MATRAS 1.2 : A Tool for Protein 3D Structure Comparison

MATRAS stands for Markovian TRAnsition of protein Structure

April 15, 2007

Developer Takeshi KAWABATA, Ken NISHIKAWA

Contact Takeshi KAWABATA
Department of Information Science
Nara Institute of Science and Technology
TEL : +81-743-72-5396
FAX : +81-743-72-5391
Email : takawaba@is.naist.jp

Reference

Kawabata T., Nishikawa K."Protein tertiary structure comparison using the Markov transition model of evolution" (2000). *Proteins*, **41**, 108-122.

1 What is MATRAS ?

MATRAS is the program sets for protein 3D structure comparison, it stands for *Markovian TRAnsition of protein Structure*. Its advantage is its score using Markov transition model of structure evolution, which is supposed to be better for detecting homologous structure similarity. At the beginning, MATRAS was developing in National Institute of Genetics (Mishima, Japan), by Takeshi Kawabata, who was a post-doctoral research fellow under the supervision of Prof. Ken Nishikawa. That was around 1998. The first purpose was basically for estimating accuracies for protein structure predictions (fold recognition) of CASP3. Its method was published in 2000, in *Proteins*. After that, the several functions, such as 3D library search and multiple 3D alignment, are added to the original MATRAS program. The WEB server of 3D structure comparison is also available (http://biunit.naist.jp/matras).

2 Method for comparing 3D structures

2.1 Definition of Similarity Scores

The structure similarity score of MATRAS is defined as a following log-odds score :

$$S(i,j) = \log \frac{P(i \to j)}{P(j)} \tag{1}$$

where $P(i \to j)$ is the transition probability that structure i changes to structure j during the evolutionary process, and $P(i)$ is the probability that structure i appears by chance. i and j can represent any kinds of 3D structural features, such as secondary structures and distance between residues. Generally speaking, estimation of the transition probability $P(j \to i)$. We estimated the transition probability by the Markov transition model, which is similar to Dayhoff's substitution model between amino acids. Matras uses the following three kinds of similarity scores.

(1) **SSE Score** S_{sse}

A secondary structure element (SSE) is a continuous residue group that is defined as an α-helix a β-strand. It is represented by a single vector defined by the principle inertial axis with the smallest moment. The spatial arrangement of a pair of SSEs is described by six parameters : the number of residues L_1, L_2, the closest distance between SSE pairs d, the bond angles θ_1,θ_2, and the dihedral angle ϕ. And we made six kinds of log-odds score corresponding six parameters, the total SSE score is represented as the sum of six terms.

(2) **Environment Score** S_{env}

This score was defined for the environment states, which are a combination of local structure and solvent accessibility. The ten kinds of "environment" are defined by combining the five local structures and the two accessibility classes.

(3) **Distance Score** S_{dis}

This score focuses on the distance between C^{β} atoms of the ith and jth residues. The distance is transformed into a discrete histogram with 1 $\AA$ width. This score is prepared for each residues separation k (=$|i - j|$). It is used in the final stage of alignment of our program, because it is the most sensitive to detect structural similarity among our three scores.

2.2 Alignment Strategy

It it difficult to find the structural corresponding residues (alignment) with the largest two-body similarity score (such as the SSE score and the distance score). We use the most popular heuristics, "hierarchical alignment", in which a rough alignment is first obtained by the SSEs, then the alignment is improved with more detailed similarity functions. Our procedures of hierarchical alignment consists of the following three stages.

(1) **Make a SSE alignment using** S_{sse}

Build-up method is used for finding the corresponding SSEs.

(2) **Preliminary DP alignment using** S_{env}

A dynamic programming alignment with S_{env} is performed, using restriction of previously aligned SSEs.

(3) **Iterative DP Alignment using** S_{dis}

A dynamic programming alignment with the distance score S_{dis} is iteratively performed using the alignment determined in the previous stage.

3 Installations

3.1 Required Environment

- **OS:** UNIX

  The original Matras was developed on SGI-IRIX workstation, and now maintained in Linux machines. We believe that it may work on any other UNIX systems, although we confirm only two environment (IRIX and Linux).

- **Programming Language:** C and Perl

  The main program of Matras is written in C, and other additional programs are written in Perl, such as making BSSP file, superimposed structures and multiple 3D alignment. To install and use Matras, you must prepare a C compiler and a Perl interpreter.

- **DSSP**

  Matras needs a program DSSP (Kabsh and Sander,1983) to assign secondary structures of proteins, for making BSSP files. You must download the source code from the site (http://swift.cmbi.ru.nl/gv/dssp/index.html), and install the DSSP program (dsspcmbi).

- **RasMol**

  A molecular graphic program is necessary to see superimposed structures obtained by Matras. We recommend the program RasMol for our purpose. The RasMol is the most popular freeware, it can work on most of UNIX platform. If you don't have RasMol in your computer, go to http://www.OpenRasMol.org and install it.

3.2 Procedures for Install

(1) Download the compressed source file "Matras[version].tar.gz"

(2) ungzip the file and extract all the files.

```
% ungzip Matras[version].tar.gz
% tar xvf Matras[version].tar
```

New directory "Matras[version]" appears.

(3) Go to src directory

```
% cd Matras[version]/src
```

(4) Edit "Makefile" for adjusting your environment.

The default makefile assumes that a user uses the gcc compiler. We also prepare a makefile for SGI ("Makefile.sgi").

(5) Make it

```
% make
```

If you succeed, an executable file “Matras” is made at the upper directory of “src”.

(6) Put the executable file “Matras” on your favorite binary directory, or add the Matras directory to your PATH variable.

3.3 Set your environment using “.matras” file

Matras reads environmental information from the file “.matras”. You must put the “.matras” files on (1)your current directory, or (2)your home directory. A sample environmental file is shown as follows, which is stored as ”dot.matras” in the base directory.

```
################################
### MATRAS ENVIRONMENT FILE ###
################################

BASE_DIR  /home/takawaba/work/Matras12
SCORE_DIR /home/takawaba/work/Matras12/data_sc/ROM-04JAN29
TMP_DIR   /home/takawaba/work/Matras12/tmpout
BSSP_DIR  /DB/BSSP
PDB_DIR   /DB/PDB
```

A line whose head is “#” is a comment that Matras skips to read. Other lines are combinations of [Variable Name] [Value of Variables]. We will explain important variables.

(1) **BASE_DIR** : a directory where Matras is installed.

(2) **SCORE_DIR** : a directory of Matras score files

(3) **BSSP_DIR** : a directory of BSSP files.
I will explain about BSSP files later.

(4) **TMP_DIR** : a directory for temporary files.
If you want to use multiple 3D alignments, you must assign TMP_DIR.

3.4 Make BSSP files for your structures

3.4.1 What is BSSP?

Matras can read only special structure file : a “BSSP” file. Unfortunately, Matras cannot read PDB file directly. A BSSP file is an extension of DSSP file, which includes XYZ coordinates of C_β atoms, not only C_α atoms. We choose the format of BSSP files because of following three reasons. (1) Secondary structure information is required to compare 3D structure. (2) XYZ coordinates of C_β atoms is necessary to align protein pairs precisely, especially for β-strand regions. (3) A file size of BSSP is 1/10 of PDB file size. It enables us to search 3D structure more quickly.

A format of BSSP file is shown in Appendix.

3.4.2 How to make BSSP files

In order to make a BSSP file, you need a program of DSSP(named “dsspcmbi”) and a Perl script “bssp.pl” (located in BASE_DIR). You must perform following two commands:

(1) `% dsspcmbi -c [pdb_file] [dssp_file]`

(2) `% bssp.pl [dssp_file] [pdb_file] > [bssp_file]`

For example, when you want to make a bssp file for myoglobin(PDBcode:1mbd), type as follows:

```
% dsspcmbi -c pdb1mbd.ent 1mbd-.dssp
% bssp.pl 1mbd-.dssp pdb1mbd.ent > 1mbd-.bssp
```

I recommend to add ”.bssp” to the end of the bssp file as a suffix. In principle, Matras assumes that one BSSP file only contains one chain of protein. For the PDB files with multi chains, you must make a new PDB file that contains only one chain you want to compare.

The BSSP files must be located in one of the following three locations.

(1) the current directory when you execute Matras.

(2) BSSP_DIR, which is defined in the “.matras” file. If you want to deal with a large amount of structures, we recommend to put in the BSSP_DIR.

(3) BSSP_DIR/subdirectory. When you put a BSSP file in BSSP_DIR, its subdirectory is named as the second and third characters of the file name. This is a similar directory system to Protein Data Bank(PDB). For example, a file “1mbd-.bssp” is located in “BSSP_DIR/mb/”, “4azuA.bssp” is located in “BSSP_DIR/az/”.

3.5 Display HELP messages

If you input a following command,

```
% Matras
```

Matras shows simple help messages. A more detailed help messages are shown using a following command:

```
% Matras H
```

4 Pairwise 3D Alignment

Comparing two structures is the basic procedures of Matras. Other structural comparisons, such as libaray search and multiple alingment, are developed based on the pairwise 3D alingment.

4.1 Basic Operation

Basically, you can compare two structures by a following command:

```
% Matras P -A [bsspfileA] -B [bsspfileB] <options>
```

For example, if you want to compare myoglobin(1mbd-.bssp) and hemoglobin α chain (4hhbA.bssp), input a following command:

```
% Matras P -A 1mbd-.bssp -B 4hhbA.bssp
```

Then you get a following output in standard output.

```
#### MATRAS VER 1.2: PROGRAM FOR PROTEIN 3D STRUCTURE COMPARISON ####
# coded by Takeshi Kawabata. Last Modified : Feb 6, 2004
#<REFERENCE>
# Takeshi Kawabata and Ken Nishikawa.
# "Protein Structure Comparison Using the Markov Transition Model of Evolution".
#   Proteins vol.41:108-122(2000).
#<COMMAND> "Matras P -A 1mbd-.bssp -B 4hhbA.bssp "
#<DATE> "May 9,2004 10:30:20"
#<MODE> P:PAIRWISE COMPARISON
#<PARAMETERS>
# ProAFile "1mbd-.bssp" ProBFile "4hhbA.bssp"
# SseAliType T EnvAliType T AlgType L
# ssescfile "3U" envscfile "T10-3U.rom" DisSc N disscfile "3U" DisScE - EnvType T Nenvstate 10
# GapExtE -6.0 GapExtD -100.0 Nkeep 35 Nrep 10 SseOffsetD 0
an->mal_sim 1
[ALIGN_RANK]  1
[PROTEIN A]   1mbd-    Naa 153 Nsse  8 "MYOGLOBIN (DEOXY, $P*H 8.4)"
[PROTEIN B]   4hhbA    Naa 141 Nsse  7 "HEMOGLOBIN (DEOXY)"
[ALIGNMENT]   Ncomp_aa 141 Ncomp_sse 6
[SIMILARITY]  Seq 27.0 %  Sec 88.7 % Exp 82.3 % CRMS 1.56 A  DRMS 1.40 A
[SCORE]       ScSSE 676.0 ScEnv 4007.8 ScDis 149312.6 Rdis 70.7 (%) Rsse 44.8 (%)
[RELIABILITY] Superfamily 100.0 %  Fold 100.0 %
     :   H1               H2              H3             H4     H5
SecA :   HHHHHHHHHHHHHHHGGHHHHHHHHHHHHHHH HHHHTT TTTTT  SHHHHHH HH
    1:VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED:60
      ***      *   * ** *     *   * * * * * *   *  *  *        *
    1:VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFD-L----SHG-SAQ:54
SecB :   HHHHHHHHHHHHHHHTTTHHHHHHHHHHHHHH GGGGGG TTS - ----STT- HH
     :   H1               H2                         - ----   - H3

     :                      H6                H7
SecA :HHHHHHHHHHHHHHHHTTTT  HHHHHHHHHHHHHTS   HHHHHHHHHHHHHHHHHH G
   61:LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP:120
       * **  *  **             *  *   ** *           *      *    *
   55:VKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLP:114
SecB :HHHHHHHHHHHHHHHHHTGGGHHHHTHHHHHHHHHTT   THHHHHHHHHHHHHHHHH T
     :                     H4   H5             H6

     :    H8
SecA :GG  HHHHHHHHHHHHHHHHHHHHHHH
```

```
  121:GDFGADAQGAMNKALELFRKDIAAKYK:147
        *         * *       **
  115:AEFTPAVHASLDKFLASVSTVLTSKYR:141
SecB :TT  HHHHHHHHHHHHHHHHHHHTTT
     :    H7
```

4.2 Measures of Structure Similarities

The head lines of the outputs contain various kinds of structure similarities.

```
[ALIGN_RANK]  1
[PROTEIN A]   1mbd-    Naa 153 Nsse  8 "MYOGLOBIN (DEOXY, $P*H 8.4)"
[PROTEIN B]   4hhbA    Naa 141 Nsse  7 "HEMOGLOBIN (DEOXY)"
[ALIGNMENT]   Ncomp_aa 141 Ncomp_sse 6
[SIMILARITY]  Seq 27.0 %  Sec 88.7 % Exp 82.3 % CRMS 1.56 A  DRMS 1.40 A
[SCORE]       ScSSE 676.0 ScEnv 4007.8 ScDis 149312.6 Rdis 70.7 (%) Rsse 44.8 (%)
[RELIABILITY] Superfamily 100.0 %  Fold 100.0 %
```

- **[ALINGMENT]**

  Ncomp_aa : Number of compared (aligned) amino acids

  Ncomp_sse : Number of compared (aligned) SSEs

- **[SIMILARIRY]**

  Seq : sequence identity (%), defined as number of identical amino acid pairs divided by Ncomp_aa.

  Sec : secondary structure identitiy(%), defined as number of identical secondary structure residue pairs divided by Ncomp_aa. 8 states secondary structure of DSSP are used.

  CRMS : Root mean square deviation ($\mathring{A}$) of C^{α} atom positions of aligned residues, after optimal superimposition.

  DRMS : Root mean square deviation ($\mathring{A}$) of distances between C^{β} atom positions of aligned residues.

- **[SCORE]**

  ScSSE: SSE score S_{sse}.

  ScDIS: Distance score S_{dis}.

  Rdis :@Normalized S_{dis} score (%) using the maximum and minimum value of the score. This is for understanding the similarity more intuitively than the raw value of S_{dis}. R_{dis} between protein A and B is defined as follow :

$$R_{dis}(A, B) = 100 \cdot \frac{S_{dis}(A, B) - S_{min}}{S_{max} - S_{min}} \tag{2}$$

where $S_{dis}(A, B)$ is a raw distance score between protein A and B, S_{max} and S_{min} is the maximum and minimum value of the score correspondingly. We set $S_{min} = 0$, S_{max} is defined as the averaged value among the self scores:

$$S_{max} = \frac{S_{dis}(A, A) + S_{dis}(B, B)}{2} \tag{3}$$

- **[RELIABILITY]**: This values shows a probability that a structure pair with a normalized score R_{dis} is classified as the same superfamily/fold relationship. This values are estimated by all-vs-all comparison of protein domains in SCOP 1.71 database.

4.3 Output files

After calculation, Matras writes three files,"1.pdb", "1.ras" and "1.mat" in the current directory.
v

- **1.pdb**

  A PDB file that contains two superimposed structures. The optimal superposition is calculated for only aligned regions, but the file contains non-aligned regions. It contains only C^{α} and C^{β} atoms. A structure assigned by "-A" has a chain identifier 'A', and one assigned by "-B" has a 'B'. If you don't want to write this file, add the option **-op -**.

- **1.ras**

  A Rasmol script for coloring aligned residues in the superimposed PDB file "1.pdb". If you input the following command:

  ```
  % rasmol 1.pdb
  RasMol> script "1.ras"
  ```

  then RasMol shows a colored structure pairs. If you don't want to write this file, add the option **-or -**.

- **1.mat**

  This file contains the values of a translation vector and a rotation matrix for superimposing two structures.

4.4 Options for Input/Output

- **-oa** [CASP style pairwise alignment file]

  Write a pairwise alignment in the CASP style format (defined in Appendix).

- **-ia** [CASP style pairwise alignment file]

  Input a pairwise alignment in the CASP style format. If this option is assigned, Matras does not calculate an alignment by himself, only calculate similarities using the input alignment and write the results.

- **-ow** [ClustalW style pairwise alignment file]

  Write a pairwise alignment in ClustalW format.

4.5 Sub-optimal Alignment

A following option is for calculating sub-optimal alignment. Its default value is 'F'alse.

- **-SO** [T or F]

  If the option ''-SO T", then Matras calculate sub-optimal alignments. The default is "F"alse. If Matras recognize more than one optimal alignment, they show all the alignments in stdout, and write correspoinding RasMol scripts are "1.ras","2.ras".."n.ras", correspoinding matrix files are "1.mat","2.mat".."n.mat".

4.6 Sequence Alignment

A following option is for calculating simply sequence alignment. For some special cases (for example, identical protein pairs with large conformational change), sequence alignment is better than 3D alignment. Its default value is 'F'alse.

- **-SQ** [T or F]

  If the option ''-SO T", then Matras calcualate an alignment considering only amino acid type, not any 3D structure features. BLOSUM 62 score and gap penalty for extension is -1, and that for open is -11.

4.7 Self 3D Alignment

Self 3D alignment is to detect structurally similar regions in one proteins. It is useful for finding repeating units of proteins.

To run the self 3D alignment, you only run the pairwise 3D alignment assigning the same bsspfile for "-A" and "-B" options,and add "-SA T" option.

```
% Matras P -A [bsspfile] -B [bsspfile] -SA T
```

For example, if you want to find repeated structures of triose phosphate isomerase (TIM), run a following command.

```
% Matras P -A 1timA.bssp -B 1timA.bssp -SA T
```

You can use any other options for pairwise 3D alignments.

5 3D Library Search (one-vs-library)

This is for finding similar structures of a query structure in many library structures. This search requires a large computational costs. For example, a search of a query structure with 200 amino acids against 3000 library structures, takes 20-30 minutes using a Intel Pentium III 800 MHz CPU. The calculation time depends on the size of query protein and the number proteins in the library.

```
% Matras L -Q [query_bsspfile] -L [library_listfile] > [result_file]
```

The file [library_listfile] contains names of BSSP files for the structure library. The format of library list file is shown in Appendix.

5.1 Score Normalization for ranking similar structures

To rank the similar score, Matras employs N_{comp}^2 -fitted Z-score for S_{dis}, as the default option. Z-score is defined as follows :

$$Z(q,l) = \frac{S(q,l) - E(q)}{\sigma(q)} \tag{4}$$

where q and l represent proteins, $S(q,l)$ is the similarity score of proteins q and l, and $E(q)$ and $\sigma(q)$ are the average value and the standard deviation of score of protein q over the database. Our distance score S_{dis} correlates with the square of N_{comp}, which is the number of compared residues. We therefore employed a quadratic normalization, in which $E(q)$ and $\sigma(q)$ in Equation (4) are determined by the least-square fitting of the similarity score. The regression line $S_q^{reg}(q,l) = A_q N_{comp}^2(q,l) + B_q$ is calculated for the score $S(q,l)$ of the query protein q against proteins l stored in the database, by the fitting parameters A_q and B_q. Then, the $E(q)$ in Equation (4) is replaced by $S_q^{reg}(q,l)$. $\sigma(q)$ is obtained by the averaged error of the regression line, $\sigma(q) = \sqrt{\sum_l^{N_{pro}} [S(q,l) - S_q^{reg}(q,l)]^2 / N_{pro}}$, where N_{pro} is the number of protein chains in the database. Because we assume that all the parameters A_q,B_q, and $\sigma(q)$ should be derived from the non-homologous proteins, we repeat the esimation two times. First, all the proteins in the library are used to estimate the parameters. Using these parameters, we calculate Z-score for each protein. To extract non-homologous protein, we chose proteins with Z-score $\leq$ 4.0, and reesimate the parameters and Z-score using the extracted non-homologous proteins.

5.2 Options

- **-R** [q or Q or S or R]

  This option is how to determine ranks for similar structures.

  - **-R q**: The default option. Matras employs N_{comp}^2 fitted Zscore of S_{dis} using protein with the plain Z-score $\leq$ 4.0, described in the previous subsection (Equation(4)).

- **-R Q**: Matras employs N^2_{comp} fitted Zscore of S_{dis} using proteins with the N^2_{comp} fitted Z-score ≤ 4.0.
- **-R S**: Matras ranks structure by the plain Z score of S_{sse}. using protein with the plain Z-score ≤ 4.0.
- **-R R**: Matras ranks structures by the score R_{dis}, which is a normalized S_{dis} score, defined in Equation (2). The score R_{dis} has a lower discrimination power than the Z-scores, however, its performance is not affected by the structural library. If your structural library is small or redundant, we recommend to use the R_{dis} (-R R) option.

- **-zt**

  A threshold Z-score value for showing the similar structure list, in the case of "-R q", "-R Q" and "-R S". The default value is 5.0.

- **-rt**

  A threshold Z-score value for showing the similar structure list, in the case of "-R R". The default value is 10.0.

5.3 Output File for 3D Library Search

An example of the result of "3D Library Search" is shown at the end of this section. Basically, the result is composed of following six parts:

- header

  This part shows basic properties of a query structure and a structure library, such as size and number of proteins.

- [BEST_SCORE_RANKING]

  This part shows library structures whose Zscore is more than threshold, with following properties.

  - rk : rank. In default, entries are sorted by Zscore.
  - entry : entry code. Normally, it is a combination of PDBcode and Chain Identifier.
  - start : Residue number string of the first aligned residue of a library structure.
  - end : Residue number string of the last aligned residue of a library structure.
  - Rdis : Normalized S_{dis} score(%), defined in Equation (2).
  - Zsc : Zscore. In default, it is N^2_{comp} fitted Z-score of S_{dis}, defined in Equation (4).
  - SqID : Sequence Identity (%)

- [BEST_SCORE_RANKING_WITH_DETAILED_INFORMATION]

  This part shows library structures whose Zscore is more than threshold, with following properties.

- rk : rank. In default, entries are sorted by Zscore.
- entry : entry code. Normally, it is a combination of PDBcode and Chain Identifier.
- Naa : Number of amino acids of a library structure.
- Ncmp : Number of compared residues,N_{comp}.
- SqID : Sequence Identity (%)
- rms : RMSD of aligned C^{α} atoms (Å)
- Ssse : Score of SSE (S_{dis}).
- Rsse : Normalized Ssse score (%), defined in Equation (2).
- Sdis : Score of SSE (S_{sse}).
- Rdis : Normalized S_{dis} score (%), defined in Equation (2).
- Zsc : Zscore. In default, it is N_{comp}^2 fitted Z-score of S_{dis}, defined in Equation (4).
- RelS : Reliability(%) that a pair with this Zsc belongs to the same SCOP superfamily.
- RelO : Reliability(%) that a pair with this Zsc belongs to the same SCOP fold.

• [BEST_RANKING_WITH_ONE_LINE_SECONDARY_STRUCTURE]

  This part shows all the alignments between a query structure and library structures in one line, using secondary structure symbols. staQ and endQ represents start and end residues for query, staL and endL represents those for library. If Matras finds partial similarity, this part clearly shows where is the aligned region.

• [CLUSTALW_STYLE_ALIGNMENT]

  This part shows all the pairwise alignments as a master-slave multiple alignment. Note that this is not a multiple alignment in the strict meaning, because all the library structures are aligned to the query structure, not aligned between library structures.

• [ALIGNMENTS]

  This part shows all the pairwise alignment one by one. Their formats are exactly similar to that of "Pairwise 3D Alignment".

5.4 An example of the result of "3D Library Search"

A following long text is an example of the result of 3D library search, obtained by a following command.

```
% Matras L -Q 4azuA.bssp -L 30scop1.71nm.list -R q -zt 5
```

The query structure is “4azuA”, and the library list file is “30scop1.71nm.list”,which is the 30 % representative list of structural domains registered in SCOP 1.71.

```
#### MATRAS VER 1.2: PROGRAM FOR PROTEIN 3D STRUCTURE COMPARISON ####
# coded by Takeshi Kawabata. Last Modified : Apr 15, 2007
#<REFERENCE>
# Takeshi Kawabata and Ken Nishikawa.
# "Protein Structure Comparison Using the Markov Transition Model of Evolution".
#   Proteins vol.41:108-122(2000).
#<COMMAND> "Matras L -Q 4azuA.bssp -L 30scop1.71nm.list -R q -zt 5 "
#<DATE> "Apr 15,2007 11:40:0"
#<MODE> L:ONE-VS-LIBRARY COMPARISON
#<PARAMETERS>
# QueFile "L" LibFile "4azuA.bssp"
# SseAliType T EnvAliType T AlgType L
# ssescfile "3U" envscfile "T10-3U.rom" DisSc N disscfile "3U" DisScE - EnvType T Nenvs
tate 10
# GapExtE -6.0 GapExtD -100.0 Nkeep 35 Nrep 10 SseOffsetD 1
[QUERY_PROTEIN]  4azuA.bssp
[QUERY_COMPND]   AZURIN (PH 5.5)
[QUERY_SIZE]     Naa 128 Nsse 10
[MAX_NAAB]       1450
[LIBRARY_FILE]   30scop1.71nm.list
[LIBRARY_SIZE]   5931
[WAY_OF_RANKING] DisSc Ncmp^2-fit Zsc after plain Zsc filter( 2.219025 *x*x + -68.313 S
D 1781.572 )
[Z_THRESHOLD]    5.000000
[NRANK]          123

BEST_SCORE_RANKING]
rk  entry  start  end    Rdis  Zsc   SqID   MOLECULAR NAME

1   1jzgA       1    128  93.7  71.54 100.0 "AZURIN"
2   1e30A      37    155  35.2  28.46  16.3 "RUSTICYANIN"
3   1fwxA1    486    579  34.4  24.13  18.5 "NITROUS OXIDE REDUCTASE"
4   1oe1A1     31    151  26.6  21.19  19.4 "DISSIMILATORY COPPER-CONTAINING NITRITE"
5   1gskA1     25    175  24.5  21.10   6.9 "SPORE COAT PROTEIN A"
6   1cyx-     126    225  23.8  19.44  10.2 "CYOA"
7   1aozA1      3    122  33.2  19.39  13.5 "ASCORBATE OXIDASE (E.C.1.10.3.3)"
8   1kv7A1     43    163  29.6  19.30  15.0 "PROBABLE BLUE-COPPER PROTEIN YACK"
9   2cuaA      78    167  30.0  18.32  17.0 "CUA"
10  1hfuA1      5    127  32.6  18.27  12.0 "LACCASE 1"
:
121 1ulvA2    689    771  18.1   5.10   6.2 "GLUCODEXTRANASE"
122 1ti6B1    196    263  15.9   5.04  13.2 "PYROGALLOL HYDROXYTRANSFERASE SMALL SUBUNI
T"
123 1wmdA1    319    434  18.0   5.02   9.5 "PROTEASE"

[BEST_SCORE_RANKING_WITH_DETAILED_INFORMATION]
rk  entry  Naa  Ncmp SqID  rms  Ssse  Rsse   Sdis  Rdis  Zsc   RelS  RelO  TAXONOMY

1   1jzgA   128  128 100.0  0.8  2618  94.7 163738  93.7  71.54  87.7  94.2 [b.6.1.1]
2   1e30A   153  104  16.3  2.7  1136  26.7  74639  35.2  28.46  83.2  92.1 [b.6.1.1]
3   1fwxA1  132   92  18.5  2.0  1015  33.6  61703  34.4  24.13  82.3  91.5 [b.6.1.4]
4   1oe1A1  159   98  19.4  3.3   669  23.6  59002  26.6  21.19  80.5  91.0 [b.6.1.3]
5   1gskA1  174  102   6.9  3.8   741  26.1  60602  24.5  21.10  80.4  91.0 [b.6.1.3]
6   1cyx-   158   88  10.2  2.3   932  22.6  51757  23.8  19.44  78.8  90.8 [b.6.1.2]
7   1aozA1  129  104  13.5  5.1   835  36.7  58476  33.2  19.39  78.8  90.8 [b.6.1.3]
8   1kv7A1  140  100  15.0  4.5   831  25.9  56499  29.6  19.30  78.7  90.7 [b.6.1.3]
9   2cuaA   122   88  17.0  2.3   896  32.7  49748  30.0  18.32  77.4  90.6 [b.6.1.2]
10  1hfuA1  131  108  12.0  5.3   805  31.7  58371  32.6  18.27  77.3  90.6 [b.6.1.3]
:
121 1ulvA2   89   81   6.2  4.1   619  28.3  23580  18.1   5.10  28.0  53.6 [b.1.18.2]
122 1ti6B1   79   68  13.2  3.9   407  18.4  19175  15.9   5.04  27.5  53.1 [b.3.5.1]
123 1wmdA1  116   95   9.5  5.9   465  21.2  28902  18.0   5.02  27.4  52.8 [b.18.1.20]
```

```
[BEST_RANKING_WITH_ONE_LINE_SECONDARY_STRUCTURE]
rk  entry  staQ |staL |    .    .    .    .    :    .    .    .    .    +    .    .    | endL| endQ
   (query )    1|  ---|ccEEEccccEEccEEEEccHHHcEEcHHHHHHccccccEEcccEEEEccccEEccccEEE|---  |128
1   1jzgA      1|    1|ccEEEccccEEccEEEEccHHHcEEcHHHHHcccccccEEcccEEEEccccEEccccEEE|128  |128
2   1e30A      2|   37|-EEEEccEEEEccEEEEcc--ccEEcccccc---ccccEEcccEEEc-ccEEEccccEEE|155  |128
3   1fwxA1     2|  486|-EEEEEcEcEEEcEEEEccccEEc--------------EEEEcEEEEc-cEEEccccEEE|579  |128
4   1oe1A1     1|   31|cEEEEEEEcEEEcEEEEcc-cccEccc----------ccccccEEEEc-cEEEcHHHEEE|151  |128
5   1gskA1     2|   25|-EEEEEEEcEEEccEEEccccccEEcc----------cccccccEEcc-cEEEcHHHEEE|175  |128
6   1cyx-      2|  126|-EEEEcEEEEEccEEEc---cEEc--------------EEEEccEEc-ccEEEcccccEE|225  |128
7   1aozA1     2|    3|-EEEEEEEcEEEcEEEEcc-cccEEccccc---ccc---ccccEEEE-ccEEEccccEEE|122  |128
8   1kv7A1     1|   43|ccEEEEEEcEEEcEEEEc--cccEEcccc----cc-----cccEEEEcccEEEcHHHcEE|163  |128
9   2cuaA      2|   78|-EEEEcEcEEEcEEEEc---cEEc--------------cEEEccEEEcccEEEccccEEE|167  |128
10  1hfuA1     3|    5|-cEEEEEEcEEEccEEEccccccEEccccc---ccc---ccccEEEccccEEEccccEEE|127  |128
:
121 1ulvA2    10|  689|-----ccEEEEEcEEEE----cEEEcE-----------EEcc-EEEE-cEEEEEccEEEE|771  |128
122 1ti6B1    28|  196|-------------cEEEEcEccEEEc-----------EEEEccEEEEEE--EEEcE-EEE|263  |128
123 1wmdA1     5|  319|--cEEEcEEEEcccEEEccccEEEEc-----------cEEcccEEEccc-EEEEcEEEE|434  |127
```

```
[CLUSTALW_STYLE_ALIGNMENT]

CLUSTAL W (1.82) multiple sequence alignment

QUERY           AECSVDIQGNDQMQFNTNAITVDKSCKQFTVNLSHPGNLPKNVMGHNWVLSTAADMQGVV
1jzgA           AECSVDIQGNDQMQFNTNAITVDKSCKQFTVNLSHPGNLPKNVMGHNWVLSTAADMQGVV
1e30A           -TVHVVAAAVPFpSFEVPTLEIPAGA-TVDVTFINTNKG----FGHSFDITKK-GPp--Y
1fwxA1          -KVRVYMSSV-ApSFSIESFTVKEGD-EVTVIVTNLDEID--DLTHGFTMGN--------
1oe1A1          KVVEFTMTIEEKMTFNGPTLVVHEGD-YVQLTLVNPATN---AMPHNVDFHGATG-----
1gskA1          -KTYYEVTMEECWGYNGPTIEVKRNE-NVYVKWMNNLPSTHPEVKTVVHLHGGVT-----
1cyx-           -PITIEVVSM-DWKWFFNEIAFPANT-PVYFKVTSNS------VMHSFFIPR--------
1aozA1          -IRHYKWEVEYMMGINGPTIRANAGD-SVVVELTNKLH----TEGVVIHWHGILQRGTPW
1kv7A1          DRNRIQLTIGAGWGYNGPAVKLQRGK-AVTVDIYNQL-----TEETTLHWHGLEVPGEVD
2cuaA           -QYTVYVLAF-AfGYQpNpIEVPQGA-EIVFKITSPD------VIHGFHVEG--------
1hfuA1          --SVDTMTLTNAILVNGPLIRGGKND-NFELNVVNDLDNPTMLRPTSIHWHGLFQRGTNW
:
1ulvA2          ---------LSSPELSVTApESTADSATAVVRGTT--------NAAKVYVSVNGT-----
1ti6B1          ---------------------------KNYVTAGILVQGDCF-EGAKVVLKSGG------
1wmdA1          ----AYVSSLSTSQKATYSFTATAGK-PLKISLVWSDAPVTLVNDLDLVITAPN------

QUERY           TDGMASGLDKDYLKPDDSRVIAHTKLIGSGEKDSVTFDVSKLKEGEQYMFFCTFPGHSAL
1jzgA           TDGMASGLDKDYLKPDDSRVIAHTKLIGSGEKDSVTFDVSKLKEGEQYMFFCTFPGHSAL
1e30A           AV-M--------PV--IDpIVAGTGFSPVPGYTNFTWH---PTA-GTYYYVCQIPGHAAG
1fwxA1          -------------------YGVAME-IGPQMTSSVTFVAAN---PGVYWYYCQWFALHME
1oe1A1          ------------------ALGGALTNVNPGEQATLRFKADR---SGTFVYHCAPMWHVVG
1gskA1          -----------------PDDSDGYAWFSKDFREVYHYPNQQ--RGAILWYHDHARLNVYG
1cyx-           -------------------LGSQIY-AMAGMQTRLHLI---ANEPGTYDGICAEIPGHSG
1aozA1          ADG-------TASI--------SQCAINPGETFFYNFT---VDNPGTFFYHGHLGMQRSG
1kv7A1          G---------GPQ-----------GIIPPGGKRSVTLNVD--QPAATCWFHPHQHRQVAG
2cuaA           -------------------TNINVE-VLPGEVSTVRYTFK--RP-GEYRIICNQYLGHQN
1hfuA1          ADG-------ADGV--------NQCPISPGHAFLYKFTPA--GHAGTFWYHSHFGTQYCG
:
1ulvA2          --------------------ATEAPVTD--GTFSLDVAL--TGAKNKVTVAAVAADG-GT
1ti6B1          ------------------KEVASAETNFF-GEFKFDALDNGE-----YTVEIDADGKS--
1wmdA1          ------------------GTQYVGNWDGRNNVENVFIN-APQS--GTYTIEVQAYNVpQT


QUERY           MKGTLTLK
1jzgA           MKGTLTLK
1e30A           QFGKIVVK
1fwxA1          MRGRMLVE
1oe1A1          MSGTLMVL
1gskA1          LVGAYIIH
1cyx-           MKFKAIAT
```

```
1aozA1          LYGSLIVD
1kv7A1          LAGLVVIE
2cuaA           MFGTIVVK
1hfuA1          LRGPMVIY
:
1ulvA2          AVEDRTVL
1ti6B1          YSDTVVID
1wmdA1          FSLAIVN-
```

```
[ALIGNMENTS]
>1 1jzgA [b.6.1.1] "AZURIN"
#Naa 128 start 1 end 128 SqID 100 % crms  0.8 Ssse  2618 Sdis 163738 Rdis  93.7 Z 71.54

     :   E1             E2       E3           H1      E4     H2
SecA : TTEEEEEB TTS BS SEEEE TT SEEEEEEE  SS  HHHH B  EEEETTTHHHHH
    1:AECSVDIQGNDQMQFNTNAITVDKSCKQFTVNLSHPGNLPKNVMGHNWVLSTAADMQGVV:60
      ************************************************************
    1:AECSVDIQGNDQMQFNTNAITVDKSCKQFTVNLSHPGNLPKNVMGHNWVLSTAADMQGVV:60
SecB :   EEEEEB TTS BS SEEEE TT SEEEEEEE  SSS HHHH B  EEEEGGGHHHHH
     :   E1             E2       E3           H1      E4     H2

     :                     E5        E6              E7
SecA :HHHHHH GGGTTS TT TT SEE   B TT EEEEEEEEGGGS TT  EEEE  STTTTTT
   61:TDGMASGLDKDYLKPDDSRVIAHTKLIGSGEKDSVTFDVSKLKEGEQYMFFCTFPGHSAL:120
      ************************************************************
   61:TDGMASGLDKDYLKPDDSRVIAHTKLIGSGEKDSVTFDVSKLKEGEQYMFFCTFPGHSAL:120
SecB :HHHHTT GGGTTS TT TT  EE   B TT EEEEEEEEGGG  TT  EEEE  STTGGGT
     :                     E5        E6              E7

     : E8
SecA :SEEEEEE
  121:MKGTLTLK:128
      ********
  121:MKGTLTLK:128
SecB :SEEEEEE
     : E8

//

>2 1e30A [b.6.1.1] "RUSTICYANIN"
#Naa 153 start 37 end 155 SqID  16 % crms  2.7 Ssse  1136 Sdis  74639 Rdis  35.2 Z 28.46

     :  E1     ----       ----- E2       E3           H1      E4
SecA :TTEEEEEB ----TTS BS -----SEEEE TT SEEEEEEE  SS  HHHH B  EEEE
    2:ECSVDIQGN----DQMQFNT-----NAITVDKSCKQFTVNLSHPGNLPKNVMGHNWVLST:52
         *             *                    *             **
   37:TVHVVAAAVLPGFPFpSFEVHDKKNPTLEIPAGA-TVDVTFINTNKG----FGHSFDITK:91
SecB :EEEEEEEES TTS SS EEETTEES EEEE TT -EEEEEEE  TT----     EEES
     :E4              E5   E6  E7      -E8         ----     E9

     :   H2                        E5       ----- E6
SecA :TTTHHHHHHHHHHHH GGGTTS TT TT SEE   B TT----- EEEEEEEEGGGS TT
   53:AADMQGVVTDGMASGLDKDYLKPDDSRVIAHTKLIGSG-----EKDSVTFDVSKLKEGEQ:107
                 *                 * *                *
   92:K-GPp--YAV-M--------PV--IDpIVAGTGFSPVPKDGKFGYTNFTWH---PTA-GT:133
SecB : - SS--  S-S--------  --  SEEEEB      BTTEEEEEEEEE ---  S-EE
     : -   --   - --------  --   E10          E11       ---   -E1

     :E7          -  E8
SecA :EEEE  STTTTT-TSEEEEEE
  108:YMFFCTFPGHSA-LMKGTLTLK:128
      *   *  *** *    *    *
```

```
  134:YYYVCQIPGHAATGQFGKIVVK:155
SecB :EEEE  STTTTTTT EEEEEE
     :2              E13

//



###                             :                               ###
###                             :                               ###
### SKIPPING THE PAIRWISE ALIGNMENTS FROM THE 3RD TO THE 122-TH ###
###                             :                               ###
###                             :                               ###


>123 1wmdA1 [b.18.1.20] "PROTEASE"
#Naa 116 start 319 end 434 SqID   9 % crms  5.9 Ssse   465 Sdis  28902 Rdis  18.0 Z 5.02

     :1  ---           E2       E3          ------ H1      E4
SecA :EEE---EB TTS BS SEEEE TT SEEEEEEE  SS ------ HHHH B  EEEETTT
    5:VDI---QGNDQMQFNTNAITVDKSCKQFTVNLSHPGNL------PKNVMGHNWVLSTAAD:55
                  *  *   *           *               *     *
  319:AYVNESSSLSTSQKATYSFTATAGK-PLKISLVWSDAPASTTASVTLVNDLDLVITAPN-:376
SecB : EEEEEEEE TT EEEEEEEE TTS- EEEEEE       TT S    SEEEEEEE TT-
     : E1          E2          - E3                    E4        -

     :H2                        E5   ---------      E6
SecA :HHHHHHHHHHH GGGTTS TT TT SEE  --------- B TT EEEEEEEGGGS TT
   56:MQGVVTDGMASGLDKDYLKPDDSRVIAHTK---------LIGSGEKDSVTFDVSKLKEGE:106
                                               *      *
     :-----------------------GTQYVGNDFTSpYNDNWDGRNNVENVFIN-APQS--G:410
SecB :-----------------------S EEETT  SSSTTS   SS SEEEEEES-S  S--E
     :-----------------------  E5                  E6     -    --E

     : E7         ---    E8
SecA : EEEE  STTT---TTTSEEEEEE
  107:QYMFFCTFPGH---SALMKGTLTL:127
       *
  411:TYTIEVQAYNVPVGpQTFSLAIVN:434
SecB :EEEEEEEEEE SS  EEEEEEEE
     :7              E8

//
```

6 All-vs-all 3D comparison

All-vs-all 3D comparison is a calculation of similarities for all the structural pairs in a library file.

```
% Matras A -L [library_listfile]
```

The format of libaray_listfile is described in the appendix. If a following listfile is used as inputs,

```
1mbd-
1ecd-
4hhbA
4hhbB
```

a following result will be obtained:

```
#### MATRAS VER 1.2: PROGRAM FOR PROTEIN 3D STRUCTURE COMPARISON ####
# coded by Takeshi Kawabata. Last Modified : Feb 6, 2004
#<REFERENCE>
# Takeshi Kawabata and Ken Nishikawa.
# "Protein Structure Comparison Using the Markov Transition Model of Evolution".
#   Proteins vol.41:108-122(2000).
#<COMMAND> "Matras A -L globinlist "
#<DATE> "May 9,2004 11:20:5"
#<MODE> A:ALL-VS-ALL COMPARISON
#<PARAMETERS>
# LibFile "globinlist"
# SseAliType T EnvAliType T AlgType L
# ssescfile "3U" envscfile "T10-3U.rom" DisSc N disscfile "3U" DisScE - EnvType T Nenvstate 10
# GapExtE -6.0 GapExtD -100.0 Nkeep 35 Nrep 10 SseOffsetD 0
#[Matras A -L globinlist ]
#Nlibrary 4 MaxNaaLib 0
#READ ALL THE STRUCTURE
#Nlib 4 Ncomb 10
#AVAbunshi/bunbo  [0]/[1]
#Npair_start  0 Npair_end 10 Npair_to_be_calculated 9
#MALLOC FOR DP:MaxNaaA 153 MaxNaaB 153
#COLDEF [proA] [proB] [NaaA] [NaaB] [Ncomp] [ScSSE] [ScEnv] [ScDis] [SqID] [CRMS] [Rdis] [Rsse]
1mbd- 1mbd-  153  153  153 1715.7 5526.4 227844.9 100.00  0.00 100.00 100.00
1mbd- 1ecd-  153  136  136 1313.6 4125.9 132476.6 20.59  1.65 65.14 76.11
1mbd- 4hhbA  153  141  141 676.0 4007.8 149312.6 26.95  1.56 70.68 44.82
1mbd- 4hhbB  153  146  145 1017.1 4428.8 157392.8 24.83  1.62 72.03 67.47
1ecd- 1ecd-  136  136  136 1736.2 4798.5 178893.6 100.00  0.00 100.00 100.00
1ecd- 4hhbA  136  141  131 610.5 3704.9 106591.4 18.32  2.40 57.07 40.21
1ecd- 4hhbB  136  146  136 889.4 3730.7 114607.0 19.12  2.25 59.07 58.61
4hhbA 4hhbA  141  141  141 1300.4 4817.5 194641.9 100.00  0.00 100.00 100.00
4hhbA 4hhbB  141  146  139 685.5 4124.9 155279.5 43.88  1.45 76.91 52.74
4hhbB 4hhbB  146  146  146 1299.1 5031.7 209173.0 100.00  0.00 100.00 100.00
```

7 Multiple 3D Alignment

Multiple 3D alignment is a comparison more than two 3D structures, and getting alignments for these multiple structures. This is done by a Perl script name "mulmat.pl", which is in the BASE_DIR directory. This script calls the Matras program several times to get pairwise alignments, and it makes a multiple alignment by assembling these pairwise alignments.

7.1 Algorithm

Getting the optimal multiple alignment for sequences is a very hard computational problem, and getting the one for 3D structure is harder. Therefore, we employed a popular heuristics, called "progressive alignment", and it it done by simply assembling pairwise alignments.

- **Step1** : Calculate pairwise alignments and similarities for all structural pairs.

  The script "mulmat.pl" executes the Matras program, and stores all the results of pairwise alignment in the TMP_DIR directory (assigned in the ".matras" file).

- **Step 2** : Make a dendrogram using these similarities.

  The scripts executes a program "TreeUN" for making a dendrogram using UPGMA method.

- **Step 3**: Starting from the leaf nodes, progressively align all nodes, in order of decreasing similarity.

7.2 Basic Operation

If you execute **mulmat.pl** without any arguments, following help messages are shown:

```
% mulmat.pl [str1] [str2]... [strN] (-options)
  for 'mul'tiple 3D alignment using 'Mat'ras
  written by Takeshi Kawabata. LastModDate :Dec 26, 2003
 <options>
 -F       : strucutre list file[]
 -TMP_DIR : temporary output dir[/home/takawaba/work/Matras12/tmpout]
 -RES_DIR : result   output dir[.]
 -ad      : alignment file directory[]
 -ow      : Outfile in ClustalW[-]
 -ov      : Outfile in Vertical style[]
 -ovp     : Outfile in Vertical style with Plain Residue Num[]
 -oh      : Outfile in Horizontal style[]
 -ohs     : Outfile in Horizontal SecStr[]
 -ohtml   : Outfile in Horizontal SecStr HTMLfile []
 -ocon    : Outfile for consensus sequence []
 -opdb    : Outfile for sup-imposed PDBs[]
```

```
-oph     : Outputfile for guided UPGMA tree[]
-ops     : Outputfile PSI-BLAST multiple alignment[]
-OS      : Output StrType 'B'ssp, 'P'db [B]
-rhead   : header of all the result outputfile[]
-thead   : header of all the temporary outputfile[]
-so      : Matras SubOptimal[F]
-QO      : seQuence Order ('T'ree)[T]
-dmat    : Output distance matrix file[]
-smat    : Output similarity matrix file[smat]
-rm      : Remove Temoporary File (T|F) [F]
-M       : do MATRAS (T|F) [T]
-T       : do Tree (T|F) [T]
```

Its basic procedure to run is as follows:

```
% mulmat.pl [bsspfile1] [bsspfile2] [bsspfile3] ....
```

We show an example for multiple alignments of 1mbd-.bssp, 1ecd-.bssp, 4hhbA.bssp and 4hhbB.bssp. You can omit their tail string ".bssp".

```
% mulmat.pl 1mbd- 1ecd- 4hhbA 4hhbB
```

If you want to compare many structures, we recommend that you make a file that contains protein names (one protein per one line), and assign the file using "-F" option. For example, firstly, you make a following file named "listfile",

```
1mbd-
1ecd-
4hhbA
4hhbB
```

and execute "mulmat.pl" using a following options.

```
% mulmat.pl -F listfile
```

7.3 Options

- **-ow** [outputfile]

  Assign an output file name in ClustalW formats.

- **-ov** [outputfile]

  Assign an output file name in vertical formats, or CASP-style multiple alignment (its format is shown in Appendix).

- **-ovp** [outputfile]

  Assign an output file name in vertical formats, or CASP-style multiple alignment with plain residue number (its format is shown in Appendix).

- **-opdb** [output pdbfile]

  This option is for output superimposed multiple structures. It is not easy to find the optimal super imposition for multiple structures. We employ a simple strategy : first find the "center" structure, and superimpose other structures to the center one.

  Simultaneously, two RasMol scripts named "mulgrp.ras" and "mulchn.ras" are written. The former is coloring aligned regions, the latter is coloring by proteins. For example, if you want to color by aligned regions, you execute following commands:

```
rasmol [output pdbfile]
RasMol>script "mulgrp.ras"
```

- **-smat** [filename]

  Assign an output file for various similarities. It contains four kinds of similarities : R_{dis}, RMS, DRMS and SqID. The following is an example.

```
[RDIS(%)]
1mbd-              0.0  65.1  70.7  72.0
1ecd-             65.1   0.0  57.1  59.1
4hhbA             70.7  57.1   0.0  76.9
4hhbB             72.0  59.1  76.9   0.0
[RMS(A)] #for aligned Calpha atoms
1mbd-            0.000 1.652 1.560 1.617
1ecd-            1.652 0.000 2.397 2.252
4hhbA            1.560 2.397 0.000 1.451
4hhbB            1.617 2.252 1.451 0.000
[DRMS(A)] #for aligned Cbeta atoms
1mbd-            0.000 1.462 1.398 1.417
1ecd-            1.462 0.000 1.916 1.875
4hhbA            1.398 1.916 0.000 1.125
4hhbB            1.417 1.875 1.125 0.000
[SqID(%)]
1mbd-            100.0  20.6  27.0  24.8
1ecd-             20.6 100.0  18.3  19.1
4hhbA             27.0  18.3 100.0  43.9
4hhbB             24.8  19.1  43.9 100.0
```

APPENDIX

A File format

A.1 BSSP

The BSSP file is very similar to the DSSP file. Only difference between them BSSP lacks the fields named “TCO”, “KAPPA” and “ALPHA” in DSSP files and has additional fields name “X-XB”, “Y-CB” and “Z-CB”, which are coordinates of C^{β} atoms.

```
<DSSP file format>
                                                                                                   1         1         1
         1         2         3         4         5         6         7         8         9         0         1         2
12345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC   N-H-->O  O-->H-N  N-H-->O  O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
    1    1 A V              0   0  155    0, 0.0   2,-0.4   0, 0.0 127,-0.1  0.000 360.0 360.0 360.0 144.8    6.9   17.8    4.6
    2    2 A L        -     0   0   20  71,-0.1 122, 0.0   1,-0.1   0, 0.0 -0.791 360.0-141.9 -92.9 121.5   10.6   17.9    4.3
    3    3 A S     >  -     0   0   44  -2,-0.4   4,-2.8   1, 0.0   5,-0.2 -0.150  29.4-103.9 -60.8-176.0   12.3   19.9    7.1
    4    4 A P  H  > S+     0   0   99    0, 0.0   4,-2.9   0, 0.0   5,-0.3  0.997 124.4  56.4-100.2  -1.8   15.0   21.9    6.2
```

```
<BSSP file format>
                                                                                                   1         1         1
         1         2         3         4         5         6         7         8         9         0         1         2
12345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678
#    RESIDUE AA STRUCTURE  BP1  BP2  ACC  N-H-->O   O-->H-N  N-H-->O  O-->H-N    X-CB  Y-CB  Z-CB  PHI  PSI     X-CA   Y-CA   Z-CA
    1    1 A V              0   0  155    0, 0.0   2,-0.4   0, 0.0 127,-0.1    6.4  19.0   5.8 360.0 144.8    6.9   17.8    4.6
    2    2 A L        -     0   0   20  71,-0.1 122, 0.0   1,-0.1   0, 0.0   11.1  18.0   2.8 -92.9 121.5   10.6   17.9    4.3
    3    3 A S     >  -     0   0   44  -2,-0.4   4,-2.8   1, 0.0   5,-0.2   12.7  19.0   8.2 -60.8-176.0   12.3   19.9    7.1
    4    4 A P  H  > S+     0   0   99    0, 0.0   4,-2.9   0, 0.0   5,-0.3   15.8  23.1   7.2-100.2  -1.8   15.0   21.9    6.2
```

A.2 CASP-style pairwise alignment

This is the format of pairwise alignment, which are used in CASP.

```
NPRO 2
PRO1 [Protein 1]
PRO2 [Protein 2]
COMMENT [Comment Line]
ALIGNMENT
[ResNum1] [ResName1] [ResNum2] [ResName2]
:
END
```

The residue names [ResName] must be written in one character way. The residue numbers [ResNum] must be identical to those in PDB files (23-27columns). If the ”RNUMPLAIN” line appears, then plain residue number (named from 1 by integer) are used. The residue name for a inserted/deleted position must be assigned as ’-’, and the residue number for indel position must be “-1”. Matras also output parameters for superimposing in this file. A following is an example.

```
NPRO 2
PRO1 1timA.bssp
PRO2 1kv8A.bssp
```

```
COMMENT Naa1  247 Naa2  213
COMMENT Ncomp 195 SqID 10.3 RMS 3.388 DRMS 2.755
COMMENT ScDis 188198.0 Rdis 34.8
PARAM_FOR_SUPERIMPOSING
#Afit=R*(A-Ga)+Gb
Ga 43.78974 29.88718  2.43385
Gb 64.59436 12.80051 25.07641
R0  0.86940 -0.47082 -0.14987
R1  0.23641  0.13004  0.96291
R2 -0.43387 -0.87259  0.22437
ALIGNMENT
K 5 L 3
F 6 P 4
F 7 M 5
V 8 L 6
G 9 Q 7
G 10 V 8
:
K 237 D 196
P 238 A 197
- -1 A 198
- -1 S 199
- -1 P 200
- -1 V 201
- -1 E 202
E 239 A 203
F 240 A 204
V 241 R 205
D 242 Q 206
I 243 F 207
I 244 K 208
- -1 R 209
N 245 S 210
A 246 I 211
K 247 A 212
H 248 E 213
END
```

A following is an example with "RNUMPLAIN".

```
NPRO 2
PRO1 1timA.bssp
PRO2 1kv8A.bssp
COMMENT Naa1  247 Naa2  213
COMMENT Ncomp 195 SqID 10.3 RMS 3.388 DRMS 2.755
```

```
COMMENT ScDis 188198.0 Rdis 34.8
RNUMPLAIN
PARAM_FOR_SUPERIMPOSING
#Afit=R*(A-Ga)+Gb
Ga 43.78974 29.88718  2.43385
Gb 64.59436 12.80051 25.07641
R0  0.86940 -0.47082 -0.14987
R1  0.23641  0.13004  0.96291
R2 -0.43387 -0.87259  0.22437
ALIGNMENT
K 4 L 1
F 5 P 2
F 6 M 3
V 7 L 4
G 8 Q 5
G 9 V 6
:
K 236 D 194
P 237 A 195
- -1 A 196
- -1 S 197
- -1 P 198
- -1 V 199
- -1 E 200
E 238 A 201
F 239 A 202
V 240 R 203
D 241 Q 204
I 242 F 205
I 243 K 206
- -1 R 207
N 244 S 208
A 245 I 209
K 246 A 210
H 247 E 211
END
```

A.3 CASP-style multiple alignment

This format is for multiple alignment, using a similar strategy of CASP-style pairwise alignment.

```
NPRO [Number of Proteins]
PRO1 [Proteine Name 1]
PRO2 [Protein Name 2]
:
PRO[N] [Protein Name N]
```

```
COMMENT [Comment]
ALIGNMENT
[ResNum1] [ResName1] [ResNum2] [ResName2].... [ResNumN] [ResNameN]
:
END
```

The residue names [ResName] must be written in one character way. The residue numbers [ResNum] must be identical to those in PDB files (23-27columns). If the "RNUMPLAIN" line appears, then plain residue number (named from 1 by integer) are used. The residue name for a inserted/deleted position must be assigned as '-', and the residue number for indel position must be "-1". A following is an example.

```
NPRO 4
PRO1 1mbd-
PRO2 1ecd-
PRO3 4hhbA
PRO4 4hhbB
ALIGNMENT
- -    - -    - -    V 1
V 1    - -    V 1    H 2
L 2    L 1    L 2    L 3
S 3    S 2    S 3    T 4
E 4    A 3    P 4    P 5
G 5    D 4    A 5    E 6
E 6    Q 5    D 6    E 7
W 7    I 6    K 7    K 8
Q 8    S 7    T 8    S 9
L 9    T 8    N 9    A 10
:
:
Y 151  - -    - -    - -
Q 152  - -    - -    - -
G 153  - -    - -    - -
END
```

A.4 Library list file

This file is for the 3D library search, contains names of BSSP files for the structure library. The format is as follows :

```
#[COMMENT]
bsspfile_head1 comment_for_bsspfile1
bsspfile_headr2 comment_for_bsspfile2
:
:
#MAXLENGTH [MAXIMUM_AA_LENGTH_IN_LIBRARY]
```

The first field splited by spaces is for a library BSSP file. The following fields are for comments of each library structures. You can put anything such a protein name and a taxonomy id in these fields. The bottom line started by "#MAXLENGTH" is the maximum length of proteins in the library. If you omit this line, Matras uses the default value (1500 amino acids).

We show an example of a list file using SCOP taxonomy ID as comments.

```
119l- d.2.1
1a02F h.1.3
1a04A c.23.1 - a.4.6
1a0aA a.38.1
1a0i- d.142.2 - b.40.4
1a0p- a.60.9 - d.163.1
:
#MAXLENGTH 1419
```

B References

(1) The original article of Matras

Kawabata T., Nishikawa K. (2000). Protein tertiary structure comparison using the Markov transition model of evolution. *Proteins*,**41**, 108-122.

(2) The article for Matras WEB server

Kawabata T. (2003). MATRAS: a program for protein 3D structure comparison. *Nucleic Acids Res.*,**31**, 3367-3369.

(3) PDB : Protein 3D Structure Database

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E., 2000. The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242.

http://www.rcsb.org/pdb/

(4) DSSP : Program for Secondary Structure Assignment

Kabsh W, Sander C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**,2577-2637.

http://swift.cmbi.ru.nl/gv/dssp/index.html

(5) RasMol : Molecular Graphics Program

http://www.openrasmol.org/

(6) SCOP : Database of Protein 3D structure Classification

Murzin A.G., Brenner S.E., Hubbard T., Chothia C. (1995). SCOP:a structural classification of protein database for the investigation of Sequences and Structures. *J. Mol. Biol.* **247**, 536-540.

http://scop.mrc-lmb.cam.ac.uk/scop/

(7) CATH : Database of Protein 3D structure Classification

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997) CATH- A Hierarchic Classification of Protein Domain Structures. *Structure.* Vol 5. No 8. p.1093-1108.

http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html

(8) DALI server : Server for automatic comparison of Protein 3D structures

Holm L, Sander C, (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**,123-138.

http://www2.ebi.ac.uk/dali/

(9) ClustalW : Programs for Multiple Sequence Alignments

Thomas J.D., Higgins D.G., Gibson T.J. (1994). CLUSTAL W:improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, **22**, 4673-4680.