AlphaFind: discover structure similarity across the proteome in AlphaFold DB

David Procházka, Terézia Slanináková, Jaroslav Olha, Adrián Rošinec, Katarína Grešová, Miriama Jánošová, Jakub Čillík, Jana Porubská, Radka Svobodová, Vlastislav Dohnal, Matej Antol

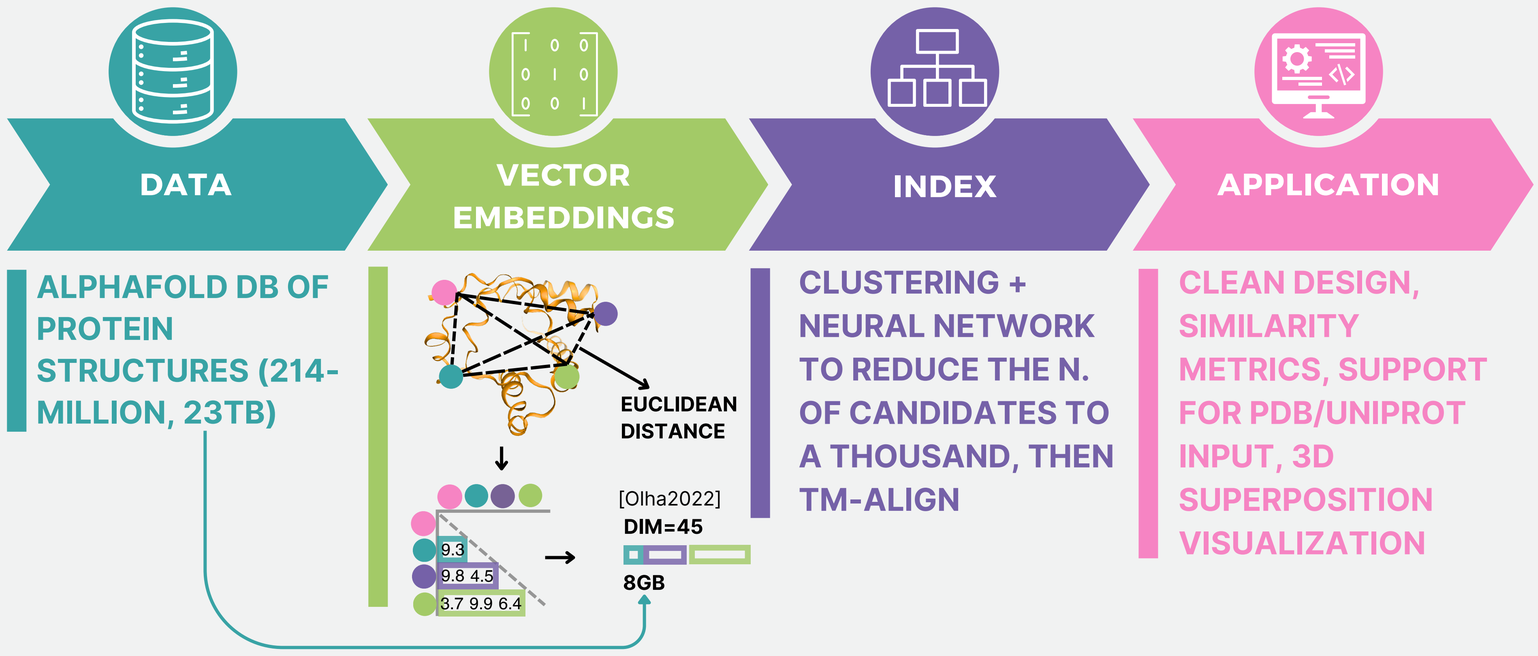
Masaryk University, Brno, Czech Republic

alphafind.fi.muni.cz doi.org/10.1093/nar/gkae397 github.com/Coda-Research-Group

INTRODUCTION

AlphaFind is a web-based search engine that allows for structure-based search of the entire AFDB. Uniprot ID, PDB ID, or Gene Symbol is accepted as input – AlphaFind will return the most similar proteins found within AlphaFold DB, with an option to perform additional search to extend and refine the results. The search results are grouped by their source organism and displayed along with their TM-Score and RMSD measures (relative to the query protein). If the result proteins are experimentally found, we provide all associated PDB IDs. The 3D visualizations of the structural overlap of the proteins are provided using NGL Viewer, and the additional filters at the top of the results table can be used to find specific organisms or Uniprot IDs.

METHOD



AlphaFind

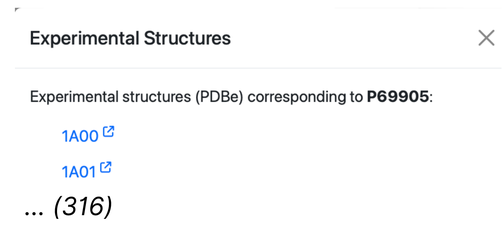
Manual → **Link to the manual**

Search with UniProt, PDB ID or Gene Symbol

P69905

Query: **P69905**
Name: **Hemoglobin subunit alpha**
Organism: **Homo sapiens**
P69905 in UniProt (316)
Search Time: **0.125 s**

Query metadata



Most similar proteins to *P69905* (showing 50 filtered out of 50)

Organism (Filter)	UniProt ID (Filter)	Global Similarity: TM-Score(?) ↓	Local Similarity: RMSD (Å)(?)	Aligned Residues	Sequence Identity(?)	Superposition
(2) Macaca mulatta	P63108	0.9999	0.050	100%		
(6) Homo sapiens	D1MGQ2	0.9999	0.050	100%		
(3) Equus caballus	P01958 (20)	0.9986	0.170	100%		
(1) Alphaproteobacteria bacterium	A0A3M1M8B2	0.8172	1.940	93.7%		

...

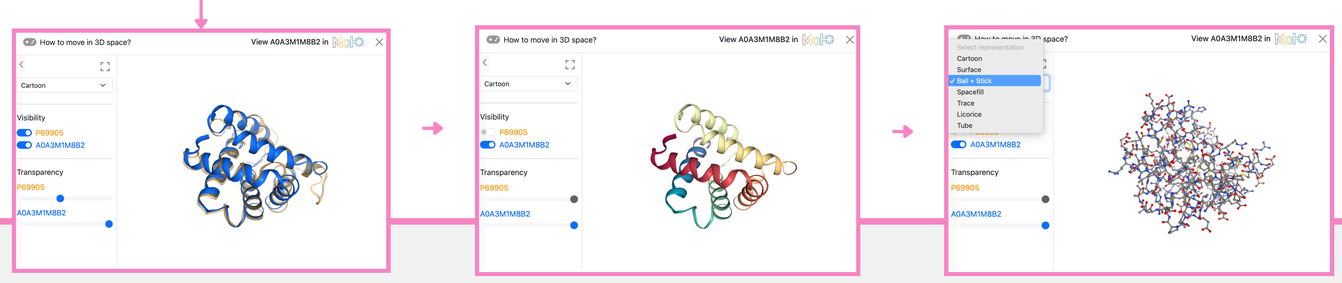
Most similar proteins to *P69905* (showing 1000 filtered out of 1000)

Export all to CSV

- ***TM-Score*** *- a global measure of similarity between two protein structures, computed by US-align [Zhang2022]*
- ***RMSD (Å)*** *- a local measure of distance between the 3D coordinates of the aligned Cα atoms -- the unaligned portions of the structures are disregarded*
- ***Aligned Residues*** *- portion of the aligned residues relative to the total length of the query protein*
- ***Sequence Identity*** *- portion of identical amino acid residues within the two aligned sequences relative to the length of the query protein.*

https://alphafind.fi.muni.cz/search?q=P69905&limit=1000

Bookmark + access anytime. Results will be accessible immediately (caching)



LIMITATIONS

- AlphaFind uses AlphaFoldDB files in version 3, before the v4 update from November 2022, which improved predictions for ~4% of structures.
- The results provided by AlphaFind are approximate; the engine can not guarantee that the best possible results for a given query will be found.
- AlphaFind processes the entire protein structure and handles protein regions with equal weight. Therefore, a high occurrence of unstructured regions in the input structure can bias the search. This phenomenon is more prevalent in coiled-coil structures but can also be observed in some small structures.

BACKLOG

- *Allow search by file upload, protein name (+autocomplete)*
- *Extend the maximum results set (currently 1000)*
- *Vector embedding method from [ADERINWALE2022]*
- *Evaluate (ROC curve) against FoldSeek*
- *AFDB v3 -> v4*
- *Extend to other complexes once AF3 is released*

REFERENCES

- [ADERINWALE2022]: Aderinwale, Tunde, et al. "Real-time structure search and structure classification for AlphaFold protein models." Communications biology 5.1 (2022): 316.
- [OLHA2022]: Olha, Jaroslav, et al. "Learned indexing in proteins: substituting complex distance calculations with embedding and clustering techniques." International Conference on Similarity Search and Applications. Cham: Springer International Publishing, 2022.
- [Zhang2022]: Zhang, Chengxin, et al. "US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes." Nature methods 19.9 (2022): 1109-1115.