



## Intermediate GCG

### (Multi-sequence analysis workshop using SeqLab)

This section duplicates the handout for the Multi-sequence analysis workshop using the command line. Comparing the two sections will give you a good idea of the differences and similarities between the two interfaces. SeqLab requires an X-windows interface and **this handout assumes that you have an X-terminal or an X-terminal emulator running on a Mac or PC**. See the Logging in on an X-terminal for more details on the logging in procedure. For more information on SeqLab features, see the SeqLab guide.

For more information on the basics of SeqLab, please see **Introduction to GCG Using SeqLab (ACHS-306b)**

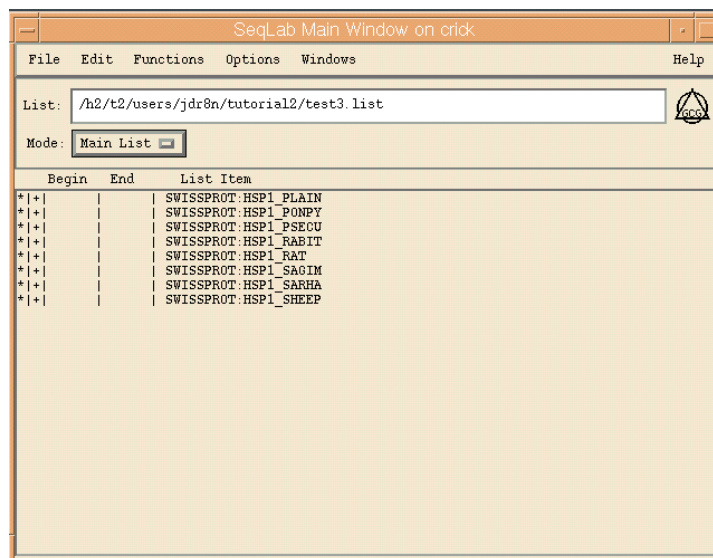
#### Prepare a directory and start GCG

To create a directory type  
`mkdir tutorial2`

Move to the directory with  
`cd tutorial2`

Start GCG with  
`gcg`

Start SeqLab with  
`seqlab`



## Creating a new list

Go back to the SeqLab window and make sure the **Mode: Main List** button is checked.

Select **File** (in the main menu bar)  
**New List...**

Type the name of the new list file in the box and click **OK**.

**Note:** Use a **.list** extension—the program will look for it by default.

## Other list operations

You can save a list at any time with the commands **File, Save List** or **File, Save List As...**

To save only selected sequences use **File, Save Selected...**

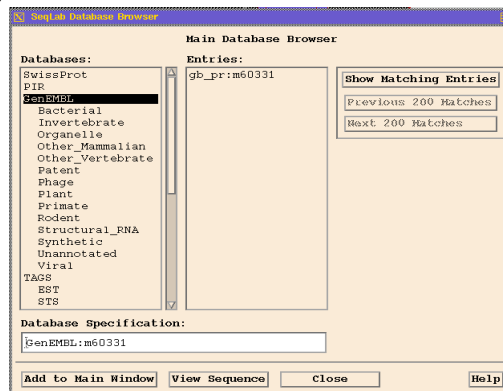
To open a previously saved list use **File, Open List...**

## Load the sequences in the main list

Go back to the SeqLab window and make sure the **Mode: Main List** button is checked.

Select **File** (in the main menu bar)  
**Add Sequences From >**  
**Databases...**

The database browser will appear.



Select **GenEmbl** in the Databases menu. Replace the \* in **GenEmbl:\*** with the accession number of the sequence for example **GenEmbl:m60331**. Press <return>.

After a while the sequence will appear under **Entries:** in the Database Browser.

**Note:** At this point you can also use the **View Sequence** button to verify that you have selected the correct sequence.

Click on **Add to Main Window** to add the sequence to the list

Repeat the previous steps to load the following sequences:

**GenEMBL:m60331**

**GenEMBL:x71334**

**GenEMBL:x71335**

**GenEMBL:x71336**

**GenEMBL:x71337**

**Close** the Main Database Browser window when you have finished.

## Multi-sequence alignment

The program PILEUP is a powerful variation of GAP. PILEUP creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments. It can also plot a tree showing the clustering relationships used to create the alignment.

### How it works

The multiple alignment procedure begins with the pairwise alignment of the two most similar sequences, producing a cluster of two aligned sequences. This cluster can then be aligned to the next most related sequence or cluster of aligned sequences. Two clusters of sequences can be aligned by a simple extension of the pairwise alignment of two individual sequences. The final alignment is achieved by a series of progressive, pairwise alignments that include increasingly dissimilar sequences and clusters, until all sequences have been included in the final pairwise alignment.

### Limitations

As shipped, PILEUP restricts each sequence in the final alignment to a maximum length of 20,000 characters. This maximum length includes the input sequence length plus the total length of all gap characters inserted into the sequence to create the final alignment. By default, each input sequence is restricted to a maximum length of 15,000. Also by default, PileUp can add a maximum of 8,000 gap characters for each sequence in the final alignment.

### Alternatives

If the sequences are too diverse for PILEUP to align, ClustalW is a good alternative. ClustalW is available at <http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html>

## A simple multi-sequence alignment

Make sure  is selected in SEQLAB.

Select all the sequences you want to work on by clicking the mouse and dragging across the names.

Select

All your sequences will appear in the editor. You can align them manually at any time.

Note: At this stage you can change the residue coloring by changing different options for Display. Features coloring will show features annotated in GenBank, such as introns and exons, in different colors.

**Select the sequences** you want to align by sweeping across the sequence names with the mouse.

Click on **Functions** in the main menu bar.

Click on **Multiple Sequence Analysis >**

Click on **PileUp...**

To run PileUp with the GCG defaults click on **GCG Defaults**

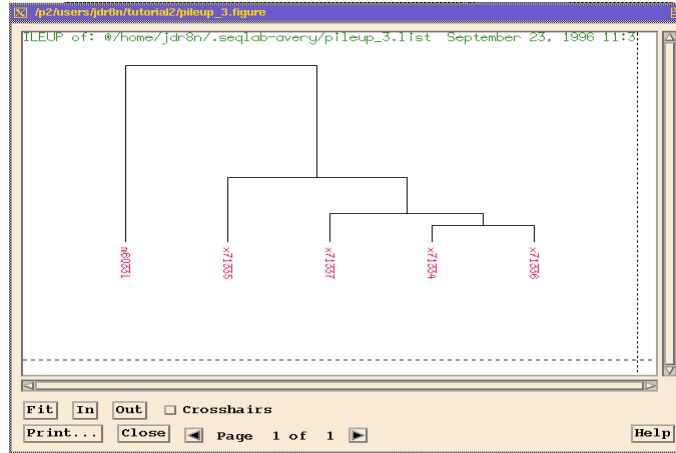
Click on **Run**

The window will disappear and after a while the results windows will appear. (This assumes you set the recommended defaults in the beginning of this document.)

**Note:** If the alignment is complex or the load on the machine high, it may take a long time for the results to appear. You can carry on with other tasks or monitor progress by clicking on **Windows** in the main menu bar and **Job Manager**. In the Job Manager window an **R** prefix indicates a program is running and an **S** indicates the job was completed successfully.

PILEUP also produces a dendrogram. The dendrogram is *not* a phylogenetic reconstruction, although the vertical branch lengths are proportional to the distances between the sequences. Its purpose is to represent the clustering

order used to create the final alignment. This order is the only information from the dendrogram used by PILEUP. In this example sequences x71336 and x71334 are first aligned and the other sequences are added progressively.



The msf file contains a text output of the alignment. Close the window if you do not need it.

```

!!NA_MULTIPLE_ALIGNMENT 1.0
PileUp of: @/home/jdr8n/.seqlab-avery/pileup_3.11st

Symbol comparison table: GenRunData:piledna.cmp  CompCheck: 6876
      GapWeight: 5
      GapLengthWeight: 1

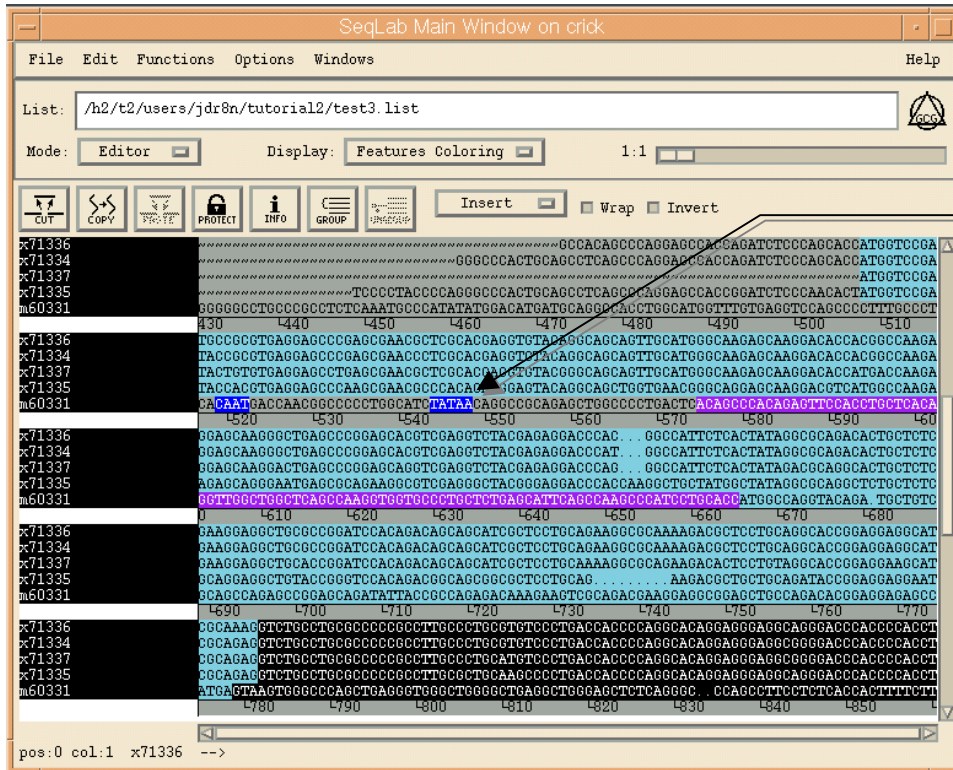
pileup_3.msf  MSF: 1309  Type: N  September 23, 1996 11:35  Check: 8484 ...
Name: x71336      Len: 1309  Check: 5517  Weight: 1.00
Name: x71334      Len: 1309  Check: 7151  Weight: 1.00
Name: x71337      Len: 1309  Check: 8890  Weight: 1.00
Name: x71335      Len: 1309  Check: 6496  Weight: 1.00
Name: m60331      Len: 1309  Check: 430   Weight: 1.00

//
      1                               50
x71336 -----
x71334 -----
x71337 -----
x71335 -----
m60331 GAGACCAAGC CTGGCCAACA TGGCGAAAGG CCATCTCTAC TAAAAATACA
  
```

The output manager contains a list of all the output files from the program.  
 Select the msf file and click on the **Add to Editor** box

A selected item appears darkened

The editor will now contain all the aligned sequences and you can proceed to refine the alignment manually.



Note the colored sequence features. Double click on the feature for more information

### Limiting the range of sequences

It is clear that the mouse sequence is much longer than all the other sequences and it will be convenient to limit the range of the sequence.

Select **Mode:**  Main List

Select **Save** or **Don't Save** if the file save window appears.

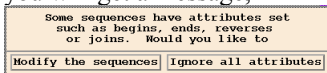
Double click on the **gb\_pr:m60331** sequence

The sequence attributes window will appear.

- Set the **Begin:** option to **470**
- Set the **End:** option to **1250**
- Click on **OK**

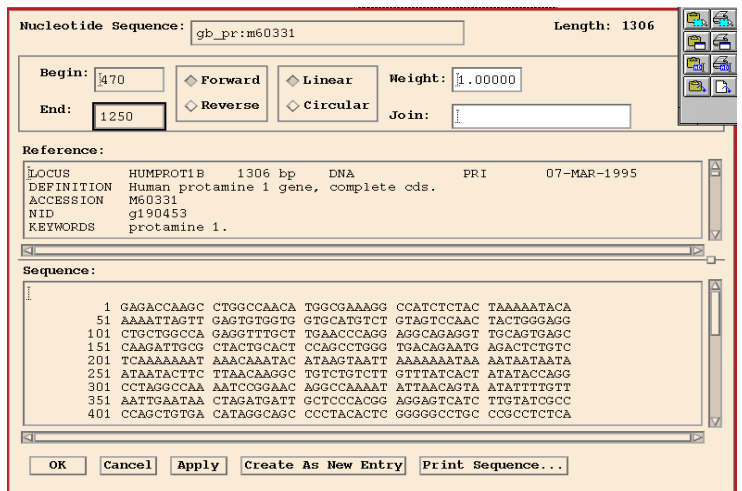
Run **PILEUP** as before.

When you move from the Main List to the Editor, you will get a message,



Select **Modify**

the **Sequences**.



## Maintaining the order of sequences

Check the “**order of sequences in alignment the same as input order**” button in the `PILEUP, options...` window.



The program functions exactly as before, i.e. the sequences are aligned pairwise starting at the closest related sequences, **only the order in which the sequences are displayed is changed.**

## Editing multiple sequences

### Aligning sequences

Within `Mode: Editor` sequences can be aligned by typing in spaces.

Note: The tilde character is interpreted as missing data, such as the unequal ends of sequences.

To insert gaps in more than one sequence at a time, first select the group of sequences, then group them by clicking on the  (**group**) button. When you insert gaps in one of the grouped sequences, the gap will appear in all of them. To ungroup sequences, select the sequences you want to ungroup and click on the  (**ungroup**) button.

### Editing sequences

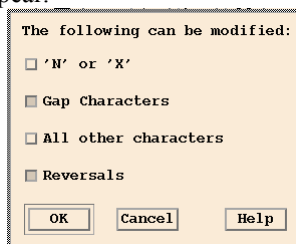
To prevent accidental modification of data, sequences are protected by default.

To remove the protection

Select the sequences you want to modify.

Click on the  button.

The Protections dialog box will appear.



Select your preferences and click **OK**.

Parts of sequences can now be cut and pasted or modified.

**Remember to restore the protection when you have finished!**

### Changing the order of sequences.

Select the sequence you want to move,  (cut) or  (copy) it and  (paste) it in the position you want.

### Creating a consensus sequence

Select the sequences you want the consensus of.

Click on **Edit** on the main menu bar.

Click on **Consensus...**

Select your options from the selection box and click **OK**.

The consensus sequence will be added as the last sequence in the editor.

## Creating list files

### Creating list files manually

You can create a list file manually as follows:

Click on **File** in the main menu bar.

Click on New **List** (If you already have a list, select Open **List**.)

Provide the name of the file you want to create and click on OK.

**Note:** The file will be created in your default directory. It is a good idea to give the same extension, e.g. **list** to all your list files.

You can now use the **File, Add Sequences from >** menu to add your own sequences from the **Sequence Files...** or database sequences from the **Databases...** menus.

When you have finished select **File, Save List** or **Save List As...**

### Programs that produce list files as output

Instead of typing in a list file, you can use the output of a number of programs as list files. If the complete database address of a sequence is included, GCG will automatically look that sequence up in the database when it is needed by the program. Therefore, the sequence does not have to be in your own directory.

Program name	Options required
Assemble.....	Click on "List file of output sequence names"
BLAST.....	Cannot create a list file, because the databases are not local.
Corrupt.....	Click on "List file of output sequence names"
FastA.....	Click on "Options..." Click on "Suppress sequence alignments in the output file"
FindPatterns..	Click on "Options..." Click on "Format output as a list file of sequence names"
FromEMBL.....	Click on "List file of output sequence names"
FromFastA....	Click on "List file of output sequence names"
FromGenBank...	Click on "List file of output sequence names"
FromIG.....	Click on "List file of output sequence names"
FromPIR.....	Click on "List file of output sequence names"
Lineup.....	Click on "List file" in the Sequence Group format: radio box.
Lookup.....	Creates a list file by default.
Motifs.....	Click on "Options..." Click on "Write output file in list file format."
Names.....	Creates a list file by default.
Pretty.....	Click on "Write the individual sequences in a PRETTY file into a separate sequence files."
ProfileSearch.	
Reformat.....	Click on "List file of output sequence names"
Sample.....	Click on "List file of output sequence names"
StringSearch..	Creates a list file by default.
TFastA.....	Click on "Options..." Click on "Suppress sequence alignments in the output file"
Translate....	Click on "List file of output sequence names"
Wordsearch	

## Exercise:

### Producing a list file with LOOKUP

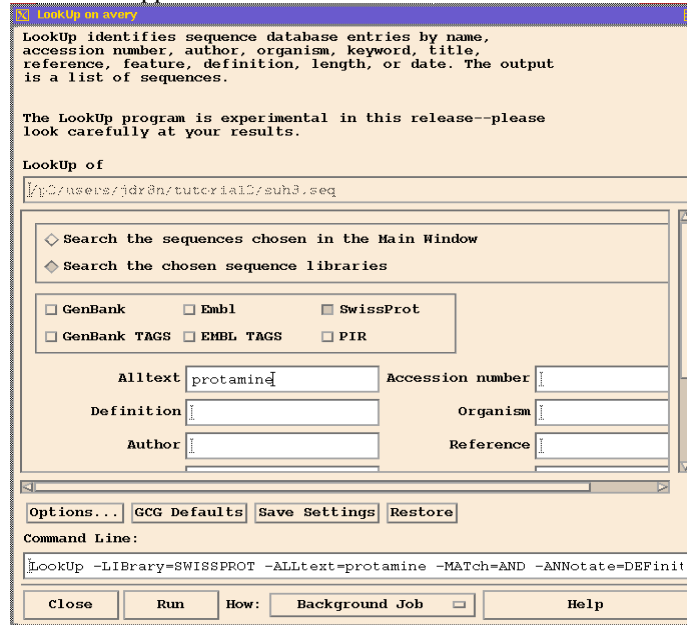
Make sure **Mode: Main List** is selected.

Click on **Functions** in the main menu bar

Click on **Database Searching >**

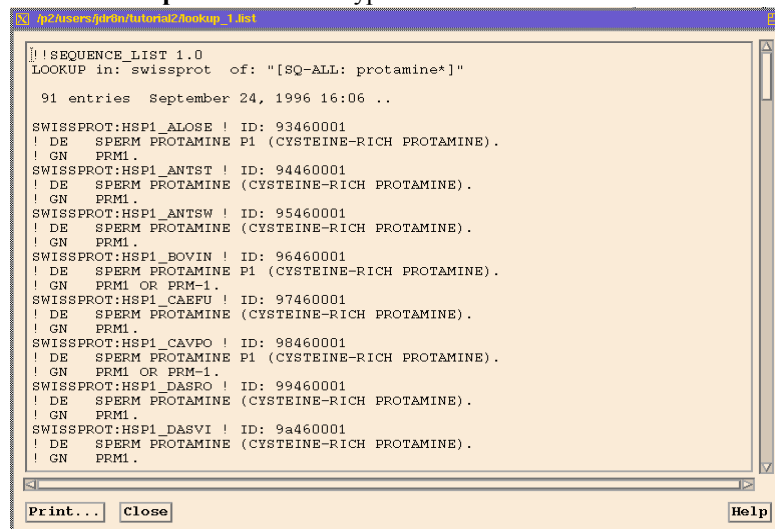
Click on **LookUp...**

The **LOOKUP** Window will appear



Select the database

Fill in the search form. In this case “**protamine**” was typed in the **All text** field.



Click on the **Close** button.

In the Job Manager window click on **Add to Main List**.

**Note:** The Job Manager window will be open if you are using the tutorial's defaults, else select **Windows, Output Manager**.

## Displaying sequence similarities

A plot of the similarities between sequences will indicate how conserved or variable the sequences are. Since functional regions are often conserved, it may indicate functional regions of the gene.

To plot a similarity profile

Select Mode: Editor □

Select all the sequences you want to include from the alignment

**Functions** (Main menu bar)  
**Multiple Sequence Analysis** >  
**Plotsimilarity...**  
**GCG Defaults**  
**Run**

The results will be displayed on your graphic screen.

**Hint:** When creating a consensus sequence with **Edit, Consensus...** (see elsewhere in this document for details) Check the “**Shade bases on similarity to consensus**” radio button. This is a useful indication of the similarity of sequences in an alignment.

## General techniques to improve alignments

### Forcing the alignment of a start codon

#### Original alignment

```
dorsa  cagtcaaaaH ATGGCAAGAT ATAGACGACA CAGCAGGAGC CGGAGT.... .AGGA GCAGATACCG
harri  cagtcaaaaH ATGGCAAGAT ATAGACGACG CAGCAGGAGC CGGAGT.... .AGGA GCAGATACCG
viver  cagtcaaaaH ATGGCAAGAT ATAGACGACG CAGCAGGAGC CGGAGT.... .AGGA GCAGATACCG
macdo  cagtcaaaaH ATGGCAAGAT ATAGACGACA CAGCAGGAGC CGGAGT.... .AGGA GCAGATACCG
stuar  cagtcaaaaH ATGGCAAGAT ATAGACGACA CAGCAGGAGC CGGAGT.... .AGGA GCAGATACCG
tapoa  cagtcaaaaH ATGGCAAGAT ATAGACGACA CAGCAGGAGC CGGAGT.... .AGGA GCAGATACCG
koala  cactgaaaaH ATGGCAAGAT ATA...GACA CAGCAGGAGC CGGAGT.... .AGGA GCAGATACCA
domestic cactgaaaaH ATGGCAAGAT ATAGAAGACG CAGCAGGAGC CGGAGT.... .AGGA GCAGATATGG
echidna cgaccacatg ttggcaccce tctgctgatt tgggaaggcca cagaaccacc taaatHATGG CAAGATTGAG
mus    cigaccacagg ttgtgtcccc tgctctgagc cagctcccgg ccaagccacg ..accHATGG CCAGATACCG
```

In this alignment the start codons are not aligned. The program does not know the difference between a start codon and a methionine. To force the alignment, the start codons are marked by the insertion of an “H” and the H is weighted in the matrix.

**Note** the “H” symbol is an ambiguous code for A, C, or T. If you need to use that symbol in one of the sequences, use one of the other ambiguity symbols.

#### Forced alignment

```
dorsa  cagtca.. aaahHATGGGCA AGATATAGAC GACACAGCAG GAGCCGGAGT AGGAGCAGAT
harri  cagtca.. aaahHATGGGCA AGATATAGAC GACGACAGCAG GAGCCGGAGT AGGAGCAGAT
viver  cagtca.. aaahHATGGGCA AGATATAGAC GACGACAGCAG GAGCCGGAGT AGGAGCAGAT
macdo  cagtca.. aaahHATGGGCA AGATATAGAC GACACAGCAG GAGCCGGAGT AGGAGCAGAT
stuar  cagtca.. aaahHATGGGCA AGATATAGAC GACACAGCAG GAGCCGGAGT AGGAGCAGAT
tapoa  cagtca.. aaahHATGGGCA AGATATAGAC GACACAGCAG GAGCCGGAGT AGGAGCAGAT
koala  cactga.. aaahHATGGGCA AGATATA... GACACAGCAG GAGCCGGAGT AGGAGCAGAT
domestic cactga.. aaahHATGGGCA AGATATAGAA GACGACAGCAG GAGCCGGAGT AGGAGCAGAT
echidna cactca.. aatHHATGGGCA AGATTCA... GGCCAGCCG GAGCCGCAGC CGCAGCCTGT
mus    aagccacg accHHATGGGCC AGATAC...C GATGCTGCCG CAGCAAAGC AGGAGCAGAT
```

**Hint:** In the above example uppercase is used to mark the coding region. The program does not differentiate between upper- and lowercase. To change the case of selected areas select Mode: Editor □ and select the regions where you want to change the case. Now select **Edit, Change Case**.

## Forcing the alignment of an intron junction

### Original alignment

```
dorsa tcta tttttgttta aacttcoccta teatccctcc ctgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
harri tcta tttttgttta aacttcocctg teatccctcc ctgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
viver tcta tttttgtttt. aacttcocctg teatccctcc ctgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
macdo tcta tttttgttta aacttcocctg teatccctcc ctgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
stuar tcta tttttgttta aacttcocctg teatccctcc ccgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
tapoa tcta tttttgttta aacttcocctg teatccctcc ccgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
koala tcta tttttaaatt tagcctcttt teattctct... ..ctcagHG GTA...TCGC AGGAGATATT C..... TCGCAGG... ..AGAA
domestic .... .tctcttta aaccttc..a ttattctctt tttctcagHG GTA...CCAC AGGAGATCTC CTCATCGTCG TCGTAGGAGA AGAAGAA
echidna cccc agHGTAGACG CAGCATGAGA TCCTCTCGCA GAAGAAGAAG GAGGAGAAGA AACTGALgag ccactctcca tgctctgct cgagaac
mus tgag aattttacca gaactcaaga gcattctgcc acattctgaa aaatgccacc gtcgatgaa aaa.....ca ggagcctgct aagHGAA
```

In this alignment the program has trouble aligning the position of the intron (marked with “H”) This can be corrected by weighing the “H” symbol to force the alignment.

### Forced alignment

```
dorsa tcta tttttgttta aacttcoccta teatccctcc ctgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
harri tcta tttttgttta aacttcocctg teatccctcc ctgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
viver tcta tttttgtttt. aacttcocctg teatccctcc ctgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
macdo tcta tttttgttta aacttcocctg teatccctcc ctgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
stuar tcta tttttgttta aacttcocctg teatccctcc ccgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
tapoa tcta tttttgttta aacttcocctg teatccctcc ccgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
koala tcta tttttaaatt tagcctcttt teattctct... ..ctcagHG GTA...TCGC AGGAGAT... ..ATTCTCG CAGG... ..AGAAGAT
domestic .... .atctcttta aaccttc..a ttattctctt tttctcagHG GTA...CCAC AGGAGATCTC CTCATCGTCG TCGTAGGAGA AGAAGAA
echidna cctt cacatctgt. .... .tctctctctc ...cccagHG TAGACGCAGC ATGAGATCCT CTCGAGAAG AAGAAGGAGG AGAAGAA
mus acca cttttct... .. .ttttctctc cttctcagHG ATGCTGCCGT CGCCGCCGCT CATAAC... ..CATAAGGTGT AAAAAAT
```

## Fetching and editing the DNA comparison table

You can fetch the DNA comparison table to your default directory. It is also good practice to rename it, so you can use the default table when needed. To fetch and redirect the file type  
`fetch pileupdna.cmp -out=fixdna.cmp`

You can now modify the file with a text processor. For this exercise change the value for a H:H match to 100.

Type `jove fixdna.cmp` and edit the text.

To save the text press `<Ctrl>x` and `s`

To exit Jove press `<Ctrl>x` and `<Ctrl>c`

**Hint:** A copy of the file called `pileupdnafix.cmp` is provided and may be chosen in the matrix browser. From the command line, you may also copy the file to your account with the **FETCH**

Add your own comments

```
!!DNA_SCORING_MATRIX_RECT 1.0
This matrix forces alignments of matching H and matching R characters
Modified scoring matrix used by PILEUP for the comparison of nucleic acid
sequences. PILEUP uses the method of Needleman/Wunsch/Sellers to make
alignments. This table scores a match for any overlap between any IUB
nucleic acid ambiguity symbols EXCEPT X/N.
February 20, 1996 14:34 ..

{
GAP_CREATE 5
GAP_EXTEND 1
}

      A   B   C   D   G   H   K   M   N   R   S   T   U   V   W   X   Y
A   1   0   0   1   0   0   0   1   1   0   0   0   0   1   1   1   0
B   0   1   1   1   1   0   1   1   1   0   1   1   1   1   1   1   1
C   0   1   1   0   0   0   0   1   1   0   1   0   0   1   0   1   1
D   1   1   1   0   1   0   1   1   1   0   1   1   1   1   1   1   1
G   0   1   0   1   1   0   1   0   1   0   1   0   0   1   0   1   0
H   0   0   0   0   0   99   0   0   0   0   0   0   0   0   0   0   0
K   0   1   0   1   1   0   1   0   1   0   1   1   1   1   1   1   1
M   1   1   1   1   1   0   0   1   1   0   1   0   0   1   1   1   1
N   1   1   1   1   1   0   1   1   1   0   1   1   1   1   1   1   1
R   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0
S   0   1   1   1   1   0   1   1   1   0   1   0   0   1   0   1   1
T   0   1   0   1   0   0   1   0   1   0   0   1   1   0   1   1   1
U   0   1   0   1   0   0   1   0   1   0   0   1   1   0   1   1   1
V   1   1   1   1   1   0   1   1   1   0   1   0   0   1   1   1   1
W   1   1   0   1   0   0   1   1   1   0   0   1   1   1   1   1   1
X   1   1   1   1   1   0   1   1   1   0   1   1   1   1   1   1   1
Y   0   1   1   1   1   0   1   1   1   0   1   1   1   1   1   1   1
```

To weight the matrix increase the value from 1 to 99

Also, set all other matching values to 0 (shown in bold)

## Practice session:

You can easily insert “H” characters in the SeqLab editor.

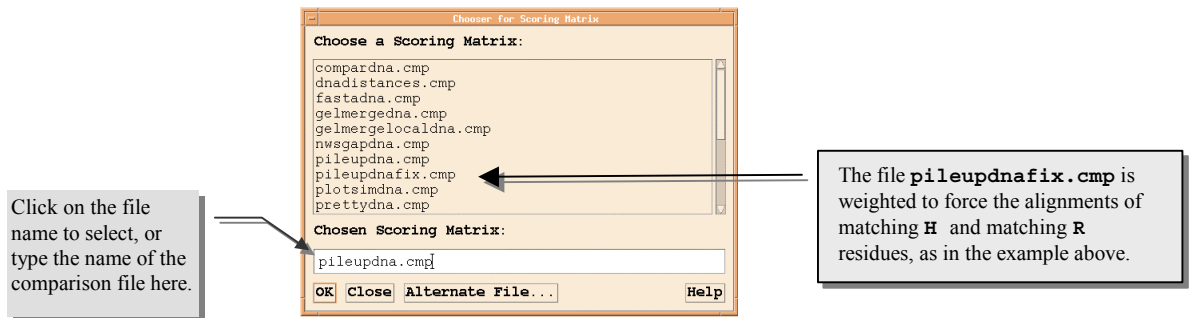
To insert characters in to sequences, you first have to remove the sequence protection, see “Editing sequences” elsewhere in this handout. (Make sure the “All other characters” button is checked.)

To inserts “H” or any other characters in the same position in more than one sequence at a time, first select the sequences you want to modify. Group the sequences. Click on the position in the sequence where you want to insert the character and type it in. The same character will appear in all the sequences. See “Editing sequences” elsewhere in this handout for more information on grouping and ungrouping sequences.

When you run PileUp, click on the Options... button

Check the  Scoring Matrix...  s/jdr8n/tutorial12/pileupdna.cmp radio button.

Click on Scoring Matrix... and type in the name of your own comparison file.



Click on OK

Run PILEUP as before

## Weighting functional residues

In this example we know that serines are involved in the phosphorylation of the protamines during spermiogenesis. Since this is a conserved, functional domain it may help the overall alignment to force the alignment of the serines. Whether this is necessary, or what weight should be assigned to them is a judgment call and can only be determined by experimentation. This type of functional alignment will vary greatly from one gene to the next

### Normal alignment

```

plat  MARF.RRSRS RSRGLY.RRR .RSRRGGR QTRSRKLSRS RRRGRSRRRK
hump1  MARY.RCCRS QRSRYRQR QRSR..... .RRRRRS CQTRRRAMRW
mus    MARY.RCCRS KSRRCRRR RRCR..... .RRRRRC CRRRR. RW
opos   MARYRRRSRS RSRRYGRRR RRSRSR... .RRRSRR RRRRGRRGR
domestic MARYRRRSRS RSRRYGRRR RRSRSR... .RRRSRR RRRRGRRGR
caen   MARY.RHSRS RSRRYRRR RRRRSYRSR RRRYR.RSR. RRRRG.RRR
austr  MVYRRHSRS RSRRYRRR RR..RLNR RRYRRSRRG RRRRRGSR
bandi  MARY.RHSRS RSRRF.RRR GRRSRVRGR DARQGRSRR RRRGKGRAHS
redkan MARY.RHSRS RSRRY.RRR RRRSYRSQ RRYRGRRR. RRSRG.RRR
tamar  MARY.RHSRS RSRRY.RRR RRRSYRSR RRSRGRRR. RSRRGRRR
koala  MARY.RHSRS RSRRY.QRR RRRSYRSQ RRYRRRGS RRRRRGRR
btpossum MARY.RHSRS RSRRYRRR RRRSYRSR RRYR.RSR. RRRRGRRR
notory MARY.RHSRS RSRRY.RRR RRRSYRSQ RRYRRHRS GRRRRGRR
dorsa  MARYRRHSRS RSRRY.RRR RRRSRGR. RRTYRSRR. HSRRRGRR
macdo  MARYRRHSRS RSRRY.RRR RRRSRHNR RTYRSRR. HSRRRGRR
crass  MARYRRHSRS RSRRY.RRR RRRSRHNR RTYRSRR. HSRRRGRR
pingra MARSRHSRS RSRRNQCQR RRRRT...YN RRTMREKPR HSRRRVRR
stuar  MARYRRHSRS RSRRYRRR RRRSRHNR RTYRSRR. HSRRRGRR

```

## Weighted alignment

```

plat  MARF.RRSRS RSRSLY.RRR ..RRSR...R .....GGRQT RSRKLSRSRR
hump1  MARY.RCCRS QSRORY..YR QRQSR.... .....RRRR RS..CQTRRR
mus    MARY.RCCRS KSRRC..RR RRRRCR.... .....RRRR RC..CRRRR
opos   MARYRRRSRS RSRRYGRRR RRSRSR.... .....RR RSRR.RRRRR
domestic MARYRRRSRS RSRRYGRRR RRSRSR.... .....RR RSRR.RRRRR
caen   MARY.RHSRS RSRRYRRR RRRRSRYRSR RRRY.....R RSRR.R.RRR
austr  MVRYYRHSRS RSRRYRRR RR...RLNR RRRY.....R RSRGRRRR
bandi  MASY.RNSRS RSRRF.RRR RGRRSRVGR . ....DARQG RS...SRRR
redkan MARY.RHSRS RSRRY.RRR RRRRSRYRSQ RRRYGRRRR RS.....RR
tamar  MARY.RHSRS RSRRY.RRR RRRRSRYRSR RRRSRGRRR RS.....RR
koala  MARY.RHSRS RSRRY.QRR RRRRSRYRSQ RRRY..RRR GSRR.R.RRR
btpossum MARY.RHSRS RSRRYRRR RRRRSRYRSR RRRY.....R RSRR.R.RRR
notory MARY.RHSRS RSRRY.RRR RRRRSRYRSQ RRRY..RRHR RSGR.R.RRR
dorsa  MARYRRHSRS RSRRY.RRR RRRRSRGR.R RRTY.....R RSRR.HSRR
macdo  MARYRRHSRS RSRRY.RRR RRRRSRHRNR RRTY.....R RSRR.HSRR
craas  MARYRRHSRS RSRRY.RRR RRRRSRHHNR RRTY.....R RSRR.HSRR
pingra MARSRRHSRS RSRRNQCQR RRRRT.Y..N RRRTMREKPR HSRRRRVRRR
stuar  MARYRRHSRS RSRRYRRR RRRRSRHHNR RRTY.....R RSRR.HSRR

```

## Amino acid comparison table

To modify the amino acid comparison table, first copy the table to your working directory. To do this Open a terminal window and start GCG.

Type `fetch blosum62.cmp -out=fixpep.cmp`

The file called "fixpep.cmp" will be copied to your directory. You can modify this file with you favorite text editor.

To edit the file with a text editor such as Jove.

**Hint:** To get more information on data files for any particular command, look under the "local data files" option.

```

!!AA_SCORING_MATRIX_RECT 1.0
BLOSUM62 amino acid substitution matrix.
Reference: Henikoff, S. and Henikoff, J. G. (1992). Amino acid
substitution matrices from protein blocks. Proc. Natl. Acad.
Sci. USA 89: 10915-10919.
February 20, 1996 14:33 ..
{
GAP_CREATE 12
GAP_EXTEND 4
}

```

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
A	4	-2	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-1	-2	-1
B	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
C	0	-3	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-1	-2	-4
D	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
E	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5
F	-2	-3	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	-1	3	-3
G	0	-1	-3	1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-1	-3	-2
H	-2	-1	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	-1	-1	-1
I	-1	-3	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1	-1	-1
K	-1	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-1	-1	-2
L	-1	-4	-1	-4	-3	-3	-2	-1	-3	2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1	-1	-1
M	-1	-3	-1	-3	-2	-3	-2	-1	-3	2	2	5	-2	-2	0	-1	-1	-1	1	-1	-1	-1	-1
N	-2	1	-3	1	0	-3	-2	-2	-1	-1	-3	-2	6	-2	0	0	1	0	-3	-4	-1	-2	0
P	-1	-1	-3	-1	1	-3	-2	-2	-1	-1	-3	-2	-2	7	-1	-2	-1	-7	-2	-4	-1	-3	-1
Q	-1	0	-3	0	2	-2	0	0	-1	5	-2	0	0	-1	5	1	0	1	-2	-2	-1	-1	2
R	-1	-2	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-1	-2	0
S	1	0	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	9	1	-2	-3	-1	-2	0
T	0	-1	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-1	-2	-1
V	0	-3	-1	-3	-2	-1	-3	-3	-3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1	-1	-2
W	-3	-4	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	-1	2	-3
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Y	-2	-3	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	-1	7	-2
Z	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5

To change the weighting of the cysteine residue, change the value here.

To change the weighting of the serine, change the value here.

## Practice session:

The following set of sequences can be used to explore the effect of weighting amino acid residues.

Go to Functions, Database Searching, Lookup...

Search SwissProt All text for the word "protamine" and definition for the word "p1".

For details, see "Producing a list file with LookUp earlier in this session.

You can now run PileUp as follows:

When you run PileUp, click on the Options... button

Check the  Scoring Matrix...  s/jdr8n/tutorial12/pileupdna.cmp radio button.

Click on Scoring Matrix... and type in the name of your own comparison file in the window.

Click on OK

Run PILEUP as before

## Weighting the ends of an alignment

In the following example the stop signals are not aligned. Since this is a functional signal, we may want to align the ends of the sequence. To do this go to the PileUp, options... window and check the "penalize end gaps like other gaps" button..

Normal alignment		End-weighted alignment	
plat	RKGWRRSRR. .SSRRSRRRN *....	plat	RKGWRRSRR. . .SSRRSRR RN....*
humpl	...W.CCRPR ...YRPRCRR H....	humpl	...W.CCRPR ..Y.....R PRCRRH.
mus	...W.CCRRR .RSYTIRCKK Y*...	mus	...W.CCRRR RSY.....T IRCKKY*
opos	GRGW.YHRR. . .SPHRRRRR RRRRA*	opos	GRGW.YHRR. . . .SPHRRR RRRRRRA*
domestic	GRGW.YHRR. . .SPHRRRRR RRR*	domestic	GRGW.YHRR. . . .SPHRRR RRRRR. *
caen	RRGW.YSRRR Y.S.RRRRRR Y*...	caen	RRGW.YSRRR ..Y.S....R RRRRRY*
austr	RRGW.YSRRR YQSRRRRRRR Y*...	austr	RRGW.YSRRR ..YQS...RR RRRRRY*
bandi	KKGWRRS... .GSRRRKRNN ENK*	bandi	KKGWRRS... . .GSRRRKR NNENK. *
redkan	RRGW.YSRRR Y.S..RRRRR Y*...	redkan	RRGW.YSRRR ..Y.S....R RRRRRY*
tamar	RRGW.YSRRR Y.S.RRRRRR Y*...	tamar	RRGW.YSRRR ..Y.S....R RRRRRY*
koala	RRGW.Y.RRR Y.S.R..RRR Y*...	koala	RRGW.Y.RRR ..Y.S....R ..RRRY*
btpossum	RRGW.YSRRR Y.S.RRGRRR Y*...	btpossum	RRGW.YSRRR ..Y.S....R RGRRRY*
notory	RRGW.Y.RRR YHS.H..RRR Y*...	notory	RRGW.Y.RRR ..YHS...H ..RRRY*
dorsa	RRGW.YSRRR Y.S.RRGRRR Y*...	dorsa	RRGW.YSRRR ..Y.S....R RGRRRY*
macdo	RRGW.YSRRR Y.S.RRGRRR Y*...	macdo	RRGW.YSRRR ..Y.S....R RGRRRY*
crass	RRGW.YSRRR Y.S.RRGRRR Y*...	crass	RRGW.YSRRR ..Y.S....R RGRRRY*
pingra	..GW.CSCRR C.SRRRRRRC *....	pingra	..GW.CSCRR ..CS....R RRRRRRC*
stuar	RRGW.YSRRR Y.S.RRGRRR Y*...	stuar	RRGW.YSRRR ..Y.S....R RGRRRY*

## Manual adjustments

Even with the best programs available today the results you obtain may not be perfect. You can only be confident of any alignment once you have carefully looked at every base or amino acids, especially where there are gaps in the alignment. The alignment of amino acids and the corresponding coding sequences must be compared and reconciled.

### Alignment using protein sequences only

Notoryc	R	S	Q	R	R	R	Y	R	R	H	R	R	S	G	R
P.ingra	Y	N	-	R	R	T	M	R	E	K	P	R	-	H	S
P.dorsa	G	R	R	R	R	T	Y	R	R	S	R	R	-	H	S
A.swain	G	R	-	R	R	T	Y	R	R	S	R	R	-	H	S
D.viver	G	R	-	R	R	T	Y	R	R	S	R	R	-	H	S
P.apica	R	N	-	-	R	T	Y	R	R	S	R	R	-	H	S
D.rosam	R	N	-	-	R	T	Y	R	-	S	R	R	-	H	S
S.harri	G	R	R	R	R	T	Y	R	-	S	R	R	-	H	S
D.hallu	G	R	R	R	R	T	Y	R	R	S	R	R	-	H	S

There are clearly many different ways to align some of the arginine residues. Their true positions only become obvious when the DNA coding sequence is taken in to account.

### Alignment using both protein and DNA sequences

```

Notoryc  CGT AGT CAG AGG AGG AGA TAC AGG AGA CAC CGG AGA AGC GGG AGG
          R  S  Q  R  R  R  Y  R  R  H  R  R  S  G  R
P.ingra  TAT AAT --- AGG AGG ACC ATG CGA GAG AAG CCG AGA --- CAT TCG
          Y  N  -  R  R  T  M  R  E  K  P  R      H  S
P.dorsa  GGA AGA CGA AGG AGG ACA TAC AGG AGA AGC CGG AGA --- CAT TCG
          G  R  R  R  R  T  Y  R  R  S  R  R      H  S
A.swain  GGA --- CGA AGG AGG ACA TAC AGG AGA AGC CGG AGA --- CAT TCG
          G  -  R  R  R  T  Y  R  R  S  R  R      H  S
D.viver  GGA --- CGA AGG AGG ACA TAC AGG AGA AGC CGG AGA --- CAT TCG
          G  -  R  R  R  T  Y  R  R  S  R  R      H  S
P.apica  CGT AAT CGA --- --- ACA TAC AGG AGA AGC CGG AGA --- CAT TCG
          R  N  R  -  -  T  Y  R  R  S  R  R      H  S
D.rosam  CGT AAT CGA --- --- ACA TAC AGG --- AGC CGG AGA --- CAT TCG
          R  N  R  -  -  T  Y  R  -  S  R  R      H  S
S.harri  GGA AGA CGA AGG AGG ACA TAC --- AGA AGC CGG AGA --- CAT TCG
          G  R  R  R  R  T  Y  -  R  S  R  R      H  S
D.hallu  GGA AGA CGA AGG AGG ACA TAC AGG AGA AGC CGG AGA --- CAT TCG
          G  R  R  R  R  T  Y  R  R  S  R  R      H  S

```

The resulting alignment is clearly improved and every arginine residue now has a unique position.

**Hint:** The Edit, Translate... function in the SEQLAB editor makes it easy to do this type of alignment. Make sure you have the "Align Translation" box checked.

### Hopeless cases

```

                ***   ***   *   *   *           Indicative sites
Macropusg  -----caaa--
Phascolod  tcttagattattggggagggga----gtgcaaat
Murexialo  tcttagatttgggggtgggg-aggagtgtgcaaat
Pseudantm  tcttagattattggggaagggggcg-gtgcaaat
Sminthopc  tcttaaattat--ggtggggggcggtgtgaaaaat
Planigali  tcttagatctagtttatgggaggagggtgcaagt
Antechist  tcttagatt-tgggggtggggaagagtgtgcaaat
Antechisw  tcttag--tattgggggtggggaggagtgtgcaaat
Phascogat  tcttagatttggggagggggaggaatgtgcaaat
Dasyurusv  tcttagattattggggaaggggggc--gtgcaaat
Dasyurush  tcttagattattggggaagggggcg-gtgcaaat
Sarcophil  tcttagattattgggga-ggg-gg--gtgcaaat
Dasykalut  tcttagattattgggga-ggg-g---gtgcaaat
Paranteca  tcttagattattgggga-ggg-gg--gtgcaaat

```

It may be impossible to align some sequences unambiguously. In the above example there are many, equally valid, alignments possible. This is part of an intron sequence, so there is no protein sequence to assist with the alignment. To make matters worse, this alignment includes a large number of indicative sites that will strongly influence a parsimony tree. It is best to exclude such regions from a phylogenetic analysis.

### Aligning very dissimilar sequences:

Pileup generally produce the best results with sequences that are fairly similar. In some cases, such as a rapidly diverging gene or in non-coding regions, it may be impossible to find an optimal alignment with Pileup. See "Limitations" earlier in this chapter. In such cases try ClustalW; it is available from the Search Launcher (<http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html>) or directly at (<http://www2.ebi.ac.uk/clustalw/>).

## Aligning small domains

When it is impossible to align a set of sequences, all is not lost. It is still possible to ask the question: are there any small domains that can be aligned? This is particularly useful in non-coding regions, where you may look for transcription factor or other binding sites. **MEME** finds conserved motifs in a group of unaligned sequences. **MEME** saves these motifs as a set of profiles. You can search a database of sequences with these profiles using the **MOTIFSEARCH** program. The following is the results of a **MEME** alignment:

```
*****
MOTIF 1 width = 27 sites = 4.0
*****
Simplified A 111115313513111171913713113
motif letter- C :::62::64::2:::26:44:8:::46
probability G 8::8:2:2:::2::88:2:2:2::8::
matrix T 199131771195791111133117151

bits 2.2
2.0
1.7
1.5
Information 1.3
content 1.1 **** * ** * *
(21.6 bits) 0.9 **** * ** * *
0.7 ***** * * ** ** * *
0.4 ***** *****
0.2 *****
0.0 -----

Multilevel GTTGCATTCAATTTGGACACCACTGTC
consensus TGAGAC AC CG TAG A CA
sequence C G GT

-----
Motif 1 in BLOCKS format
-----
BL MOTIF 1 width=27 seqs=4
/home/jdr8n/.seqlab-crick/input_8.rsrf{GOLD_GEN} ( 763) GTTGTCATAATTTGGCCACCACTGCA 1
/home/jdr8n/.seqlab-crick/input_8.rsrf{BOVINE_INTRON3} ( 1125) GTTGCATTCCCTATTGGACATCACTGTC 1
/home/jdr8n/.seqlab-crick/input_8.rsrf{HUMAN_INTRON3} ( 1189) GTTGCATTCCCTGTTGGACAGTGCTGTC 1
/home/jdr8n/.seqlab-crick/input_8.rsrf{CHICKEN_INTRON} ( 128) GTTGCCTGCATTCTGGAGACAACAGCC 1
```

## Displaying aligned sequences

The program pretty can help to display sequences, change the format, show differences, and calculate a consensus sequence.

ALSCRIPT is a program that will automatically box and color sequence alignments. The program is flexible and the resulting output is very impressive. To convert sequences from msf format to the block format required by ALSCRIPT, use the MSF2BLC program. ALSCRIPT is *very* difficult to use and not at all user friendly. *We can only offer limited support for this program, so be prepared to spend a significant amount of time on it.* More information on Alscript, including the instruction manual, is available at the Uva molecular biology web site <http://www.med.virginia.edu/achs/molbio/software/software.html>

## Publishing aligned sequences

The sequence alignments can have a dramatic effect of the results of subsequent phylogenetic analysis. It is therefore very important that any deviations from the standard program usage be carefully considered and documented, otherwise the results will not be reproducible by other scientists.

---

## Exercises

1. Use fasta to find the 10 protein sequences in the PIR1 database that are most similar to L35341 and align these sequences.
  2. Repeat the alignment from the previous exercise, but weigh the arginine-arginine matches to 9.5.
-

## APPENDIX I

The alignment tools we have discussed are not able to determine more than one alternative position of alignment, such as in repeats. Dotmatrix analysis provides a useful tool to look for repeats, or internal repeats, or alignments that are separated by large gaps or inserts.

### Dotmatrix analysis

Programs such as GAP or BESTFIT are unable to find repeats or unusual structures within a sequence. A sensitive way to find such features is with the dotmatrix analysis

Open a terminal and start GCG.

To fetch an example sequence with repeats type

```
fetch winchester.seq
```

Optional: Close the window by typing  
exit

Load the sequences into the main list.

Make sure **Mode: Main List** is selected in SeqLab.

Click on **File** in the main menu

Click on **Add Sequences From >**

Click on **Sequence Files...**

Select `winchester.seq` from the files menu

Click on **Add** at the bottom of the box.

**Important:** We need two copies so select `winchester.seq` again and click on **Add** at the bottom of the box.

Compare the sequences

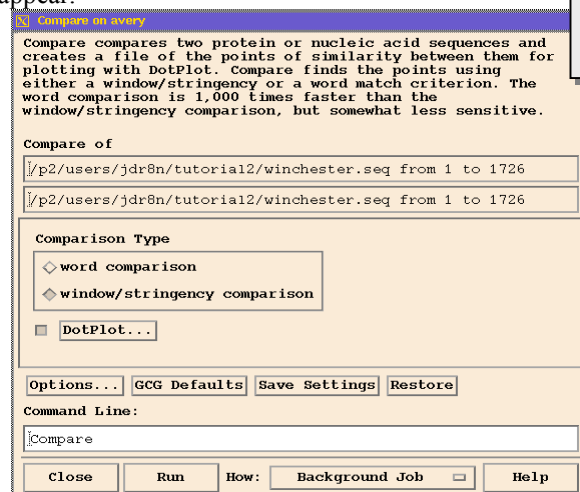
Select both copies of `winchester.seq` by clicking and sweeping across them with the mouse.

Click on **Functions** in the main menu bar.

Click on **Comparison**

Click on **Compare...**

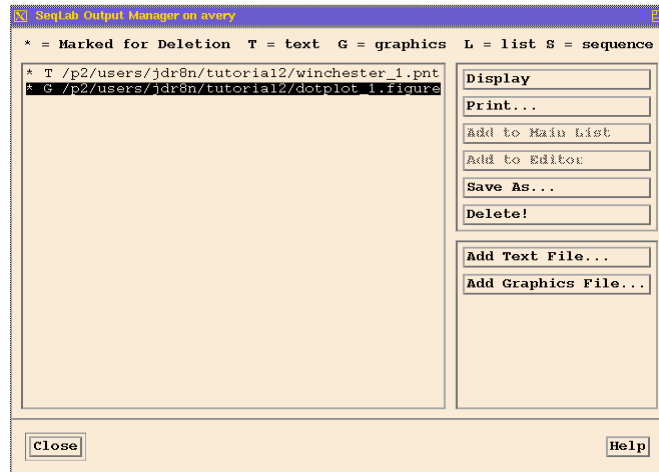
The Compare dialog box will appear.



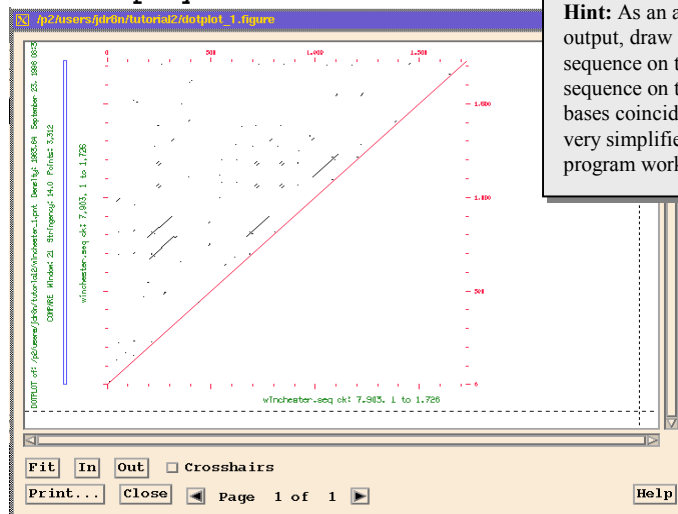
**Hint:** The programs **Lalign** and **Palign** will generally produce output that is easier to interpret. Lalign and Palign are available on the FASTA server at <http://fasta.bioch.virginia.edu/>

We can accept all the defaults by clicking on **GCG Defaults**

Click on **Run**



The following dialog box will appear. If you selected the defaults suggested in the tutorial the dotplot will appear automatically, otherwise click on **Display**.



**Hint:** As an aid to understanding the output, draw a square and write a short sequence on the x axis and another sequence on the Y axis. Where the bases coincide, mark an x. This is a very simplified model of how the program works

You will see a number of repeats.

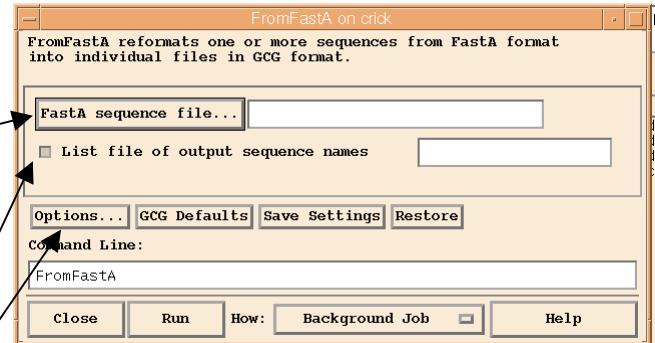
An Xwindows display of a dot-matrix analysis.

To search for reverse complemented repeats, run the program again, but first double click in one of the sequences and check the reverse button to reverse the sequence. What do you expect?

## APPENDIX II

Import a multiple sequence alignment in FASTA format into SeqLab:

1. Open SeqLab and go to the **Main List** window.
2. Choose **Functions**  
**Importing/Exporting**  
**FromFastA...**
3. Go to the **Fasta Sequence File...** box and type the name of the FastA sequence file or use the file browser.
4. Make sure the **List file of output sequence** button is checked and type in the name of the list file.
5. Click on **Options...** and check **nucleic acid** or **protein** to match the sequence file. Close the dialog box.
6. Click on **run**.
7. When the conversion is complete the list of files will be displayed (You will have to click on **windows** and **Output manager** if your automatic display option is not set.) Close the display box.
8. In the **Output Manager** make sure the list file is selected and click on **Add to main list**. Close the **Output Manager**.
9. In Seqlab (Main list) make sure the file is selected. Change the window toggle from **Main list** to **Editor**.
10. Your sequences will now be read and displayed in the sequence editor.

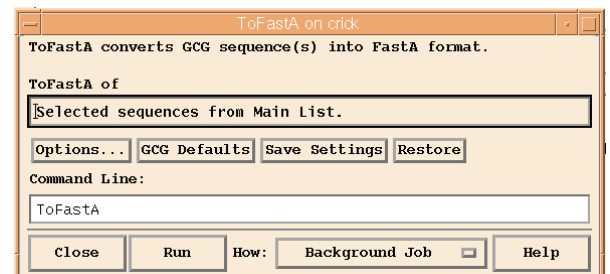


**Note:** The actual sequences will be stored as individual sequence file in your SeqLab default directory.

---

Export a multiple sequence alignment in FastA format from Seqlab:

1. Make sure you are in the SeqLab Editor window.
2. Select all the sequences you want to export.
3. Choose **Functions**  
**Importing/Exporting**  
**ToFastA...**
4. Click on **Run**.
5. When the conversion is complete the list of files will be displayed (You will have to click on **windows** and **Output manager** if your automatic display option is not set.) Close the display box.
6. Click on **Save As...**
7. In the Save As dialog box type in the new sequence name and click on **Save**.



**Note:** If you do not specify a path, the file will be saved in the SeaLab default directory.

---

## APPENDIX III

### Nucleotides

IUB/GCG	Meaning	Complement	Staden/Sanger
A	A	T	A
C	C	G	C
G	G	C	G
T/U	T	A	T
M	A or C	K	M
R	A or G	Y	R
W	A or T	W	W
S	C or G	S	S
Y	C or T	R	Y
K	G or T	M	K
V	A or C or G	B	V
H	A or C or T	D	H
D	A or G or T	H	D
B	C or G or T	V	B
X/N	G or A or T or C	X/N	N
./~	gap character	./~	-

### Amino Acids

Symbol	3-letter	Meaning	Codons	IUB Depiction
A	Ala	Alanine	GCT, GCC, GCA, GCG	!GCX
B	Asp, Asn	Aspartic, Asparagine	GAT, GAC, AAT, AAC	!RAY
C	Cys	Cysteine	TGT, TGC	!TGY
D	Asp	Aspartic	GAT, GAC	!GAY
E	Glu	Glutamic	GAA, GAG	!GAR
F	Phe	Phenylalanine	TTT, TTC	!TTY
G	Gly	Glycine	GGT, GGC, GGA, GGG	!GGX
H	His	Histidine	CAT, CAC	!CAY
I	Ile	Isoleucine	ATT, ATC, ATA	!ATH
K	Lys	Lysine	AAA, AAG	!AAR
L	Leu	Leucine	TTG, TTA, CTT, CTC, CTA, CTG	!TTR, CTX, YTR, YTX
M	Met	Methionine	ATG	!ATG
N	Asn	Asparagine	AAT, AAC	!AAY
P	Pro	Proline	CCT, CCC, CCA, CCG	!CCX
Q	Gln	Glutamine	CAA, CAG	!CAR
R	Arg	Arginine	CGT, CGC, CGA, CGG, AGA, AGG	!CGX, AGR, MGR; MGX
S	Ser	Serine	TCT, TCC, TCA, TCG, AGT, AGC	!TCX, AGY; WSX
T	Thr	Threonine	ACT, ACC, ACA, ACG	!ACX
V	Val	Valine	GTT, GTC, GTA, GTG	!GTX
W	Trp	Tryptophan	TGG	!TGG
X	Xxx	Unknown		!XXX
Y	Tyr	Tyrosine	TAT, TAC	!TAY
Z	Glu, Gln	Glutamic, Glutamine	GAA, GAG, CAA, CAG	!SAR
*	End	Terminator	TAA, TAG, TGA	!TAR, TRA; TRR