NAME

pima.sh - pattern-induced multi-sequence alignment program

SYNOPSIS

pima.sh *cluster_name cluster_score_cutoff seq_filename*
[*ref_seq_name sec_struct_seq_filename* [*ss_gap_penalty*]]

EXAMPLES

pima.sh SAMPLE 25.0 sample-family.pep
pima.sh SAMPLE-STRUCT 25.0 sample-struct.pep 1ldm pdb-dssp.ss

DESCRIPTION

pima.sh performs a multi-sequence alignment of a set of (presumably related) sequences using an extension of our covering pattern construction algorithm (Smith and Smith 1990, 1992). All pairwise comparisons between sequences in the set are performed and the resulting scores clustered into one or more families using using two different linkage rules: 1) maximal linkage (Smith and Smith, 1990) and 2) sequential branching (see Smith and Smith, 1992). For the latter, all pairwise scores are sorted high-to-low, the first sequence from the highest scoring pair is chosen as the "reference sequence", and the sequences clustered based strictly on the order of similarity to the reference sequence. Each cluster is then multiply-aligned using a pattern-based alignment algorithm (Smith and Smith, 1992). Patterns are constructed using the amino acid class hierarchy shown below.

If secondary structure sequences are provided for one or more of the primary sequences (one of which must be designated as a "reference sequence") then the sequences are clustered using the sequentially branching rule and the set multiply-aligned using a secondary structure- dependent gap penalty algorithm (Smith and Smith, 1992).

```
                  Amino Acid Classes                     Match score

                                                              -2
          _______________ X _________________                  0
         /           /            \               \
      _ f _         /          ______r ________    \           1
     /  /    \     /          /    /      \      \     \
    /  c      \   e          /   m        p      \  _ j __     2
   /  /  \     \ / \        /   / \      / \      \ /   \  \
  /  a    b     d   \      /   l   k   o   n      i      h  \   3
 /  / \  / \   /|\   \    /   / \ / \ / \  /\    / \    / \  \
C   I V  L M  F W Y   H   N   D   E   Q  K R   S T   A G   P    5
```

Note: During clustering, matches between unrelated amino acids, i.e. connected only through the 'X' (wild-card) class, score -2; also the final (summed) alignment scores are length-normalized by multiplying by $\log(.05)/\log(n*m)$, where n and m are the lengths of the two aligned sequences. During pattern construction/multi-alignment, matches between unrelated residues or matches between any amino acid (or any class) and 'X' score 0; the alignment score between any two sequences will thus equal the score of aligning the resulting "child" pattern (the pattern produced by the alignment) back against either (both) of the two parental sequences.

PARAMETERS

cluster_name

An arbitrary name used to label the cluster.

cluster_score_cutoff

The lowest match score to be used to incorporate a sequence into a cluster. Lowest recommended value: **25.0** ; a value of 0.0 will force all input seqs into 1 cluster, but the final pattern may be completely degenerate.

sequence_filename

Name of the input file containing the sequences to be clustered and multi-aligned. Sequences must be in a table format: LOCUS_NAME<tab>SEQUENCE (one seq/line !!). LOCUS_NAMES **must** start with a letter [A-Z or a-z].

ref_seq_name

[optional; if specified, then sec_struct_seq_file must also be specified] is the locus name of one of the primary sequences for which the secondary structure is in the file seq_struct_seq_filename.

sec_struct_seq_filename

[optional; if specified, then REF_SEQ_NAME must also be specified] contains secondary structure sequences for one or more of the primary sequences in the set. The secondary structure sequences in the file sec_struct_seq_filename must be in table format and the locus name of each sequence must be the locus name of it's corresponding primary sequence with the suffix '.ss' (e.g. 1ldm.ss). An alpha-helix, 3-10 helix and beta-strand must be designated 'h', 'g', and 'e', repectively. All other characters in the secondary structure sequences will be ignored with respect to the the structure-dependent gap penalty. To allow gaps to be placed between the first and the second and the last elements of these structures, the first and last 2 elements of each should be changed to another character designation. In the secondary structure sequence file **pdb-dssp.ss** provided with this package, these end cap elements are designated 'i', 'f', and 'd', for alpha-helices, 3-10 helices and beta-strands, respectfully.

ss_gap_penalty

[optional; default: 66.7] is the penalty for placing a gap within a alpha- or 3-10 helix or a beta-strand.

OUTPUT FILES CREATED

cluster_name--ML|SB][.ext].cluster

The cluster tree(s)s created by the clustering algorithm(s): maximal linkage clusters are labelled with '-ML' appended to the cluster_name; sequential branching clusters are labeled '-SB'. If more than one cluster is generated from the input sequence set, each cluster is given an extension (*cluster_name*-ML.1, *cluster_name*-ML.2, etc). Each cluster in a cluster file is represented as a nested list with sequence names separated by a match score, e.g.:
CLUSTER_NAME-ML<tab>((A 200.0 B) 150.0 C)
File format: *cluster_name*-[ML|SB][.ext]<tab>cluster_nested_list

cluster_name[-ML|-SB][.ext].pattern

The "root" AACC pattern constructed from each cluster.
File format: *cluster_name*-[ML|SB][.ext]<tab>AACC_sequence

cluster_name[-ML|-SB][.ext].pima

The pattern-induced multiple-sequence alignment of each clustered sequence set; includes the "nodal" patterns used to align the sequences (the nodal patterns have the locus name *cluster_name*-[ML|SB].ext -- extensions added to the sequence names match the extension of the nodal-pattern used to align the corresponding sequence subset, e.g. seq_1-ML.1 and seq_2-ML.1 would be aligned by nodal-pattern *cluster_name*-ML.1 .
File format: name<tab>AACC_sequence

cluster_name-[ML|SB][.ext].mase

The above file re-formatted into IntelliGenetics sequence file format; this file can then be used directly with **MASE**, MBCRR's Mutiple-Aligned Sequence Editor.

REQUIRED AUXILLARY PROGRAMS/SCRIPTS/FILES

Programs: **cluster, pima-mso, pima-pm, extract-seqs, IG-to-tbl, tbl-to-IG**
Scripts: **make-cluster.sh, make-pattern.sh**
Files: weight-pm.mat, class-pm.mat, weight2-.mat

REFERENCES

Smith, Randall F. and Smith, Temple F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. PNAS 87:118-122.

Smith, Randall F. and Temple F. Smith (1992). Pattern-Induced Multi-sequence Alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for comparitive protein modelling. Protein Engineering (In Press, vol 5., num 1).

Randall F. Smith
Human Genome Center and Dept. of Cell Biology
Baylor College of Medicine, Houston TX 77096
rsmith@bcm.tmc.edu

Temple F. Smith
Molecular Bio-Enginnering Research Center
Boston Univ., 36 Cummington St, Boston, MA 02115
tsmith@darwin.bu.edu

NAME

print-pima.sh - re-format a pima file into a user-readable/printable form

SYNOPSIS

print-pima.sh [-h] *pima-file(s)* > *output-file*

EXAMPLE

print-pima.sh SAMPLE-ML.pima > SAMPLE-ML.print-pima

DESCRIPTION

Pretty-prints a pima output file (e.g. a ∗.pima file created by 'pima.sh') using the "OUTPUT-ALIGNED" function of MASE (Faulkner and Jurka's Multiple Aligned Sequence Editor). MASE must be installed on your system to use this function. MASE is available from the MBCRR via anonymous ftp to: mbcrr.harvard.edu

OPTIONS

-h Prints the usage syntax then exits.

REFERENCE

Faulkner, Donald V., and Jerzy Jurka (1988).
Multiple aligned sequence editor (MASE).
Trends in Biochem. Sci. (TIBS) 13:321-322.

Smith, Randall F. and Temple F. Smith (1991).
Molecular Biology Computer Research Resource, Galleria Level 1
Dana-Farber Cancer Institute and School of Public Health
Harvard University
44 Binney St., Boston MA 02115 USA (617)732-3746
Internet: rsmith@mbcrr.harvard.edu
BITNET: rsmith%mbcrr@husc6.bitnet

NAMES

IG-to-tbl and **tbl-to-IG** - sequence-file-format conversion programs

SYNOPSIS

IG-to-tbl *IG_formatted_seq_file(s)* > *table_formatted_seq_file*
tbl-to-IG *table_formatted_seq_file(s)* > *IG_formatted_seq_file*

DESCRIPTION

IG-to-tbl

Converts IntelliGenetics-formatted sequence files into the table file format used by programs in the AACC Pattern Utilities Package.

tbl-to-IG

Converts table-formatted sequence files into the IG sequence format (e.g. input sequence files for **MASE**, MBCRR's Multiple-Aligned Sequence Editor, which must be in IG-format).

PARAMETERS

IG_formatted_seq_files

IG file format:

```
; at least one comment line (a ';' is required in column 1)
; any number of additional comment lines
LOCUS_NAME
VVYTDCTESGQNLCLCEGSNVCGQGNKCILGSDGEKNQ
CVTGEGTPKPQSHNDGDFEEIPEEYLQILVMAGDENS1
; comment line(s) of next sequence, etc.
```

The sequence field may contain any number of characters per line (standard format = 70 cpl); an optional '1' is allowed at end of last sequence line; no other record delimiters are required. The file 'SAMPLE.mase' included in this package is an example of a IG-formatted sequence file.

table_formatted_seq_files

File format: LOCUS_NAME<tab>SEQUENCE
where the sequence field is all on one line (No internal line breaks !!). A single <return> ends each record; no other record delimiters are required. The file 'sample-family.pep' included in this package is an example of a table-formatted sequence file.

REFERENCE

Smith, Randall F. and Temple F. Smith (1989).
Molecular Biology Computer Research Resource, LG-127
Dana-Farber Cancer Institute and School of Public Health
Harvard University
44 Binney St., Boston MA 02115 USA (617)732-3746
Internet: rsmith@mbcrr.harvard.edu
BITNET: rsmith%mbcrr@husc6.bitnet

NAME

extract-root-pat.sh - extract "root" AACC pattern from pima-files

SYNOPSIS

extract-root-pat.sh *pima-file(s) > output-file*

EXAMPLE

extract-root-pat.sh SAMPLE∗.pima > SAMPLE.pattern

DESCRIPTION

Extracts the final "root" AACC pattern from one or more ’CLUSTER_NAME.pima’ files previously created by the program **make-pattern.sh**.

PARAMETERS

pima-file

One or more multi-alignment sets created by **pima.sh** or **make-pattern.sh**.

REFERENCE

Smith, Randall F. and Temple F. Smith (1990).
Molecular Biology Computer Research Resource, Galleria Level 1
Dana-Farber Cancer Institute and School of Public Health
Harvard University
44 Binney St., Boston MA 02115 USA (617)732-3746
Internet: rsmith@mbcrr.harvard.edu
BITNET: rsmith%mbcrr@husc6.bitnet

NAME

install.sh - install and test AACC Pattern Utilities Package

SYNOPSIS

install.sh

DESCRIPTION

For those sites running under either the Unix or Ultrix operating systems, this shell script 1) automatically compiles the source code for all of the C-language programs used in the PIMA Package then 2) tests these programs by running the test program:
pima.sh TEST 25.0 sample-family.pep

The TEST.∗ output files are then compared to the sample output files provided with this package (the SAMPLE.∗ files) using the Unix *diff* program.

REQUIRED AUXILLARY PROGRAMS/SCRIPTS/FILES

Scripts:

pima.sh, make-cluster.sh, make-pattern.sh

Programs:

cluster, pima-mso, pima-pm, IG-to-tbl, tbl-to-IG, extract-seqs

Files: sample-family.pep, weight-pm.mat, class-pm.mat, weight2-.mat, SAMPLE.cluster, SAMPLE.pima

REFERENCE

Smith, Randall F. and Temple F. Smith (1991).
Molecular Biology Computer Research Resource, LG-127
Dana-Farber Cancer Institute and School of Public Health
Harvard University
44 Binney St., Boston MA 02115 USA (617)732-3746
Internet: rsmith@mbcrr.harvard.edu
BITNET: rsmith%mbcrr@husc6.bitnet