PDBeFold: Searching for structural homologues of a protein

Gary Battle [1]

- Proteins
- Structures
- Beginner
- 1 hour

PDBeFold can be used to identify structural homologues in the PDB. PDBe's Secondary Structure Matching service (SSM) allows you to interactively compare, align and superimpose protein structures in 3D. This course will show you how to use PDBeFold and what you can do with it.

An undergraduate degree in a life science subject would be an advantage. You may wish to look at the PDBe quick tour [2] and Biomacromolecular structures [3] course as an introduction to this tutorial.

Learning objectives:

- Evaluate what PDBeFold is and what it can do
- Be able to launch PDBeFold and search for 3D structures that are similar to your protein of interest
- Interpret structural similarity of your results based on Q-scores
- Be able to view and analyse structures from your returned results

What is PDBeFold?

PDBeFold [4] is an interactive service that allows you to identify structures that are similar to that of your reference protein. It is a very powerful structure alignment tool which can perform both pairwise and multiple three-dimensional alignment. In addition, PDBeFold gives you various options to sort the results of your structural alignment query.

When to use PDBeFold

You can use PDBeFold to answer the following questions:
• Are there any structures in the PDB [5] that are similar to mine?
• What SCOP [6] and/or CATH [7] family could my structure belong to?
• Can I get some idea about the possible function of my protein based on structural similarity with others proteins in PDB?
• Can I perform a multiple alignment of many of my structures?

PDBeFold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.

PDBefold structure alignment is based on identification of residues occupying 'equivalent' geometrical positions. In other words, unlike sequence alignment, residue type is neglected.
Figure 1 Accessing PDBeFold form the PDBe homepage.

Try it for yourself....

1. Click on 'Structure similarity' arrowed in Figure 1.

**PDBeFold start page**

The introductory explanation tells you that PDBeFold expects you to supply a starting structure that will be analysed first and then used as a query to search the PDB [5] archive (Figure 2). After the search, PDBeFold will report any structures with a similar arrangement of secondary structure [9] elements (SSEs). It will rank these by how many elements it believes are significantly similar to those in your query structure, and it will then present the matches for you to look at and assess.
Try it for yourself....

1. Click the 'Launch PDBefold' button (Figure 2).

**PDBefold submission form**

The submission form also allows you to upload your own coordinate file, or multiple files - if for example you want to submit a family of proteins that share a particular fold. The submission form can be used to select a subset of the chains available in the query PDB [5] entries (or just a fragment of a chain) and also to alter any program parameters before running the search.
Try it for yourself....

1. Type in '1pvl' in the search query box labelled 'PDB code' in (circled in red in Figure 3). This will be used to search the entire PDB archive indicated as 'Target'.
2. For now just accept the defaults and click 'submit your query' (arrowed in Figure 3).
   You should then get a message that 'your request is being processed'. The results are shown in Figure 4.

Search results for PDB entry 1pvl

After the search has ended, matching structures ('hits') in the PDB [5] archive are presented in a table which may continue over several pages (Figure 4).

By default the results are ordered by Q-score [11] - an estimate of the quality of each match. The Q-score takes into account the number of residues in the matched SSEs and their positions in space. High Q-scores are obtained for structures where a large number of residues (in equivalent structural elements) superimpose well in three-dimensional space.
Figure 4 PDBeFold search results for 1pvl. Each line in the table is a hit in the archive. Some entries appear more than once owing to them having multiple chains.

For the 1pvl search run here, the PDBeFold results table gives you a number of well-matched structures. The actual number of presented structures can vary depending on the search parameters and the actual content of the PDB archive when the search is run. In your own work, you may want to go back and alter the initial search settings to suit your particular query.

The top hit in our search here (labelled 1 in the first column of the table) can be found in the 'Match' column. Unsurprisingly it is 1pvl with 100% of the SSEs matched.

Click on the top S component hit (mouse arrow in Figure 4) to see details of match (shown in Figure 5).

Displaying the results

Displaying your results
on the very last page of the results!

The interface provides you with a 'last page' button to go there quickly. Another good tip is to change the 'Sort by' pull-down (at the bottom of the table) to 'Seq %' instead of 'Q score'. The last entries on the last page will then have hits with little sequence identity to your query but high structural similarity - these may be homologues that you may have not found with a simple sequence search.

Although it may seem a bit odd to see our query returned from the search, PDBFold is searching the whole PDB[5] archive for us - so as long as 1pvl is in the archive, it should be found and be a perfect hit!

In fact several other hits appear in the top half of the table that are related structures of the Panton Valentine Leukotoxin[12] F component. This often happens as the PDB will typically contain multiple entries of the same protein from different experiments, crystal forms, or with different small molecules[13] bound. From its title you might guess that 2qk7 also has the S component in it; however, interestingly PDBFold has in fact only matched the F component (which is chain B in this entry)!

Although the top hits are not very interesting, our search for structural similarity between Panton Valentine Leukotoxin components has in fact succeeded in finding non-trivial hits, since four chains from 1t5r are returned in the top 20 matches.

**Getting more detail about your results**

To look in detail at any returned structural match just click on the underlined number in the first column (arrowed in Figure 4[14]).

The summary for the top 1t5r hit in our search (match 9) is shown in Figure 5. You can see that our Panton Valentine Leukotoxin[12] F component 1pvl query is on the left and the S component 1t5r match is on the right. In the middle are the details of the superimposition from the structural alignment that PDBFold has done. A total of 19 SSEs were matched satisfactorily between the two structures corresponding to 245 residues (Nalign) and with an RMSD[15] of 1.419Å.

Below the table is a quick summary of the secondary structure[9] alignment - ‘S’ represents aligned strands and ‘H’ represent aligned helices (‘s’ and ‘h’ are for elements that don’t align well or which correspond to a loop in one of the structures). The superimposed coordinates can also be downloaded from this page.
At the bottom of the results page is a detailed matching of residue ranges. Although complicated, this can be very helpful if, for example, you want to write out a residue selection to make an image with equivalent colouring between matched parts of two structures.

Finally, below the residue ranges summary is a detailed table of matches between the query and the hit on a residue-by-residue basis. This can be helpful in preparing a structure-based alignment of the protein sequences. Or it may highlight significant differences where a residue in the equivalent three-dimensional position in a domain has changed its character markedly - perhaps to adapt the fold to a new function.

Click on the 'view superposed' button (mouse arrow in Figure 5).

### Comparing the matches with your query

The 'view superposed' button (arrowed in Figure 5) launches a molecular viewer window showing the superimposed domains, similar to that in Figure 6. This is an interactive viewer so you can rotate the superimposition to get a good view of the toxin functional subdomains.
Figure 6 Jmol graphic display of superimposed 1pvl and 1t5r. Bracketting shows regions of ?-sandwich (blue), stem (red), and rim (yellow) subdomains.

In the view here we have also used coloured brackets alongside to show the positions of the stem (red), rim (yellow), and ?-sandwich (blue) subdomains. By rotating the superimposition in your own session, you should see that the elements of the query 1pvl are in darker colours whilst those of the matched 1tr5 structure are in lighter colours. Elements that could not be matched in the superimposition are shown in dark and light grey (for the query and the hit, respectively). Loops connecting matched SSEs are also coloured provided that their backbone atoms superimpose well.

Most of the time you can inspect the superimposition and recognise ?-strands or ?-helices that are most likely equivalent but have not been matched by PDBeFold because they have altered position too much between the two structures. The program has left them out to improve its Q-score [11] for the match, but sometimes looking by eye you will realise that the local structural differences are biologically significant.

Summary

- PDBeFold is an interactive service that allows you to identify structures that are similar to that of your reference protein.
- You can perform pairwise or multiple comparisons as well as 3D alignments of structures.
- Results are ordered by the PDBeFold program's estimate of the quality of each match, known as the Q-score [11].
- The most interesting results are generally displayed at the bottom of the table as they have the most structural similarity with your reference protein.
Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Learn more

Recommended courses

Train online courses available:

- [Biomacromolecular structures](#): An introduction to EBI resources;
- [PDBePISA](#) [17] tutorial - explore macromolecular interfaces and predict the quaternary structure of your protein;

Get help and support on PDBeFold

The EBI's [PDBe team](#) [20] develops and maintains the EBI's [Protein Data Bank in Europe](#) [21].

Support

- A list of frequently asked questions about PDBeFold can be found on our [website](#) [22].
- For general enquiries about the [PDB](#) [5], contact the pdbhelp [at] ebi.ac.uk (PDB helpdesk).

Collaborators

PDBe collaborates with the [X-ray crystallography](#) [23], [Nuclear Magnetic Resonance](#) [24] (NMR) spectroscopy and cryo-[Electron Microscopy](#) [25] (EM) communities. To keep abreast of new developments in the NMR community, PDBe has participated in EU projects and continues to contribute to the [Collaborative Computational Project for the NMR community](#) [26] (CCPN). PDBe also operates [EMDB](#) [27], the international repository for density maps, which are created using high-resolution biological transmission electron microscopy in collaboration with [RCSB](#) [28] and [Baylor College of Medicine](#) [29]. EMDB contains both macromolecular images and structures reconstructed using the single-particle method and images of sub-cellular regions from [electron tomography](#) [30].

Funding

Core support for PDBe is provided by the [European Molecular Biology Laboratory](#) [31] and [The Wellcome Trust](#) [32]. Specific projects are funded by the [European Union](#) [33], the [UK Biotechnology and Biosciences Research Council](#) [34] (BBSRC) and the [National Institutes of Health](#) [35] (NIH).
Contributors

Gary Battle [1]

EMBL-EBI
Project Leader Outreach - Kleywegt team: Protein Data Bank in Europe (EMBL-EBI alumni)

Gary was the outreach coordinator for the Protein Data Bank in Europe (PDBe). He was responsible for helping users with a broad range of backgrounds and interests to make the most of macromolecular structural data. Gary has a PhD in synthetic organic chemistry from the University of Warwick. Before joining the EMBL-EBI, he worked for over 10 years at the Cambridge Crystallographic Data Centre where he gained a wealth of experience in supporting software tools and data resources for pharmaceutical discovery, life science research and materials design.

Abhik Mukhopadhyay [36]

EMBL-EBI
Scientific Database Curator - Kleywegt team: Protein Data Bank in Europe

Source URL: https://www.ebi.ac.uk/training/online/course/pdbefold-searching-structural-homologues-protein

Links
[1] https://www.ebi.ac.uk/training/online/trainers/battle
[2] https://www.ebi.ac.uk/training/online/course/pdbe-quick-tour
[3] https://www.ebi.ac.uk/training/online/course/biomacromolecular-structures-introduction-ebi-reso
[5] https://www.ebi.ac.uk/training/online/glossary/pdb
[6] https://www.ebi.ac.uk/training/online/glossary/scop
[7] https://www.ebi.ac.uk/training/online/glossary/cath
[8] https://www.ebi.ac.uk/training/online/glossary/leukotoxin
[9] https://www.ebi.ac.uk/training/online/glossary/secondary-structure
[10] https://www.ebi.ac.uk/training/online/glossary/pdbe
[11] https://www.ebi.ac.uk/training/online/glossary/q-score