What is hidden in the darkness? A large-scale approach to make sense of all natural unknown proteins

Joana Pereira, Torsten Schwede
Biozentrum and SIB Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Switzerland

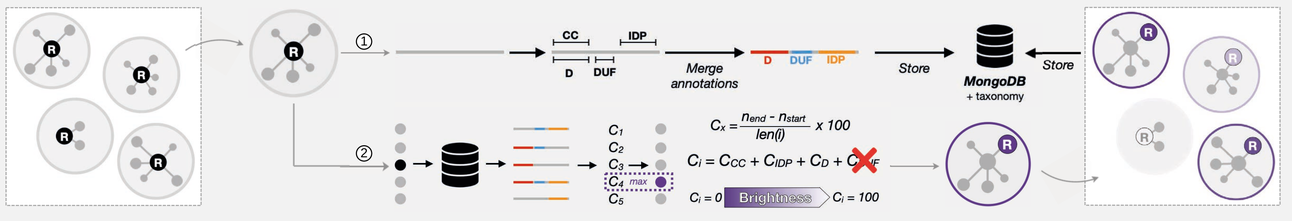
Large-scale genomic projects are promoting an exponential increase in the number of protein sequences deposited in protein repositories every year and the number of **“hypothetical proteins” and “proteins of unknown function”** is increasing proportionally. This can be due to:

- low sensitivity of the methods behind their annotationn and classificationn
- the presence of sequences belonging to novel not hitherto described biological systems

`How many represent never-before-seen protein families? How much novelty is hidden in these ”dark” proteins?`

How we define “brightness”:

The brightness of a protein sequence corresponds to the **full-length coverage with annotations (domains, predicted disorder and coiled coils)** of the best annotated sequence in the same UniRef50 cluster.

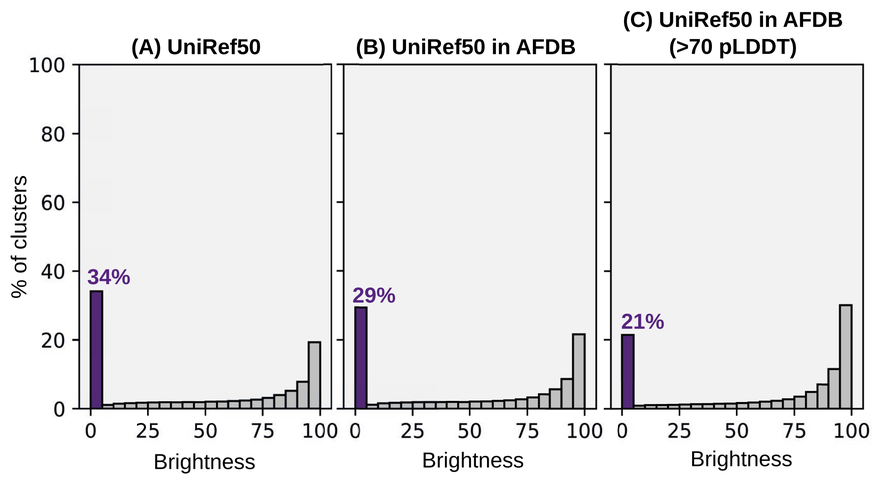


A large fraction of UniRef50 and UniProt remains in the dark

We now found that:

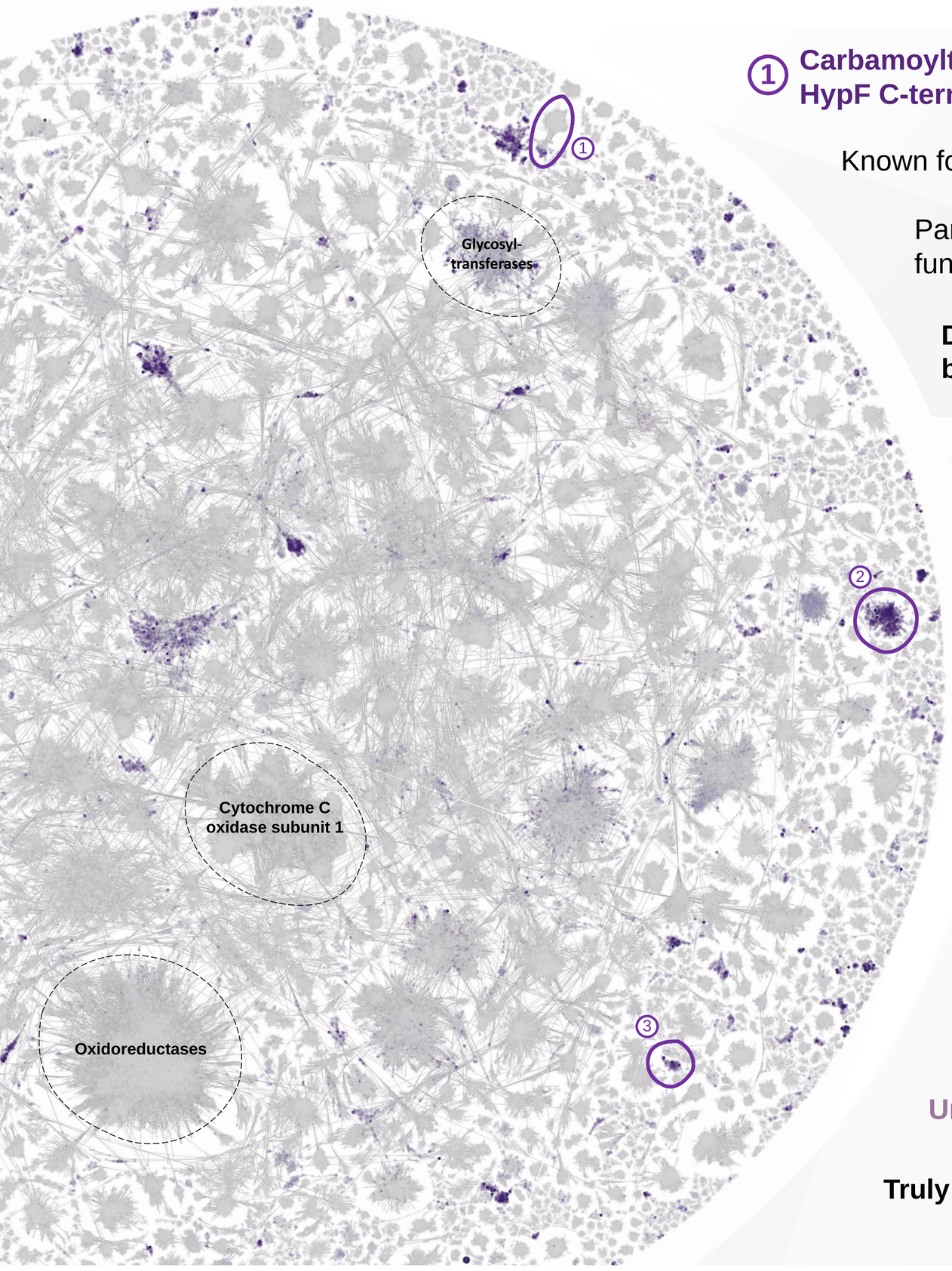
1. **34% of all UniRef50 clusters** are made of sequences with less than 5% of their full sequence annotated.
2. This corresponds to **10% of all non-redundant UniProt and UniParc sequences**.
3. The same proportion of **darkness is found in the AlphaFold database v3**.
4. Most clusters are small in size, but some with >1000 sequences were identified.
5. These sequences are widespread and the most common are from **marine sediment metagenomes**.

`In all proteomes, there are pitch dark proteins whose folds are predicted with very high confidence!`



These proteins are dispersed throughout the protein universe

We compared all those UniRef50 representatives with a pLDDT > 95 (~1 million UniRef50 entries), modelling their sequence landscape as a protein sequence similarity network. Pitch dark proteins correspond either (1) to **fully dark galaxies**, or (2) to **points on the edges of bright galaxies**.

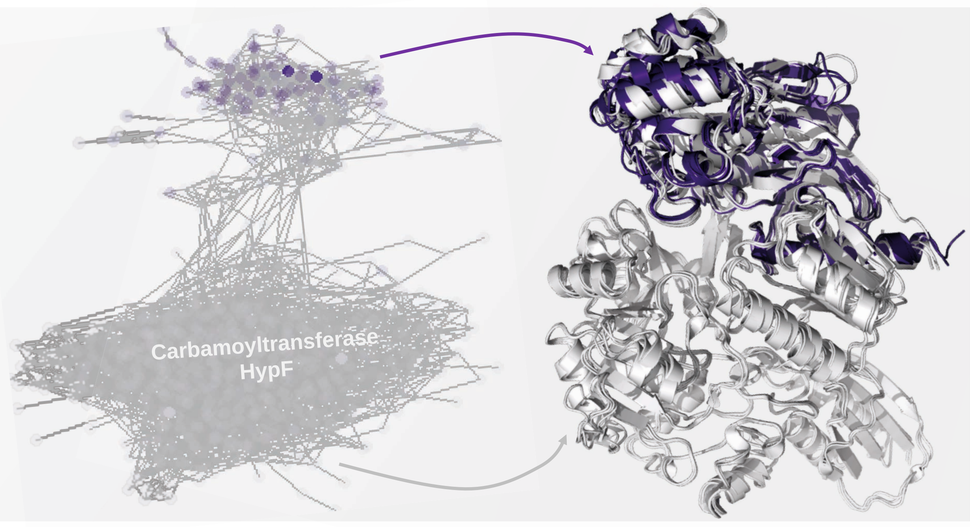


① Carbamoyltransferase HypF C-terminal-like

Known fold

Partially known function

Divergent form of bright family

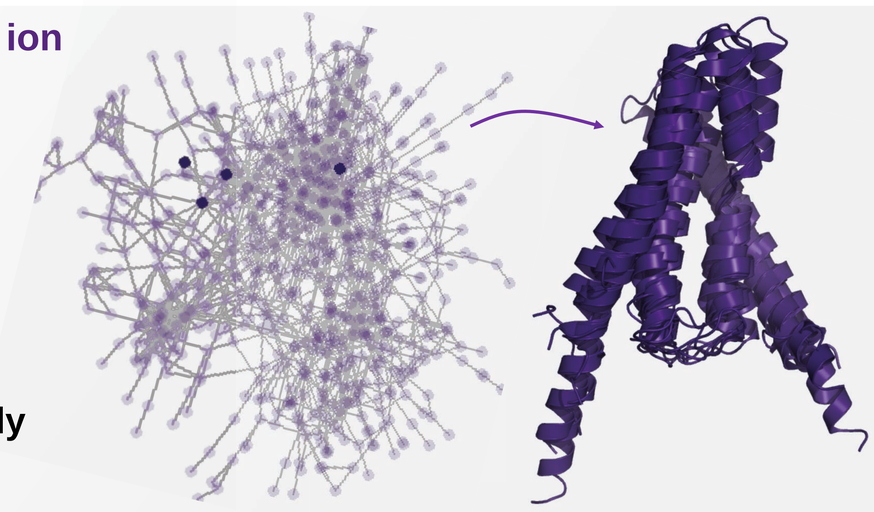


② Putative fluoride ion transporter CrcB

Unknown fold

Known function

Partially dark family



③ Uncharacterized protein

Unknown fold

Unknown function

Truly dark family

Where are we going from this:

- Expand the landscape to all catalogued proteins and explore different distance representations.
- Study how different sequence and structure features shape the local and overall structure of the modelled landscape.
- Make all of this available as an interactive **Protein Universe Atlas** that can be used to navigate through the darkness of the protein universe.

`Check also`

Janani Durairaj *et al.*

“Characterization of rare and novel AlphaFold structural space”

joana.pereira@unibas.ch @Joana_MPereira