

Intermediate GCG

(Multi-sequence analysis workshop)

Multi-sequence analysis is a powerful tool to study the relationships between sequences and the evolution of genes. The basis of all such analysis is the alignment of multiple sequences. GCG provides powerful tool for aligning and editing multiple sequences. A good alignment, however, does not only rely on a good algorithm, but also on our knowledge of a particular gene.

Prepare a directory and start GCG

To create a directory type
`mkdir tutorial2`

Move to the directory with
`cd tutorial2`

Start GCG with
`gcg`

Fetch five sequences as follows:

```
fetch gb:m60331 -out=m60331.seq
fetch gb:x71334 -out=x71334.seq
fetch gb:x71335 -out=x71335.seq
fetch gb:x71336 -out=x71336.seq
fetch gb:x71337 -out=x71337.seq
```

As a shortcut, for the purposes of this tutorial use:
`fetch x*.seq`

Check your directory to see if all the sequences are there by typing
`ls`

Aligning two sequences

The program used to align two sequences, is “gap.” Gap uses the algorithm of Needleman and Wunsch to find the alignment of two complete sequences that maximizes the number of matches and minimizes the number of gaps. Gap considers all possible alignments and gap positions and creates the alignment with the largest number of matched bases and the fewest gaps.

To run gap type
`gap`

The program will ask for the names of the sequences, the ranges, and the strands and will then write the result to a “.pair” file. The alignment process involves the introduction of gaps until an optimal alignment is achieved. “Optimal” in this case is relative, since the alignment depends on the gap and gap length penalties.

This program is relatively simple and based on the widely accepted algorithm of Needleman and Wunsch (J. Mol. Biol. 48; 443-453 (1970)). However, it has the obvious disadvantage that it is limited to analyzing two sequences and there is no way to adjust the result manually. This remains a useful tool to check alignments of adjacent sequences.

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

To fetch an example sequence with repeats type
`fetch winchester.seq`

First you have to compare the sequences with
`compare winchester.seq`

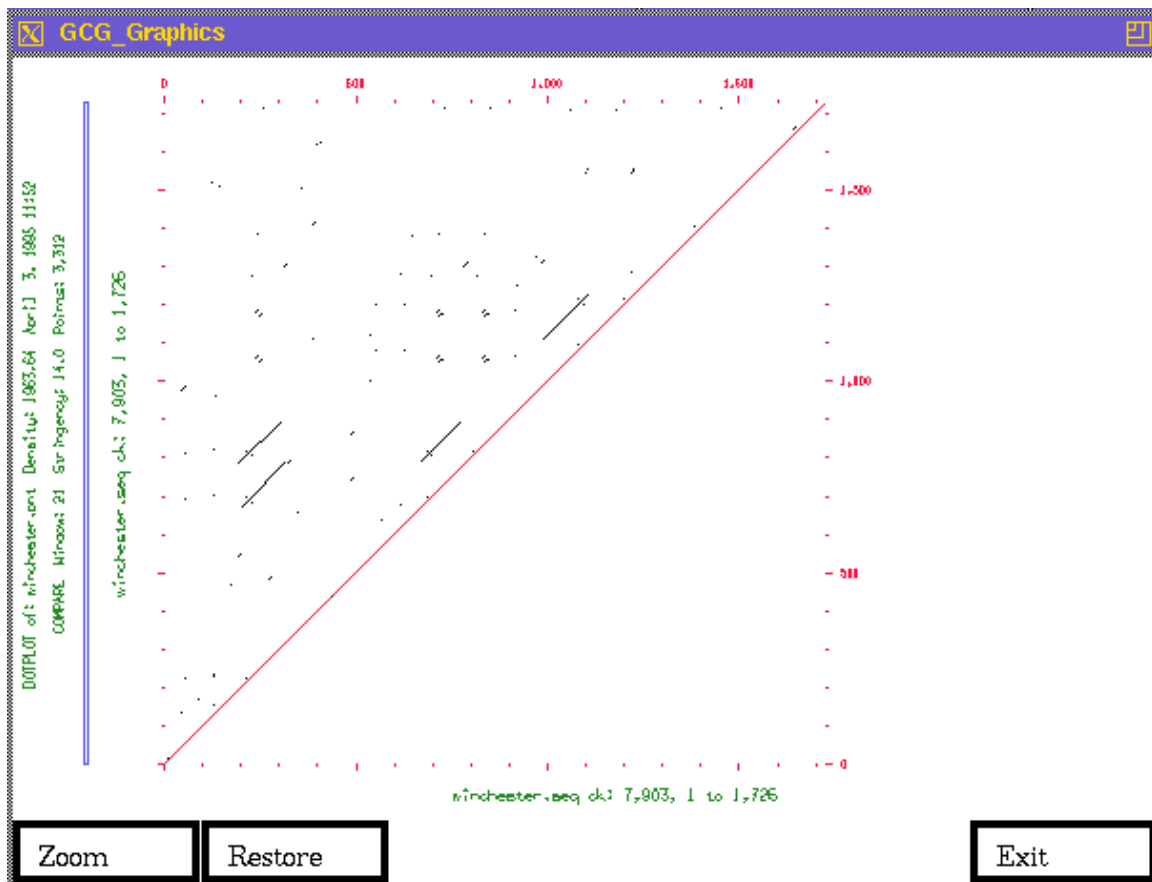
Accept all the defaults.

This will compare the sequence with itself and create a `winchester.pnt` output file.

Now plot the dotmatrix with
`dotplot winchester.pnt`

You will see a number of repeats.

Hint: 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 and helps to explain the output.



An Xwindows display of a dot-matrix analysis.

To search for reverse complemented repeats, run the program again, but type "y" when it asks to reverse one of the sequences. What do you expect?

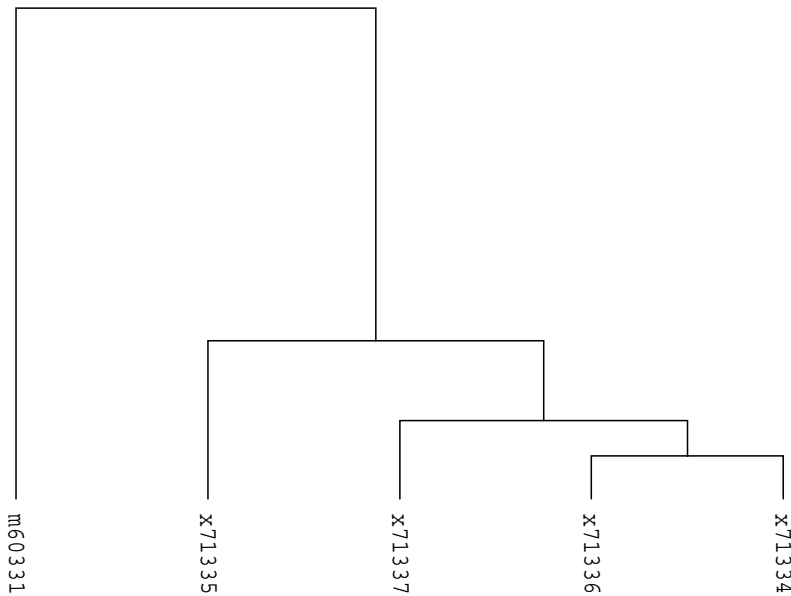
Multi-sequence alignment

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.

Function

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.

PILEUP of: @monkey.fil February 15, 1995 15:08



In this example sequences x71336 and x71334 are first aligned and the other sequences are added progressively.

Limitations

As shipped, PileUp restricts each sequence in the final alignment to a maximum length of 7,000 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 5,000 gap characters for each sequence in the final alignment.

If you wish to align longer sequences, then you can specify a maximum sequence length of up to 7,000 with the `-MAXSeg` command-line qualifier (e.g. `-MAXSeg=6000`). If you increase the maximum sequence length in this way, then the maximum amount of allowed gapping is automatically reduced so that the final aligned sequence length cannot exceed 7,000 for any sequence.

A simple multi-sequence alignment

You can do an alignment directly by typing

```
pileup *.seq
```

The result is written to an “msf” (multiple sequence format) file (For more information on the msf file format see later in this document.) To see the file type

```
more *.msf
```

The advantage of this procedure is that is simple. However, it does not allow you to specify parts of the sequences or the order in which they appear. All sequences in your directory that satisfies the wild card will also be included. This can become a problem in a big project. A better way to specify the sequences is through the use of a names file. In its simplest form, a names file is a list of file names.

Pileup also produces a dendrogram. The dendrogram is *not* a phylogenetic reconstruction, although the vertical branch lengths are proportional to the distance 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.

Create a list file manually

You can use the text editor of your choice to create a names file. In this exercise we use `jove` (See the “Jove, short descriptions” handout for more information.).

To create a names file called “monkey.fil” type

```
jove monkey.fil
```

Enter the names, one per line as follows

```
m60331.seq begin:470 end:1250
x71334.seq
x71335.seq
x71336.seq
x71337.seq
```

When you are finished save the file by typing

```
<control>x s
```

On some systems, pressing <control>s by accident will freeze the screen. Restore the screen with <control>q

Exit the program with

```
<control>x<control>c
```

Running pileup with a list file

You can run pileup now by typing

```
pileup @monkey.fil
```

The @ symbol is specifies that sample.fil is a names list. If the @ symbol is omitted, the program will regard sample.fil as a sequence file and produce the error message "no sequence in sample.fil"

The program will respond with

```
What is the gap creation penalty (* 5.00 *) ?
```

Press <return> to accept the default.

The gap creation penalty is a function of how many more bases must be aligned before a gap is introduced. A low value will produce many gaps and a high value fewer gaps.

The program will respond with

```
What is the gap extension penalty (* .30 *) ?
```

Press <return> to accept the default.

The gap extension penalty controls the size of the gaps. A low value will produce large gaps and a high value will produce smaller gaps.

The program will respond with

```
This program can display the clustering relationships graphically.
```

```
Do you want to:
```

- A) Plot to a FIGURE file called "pileup.figure"
- B) Plot graphics on LASERWRITER attached to |lpr -h -P achs_12
- C) Suppress the plot

```
Please choose one (* A *):
```

Press b<return> to plot or c<return> to suppress the plot.

The program will respond with

```
What should I call the output file name (* monkey.msf *) ?
```

Press <return> to accept the default.

The resulting file should look as follows:

```
FileUp of: @monkey.fil

Symbol comparison table: GenRunData:pileupdna.cmp  CompCheck: 6876

          GapWeight: 5.000
          GapLengthWeight: .300

monkey.msf  MSF: 782  Type: N  February 10, 1995 14:29  Check: 5893 ..

Name: x71334      Len:   782  Check: 2890  Weight:  1.00
Name: x71336      Len:   782  Check: 1594  Weight:  1.00
Name: x71337      Len:   782  Check: 5493  Weight:  1.00
Name: x71335      Len:   782  Check: 4085  Weight:  1.00
Name: m60331      Len:   782  Check: 1831  Weight:  1.00

//

          1                               50
x71334  ..... ..GGGCCAC TGCAGCCTCA GCCCAGGAGC CACCAGATCT
x71336  ..... ..GCCACA GCCCAGGAGC CACCAGATCT
x71337  ..... ..GCTGAGGAGC CACCAGATCT
x71335  TCCCTACCC CAGGGCCAC TGCAGCCTCA GCCCAGGAGC CACCGGATCT
m60331  .GATGCAGGC CACCTGGCAT GTTTGTGAG GTCCAGCCC TTTGCCCTCA

          51                               100
x71334  CCCAGACCA TGGTCCGATA CCGCGTGAGG AGCCCGAGCG AACCTCGCA
x71336  CCCAGACCA TGGTCCGATG CCGCGTGAGG AGCCCGAGCG AACGCTCGCA
x71337  .....A TGGTCCGATA CTGTGTGAGG AGCTGAGCG AACGCTCGCA
x71335  CCCAACACTA TGGTCCGATA CCACGTGAGG AGCCCAAGCG AACGCCACA
m60331  CAATGACCAA CGGCCCTG GCATCTATAA CAGGCCGAG AGCTGGCCCC
```

Maintaining the order or sequences

To maintain the order in which the sequences appear in the .msf file the same as in the list file include the -nosort option in the command line by typing

```
pileup @monkey.fil -nosort
```

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

LineUp is a screen editor for editing multiple sequence alignments. You can edit up to 30 sequences simultaneously. New sequences can be typed in by hand or added from existing sequence files. A consensus sequence identifies places where the sequences are in conflict.

To edit the monkey.msf multiple sequence file (msf) type

```
lineup monkey.msf
```

As in SeqEd, you can move the cursor with the arrow keys and insert or delete symbols or gaps in the sequences. In LineUp, the cursor can travel from one sequence to another. You can add new sequences by hand or from existing sequence files, and you can move sequences from one position to another.

LineUp provides a surface on which you can arrange and edit many sequences. This surface resembles a piece of graph paper with 31 rows and as many columns as you need. The screen acts as a window behind which the LineUp surface is scrolled.

When asked for a group name, enter "monkey".

Note: The file is in the FOSN format. To change it to the more useful MSF format, type msf at the command line

The screen should look as follows:

```

x71335      Column: 1  Row: 12      No AutoCons      FOSN: monkey      Nucleotide

15: .....GGGCCACTGCAGCCTCAGCCAGGAGCCACCAGATCTCCAGCACCATGGTCCGATACCGG
14: .....GCCACAGCCAGGAGCCACCAGATCTCCAGCACCATGGTCCGATGCCGCG
13: .....ATGGTCCGATACTGTG
12: TCCCCTACCCAGGGCCCACTGCAGCCTCAGCCAGGAGCCACCGGATCTCCCAACACTATGGTCCGATACCAG

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....
      0          10         20         30         40         50         60         70

"monkey.msf{*)" successfully loaded.
  
```

Here is the summary of Screen Mode commands you would see in the on-line help:

```

Screen Mode
[n] is an optional numeric parameter.

G, A, T, C .... - inserts a sequence character
<Delete>       - deletes a sequence character, "drags" a
                 sequence to the left if cursor is at its start
<Space bar>    - "pushes" a sequence to the right if cursor is at its start
/TAACG<Return> - finds the next occurrence of "TAACG", last
                 pattern is the default when none is specified
[n]<Right-arrow> - move ahead [n characters]
[n]<Left-arrow>  - move back [n characters]
[n]<Up-arrow>    - move up to next sequence [or to row specified]
[n]<Down-arrow> - move down to next sequence [or to row specified]
[n]<Return>     - move to column n
<Ctrl>H       - move to start of current sequence
<Ctrl>E       - move to end of current sequence
<Ctrl>R       - redraw the screen
<Ctrl>Z       - enter Command Mode      Use <Ctrl>D to enter command mode
<Ctrl>I       - push over all seqs starting past current column
<Ctrl>D       - pull over all seqs starting past current column
[n]<          - move 50 [or n] positions to left
[n]>          - move 50 [or n] positions to right
  
```

Entering Command Mode

Use <Ctrl>D to leave Screen Mode and enter Command Mode.

Returning to the Screen Mode

If you simply press <Return>, LineUp returns to Screen Mode described above. If you have -SINGLE command on the command line or in your command-line initializing file, LineUp returns to Screen Mode immediately after executing each command.

Command Mode

x and y represent numbers for column and row.
Only the capitalized part of the command is necessary.

| | | |
|-------|--------------------|--|
| [x,y] | Get [filename] | - add sequence [at position x,y] [from filename] |
| [x,y] | New [seqname] | - add empty sequence [at position x,y] [named seqname] |
| [x,y] | Move [seqname] | - move current or specified sequence [to x,y] |
| | REMove [seqname] | - delete current or specified sequence entirely |
| | REName [old] [new] | - change sequence name (changing consensus name changes the group) |
| | REDraw | - redraw the screen |
| | HEAding [seqname] | - edit documentary heading of current or specified sequence |
| | screen | - enter screen mode (pressing <Return> is sufficient) |
| | NUCleotide | - use nucleotide ambiguity codes in find and consensus |
| | PROtein | - do not use nucleotide ambiguity codes |
| | SPOacewalk | - use spacewalk to position sequences |
| | NOspacewalk | - DO NOT use spacewalk to position sequences |
| | FOSN | - use list file format when writing |
| | MSF | - use multiple sequence format files when writing |
| [n] | S�ide | - add n to all sequence columns |
| [s,f] | ROWMove [n] | - move a set of rows (s to f) up or down [n rows] |
| [s,f] | PRint [filename] | - write the sequence group to a Pretty format file |
| | SUMmary [filename] | - write the sequence names and positions in a file or on the terminal screen |
| | GOto [seqname] | - put cursor on start of named sequence |
| [s,f] | CONSeensus | - calculate consensus [from s to f] |
| | AUTOconsensus | - automatically calculate consensus (slow, may cause the program to crash) |
| | NOAUTOconsensus | - turn off automatic consensus |
| | FLip | - reverse complement the current group |
| | ZIP [filename] | - align and gap a sequence to the current group (do NOT use with autocon) |
| | Write [filename] | - write the current sequence group to a file |
| | EXit [filename] | - write the current group to a file and stop |
| | Quit | - quit the editor without writing out the group |

LineUp lets you modify and execute previous commands.

The <Up-arrow> key displays previous commands.

Creating a consensus sequence

1. Go to the command line with <ctrl>d
2. Create a new sequence by typing n<return> and give the sequence name the same as the group name, "monkey" in this case. Move to an empty line and press <return>
3. Type autocon<return>

You can have a consensus sequence display the dominant character at each column where sequences overlap. The consensus uses uppercase where overlapping sequences are in agreement; it uses lowercase to show disagreement and periods to show where there is no consensus at all.

NOTE: The PRETTY program provides a simpler way to add a consensus sequence to an alignment.

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 | Optional parameter |
|---------------|--------------------|
| Assemble | -LISTfile |
| Corrupt | -LISTfile |
| FastA | -NOALIGN |
| FindPatterns | -NAMEs |
| FromEMBL | -LISTfile |
| FromFastA | -LISTfile |
| FromGenBank | -LISTfile |
| FromIG | -LISTfile |
| FromPIR | -LISTfile |
| Lineup | |
| Motifs | -NAMEs |
| Names | |
| Pretty | -UGLY |
| ProfileSearch | |
| Reformat | -LISTfile |
| Sample | -LISTfile |
| Simplify | -LISTfile |
| StringSearch | |
| Lookup | |
| TFastA | -noalign |
| Translate | -LISTfile |
| Wordsearch | |

Producing a list file with Lookup

To search the database with lookup type
 lookup<return>

Select the database

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

Write te results to a file, “lookup.list”

For more information on lookup see *Introduction to GCG*
 (Document ACHS-306)

```
LOOKUP in: genbank of: "[SQ-ALL: protamine*]"
147 entries January 30, 1996 10:36 ..
GB_BA:ECOTGY1 ! ID: c51e0005
! DEFINITION E.coli tyrT locus containing two Tyr-tRNA-1 genes.
GB_IN:AGPROT ! ID: 2d640005
! DEFINITION Boll Weevil mRNA for protamine.
GB_IN:CFDC2G ! ID: 536e0005
! DEFINITION C.fasciculata cdc2 gene homologue.
GB_IN:DMMST35BA ! ID: 87750005
! DEFINITION D.melanogaster mRNA for protamine (mst35Ba).
GB_IN:DMMST35BB ! ID: 88750005
! DEFINITION D.melanogaster mRNA for protamine (mst35Bb).
GB_IN:DMMSTBAGE ! ID: 96750005
! DEFINITION D.melanogaster gene for protamine (mst35Bb).
GB_IN:DMMSTBBGE ! ID: 97750005
! DEFINITION D.melanogaster gene for protamine (mst35Bb).
GB_IN:HTPHIO ! ID: 5d8a0005
! DEFINITION Holothuria tubulosa mRNA for sperm-specific protein phi-0.
GB_IN:MESSPP1A ! ID: f48d0005
! DEFINITION M.edulis mRNA for sperm-specific protein Phi-1.
GB_OM:ASWPRP1A ! ID: b4a90005
! DEFINITION Antechinus stuartii protamine P1 gene, complete cds.
```

You can now run pileup using this listfile as follows:

```
pileup @lookup.fil
```

Since all these sequences have a complete database address, they do not have to be in your local directory. The program will automatically find them in the database.

Producing a list file with FastA

To produce a list file using fasta type
 fasta -noalign

Respond with the sequence name “x71334.seq”

You can now edit the x71334.fasta output file and use it as a list file for pileup.

Displaying sequence similarities

To plot 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 type

```
plotsimilarity monkey.msff{*}
```

Note: Run this command only if you have your default printer set.

The results will be displayed on your graphic screen.

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 ttggcaccct tctgctgatt tggaaaggcca cagaccacc taaatHATGG CAAGATTCAG
mus cigaccacagc ttggtgtcccc tgctctgagc cagctccccg ccaagccacc ..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 that 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.. aaahHATGGCA AGATATAGAC GACACAGCAG GAGCCGGAGT AGGAGCAGAT
harri cagtca.. aaahHATGGCA AGATATAGAC GACGCAGCAG GAGCCGGAGT AGGAGCAGAT
viver cagtca.. aaahHATGGCA AGATATAGAC GACGCAGCAG GAGCCGGAGT AGGAGCAGAT
macdo cagtca.. aaahHATGGCA AGATATAGAC GACACAGCAG GAGCCGGAGT AGGAGCAGAT
stuar cagtca.. aaahHATGGCA AGATATAGAC GACACAGCAG GAGCCGGAGT AGGAGCAGAT
tapoa cagtca.. aaahHATGGCA AGATATAGAC GACACAGCAG GAGCCGGAGT AGGAGCAGAT
koala cactga.. aaahHATGGCA AGATATA... GACACAGCAG GAGCCGGAGT AGGAGCAGAT
domestic cactga.. aaahHATGGCA AGATATAGAA GACGCAGCAG GAGCCGGAGT AGGAGCAGAT
echidna caccta.. aathHATGGCA AGATTCA... GGCACAGCCG GAGCCGGAGC CGCAGCCTGT
mus aagccagc acchHATGGCC AGATAC...C GATGCTGCCC CAGCAAAGC AGGAGCAGAT
```

Hint: In the above example uppercase is used to mark the coding region. The program does not differentiate between upper- and lowercase.

Forcing the alignment of an intron junction

Original alignment

```
dorsa  tcta tttttgttta aacttcoccta tcatccctcc ctgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
harri  tcta tttttgttta aacttcocctg tcatccctcc ctgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
viver  tcta tttttgttt. aacttcocctg tcatccctcc ctgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
macdo  tcta tttttgttta aacttcocctg tcatccctcc ctgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
stuar  tcta tttttgttta aacttcocctg tcatccctcc ccgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
tapoa  tcta tttttgttta aacttcocctg tcatccctcc ccgctcagHG GTATTCTCGC AGGAGATATT C..... TCGCAGGGGA AGAAGAA
koala  tcta tttttaaatt tagccttctt tcatctct.. ...ctcagHG GTA...TCGC AGGAGATATT C..... TCGCAGG... ..AGAA
domestic ... .tctcttta aaccttc.a ttattctct tttctcagHG GTA...CCAC AGGAGATCTC CTCATCGTCG TCGTAGGAGA AGAAGAA
echidna cccc agHGTAGAGC CAGCATGAGA TCCTCTGCA GAAGAAGAAG GAGGAGAAGA AACTGALgag ccactctcca tgctctctct cgagaac
mus    tgag aattttacca gaactcaaga gcatctcgcc acatcttgaa aaatgccacc gtcogatgaa 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 tcatccctcc ctgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
harri  tcta tttttgttta aacttcocctg tcatccctcc ctgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
viver  tcta tttttgttt. aacttcocctg tcatccctcc ctgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
macdo  tcta tttttgttta aacttcocctg tcatccctcc ctgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
stuar  tcta tttttgttta aacttcocctg tcatccctcc ccgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
tapoa  tcta tttttgttta aacttcocctg tcatccctcc ccgctcagHG GTATTCTCGC AGGAGAT... ..ATTCTCG CAGGGGAAGA AGAAGAT
koala  tcta tttttaaatt tagccttctt tcatctct.. ...ctcagHG GTA...TCGC AGGAGAT... ..ATTCTCG CAGG..... AGAAGAT
domestic ... .atctcttta aaccttc.a ttattctct tttctcagHG GTA...CCAC AGGAGATCTC CTCATCGTCG TCGTAGGAGA AGAAGAA
echidna cctt cacatctgt. .... tctctctcc ...cccagHG TAGACGCAGC ATGAGATCCT CTCGAGAAG AAGAAGGAGG AGAAGAA
mus    acca cttttct... ..tac cttctcagHG ATGCTGCCCT CCGCCCGCCT CATAAC... CATAAGGTGT AAAAAAT
```

Fetching and editing the DNA comparison table

You can fetch the DNA comparison table to your home 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 your favorite text processor. For this exercise change the value for a H:H match to 100.

pileupdna.cmp

Default 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.

January 23, 1991 14:09

| A | B | C | D | G | H | K | M | N | R | S | T | U | V | W | X | Y | .. |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|
| 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | A |
| | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | B |
| | | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | C |
| | | | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | D |
| | | | | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | G |
| | | | | | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | H |
| | | | | | | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | K |
| | | | | | | | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | M |
| | | | | | | | | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | N |
| | | | | | | | | | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | R |
| | | | | | | | | | | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | S |
| | | | | | | | | | | | 1.0 | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | T |
| | | | | | | | | | | | | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | U |
| | | | | | | | | | | | | | 1.0 | 1.0 | 1.0 | 1.0 | V |
| | | | | | | | | | | | | | | 1.0 | 1.0 | 1.0 | W |
| | | | | | | | | | | | | | | | 1.0 | 1.0 | X |
| | | | | | | | | | | | | | | | | 1.0 | Y |

To weight the matrix
increase the value from
1.0 to 100

Practice session:

The following set of sequences have an "H" character inserted to mark the positions of the start and stop codons as well as the intron junctions. To fetch the sequences type

```
fetch *.fix
```

You can now run pileup as follows:

Without weighting

```
pileup @tutorial2.fix
```

With weighting

```
pileup @tutorial2.fix -dat=fixdna.cmp
```

Where fixdna.cmp is the modified comparison table

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 RSRSLY.RRR ..RRSRGGR QTRSRKLSRS RRRGRSRRRK
hump1  MARY.RCCRS QSRORYRQR QRSR..... .RRRRS CQTRRRAMRW
mus    MARY.RCCRS KSRRCRRR RRRCR..... .RRRRRC CRRRR..RW
opos   MARYRRRSRS RSRRYGRRR RRSRSR.... .RRRSRR RRRRRGRRGR
domestic MARYRRRSRS RSRRYGRRR RRSRSR.... .RRRSRR RRRRRGRRGR
caen   MARY.RHSRS RSRRYRRR RRRRSRYRSR RRRY.RSR. RRRRG.RRR
austr  MVRYRRHSRS RSRRYRRR RR...RLNR RRYRRSRG RRRRRGSR
bandi  MASY.RNSRS RSRRF.RRR RRRSRVRGR DARQGRSRR RRRGKGRAHS
redkan MARY.RHSRS RSRRY.RRR RRRRSYRSQ RRYRGRRR. RSRRG.RRR
tamar  MARY.RHSRS RSRRY.RRR RRRRSYRSR RRSRGRRR. RSRRGRRR
koala  MARY.RHSRS RSRRY.QRR RRRRSYRSQ RRYRRRGS RRRRRGRR
btpossum MARY.RHSRS RSRRYRRR RRRRSYRSR RRYR.RSR. RRRRGRRR
notory MARY.RHSRS RSRRY.RRR RRRRSYRSQ RRYRRHRS GRRRRGRR
dorsa  MARYRRHSRS RSRRY.RRR RRRSRGR.R RRYRRSR. HSRRRGRR
macdo  MARYRRHSRS RSRRY.RRR RRRSRHRNR RRYRRSR. HSRRRGRR
crass  MARYRRHSRS RSRRY.RRR RRRSRHNR RRYRRSR. HSRRRGRR
pingra MARSRRHSRS RSRNQCQR RRRRT..YN RRRTMREKPR HSRRRVRR
stuar  MARYRRHSRS RSRRYRRR RRRSRHNR RRYRRSR. HSRRRGRR

```

Weighted alignment

```

plat  MARF.RRSRS RSRSLY.RRR ..RRSR...R .....GGRQT RSRKLSRSRR
hump1  MARY.RCCRS QSRORY..YR QRQRSR.... .RRR RS..CQTRRR
mus    MARY.RCCRS KSRRC..RR RRRRCR.... .RRR RC..CRRRR
opos   MARYRRRSRS RSRRYGRRR RRSRSR.... .RR RSR.RRRR
domestic MARYRRRSRS RSRRYGRRR RRSRSR.... .RR RSR.RRRR
caen   MARY.RHSRS RSRRYRRR RRRRSYRSR RRY.....R RSR.R.RRR
austr  MVRYRRHSRS RSRRYRRR RR...RLNR RRY.....R RSRGRRRR
bandi  MASY.RNSRS RSRRF.RRR RRRSRVRGR .....DARQG RS...SRRR
redkan MARY.RHSRS RSRRY.RRR RRRRSYRSQ RRYRGRRR RS.....RR
tamar  MARY.RHSRS RSRRY.RRR RRRRSYRSR RRSRGRRR RS.....RR
koala  MARY.RHSRS RSRRY.QRR RRRRSYRSQ RRY..RRR GSRR.R.RR
btpossum MARY.RHSRS RSRRYRRR RRRRSYRSR RRY.....R RSR.R.RRR
notory MARY.RHSRS RSRRY.RRR RRRRSYRSQ RRY..RRH RSGR.R.RR
dorsa  MARYRRHSRS RSRRY.RRR RRRSRGR.R RRY.....R RSR.HSRR
macdo  MARYRRHSRS RSRRY.RRR RRRSRHRNR RRY.....R RSR.HSRR
crass  MARYRRHSRS RSRRY.RRR RRRSRHNR RRY.....R RSR.HSRR
pingra MARSRRHSRS RSRNQCQR RRRRT.Y..N RRRTMREKPR HSRRRVRR
stuar  MARYRRHSRS RSRRYRRR RRRSRHNR RRY.....R RSR.HSRR

```

Amino acid comparison table

To modify the amino acid comparison table, first copy the table to your working directory. To do this type
`fetch pileuppep.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.

Note: GCG will first look in your working directory for your .cmp file. If you have a file called pileuppep.cmp in your working directory, GCG will read that one first.

To use your own comparison table, called “fixpep.cmp” use the `-dat=fixpep.cmp` option in the command line e.g.

```
pileup @yourlisfile.fil -dat=fixpep.cmp
```

To get more information on data files for any particular command, look under the genhelp, “local data files” option.

pileuppep.cmp

Default scoring matrix used by PILEUP for the comparison of protein sequences. PILEUP uses the method of Needleman/Wunsch/Sellers to make alignments.

Dayhoff table (Schwartz, R. M. and Dayhoff, M. O. [1979] in Atlas of Protein Sequence and Structure, Dayhoff, M. O. Ed, pp. 353-358, National Biomedical Research Foundation, Washington D.C.) rescaled by dividing each value by the sum of its row and column, and normalizing to a mean of 0 and standard deviation of 1.0. The value for FY (Phe-Tyr) = RW = 1.425. Perfect matches are set to 1.5 and no matches on any row are better than perfect matches.

Table used by Gribskov and Burgess NAR 14(16) 6745-6763

December 29, 1986 12:46

| A | B | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y | Z | .. |
|-----|-----|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|----|
| 1.5 | 0.2 | 0.3 | 0.3 | 0.3 | -0.5 | 0.7 | -0.1 | 0.0 | 0.0 | -0.1 | 0.0 | 0.2 | 0.5 | 0.2 | -0.3 | 0.4 | 0.4 | 0.2 | -0.8 | -0.3 | 0.2 | A |
| | 1.1 | -0.4 | 1.1 | 0.7 | -0.7 | 0.6 | 0.4 | -0.2 | 0.4 | -0.5 | -0.3 | 1.1 | 0.1 | 0.5 | 0.1 | 0.3 | 0.2 | -0.2 | -0.7 | -0.3 | 0.6 | B |
| | | 1.5 | -0.5 | -0.6 | -0.1 | 0.2 | -0.1 | 0.2 | -0.6 | -0.8 | -0.6 | -0.3 | 0.1 | -0.6 | -0.3 | 0.7 | 0.2 | 0.2 | -1.2 | 1.0 | -0.6 | C |
| | | | 1.5 | 1.0 | -1.0 | 0.7 | 0.4 | -0.2 | 0.3 | -0.5 | -0.4 | 0.7 | 0.1 | 0.7 | 0.0 | 0.2 | 0.2 | -0.2 | -1.1 | -0.5 | 0.9 | D |
| | | | | 1.5 | -0.7 | 0.5 | 0.4 | -0.2 | 0.3 | -0.3 | -0.2 | 0.5 | 0.1 | 0.7 | 0.0 | 0.2 | 0.2 | -0.2 | -1.1 | -0.5 | 1.1 | E |
| | | | | | 1.5 | -0.6 | -0.1 | 0.7 | -0.7 | 1.2 | 0.5 | -0.5 | -0.7 | -0.8 | -0.5 | -0.3 | -0.3 | 0.2 | 1.3 | 1.4 | -0.7 | F |
| | | | | | | 1.5 | -0.2 | -0.3 | -0.1 | -0.5 | -0.3 | 0.4 | 0.3 | 0.2 | -0.3 | 0.6 | 0.4 | 0.2 | -1.0 | -0.7 | 0.3 | G |
| | | | | | | | 1.5 | -0.3 | 0.1 | -0.2 | -0.3 | 0.5 | 0.2 | 0.7 | 0.5 | -0.2 | -0.1 | -0.3 | -0.1 | 0.3 | 0.5 | H |
| | | | | | | | | 1.5 | -0.2 | 0.8 | 0.6 | -0.3 | -0.2 | -0.3 | -0.3 | -0.1 | 0.2 | 1.1 | -0.5 | 0.1 | -0.2 | I |
| | | | | | | | | | 1.5 | -0.3 | 0.2 | 0.4 | 0.1 | 0.4 | 0.8 | 0.2 | 0.2 | -0.2 | 0.1 | -0.6 | 0.4 | K |
| | | | | | | | | | | 1.5 | 1.3 | -0.4 | -0.3 | -0.1 | -0.4 | -0.4 | -0.1 | 0.8 | 0.5 | 0.3 | -0.2 | L |
| | | | | | | | | | | | 1.5 | -0.3 | -0.2 | 0.0 | 0.2 | -0.3 | 0.0 | 0.6 | -0.3 | -0.1 | -0.1 | M |
| | | | | | | | | | | | | 1.5 | 0.0 | 0.4 | 0.1 | 0.3 | 0.2 | -0.3 | -0.3 | -0.1 | 0.4 | N |
| | | | | | | | | | | | | | 1.5 | 0.3 | 0.3 | 0.4 | 0.3 | 0.1 | -0.8 | -0.8 | 0.2 | P |
| | | | | | | | | | | | | | | 1.5 | 0.4 | -0.1 | -0.1 | -0.2 | -0.5 | -0.6 | 1.1 | Q |
| | | | | | | | | | | | | | | | 1.5 | 0.1 | -0.1 | -0.3 | 1.4 | -0.6 | 0.2 | R |
| | | | | | | | | | | | | | | | | 1.5 | 0.3 | -0.1 | 0.3 | -0.4 | 0.0 | S |
| | | | | | | | | | | | | | | | | | 1.5 | 0.2 | -0.6 | -0.3 | 0.1 | T |
| | | | | | | | | | | | | | | | | | | 1.5 | -0.8 | -0.1 | -0.2 | V |
| | | | | | | | | | | | | | | | | | | | 1.5 | 1.1 | -0.8 | W |
| | | | | | | | | | | | | | | | | | | | | 1.5 | -0.6 | Y |
| | | | | | | | | | | | | | | | | | | | | | 1.1 | Z |

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. To fetch the sequences type

```
fetch *.pro
```

You can now run pileup as follows:

Without weighting

```
pileup @tutorial2.pro
```

With weighting

```
pileup @tutorial2.pro -dat=fixpep.cmp
```

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 use the `-endw` option in the command line.

Normal alignment

```

plat  RKGWRRSRR. .SSRRSRRRN *....
humpl . . .W.CCRPR . . .YRPRCRR H....
mus   . . .W.CCRRR .RSYTIRCKK Y*...
opos  GRGW.YHRR. . .SPHRRRRR RRRRA*
domestic GRGW.YHRR. . .SPHRRRRR RRR*..
caen  RRGW.YSRRR Y.S.RRRRRR Y*...
austr RRGW.YSRRR YQSRRRRRRR Y*...
bandi KKGWRRS... .GSRRRKRNN ENK*..
redkan RRGW.YSRRR Y.S..RRRRR Y*...
tamar RRGW.YSRRR Y.S.RRRRRR Y*...
koala RRGW.Y.RRR Y.S.R..RRR Y*...
btpossum RRGW.YSRRR Y.S.RRGRRR Y*...
notory RRGW.Y.RRR YHS.H..RRR Y*...
dorsa  RRGW.YSRRR Y.S.RRGRRR Y*...
macdo  RRGW.YSRRR Y.S.RRGRRR Y*...
crass  RRGW.YSRRR Y.S.RRGRRR Y*...
pingra . .GW.CSCRR C.SRRRRRRC *....
stuar  RRGW.YSRRR Y.S.RRGRRR Y*...

```

End-weighted alignment

```

plat  RKGWRRSRR. . .SSRRSRR RN....*
humpl . . .W.CCRPR . . .Y.....R PRCRRH.
mus   . . .W.CCRRR RSY.....T IRCKKY*
opos  GRGW.YHRR. . . .SPHRRR RRRRRRA*
domestic GRGW.YHRR. . . .SPHRRR RRRRR. *
caen  RRGW.YSRRR . .Y.S.....R RRRRRY*
austr RRGW.YSRRR . .YQS.....R RRRRRY*
bandi KKGWRRS... . .GSRRRKR NNENK. *
redkan RRGW.YSRRR . .Y.S..... RRRRRY*
tamar RRGW.YSRRR . .Y.S.....R RRRRRY*
koala RRGW.Y.RRR . .Y.S.....R . .RRRY*
btpossum RRGW.YSRRR . .Y.S.....R RGRRRY*
notory RRGW.Y.RRR . .YHS.....H . .RRRY*
dorsa  RRGW.YSRRR . .Y.S.....R RGRRRY*
macdo  RRGW.YSRRR . .Y.S.....R RGRRRY*
crass  RRGW.YSRRR . .Y.S.....R RGRRRY*
pingra . .GW.CSCRR . . .CS.....R RRRRRC*
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.

Hopeless cases

*** *** * * * Indicative sites

```

Macropusg -----caa--
Phascolod tcttagattattggggagggga----gtgcaaat
Murexialo tcttagatttgggggtgggg-aggagtgtgcaaat
Pseudantm tcttagattattggggaagggggcg-gtgcaaat
Sminthopc tcttaaattat--ggtgggggcggtgtgaaaat
Planigali tcttagatctagtttatgggaggggggtgcaagt
Antechist tcttagatt-tgggggtggggaagagtgtgcaaat
Antechisw tcttag--tattgggggtgggagggagtgtgcaaat
Phascogat tcttagatttggggaggggaggaatgtgcaaat
Dasyurusv tcttagattattggggaaggggggc--gtgcaaat
Dasyurush tcttagattattggggaagggggcg-gtgcaaat
Sarcophil tcttagattattgggga-ggg-gg--gtgcaaat
Dasykalut tcttagattattgggga-ggg-g---gtgcaaat
Paranteca tcttagattattgggga-ggg-gg---gtgcaaat

```

In spite of our best efforts, some parts of a group of sequences can sometimes not be aligned 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.

Displaying aligned sequences

The program pretty can help to display sequences, change the format, show differences, and calculate a consensus sequence. The consensus sequence calculated by PRETTY is different, and more sophisticated, than the one calculated by LINEUP.

Manual boxing

To manually box conserved domains, we recommend personal computer program, PageMaker.

Automatic boxing

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.