Rapid and Sensitive Protein Complex Alignment with Foldseek-Multimer

Woosub Kim[1], Milot Mirdita[2], Eli Levy Karin[3], Cameron Gilchrist[2], Hugo Schweke[4], Johannes Söding[5,6], Emmanuel Levy[4,✉], and Martin Steinegger[1,2,7,8,✉]

Advances in computational structure prediction will vastly augment the hundreds of thousands of currently-available protein complex structures. Translating these into discoveries requires aligning them, which is computationally prohibitive. Foldseek-Multimer computes complex alignments from compatible chain-to-chain alignments, identified by efficiently clustering their superposition vectors. Foldseek-Multimer is 3-4 orders of magnitudes faster than the gold standard, while producing comparable alignments; allowing it to compare dozens of billions of complex-pairs in a day. Foldseek-Multimer is open-source software: github.com/steineggerlab/foldseek and webserver: search.foldseek.com.

Contact: ***emmanuel.levy@gmail.com, martin.steinegger@snu.ac.kr***

The similarity between two protein complexes is reflected in their optimal structural alignment, which also dictates a pairing of their chains. Aligning and comparing quaternary structures is essential for quantifying their structural diversity and identifying structural similarities and changes across different conformations or homologs. Furthermore, it is important to understanding protein function because many proteins operate as complexes (1).

Recently, Foldseek (2) has been developed as a fast structural aligner to detect similarity between two single-chain proteins, expressed using 3Di, a designated alphabet for describing tertiary amino acid interactions. Using Foldseek allows searching for similar single-chain structures in large databases, such as the AFDB (3). However, since aligning two complexes requires knowing the correct pairing of their chains, Foldseek cannot be used directly to find the alignment between them.

US-align (4) is a structural aligner for various types of molecules, including protein complexes. Its strategy for complex alignment is TM-score maximization. As there is a factorial number of possible assignments of chain pairings, US-align employs a greedy search heuristic for proposing candidate assignments, which are refined by dynamic programming. This heuristic was shown to make US-align up to five times faster than the long-standing state-of-the-art MM-align (5), while producing higher scoring alignments, making US-align the gold-standard for pairwise complex alignment.

Aiming to discover pairs of structurally conserved interfaces in large databases, Dey et al. developed QSalign (6). QSalign saves computation time by performing the full pairwise structural alignment only on complex pairs prefiltered based on their sequence similarity, retaining pairs with at least ca. 25% sequence identity. This speed-up comes at the expense of sensitivity, limiting its ability to discover structurally similar pairs in the twilight zone or below. Despite this speed-up, QSalign still took several months to conduct an all-vs.-all search encompassing about a hundred thousand complexes in the 3DComplex DB V5 (7) on 100 CPUs.

The challenge of sensitively searching large databases is expected to intensify as the computational prediction of protein complexes using tools like AlphaFold-Multimer (8) can now be performed on entire proteomes to systematically predict complexes (9–11) and to metagenomic samples. This will enrich our databases with a plethora of structures, potentially in the millions, in the coming years.

To address the need for large-scale structural comparisons between complexes, we developed Foldseek-Multimer (**Fig. 1**). Three factors contribute to its speed: 1) using Foldseek for fast chain-to-chain comparison, 2) describing chain-to-chain alignments as superposition vectors, and using them to identify complex alignments by efficient clustering, and 3) utilizing clustered databases during search. Through benchmarks, we show that Foldseek-Multimer is: 1) nearly as accurate as US-align, while being orders of magnitude faster, 2) sensitive and suitable for metagenomic studies of complexes with low sequence identity to others, 3) capable of all-vs.-all searches, examining billions of complex-pairs in 24h.

The quality of Foldseek-Multimer's alignments was compared to that of US-align on a benchmark of 931 pairs of protein complexes, known to be structurally similar, using either tool to align them. Foldseek-Multimer was run in two modes, differing in the algorithm used for chain-to-chain alignment: 3Di+AA (*Foldseek-MM*) or TM-align (12) (*Foldseek-MM-TM*). Both tools detected the vast majority (> 95%) of pairs as similar, aligning them with a TM-score >= 0.65, which is a cutoff found to be optimal for detecting structural homology among complexes (6) (US-align: 97.6%, *Foldseek-MM-TM*: 97.4% and *Foldseek-MM*: 95.8%). Using either mode, Foldseek-Multimer computed highly correlated TM-scores to those of US-align (**Fig. 2a**). We measured the runtime of the tools, breaking down the contribution of Foldseek-Multimer's components to its speed. First, given the task of computing 931 pairwise alignments, we observed a speedup of 1-2 orders of magnitude over US-align (**Fig. 2b top**), reflecting the efficiency of the chain-to-chain alignment (*Foldseek-MM*) and superposition clustering (*Foldseek-MM* and *Foldseek-MM-TM*). The performance of *Foldseek-MM-TM* thus highlights the key contribution of Foldseek-Multimer's innovative use of superpositions as an alternative to US-align's global alignment. Next, the tools queried each of the 667 complexes in the benchmark (Online Methods) against the 3DComplexV7 database (7). Here, Foldseek-

[1] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea. [2] School of Biological Sciences, Seoul National University, Seoul, Republic of Korea. [3] ELKMO, Copenhagen, Denmark. [4] Department of Chemical and Structural Biology, Weizmann Institute of Science, Rehovot, Israel. [5] Quantitative and Computational Biology, Max-Planck Institute for Multidisciplinary Sciences, Göttingen, Germany. [6] Campus Institute Data Science (CIDAS), University of Göttingen, Germany. [7] Institute of Molecular Biology and Genetics, Seoul National University, Seoul, Republic of Korea. [8] Artificial Intelligence Institute, Seoul National University, Seoul, Republic of Korea.

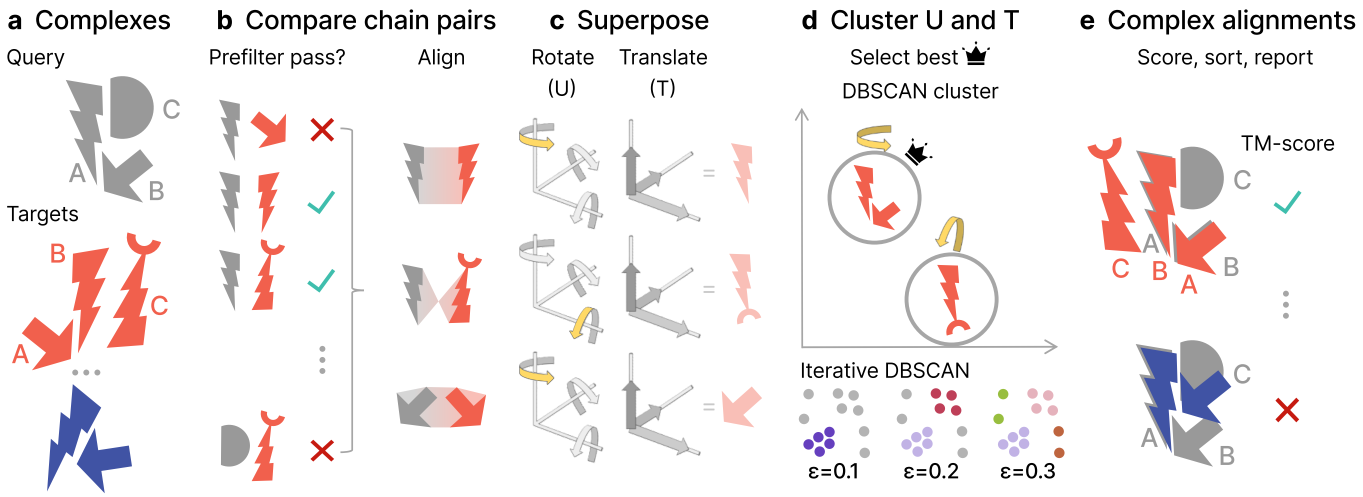


Fig. 1 Schematic description of Foldseek-Multimer principles. **a**, Foldseek-Multimer allows fast querying of input complex(es) against a large database, potentially containing millions of targets. **b**, All chains from the query (gray) are compared to those of each target (red). A prefilter allows to quickly reject non-matching chain pairs so the full alignment procedure is only applied to promising complex pairs. **c**, Foldseek-Multimer represents each chain-to-chain alignment as a superposition, described by rotations and translations, required for superposing the target chain onto the query. In this simplified example, two chain-to-chain alignments (top, bottom) are a rotation along one of the axes (yellow highlight), while one (middle) is a rotation along a different axis. **d**, Two or more chain-to-chain alignments, which belong to the same complex-to-complex alignment will have the same superposition. Foldseek-Multimer uses the DBSCAN algorithm iteratively, with increasing radii, to identify superposition clusters and selects the best-scoring valid cluster for computing the complex alignment. **e**, Based on the best-scoring cluster, the total TM-score is computed over all chain-pairs between the query and each of the targets.

Multimer was 3-4 orders of magnitude faster than US-align (**Fig. 2b bottom**) due to an additional speedup by its prefilter. Recently, Altae-Tran et al. (13) discovered the first CRISPR-Cas type IV-A system with a specified interference mechanism in an environmental sample of *Sulfitobacter* sp. JL08. Intrigued by how evolutionarily distant this system was to known proteins (<65% sequence identity to any entry in the nr database (14)), we predicted a part of its ribonucleoprotein complex structure using ColabFold-AlphaFold-Multimer (8, 15). The prediction was of acceptable quality (pTM=0.564) and we provided it as a query to Foldseek-Multimer and US-align in a search against the PDB100 (Online Methods). Despite having six chains and spanning 1,843 amino-acids, it took Foldseek-Multimer only 55 seconds in FS-MM mode and 7 minutes in FS-MM-TM mode to compare this query to the 426,347 entries of PDB100. By contrast, it took US-align 13 days.

Here, in addition to its fast core-algorithm (**Fig. 1**), Foldseek-Multimer gained further acceleration since PDB100 is a clustered database, allowing it to search against the 343,785 representatives, instead of all entries, and to expand the search only within promising clusters (Online Methods). *Foldseek-MM-TM* and US-align scored five entries above 0.65. These entries were the top ranks by *Foldseek-MM*, scoring above 0.5 but below 0.65 (**Fig. 2c**). All five hits were from a recently reported type IV-A system in *Pseudomonas aeruginosa* (16), which belongs to a different class (Gammaproteobacteria) than that of the query (Alphaproteobacteria). When examining the best match, 7xg4, we found that Foldseek-Multimer could identify similarity, despite extremely low sequence identity (11.1-19.8%) between the six subunit pairs of *Sulfitobacter* sp. JL08 and those of 7xg4. This provides further support for the previous identification of the *Sulfitobacter* sp. JL08 system as type IV-A and highlights the potential of Foldseek-Multimer for investigating protein complex structures predicted in distant organisms from environmental samples.

Next, we examined Foldseek-Multimer in an all-vs.-all setting, using the 3DComplexV7 database (7) as it had been previously analyzed in this setting using QSalign (Online Methods). QSalign relies on the time-consuming Kpax (17) structural alignment method, which prohibits it from conducting an exhaustive structural search. Thus, it first identified ca. 58 million pairs, which shared sequence similarity and then applied Kpax only to them, detecting ca. 11,2 million as similar (Online Methods: "QSalign Pairs"). Using 128 cores, *Foldseek-MM* then queried the clustered 3DComplexV7 (Online Methods) against itself, examining 57 billion pairs in 24h. Applying the same TM-score >= 0.65 cutoff as QSalign, *Foldseek-MM* identified 93.9% of the pairs previously identified by QSalign and found an additional 16 million similar pairs: "Foldseek-MM Pairs" (**Fig. 2d**). We used US-align for evaluating a randomly-selected sample of 1% of these pairs (Online Methods). US-align confirmed 98.2% of the sampled pairs and rejected 1.8% (TM-score < 0.65). We thus conclude that over 15.7 million pairs are new discoveries by Foldseek-Multimer, owed to its ability to detect similar complex structures, even when they share little sequence similarity.

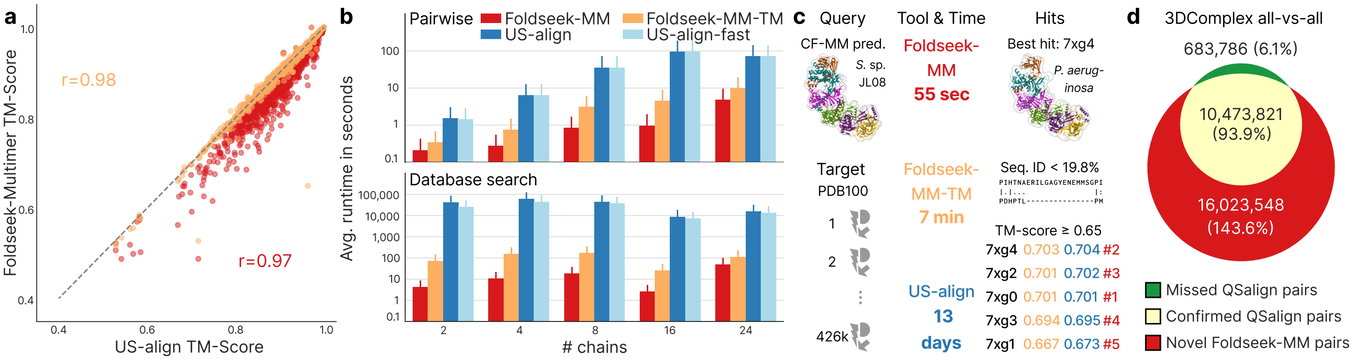


Fig. 2 Performance of Foldseek-Multimer. **a**, Query-length normalized TM-scores (target-normalized: Supp. Fig. 1) computed for 931 pairs of structurally similar complexes by US-align (x-axis) or Foldseek-Multimer (y-axis). Both measures correlate highly (Pearson's *r*). **b**, Execution time based on the dataset used for panel a. Complexes were binned by their number of chains, selected bins are shown (all bins: Supp. Fig. 2). Components' contribution to speed: in pairwise mode (top), alignment (*Foldseek-MM*) and superposition clustering (*Foldseek-MM* and *Foldseek-MM-TM*) make Foldseek-Multimer 10-100 times faster than US-align. In database search (bottom), complexes were queried against 3DComplexV7. Foldseek-Multimer is further accelerated by using its prefilter, making it 10^3-10^4 times faster. **c**, An AlphaFold-Multimer prediction of a part of a CRISPR-Cas ribonucleoprotein from an environmental sample (top-left) was queried by Foldseek-Multimer and US-align against PDB100. *Foldseek-MM-TM* identified the same hits as US-align, while being >2,300 times faster. These hits were top-ranked by *Foldseek-MM* (red) with TM-score > 0.5. Non-aligned components of 7xg4 (top-right) are set as transparent. **d**, Foldseek-Multimer was run on 57 billion pairs of complexes from 3DComplexV7. It discovered most pairs previously identified as similar by QSalign, and found an additional 16M pairs.

In addition to developing a command-line tool, we extended the Foldseek webserver to support Foldseek-Multimer and visualize its search results, using the NGL viewer library (18). The webserver overlays chain-to-chain assignments by using translucently colored protein surfaces. Users can choose between the two Foldseek-Multimer alignment modes, and apply taxonomic filters, restricting the search to specific clades.

In summary, the unprecedented combination of sensitivity and speed offered by Foldseek-Multimer makes it an essential tool for investigating protein complex structures in the AlphaFold2 era.

References

1. Levy, E. D. & Teichmann, S. Structural, evolutionary, and assembly principles of protein oligomerization. *Prog. Mol. Biol. Transl. Sci.* **117**, 25–51 (2013).
2. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
3. Varadi, M. *et al.* AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).
4. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* **19**, 1109–1115 (2022).
5. Mukherjee, S. & Zhang, Y. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res.* **37**, e83 (2009).
6. Dey, S., Ritchie, D. W. & Levy, E. D. PDB-wide identification of biological assemblies from conserved quaternary structure geometry. *Nat. Methods* **15**, 67–72 (2018).
7. Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* **2**, e155 (2006).
8. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 (2022).
9. Schweke, H. *et al.* An atlas of protein homo-oligomerization across domains of life. *Cell* **187**, 999–1010.e15 (2024).
10. Burke, D. F. *et al.* Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* 1–10 (2023).
11. Humphreys, I. R. *et al.* Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).
12. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
13. Altae-Tran, H. *et al.* Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering. *Science* **382**, eadi1910 (2023).
14. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information in 2023. *Nucleic Acids Res.* **51**, D29–D38 (2023).
15. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
16. Cui, N. *et al.* Type IV-A CRISPR-Csf complex: Assembly, dsDNA targeting, and CasDinG recruitment. *Mol. Cell* **83**, 2493–2508.e5 (2023).
17. Ritchie, D. W., Ghoorah, A. W., Mavridis, L. & Venkatraman, V. Fast protein structure alignment using gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics* **28**, 3274–3281 (2012).
18. Rose, A. S. *et al.* NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics* **34**, 3755–3758 (2018).

ONLINE METHODS

Algorithm: Overview

Foldseek-Multimer examines all possible chain-to-chain pairings between the compared complexes, using Foldseek (Fig. 1b). It then uses the fact that a structural alignment between two complexes implies a superposition: a set of rotations and translations, which minimize the sum of squared distances between the aligned residue pairs (19). Foldseek-Multimer therefore computes for each chain-to-chain alignment a vector, representing its superposition (Fig. 1c). Next, it uses DBSCAN (20) for clustering these vectors to identify compatible sets of chain-to-chain alignments, which share the same superposition and define valid complex alignments (Fig. 1d). Once complex alignments are identified, Foldseek-Multimer computes their TM-score (21) and reports them (Fig. 1e).

Algorithm: Input

Foldseek-Multimer allows for searching one or more query protein complex structures against a target complex structure, a database of complex structures or a database of clustered structures. Structures can be provided in PDB/mmCIF format or as a Foldseek-formatted database. Formatting structures is possible using the *createdb* command.

Algorithm: Chain-to-chain alignments

By utilizing Foldseek, Foldseek-Multimer offers two main modes for chain-to-chain structure comparison. The default mode, 3Di+AA, encodes structures as sequences over a 20-state 3Di alphabet, as fully described by van Kempen et al. [2]. Additionally, chain-to-chain alignments can be computed using TM-align [12], which is a global, albeit slower alignment method. During database search, a prefilter, which is based on the 3Di+AA mode, allows for a fast removal of most chain pairs, continuing to compute chain-to-chain alignments only on promising candidates.

Algorithm: Chain-to-chain superposition vectors

Given a chain-to-chain alignment, Foldseek-Multimer computes the superposition of the target chain onto the query chain, using nine rotations (U) and three translations (T). In preparation for aligning complex structure Q and complex structure T, Foldseek-Multimer creates a matrix with 12 columns, whose rows are the superposition vectors, computed from all chain-to-chain alignments, belonging to Q and T. The mean and the standard deviation (SD) of each column are then used to compute the coefficient of variation (CV = SD/mean) of the column and exclude less-informative columns (CV < 0.1). If the mean value of the column is < 1, the SD value is used instead of the CV for the exclusion criterion. Finally, the retained columns undergo normalization since they can have different scales. To that end, Foldseek-Multimer subtracts from each column its mean and divides it by its SD. We denote the resulting reduced and normalized matrix as *supQT*.

Algorithm: Chain-to-chain clustering

DBSCAN is used for clustering the rows of *supQT* as it doesn't require knowing the number of clusters *a-priori*. First, for each row of *supQT*, all rows within a radius of epsilon (initialized to 0.1) from it, are defined as its *neighbors*. To that end, the Euclidean distance between the row and each of the other rows is computed and compared to epsilon. Then, all rows, which have at least two neighbors are considered as "core-points" and the rest as "non-core-points". Next, a core-point is selected at random to start the first cluster. All its core-point neighbors are added to the first cluster. Each added core-point neighbor also adds its core-point neighbors and so on, until no more core-points can be added to the first cluster. Then all non-core-points, which are neighbors of members of the first cluster are added to it as well (without adding their neighbors). The second cluster is constructed similarly, operating on the remaining unclustered points.

After each cluster is computed, Foldseek-Multimer evaluates its validity, dismissing clusters that contain only one chain-to-chain alignment, or which include the same chain in multiple chain-to-chain alignments.

Cluster rescuing by Nearest Neighbors. Foldseek-Multimer attempts to rescue clusters dismissed due to including the same chain in multiple chain-to-chain alignments by selecting a compatible subgroup of points (i.e., chain-to-chain alignments) from each such cluster. To that end, the cluster center is computed and points in the cluster are selected for the subgroup in the order of their distance to it (closest point to the center is selected first). Selection for the subgroup is stopped once the process encounters a point that includes a chain, which was already added by a previous point.

If no valid clusters are found, the value of the radius epsilon is increased by 0.1 and the procedure is repeated. Each of the resulting valid clusters is equivalent to a set of compatible chain-to-chain alignments with a similar superposition that together define a complex alignment between Q and T.

Algorithm: TM-score computation

TM-scores are computed for the complex alignment derived from each of the valid clusters found for a Q-T complex pair as follows. First, the chains of complex Q are concatenated to each other in some order. Given the concatenation order of the chains in Q, Foldseek-Multimer concatenates the chains of complex T, in the order of their pairwise matches to the chains of Q, as defined by the cluster. Then, the TM-score between the concatenated Q and concatenated T is computed the same way Foldseek computes it for single-chain pairwise alignments, using the Cα coordinate vectors of both chains (concatenated chains in this case). Using this computation, all complex alignments a given query complex Q has with a specific target T and with all other target complexes can be ranked and reported by their TM-score.

Algorithm: Utilizing clustered databases

In order to further accelerate Foldseek-Multimer, we aimed to reduce the redundancy in the target database, an approach, which is also adopted by TM-search (22). To that end, we introduced a new capability to Foldseek, which allows it to efficiently search through clustered databases in MMseqs2 or Foldseek format (e.g., PDB100, see section). If the input has M cluster representatives and N cluster members (M < N), Foldseek will first search (prefilter + alignment) against the M representatives, finding candidates below a specific E-value threshold (the default value of 10 was used in this study). Extending to promising clusters only, the alignment step will then be carried out on all cluster members of the candidates. Foldseek-Multimer will use the alignment results of all extended clusters for computing superposition matrices and the following procedure steps, as described above.

The 3DComplex database

For the analyses presented in Fig. 2a, 2b and 2d, we downloaded the 3DComplex database version 7 (3DComplexV7 DB; **Data Availability**). Briefly, this database holds 238,966 structures, consisting of 539,146 chains and was created from the "Biological Units/Assemblies" downloaded from the PDB by the method described by Levy et al. [7].

QSalign Pairs

Prior to this study, QSalign had been applied to 3DComplexV7 DB and yielded a list of 57,953,513 compared structural pairs (15,647,147 heteromers + 43,120,560 homomers) in SQL format. Selecting high-scoring pairs (max TM-score >= 0.65) resulted in a list of 15,180,364 structurally-similar unique pairs. We removed 4,022,757 pairs that were either single-chain alignments (3,593,300) or identified as false-positives by US-align (429,457 with TM-score < 0.65). The remaining 11,157,607 are denoted here as "QSalign Pairs".

Pairwise benchmark

Starting with the list of 57,953,513 QSalign-compared 3DComplexV7 structures, pairs of complexes were selected per number of subunits, with that number ranging from 2 to 24. For each size, the criteria for selection were that the TM-score computed by Kpax [17] was greater than 0.8, and that pairs of homomers had less than 80% sequence identity. If more than 100 pairs matched the criteria, only the first 100 were selected, resulting in a total of 931 complex pairs included in the benchmark.

The PDB100 database

A version of the PDB, termed PDB100 was used to search for structural homologs of an environmental CRISPR-Cas as well as to measure the runtimes of Foldseek-Multimer and US-align. PDB100 was first introduced by van Kempen et al. [2], but further developed in this study, as described here. First, PDB, containing the asymmetric unit of 207,937 entries, consisting of 1,047,615 chains, was downloaded in November 2023 (**Data Availability**). Of these, 11,901 entries were associated with more than one structural model (e.g., the NMR experiment 2KOX). In total, 426,347 structural models were associated with the PDB entries. Next, all chains were clustered using Foldseek (parameters: `-c 0.95 --min-seq-id 1.0`), resulting in 343,785 redundancy-reduced representatives. In contrast to van Kempen et al., PDB100 is now a cluster database, which holds the representatives alongside information to associate them to their cluster chains and structural models. PDB100 is updated regularly and is available through the Foldseek webserver and can be downloaded using the *databases* command.

Environmental CRISPR-Cas

Four *Sulfitobacter* sp. JL08 protein sequences, identified as CRISPR-Cas type IV-A components by Altae-Tran et al. [13]: Csf1, Csf2, Csf3 and Cas6 were obtained from the plasmid map "pHS1068 NZ_CP025815 DinG HNH proteins (E. coli codon optimized) CRISPR array in pACYCDuet-1 with Lac promoters.gb", released by the authors. Following the reported stoichiometry of the CRISPR-Cas type IV-A core complex (23), we constructed an input file for ColabFold-AlphaFold-Multimer [15] with eight chains: 1xCsf1 + 5xCsf2 + 1xCsf3 + 1xCas6. When examining the structure, we noticed that AlphaFold-Multimer did not predict an interaction between Csf1 and Cas6 and the rest of the complex, so we omitted them and re-predicted the structure: 5xCsf2 + 1xCsf3. Comparing the four sequences of *Sulfitobacter* sp. JL08 to protein nr [14] was performed using the blastp webserver (Feb. 2024).

A clustered 3DComplex V7

For the all-vs.-all analysis presented in Fig. 2d, we used a clustered version of the 3DComplex V7 database. Its 539,146 chains were clustered using MMseqs2 (parameters: `-c 0.99 --min-seq-id 0.9`), resulting in 146,288 redundancy-reduced representatives. This procedure took 18 seconds, using 64 threads.

Evaluation of "Foldseek-MM Pairs"

About 16 million pairs of complexes were detected only by *Foldseek-MM* as similar. Since running US-align over all pairs is prohibitively slow, we randomly selected 160,252 pairs (ca. 1% of all pairs) and computed their alignment using US-align. For 2,844 of these (1.8%), US-align reported a TM-score < 0.65, which we use as an estimate for the false-positive rate among the 16 million. 157,391 pairs (98.2%) were confirmed as matches by US-align and the rest (17 pairs, <0.0001%) were aligned as monomers.

Runtime evaluation

A server with a 1x AMD EPYC 7702P 64-core CPU and 1 TB RAM memory was used in all benchmarks, using a single core for runtime measurements. The queries for the time measurements in Fig. 2b were the 677 unique complexes associated with the 931 pairs used in the benchmark. Due to its high computational demand, the runtime of US-align on these 677 complexes against the 238,966 3DcomplexV7 entries was extrapolated from running against 1,000 randomly-sampled 3DcomplexV7 entries. Reporting the average over the number of cases Nc

= 142,109,124,18,101,7,42,8,41,44,17,5,5,14 for each number of chains c = 2,3,4,5,6,7,8,9,10,12,14,16,18,24: $avg = \frac{1}{N_c} \sum_{i=1}^{N_c} t(q_i, \text{sample}) \frac{238{,}966}{1000}$. For the same reason, the total runtime was also extrapolated when measuring US-align on the *Sulfitobacter* sp. JL08 structure against the PDB100, using five samples: $avg = \frac{1}{5} \sum_{i=1}^{5} t(q, \text{sample}_i) \frac{426{,}347}{1000}$. All Foldseek-Multimer runtime measurements were performed against the full database, without extrapolation.

Tool commands and arguments

Foldseek-MM commit 54865b (default, using 3Di+AA):

```
foldseek easy-complexsearch query.pdb
target.pdb/targetDB result tmp --threads 1
--exhaustive-search 1
```

Foldseek-MM-TM commit 54865b (using tmalign):

```
foldseek easy-complexsearch query.pdb
target.pdb/targetDB result tmp --threads 1
--exhaustive-search 1 --alignment-mode 1
```

US-align version 20220924:

```
US-align query.pdb target.pdb -mm 1 -ter 0 -mol prot
```

Additionally, the flag '`-fast`' was set for during runtime assessments in Fig. 2b).

Data availability

3DComplexV7 DB:
shmoo.weizmann.ac.il/elevy/3dcomplexV6/Home.cgi
PDB:
files.wwpdb.org/pub/pdb/data/structures/all

Code availability

Foldseek-Multimer and its webserver are GPLv3-licensed free open-source software. The source code and binaries for Foldseek-Multimer can be downloaded at github.com/steineggerlab/foldseek. The analysis scripts are available at github.com/steineggerlab/foldseek-multimer-analysis. The webserver is available at search.foldseek.com and its source-code at github.com/soedinglab/mmseqs2-app.

References

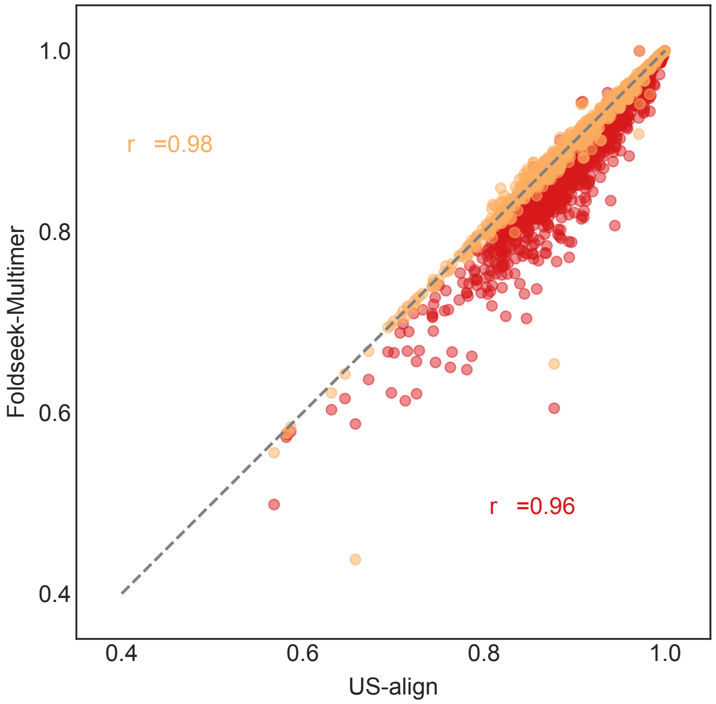
19. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **32**, 922–923 (1976).
20. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A density-based algorithm for discovering clusters in large spatial databases with noise*, KDD'96, 226–231 (AAAI Press, 1996).
21. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
22. Liu, Z., Zhang, C., Zhang, Q., Zhang, Y. & Yu, D.-J. TM-search: An efficient and effective tool for protein structure database search. *J. Chem. Inf. Model.* **64**, 1043–1049 (2024).
23. Taylor, H. N. *et al.* Positioning diverse type IV structures and functions within class 1 CRISPR-Cas systems. *Front. Microbiol.* **12**, 671522 (2021).

Acknowledgments

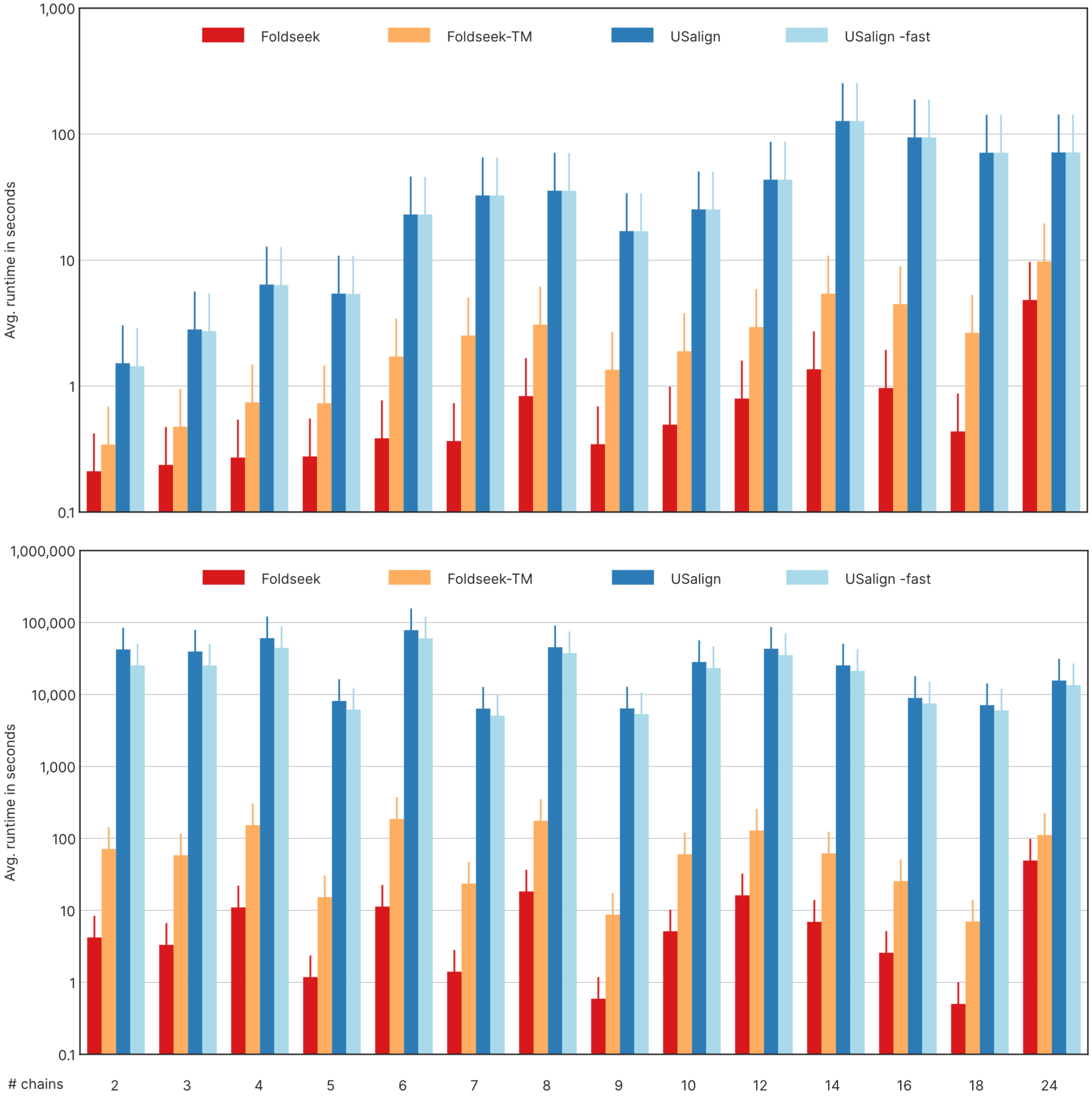
We thank Hyunbin Kim for helping with designing Figure 1. We thank the EMBL-EBI for releasing the "EMBL-EBI Icon fonts for the life sciences" that are used in the webserver. M.S. acknowledges support by the National Research Foundation of Korea grants (2020M3-A9G7-103933, 2021-R1C1-C102065 and 2021-M3A9-I4021220), Samsung DS research fund, Creative-Pioneering Researchers Program and AI-Bio Research Grant through Seoul National University. M.M. acknowledges support from the National Research Foundation of Korea (grant RS-2023-00250470).

Author contributions

W.K., J.S. and M.S. designed the Foldseek-Multimer algorithm. W.K., M.M. and M.S. developed the software. M.M. and C.G. developed the Foldseek-Multimer webserver. E.L. and H.S. designed the comparison to QSalign, provided 3DComplexV7 and applied QSalign to it. E.L.K. developed the Crispr-Cas example. E.L.K., W.K. and M.S. designed the figures, performed the benchmarks and wrote the manuscript, with contributions from all authors.



Supplementary Figure 1. Target-length normalized TM-scores of US-align and Foldseek-MM. Target-length normalized TM-scores (query-normalized: Fig. 2a) computed for 931 pairs of structurally similar complexes by US-align (x-axis) or Foldseek-Multimer (y-axis). Both measures correlate highly (Pearson's *r*).



Supplementary Figure 2. Speed comparison of Foldseek-Multimer to US-align. Execution time based on the dataset used for Fig. 2a. Complexes were binned by their number of chains. Speed comparison of pairwise alignment (top): bar height depicts the average and standard error computed for each bin over the following number of cases: 100, 100, 100, 43, 100, 22, 100, 18, 82, 100, 51, 18, 12, and 85. Speed comparison of Database search (bottom), complexes were queried against 3DComplexV7: bar height depicts the average and standard error computed for each bin over the following number of cases: 142, 109, 124, 18, 101, 7, 42, 8, 41, 44, 17, 5, 5, and 14.