Structural Motif Detection on the Scale of the Protein Universe with

FOLDDISCO

Hyunbin Kim[1,2], Rachel Seongeun Kim[1,2], Milot Mirdita[2] and Martin Steinegger[1,2,*]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
[2]School of Biological Sciences, Seoul National University, Seoul, Republic of Korea

Protein structural motifs are short, evolutionarily-conserved patterns of atoms involved in protein functions. These motifs are usually discontinuous in sequence making them difficult to detect by structural alignment methods like Foldseek. Graph-based, disjoint segment-utilizing methods are more sensitive but computationally intense. Inverted index-based motif search, such as offered by the RCSB, provides constant search time but would require substantial storage for large predicted protein structure databases, such as the AlphaFoldDB and ESMAtlas.

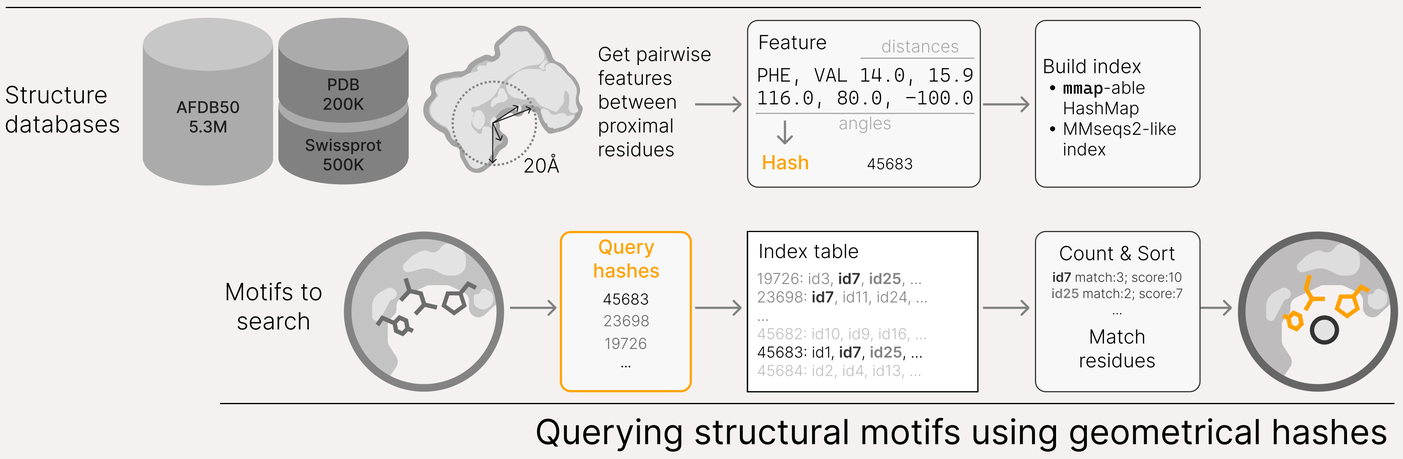
Here, we present **Folddisco**, a novel inverted-index based method that overcomes these limitations.

For the first time, Folddisco allows to detect structural motifs within databases representing the whole protein universe by efficiently compressing the inverted index, allowing the full AlphaFoldDB to fit on a single disk and enabling for the first time protein-universe-scale motif search on a single machine.
To do so, Folddisco introduces several innovations by

(1) reducing inverted-index storage space by 70% by omitting location information
(2) improving precision with a novel feature for capturing side-chain orientation
(3) offering fast searching speed with a highly optimized index structure.

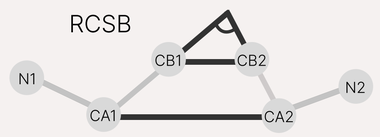
Folddisco is free and open source software written in Rust available at **https://folddisco.foldseek.com.**

WORKFLOW

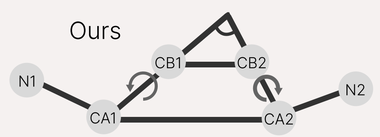


FEATURE

RCSB suggest structural motifs consisting of Cα-Cα and Cβ-Cβ distances and their angles between.



However, this ignores side-chain orientation and increases false-positives.



Folddisco uses **two side-chain torsion angles** in addition to RCSB's feature, improving search accuracy.
To capture the orientation, all **angles are encoded with both sine & cosine.**

INDEX

Folddisco efficiently compresses the inverted index to fit the whole AlphaFold database on a single disk and enable protein-universe searches with a single machine.

Our main approach is building record-level index, which does **not save information within each protein.**

BEFORE	Folddisco
Feature	Feature
PHE-VAL-14-16-116	**456838625 (32bit)**
Index	Index
PDB1-RES1-RES2	**PDBID (~8-16bit)**

Omitting residue information deduplicates observed hashes. By applying **delta-encoding**, we required storage space by 70%.

Indexing 53 million AFDB50 proteins requires only 1.6TB compared to ~18TB for pyScoMotif (extrapolated).

QUERY

We introduce a modified **Inverse Document Frequency (IDF)** as the ranking metrics for the queried results. By giving more weight to rare features, motif ranking is improved.
A length-based penalty avoids favoring random matches in long proteins.

$$\mathrm{IDF}_h = \frac{\#\text{ of structures containing hash } h}{\#\text{ of structures in the set}}$$

$$Score = \sum_{i=1}^{n} (\mathrm{IDF}_{h_i} - \log_2(N_{\text{residue}}) + C_{\text{length penalty}})$$

Folddisco accelerates residue matching with an **amino acid-based pre-filter**.

Matching relies on two graphs: a **primary feature hash graph** and a secondary Cα distance graph.
Residue matches are determined by finding **connected components** in the feature graph.

AVAILABILITY

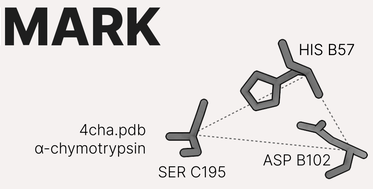
#recomb2025

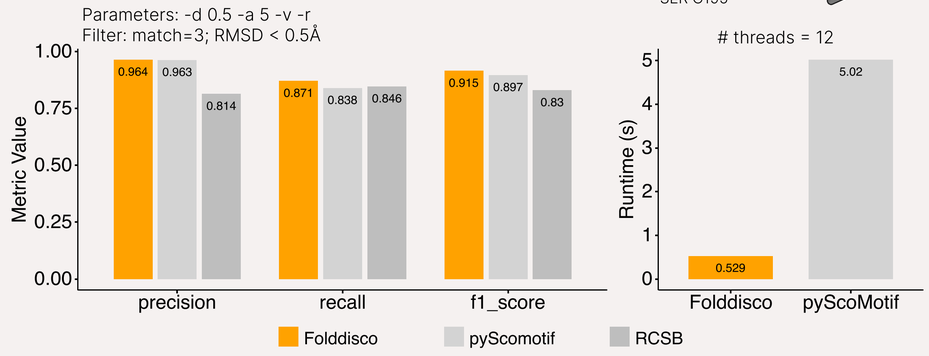
Folddisco is free and open source Rust software available at
https://folddisco.foldseek.com
We are also building a webserver :)

folddisco

QUERYING BENCHMARK

S01 serine peptidase motif

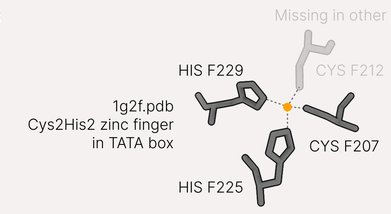


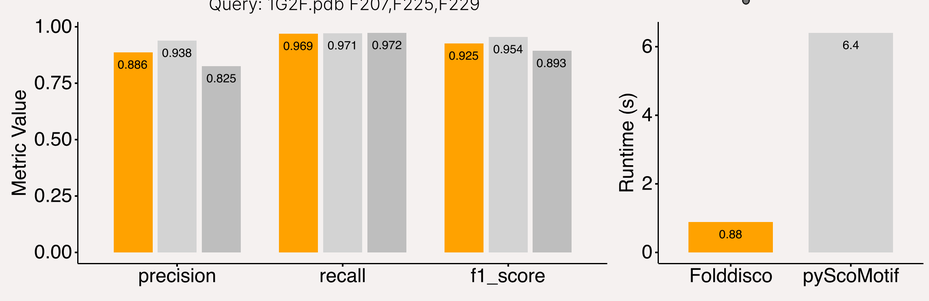


In a serine peptidase motif search benchmark against 23,391 human AlphaFold proteome, Folddisco achieved the most TPs with least FPs (108 TP, 4 FP). TPs are MEROPS-annotated S01 family serine peptidases.

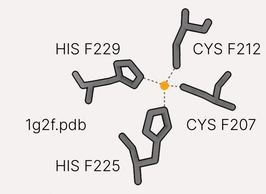
C2H2-type zinc finger motif

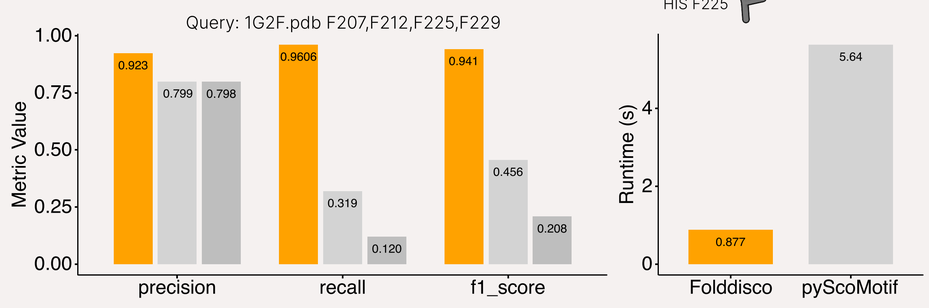
Due to false-negative issue in pyScoMotif if searching with 4 residues, we benchmark with both 3 and 4 residues for a fair comparison.



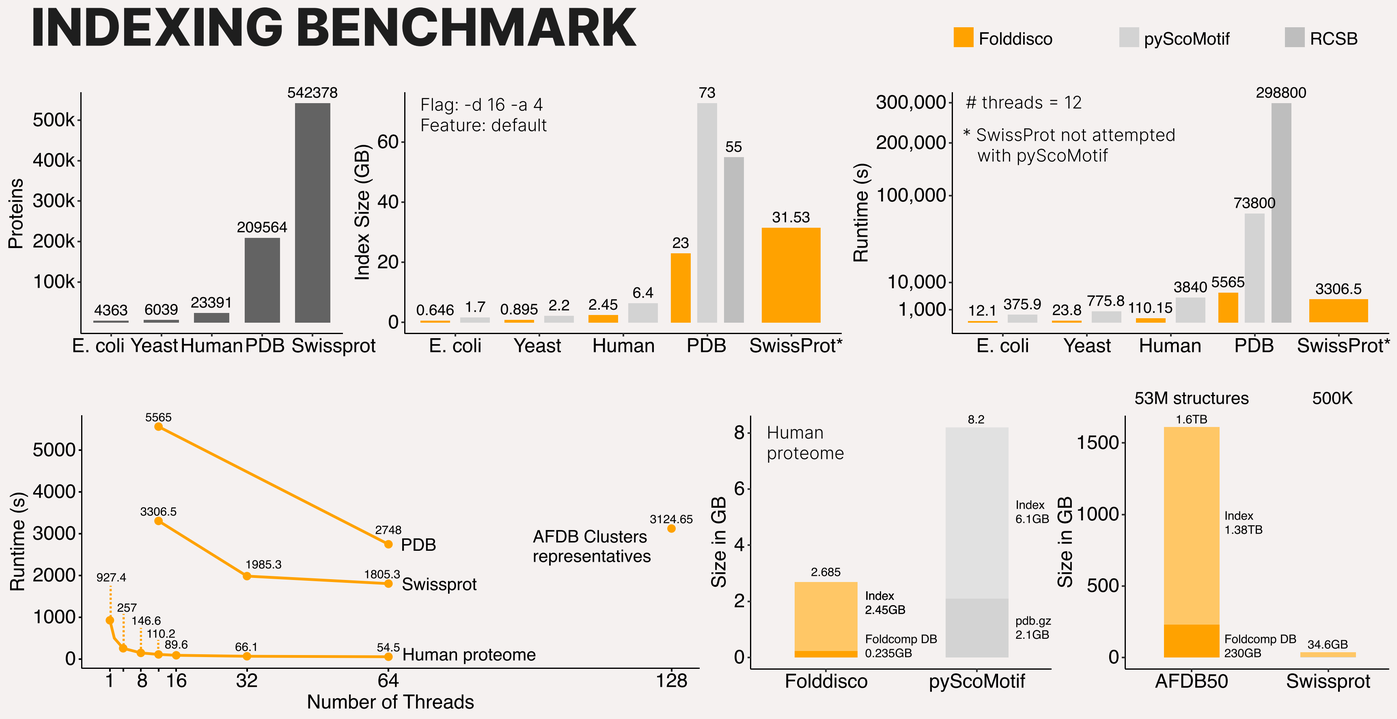


Folddisco does not suffer query-dependent performance drops with zinc finger motifs.
With full motif, Folddisco doubles F1-score and is 6.4 times faster compared to pyScoMotif.



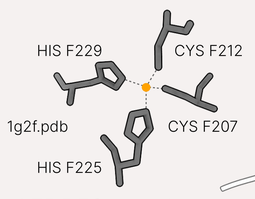


INDEXING BENCHMARK

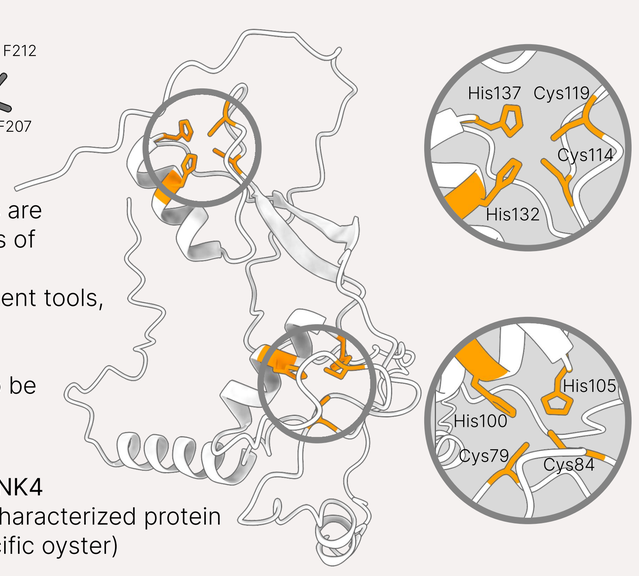


Indexing the human proteome took 110 seconds resulting in a 2.45GB index, 34.86 times faster compared to pyScoMotif. PDB takes 23GB storage space, a 68% reduction from pyScoMotif and 58% from RCSB's index. Additionally, we have built a 1.6 TB index of 53 million AFDB50 representatives.

ANNOTATING UNCHARACTERIZED



Most AFDB50 matches are concordant with results of sequence/profile-based domain assignment tools, such as InterPro.
Approximately, 3% of matches were found to be uncharacterized.



Example uncharacterized protein with two zinc-fingers

K1RNK4
Uncharacterized protein
(pacific oyster)

Reference
- Bittrich, Sebastian, Stephen K. Burley, and Alexander S. Rose. "Real-time structural motif searching in proteins using an inverted index strategy." PLoS computational biology 16.12 (2020)
- Cia, Gabriel, et al. "pyScoMotif: Discovery of similar 3D structural motifs across proteins." Bioinformatics Advances 3.1 (2023)

Acknowledgement
This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00396026, 2021R1C1C1012065, 2021M3A9I4021220, 2020M3A9G7103933)