Functional annotation of dinoflagellate protein sequences and structural prediction by deep learning approach

Jérémy Rousseau[1], Lucie Bittner[1,2], Mathilde Carpentier[1]

[1] Muséum National d'Histoire Naturelle – Atelier de Bio-Informatique, 45 rue Buffon, 75005, Paris, France
[2] Institut Universitaire de France, 1 rue Descartes, 75231, Paris, France
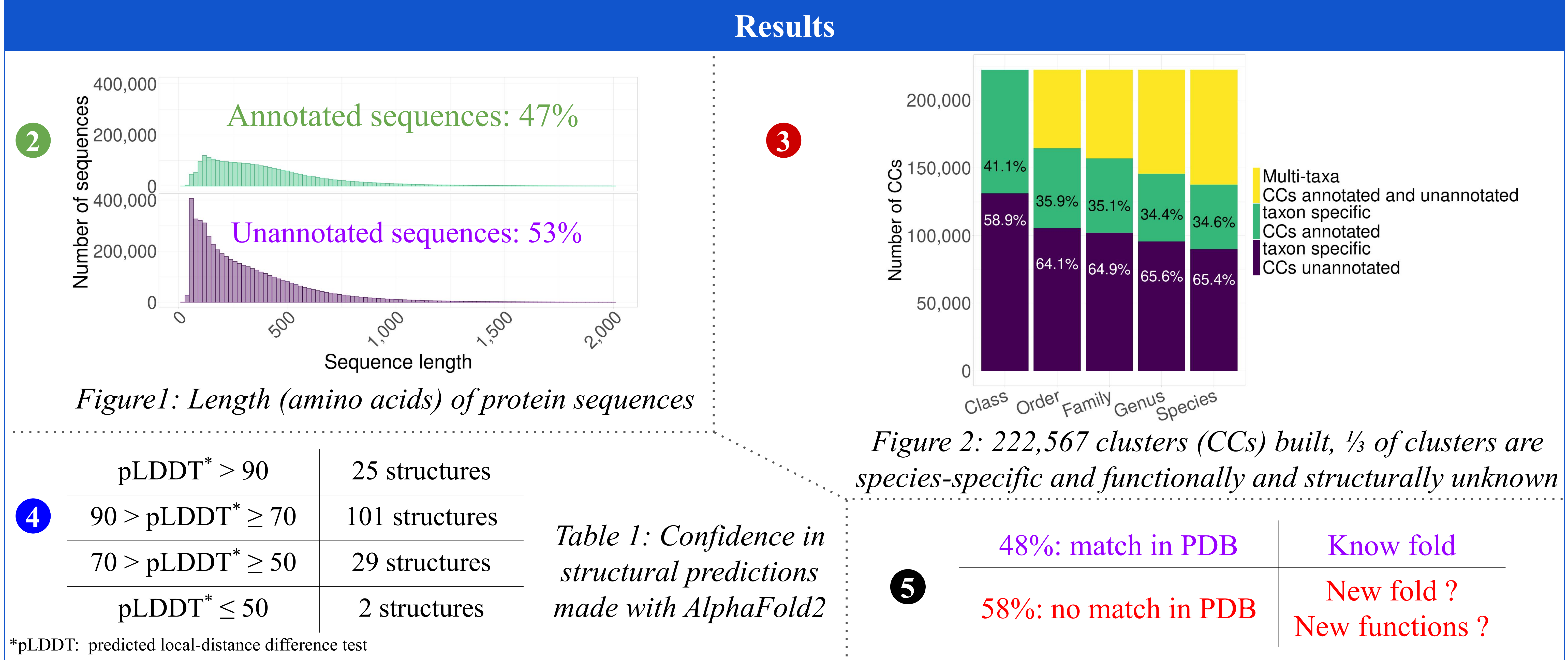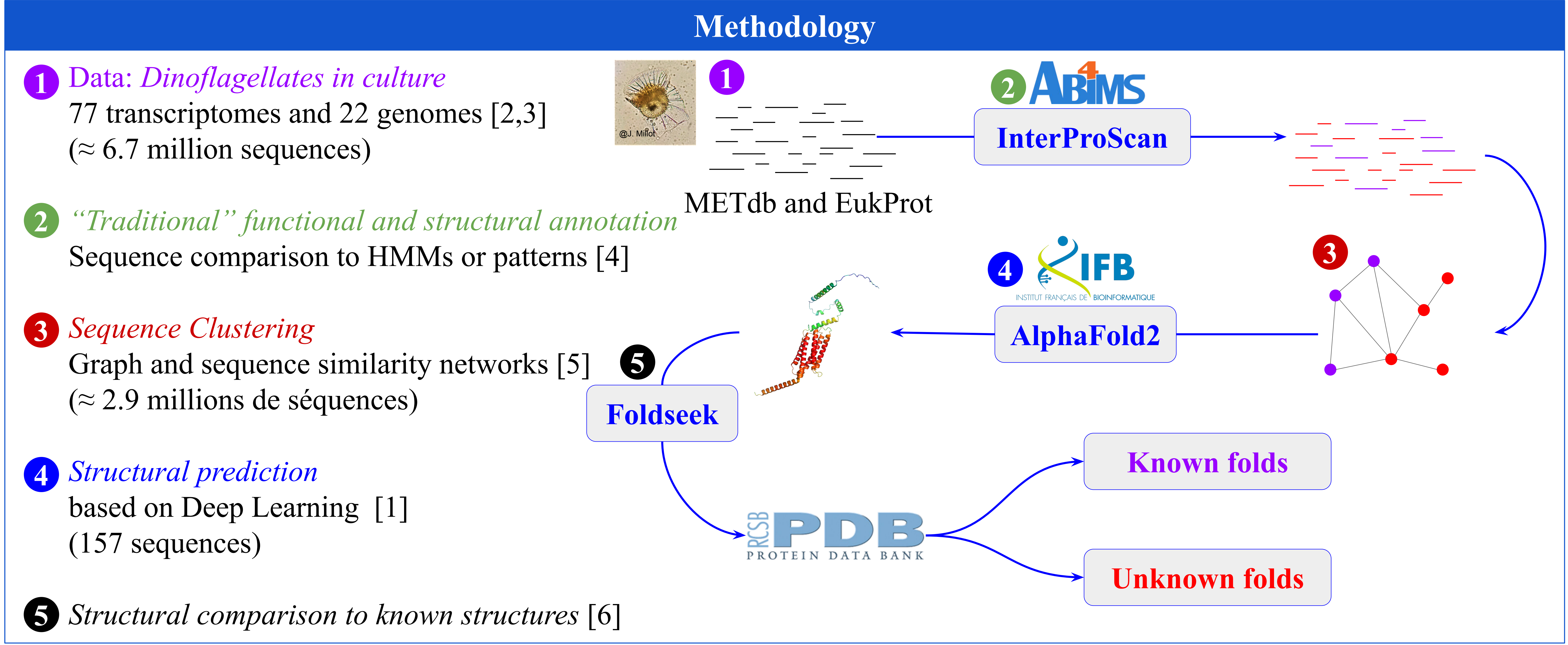
Corresponding Author: jeremy.rousseau2@outlook.fr

Introduction

Functional annotation of non-model organisms is a challenge. Protein structure help us to understand/predict the function of unknown proteins. Structure prediction was revolutionized in 2020 by AlphaFold2 [1], making it possible to obtain the structure of many proteins. We focused on dinoflagellates to develop a methodology for functional sequence annotation and explore structural innovations.

Goals

⇨ To develop a methodology for structural and functional annotation of functionally unknown genomic sequences.

⇨ To determine whether structural prediction is relevant for exploring structural innovation.

Methodology

① Data: *Dinoflagellates in culture*
77 transcriptomes and 22 genomes [2,3]
(≈ 6.7 million sequences)

② *"Traditional" functional and structural annotation*
Sequence comparison to HMMs or patterns [4]

③ *Sequence Clustering*
Graph and sequence similarity networks [5]
(≈ 2.9 millions de séquences)

④ *Structural prediction*
based on Deep Learning [1]
(157 sequences)

⑤ *Structural comparison to known structures* [6]

METdb and EukProt — InterProScan — AlphaFold2 — Foldseek — PDB RCSB PROTEIN DATA BANK — Known folds / Unknown folds

Results

② Annotated sequences: 47%
Unannotated sequences: 53%

Figure1: Length (amino acids) of protein sequences

③

Multi-taxa CCs annotated and unannotated taxon specific
CCs annotated taxon specific
CCs unannotated

41.1% / 58.9%
35.9% / 64.1%
35.1% / 64.9%
34.4% / 65.6%
34.6% / 65.4%

Class · Order · Family · Genus · Species

Figure 2: 222,567 clusters (CCs) built, ⅓ of clusters are species-specific and functionally and structurally unknown

④

pLDDT* > 90	25 structures
90 > pLDDT* ≥ 70	101 structures
70 > pLDDT* ≥ 50	29 structures
pLDDT* ≤ 50	2 structures

*pLDDT: predicted local-distance difference test

Table 1: Confidence in structural predictions made with AlphaFold2

⑤

48%: match in PDB	Know fold
58%: no match in PDB	New fold ? New functions ?

Conclusion/Perspectives

In dinoflagellates, only half the sequences can be annotated functionally and/or structurally. Moreover, the use of graphs enables sequences to be grouped together, thus reducing the number of unknown sequences.
Structural prediction is an interesting approach for studying potential innovations.

References and Acknowledgments

[1] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, Art. no. 7873, Aug. 2021, doi:10.1038/s41586-021-03819-2.
[2] G. Niang *et al.*, "METdb: A GENOMIC REFERENCE DATABASE FOR MARINE SPECIES," *F1000Research*, vol. 9, Jun. 2020, doi: 10.7490/f1000research.1118000.1.
[3] D. J. Richter *et al.*, "EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes," *Peer Community J.*, vol. 2, 2022, doi: 10.24072/pcjournal.173.
[4] P. Jones *et al.*, "InterProScan 5: genome-scale protein function classification," *Bioinforma. Oxf. Engl.*, vol. 30, no. 9, pp. 1236–1240, May 2014, doi: 10.1093/bioinformatics/btu031.
[5] E. Faure *et al.*, "Mixotrophic protists display contrasted biogeographies in the global ocean," *ISME J.*, vol. 13, no. 4, Art. no. 4, Apr. 2019, doi: 10.1038/s41396-018-0340-5.
[6] H. M. Berman *et al.*, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/nar/28.1.235.