Cross-phyla protein annotation by structural prediction and alignment

F Ruperti, N Papadopoulos, JM Musser, M Mirdita, M Steinegger, and D Arendt

Genome Biology, 2023

BCBB Omics Journal Club

Apr 1 2025

Presenter: Margaret Ho

Mmseqs2 + colabfold

FoldSeek

Ruperti *et al. Genome Biology* (2023) 24:113
https://doi.org/10.1186/s13059-023-02942-9

Genome Biology

Uncharacterized proteins
Hypothetical proteins

RESEARCH Open Access

Cross-phyla protein annotation by structural prediction and alignment

Fabian Ruperti[1,2†], Nikolaos Papadopoulos[1,3†], Jacob M. Musser[1], Milot Mirdita[4], Martin Steinegger[4,5] and Detlev Arendt[1,6*]

Porifera - freshwater sponge – early diverged metazoan

Spongilla lacustris

Svedberg *et al. BMC Genomics* (2024) 25:6
https://doi.org/10.1186/s12864-023-09924-y

BMC Genomics

RESEARCH Open Access

Functional annotation of a divergent genome using sequence and structure-based similarity

Dennis Svedberg[1,2†], Rahel R. Winiger[1†], Alexandra Berg[1,2†], Himanshu Sharma[1,2], Christian Tellgren-Roth[3], Bettina A. Debrunner-Vossbrinck[4], Charles R. Vossbrinck[5] and Jonas Barandun[1*]

Microsporidia - unicellular obligate intracellular parasites related to fungi

Found to infect almost all taxa of animals – insects, crustaceans, fish, humans

Vairimorpha necatrix

MorF: morpholog finder workflow

https://git.embl.de/grp-arendt/MorF

- Experimentally determined knowledge is only available to a few model orgs
- For non-model species, sequence-based prediction of gene orthology can be used to infer protein identity, but this approach loses predictive power at longer evolutionary distances
- Ruperti et al. propose a **MorF workflow for protein annotation using structural similarity**, exploiting that **similar protein structures often reflect homology and are more conserved that protein sequences**

mmseqs2 is super fast k-mer based similarity sequence alignment and clustering

- uses k-mer similarity for really good prefilter to reduce search space + clever heuristics

Mmseqs2 + colabfold

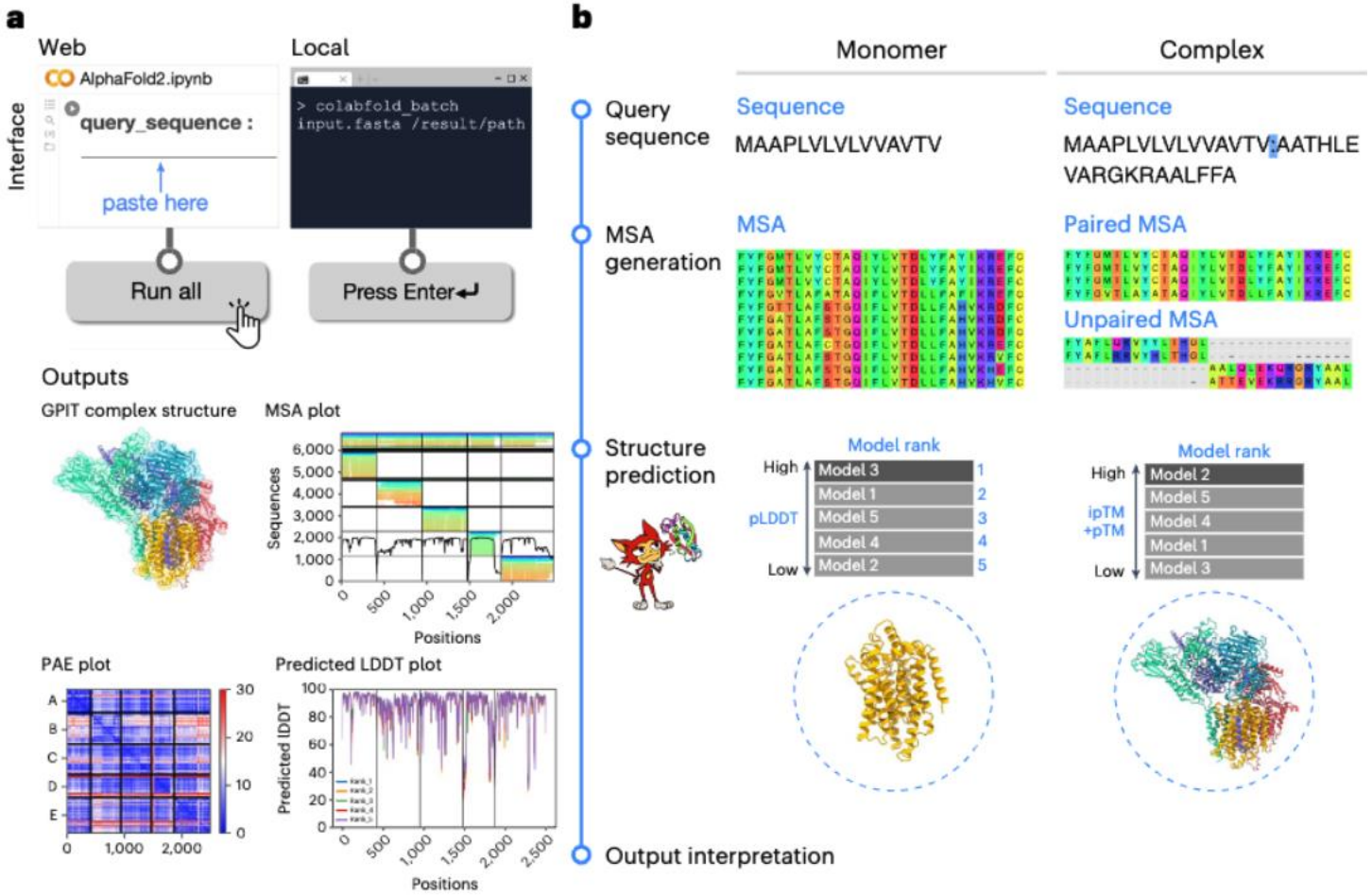
https://github.com/steineggerlab/colabfold-protocol

FoldSeek

https://github.com/steineggerlab/foldseek

(1) CoLabFold for protein structure prediction

Automates MSA generation and running of AlphaFold2 to get protein structures



Upon receiving input protein sequence, ColabFold-AF2:

1) **generates MSA using MMseqs2** against UniRef30, PDB databases

2) **performs structure prediction using Alphafold2** models and generates structure models, which are iteratively refined and ranked by confidence scores

pLDDT = per-residue measure of local confidence

ipTM = accuracy of the predicted relative positions of the subunits forming protein-protein complex

Kim, G., Lee, S., Levy Karin, E., Kim, H., Moriwaki, Y., Ovchinnikov, S., Steinegger, M., & Mirdita, M. (2025). Easy and accurate protein structure prediction using ColabFold. *Nature Protocols*, *20*(3), 620–642.

Great CoLabFold talk from M Steinegger and S Ovchinnikov
https://www.youtube.com/watch?v=Rfw7thgGTwI

CoLabFold pipeline on Skyline

Automates MSA generation and running of AlphaFold2 to get protein structures

```
(base) [homc@ai-hpcsubmit1 ~]$ module load colabfold
-----------------------------------------------------------

This module creates shell functions to help run programs
available inside this container using 'apptainer exec'

This container requests a GPU using the --nv flag and
binds the colabfold cache at /data/bio_db/colabfold_cache as
'/cache' inside the container

Available Commands:
colabfold_batch, colabfold_search, colabfold_split_msas
-----------------------------------------------------------
```

CoLabFold on Biowulf as well
https://hpc.nih.gov/apps/colabfold.html

Colabfold_batch automatically submits queries to the public MSA server, but rate limited

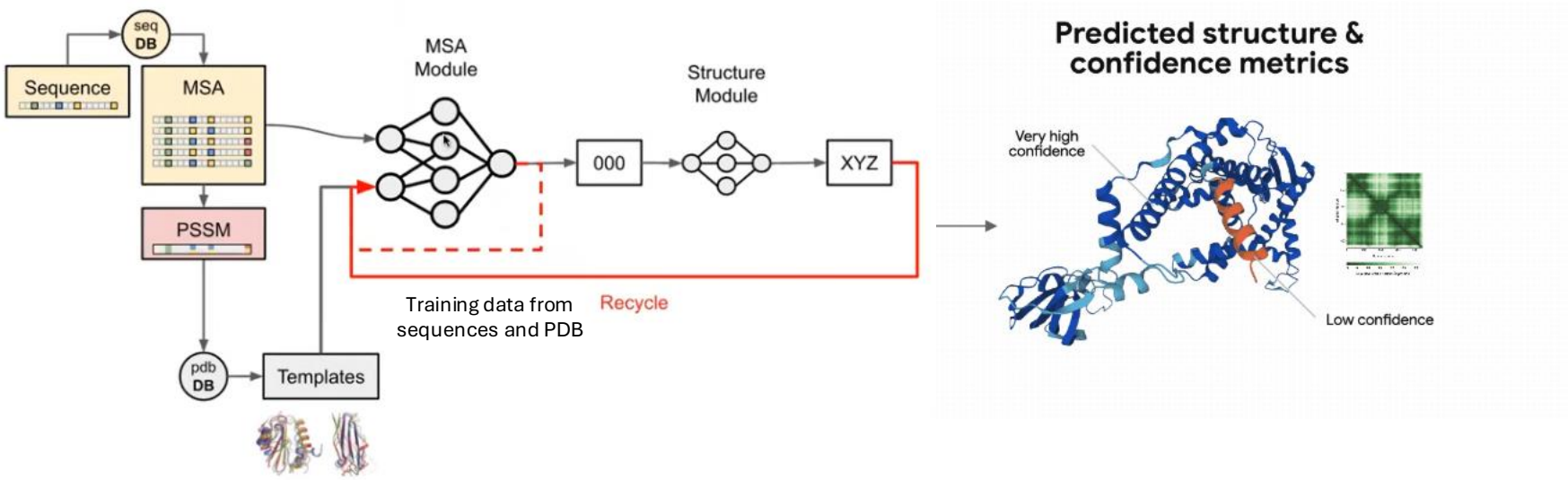
```
(base) [homc@ai-hpcsubmit1 Mar2025_proteinstructpred]$ colabfold_batch Tritrichomonas_casperi_120423.HYPOTHETICALPROTEIN.fa colabfold_results/
2025-03-28 18:42:05,408 Running colabfold 1.5.3

WARNING: You are welcome to use the default MSA server, however keep in mind that it's a
limited shared resource only capable of processing a few thousand MSAs per day. Please
submit jobs only from a single IP address. We reserve the right to limit access to the
server case-by-case when usage exceeds fair use. If you require more MSAs: You can
precompute all MSAs with `colabfold_search` or host your own API and pass it to `--host-url`
```

Colabfold_search runs with mmseqs2 to generate MSAs locally

AlphaFold2 summary

Alphafold2 release in 2021
Since then, there have been newer additions like AlphaFold-Multimer, and **AlphaFold3 (2024)**

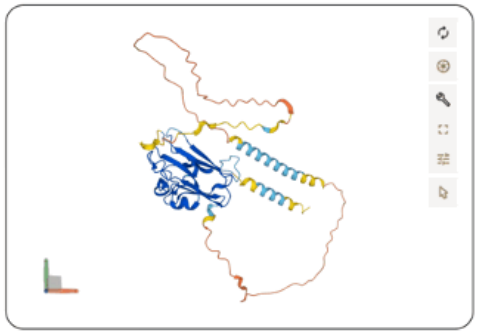


AlphaFold2 predicts	AlphaFold2 struggles to predict	AlphaFold2 doesn't predict
• Single protein chains	• Multiple conformations for the same sequence	• Protein-DNA and protein-RNA complexes
• Protein multimers	• Effects of point mutations	• Nucleic acid structure
• Multisubunit protein-protein complexes	• Antigen-antibody interactions	• Ligand and ion binding
		• Post-translational modifications
		• Membrane plane for transmembrane domains

Figure from https://www.youtube.com/watch?v=Rfw7thgGTwI

https://www.ebi.ac.uk/training/online/courses/alphafold/an-introductory-guide-to-its-strengths-and-limitations/strengths-and-limitations-of-alphafold/

AlphaFold Confidence Scores

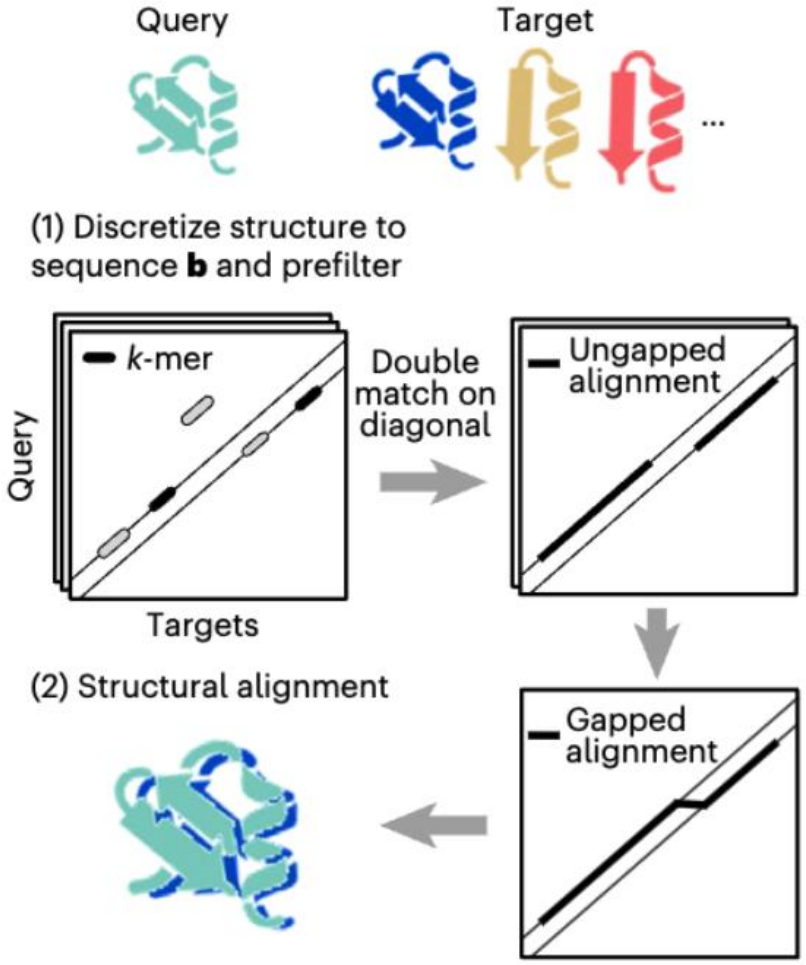


- **pLDDT** - **per-residue measure of local confidence**
  - [predicted local distance difference test]
  - If pLDDT below 50
    - Could be due to highly flexible or intrinsically disordered structure
    - Alphafold may just not have enough information to predict with confidence
  - **High pLDDT indicates well-predicted, stable regions, important for homology detection**
- **ipTM** - **accuracy of predicted relative positions of subunits forming protein-protein complex**
  - [interface predicated template modeling]
  - **Confidence of interfaces between subunits or domains, which are important for homology detection**
- **PAE** – measure of confidence in relative position of 2 residues within predicted structure [predicted aligned error]
  - **how confident AlphaFold is that domains are well packed** and that relative placement of the domains in the predicted structure is correct
- **PTM** – integrated measure **of how well predicted overall structure of complex** [predicted template modeling]
  - Predicted TM score for superposition between predicted and hypothetical true structure
  - TM above 0.5 means overall predicted fold for complex similar to true structure
  - TM below 0.5 means predicted structure likely wrong

pLDDT and ipTM are the most relevant metrics for structure-based homology

Great explainer on Alphafold by EMBL-EBI: https://www.ebi.ac.uk/training/online/courses/alphafold/

(2) FoldSeek for protein structure search



- Foldseek enables fast and sensitive comparisons of large protein structure sets
  - supporting monomer and multimer searches, and clustering
  - runs on CPU, supports GPU acceleration for faster searches

- FoldSeek is over 4000 times faster than TM-align
  - Key insight is using a structural alphabet not for AA backbone (as other methods do) but tertiary interactions
  - 20 states of 3D interaction (3Di) alphabet that describes for given residue the geometric conformation with its spatially closest residue
  - Weaker dependency b/w consecutive AA letters
  - Highest information density between conserved protein cores and lowest in non-conserved coil/loop regions

- In studies of human proteins, FoldSeek identified structural homology with <10% sequence identity across a range of species, including fungi, plants, and animals

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, *42*(2), 243–246. https://doi.org/10.1038/s41587-023-01773-0

(3) “Best hit” FoldSeek input into EggNOG mapper for annotation

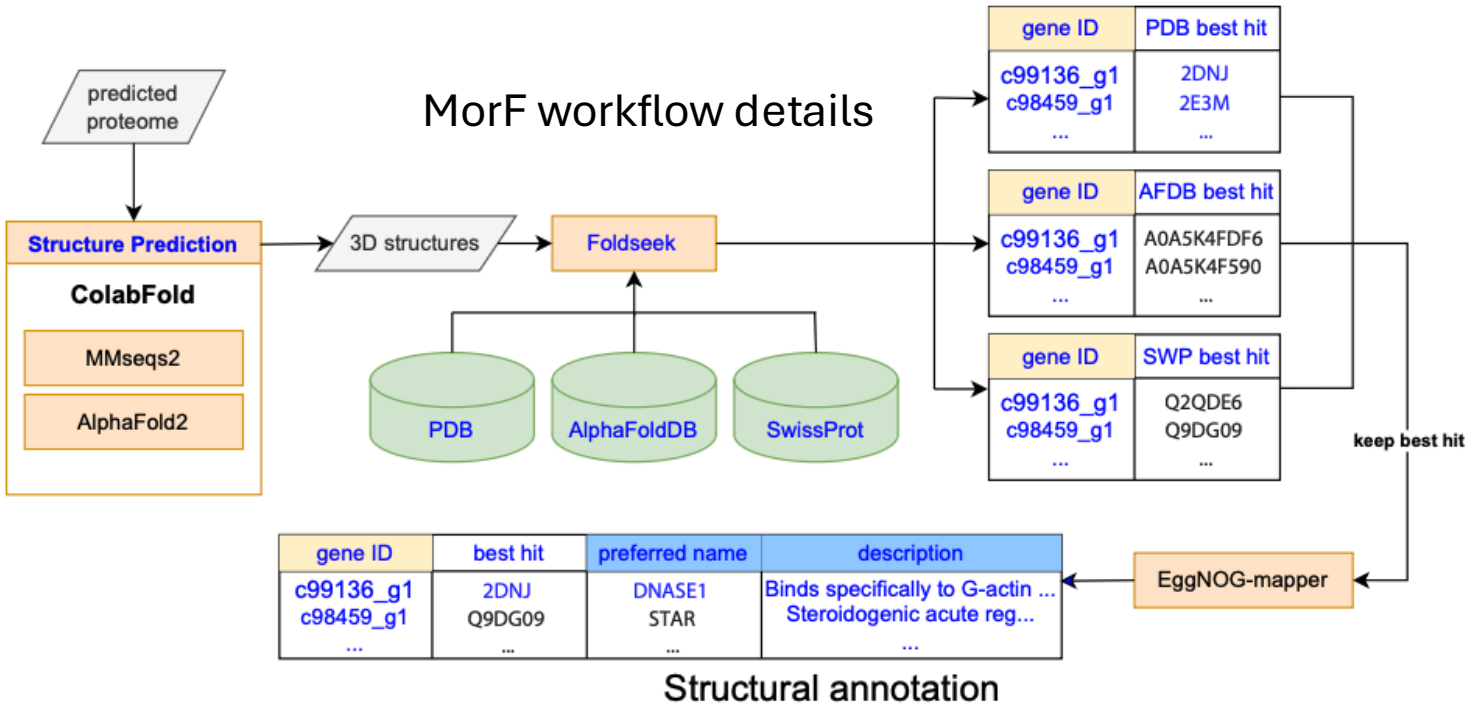
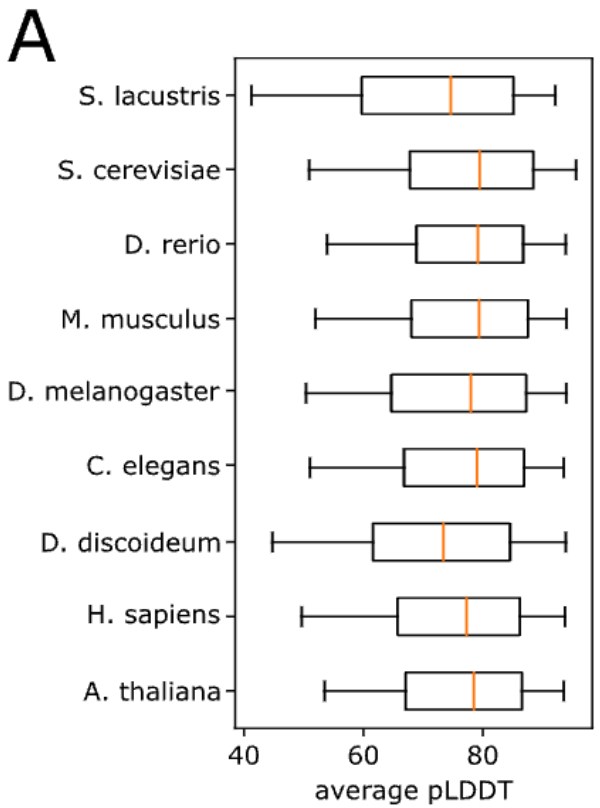
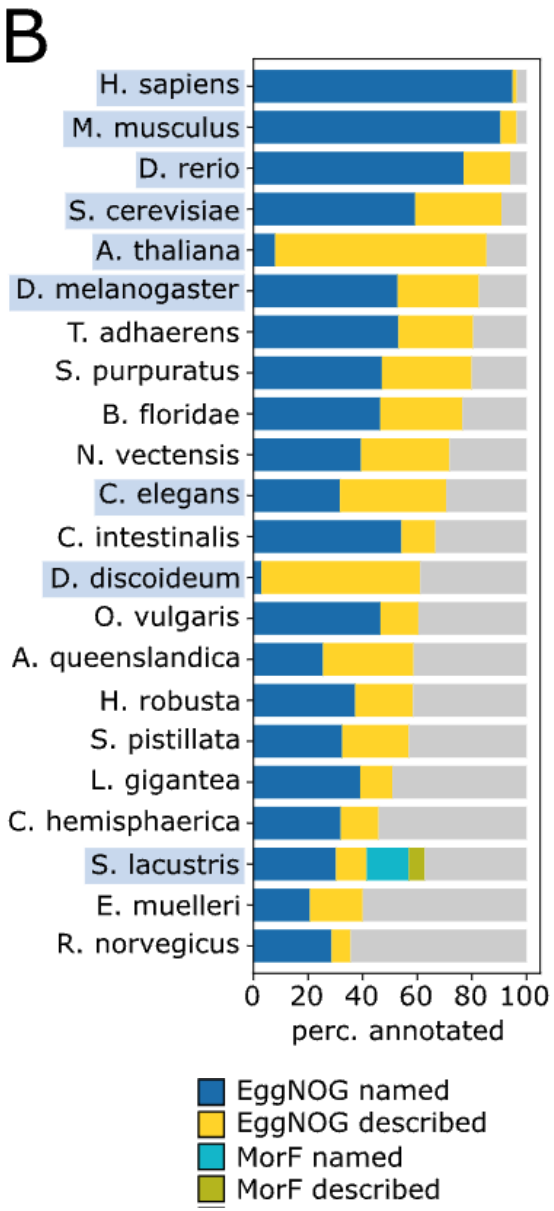


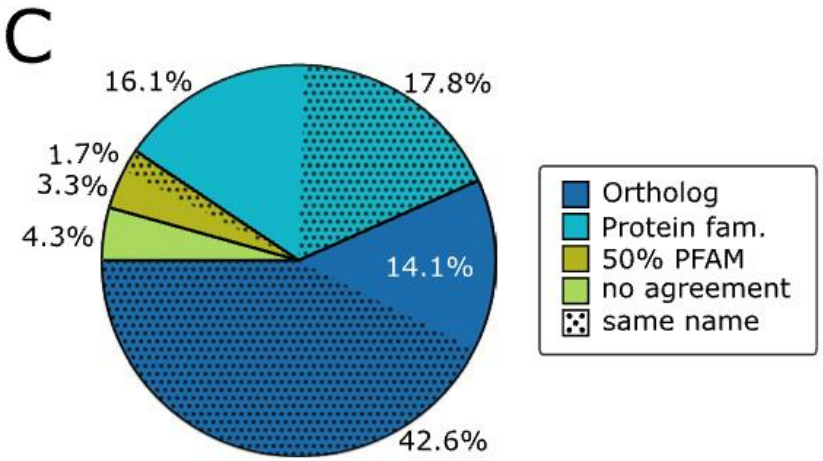
Figure S1 Overview of the MorF workflow ColabFold is called with a proteome as input. It produces multiple sequence alignments for each protein using MMseqs2 and predicts structures for each structure using AlphaFold2. The best model per protein is kept and used as query in Foldseek to search against PDB, AlphaFoldDB, and SwissProt. This creates an m8-formatted table for each database, where the significant hits for each query protein are kept. The tables are merged by keeping the target with the highest bit score (= highest scoring morpholog). The sequence of the morpholog is then retrieved from UniProt and used as a query for emapper. This finds the target sequence itself and retrieves the EggNOG entry, including preferred name and description.

MorF annotation with *S. lacustris* (sponge)



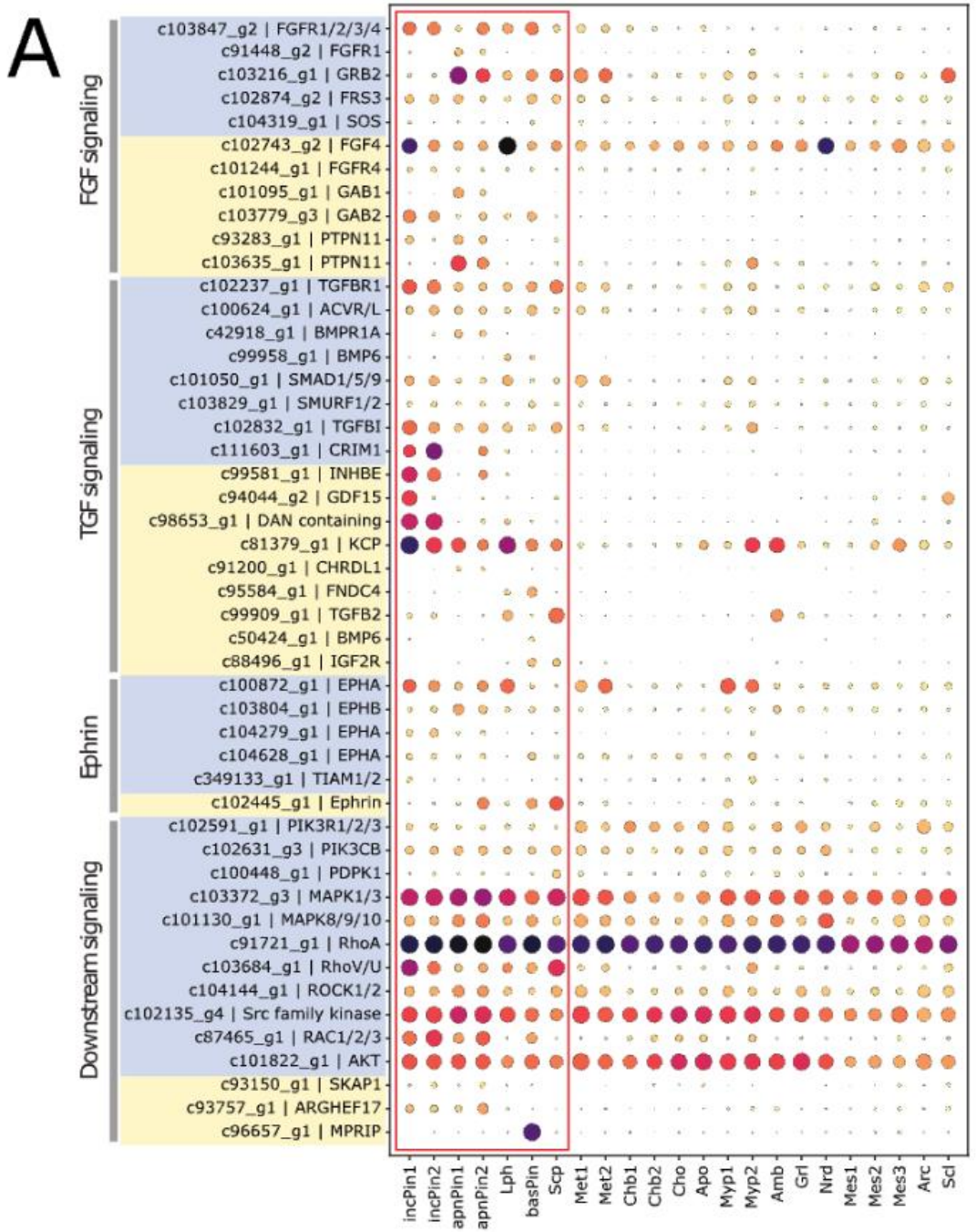
S. lacustris has lower average pLDDT (4-6% lower) than of well characterized animal models

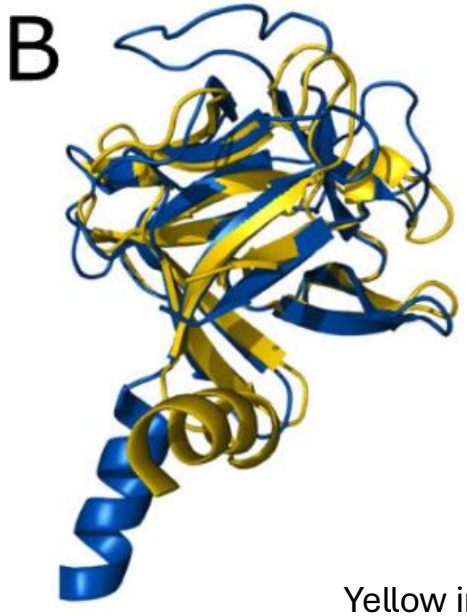




Adding MorF annotations increased annotation by 50%, and pie chart shows agreement between EGGNOG, PFAM, and MorF

Ruperti et al. Genome Biology 2023

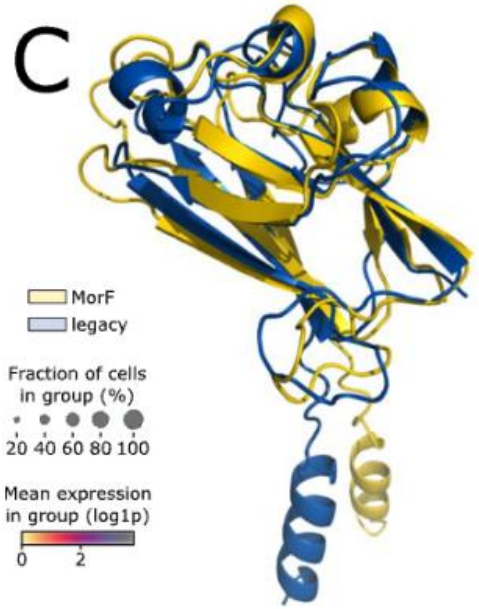




Spongilla FGF (blue) and chicken FGF4 (yellow)

Sequence identity of 11.8%

Yellow indicates new members of protein families discovered (blue are legacy annotations based on sequence)



Spongilla ephrin (blue) and *C. elegans* efn-3

Sequence identity of 22%

Ruperti et al. Genome Biology 2023

MorF approach in sponges - summary

- MorF approach accurately predicted protein function with known homology in 90% of cases, and **annotates additional 50% of the proteome beyond standard sequence-based methods**
  - Showed new functions for sponge cell types including FGF, TGF, ephrin signaling in epithelial cells and redox metabolism in myopeptidocytes. Found new genes specific to mesocytes and propose they may function to digest cell walls

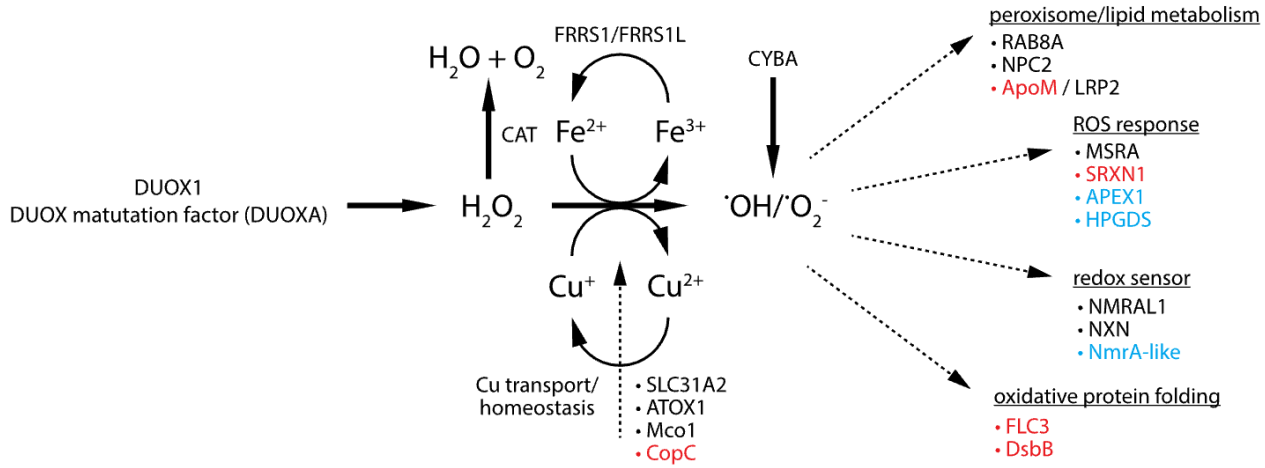
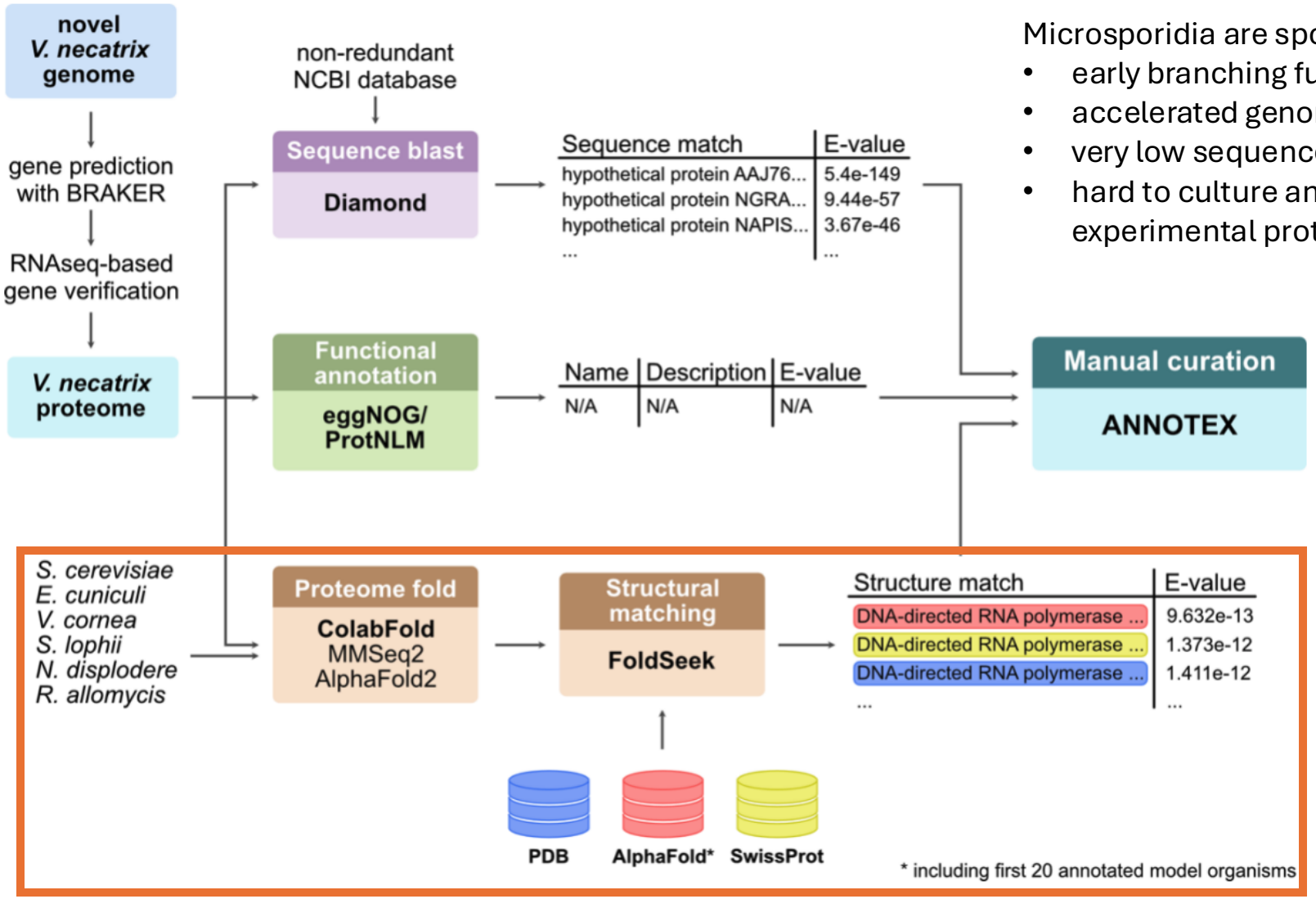


Fig. 3 ROS metabolism and redox-control in myopeptidocytes. Myopeptidocytes differentially express multiple genes involved in redox control and ROS defense. Genes in black have been annotated using sequence based methods. Blue proteins have protein family level sequence-based annotation with updated functions inferred by MorF. Genes in red have been functionally annotated using MorF

Ruperti et al. Genome Biology 2023

Another example: similar approach in microsporidia



Microsporidia are spore-forming unicellular parasites / fungi

- early branching fungal kingdom
- accelerated genome evolution due to obligate parasite lifestyle
- very low sequence similarity to other fungi and within clade
- hard to culture and genetically manipulate, AT-rich + codon bias; experimental protein characterization in E. coli / S. cerevisiae is tricky

Svedberg, D et al (2024). Functional annotation of a divergent genome using sequence and structure-based similarity. *BMC Genomics*, *25*(1), 6. https://doi.org/10.1186/s12864-023-09924-y

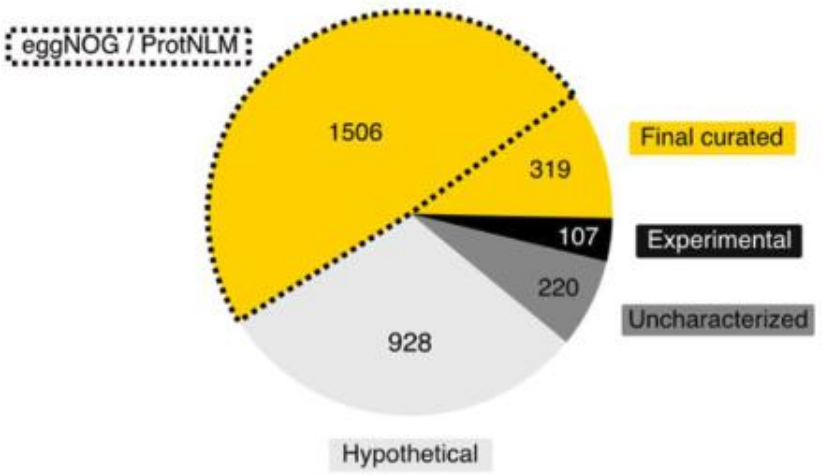
Added value of structural homology-based annotation

- Compared to eggNOG / ProtNLM (50% annotated), additional 319 genes annotated by structural approach
  - 92% of annotated genes confirmed by RNA reads
- Uncharacterized = genes conserved in several microsporidian species
- Hypothetical = genes only conserved within order Nosematida
  - RNA reads covered 87% of hypothetical genes, suggesting that they are actually expressed

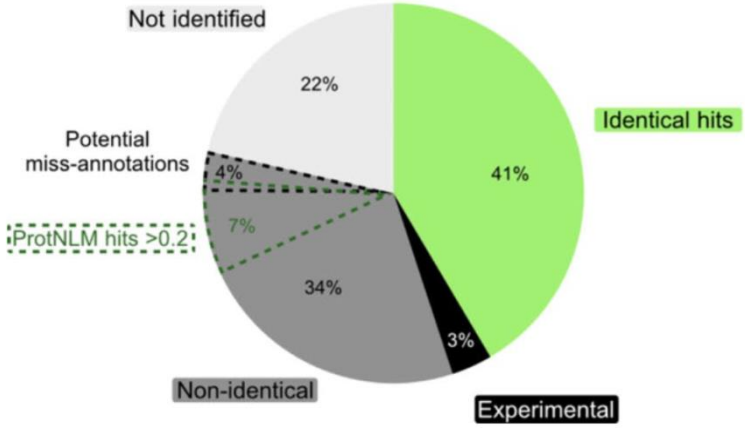
- 41% of ProtNLM hits have identical match in structure based homology annotation

- Manual investigation identified a few potential misannotations within ProtNLM, but some ProtNLMs that are non-identical have esp high score

Final Annotation



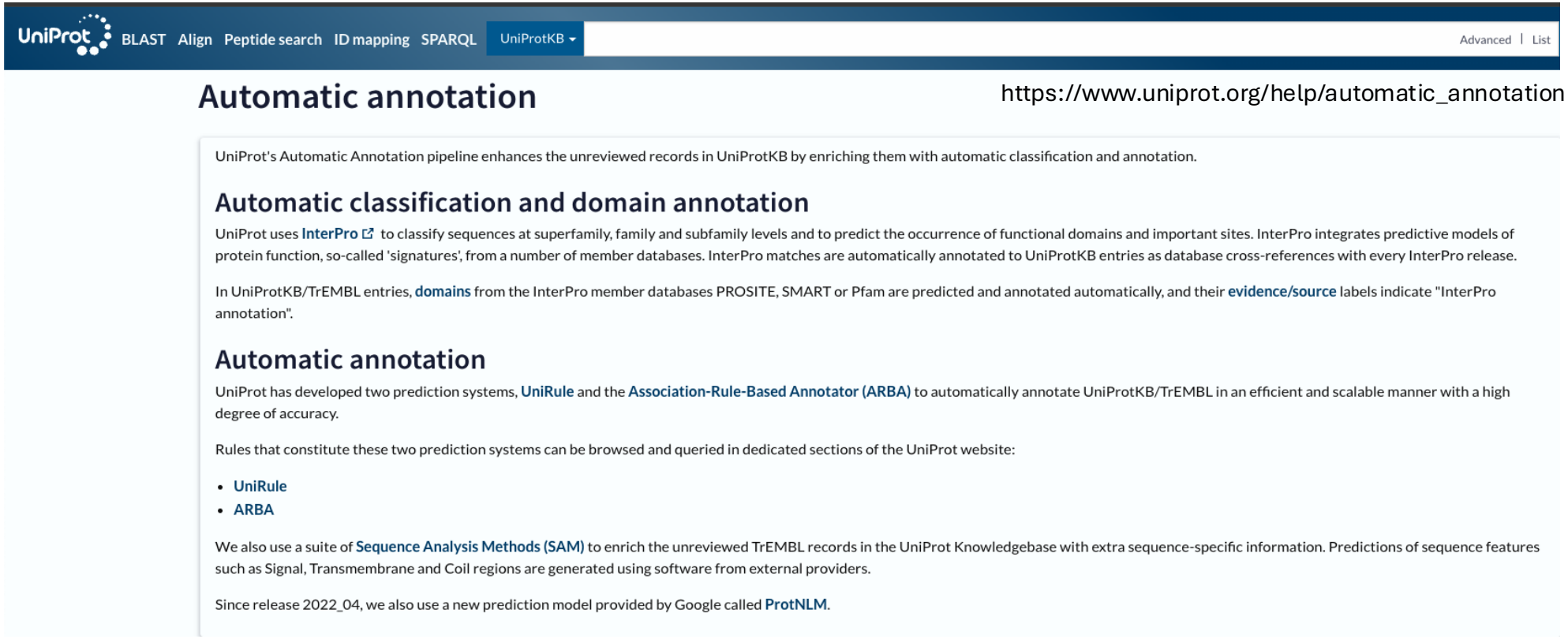
Comparison with ProtNLM



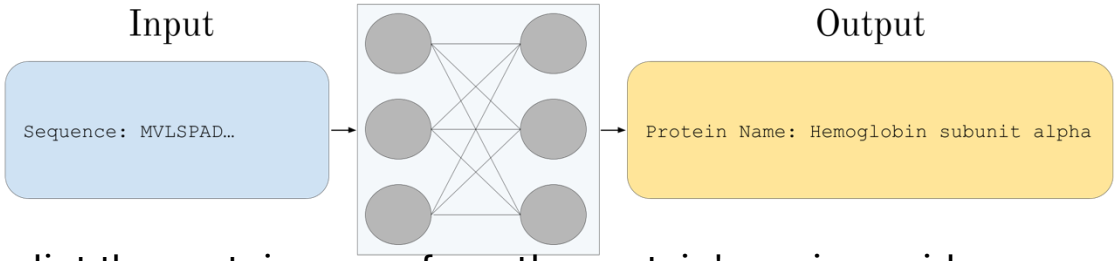
Svedberg et al. BMC Genomics 2024

UniProt incorporates ProtNLM

in addition to usual rule-based classification methods UniRule and ARBA



Original ProtNLM model in 2022
(developed by Google)

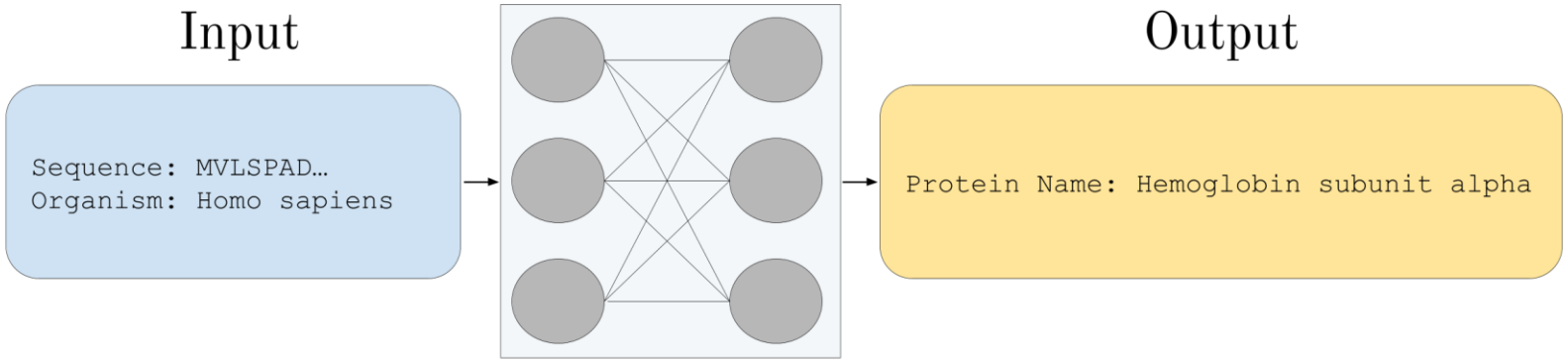


ProtNLM is a transformer model trained to predict the protein name from the protein's amino acid sequence. This method works by predicting a short textual description for proteins based solely on their amino acid sequence, using a sequence-to-sequence model.

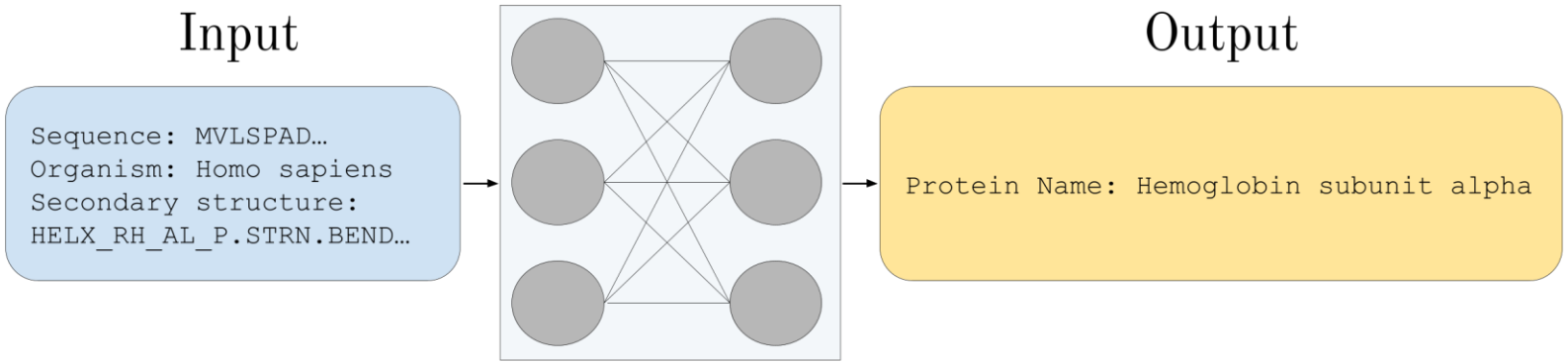
https://www.uniprot.org/help/ProtNLM

ProtNLM now also incorporates sequence alignment and predicted secondary structure for corroboration

In UniProt 2022_04, all ProtNLM annotations were produced by a single model of this form.



Later, they consider models that additionally take as input the protein's secondary structure, which we extract from the predicted AlphaFold structure included in UniProt.

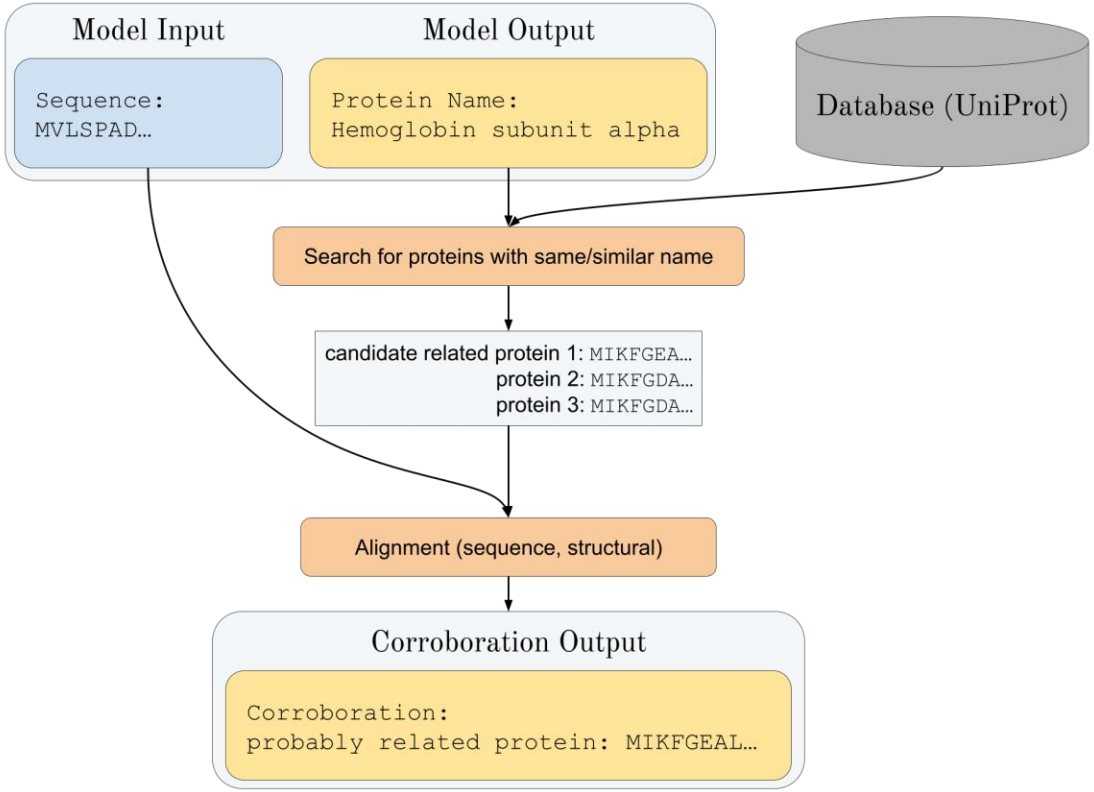


Google's preprint on ProtNLM here

https://www.uniprot.org/help/ProtNLM

ProtNLM now also incorporates sequence alignment and predicted secondary structure for corroboration

2023 ProtNLM
Ensemble Model



"Starting with UniProt 2023_02, **when UniProt supplies an AlphaFold prediction for a given accession, we also include in the ensemble a model that takes the amino acid sequence, organism and predicted secondary structure as inputs**. The 7 models that form the new ensemble were re-trained from scratch on the original training data, from which we also held out a small amount of random sequences for in-distribution validation. To incorporate new biocurator feedback and recent updates to some ground truth names in UniProt, we also fine-tuned the models on updated training data with additional protein name filtering proposed by the biocurators."

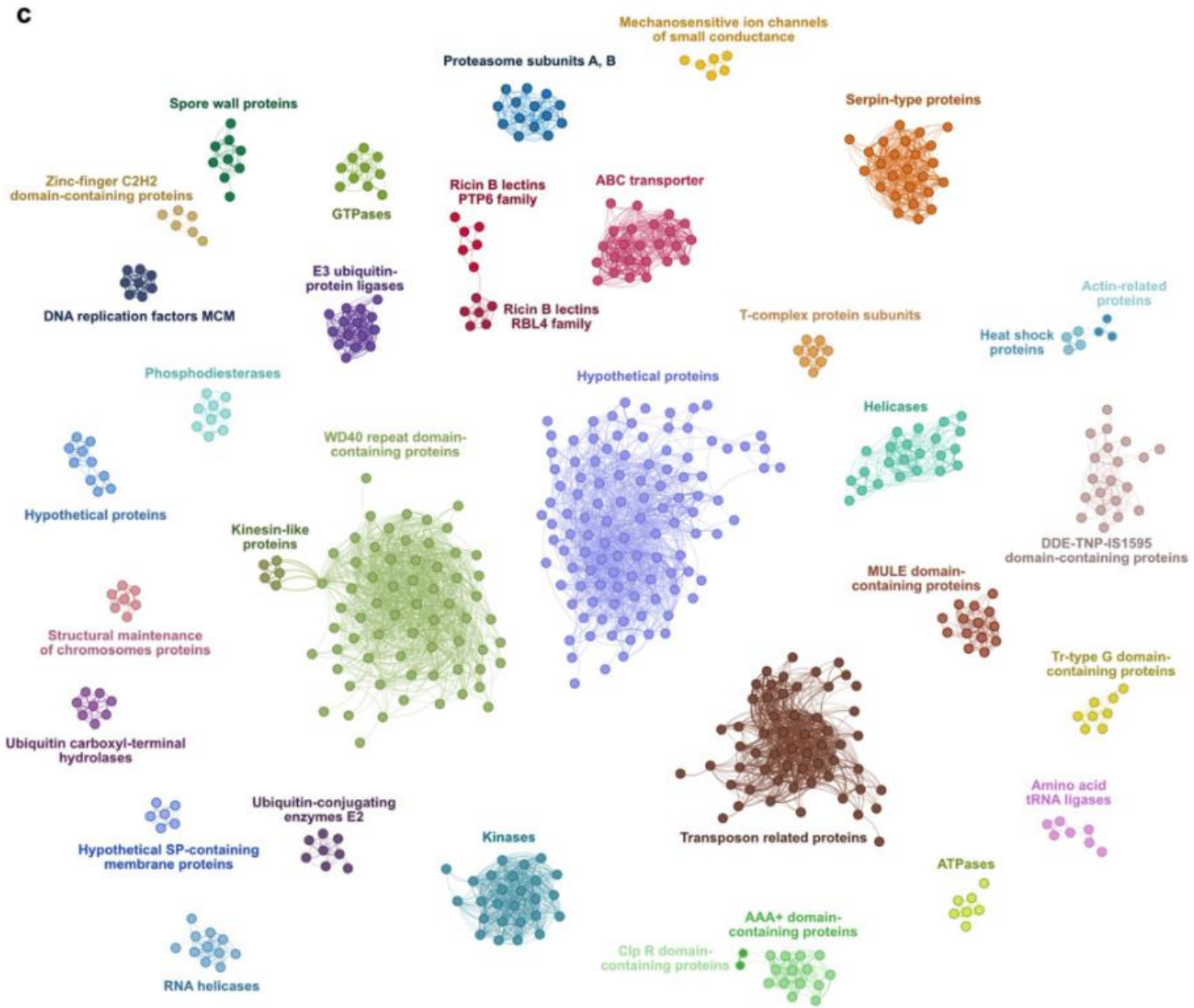
https://www.uniprot.org/help/ProtNLM

Structure-based network of highly abundant protein-fold families encoded by the analyzed V. necatrix genome

AlphaFold-predicted protein models were analyzed for structural relatedness in a *Foldseek all-against-all search*

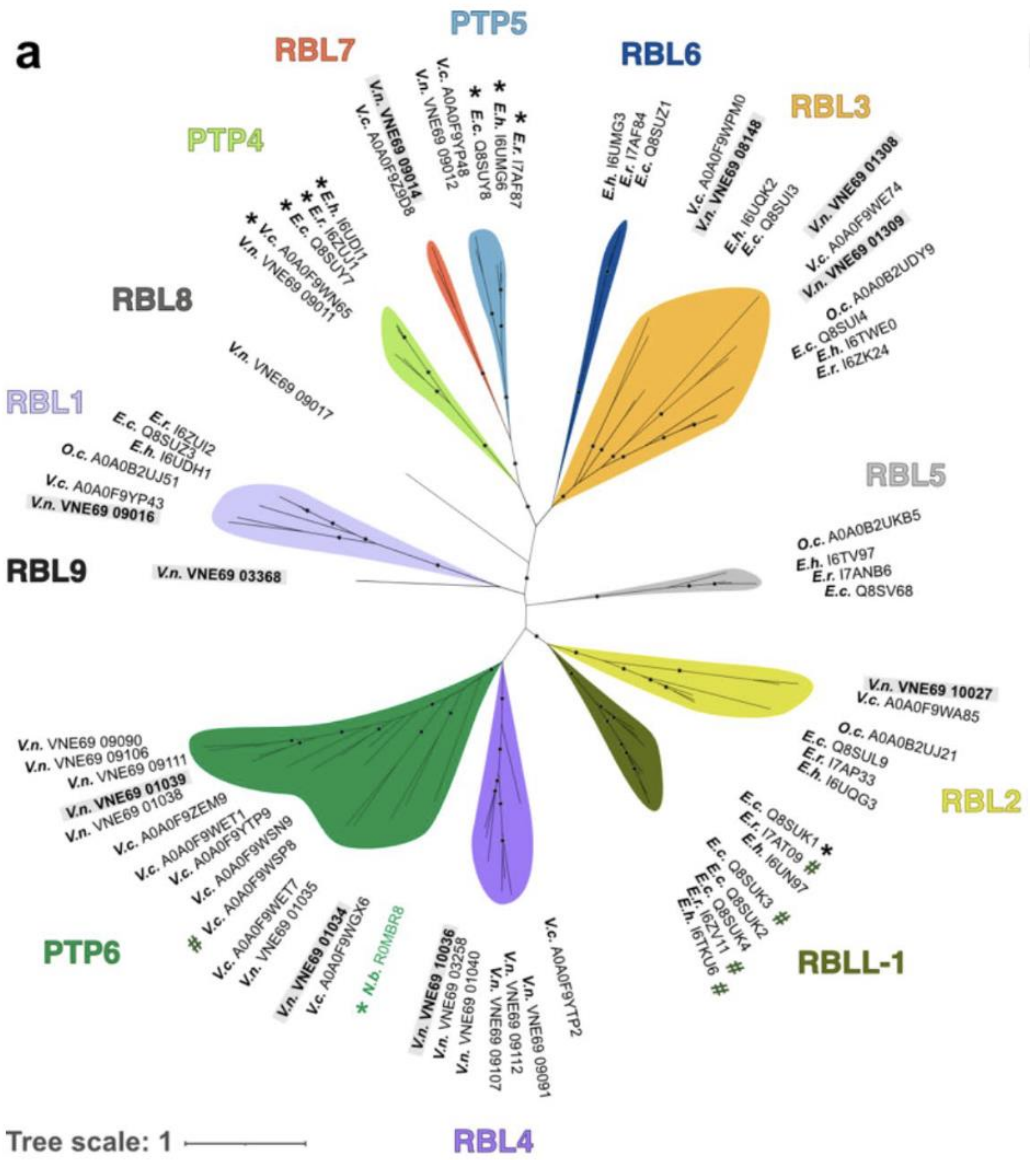
Structural similarity is represented by TM score (used as a measure for protein network graph generated in Gephi)

Each node represents a protein colored according to its fold family

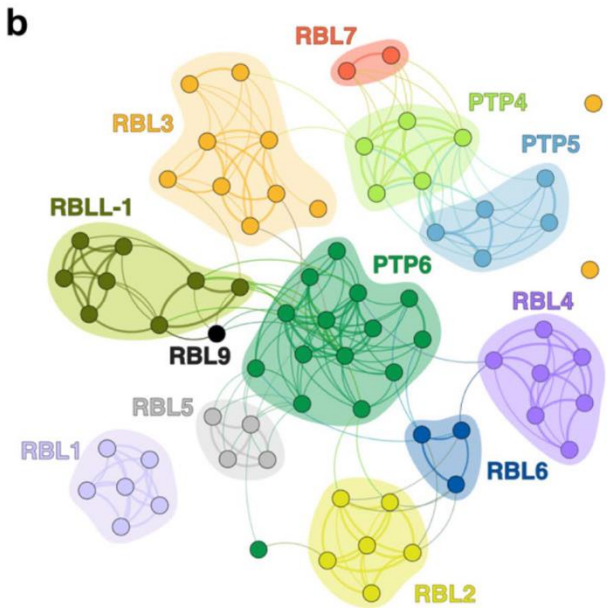


Svedberg et al. BMC Genomics 2024

Lineage specific expansion of RBL family in microsporidia



- RBL proteins are beta-trefoil fold lectins, a class of carbohydrate binding protein that aids in pathogen adherence to host cells
- Cladogram of **~70 RBLs identified using structure-based similarity**
  - 22 found in V. necatrix and others previously not identified in other species
  - Cladogram is from sequences, analyzed by IQTREE
  - All localize to the microsporidian polar tube (infection organelle) or spore wall from expression data
- 13 different RBL clades, of which 4 contain previously characterized proteins, the rest are new
- Structure-based clustering of RBL proteins supports clade grouping

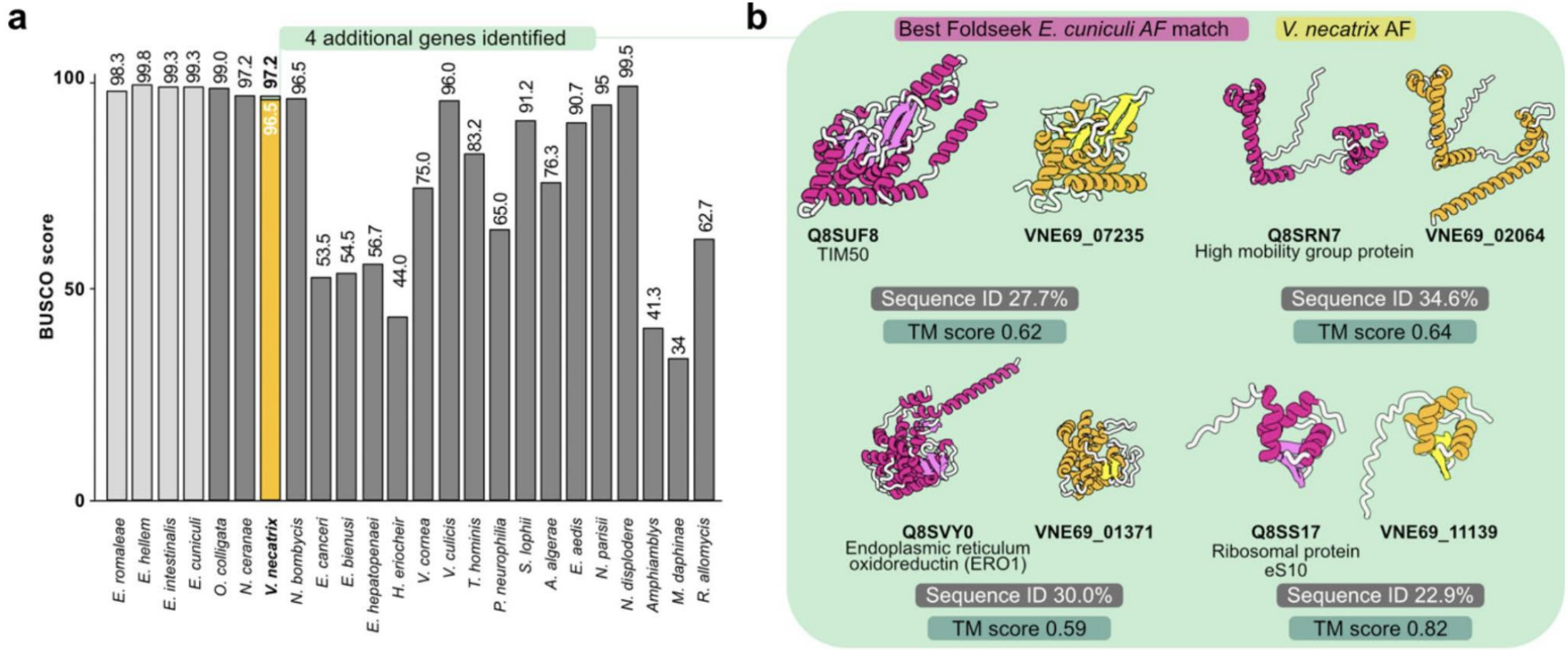


Svedberg et al. BMC Genomics 2024

4 of 11 missing BUSCO genes could be retrieved using protein structural similarity approach

BUSCO = benchmarking universal single copy orthologs

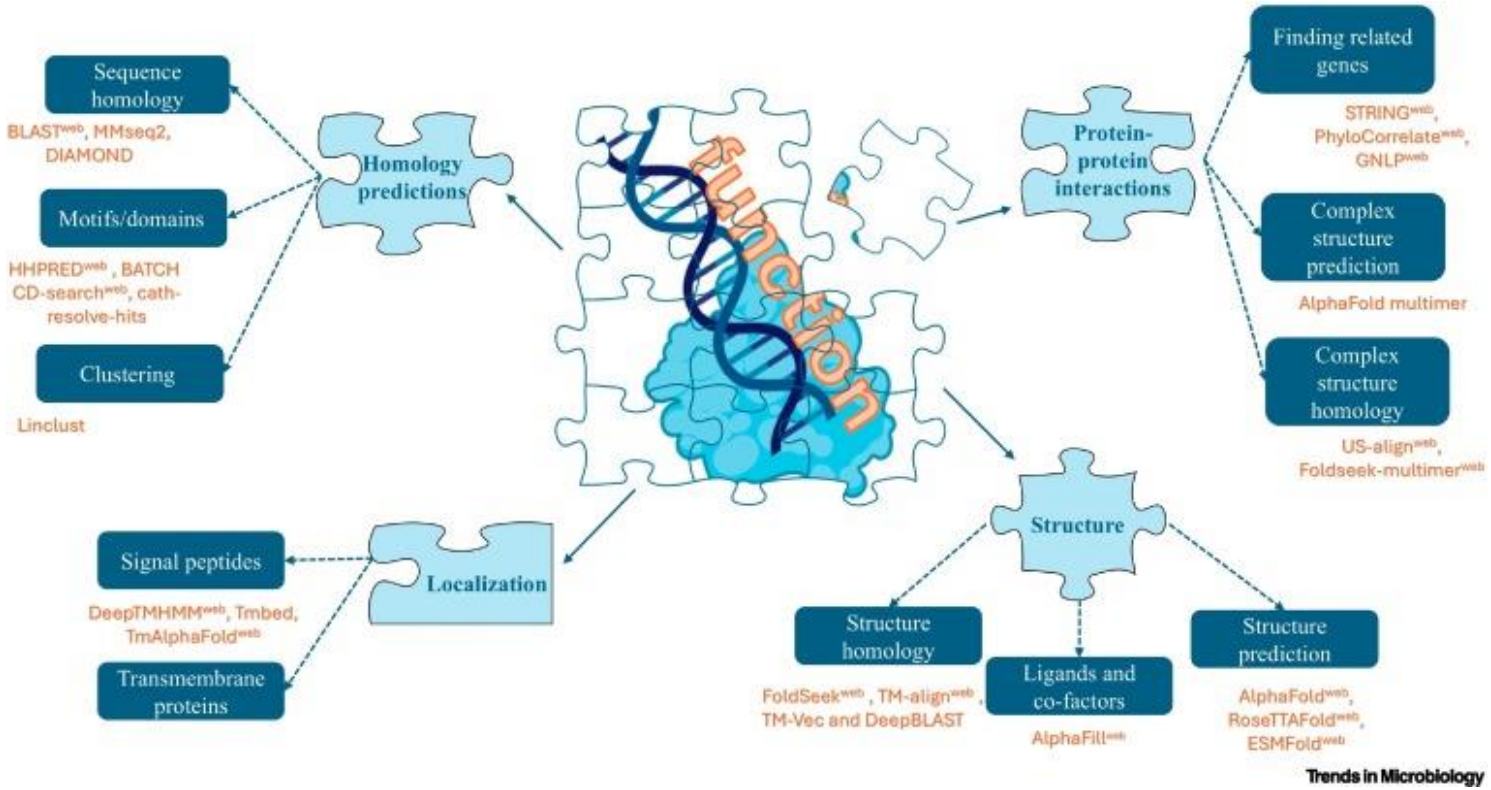
BUSCO relies entirely on sequence similarity to estimate genome completeness



TM score of >0.5 is typically strong indication of structural similarity, with good chance that they are homologous (likely share common evolutionary origin)

Svedberg et al. BMC Genomics 2024

Summary: Another piece in the puzzle for protein-coding gene annotation



Mmseqs2 + colabfold
https://github.com/steineggerlab/colabfold-protocol

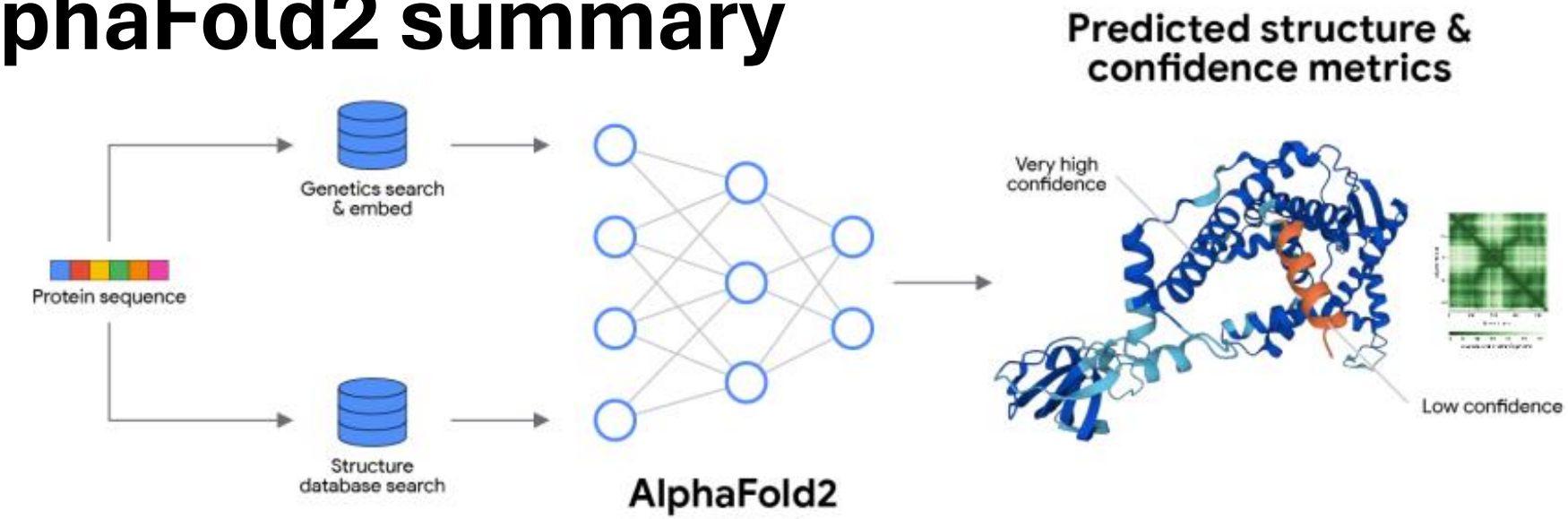
FoldSeek
https://github.com/steineggerlab/foldseek

- Structure-based protein homology can potentially fill in gaps in annotation where sequence-based approach falters
- Other approaches such as PPI, ligand binding , localization, etc probably hold key to further insights into protein function for annotation
- Even Uniprot is making use of some secondary structure alignment information to annotations
- While papers showed some nice agreement with existing methods, further experimental validation would be useful
- I'm in the process of testing on some of our protist "hypothetical proteins" to see if it works well in our hands

Figure from Cheskis et al. Trends in Microbiology 2024

Extra slides

AlphaFold2 summary



The role of multiple sequence alignment (MSA)

From the user's perspective, the only input AlphaFold2 needs is protein sequence(s). However, AlphaFold2 works by building a multiple sequence alignment (MSA), in which multiple similar protein sequences are set alongside each other. The MSA is generated by querying several protein sequence databases with the input sequence.

The primary input for AlphaFold2's neural network is then the MSA. AlphaFold2 uses MSAs to compare and analyse the sequences of similar proteins from different organisms. It highlights similarities and differences, which helps understand the evolutionary relationships between the proteins.

If two amino acids in a protein are in close contact, mutations in one of them will probably be followed by mutations of the other. This preserves the structure of the protein, and is known as co-evolution or covariation. The opposite is also true: if two regions of a protein are changing and evolving independently from each other, it is likely that they are not in direct contact (Benner & Gerloff, 1991; Göbel et al., 1994; Korber et al., 1993; Taylor & Hatrick, 1994).

A high-quality MSA is essential for AlphaFold2 to produce an accurate prediction of protein structure. A diverse and deep MSA, with hundreds or thousands of sequences in the alignment, will help AlphaFold2 to identify co-evolutionary signals and use them to figure out the protein's 3D structure. Conversely, a shallow MSA, with only tens of sequences and low variability among them, is the most common reason for failing, non-confident and inaccurate AlphaFold2 predictions.

The role of pair representations

When AlphaFold predicts the 3D structure of a protein, it creates a set of "pair representations". Every pair of amino acid residues in the protein, no matter how distant, is represented separately. This enables the software to encode the co-evolutionary relationships between them based on the MSA. This information can ultimately be interpreted as the relative positions of amino acid residues and distances between them.

AlphaFold2 uses a neural network called Evoformer. This interprets and updates both the MSA and the pair representations. The important aspect of this network is the continuous flow of information between the MSA and the pair representations. This enables reasoning about spatial and evolutionary relationships, which refines the structural hypothesis.

If available, AlphaFold2 can use supplied protein structures (e.g. structures derived from experiment) as templates. However, AlphaFold2 tends to ignore such templates if there is enough information coming from the MSA.

How do we get a structure?

AlphaFold2's structure module takes both the updated pair representation and the original sequence (which is the first row of the updated MSA) from the Evoformer. The structure module first turns this into a backbone of the 3D structure. It then finishes the modelling by placing the amino acid side chains and refining their positions.

AlphaFold2 then performs an iterative process called "recycling". It feeds the MSA, the pair representations and the 3D structure back to the neural network, and generates a new 3D structure. This process is repeated three times, allowing AlphaFold2 to improve the accuracy of the final structure.

https://www.ebi.ac.uk/training/online/courses/alphafold/inputs-and-outputs/a-high-level-overview/