

ACTPDBCMP (Active Part of PDB Comparison) is a software tool for comparison and clustering PDB conformations of enzyme active sites created at the V.P. Kukhar Institute of Bioorganic Chemistry and Petrochemistry (IBOPC) of the National Academy of Sciences of Ukraine (<https://bpci.kiev.ua/en/>).

It has been already used to investigate binding site conformations of protein tyrosine phosphatase 1B.

<https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1747-0285.2012.01370.x>

Analyses of crystal structures of PTP1B were carried out using a special software tool created by the authors. The algorithm of ACTPDBCMP (Active Part of PDB Comparison) consists of the following steps.

- 1) Reading and ordering of enzyme chains.
 - a. All chains are read from specified PDB files. There may be more than 1 chain in a PDB file.
 - b. Chains are sorted alphabetically, e.g., 1Q6M-A, goes before 1Q6T-B and numbered.
- 2) Finding a fragment in all enzyme chains. The fragment is represented as a set of residues or as a separate PDB file.
 - a. A fragment is found in a chain if all residues are found (atoms may be missing). Residue numbering may be shifted, but relative distances must be the same (e.g., Arg24, Arg47, Phe181 (present in most chains) correspond to Arg524, Arg547, Phe681 or even to Arg1024, Arg1047, Phe1181 in some chains). A fragment found in a chain will be called chain fragment.

- b. If some residues are not found, the chain is removed from further analysis.
- 3) Pairwise comparison of all found chain fragments (in our case $102 \cdot (102 - 1) / 2 = 5,151$ comparisons). This includes the following steps for each pair of fragments.
 - a. Atom mapping.
 - i. For each atom in Fragment 1 a corresponding atom is found in Fragment 2. For example, CB of Arg47 corresponds to CB of Arg547. If we try to find corresponding atom for NH1 of Arg47, the situation is more complicated. We have to take into account the symmetry of Arg. In this case we have to consider both possible mappings (NH1-NH1, NH2-NH2) and (NH1-NH2, NH2-NH1) and choose the one that gives smaller sum of distances. The same thing happens in case of Leu, Val, Glu, Asp. Aromatic residues Tyr and Phe have two symmetric atom pairs, CD1/CD2 and CE1/CE2, which must be mapped simultaneously. A list of alternative atom mappings is formed for residues with symmetry.
 - ii. The number of corresponding atom pairs is determined. If an atom does not have a corresponding atom or there are several alternative ones, atom pair is not formed.
 - b. Structure fitting and calculation of root mean square deviation (RMSD).
 - i. Geometric centers of the chain fragments are superimposed.
 - ii. Quasi-random orientations (3-dimensional Sobol' sequence (34) are generated (the number depends on required fineness).

This is an attempt to make a global optimization. Otherwise we can find only the first local minimum.

- iii. For each starting orientation two structures are fitted, so that RMSD is minimized (both rotations and translations are used). RMSD between fragments is calculated over all pairs of corresponding atoms. We have tried different optimization techniques of which Powell's direction set method appeared to be the best (34). At each optimization step alternative mapping is considered for symmetric atoms.
 - iv. Minimized RMSD becomes RMSD between fragments.
 - v. RMSDs are calculated for each residue in a fragment. They will be used to determine mobility of each residue.
- 4) Matrix of RMSDs between residues is formed. For example, $D[1, 2]$ is RMSD between the chains 1 and 2 (ordered at step 1). This is called a similarity matrix. It is used for cluster analysis.
 - 5) Cluster analysis. Cluster analysis routine is based on OCLINK from the IMSL (International Mathematical and Statistical Libraries) (35). It performs complete-linkage hierarchical cluster analysis. A predefined number of clusters is selected.
 - 6) Mobility of each residue is determined. Mobility of a residue is an average residue's RMSD (step 3.b.v). It is averaged over all pairs of chain fragments (in our case 5,151 pairs).
 - 7) PDB files for each pair of overlapped fragments are optionally generated. We have done this only for cluster centroids.

ACTPDBCMP is a Visual Studio 2022 solution that has 3 projects.

- 1) ACTPDBCMP is a C++ project implementing the described above approach.

Usage:

There is one command line parameter. This is a path to config file containing following parameters:

Name	Default value	meaning
Template	-	Fragment being searched for (active part of an enzyme)
CA	0	If 1 only CA atoms are aligned, fast and dirty method
PI_steps	3	Defines precision, PI_steps * 5 random positions are taken for alignment optimization, though default value is 3, 1 could be enough.
OutPDB	0	If 1, PDB files with all overlapping sites are created in a subfolder pdbout. If there are 100 sites there will be $(100 - 1) * 100 / 2 = 45$ overlapping pairs.
Threads	4	Number of threads
OnlyBackbone	0	Use only backbone atoms for structure alignment.
pdbext	PDB	Extension of PDB files

ACTPDBCMP takes all PDB files in current directory and produces the following output files:

bad_files.txt – list of files where the site has not been found.

sites.txt – list of sites found in all PDB files, file name, chain, number of atoms are included.

dist_matr.txt – distance matrix for all sites, sites go in the same order as in sites.txt

cluster.txt – results of hierarchical cluster analysis, aggregated clusters start from N +1, where N is the number of sites.

res_impact.txt – a list of residues sorted by their mobility.

Residue name	RMS mobility	Relative mobility, %	Total no. of atom pairs compared	Atom pairs per site comparison	Max dist	Between		Min distance	between	
PHE682	2.87	100.0	1001	11.0	7.88	05qdiC	01bzj	0.00	05qdeB	05qdeA

- 2) ACLUST is a C++ project to select a predefined number of clusters from the results of hierarchical cluster analysis (cluster.txt produced by ACTPDBCMP)

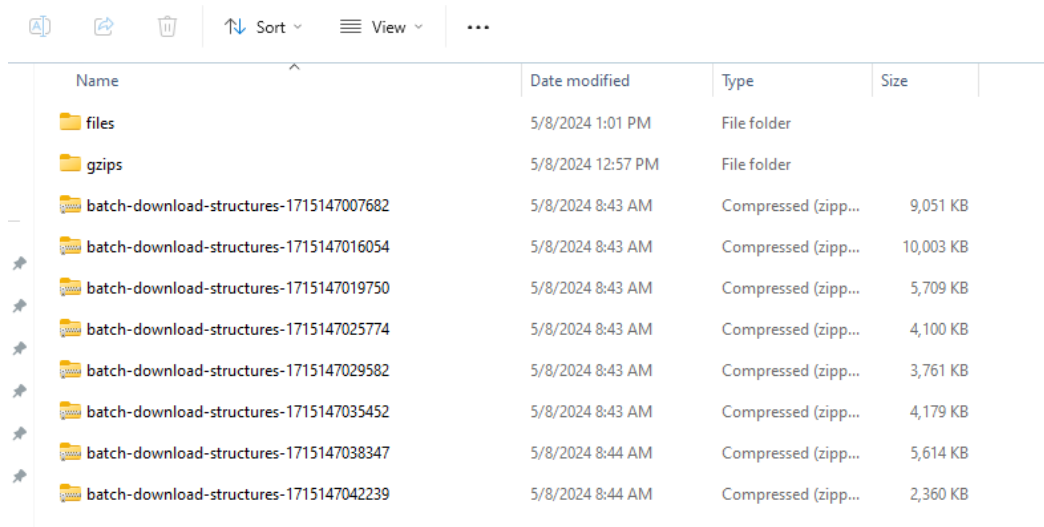
Usage:

There is one command line parameter. This is a path to config file containing following parameters:

Name	Default value	meaning
Cluster	cluster.txt	The name of cluster file produced by ACTPDBCMP
Sites	sites.txt	The name of sites file produced by ACTPDBCMP

Matrix	dist_matr.txt	The name of distance matrix file produced by ACTPDBCMP
Output	stdout	Output file name, if empty stdout (console) is used
Level	0.0	Minimal distance between output clusters
Number	5	Number of clusters to select, has effect only if Level not defined or < 0

3) UnpackPDB is a C# .net framework 5.0 utility to unpack multiple PDB files downloaded from <https://www.rcsb.org/>



Name	Date modified	Type	Size
files	5/8/2024 1:01 PM	File folder	
gzipts	5/8/2024 12:57 PM	File folder	
batch-download-structures-1715147007682	5/8/2024 8:43 AM	Compressed (zipp...	9,051 KB
batch-download-structures-1715147016054	5/8/2024 8:43 AM	Compressed (zipp...	10,003 KB
batch-download-structures-1715147019750	5/8/2024 8:43 AM	Compressed (zipp...	5,709 KB
batch-download-structures-1715147025774	5/8/2024 8:43 AM	Compressed (zipp...	4,100 KB
batch-download-structures-1715147029582	5/8/2024 8:43 AM	Compressed (zipp...	3,761 KB
batch-download-structures-1715147035452	5/8/2024 8:43 AM	Compressed (zipp...	4,179 KB
batch-download-structures-1715147038347	5/8/2024 8:44 AM	Compressed (zipp...	5,614 KB
batch-download-structures-1715147042239	5/8/2024 8:44 AM	Compressed (zipp...	2,360 KB

This is the example. 374 PDB files selected by “PTP1B” query are downloaded as 8 zip files containing separate *.pdb.zip files for each PDB

file. If you run UnpackPDB in this directory you will get all *.pdb.gzip files in “gzs” subfolder and all unpacked PDB files in “files” subfolder.

ACTBDBCMP and ACLUST can be compiled by g++ in Linux/Unix environment. Each project folder contains a makefile.

Test data can be found in the tst folder:

- 1) pdb subfolder – a few PTP1B files for testing, 1b.ent is an active site to be found in all other files. There is also a conig file config.txt with parameters for ACTPDBCMP and ACLUST
- 2) unpack subfolder – test data for UnpackPDB.