

## NAME

ExtractFromPDBFiles.pl - Extract specific data from PDBFile(s)

## SYNOPSIS

ExtractFromPDBFiles.pl PDBFile(s)...

```
ExtractFromPDBFiles.pl [-c, --chains First | All | "ChainID, [ChainID,...]"] [--CombineChains> yes | no] [-d, --distance number]
[--DistanceMode Atom | Hetatm | Residue | XYZ] [--DistanceOrigin "AtomNumber, AtomName" | "HetatmNumber,
HetAtmName" | "ResidueName, ResidueNumber, [ChainID]" | "X,Y,Z">] [-h, --help] [-k, --KeepOldRecords yes | no] [-m,
--mode Chains | Sequences | Atoms | CAlphas | ResidueNums | ResiduesRange | Distance | NonWater | NonHydrogens] [
--ModifyHeader yes | no] [--NonStandardKeep yes | no] [--NonStandardCode character] [-o, --overwrite] [-r, --root
rootname] [--Residues "ResidueNum,[ResidueNum...]" | StartResidueNum,EndResiduNum ] [--SequenceLength number] [
--SequenceRecords Atom | SeqRes] [--SequenceIDPrefix FileName | HeaderRecord | Automatic] [--WaterResidueNames
Automatic | "ResidueName, [ResidueName,...]"] [-w, --WorkingDir dirname] PDBFile(s)...
```

## DESCRIPTION

Extract specific data from *PDBFile(s)* and generate appropriate PDB or sequence file(s). Multiple PDBFile names are separated by spaces. The valid file extension is *.pdb*. All other file name extensions are ignored during the wild card expansion. All the PDB files in a current directory can be specified either by *\*.pdb* or the current directory name.

During *Chains* and *Sequences* values of *-m, --mode* option, all ATOM/HETAM records for chains after the first model in PDB files containing data for multiple models are ignored.

## OPTIONS

*-c, --chains First | All | ChainID,[ChainID,...]*

Specify which chains to extract from *PDBFile(s)* during *Chains | Sequences* value of *-m, --mode* option: first chain, all chains, or a specific list of comma delimited chain IDs. Possible values: *First | All | ChainID,[ChainID,...]*. Default: *First*. Examples:

```
A
A,B
All
```

*--CombineChains yes | no*

Specify whether to combine extracted chains data into a single file during *Chains* or *Sequences* value of *-m, --mode* option. Possible values: *yes | no*. Default: *no*.

During *Chains* value of *<-m, --mode>* option with *Yes* value of *<--CombineChains>*, extracted data for specified chains is written into a single file instead of individual file for each chain.

During *Sequences* value of *<-m, --mode>* option with *Yes* value of *<--CombineChains>*, residues sequences for specified chains are extracted and concatenated into a single sequence file instead of individual file for each chain.

*-d, --distance number*

Specify distance used to extract ATOM and HETATM records during *Distance* value of *-m, --mode* option. Default: *10.0* angstroms.

*--DistanceMode Atom | Hetatm | Residue | XYZ*

Specify how to extract ATOM and HETATM records from *PDBFile(s)* during *Distance* value of *-m, --mode* option: extract all the records within a certain distance specified by *-d, --distance* from an atom or hetroatom record, a residue, or any arbitrary point. Possible values: *Atom | Hetatm | Residue | XYZ*. Default: *XYZ*.

During *Residue* value of *--distancemode*, distance of ATOM/HETATM records is calculated from all the atoms in the residue and the records are selected as long as any atom of the residue lies within the distance specified using *-d, --distance* option.

*--DistanceOrigin "AtomNumber,AtomName" | "HetatmNumber,HetAtmName" | "ResidueName,ResidueNumber,[ChainID]" | "X,Y,Z"*

This value is *--distancemode* specific. In general, it identifies a point used to select other ATOM/HETATMS within a specific distance from this point.

For *Atom* value of *--distancemode*, this option corresponds to an atom specification. Format: *AtomNumber,AtomName*. Example:

```
455,CA
```

For *Hetatm* value of *--distancemode*, this option corresponds to a hetatm specification. Format: *HetatmNumber,HetAtmName*. Example:

```
5295,C1
```

For *Residue* value of *--distancemode*, this option corresponds to a residue specification. Format: *ResidueNumber, ResidueName[,ChainID]*. Example:

```
78,MSE
977,RET,A
978,RET,B
```

For *XYZ* value of `--distancemode`, this option corresponds to a coordinate of an arbitrary point. Format: *X,Y,X*. Example:

10.044,19.261,-4.292

`-h, --help`

Print this help message.

`-k, --KeepOldRecords yes | no`

Specify whether to transfer old non ATOM and HETATM records from input PDBFile(s) to new PDBFile(s) during *Chains* | *Atoms* | *HetAtms* | *CAalphas* | *Distance* | *NonWater* | *NonHydrogens* value of `-m --mode` option. By default, except for the HEADER record, all other unnecessary non ATOM/HETATM records are dropped during the generation of new PDB files. Possible values: *yes* | *no*. Default: *no*.

`-m, --mode Chains | Sequences | Atoms | CAalphas | ResidueNums | ResiduesRange | Distance | NonWater | NonHydrogens`

Specify what to extract from *PDBFile(s)*: *Chains* - retrieve records for specified chains; *Sequences* - generate sequence files for specific chains; *Atoms* - extract atom records; *CAalphas* - extract records for alpha carbon atoms corresponding to amino acids; *ResidueNums* - extract records for specified residue numbers; *ResiduesRange* - extract records for residues between specified residue number range; *Distance* - extract records with in a certain distance from a specific position; *NonWater* - extract records corresponding to residues other than water; *NonHydrogens* - extract non-hydrogen records.

Possible values: *Chains, Sequences Atoms, CAalphas, ResidueNums, ResiduesRange, Distance*. Default value: *NonWater*

During the generation of new PDB files, unnecessary CONECT records are dropped.

For *Chains* mode, data for appropriate chains specified by `--c --chains` option is extracted from *PDBFile(s)* and placed into new PDB file(s).

For *Sequences* mode, residues names using various sequence related options are extracted for chains specified by `--c --chains` option from *PDBFile(s)* and FASTA sequence file(s) are generated.

For *Distance* mode, all ATOM and HETATM records with in a distance specified by `-d --distance` option from a specific atom, residue or a point indicated by `--distancemode` are extracted and placed into new PDB file(s).

For *NonWater* mode, non water ATOM and HETATM record lines, identified using value of `--WaterResidueNames`, are extracted and written to new PDB file(s).

For *NonHydrogens* mode, ATOM and HETATOM record lines containing element symbol other than *H* are extracted and written to new PDB file(s).

For all other options, appropriate ATOM and HETATM records are extracted to generate new PDB file(s).

`--ModifyHeader yes | no`

Specify whether to modify HEADER record during the generation of new PDB files for `-m, --mode` values of *Chains* | *Atoms* | *CAalphas* | *Distance*. Possible values: *yes* | *no*. Default: *yes*. By default, Classification data is replaced by *Data extracted using MayaChemTools* before writing out HEADER record.

`--NonStandardKeep yes | no`

Specify whether to include and convert non-standard three letter residue codes into a code specified using `--nonstandardcode` option and include them into sequence file(s) generated during *Sequences* value of `-m, --mode` option. Possible values: *yes* | *no*. Default: *yes*.

A warning is also printed about the presence of non-standard residues. Any residue other than standard 20 amino acids and 5 nucleic acid is considered non-standard; additionally, HETATM residues in chains also tagged as non-standard.

`--NonStandardCode character`

A single character code to use for non-standard residues. Default: *X*. Possible values: *?, -, or X*.

`-o, --overwrite`

Overwrite existing files.

`-r, --root rootname`

New PDB and sequence file name is generated using the root: `<Root><Mode>.<Ext>`. Default new file name: `<PDBFileName>Chain<ChainID>.pdb` for *Chains* mode; `<PDBFileName>SequenceChain<ChainID>.fasta` for *Sequences* mode; `<PDBFileName>DistanceBy<DistanceMode>.pdb` for *Distance* `-m, --mode` `<PDBFileName><Mode>.pdb` for *Atoms* | *CAalphas* | *NonWater* | *NonHydrogens* `-m, --mode` values. This option is ignored for multiple input files.

`--Residues "ResidueNum,[ResidueNum...]" | StartResidueNum,EndResiduNum`

Specify which residue records to extract from *PDBFiles(s)* during *ResidueNums* and *ResiduesRange* value of `-m, --mode` option: extract records corresponding to residue numbers specified in a comma delimited list of residue numbers or with in the range of start and end residue numbers. Possible values: "*ResidueNum*,[*ResidueNum*,...]" or *StartResidueNum*,*EndResiduNum*. Default: *None*.

`--SequenceLength number`

Maximum sequence length per line in sequence file(s). Default: *80*.

`--SequenceRecords Atom | SeqRes`

Specify which records to use for extracting residue names from *PDBFiles(s)* during *Sequences* value of *-m*, *--mode* option: use ATOM records to compile a list of residues in a chain or parse SEQRES record to get a list of residues. Possible values: *Atom* / *SeqRes*. Default: *Atom*.

*--SequenceIDPrefix* *FileName* / *HeaderRecord* / *Automatic*

Specify how to generate a prefix for sequence IDs during *Sequences* value of *-m*, *--mode* option: use input file name prefix; retrieve PDB ID from HEADER record; or automatically decide the method for generating the prefix. The chain IDs are also appended to the prefix. Possible values: *FileName* / *HeaderRecord* / *Automatic*. Default: *Automatic*

*--WaterResidueNames* *Automatic* / "*ResidueName*,*[ResidueName,...]*"

Identification of water residues during *NonWater* value of *-m*, *--mode* option. Possible values: *Automatic* / "*ResidueName*,*[ResidueName,...]*". Default: *Automatic* - corresponds to "HOH,WAT,H2O". You can also specify a different comma delimited list of residue names to use for water.

*-w*, *--WorkingDir* *dirname*

Location of working directory. Default: current directory.

## EXAMPLES

To extract non-water records from Sample2.pdb file and generate Sample2NonWater.pdb file, type:

```
% ExtractFromPDBFiles.pl Sample2.pdb
```

To extract non-water records from Sample2.pdb file using HOH or WAT residue name for water along with all old non-coordinate records and generate Sample2NewNonWater.pdb file, type:

```
% ExtractFromPDBFiles.pl -m NonWater --WaterResidueNames "HOH,WAT"
-KeepOldRecords Yes -r Sample2New -o Sample2.pdb
```

To extract non-hydrogens records from Sample2.pdb file and generate Sample2NonHydrogen.pdb file, type:

```
% ExtractFromPDBFiles.pl -m NonHydrogens Sample2.pdb
```

To extract data for first chain in Sample2.pdb and generate Sample2ChainA.pdb, type file, type:

```
% ExtractFromPDBFiles.pl -m chains -o Sample2.pdb
```

To extract data for both chains in Sample2.pdb and generate Sample2ChainA.pdb and Sample2ChainB.pdb, type:

```
% ExtractFromPDBFiles.pl -m chains -c All -o Sample2.pdb
```

To extract data for alpha carbons in Sample2.pdb and generate Sample2CAlphas.pdb, type:

```
% ExtractFromPDBFiles.pl -m CAlphas -o Sample2.pdb
```

To extract records for specific residue numbers in all chains from Sample2.pdb file and generate Sample2ResidueNums.pdb file, type:

```
% ExtractFromPDBFiles.pl -m ResidueNums --Residues "3,6"
Sample2.pdb
```

To extract records for a specific range of residue number in all chains from Sample2.pdb file and generate Sample2ResiduesRange.pdb file, type:

```
% ExtractFromPDBFiles.pl -m ResiduesRange --Residues "10,30"
Sample2.pdb
```

To extract data for all ATOM and HETATM records with in 10 angstrom of an atom specified by atom serial number and name "1,N" in Sample2.pdb file and generate Sample2DistanceByAtom.pdb, type:

```
% ExtractFromPDBFiles.pl -m Distance --DistanceMode Atom
--DistanceOrigin "1,N" -k No --distance 10 -o Sample2.pdb
```

To extract data for all ATOM and HETATM records with in 25 angstrom of an arbitrary point "0,0,0" in Sample2.pdb file and generate Sample2DistanceByXYZ.pdb, type:

```
% ExtractFromPDBFiles.pl -m Distance --DistanceMode XYZ
--DistanceOrigin "0,0,0" -k No --distance 25 -o Sample2.pdb
```

AUTHOR

Manish Sud <msud@san.rr.com>

SEE ALSO

InfoPDBFiles.pl, ModifyPDBFiles.pl

COPYRIGHT

Copyright (C) 2004-2012 Manish Sud. All rights reserved.

This file is part of MayaChemTools.

MayaChemTools is free software; you can redistribute it and/or modify it under the terms of the GNU Lesser General Public License as published by the Free Software Foundation; either version 3 of the License, or (at your option) any later version.