Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Web Server**

DECOMP: A PDB decomposition tool on the web

Rafael Ordog[1,2], Zoltán Szabadka[1,2,3], Vince Grolmusz[1,2]*

[1]Protein Information Technology Group, Eötvös University, 1117 Budapest, Hungary; [2]Uratim Ltd. 4400 Nyíregyháza, Hungary; [3]Google Inc, Europe; Vince Grolmusz – Email: grolmusz@cs.elte.hu; * Corresponding author

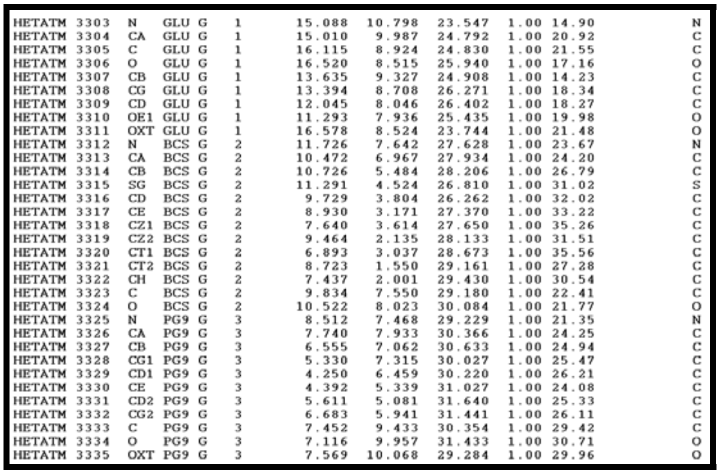
Received February 22, 2009; accepted May 22, 2009; published July 27, 2009

Abstract:
The protein databank (PDB) contains high quality structural data for computational structural biology investigations. We have earlier described a fast tool (the decomp_pdb tool) for identifying and marking missing atoms and residues in PDB files. The tool also automatically decomposes PDB entries into separate files describing ligands and polypeptide chains. Here, we describe a web interface named DECOMP for the tool. Our program correctly identifies multi-monomer ligands, and the server also offers the preprocessed ligand-protein decomposition of the complete PDB for downloading (up to size: 5GB)

Availability: http://decomp.pitgroup.org

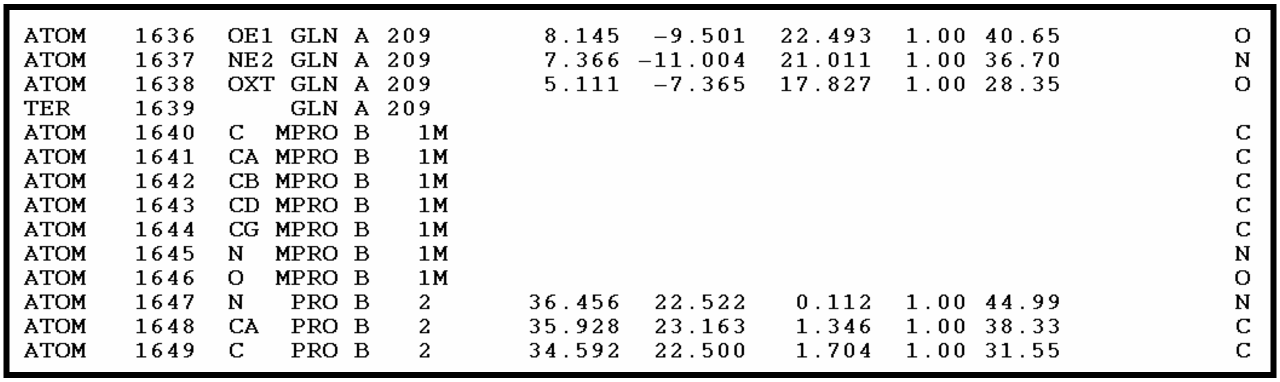
Keywords: PDB, web tool, decomposition, server, ligands, SEQRES

Background:
The Protein Data Bank **[1]** started to function as the depository of the crystallographic data, complementing journal publications: researchers solved the structure of a protein, wrote a paper on the result, and deposited the data of the solution in the publicly available PDB. The irregularities of the structure deposited (such as lacking atomic coordinates, broken chains, unidentified substructures) are mostly remarked in the cited publications and also in the remark-fields of the PDB file. The textual annotations in the scientific publication elsewhere or in the remark-fields in the very same PDB-file, however, make the automatic processing of the protein-structures very difficult. This statement may be a little bit confusing, since atoms, carrying the HET label are not supposed to be in the peptide-chain, so those structures that contains HET atoms other than the oxygen of the water would qualify for being a complex. Unfortunately, this is not the case. Metal ions, modified residues (in a surprisingly large number), and small molecules added in the crystallization all contain heteroatoms, and they are frequently not considered to be ligands. With our decomp_pdb program **[2]** protein-ligand complexes are identified reliably, and the ligands are deposited in separate files. Missing residues and atoms in chains are handled properly, that is, even if several atoms are missing from a chain our algorithm will still not recognize the parts as distinct chains. Placeholders are inserted into chains for missing residues/atoms (an example is given in **Figure 2**), denoting that the objects were not measured crystallographically, but - according to the more reliable sequence information - they should be there. This way our algorithm "repairs" faulty PDB's, or recognizes that flexible chain sequences are present. We should remark, that missing atoms are usually a sign of mobile loop or string in the protein-crystal, since flexible atoms will not give usable electron density maps. Consequently, mapping missing atoms this way may help to automatically identify flexible protein parts. Ligands are identified without using the HET-atom labels, properly handling modified residues and small artifacts, due to crystallization protocols. CONECT records of the ligand-atoms are computed automatically (these records for the ligands generally are not present in the PDB file).



```
HETATM 3303  N   GLU G   1      15.088  10.798  23.547  1.00 14.90           N
HETATM 3304  CA  GLU G   1      15.010   9.987  24.792  1.00 20.92           C
HETATM 3305  C   GLU G   1      16.115   8.924  24.830  1.00 21.55           C
HETATM 3306  O   GLU G   1      16.520   8.515  25.940  1.00 17.16           O
HETATM 3307  CB  GLU G   1      13.635   9.327  24.908  1.00 14.23           C
HETATM 3308  CG  GLU G   1      13.394   8.708  26.271  1.00 18.34           C
HETATM 3309  CD  GLU G   1      12.045   8.046  26.402  1.00 18.27           C
HETATM 3310  OE1 GLU G   1      11.293   7.936  25.435  1.00 19.98           O
HETATM 3311  OXT GLU G   1      16.578   8.524  23.744  1.00 21.48           O
HETATM 3312  N   BCS G   2      11.726   7.642  27.628  1.00 23.67           N
HETATM 3313  CA  BCS G   2      10.472   6.967  27.934  1.00 24.20           C
HETATM 3314  CB  BCS G   2      10.726   5.484  28.206  1.00 26.79           C
HETATM 3315  SG  BCS G   2      11.291   4.524  26.810  1.00 31.02           S
HETATM 3316  CD  BCS G   2       9.729   3.804  26.262  1.00 32.02           C
HETATM 3317  CE  BCS G   2       8.930   3.171  27.370  1.00 33.22           C
HETATM 3318  CZ1 BCS G   2       7.640   3.614  27.650  1.00 35.26           C
HETATM 3319  CZ2 BCS G   2       9.464   2.135  28.133  1.00 31.51           C
HETATM 3320  CT1 BCS G   2       6.893   3.037  28.673  1.00 35.56           C
HETATM 3321  CT2 BCS G   2       8.723   1.550  29.161  1.00 27.28           C
HETATM 3322  CH  BCS G   2       7.437   2.001  29.430  1.00 30.54           C
HETATM 3323  C   BCS G   2       9.834   7.550  29.180  1.00 22.41           C
HETATM 3324  O   BCS G   2      10.522   8.023  30.084  1.00 21.77           O
HETATM 3325  N   PG9 G   3       8.512   7.468  29.229  1.00 21.35           N
HETATM 3326  CA  PG9 G   3       7.740   7.933  30.366  1.00 24.25           C
HETATM 3327  CB  PG9 G   3       6.555   7.062  30.633  1.00 24.94           C
HETATM 3328  CG1 PG9 G   3       5.330   7.315  30.027  1.00 25.47           C
HETATM 3329  CD1 PG9 G   3       4.250   6.459  30.220  1.00 26.21           C
HETATM 3330  CE  PG9 G   3       4.392   5.339  31.027  1.00 24.08           C
HETATM 3331  CD2 PG9 G   3       5.611   5.081  31.640  1.00 25.33           C
HETATM 3332  CG2 PG9 G   3       6.683   5.941  31.441  1.00 26.11           C
HETATM 3333  C   PG9 G   3       7.452   9.433  30.354  1.00 29.42           C
HETATM 3334  O   PG9 G   3       7.116   9.957  31.433  1.00 30.71           O
HETATM 3335  OXT PG9 G   3       7.569  10.068  29.284  1.00 29.96           O
```

Figure 1: The DECOMP_PDB output-ligand 10gs.pdb.out.lig.3 contains the 3-monomer GLU-BCS-PG9 molecule correctly, in one single file, even if it contains three monomer ID’s.



```
ATOM   1636  OE1 GLN A 209       8.145  -9.501  22.493  1.00 40.65           O
ATOM   1637  NE2 GLN A 209       7.366 -11.004  21.011  1.00 36.70           N
ATOM   1638  OXT GLN A 209       5.111  -7.365  17.827  1.00 28.35           O
TER    1639      GLN A 209
ATOM   1640  C   MPRO B   1M                                                 C
ATOM   1641  CA  MPRO B   1M                                                 C
ATOM   1642  CB  MPRO B   1M                                                 C
ATOM   1643  CD  MPRO B   1M                                                 C
ATOM   1644  CG  MPRO B   1M                                                 C
ATOM   1645  N   MPRO B   1M                                                 N
ATOM   1646  O   MPRO B   1M                                                 O
ATOM   1647  N    PRO B   2       36.456  22.522   0.112  1.00 44.99           N
ATOM   1648  CA   PRO B   2       35.928  23.163   1.346  1.00 38.33           C
ATOM   1649  C    PRO B   2       34.592  22.500   1.704  1.00 31.55           C
```

Figure 2: Atoms or residues are frequently missing at the beginning or at the end of polypeptide chains. In this example a missing residue and six missing atoms are identified at the beginning of chain B of pdb entry 10gs.

Methodology:

Our program selects atoms from the PDB entry that are part of a protein or DNA chain. We do not use the chain-identifier for this purpose. However, we use SEQRES data and refined graph-theoretical algorithms described elsewhere **[2].** It selects the water molecules, and removes them from the set of possible ligand atoms. Then metal and other small ions are selected, that will not be considered as ligands. A complete list of residue names that were considered as ions (so not as ligands) is given in the file ion_list.txt. All the remaining atoms will form the set of ligand atoms. Within this set, we use a graph-theory component detecting algorithm, so a ligand is defined as a connected component of the graph formed by the ligand atoms as vertices and the covalent bonds between the ligand atoms as the edges.

Functionality:

The DECOMP tool correctly identifies ligand molecules, even if they are composed of more than one monomers. For example, when decomposing PDB entry 10GS with options "Export ligands", the file 10gs.pdb.out.lig.3 contains the 3-monomer GLU-BCS-PG9 molecule correctly **(Figure 1)**.

Utility:

Provide a list of PDB codes in the appropriate box at the web server and check the desired options. The PDB codes should be separated either by "spaces" or "new line" characters. Press the "schedule job" button and the request will be inserted into a queue. Progress is monitored in the "Log window". The result will be a link in the "Log window" to a tar.gz file. The result file contains one directory for each of the pdb's listed. Each of these directories contains an error log with ".pdb.error" extension, the decomposed pdb file with ".pdb" extension, and if "Export ligands" or "Export ions" option was specified, than a separate file is present for each of the ligands or ions. An error file is presented if there was a fatal error while processing the PDB file. The result files are usually viewed by popular PDB viewer tools. A pre-processed, constantly updated compressed file can be downloaded with the results when the entire PDB file has be decomposed. The result files are stored for 3 days, and log files are stored for 30 days in the server.

Acknowledgements:

The authors acknowledge the partial funding of the following agencies: NKTH: OMFB-01295/2006 and OM-00219/2007 through Uratim Ltd. and OTKA NK-67867 through Eotvos University.

References:

[1] HM Berman *et al., Nucl. Acids Res.,* (2000) 28: 235
[2] Z Szabadka, V Grolmusz, *J. Mol. Graph. Mod.* (2007) 25: 831

Edited by P. Kangueane
Citation: Ordog *et al*, Bioinformation 3(10): 413-414 (2009)