Clustering of Large Protein Databases for Functional Family Inference

M2 Internship
February–August 2026

Defne Ozguven, Hubert Naacke, Mathilde Carpentier, Stéphane Gançarski, Lucie Bittner

CONTEXT

Advances in metagenomic sequencing technologies produce millions of protein sequences from environmental organisms, yet experimental functional annotation remains extremely rare.
The challenge is to automatically group these proteins into homogeneous functional families to propagate known annotations to unknown proteins at a very large scale.

⚠ Only a small fraction of proteins have a known annotation. Ideal clusters must be homogeneous (same function) and minimal number of singletons.

CLUSTERING

Markov Clustering Algorithm (MCL): Simulates random walks on the similarity graph via Markov chains.
→ **Inflation** parameter strengthens strong connections, attenuates weak ones.

We can also consider the evolutionary relationships of the sequences during clustering:

- **OrthoMCL**
- **OrthoFinder**

These methods identify orthologs (genes related by speciation) rather than just homologs (genes related by any common ancestry), enabling more biologically meaningful cluster boundaries.

DATA

7.4M protein sequences

Dinoflagellate amino acid sequences in FASTA format

89M alignments

Sequence Similarity Networks (SSN) obtained by BLAST/DIAMOND

48M annotations

InterProScan annotations from 19 data banks: mapped to **Gene Ontology**

CLUSTERING RESULTS

	MCL	OrthoMCL	OrthoFinder
Total clusters	493K	347K	478K
Total sequences	5.3M	2.5M	5.8M
Fully dark	452K	276K	415K
Mixed	33K	41K	46K
Singletons	11K	18K	0
Homogeneity	0.930	0.947	0.929

THREE APPROACHES

PRE-CLUSTERING:
Cluster independently on SSN and on embeddings, then align the two partitions

POST-CLUSTERING:
Reweight the SSN edges by combining alignment scores with cosine similarity between embedding vectors

INTEGRATED:
Exploit embedding similarity iteratively during the graph clustering process itself

EMBEDDINGS

Language models trained on hundreds of millions of sequences produce dense vectors encoding **function** and **structure** implicitly, without explicit labels.
Key advantage: twilight zone (<30% sequence similarity), less complexity

ESM-2
650M parameters

ProtBert
420M parameters

EVALUATION

Homogeneity Score: Fraction of annotated proteins in a cluster sharing the dominant annotation. Computed via **Jaccard** on GO Molecular Function terms to handle multi-label settings.
Singleton Rate: Clusters containing a single member
Annotation potential: Number of sequences that can receive an annotation

FUTURE WORK

Knowledge graphs containing:

- Sequence embeddings
- Protein 3D structure
- Environmental metadata

Addition of data from:

- TARA Oceans (74M sequences)
- ATLASea (100M+ sequences)

Necessitating **scalable methods.**

MaDICS Symposium
June 2026