Sequence Analysis Spring 2000 Dr. Richard Friedman
(212)305-6901 (76901) friedman@cuccfa.ccc.columbia.edu 130BB

Lesson 8 Multiple Sequence Alignment

Assignment

1. Read GCG tutorial Chapter 7.3 (Work through it on the computer if you have time and computer access).

Summary of Commands:

Note: In this document different fonts have different meanings:

Times is used to explain commands.

`Courier is used to indicate commands and command options.`

`Courier italics are used to indicate command parameters, for example, filenames.`

`Courier bold is used to indicate commands that are not displayed.`

`Courier bold italics are used to indicate computer- generated output.`

Helvetica is used to indicate menu items.

`tk`	Sets GCG so that graphics are output to the screen.
`lw`	Sets GCG so that graphics are output to a postscript file.
`lz`	Sets GCG so that graphics are output to cuccfa's laser printer.
`pileup`	Performs a progressive alignment without weights. Uses a list file as input. Outputs to an msf file.
`pretty -con -case`	Generates a *.pretty file from an msf file. In the pretty file, a consensus sequence is given. Residues that match the consensus sequence are displayed as capital letters. Residues that don't match the consensus sequence are displayed as lower case letters. To include all of the sequences in the msf file a "{*}" must be specified at the end of the file name, for example: "sh2.msf{*}"

`landscape` *`filename`*

Prints out a text file on cuccfa's laser printer in "landscape mode", i.e. the long side of the paper is sideways.

`portrait` *`filename`*

Prints out a text file on cuccfa's laser printer in "portrait mode", i.e. the short side of the paper is sideways.

`prettybox –con`

Generates a *.prettybox.ps file from an msf file. To include all of the sequences in the msf file a "{*}" must be specified at the end of the file name, for example: "sh2.msf{*}". The *.prettybox.ps file is in the Postscript lanquage, and as such cannot be displayed on an ASCII terminal. In the *.prettybox.ps file, a consensus sequence is given. Residues are shaded depending on their resemblance to the consensus sequence according to the following shading convention:

Match	Foreground	Background
Identical	White	Black
Non-similar	Black	White
Similar	Black	Dark Gray
Somewhat Similar	Black	Light Gray

`lpr` *`name.ps`*

Prints out the Prettybox Postscript file on the Cuccfa printer. Postscript files can also be downloaded to a Mac with the Macintosh utility Fetch and printed out with the Macintosh program Drop-Ps on a local Postscript Printer. Postscript files can also be downloaded to a PC with WS-FTP and printed out with the Windows program GS-view on a local Postscript Printer.

`ghostview` *`name.ps`*

Displays Prettybox output in an X-window. Must be executed from an X-terminal.

`ps2pdf` *`name.ps`*

Converts a postcript file to a pdf file. PDF files can either be downloaded to a Mac or a PC and displayed with Adobe Acrobat Reader. Display on cuccfa by typing:

`acroread` *`name.pdf`*

Displays Prettybox output, that has been translated into a PDF file. in an X-window on cuccfa. Must be typed from an X-terminal.

`clustalw`

Performs a progressive alignment weighted inversely by evolutionary distance. Can use an msf file as input., but this may lead to extra gaps. It is better to use a Fasta multiple sequence file as input. Can be set to output an msf file. Generally more accurate than pileup for proteins.

`msa`	Performs a multiple sequence alignment based on optimizing the score of all pairwise alignments simultaneously. Requires as input a file in which sequences in Fasta format are listed in succession. Outputs a file in msa format.
`grep SH2 filename.pep`	Lists every line that contains "SH2" in the sequence file "filename.pep". Useful for finding the range of an SH2 domain in a sequence.
`assemble`	Extracts a range of residues from a sequence file in GCG format and places them in another file in GCG format.
`tofasta`	Translates a sequence file in GCG format to a sequence file in Fasta format. Can be used with a list file to automatically extract sequences of different ranges.
`cat *tfa > sh2.5.limits.tfa`	Places the contents of all the *tfa files in a file called "sh2.5.limits.tfa". A way of making a multiple sequence file in Fasta format.
`msa sh2.5.limits.tfa > sh2.5.limits.msa`	Performs a multiple sequence alignment based on optimizing the score of all pairwise alignments simultaneously. Uses "sh2.5.limits.tfa" as an input file. The output is placed in "sh2.5.limits.msa".
`msa -e eps sh2.5.limits.tfa > sh2.5.limits.msa`	Does an msa alignment using a file of epsilon parameters "eps". See below for explanation.
`msagcg`	Translates an msa output file into a msf format. When it prompts you for sequence names, the sequence names that you type in must be of 5 characters or fewer.
`meme`	Generates a set of short ungapped alignments between a set of sequences.
http://www.ncbi.nlm.nih.gov/BLAST/	NCBI Blast page.
flat master-slave without identities	Formats, or reformats, Blast output page to display the alignment of the query sequence with all target sequences with E≤ .01.

Lab (or homework):

1. Work through the GCG tutorial chapter 7.3. The following testpile.list file should be given, rather than the one in the text:

..
pr:humighae2 Begin:680 End:840
pr:humighad Begin:840 End:1000
test2.gcg Begin:185
pr:ggige2c1 Begin:420 End:580

(test2.gcg is available at /usr2/seq/seqclass/lab8)

Note: In the following exercises it may be hard to compare two alignments on the screen. If this is the case print out the alignments either on cuccfa's printer or on you mac or PC after the lab, and do the comparison at that time. If you type Pretty files out on cuccfa you should use the command

`landscape` *`filename`*

2. Use Pileup to align the following proteins which contain SH2 domains. The protein names in SwissProt are:

ABL1_HUMAN
FYN_HUMAN
HCK_HUMAN
SHC_HUMAN
SRC_HUMAN
TEC_HUMAN

Hint: to get the range of the SH2 domains in each sequence fetch the six sequences and then type:

`grep FT *sw | grep SH2`

3. Use Clustalw to align the same sequences. . Put your sequences into a Fasta multiple sequence file firstand use this file as input to Clustalw. You can do this this with the `tofasta` command using the list file from the previous exercise as input. Then to run Clustalw type:

`clustalw`

The machine responds:

```
*************************************************************
 ******** CLUSTAL W (1.7) Multiple Sequence Alignments   ********
 *************************************************************

     1. Sequence Input From Disc
     2. Multiple Alignments
     3. Profile / Structure Alignments
     4. Phylogenetic trees

     S. Execute a system command
     H. HELP
     X. EXIT (leave program)
```

Your choice:

You select the menu item corresponding to sequence input:

Your choice: 1

To which the machine responds:

Sequences should all be in 1 file.

7 formats accepted:
NBRF/PIR, EMBL/SwissProt, Pearson (Fasta), GDE, Clustal, GCG/MSF, RSF.

Enter the name of the sequence file:

You use the output of your Pileup run, which is in GCG/MSF format:

Enter the name of the sequence file: *sh2.5.limits.tfa*

The program responds:

```
*************************************************************
 ******** CLUSTAL W (1.7) Multiple Sequence Alignments   ********
 *************************************************************


     1. Sequence Input From Disc
     2. Multiple Alignments
     3. Profile / Structure Alignments
     4. Phylogenetic trees

     S. Execute a system command
     H. HELP
     X. EXIT (leave program)
```

Your choice:

You select the multiple sequence alignment option

Your choice: 2

The program responds:

******* MULTIPLE ALIGNMENT MENU *********

```
     1.   Do complete multiple alignment now (Slow/Accurate)
     2.   Produce guide tree file only
     3.   Do alignment using old guide tree file

     4.   Toggle Slow/Fast pairwise alignments = SLOW

     5.   Pairwise alignment parameters
     6.   Multiple alignment parameters

     7.   Reset gaps between alignments? = OFF
     8.   Toggle screen display     = ON
     9.   Output format options

     S.   Execute a system command
     H.   HELP
     or press [RETURN] to go back to main menu


Your choice:
```

You must first set the desired output format:

```
Your choice: 9
```

The machine responds:

```
 ********* Format of Alignment Output *********


     1. Toggle CLUSTAL format output      =    ON
     2. Toggle NBRF/PIR format output     =    OFF
     3. Toggle GCG/MSF format output      =    OFF
     4. Toggle PHYLIP format output =    OFF
     5. Toggle GDE format output    =    OFF

     6. Toggle GDE output case =    LOWER
     7. Toggle CLUSTALW sequence numbers =    OFF
     8. Toggle output order    =    ALIGNED

     9. Create alignment output file(s) now?

     0. Toggle parameter output     = OFF

     H. HELP


Enter number (or [RETURN] to exit):
```

Entering 1 will turn the clustal output option off, then entering 3 will turn the GCG/MSF option on, so that you get the following configuration:

```
********* Format of Alignment Output *********
```

```
     1. Toggle CLUSTAL format output     =    OFF
     2. Toggle NBRF/PIR format output =  OFF
     3. Toggle GCG/MSF format output     =    ON
     4. Toggle PHYLIP format output =    OFF
     5. Toggle GDE format output    =    OFF

     6. Toggle GDE output case =    LOWER
     7. Toggle CLUSTALW sequence numbers =    OFF
     8. Toggle output order    =    ALIGNED

     9. Create alignment output file(s) now?

     0. Toggle parameter output     = OFF

     H. HELP
```

Enter number (or [RETURN] to exit):

Go back to main menu by typing <Return>. (WARNING, typing 9 will read out the input file, in the desired format, WITHOUT DOING AN ALIGNMENT).

Enter number (or [RETURN] to exit): [RETURN]

```
****** MULTIPLE ALIGNMENT MENU ******


     1.   Do complete multiple alignment now (Slow/Accurate)
     2.   Produce guide tree file only
     3.   Do alignment using old guide tree file

     4.   Toggle Slow/Fast pairwise alignments = SLOW

     5.   Pairwise alignment parameters
     6.   Multiple alignment parameters

     7.   Reset gaps between alignments? = OFF
     8.   Toggle screen display     = ON
     9.   Output format options

     S.   Execute a system command
     H.   HELP
     or press [RETURN] to go back to main menu
```

Your choice:

Hit the key that will produce a complete mutiple alignment.

Your choice:**1**

The program echoes,

```
WARNING: Output file name is the same as input file.


Enter a name for the GCG output file      [sh2.5.limits.msf]:
sh2.5.limits.clus.msf

Enter name for new GUIDE TREE           file   [sh2.5.limits.dnd]:
Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score:     69
Sequences (1:3) Aligned. Score:     58
Sequences (1:4) Aligned. Score:     35
Sequences (1:5) Aligned. Score:     32
Sequences (1:6) Aligned. Score:     4
Sequences (2:3) Aligned. Score:     55
Sequences (2:4) Aligned. Score:     36
Sequences (2:5) Aligned. Score:     31
Sequences (2:6) Aligned. Score:     9
Sequences (3:4) Aligned. Score:     35
Sequences (3:5) Aligned. Score:     34
Sequences (3:6) Aligned. Score:     6
Sequences (4:5) Aligned. Score:     28
Sequences (4:6) Aligned. Score:     12
Sequences (5:6) Aligned. Score:     5
Guide treefile created:   [sh2.5.pileup.dnd]
Start of Multiple Alignment
There are 5 groups
Aligning...
Group 1: Sequences:  2    Score:8188
Group 2: Sequences:  3    Score:4515
Group 3:  Delayed
Group 4:  Delayed
Group 5:  Delayed
Sequence:4Score:3738
Sequence:5Score:3637
Sequence:6Score:3613
Alignment Score 11995


Consensus length = 2623
GCG-Alignment file created     [sh2.5.clustal.msf]

PileUp


     MSF: 2623  Type: P    Check:     4876 ..
```

```
 Name: FYN_HUMAN oo  Len: 2623  Check:    226  Weight:   1.00
 Name: SRC_HUMAN oo  Len: 2623  Check:    7538 Weight:   1.00
 Name: HCK_HUMAN oo  Len: 2623  Check:    9644 Weight:   1.00
 Name: ABL1_HUMAN oo Len: 2623  Check:    6040 Weight:   1.00
 Name: TEC_HUMAN oo  Len: 2623  Check:    1643 Weight:   1.00
 Name: SHC_HUMAN oo  Len: 2623  Check:    9785 Weight:   1.00

//


FYN_HUMAN ..........  ..........  ..........  ..........  ..........
SRC_HUMAN ..........  ..........  ..........  ..........  ..........
HCK_HUMAN ..........  ..........  ..........  ..........  ..........
ABL1_HUMAN..........  ..........  ..........  ..........  ..........
TEC_HUMAN ..........  ..........  ..........  ..........  ..........

Press [RETURN] to continue or  X    to stop:
```

Keep pressing [Return] until the alignment is done. Follow the program's instructions to get out of Clustalw.

4. Compare the alignments made by Pileup and Clustalw with the programs Pretty and/or Prettybox. To execute the program pretty, type

```
pretty -con -cas
```

and answer the questions that the program asks you. To execute Prettybox type

```
prettybox –con
```

and answer the questions that the program asks you.

IMPORTANT: When specifying the sequences in an msf file, specify them as:

sequencename.msf{}*

(The "con" option calculates a consensus sequences. The "cas" option portrays amino acids that match the consensus in upper case and those that don't in lower case).

Display these files using Ghostview or Acrobat Reader. Which do you think gives a better alignment, Clustalw or Pileup? (Note, you might have to print these files on cuccfa's printer or your Mac or PC printer out after class to be able to compare them).

5. Display the pseudo-multiple sequence alignment of the human rad9 protein, human.tfa, with its statistically significant homologs using the NCBI Blast server.

6. It is also useful to align the SH2 domains using MSA.

A. Run msa by typing

msa*sh2.5.limits.tfa > sh2.5.limits.msa*

B. If in the output file for MSA there are instances in which ***epsilon*** > ***maximum epsilon***, you can improve the alignment by writing a file called eps which contains epsilon values in the order that they occur in the msa output. The epsilon value for each sequence pair should be

   max(epsilon + 1 , maximum epsilon)

   i.e., either epsilon + 1, or maximum epsilon, whichever is higher. For example, the first line of the eps file in the present example should read:

   9 6 37 7 19

   The epsilon file gives the new maximum epsilon values.

C. Rerun MSA with the new epsilon file using the following command:

   `msa -e eps sh2.5.limits.tfa > sh2.5.limits.msa`

   Repeat this process until ***maximum epsilon*** > ***epsilon*** in all instances.

D. Translate the MSA output file into GCG format by typing

   `msagcg`

   and answering the questions that the program asks you. When it prompts you for sequence names, the sequence names that you type in must be of 5 characters or fewer.

E. Use Pretty and/or Prettybox to find the consensus sequence of the alignment obtained by msa. Which alignment is best? The one obtained by Pileup, MSA, or Clustalw? KEEP THE BEST FILE FOR THE NEXT LAB. You May delete the others if you are not a SeqLab user. If you are a SeqLab user, read below.

7. Align the SH2 sequences using Meme.

SeqLab Users: Read and work through the SeqLab Tutorial Lesson 3 and 4. Repeat the exercises involving Pileup with SeqLab and print out your with a consensus sequence and the residues in the alignment shaded according to how well they match the consensus sequence. Do the same for the alignments generated with Clustalw and MSA.

Web sites:
http://dot.imgen.bcm.tmc.edu:9331/multi-align/
Baylor Multiple Sequence Alignment page

For Clustalw:
http://www2.ebi.ac.uk/clustalw/
European Bioinformatics Institute Clustalw

For Msa:
http://www.ibc.wustl.edu/ibc/msa.html
Washington University MSA

http://www.ncbi.nlm.nih.gov/BLAST/
NCBI Blast page - can be used for pseudo-multiple sequence alignments.

Documentation:
More information about Clustalw can be found in the files in
/usr2/seq/doc/clustalw

For Meme:
http://www.sdsc.edu/MEME/meme/website/

More information on MSA can be found by typing:

```
nroff -man /usr2/seq/doc/msa.1
```

This file can be printed out by typing

```
nroff -man /usr2/seq/doc/msa.1 | lpr
```

Bibliography:

1. D. Feng and R. Doolitle, Progressive alignment and phylogentic tree construction of protein sequences, Advances in Enzymology, Vol. 183, 375-387.
2. D.G. Higgins, J.D. Thompson, and T. Gibson, Using CLUSTAL for multiple sequence alignment. Advances in Enzymology, Vol. 266, 383-401.
3. References in the MSA man page cited above.

Answer key to lab 7:

6. Translate the following into motif notation: “A peptide begins with an alanine, a glycine, and an arginine, followed by between 20 and 30 amino acids of any kind, followed by a glutamine and arginine pair repeated between 15 and 92 times and terminating in a tryptophan”.
<AGRX{20,30}(Q,R){15,92}w>

7. ***AC000348) T7N9.18 [Arabidopsis thaliana] with E=0.037 .***