NAME

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

SYNOPSIS

ExtractFromPDBFiles.pl PDBFile(s)...

ExtractFromPDBFiles.pl [-a, --Atoms "AtomNum, [AtomNum...]" | "StartAtomNum, EndAtomNum" | "AtomName, [AtomName...]"] [-c, --chains First | All | "ChainID, [ChainID,...]"] [--ChainsRecordMode *AcrossTER* | *NotAcrossTER*] [<--CombineChains> yes | no] [-d, --distance number] [--DistanceMode Atom | Hetatm | Residue | XYZ] [--DistanceOrigin "AtomNumber, AtomName" | "HetatmNumber, HetAtmName" | "ResidueNumber, ResidueName, [ChainID]" | "X,Y,Z">] [<--DistanceSelectionMode> ByAtom | ByResidue] [-h, --help] [-k, --KeepOldRecords yes | no] [-m, --mode Chains | Sequences | Atoms | CAlphas | AtomNums | AtomsRange | AtomNames | ResidueNums | ResiduesRange | ResidueNames | Distance | NonWater | NonHydrogens] [--ModifyHeader yes | no] [--NonStandardKeep yes | no] [--NonStandardCode character] [-o, --overwrite] [-r, --root rootname] --RecordMode *Atom* | *Hetatm* | *AtomAndHetatm*] [--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 fils containing data for multiple models are ignored.

OPTIONS

-a, --Atoms *"AtomNum,[AtomNum...]"* | *"StartAtomNum,EndAtomNum"* | *"AtomName,[AtomName...]"*

Specify which atom records to extract from *PDBFiles(s)* during *AtomNums*, *AtomsRange*, and *AtomNames* value of -m, --mode option: extract records corresponding to atom numbers specified in a comma delimited list of atom numbers/names, or with in the range of start and end atom numbers. Possible values: *"AtomNum[,AtomNum,..]"*, *StartAtomNum,EndAtomNum*, or *"AtomName[,AtomName,..]"*. Default: *None*. Examples:

```
10
15,20
N,CA,C,O
```

-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
```

--ChainsRecordMode *AcrossTER* | *NotAcrossTER*

Specify whether to extract ATOM and HETATM record lines across TER records from *PDBFile(s)* during *Chains* value of -m, --mode option. Possible values: *AcrossTER* | *NotAcrossTER*. Defaul value: *NotAcrossTER*.

This option allows retrieval ATOM and HETATM record lines for a specific chain which spread across TER record in *PDBFile(s)*.

--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/HETATM recods during *Distance* value of -m, --mode option. Default: *10.0* angstroms.

--RecordMode option controls type of record lines to extract from *PDBFile(s)*: ATOM, HETATM or both.

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

Specify how to extract ATOM/HETATM records from *PDBFile(s)* during *Distance* value of -m, --mode option: extract all the records within a certain distance specifed by -d, --distance from an atom or hetro atom record, a residue, or any artbitrary 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 with in the distace specified using -d, --distance option.

--RecordMode option controls type of record lines to extract from *PDBFile(s)*: ATOM, HETATM or both.

--DistanceSelectionMode *ByAtom* | *ByResidue*

Specify how how to extract ATOM/HETATM records from *PDBFile(s)* during *Distance* value of -m, --mode option for all values of --DistanceMode option: extract only those ATOM/HETATM records that meet specified distance criterion; extract all records corresponding to a residue as long as one of the ATOM/HETATM record in the residue satisfies specified distance criterion. Possible values: *ByAtom, ByResidue*. Default value: *ByAtom*.

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

This value is --distancemode specific. In general, it identifies a point used to select other ATOM/HETATMS with in 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
```

--RecordMode option controls type of record lines to extract from *PDBFile(s)*: ATOM, HETATM or both.

-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* | *CAlphas* | *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* | *CAlphas* | *AtomNums* | *AtomsRange* | *AtomNames* | *ResidueNums* | *ResiduesRange* | *ResidueNames* | *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; *CAlphas* - extract atom records for alpha carbon atoms; *AtomNums* - extract atom records for specified atom numbers; *AtomsRange* - extract atom records between specified atom number range; *AtomNames* - extract atom records for specified atom names; *ResidueNums* - extract records for specified residue numbers; *ResiduesRange* - extract records for residues between specified residue number range; *ResidueNames* - extract records for specified residue names; *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, CAlphas, AtomNums, AtomsRange, AtomNames, ResidueNums, ResiduesRange, ResidueNames, Distance, NonWater, NonHydrogens*. Default value: *NonWater*

During the generation of new PDB files, unnecessay 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/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/HETATM record lines, identified using value of --WaterResidueNames, are extracted and written to new PDB file(s).

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

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

--RecordMode option controls type of record lines to extract and process from *PDBFile(s)*: ATOM, HETATM or both.

--ModifyHeader *yes* | *no*

Specify whether to modify HEADER record during the generation of new PDB files for -m, --mode values of *Chains* | *Atoms* | *CAlphas* | *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* | *CAlphas* | *NonWater* | *NonHydrogens* -m, --mode values. This option is ignored for multiple input files.

--RecordMode *Atom* | *Hetatm* | *AtomAndHetatm*

Specify type of record lines to extract and process from *PDBFile(s)* during various values of -m, --mode option: extract only ATOM record lines; extract only HETATM record lines; extract both ATOM and HETATM lines. Possible values: *Atom* | *Hetatm* | *AtomAndHetatm* | *XYZ*. Default during *Atoms, CAlphas, AtomNums, AtomsRange, AtomNames* values of -m, --mode option: *Atom*; otherwise: *AtomAndHetatm*.

This option is ignored during *Sequences* values of -m, --mode option.

--Residues *"ResidueNum,[ResidueNum...]"* | *"StartResidueNum,EndResiduNum"* | *"ResidueName,[ResidueName...]"*

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

```
20
5,10
```

TYR,SER,THR

--RecordMode option controls type of record lines to extract from *PDBFile(s)*: ATOM, HETATM or both.

--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,H20". 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 corresponding to only ATOM records from Sample2.pdb file and generate Sample2NonWater.pdb file, type:

```
% ExtractFromPDBFiles.pl --RecordMode Atom 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 specifed 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 for complete residues with any atom or hetatm less than 10 angstrom of an atom specifed 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" --DistanceSelectionMode ByResidue
  -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