CLUSTER
Cluster analysis of UNRES simulation results

Department of Molecular Modeling
Faculty of Chemistry
University of Gdansk
Sobieskiego 18
80-952 Gdansk, Poland

Scheraga Group
Baker Laboratory of Chemistry
and Chemical Biology
Cornell University
Ithaca, NY 14853-1303, USA

September 28, 2012

Contents

1 LICENSE TERMS

- This software is provided free of charge to academic users, subject to the condition that no part of it be sold or used otherwise for commercial purposes, including, but not limited to its incorporation into commercial software packages, without written consent from the authors. For permission contact Prof. H. A. Scheraga, Cornell University.
- This software package is provided on an "as is" basis. We in no way warrant either this software or results it may produce.
- Reports or publications using this software package must contain an acknowledgment to the authors and the NIH Resource in the form commonly used in academic research.

2 REFERENCES

The program incorporates the hierarchical-clustering subroutine, hc.f written by G. Murtagh (refs 1 and 2). The subroutine contains seven methods of hierarchical clustering.

[1] Murtagh. Multidimensional clustering algorithms; Physica-Verlag: Vienna, Austria, 1985.

[2] F. Murtagh, A. Heck. MultiVariate data analysis; Kluwer Academic: Dordrecht, Holland, 1987.

[3] A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Oldziej, K. Wachucik, H.A. Scheraga. Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B*, **2007**, 111, 260-285.

[4] S. Oldziej, A. Liwo, C. Czaplewski, J. Pillardy, H.A. Scheraga. Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 2. Off-lattice tests of the method with single proteins. *J. Phys. Chem. B.*, **2004**, 108, 16934-16949.

3 FUNCTIONS OF THE PROGRAM

The program runs cluster analysis of UNRES simulation results. There are two versions of the program depending on the origin of input conformation:

1. CLUST-UNRES: performs cluster analysis of conformations that are obtained directly from UNRES runs (CSA, MCM, MD, (M)REMD, multiple-conformation energy minimization). The source code and other important files are deposited in CLUST-UNRES subdirectory

   The source code of this version is deposited in clust-unres/src

2. CLUST-WHAM: performs cluster analysis of conformations obtained in UNRES MREMD simulations and then processed with WHAM (weighted histogram analysis method). This enables the user to obtain clusters as conformational ensembles at a given temperature and to compute their probabilities (section 2.5 of ref 3). This version is deposited in the CLUST-WHAM subdirectory. This version has single- and multichain variants, whose source codes are deposited in the following subdirectories:

   (a) clust-wham/src single-chain proteins

   (b) clust-wham/src-M oligomeric proteins

The version developed for oligomeric proteins treats whole system as a single chain with dummy residues inserted. It also works for single chains but is not fully checked and it is recommended to use single-chain version for single-chain proteins.

4 INSTALLATION

Customize Makefile to your system. See section 7 of the description of UNRES for compiler flags that are used to created executables for a particular force field. There are already several Makefiles prepared for various systems and force fields.

Run make in the appropriate source directory version. CLUST-UNRES runs only in single-processor mode an CLUST-WHAM runs in both serial and parallel mode [only conformation-distance (rmsd) calculations are parallelized]. The parallel version uses MPI.

5 RUNNING THE PROGRAM

The program requires a parallel system to run. Depending on system, either the wham.csh C-shell script (in WHAM/bin directory) can be started using mpirun or the binary in the C-shell script must be executed through mpirun. See the wham.csh C-shell script and section 6 for the files processed by the program.

6 INPUT AND OUTPUT FILES

6.1 Summary of files

The C-shell script wham.csh is used to run the program (see the bin/WHAM directory). The data files that the script needs are mostly the same as for UNRES (see section 6 of UNRES description). In addition, the environmental variable CONTFUN specifies the method to assess whether two side chains are at contact; if EONTFUN=GB, the criterion defined by eq 8 of ref 4 is used to assess whether two side chains are at contact. Also, the parameter files from the C-shell scripts are overridden if the data from Hamiltonian MREMD are processed; if so, the parameter files are defined in the main input file.

The main input file must have inp extension. If it is INPUT.inp, the output files are as follows:

Coordinate input file COORD.ext, where ext denotes file extension in one of the following formats:

int (extension int; UNRES angles theta, gamma, alpha, and beta),

x (extension x; UNRES Cartesian coordinate format; from MD),

pdb (extension pdb; Protein Data Bank format; fro MD),

cx (extension cx; xdrf format; from WHAM).

INPUT_clust.out (single-processor mode) or INPUT_clust.out_xxx (parallel mode) – output file(s) (INPUT.out_000 is the main output file for parallel mode).

COORD_clust.int – leading (lowest-energy) members of the families. in internal-coordinate format.

COORD_clust.x – leading members of the families in UNRES Cartesian coordinate format.

COORD_xxxx.pdb or COORD_xxxx_yyy.pdb (CLUST-UNRES) – PDB file of member yyy of family xxxx; yyy is omitted if the family contains only one member within a given energy cut-off.

COORD_TxxxK_yyyy.pdb – concatenated conformations in PDB format of the members of family yyyy clustered at T=xxxK ranked by probabilities in descending order at this temperature (CLUST-WHAM).

COORD_T_xxxK_ave.pdb – cluster-averaged coordinates and coordinates of a member of each family that is closest to the cluster average in PDB format, concatenated in a single file (CLUST-WHAM).

INPUT_clust.tex – PicTeX code of the cluster tree.

INPUT.rms – rmsds between conformations.

6.2 Main input file

This file has the same structure as the UNRES input file; most of the data are input in a keyword-based form (see section 7.1 of UNRES description). The data are grouped into records, referred to as lines. Each record, except for the records that are input in non-keyword based form, can be continued by placing an ampersand (&) in column 80. Such a format is referred to as the data list format.

In the following description, the default values are given in parentheses.

6.2.1 Title

An 80-character string from the first line is input.

6.2.2 General data

(Data list format.)

NRES (0) – the number of residues.

ONE_LETTER – if present, the sequence is input in one-letter code.

SYM (1) – number of chains with same sequence (for oligomeric proteins only).

WITH_DIHED_CONSTR – if present, dihedral-angle restraints were imposed in the processed MREMD simulations

RESCALE (1) – Choice of the type of temperature dependence of the force field.

0 – no temperature dependence,

1 – homographic dependence (not implemented yet with any force field)

2 – hyperbolic tangent dependence [3].

DISTCHAINMAX (50.0) – for oligomeric proteins, distance between the chains above which restraints will be switched on to keep the chains at a reasonable distance.

PDBOUT – clusters will be printed in PDB format.

ECUT – energy cut-off criterion to print conformations (UNRES-CLUST runs). Only those families will be output the energy of the lowest-energy conformation of which is within ECUT kcal/mol above that of the lowest-energy conformation and for a family only those members will be output which have energy within ECUT kcal/mol above the energy of the lowest-energy member of the family.

PRINT_CART – output leading members of the families in UNRES x format.

PRINT_INT – output leading members of the families in UNRES int format.

REF_STR – if present, reference structure is input and rmsd will be computed with respect to it (CLUST-UNRES only; rmsd is provided in the cx file from WHAM for CLUST-WHAM runs).

PDBREF – if present, reference structure will be read in from a pdb file.

SIDE – side chains will be considered in superposition when calculating rmsd.

CA_ONLY – only the Calpha atoms will be used in rmsd calculation.

NSTART (0) – first residue to superpose.

NEND (0) – last residue to superpose.

NTEMP (1) – number of temperatures at which probabilities will be calculated and clustering performed (CLUST-WHAM).

TEMPER (NTEMP tiles) – temperatures at which clustering will be performed (CLUST-WHAM).

EFREE – if present, conformation entropy factor is read if the conformation is input from an x or pdb file.

PROB (0.99) – cut-off on the summary probability of the conformations that are clustered at a given temperature (CLUST-WHAM).

IOPT (2) - clustering algorithm:

1 – Ward's minimum variance method.

2 – single link method.

3 – complete link method.

4 – average link (or group average) method.

5 – McQuitty's method.

6 – Median (Gower's) method.

7 – centroid method.

Instead of IOPT=1, MINTREE and instead of IOPT=2 MINVAR can be specified

NCUT (1) – number of cut-offs in clustering.

CUTOFF (-1.0; NCUT values) cut-offs at which clustering will be performed; at the cut-off flagged by a "-" sign clustering will be performed with cutoff value=abs(cutoff(i)) and conformations corresponding to clusters will be output in the desired format.

MAKE_TREE – if present, produce a clustering-tree graph.

PLOT_TREE – if present, the tree is written in PicTeX format to a file.

PRINT_DIST – if present, distance (rmsd) matrix is printed to main output file.

PUNCH_DIST – if present, the upper-triangle of the distance matrix will be printed to a file.

6.2.3 Energy-term weights and parameter files

WSC (1.0) – side-chain-side-chain interaction energy.

WSCP (1.0) – side chain-peptide group interaction energya.

WELEC (1.0) – peptide-group-peptide group interaction energy.

WEL_LOC (1.0) – third-order backbone-local correlation energy.

WCORR (1.0) – fourth-order backbone-local correlation energy.

WCORR5 (1.0) – fifth-order backbone-local correlation energy.

WCORR6 (1.0) – sixth-order backbone-local correlation energy.

WTURN3 (1.0) – third-order backbone-local correlation energy of pairs of peptide groups separated by a single peptide group.

WTURN4 (1.0) – fourth-order backbone-local correlation energy of pairs of peptide groups separated by two peptide groups.

WTURN6 (1.0) – sixth-order backbone-local correlation energy for pairs of peptide groups separated by four peptide groups.

WBOND (1.0) – virtual-bond-stretching energy.

WANG (1.0) – virtual-bond-angle-bending energy.

WTOR (1.0) – virtual-bond-torsional energy.

WTORD (1.0) – virtual-bond-double-torsional energy.

WSCCOR (1.0) – sequence-specific virtual-bond-torsional energy.

WDIHC (0.0) – dihedral-angle-restraint energy.

WHPB (1.0) – distance-restraint energy.

SCAL14 (0.4) – scaling factor of 1,4-interactions

6.2.4 Molecule information

6.2.4.1 Sequence information

Amino-acid sequence

3-letter code: Sequence is input in format 20(1X,A3)

1-letter code: Sequence is input in format 80A1

6.2.4.2 Dihedral angle restraint information

This is the information about dihedral-angle restraints, if any are present. It is specified only when WITH_DIHED_CONSTR is present in the first record.

1st line: ndih_constr – number of restraints (free format)

2nd line: ftors – force constant (free format)

Each of the following ndih_constr lines:

idih_constr(i),phi0(i),drange(i) (free format)

idih_constr(i) – the number of the dihedral angle gamma corresponding to the ith restraint

phi0(i) – center of dihedral-angle restraint

drange(i) – range of flat well (no restraints for phi0(i) +/- drange(i))

6.2.4.3 Disulfide-bridge data

1st line: NS, (ISS(I),I=1,NS) (free format)

NS – number of cystine residues forming disulfide bridges.

ISS(I) – the number of the Ith disulfide-bonding cystine in the sequence.

2nd line: NSS, (IHPB(I),JHPB(I),I=1,NSS) (free format)

NSS – number of disulfide bridges

IHPB(I),JHPB(I) – the first and the second residue of ith disulfide link.
Because the input is in free format, each line can be split

6.2.5 Reference structure

If PDBREF is specified, filename with reference (experimental) structure, otherwise UNRES internal coordinates as the theta, gamma, alpha, and beta angles.

6.3 Main output file

The main (with name INPUT_clust.out or INPUT_clust.out_000 for parallel runs) output file contains the results of clustering (numbers of families at different cut-off values, probabilities of clusters, composition of families, and rmsd values corresponding to families (0 if rmsd was not computed or read from WHAM-generated cx file).

The output files corresponding to non-master processors (INPUT_clust.out_xxx where xxx>0 contain only the information up to the clustering protocol. These files can be deleted right after the run.

Excerpts from the a sample output file are given below:

CLUST-UNRES:

```
THERE ARE   20 FAMILIES OF CONFORMATIONS

FAMILY    1 CONTAINS    2 CONFORMATION(S):
  42 -2.9384E+03  50 -2.9134E+03


Max. distance in the family:    14.0; average distance in the family:    14.0

FAMILY    2 CONTAINS    3 CONFORMATION(S):
  13 -2.9342E+03   7 -2.8827E+03  10 -2.8682E+03
```

CLUST-WHAM:

```
AT CUTOFF: 200.00000
Maximum distance found:  137.82
Free energies and probabilities of clusters at 325.0 K
clust   efree    prob sumprob
    1   -76.5 0.25035 0.25035
    2   -76.5 0.24449 0.49484
    3   -76.4 0.21645 0.71129
    4   -76.4 0.20045 0.91174
    5   -75.8 0.08826 1.00000


THERE ARE    5 FAMILIES OF CONFORMATIONS

FAMILY    1 WITH TOTAL FREE ENERGY   -7.65228E+01 CONTAINS  548 CONFORMATION(S):
8363  -7.332E+013939  -7.332E+012583  -7.332E+017395  -7.332E+019932  -7.332E+01
5816  -7.332E+013096  -7.332E+012663  -7.332E+014099  -7.332E+016822  -7.332E+01
3176  -7.332E+017542  -7.332E+018933  -7.332E+017315  -7.332E+01 200  -7.332E+01.
```

```
.
5637  -7.062E+018060  -7.061E+013797  -7.060E+018800  -7.057E+016295  -7.057E+01
6298  -7.057E+012332  -7.057E+012709  -7.057E+01

Max. distance in the family:    16.5; average distance in the family:     8.8
Average RMSD 8.22 A
```

6.4 Output coordinate files

6.4.1 The internal coordinate (int) files

The file with name COORD_clust.int contains the angles theta, gamma, alpha, and beta of all residues of the leaders (lowest UNRES energy conformations from consecutive families for CLUST-UNRES runs and lowest free energy conformations for CLUST-WHAM runs). The format is the same as that of the file output by UNRES; see section 9.1.1 of UNRES description.

For CLUST-WHAM runs, the first line contains more items:

number of family	(format i5)
UNRES free energy of the conformation	(format f12.3)
Free energy of the entire family	(format f12.3)
number of disulfide bonds	(format i2)
list disulfide-bonded pairs	(format 2i3)
conformation class number (0 if not provided)	(format i10)

6.4.2 The Cartesian coordinate (x) files

The file with name COORD_clust.x contains the Cartesian coordinates of the alpha-carbon and side-chain-center coordinates. The coordinate format is as in section 9.1.2 of UNRES description and the first line contains the following items:

Number of the family	(format I5)
UNRES free energy of the conformation	(format f12.3)
Free energy of the entire family	(format f12.3)
number of disulfide bonds	(format i2)
list disulfide-bonded pairs	(format 2i3)
conformation class number (0 if not provided)	(format i10)

6.4.3 The PDB files

The PDB files are in standard format (see ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions). The ATOM records contain Calpha coordinates (CA) or UNRES side-chain-center coordinates (CB). For oligomeric proteins chain identifiers are present (A, B, ..., etc.) and each chain ends with a TER

record. Coordinates of a single conformation or multiple conformations The header (REMARK) records and the contents depends on cluster run type. The next subsections are devoted to different run types.

6.4.3.1 CLUST-UNRES runs

The files contain the members of the families obtained from clustering such that the lowest-energy conformation of a family is within ECUT kcal/mol higher in energy than the lowest-energy conformation. Again, within a family, only those conformations are output whose energy is within ECUT kcal/mol above that of the lowest-energy member of the family. Families and the members of a family within a family are ranked by increasing energy. The file names are:

COORD_xxxx.pdb where xxxx is the number of the family, if the family contains only one member of if only one member is output.

COORD_xxxx_yyy.pdb where xxxx is the number of the family and yyy is the number of the member of this family.

An example is the following:

```
REMARK R0001                              ENERGY    -2.93843E+03
ATOM      1  CA  GLY     1       0.000   0.000   0.000
ATOM      2  CA  HIS     2       3.800   0.000   0.000
ATOM      3  CB  HIS     2       5.113   1.656   0.015
ATOM      4  CA  VAL     3       5.927  -3.149   0.000
.
.
.
ATOM    346  CB  GLU   183     -43.669 -32.853  -7.320
TER
CONECT    1    2
CONECT    2    4    3
.
.
.
CONECT  341  343  342
CONECT  343  344
CONECT  345  346
```

where ENERGY is the UNRES energy. The CONECT records defined the Calpha-Calpha and Calpha-SC connection.

6.4.3.2 CLUST-WHAM runs

The program generates a file for each family with its members and a summary file with ensemble-averaged conformations for all families. These are described in the two next sections.

6.4.3.2.1 Conformation family files

For each family, the file name is COORD_TxxxK_yyyy.pdb, where yyyy is the number of the family and xxx is the integer part of the temperature (K). The first REMARK line in the file contains the information about the free energy and average rmsd of the entire cluster and, for each conformation, the initial REMARK line contains these quantities for this conformation. Same applies to oligomeric proteins, for which the TER records separate the chains and the ENDMDL record separates conformations. An example is given below.

```
REMARK CLUSTER    1 FREE ENERGY  -7.65228E+01 AVE RMSD 8.22
REMARK 1BDD L18G full clust ENERGY    -7.33241E+01 RMS  10.40
ATOM      1  CA  VAL     1      18.059 -33.585   4.616  1.00  5.00
ATOM      2  CB  VAL     1      18.720 -32.797   3.592  1.00  5.00
.
.
.
ATOM    115  CA  LYS    58      29.641 -44.596  -8.159  1.00  5.00
ATOM    116  CB  LYS    58      27.593 -45.927  -8.930  1.00  5.00
TER
CONECT    1    3    2
CONECT    3    5    4
.
.
CONECT  113  114
CONECT  115  116
TER
REMARK 1BDD L18G full clust ENERGY    -7.33240E+01 RMS  10.04
ATOM      1  CA  VAL     1       3.174   2.833 -34.386  1.00  5.00
ATOM      2  CB  VAL     1       3.887   2.811 -33.168  1.00  5.00
.
.
ATOM    115  CA  LYS    58      16.682   6.695 -20.438  1.00  5.00
ATOM    116  CB  LYS    58      18.925   5.540 -20.776  1.00  5.00
TER
CONECT    1    3    2
CONECT    3    5    4
CONECT  113  114
CONECT  115  116
TER
```

6.4.3.2.2 Average-structure file

The file name is COORD_T_xxxK_ave.pdb. The entries are in pairs; the first one is cluster-averaged conformation and the second is a family member which has the lowest rmsd from this average conformation. Computing average conformations is explained in section 2.5 of ref 3. Example excerpts from an entry corresponding to a given family are shown below.

```
REMAR AVERAGE CONFORMATIONS AT TEMPERATURE  300.00
REMARK CLUSTER    1
REMARK 2HEP clustering 300K ENERGY    -8.22572E+01 RMS   3.29
ATOM      1  CA  MET     1     -17.748  48.148 -19.284  1.00  5.96
ATOM      2  CB  MET     1     -17.373  47.911 -19.294  1.00  6.34
ATOM      3  CA  ILE     2     -18.770  49.138 -18.133  1.00  3.98
.
.
.
ATOM     80  CB  PHE    41     -14.353  44.680 -15.642  1.00  2.62
ATOM     81  CA  ARG    42     -11.619  41.645 -13.117  1.00  4.06
ATOM     82  CB  ARG    42     -11.330  40.378 -13.313  1.00  5.19
TER
CONECT    1    3    2
CONECT    3    5    4
.
.
.
CONECT   76   78   77
CONECT   78   79
CONECT   79   80
CONECT   81   82
TER
REMARK 2HEP clustering 300K ENERGY    -8.22572E+01 RMS   3.29
ATOM      1  CA  MET     1     -37.698  40.489 -32.408  1.00  5.96
ATOM      2  CB  MET     1     -38.477  39.426 -34.159  1.00  6.34
.
.
.
ATOM     80  CB  PHE    41     -35.345  50.342 -31.371  1.00  2.62
ATOM     81  CA  ARG    42     -33.603  54.332 -27.130  1.00  4.06
ATOM     82  CB  ARG    42     -33.832  53.074 -24.415  1.00  5.19
TER
CONECT    1    3    2
CONECT    3    5    4
.
.
.
CONECT   76   78   77
CONECT   78   79
CONECT   79   80
CONECT   81   82
```

```
TER
```

6.5 The conformation-distance file

The file name is INPUT_clust.rms. It contains the upper-diagonal part of the matrix of rmsds between conformations and differences between their energies:

i,j,rmsd,energy(j)-energy(i) (format 2i5,2f10.5)

where i and j, j>i are the numbers of the conformations, rmsd is the rmsd between conformation i and conformation j and energy(i) and energy(j) are the UNRES energies of conformations i and j, respectively.

6.6 The clustering-tree PicTeX file

This file contains the PicTeX code of the clustering tree. The file name is INPUT_clust.tex. It should be supplemented with LaTeX preamble and final commands or incorporated into a LaTeX source and compiled with LaTeX. The picture is produced by running LaTeX followed by dvips, dvipdf or other command to convert LaTeX-generated dvi files into a human-readable files.

7 SUPPORT

Dr. Adam Liwo
Faculty of Chemistry, University of Gdansk
ul. Sobieskiego 18, 80-952 Gdansk Poland.
phone: +48 58 523 5430
fax: +48 58 523 5472
e-mail: adam@chem.univ.gda.pl

Dr. Cezary Czaplewski
Faculty of Chemistry, University of Gdansk
ul. Sobieskiego 18, 80-952 Gdansk Poland.
phone: +48 58 523 5430
fax: +48 58 523 5472
e-mail: czarek@chem.univ.gda.pl

Prepared by Adam Liwo, 02/19/12

LaTeXversioin, 09/28/12