliography

1. Strodthoff, N., Wagner, P., Wenzel, M., and Samek, W. (2020). UDSMProt: universal deep sequence models for protein classification. Bioinformatics , 36 (8), 2401–2409
2. Dalkiran, A., Rifaioglu, A. S., Martin, M. J., Cetin-Atalay, R., Atalay, V., and Do ğan, T. (2018). ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. BMC Bioinformatics , 19 (1), 334