# *myPresto* 5.0

*- clustering -*

USER MANUAL

2018/01/12

## 本ドキュメントについて

本ドキュメントは、「***myPresto* 5.0** USER MANUAL」の別冊です。コピーライト、プログラム使用許諾条件、著者および引用文献については、「***myPresto* 5.0** USER MANUAL」の記述に準じます。

## 謝辞

本ソフトウェアの研究開発は、国立研究開発法人日本医療研究開発機構(AMED)の援助によって行われました。ここに感謝の意を記します。

本ソフトウェアは、故・京極好正博士の始められた研究の中で開発されました。

# 1. 概要

クラスタリングツール clustering は、トラジェクトリに含まれる構造をクラスタリングし、代表的な構造を PDB ファイルとして出力します。clustering では、非常に多くの構造を扱うことができないため、トラジェクトリ抽出ツール selection で構造数を減らしてから、clustering を実行するとよいでしょう。clustering では 1000 個以内の構造を取り扱うことを想定しています。selection は、入力したポテンシャルエネルギーの確率分布に従って、座標トラジェクトリの構造を抽出し、再構成するツールです。

# 2. インストール

(1) インストールに必要な環境

・UNIX (Linux) 環境 ：selection, clustering の実行環境です。

・Fortran90 コンパイラ ： selection, clustering の構築に使用します。

( GNU FORTRAN コンパイラ(gfortran)、または、Intel FORTRAN コンパイラ(ifort) )

(2) インストール方法

CLUSTERINGyymmdd.tar.gz を書き込み可能権限のあるディレクトリ（例えば、ホームディレクトリ）に配置して、次のコマンドを実行します。(yymmdd は年月日の数字です。)

```
% tar -xzvf CLUSTERINGyymmdd.tar.gz
```

次のコマンドを実行することにより、プログラムをインストールします。

```
% cd CLUSTERINGyymmdd

次のコマンドは、どちらか一方を実行します。
% bin/install.sh          (GNU のコンパイラを使用する場合)
% bin/install.sh intel    (Intel のコンパイラを使用する場合)
```

インストール後のディレクトリ構成:

```
CLUSTERING
 |--OREADME
 |--bin/
 |    |--install.sh
 |    |--test_selection.sh
 |    |--test_clustering.sh
 |    |--clustering (install.sh 実行後に出現)
 |    `--selection  (install.sh 実行後に出現)
 |--doc/
 |--sample/
 `--src/
```

(3) テストプログラムの実行

次のコマンドで、トラジェクトリ抽出ツールのテストプログラムを実行します。

```
% bin/test_selection.sh
```

このテストプログラムの出力先は、CLUSTERINGyymmdd/test_selection/です。このテストプログラムを実行することにより、selection が正常に動作することを確認することができます。

次のコマンドで、クラスタリングツールのテストプログラムを実行します。

```
% bin/test_clustering.sh
```

このテストプログラムの出力先は、CLUSTERINGyymmdd/test_clustering/です。このテストプログラムを実行することにより、clustering が正常に動作することを確認することができます。

# 3. selectionの実行方法

入力データ

トラジェクトリ抽出ツールselectionの入力を以下に示します。

(1)座標トラジェクトリ

cosgeneの出力トラジェクトリ

(2)エネルギー確率分布

reweightingツールの出力ファイル

(3)トラジェクトリ抽出ツールの制御ファイル

(3-1)エネルギー確率分布ファイル名

(3-2)座標トラジェクトリファイル名

(3-3)トラジェクトリファイルの型(Single | Double)

(3-4)サンプリング区間先頭

(3-5)サンプリング区間最後

(3-6)確率分布にかける係数(抽出する座標数はこの係数に比例する)

(3-7)出力トラジェクトリファイル名

(3-8)原子数

制御ファイル例)

bestfit対象原子指定ファイル例(水素以外の蛋白質原子のみbestfitする)

```
pdf.total
ala8.cor_ST
S
0
10000000
100.0
select.cor
```

標準出力例）

```
***** COORDINATE TRAJCETORY SELECT TOOL FOR COSGENE (2005/08/31) *****
   FUNCTION : SELECT TRAJECTORY AND OUTPUT TRAJECTORY FILE

   INPUT :
     (1) ENERGY PROBABILITY DENCTY FUNCTION FILE NAME
     (2) COSGENE TRAJCECTORY FILE NAME
     (3) TRAJECTORY FORMAT
     (4) START LOOP NUMBER
     (5) END LOOP NUMBER
     (6) SELECTION RATE
     (7) OUTPUT TRAJECTORY FILE NAME
   OUTPUT :
     (1)SELECTED TRAJECTORY
 ****************************************************************

 INPUT ENERGY PROBABILITY DENCITY FUNCTION FILE NAME
 INPUT TRAJECTORY FILE NAME
 INPUT COORDINATE TRAJECTORY FORMAT ("S"ingle | "D"ouble)
 INPUT START LOOP NUMBER
 INPUT END LOOP NUMBER
 SELECTION RATE (0.0 < RATE
 OUTPUT NEW TRAJECTORY FILE NAME
 ***** SELECT TRAJECTORY RESULT                          *****
  1) DISTRIBUTION
   POTENTIAL-ENERGY  PROBABILITY(%)  TRAJECTORIES   SAMPLES  SAMPLE-RATE(%)
     -0.55000E+01            0.933              0         0   -----
     -0.45000E+01            1.000              1         1  100.000
     -0.35000E+01            0.987              0         0   -----
     -0.25000E+01            0.970              0         0   -----
     -0.15000E+01            1.007              0         0   -----
     -0.50000E+00            1.023              4         4  100.000
      0.50000E+00            1.039              2         2  100.000
      0.15000E+01            1.031              4         4  100.000
      0.25000E+01            1.022             10        10  100.000
      0.35000E+01            0.952             13        12   92.308
      0.45000E+01            0.905             10        10  100.000
      0.55000E+01            0.837             12         9   75.000
      0.65000E+01            0.796             15        11   73.333
      0.75000E+01            0.776             13        11   84.615
    P.D.F. SUM=   99.9852000000000
    STRUCT SUM=        2000
    SAMPLE SUM=        1056

  2) INPUT FILES
       TRAJECTORY FILE        :
         ala8.cor_ST

  3) SELECTION
      TOTAL TRAJECTORY NUMBER :        2000
      SAMPLING BOUND          :           0 -    10000000
      SAMPLING NUMBER         :        2000
      RATE                    :   100.000000000000
      OUTPUT NUMBER           :        1056
      TRAJECTORY FILE         :
         select.cor
 ****************************************************************
```

# 4. clustering の実行

入力データ

（１）制御ファイル

（１－１）トポロジーファイル名

（１－２）ベストフィットの適用（"Y" | "N"）

（１－２－１）ベストフィット対象原子指定ファイル名

（１－３）RMSD 計算対象原子の指定（"Y" | "N"）

（１－３－１）RMSD 計算対象原子指定ファイル名

（１－４）サンプル数

（１－５）クラスタ数

（１－６）サンプリング区間先頭

（１－７）サンプリング区間最後

（１－８）入力トラジェクトリファイル名

（１－９）トラジェクトリファイルの型（"S" | "D"）

（１－１０）クラスタリング方法（"nearest " | "furthest" | "median " | "centroid" | "average " | "flexible" | "ward " ）

（１－１０－１）flexible 指定時のβ値

（１－１１）出力 PDB 名先頭

（１－１２）デンドログラムファイル名

（２）cosgene の入力トポロジーファイル

（３）cosgene の出力トラジェクトリファイル

（４）ベストフィット対象原子指定ファイル（ベストフィット対象原子指定時）

ベストフィット対象原子を cosgene の「系の重心合わせ指定用ファイル」と同じ書式で指定します。(p.161「A.2.11 系の重心合わせ指定用ファイル」を参照)

（５）RMSD 計算対象原子指定ファイル（RMSD 計算対象原子指定時）

RMSD 計算対象原子を cosgene の「系の重心合わせ指定用ファイル」と同じ書式で指定します。(p.161「A.2.11 系の重心合わせ指定用ファイル」を参照)

【注意】

メモリ使用量の観点から、サンプリングする構造は 1000 個以内が望ましいです。

【注意】

サンプリングする構造を 1000 個以上に指定することは可能です。メモリ確保に失敗したときは以下のエラーメッセージを出力し、プログラムの実行を停止します。

"CANNOT ALLOCATE MEMORY, DECREASE SAMPLING NUMBER"

■制御ファイル例

```
ala8.tpl       ; トポロジーファイル名
n              ; ベストフィットの適用
n              ; RMSD 計算対象原子の指定
10             ; サンプル数
10             ; クラスタ数
10             ; サンプリング区間先頭
40             ; サンプリング区間最後
ala8.cor_ST    ; 入力トラジェクトリファイル名
S              ; トラジェクトリファイルの型
nearest        ; クラスタリング方法
ala8.cls       ; 出力 PDB 名先頭
ala8.tree      ; デンドログラムファイル名
```

■ベストフィット対象原子指定ファイル例

（水素以外の蛋白質原子のみベストフィットする例）

```
SETBST> LIST
  FIX 1    1  1 32 H*  YES  ; 蛋白(チェイン 1)の 1～32 残基の"H*"は bestfit 対象外
  FIX 2    2  1  1 *   YES  ; リガンド(チェイン 2)の全原子は bestfit 対象外
  FIX 3 1000  1  1 *   YES  ; 水分子(チェイン 3～1000)の全原子は bestfit 対象外
```

■RMSD 計算対象指定ファイル

（水素以外のリガンド原子のみ RMSD 計算する例）

```
SETBST> LIST
  FIX 1    1  1 32  *  YES ; 蛋白(チェイン 1)の 1～32 残基の全原子は bestfit 対象外
  FIX 2    2  1  1 H*  YES ; リガンド(チェイン 2)の"H*"は bestfit 対象外
  FIX 3 1000  1  1  *  YES ; 水分子(チェイン 3～1000)の全原子は bestfit 対象外
```

出力データ

（1）ログ（標準出力に出力）

（1－1）ツールの使用方法

（1－2）データ入力問い合わせ

（1－3）クラスタリング条件

（1－4）入力トポロジーファイル情報

（1－5）ベストフィット対象原子一覧（ベストフィット対象原子指定ファイルに表示指定がある場合）

（1－6）RMSD 計算対象原子一覧（RMSD 計算対象指定ファイルに表示指定がある場合）

（1－7）クラスタリング進行状況

（1－8）出力 PDB ファイル名

（2）代表構造の PDB ファイル

クラスタ番号、構造数、エネルギーおよびループ回数をコメント出力し、原子情報を出力します。出力ファイル名は「"出力 PDB 名先頭"+"."+ループ回数」となります。

（3）デンドログラムファイル

ループ回数とポテンシャルを葉の名称としたデンドログラムを出力します。

■代表構造の PDB ファイル例

```
REMARK     CLUSTER   :           1
REMARK     STRUCTURE NUMBER:           140
REMARK     LOOP      :       10000
REMARK     POTENTIAL :   176.955627441406
ATOM      1  CA  ACE     1       2.508   1.314  -3.948 12
ATOM      2 HH31 ACE     1       2.771   1.634  -4.954  1.01  0.11
ATOM      3 HH32 ACE     1       2.166   0.280  -3.974  1.01  0.11
ATOM      4 HH33 ACE     1       1.718   1.947  -3.546  1.01  0.11
ATOM      5  C   ACE     1       3.771   1.408  -3.102 12.01  0.60
```

■デンドログラムファイル例

```
(
(
"10000 176.96 KCAL/MOL"
,
(
"13000 174.52 KCAL/MOL "
,
"16000 184.61 KCAL/MOL "
)
)
,
(
(
"19000 163.18 KCAL/MOL "
,

(
"22000 162.05 KCAL/MOL "
,
"28000 147.56 KCAL/MOL "
)
)
,
"25000 146.70 KCAL/MOL "
)
)
;
```

# 5. 実行例

### （1）再構成したカノニカル分布を用いた構造の抽出

ここでは、cosgene_sample_pack に含まれる実行例 Sample-4(F.B.McMD), Sample-5(S.T.McMD)およびSample-6(G.S.T.McMD)においてreweightにより任意の温度でのエネルギー分布を計算したものを使います。ここではそのエネルギー分布を満たすように、トラジェクトリファイルから座標を抽出します。

構造の抽出には解析ツール selection を用います。次の入力ファイルを用意します(select.inp)。

```
pdf363_st
ala_st.cor
S
1000
30000000
0.5
ala_st_363.cor
```

第１行目に、Sample-4, Sample-5 あるいは Sample-6 の最後で作成したエネルギー確率分布ファイルを指定します。

第２行目は MD 実行時に出力されたトラジェクトリファイルを指定します。

第３行目はトラジェクトリファイルの型(Single｜Double)です。MD 実行時の指定と一致させます。

第４、５行目でサンプリング区間を指定します。

第６行目で抽出する構造の割合（%）を指定します。サンプリング区間の構造のうちここで指定した割合の構造が出力されます。

第７行目で出力トラジェクトリファイル名を指定します。

第８行目で原子数を指定します。

```
% selection  < select.inp
```

で実行します。

### （2）構造のクラスタリング

次に、（1）でカノニカル分布を構成するように抽出した多数の構造からクラスタ解析により代表構造を取り出す方法を説明します。構造間の RMSD を用いてクラスタリングを行います。

クラスタ解析には解析ツール clustering を使用します。

まず、以下のような制御ファイル(clustering.inp)を用意します。

```
ala_ala.tpl
y
ala_ala.fit
y
ala_ala.rmsd
400
10
1
500
ala_st_363.cor
S
average
ala_st_363.cls
ala_st_363.tree
```

第１行目にトポロジーファイルを指定します。

第２行目には RMSD 計算時の bestfit 適用の有無を指定します(y ｜ n)。

bestfit の実行を指定したときは次行に bestfit に使用する原子を指定するファイルの名前を記述します。

ここでは ala_ala.fit として下記の内容のファイルを用意します。この例ではチェイン１の残基１－４の水素を bestfit に使用しないよう設定しています。この記述形式は cosgene の系の重心合わせ指定用ファイルと同じです。

```
SETBST> LIST
FIX 1 1 1 4 H* YES;
```

第４行目には RMSD の計算に使用する原子の指定の有無を記述します(y | n)。指定する場合には次行に RMSD の計算に用いる原子の指定ファイル名を記述します。今回は ala_ala.fit と同内容のものを ala_ala.rmsd として用います。bestfit、RMSD に使用された原子は実行時のログで確認できます。

第６行目にはクラスタリングに使用する構造数を指定します。系にもよりますが 1000 個以内程度で指定します。第７行目には最終的なクラスタの数を指定します。

第８、９行目で使用する座標トラジェクトリの範囲の開始・終了位置を指定します。このトラジェクトリの範囲は第６行目に指定した構造数を確保できるように指定します。また、先ほど select ツールにより取り出した構造数の範囲内で指定します。

第１０行目に座標トラジェクトリのファイル名を指定します。今回は select ツールによりカノニカル分布を再現したトラジェクトリファイルを使用します。

第１１行目にはトラジェクトリファイルの形式を指定します。(S | D)

第１２行目にはクラスタリングの方法を記述します（”nearest”| “furthest”| “median” | “centroid”| “average”| “flexible”| “ward”)。

ここで flexible を指定した場合は次行にβ値を設定します。

第１３行目には出力 PDB ファイルの先頭名を指定します。

第１４行目には出力デンドログラムのファイル名を指定します。

```
% clustering  < clustering.inp
```

で実行します。ログ出力のうちクラスタごとに出力された構造数、平均構造との RMSD がクラスタの特徴をよく表します。

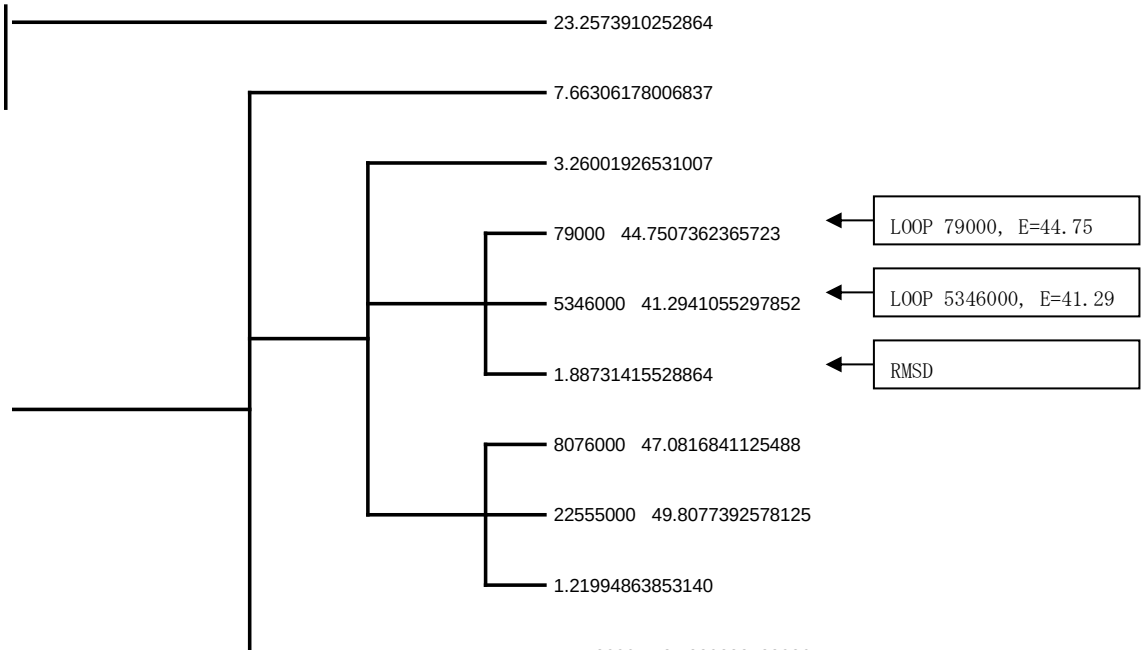
```
        （前略）

CLUSTER ID      :           1
STRUCTURE COUNT :          32
LOOP NUMBER     :    16969000
RMSD OF AVERAGE :   1.44551613437161
OUTPUT PDB FILE :ala_st_363.cls.16969000

CLUSTER ID      :           2
STRUCTURE COUNT :          32
LOOP NUMBER     :     8055000
RMSD OF AVERAGE :   1.87505849035797
OUTPUT PDB FILE :ala_st_363.cls.8055000

          （後略）
```

指定したクラスタ個数の構造（PDB 形式）のほか、デンドログラムファイル（PHYLIP 形式）が出力され汎用的なツールを用いて表示することができます。デンドログラムは二つの構造（またはクラスター）とその間の RMSD からなる３分木で表示されます。



***myPresto* 5.0**