Do deep neural networks improve functional annotation of alpha-cyanobacteria?

Juliana SILVA BERNARDES

https://scholar.google.com/citations?user=a-ZYwhEAAAAJ&hl=en

https://www.lcqb.upmc.fr/julianab/

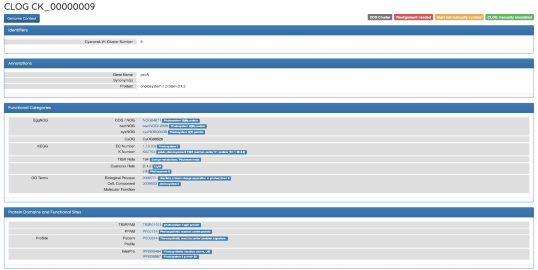
CNRS • SORBONNE UNIVERSITÉ
Station Biologique
de Roscoff

jusilvabernardes@sb-roscoff.fr;
juliana.silva_bernardes@sorbonne-universite.fr

Why α-Cyanobacteria ?

- Phototrophic **microorganisms**, appeared **3.5 billion** years ago
- **Responsible** for the apparition of **Oxygen** on Earth
- One of the **most diverse/widely distributed** prokaryotic phyla
- Colonize both **terrestrial** and **aquatic environments**
- Origin of land **plants /marine algae** over the next millions of years
- They are **most abundant photosynthetic organisms** in the ocean and large lakes.

Cyanorak



CLOG CK_00000009

Genome Context

CDS Cluster | Realignment needed | Start not manually curated | CLOG manually annotated

Identifiers

Cyanorak V1 Cluster Number	9

Annotations

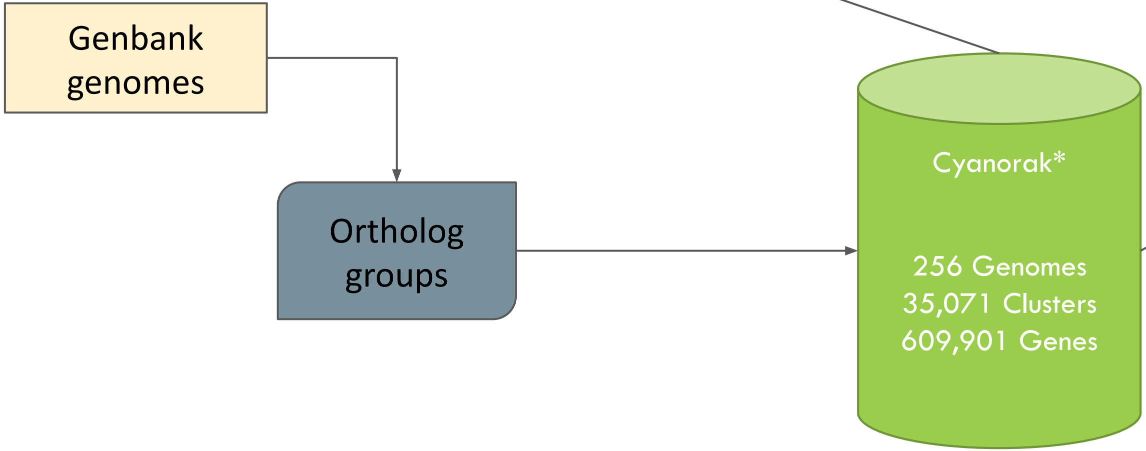
Gene Name	psbA
Synonym(s)	
Product	photosystem II protein D1.2

Functional Categories

EggNOG	COG / NOG	NOG04871 Photosystem Q(B) protein
	bactNOG	bactNOG12939 Photosystem Q(B) protein
	cyaNOG	cyaNOG00439 Photosystem Q(B) protein
	CyOG	CyOG00526
KEGG	EC Number	1.10.3.9 Photosystem II
	K Number	K02703 psbA; photosystem II P680 reaction center D1 protein [EC:1.10.3.9]
	TIGR Role	164 Energy metabolism / Photosynthesis
	Cyanorak Role	D.1.2 Light J.8 Photosystem II
GO Terms	Biological Process	0009771 obsolete primary charge separation in photosystem II
	Cell. Component	0009523 photosystem II
	Molecular Function	

Protein Domains and Functional Sites

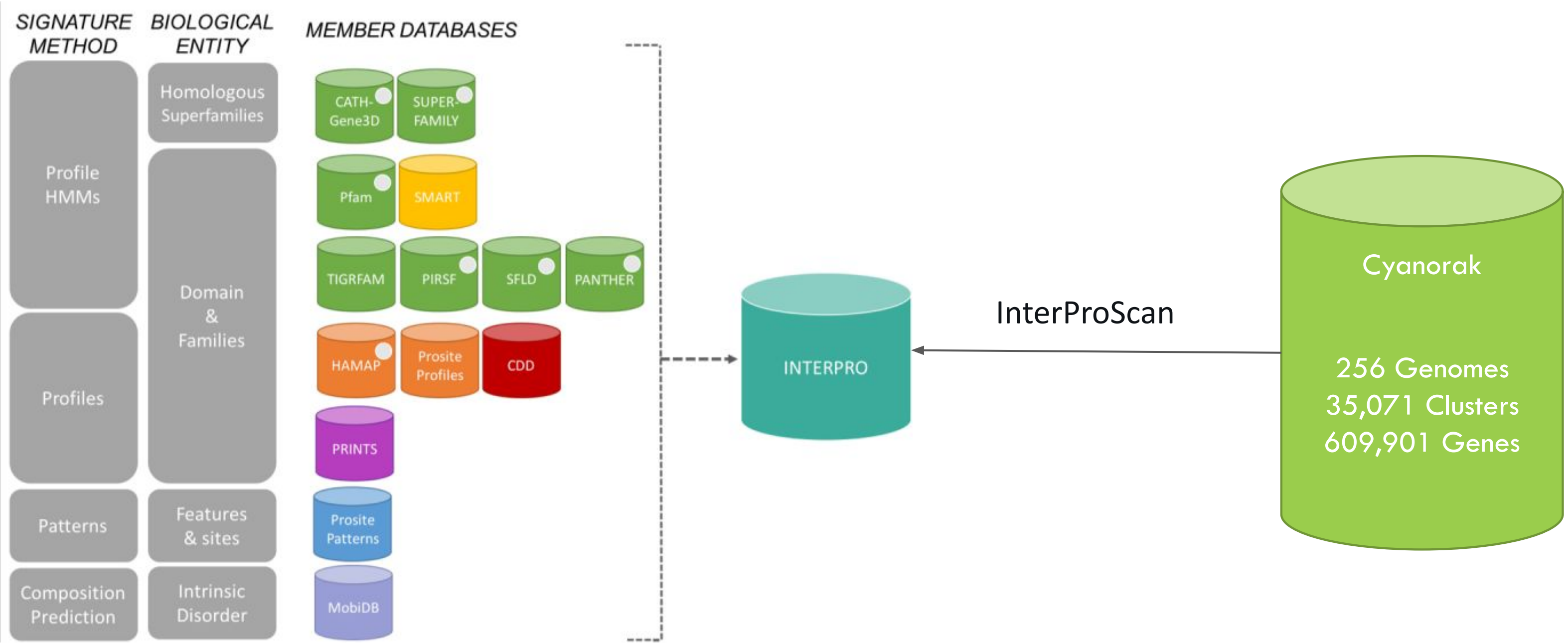
	TIGRFAM	TIGR01151 photosystem II q(b) protein
	PFAM	PF00124 Photosynthetic reaction centre protein
ProSite	Pattern	PS00244 Photosynthetic reaction center proteins signature.
	Profile	
	InterPro	IPR000484 Photosynthetic reaction centre, L/M IPR005867 Photosystem II protein D1



*https://academic.oup.com/nar/article/49/D1/D667/5943826

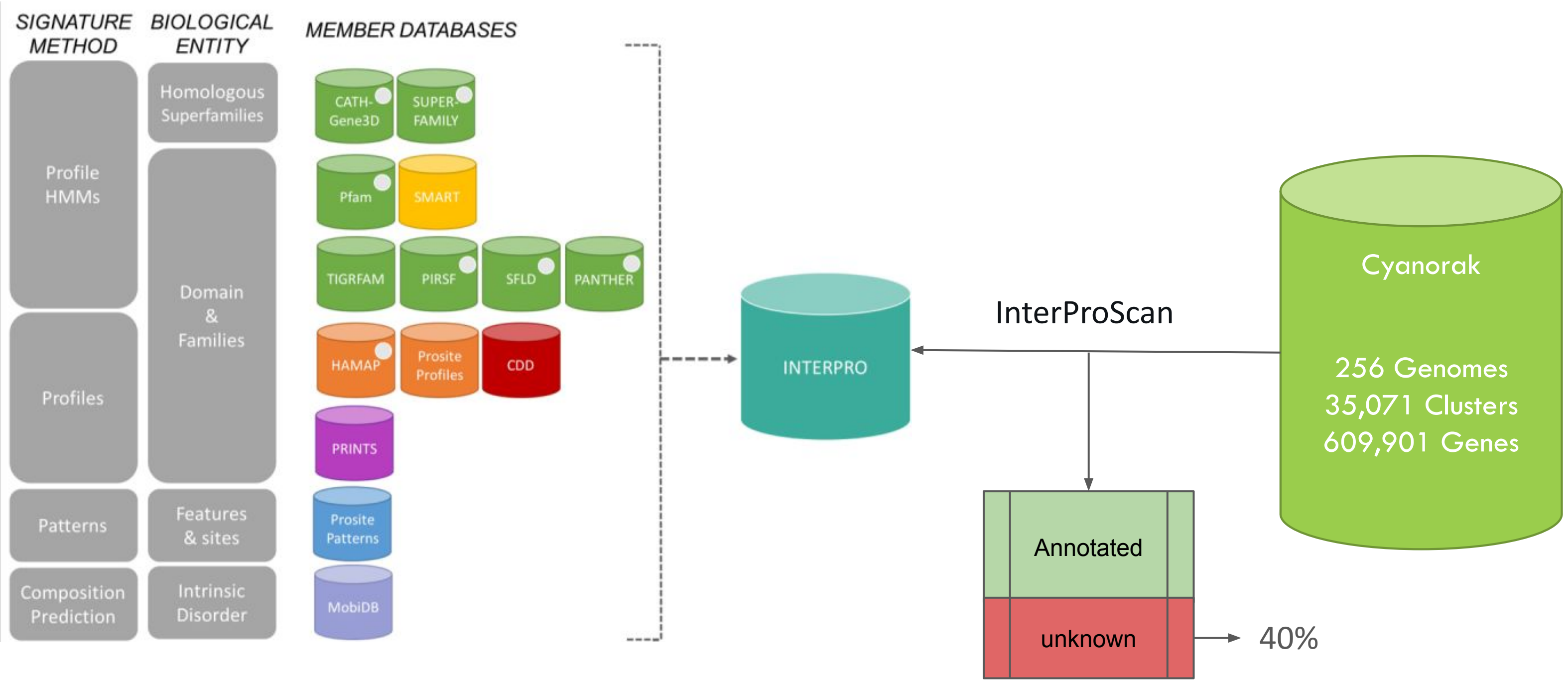
Functional Annotation

- Functional annotation of α-Cyanobacteria is based on Interpro database



Functional Annotation

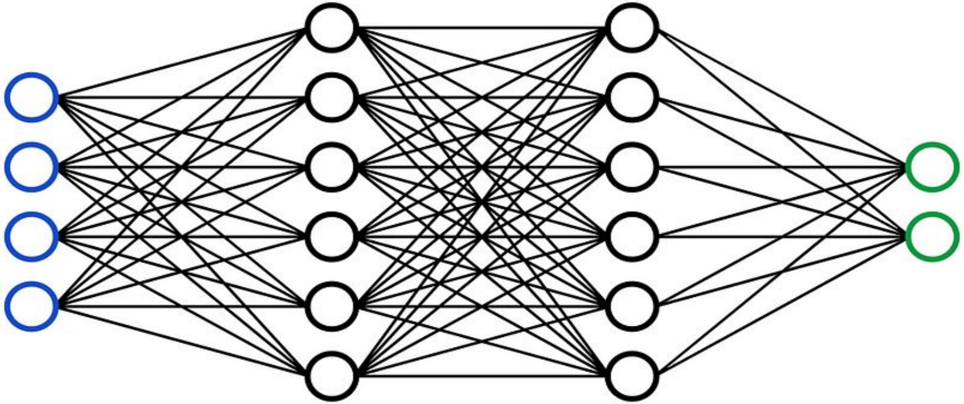
- Functional annotation of α-Cyanobacteria is based on Interpro database



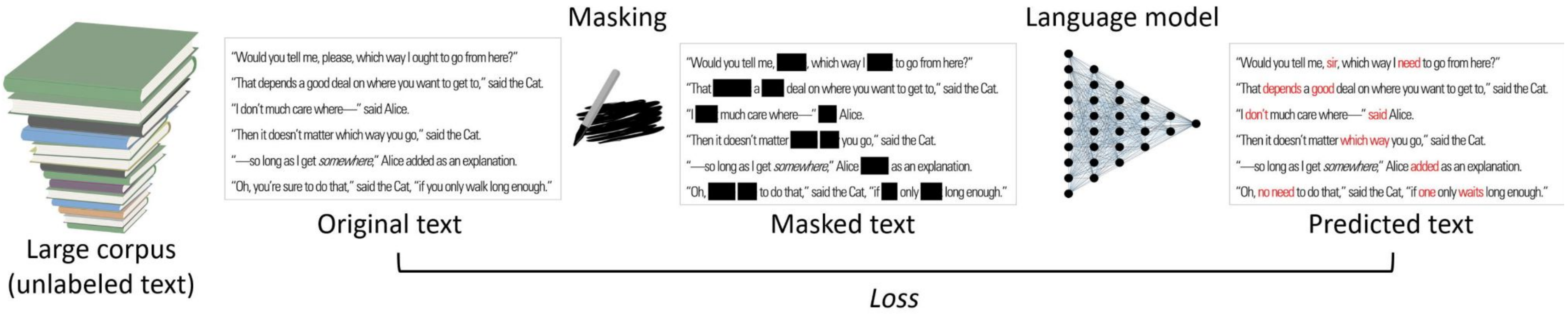
Functional Annotation

- Deep neural networks have already reached impressive results in protein three-dimensional structure prediction.

- Here, we investigated if a such methodology could improve function annotation in α-Cyanobacteria



Large language models

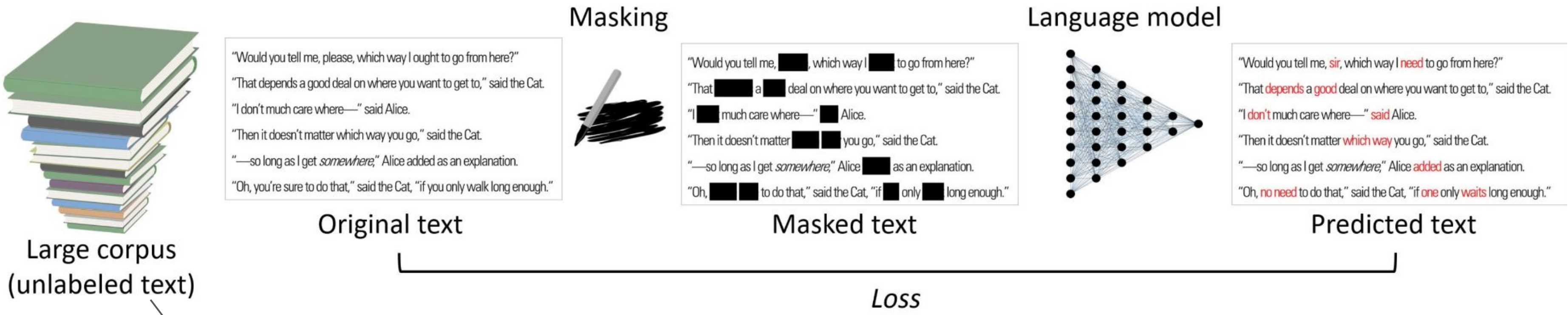


"Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where—" said Alice.
"Then it doesn't matter which way you go," said the Cat.
"—so long as I get *somewhere*," Alice added as an explanation.
"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

"Would you tell me, [redacted], which way I [redacted] to go from here?"
"That [redacted] a [redacted] deal on where you want to get to," said the Cat.
"I [redacted] much care where—" [redacted] Alice.
"Then it doesn't matter [redacted] [redacted] you go," said the Cat.
"—so long as I get *somewhere*," Alice [redacted] as an explanation.
"Oh, [redacted] [redacted] to do that," said the Cat, "if [redacted] only [redacted] long enough."

"Would you tell me, sir, which way I need to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where—" said Alice.
"Then it doesn't matter which way you go," said the Cat.
"—so long as I get *somewhere*," Alice added as an explanation.
"Oh, no need to do that," said the Cat, "if one only waits long enough."

Protein large language models



"Would you tell me, please, which way I ought to go from here?"

"That depends a good deal on where you want to get to," said the Cat.

"I don't much care where—" said Alice.

"Then it doesn't matter which way you go," said the Cat.

"—so long as I get *somewhere*," Alice added as an explanation.

"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

"Would you tell me, [redacted], which way I [redacted] to go from here?"

"That [redacted] a [redacted] deal on where you want to get to," said the Cat.

"I [redacted] much care where—" [redacted] Alice.

"Then it doesn't matter [redacted] [redacted] you go," said the Cat.

"—so long as I get *somewhere*," Alice [redacted] as an explanation.

"Oh, [redacted] [redacted] to do that," said the Cat, "if [redacted] only [redacted] long enough."

"Would you tell me, sir, which way I need to go from here?"

"That depends a good deal on where you want to get to," said the Cat.

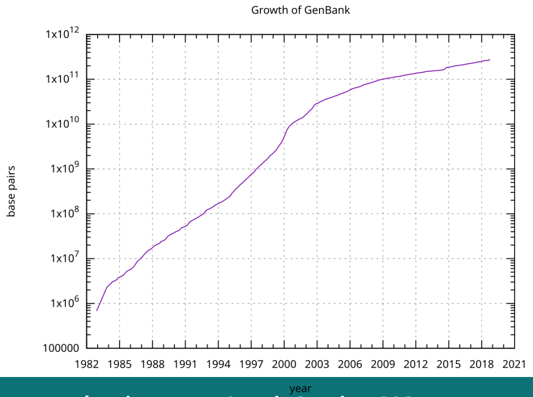
"I don't much care where—" said Alice.

"Then it doesn't matter which way you go," said the Cat.

"—so long as I get *somewhere*," Alice added as an explanation.

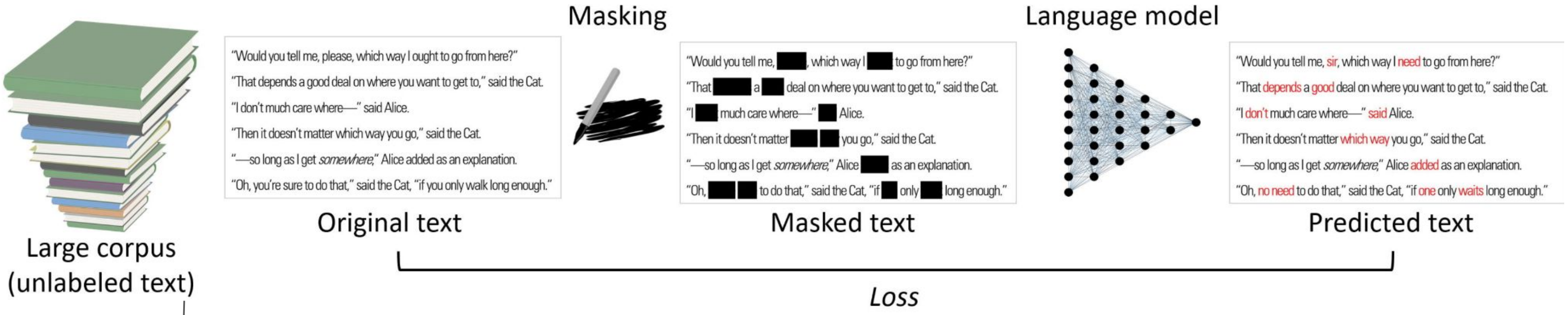
"Oh, no need to do that," said the Cat, "if one only waits long enough."

Protein sequences



As of release 228, retrieved 2018-11-30

Protein language models

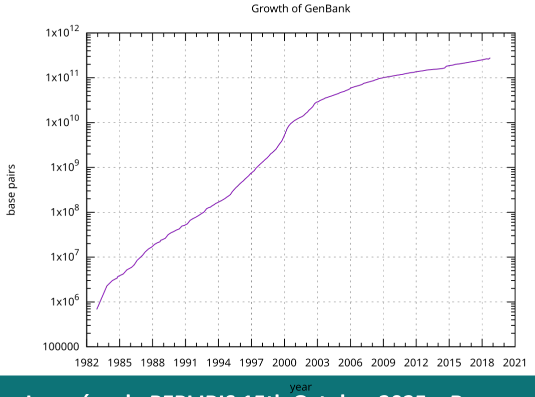


"Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where—" said Alice.
"Then it doesn't matter which way you go," said the Cat.
"—so long as I get *somewhere*," Alice added as an explanation.
"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

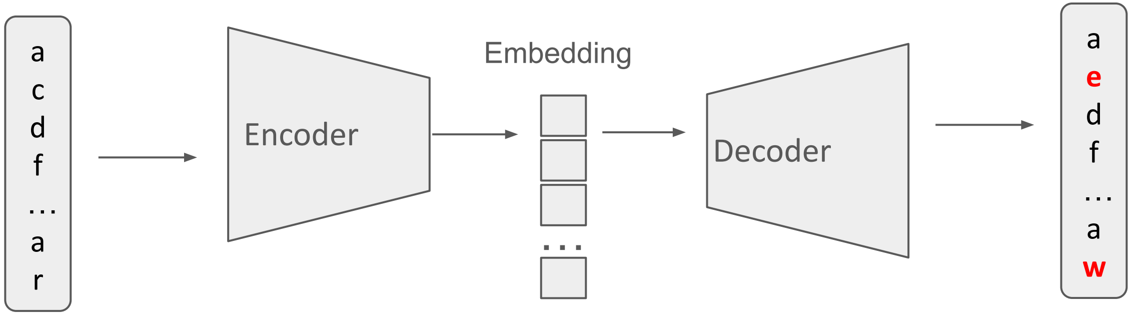
"Would you tell me, [redacted], which way I [redacted] to go from here?"
"That [redacted] a [redacted] deal on where you want to get to," said the Cat.
"I [redacted] much care where—" [redacted] Alice.
"Then it doesn't matter [redacted] you go," said the Cat.
"—so long as I get *somewhere*," Alice [redacted] as an explanation.
"Oh, [redacted] to do that," said the Cat, "if [redacted] only [redacted] long enough."

"Would you tell me, sir, which way I need to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where—" said Alice.
"Then it doesn't matter which way you go," said the Cat.
"—so long as I get *somewhere*," Alice added as an explanation.
"Oh, no need to do that," said the Cat, "if one only waits long enough."

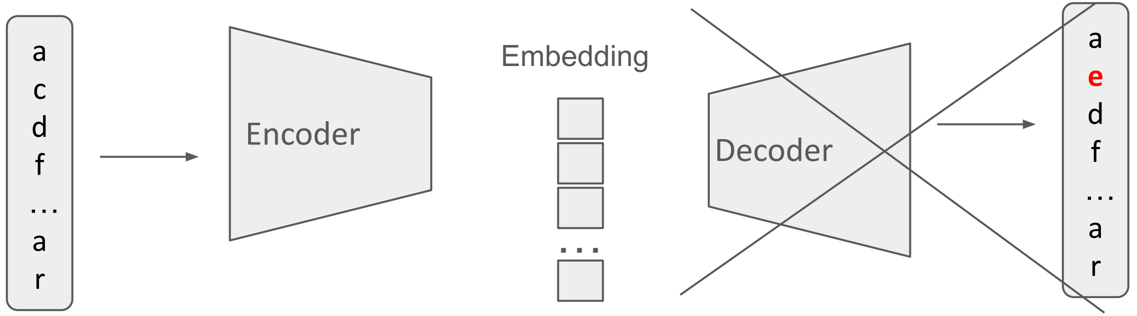
Protein sequences

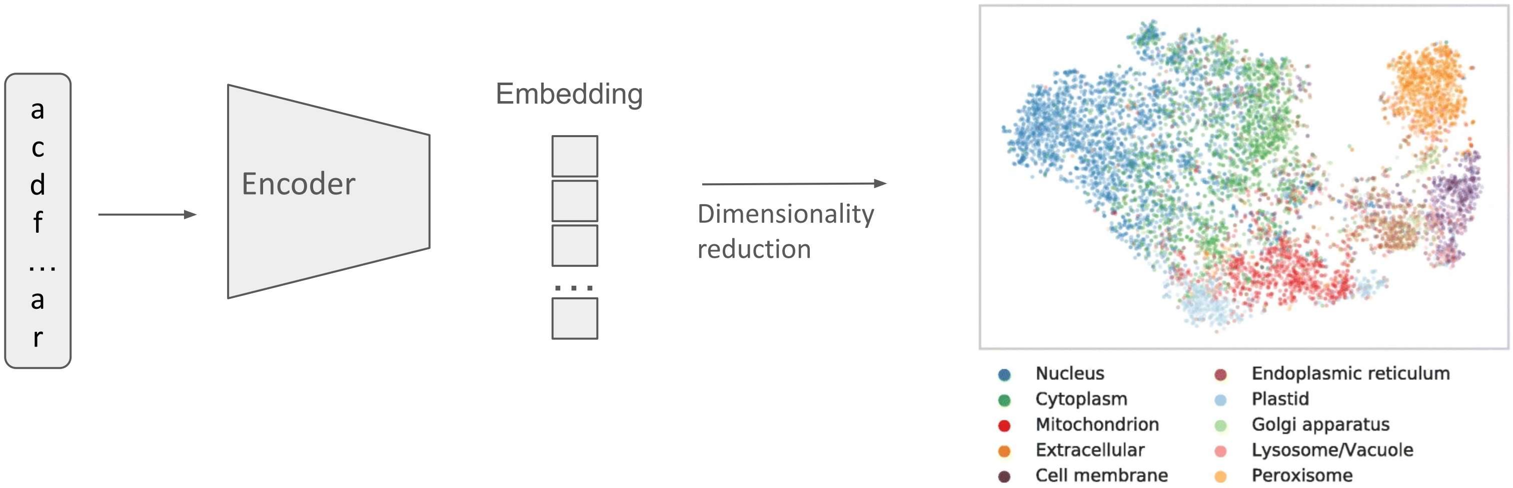


As of release 228, retrieved 2018-11-30

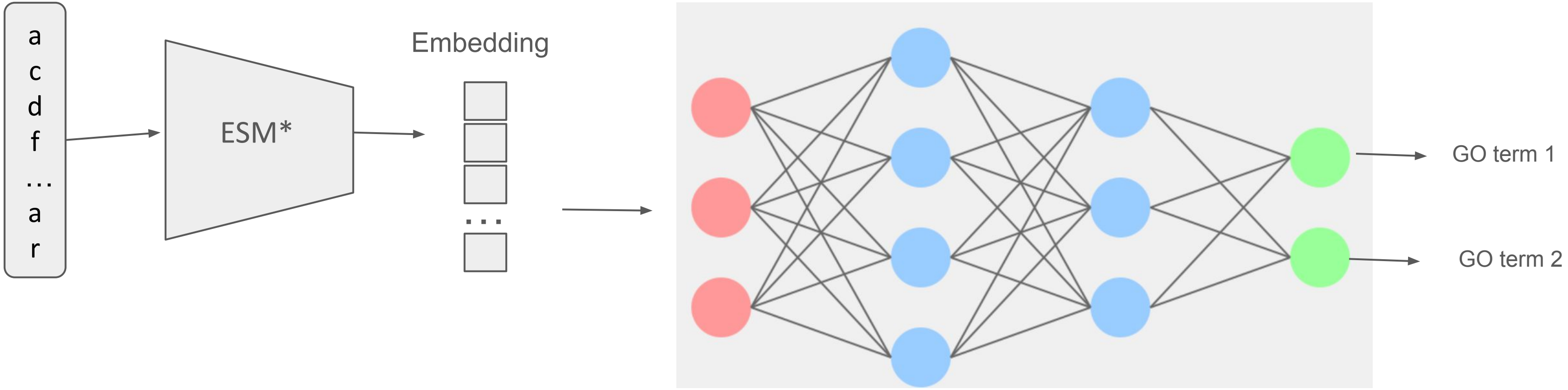


Protein language models





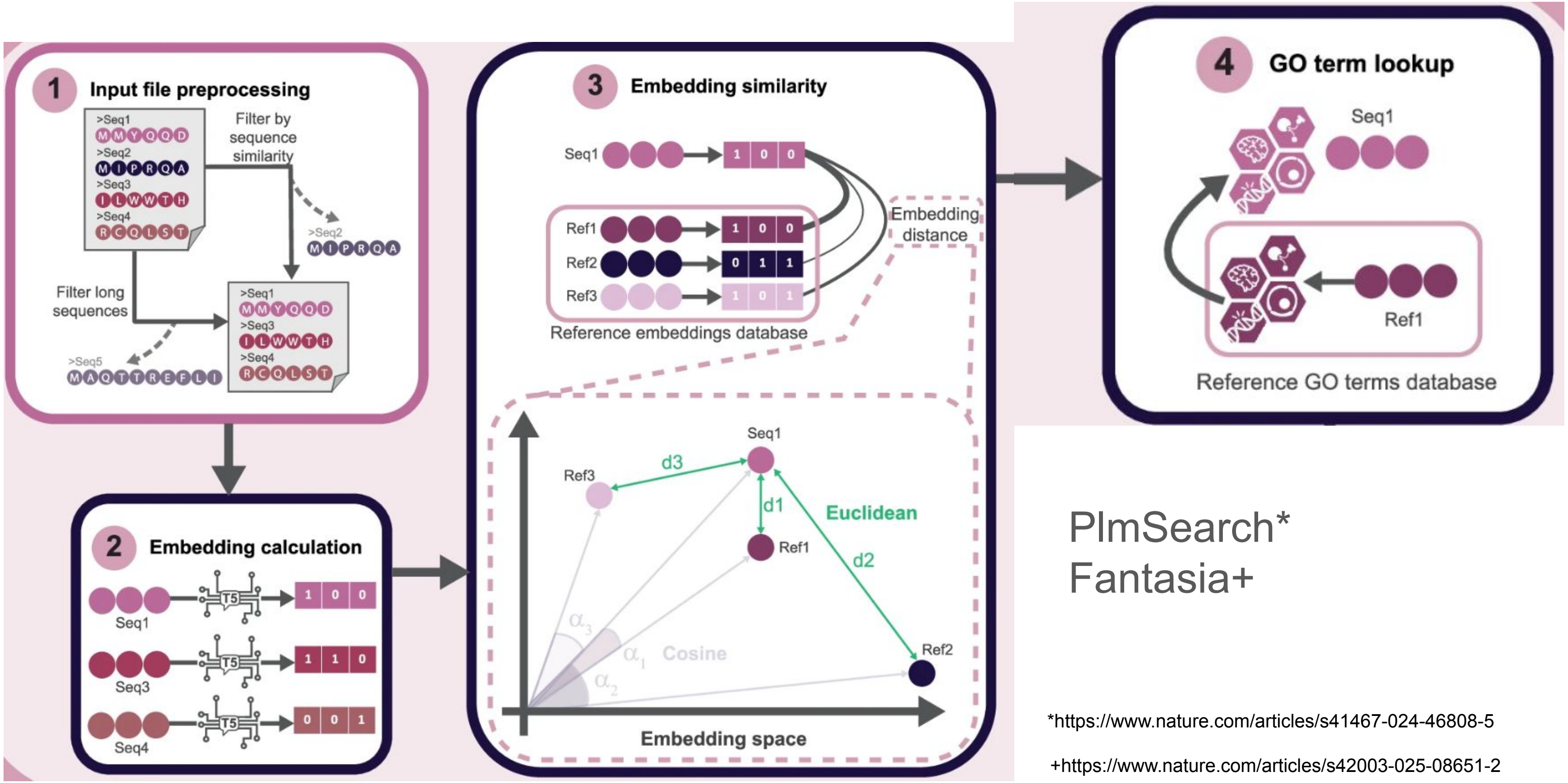
Functional Annotation with Deep-learning



ML classifier

*https://www.pnas.org/doi/abs/10.1073/pnas.2016239118

Functional Annotation with Deep-learning

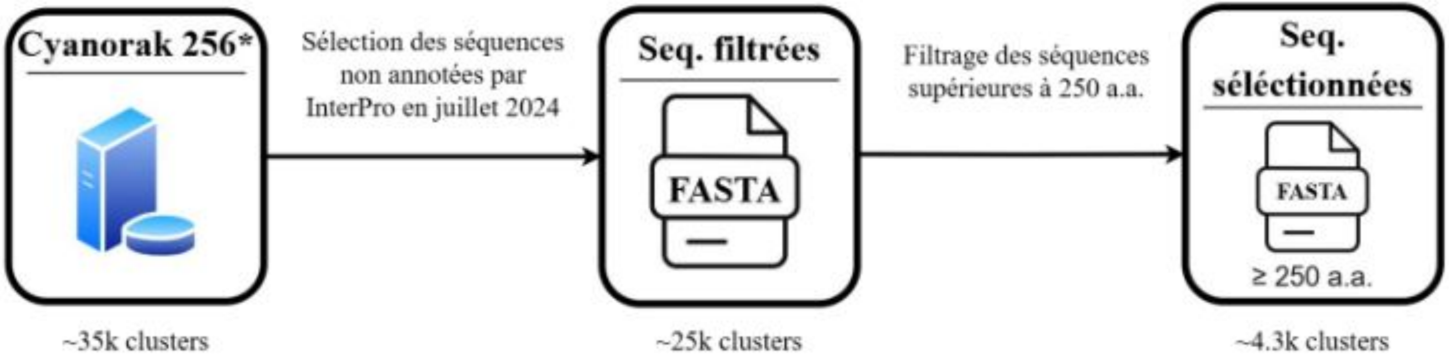


PlmSearch*
Fantasia+

*https://www.nature.com/articles/s41467-024-46808-5

+https://www.nature.com/articles/s42003-025-08651-2

Data preparation



Cyanorak 256*

~35k clusters

Sélection des séquences non annotées par InterPro en juillet 2024

Seq. filtrées

FASTA

~25k clusters

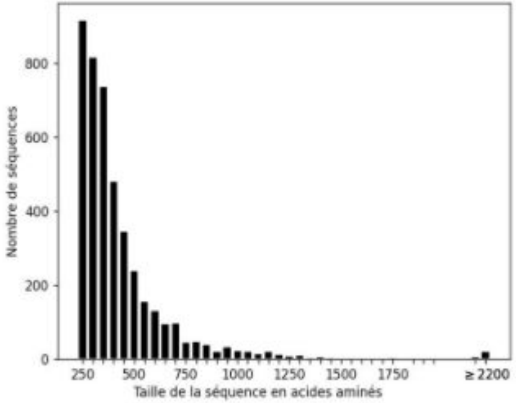
Filtrage des séquences supérieures à 250 a.a.

Seq. séléctionnées

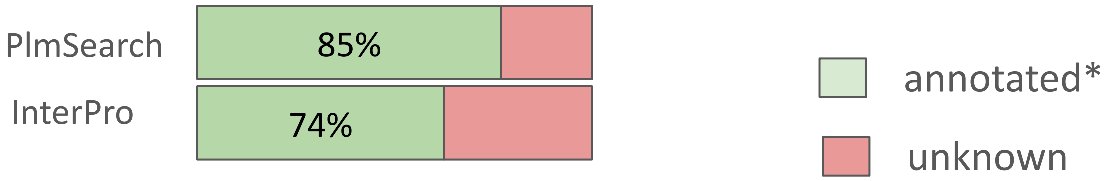
FASTA

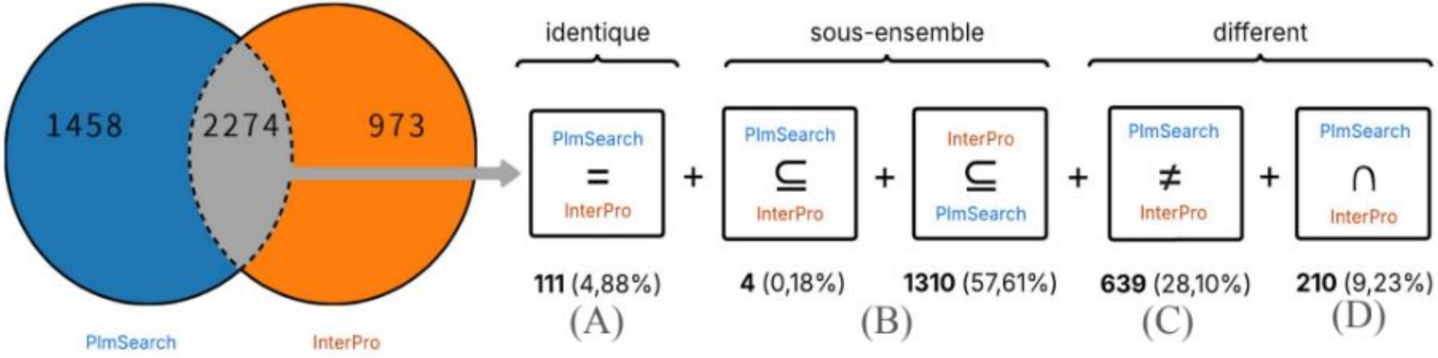
≥ 250 a.a.

~4.3k clusters

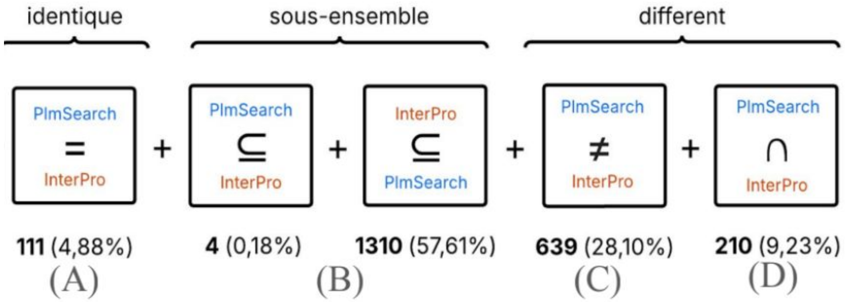


Results : comparing predictions





Results : comparing predictions



identique

sous-ensemble

different

PlmSearch = InterPro

+

PlmSearch ⊆ InterPro

+

InterPro ⊆ PlmSearch

+

PlmSearch ≠ InterPro

+

PlmSearch ∩ InterPro

111 (4,88%)

(A)

4 (0,18%)

1310 (57,61%)

(B)

639 (28,10%)

(C)

210 (9,23%)

(D)

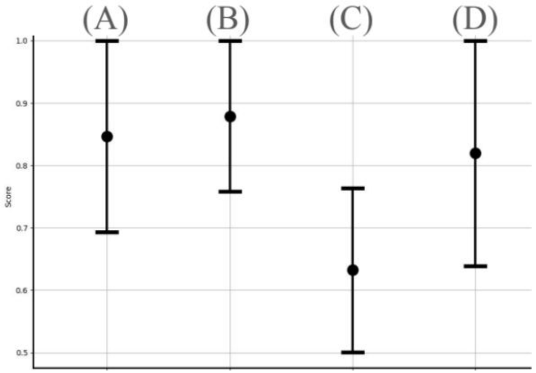
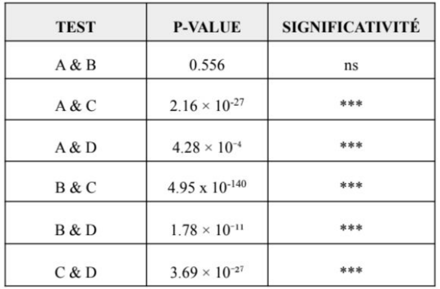


Tableau récapitulatif des tests statistiques de Mann-Whitney
*(ns : non significatif ; *** : < au risque α = 1‰)*



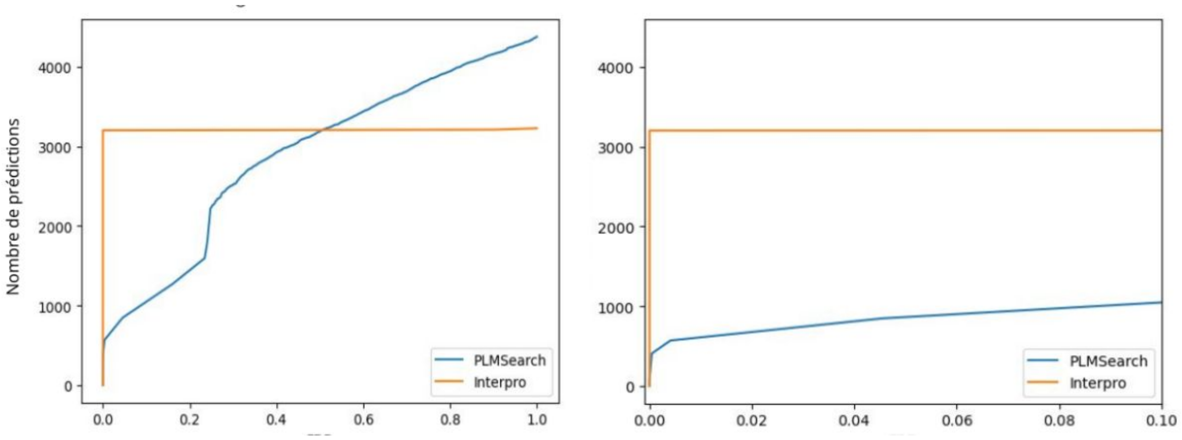
TEST	P-VALUE	SIGNIFICATIVITÉ
A & B	0.556	ns
A & C	2.16×10^{-27}	***
A & D	4.28×10^{-4}	***
B & C	$4.95 \text{ x } 10^{-140}$	***
B & D	1.78×10^{-11}	***
C & D	3.69×10^{-27}	***

Results : False discovery rates

- We shuffled the amino acid order of each sequence with original dataset (D) to create an artificial database (R)

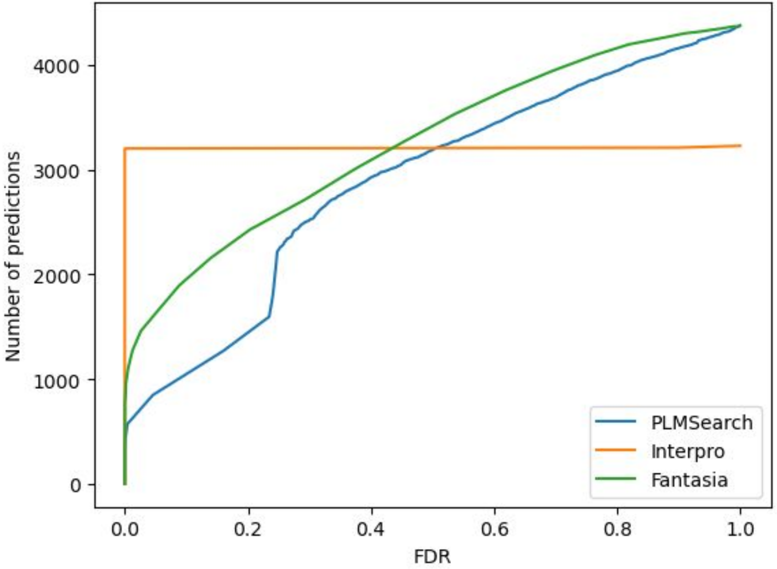
- FDR for a given score S is :

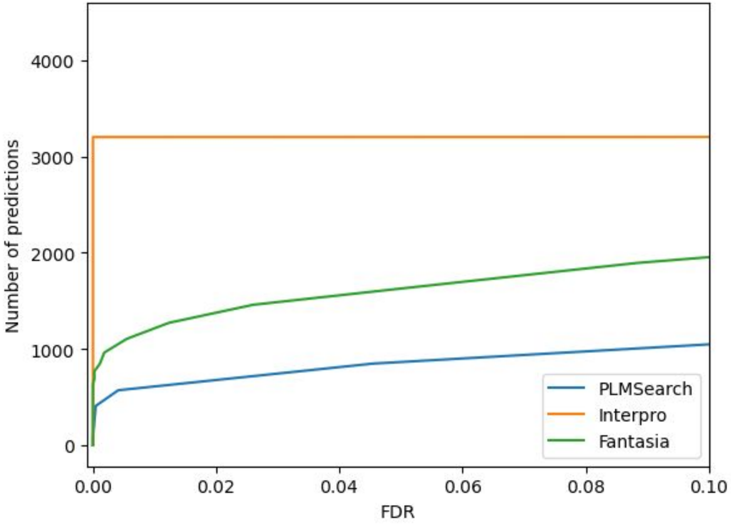
$$\frac{\text{number of predictions in R}}{\text{number of predictions in R + number of predictions in D}}$$



Results : False discovery rates

- We also tested Fantasia and measured FDR





Conclusion

- PLMSearch seems to complete/enrich the InterPro annotations.
- However, when it finds completely different annotations, the scores are significantly lower, indicating low accuracy.
- For a given FDR PLMSearch annotated less proteins than Interpro, the difference is even more evident for an FDR bounded to 10%.
- Higher FDR were also observed in Fantasia results.

Acknowledges

- Emile Hembert - L2 student - Sorbonne Université
- Dorian Le Roux - L2 student - Sorbonne Université
- Fabio RJ Vieira - IR - Sorbonne Université
- Laurence Garczarek - CNRS-DR
- Frédéric Partensky, CNRS-DR

Thanks for your attention !