****
Installation Notes - PLSEARCH 4.0 - June 1991
(Pattern Library rel. 4.0 generated from SWISS-PROT rel. 13)
****

These files contain a library of diagnostic sequence patterns for protein families in rel. 13 of the EMBL SWISS-PROT database, the source code for **plsearch**, a dynamic programming search tool which can be used search the pattern library, and auxiliary files for use with plsearch. PLSEARCH is written in 'C' and has been implemented under both Unix and VMS operating systems. The user manual for PLSEARCH is in the file 'PLSEARCH-MANUAL.doc' (A PostScript formatted copy of the user manual is in the file 'PLSEARCH-MANUAL.ps').

Compilation considerations are described below.

****
PLSEARCH 4.0 - COMPILATION NOTES
****

FOR UNIX USERS ONLY:

1- Decide on a permanent location for the package (e.g. /usr/local/src/MBCRR-Package/plsearch) then create an environmental variable called "PLSEARCH_HOME" that points to that location. On our system, for example, the default shell is the C-shell and we have added the following line to our user's **~/.cshrc** file:
setenv PIMA_HOME /usr/local/src/MBCRR-Package/plsearch

2- cd into the plsearch/bin dir and type: install.sh
This will automatically compile the package and will run a test of plsearch using sample input and output files provided with the program.
Note: There may be minor floating-point format differences between your sample output files and ours.

3- Either move all of the files in plsearch/bin dir to /usr/local/bin or add the plsearch/bin dir to your path.

FOR NON-UNIX USERS:

1- The default pathname of the "*.pls" files can be set by changing the variable 'PLS_DIR' in the file **plsearch.h**. For example, if the plsearch package is to be kept in '/home/plsearch',
change '#define PLS_DIR ""' to '#define PLS_DIR "/home/plsearch"' in the **plsearch.h file.**

2- If your 'C' compiler lacks a Unix C-library package change
"#define UNIX TRUE" to "#define UNIX FALSE" in **plsearch.h**.

3- Compile the src code for IG-to-tbl.c and tbl-to-IG.c in plsearch/src. Compile the source code for plsearch (in the directory plsearch/src/plsearch) using the math library. For VAX/VMS systems, DON'T use the g-float math option; it is known cause floating point problems with this software. Turn off the 'C' Optimizer if you're having problems -- the optimizer may cause problems with VAX/VMS 4.6 systems.
Note: There may be minor floating-point format differences between your sample output files and ours.

MISC. COMPILATION NOTES:

1- Four parameters in the header file 'src/plsearch.h' might need to be changed to match your operating system, the memory configuration of your system, and/or your 'C' library:
MAXV sets the size of the largest pattern line (PAT_ID + PATTERN) that can be read in from the pattern file -- current default = 2500 chars. This value can be reduced, but that part of a pattern line that is over this limit will not read in.
MAXH sets the size of the largest query sequence that can be read in -- current default = 2500 chars. This value can also be reduced, but that part of a query seq. that is over this limit will not read in.
MAX_H_SIZE sets the size limit on the largest trace-back array allowable. In order to perform an alignment of 2 seqs of length n and m, an n X m trace-back array must be allocated (at 2 bytes per point). -- current default is 1500 * 1500 (ie. a 4.5 Mbyte limit). However, a much smaller (query_sequence_length X 3 X 4-byte) matrix is required to find the best match score between a pattern of any length and a given query sequence. If n X m does exceed MAX_H_SIZE, the program will calculate the best match score but the trace-back and alignment will not be performed. Therefore this parameter could be set at a much lower limit (even to 0) for machines with no virtual memory.
The calc_H() and paths() subroutines each contain a single bcopy() call. Change "#define BCOPY TRUE" to "#define BCOPY FALSE" if your C-library does not contain bcopy(). 2- The qsort_scores() routine in main-pls.c might have to be deleted it your C-library does not contain the qsort() function (see source).

OTHER NOTES:

1- The **PLSEARCH-MANUAL.ps/fR file is the** PLSEARCH-MANUAL.nroff-man file formatted for printing on PostScript laser printers (we use 'ptroff' from Adobe's Transcript Package to create nroff-formatted PostScript files: ptroff -man -t PLSEARCH-MANUAL.nroff-man > PLSEARCH-MANUAL.ps) On Macs, PostScript files can be sent directly to a PostScript laser printer using the public-domain 'SendPS' utility.

2- PLSEARCH allows patterns constructed using the Pattern Utilities toolkit (see below) to be run against all sequences in any protein sequence database if the database sequences and sequence title lines are each in a separate file in table format (LOCUS_NAME<tab>SEQUENCE and LOCUS_NAME<tab>TITLE_LINE, respectively, with NO line breaks within the SEQUENCE or TITLE lines; the PLSEARCH "patterns.pls" and "titles.pls" files are in this format). To set up this search option, 3 environmental variables need to be defined (as described above), PLSEARCH_SEQ_DB_NAME, PLSEARCH_SEQ_DB_FILE, and PLSEARCH_SEQ_DB_TITLES, e.g.:
setenv PLSEARCH_SEQ_DB_NAME "SWISS-PROT Rel. 13"
setenv PLSEARCH_SEQ_DB_FILE /seq/swissprot-13/all.seq-tbl
setenv PLSEARCH_SEQ_DB_TITLES /seq/swissprot-13/all.titles
Alternately, the corresponding variables (SEQ_DB_NAME, SEQ_DB_FILE, and SEQ_DB_TITLES) can set in the 'plsearch.h' file as described above. 3- the PATTERN_IDs (the key field) in the *.pls files are in ASCII order (e.g. 1, 10, 11 ...) rather than numeric order (1, 2, 3, ...) since we use awk(1) and join(1) to perform relational queries involving these tables (the Unix join(1) command requires that the records within the joined files be pre-sorted in increasing ASCII order).

4- PATTERN_IDs ending in a '-F' are from clusters containing sequence fragments only. PATTERN_IDs ending in an '-O' are from a small number of mis-grouped clusters that initially contained more than one sequence family; such mis-clustered families have been separated and put into separate '-O' clusters.

5- On our Sun-4, running sample.pep against the current pattern database takes about 35 min (real time).

****

Randall F. Smith
Temple F. Smith
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
****