MCB 5472

Sequence alignment

Peter Gogarten
Office: *BSP 404*
phone: *860 486-4061,*
Email: *gogarten@uconn.edu*

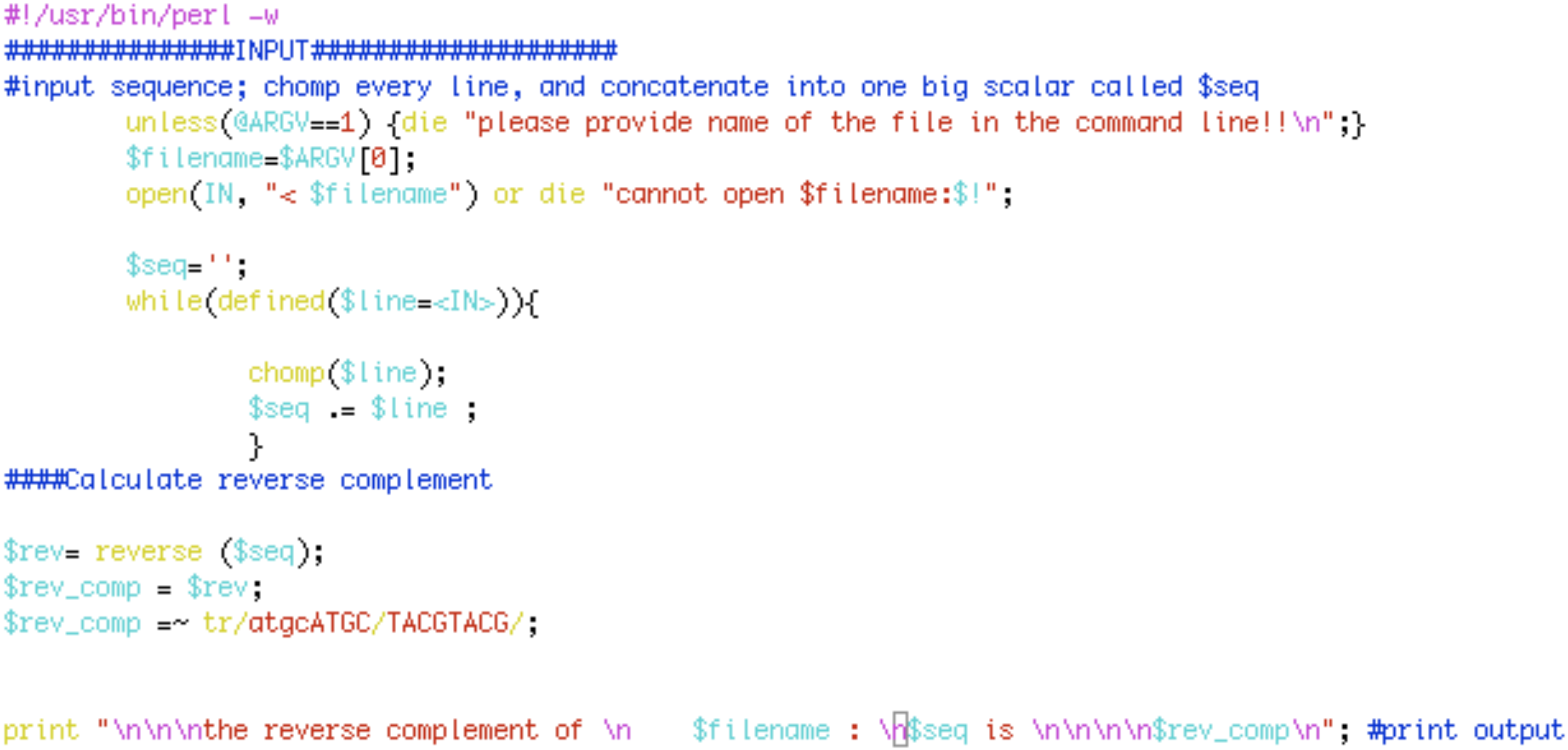
OLD ASSIGNMENTS

Write a script that reads in a sequence and prints out the reverse complement.
Modify your script to that it can handle a sequence that goes over several lines.

- Background: `$comp =~ tr/ATGC/TACG/;` #translates every A in $comp into a T; every T into an A; every G into a C and every C into a G

- Read P 14 on hashes, write the program suggested in the chapter.

- Write a script reads in a sequence and prints out the reverse complement.
- Modify your script to that it can handle a sequence that goes over several lines?



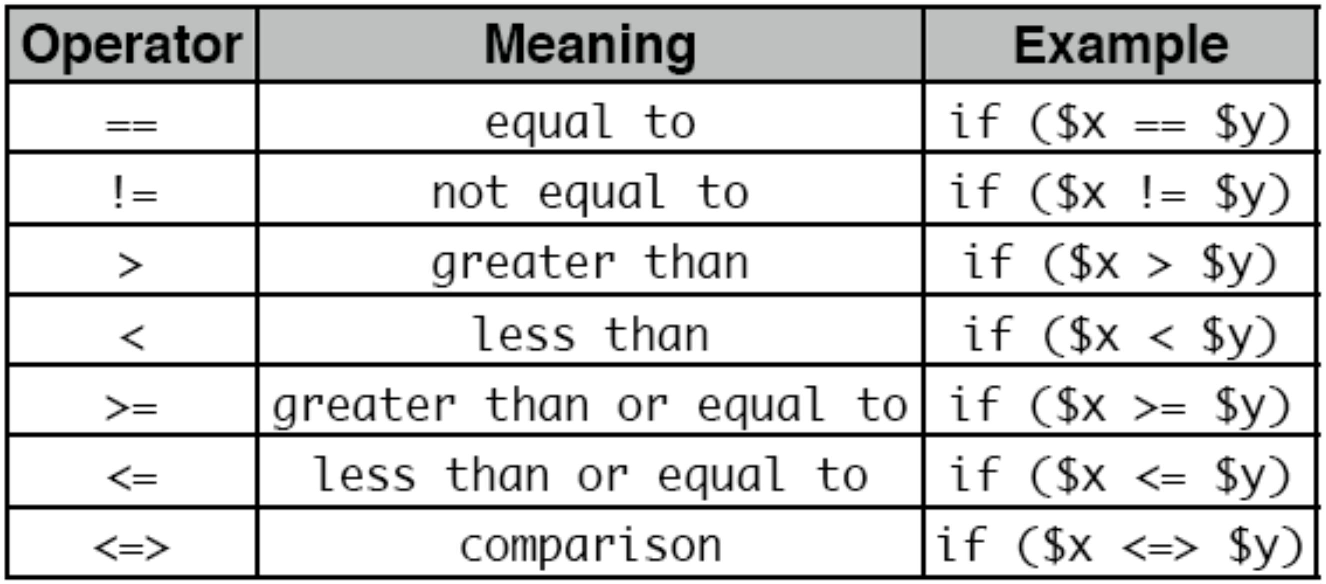
```
#!/usr/bin/perl -w
###############INPUT####################
#input sequence; chomp every line, and concatenate into one big scalar called $seq
        unless(@ARGV==1) {die "please provide name of the file in the command line!!\n";}
        $filename=$ARGV[0];
        open(IN, "< $filename") or die "cannot open $filename:$!";

        $seq='';
        while(defined($line=<IN>)){

                chomp($line);
                $seq .= $line ;
                }
####Calculate reverse complement

$rev= reverse ($seq);
$rev_comp = $rev;
$rev_comp =~ tr/atgcATGC/TACGTACG/;


print "\n\n\nthe reverse complement of \n    $filename : \n$seq is \n\n\n\n$rev_comp\n"; #print output
```

Go through class4Answers.pl

Go through sort_example1.pl and sort_example2.pl

Do the following statements evaluate to true or false? (Check P5)

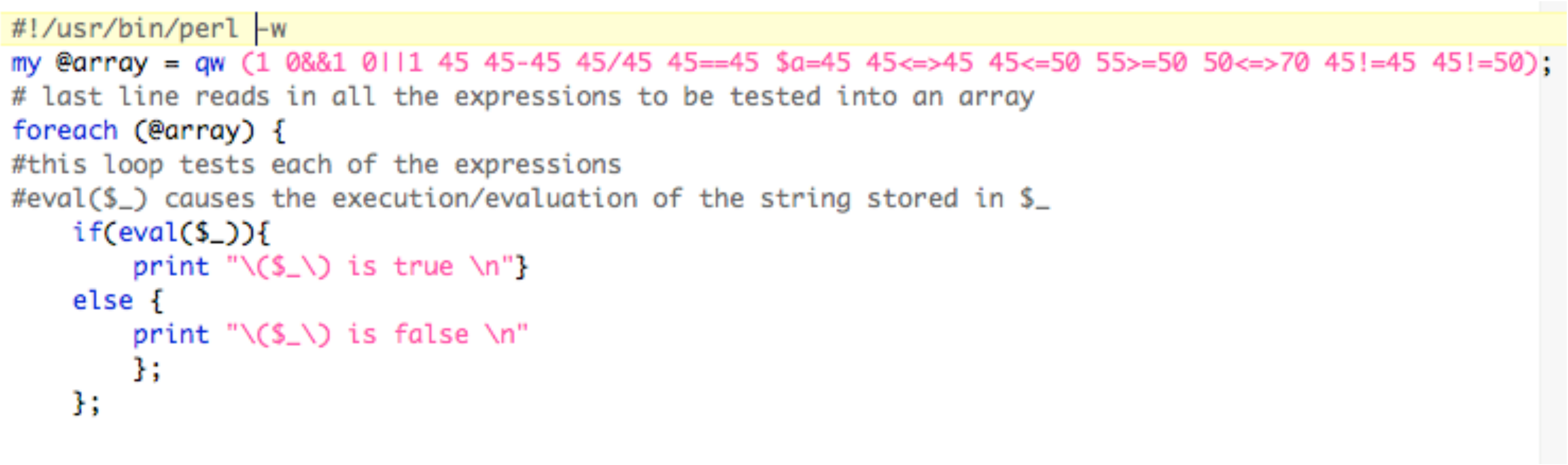
- 1
- 0 && 1
- 0||1
- 45
- 45-45
- 45/45
- 45==45
- 45<=>45
- 45<=50
- 55>=50
- 50<=>70
- 45!=45
- 45!=50



Operator	Meaning	Example
==	equal to	if ($x == $y)
!=	not equal to	if ($x != $y)
>	greater than	if ($x > $y)
<	less than	if ($x < $y)
>=	greater than or equal to	if ($x >= $y)
<=	less than or equal to	if ($x <= $y)
<=>	comparison	if ($x <=> $y)

from http://korflab.ucdavis.edu/Unix_and_Perl/unix_and_perl_v2.3.3.pdf

True or False?



```
#!/usr/bin/perl -w
my @array = qw (1 0&&1 0||1 45 45-45 45/45 45==45 $a=45 45<=>45 45<=50 55>=50 50<=>70 45!=45 45!=50);
# last line reads in all the expressions to be tested into an array
foreach (@array) {
#this loop tests each of the expressions
#eval($_) causes the execution/evaluation of the string stored in $_
    if(eval($_)){
        print "\($_\) is true \n"}
    else {
        print "\($_\) is false \n"
        };
    };
```

NEW ASSIGNMENTS

Read through P20. Functions (subroutines)

Turn your script that calculates the reverse complement of a sequence into a subroutine

Write a script that takes all files with the extension .fa (containing a single fasta formated sequence) and writes their contents in a single multiple sequence file.

Read through class5.pl

assignments continued (use class 5 as sample)

Assume that you have the following non-aligned multiple sequence files in a directory:

A.fa : vacuolar/archaeal ATPase catalytic subunits ;
B.fa : vacuolar/archaeal ATPase non-catalytic subunits;
alpha.fa : F-ATPases non-catalytic subunits,
beta.fa : F-ATPases catalytic subunits,
F.fa : ATPase involved in the assembly of the bacterial flagella.

Write a perl script that executes muscle or clustalw and
1) aligns the sequences within each file
2) successively calculates profile alignments between all aligned sequences.

Hints:

`system (command);`# executes "command" as if you had typed `command` in the command line

global *vs* local

Alignments can be global or local.
BLAST calculates **local** alignments, for databank searches and to find pairwise similarities local alignments are preferred.

Example using bl2seq with GIs : **137464** *versus* **6319974**
and **137464** *versus* **254565713**

However, for multiple sequences to be used in phylogenetic reconstruction, global alignments are the usual choice.
We will use two programs: MUSCLE and CLUSTALW

Note: Multiple alignments are more accurate than pairwise alignments! (see Fig 12.2. in Higgs and Atwood). The more sequences one includes, the more reliable the result. Same for phylogenetic reconstruction (taxon sampling).

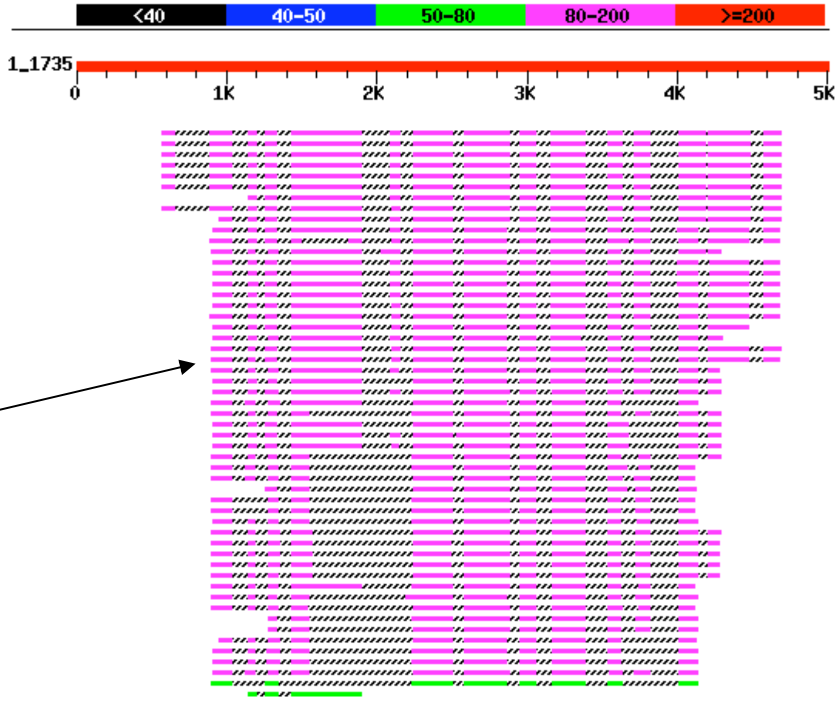
dotlet

The Swiss Institute for Bioinformatics provides a JAVA applet that perform interactive dot plots. It is called Dotlet. The main use of dot plots is to detect domains, duplications, insertions, deletions, and, if you work at the DNA level, inversions (excellent illustrations of the use of dot plots are given on the examples page).

One application of this program is to find internal duplications and to locate exons.

Example: this sequence against itself
(if time do in bl2seq as well)
genomic sequence against Protein

As similar result can be obtained using blastx against a protein databank



The Needlemann Wunsch Algorithm

a step by step illustration is here

a) **fill in scoring matrix**
b) **calculate max. possible score for each field**
c) **trace back alignment through matrix**

see http://en.wikipedia.org/wiki/Needleman–Wunsch_algorithm and http://snowedin.net/ideas/Analogies+in+Alignment for multiple paths.

Caution

NOTE that clustalw and other multiple sequence alignment programs do NOT necessarily find an alignment that is optimal by any given criterion.

Even if an alignment is optimal (like in the Needleman-Wunsch algorithm), it usually is not UNIQUE. It often is a good idea to take different extreme pathways through the alignment matrix, or to use a program like tcoffee that uses many different alignment programs.

clustalw

runs on all possible platforms (unix, mac, pc), and it is part of most multiprogram packages, and it is also available via different web interfaces. Examples: here, and here.

Clustalw uses a very simple menu driven command-line interface, and you also can run it from the command line only (i.e., it is easy to incorporate into scripts for repeated analyses – to get info on the commanline options type clustalw –options and clustalw -help.)

Clustalx uses the same algorithms as clustalw. However, it has a much nicer interface, it displays information on the level of similarity, and it uses color in the alignment. Especially for amino acids the use of color greatly enhances the ability to recognize conservative replacements. Clustalx is available for different platforms at the ebi's ftp site (follow your platform, clustalx is stored in the clustalw folders)
Clustal reads and writes most formats used by different programs. The easiest format is the FASTA format:

Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73:237-244;
Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research ***22****, 4673-4680*

clustal

To align sequences clustal performs the following steps:

1) Pairwise distance calculation
2) Clustering analysis of the sequences
3) Iterated alignment of two most similar sequences or groups of sequences.

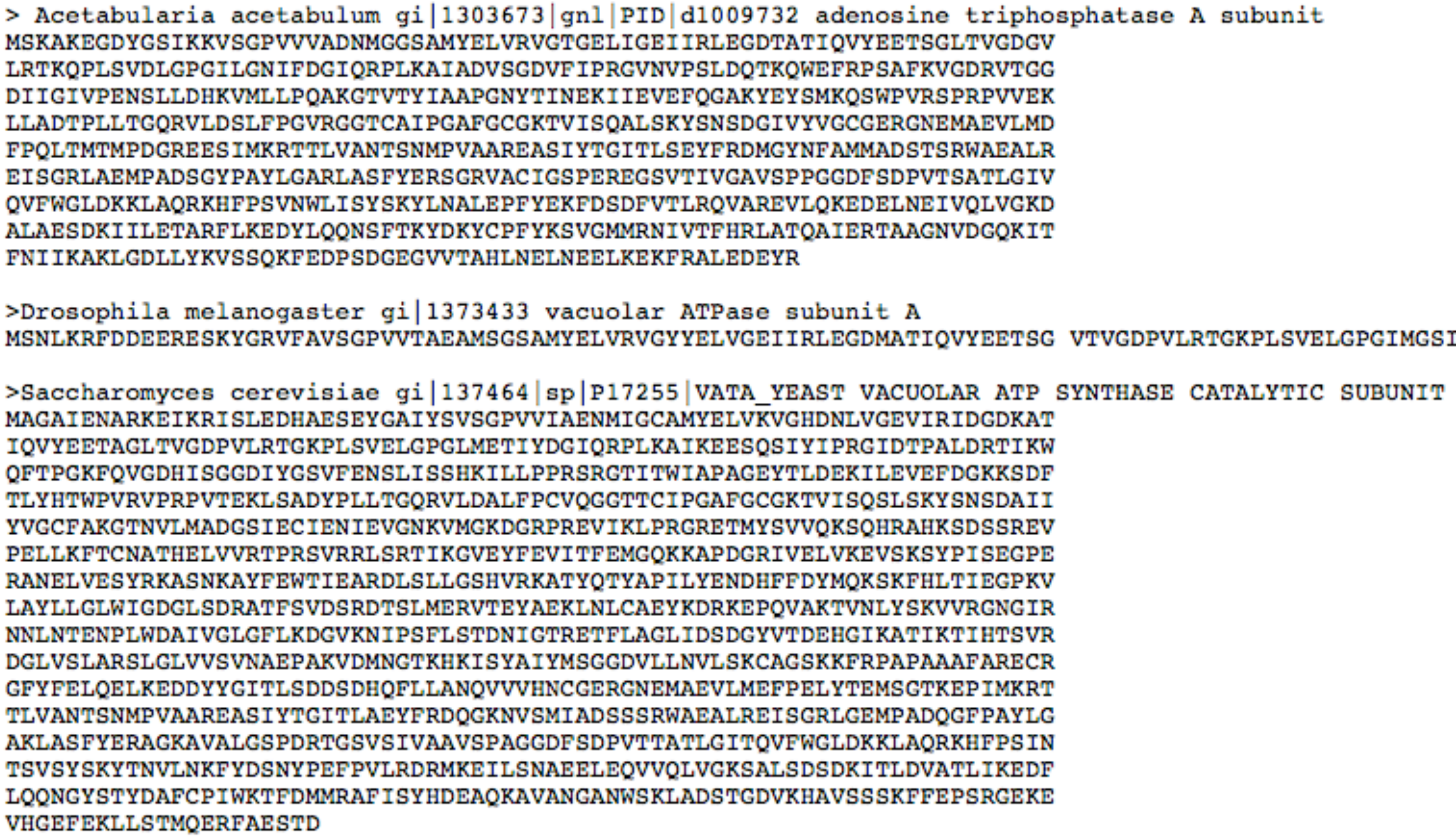
It is important to realize that the second step is the most important. The relationships found here will create a serious bias in the final alignment. The better your guide tree, the better your final alignment.

You can load a guide tree into clustal. This tree will then be used instead of the neighbor joining tree calculated by clustalw as a default. (The guide tree needs to be in normal parenthesis notation WITH branch lengths).

Sample input file   Sample output file

clustal

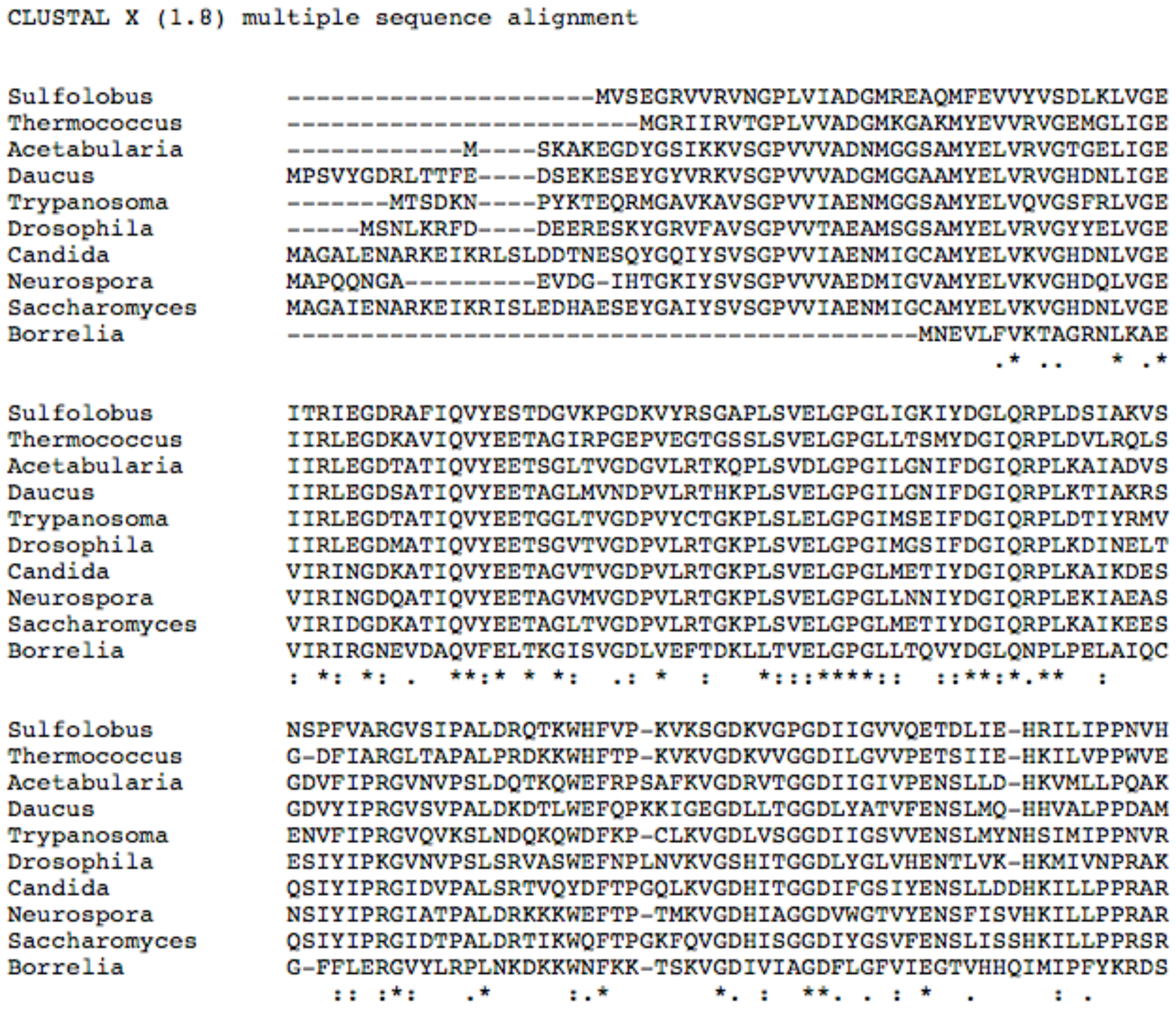
Sample input



```
> Acetabularia acetabulum gi|1303673|gnl|PID|d1009732 adenosine triphosphatase A subunit
MSKAKEGDYGSIKKVSGPVVVADNMGGSAMYELVRVGTGELIGEIIRLEGDTATIQVYEETSGLTVGDGV
LRTKQPLSVDLGPGILGNIFDGIQRPLKAIADVSGDVFIPRGVNVPSLDQTKQWEFRPSAFKVGDRVTGG
DIIGIVPENSLLDHKVMLLPQAKGTVTYIAAPGNYTINEKIIEVEFQGAKYEYSMKQSWPVRSPRPVVEK
LLADTPLLTGQRVLDSLFPGVRGGTCAIPGAFGCGKTVISQALSKYSNSDGIVYVGCGERGNEMAEVLMD
FPQLTMTMPDGREESIMKRTTLVANTSNMPVAAREASIYTGITLSEYFRDMGYNFAMMADSTSRWAEALR
EISGRLAEMPADSGYPAYLGARLASFYERSGRVACIGSPEREGSVTIVGAVSPPGGDFSDPVTSATLGIV
QVFWGLDKKLAQRKHFPSVNWLISYSKYLNALEPFYEKFDSDFVTLRQVAREVLQKEDELNEIVQLVGKD
ALAESDKIILETARFLKEDYLQQNSFTKYDKYCPFYKSVGMMRNIVTFHRLATQAIERTAAGNVDGQKIT
FNIIKAKLGDLLYKVSSQKFEDPSDGEGVVTAHLNELNEELKEKFRALEDEYR

>Drosophila melanogaster gi|1373433 vacuolar ATPase subunit A
MSNLKRFDDEERESKYGRVFAVSGPVVTAEAMSGSAMYELVRVGYYELVGEIIRLEGDMATIQVYEETSG VTVGDPVLRTGKPLSVELGPGIMGSI

>Saccharomyces cerevisiae gi|137464|sp|P17255|VATA_YEAST VACUOLAR ATP SYNTHASE CATALYTIC SUBUNIT
MAGAIENARKEIKRISLEDHAESEYGAIYSVSGPVVIAENMIGCAMYELVKVGHDNLVGEVIRIDGDKAT
IQVYEETAGLTVGDPVLRTGKPLSVELGPGLMETIYDGIQRPLKAIKEESQSIYIPRGIDTPALDRTIKW
QFTPGKFQVGDHISGGDIYGSVFENSLISSHKILLPPRSRGTITWIAPAGEYTLDEKILEVEFDGKKSDF
TLYHTWPVRVPRPVTEKLSADYPLLTGQRVLDALFPCVQGGTTCIPGAFGCGKTVISQSLSKYSNSDAII
YVGCFAKGTNVLMADGSIECIENIEVGNKVMGKDGRPREVIKLPRGRETMYSVVQKSQHRAHKSDSSREV
PELLKFTCNATHELVVRTPRSVRRLSRTIKGVEYFEVITFEMGQKKAPDGRIVELVKEVSKSYPISEGPE
RANELVESYRKASNKAYFEWTIEARDLSLLGSHVRKATYQTYAPILYENDHFFDYMQKSKFHLTIEGPKV
LAYLLGLWIGDGLSDRATFSVDSRDTSLMERVTEYAEKLNLCAEYKDRKEPQVAKTVNLYSKVVRGNGIR
NNLNTENPLWDAIVGLGFLKDGVKNIPSFLSTDNIGTRETFLAGLIDSDGYVTDEHGIKATIKTIHTSVR
DGLVSLARSLGLVVSVNAEPAKVDMNGTKHKISYAIYMSGGDVLLNVLSKCAGSKKFRPAPAAAFARECR
GFYFELQELKEDDYYGITLSDDSDHQFLLANQVVVHNCGERGNEMAEVLMEFPELYTEMSGTKEPIMKRT
TLVANTSNMPVAAREASIYTGITLAEYFRDQGKNVSMIADSSSRWAEALREISGRLGEMPADQGFPAYLG
AKLASFYERAGKAVALGSPDRTGSVSIVAAVSPAGGDFSDPVTTATLGITQVFWGLDKKLAQRKHFPSIN
TSVSYSKYTNVLNKFYDSNYPEFPVLRDRMKEILSNAEELEQVVQLVGKSALSDSDKITLDVATLIKEDF
LQQNGYSTYDAFCPIWKTFDMMRAFISYHDEAQKAVANGANWSKLADSTGDVKHAVSSSKFFEPSRGEKE
VHGEFEKLLSTMQERFAESTD
```

clustal

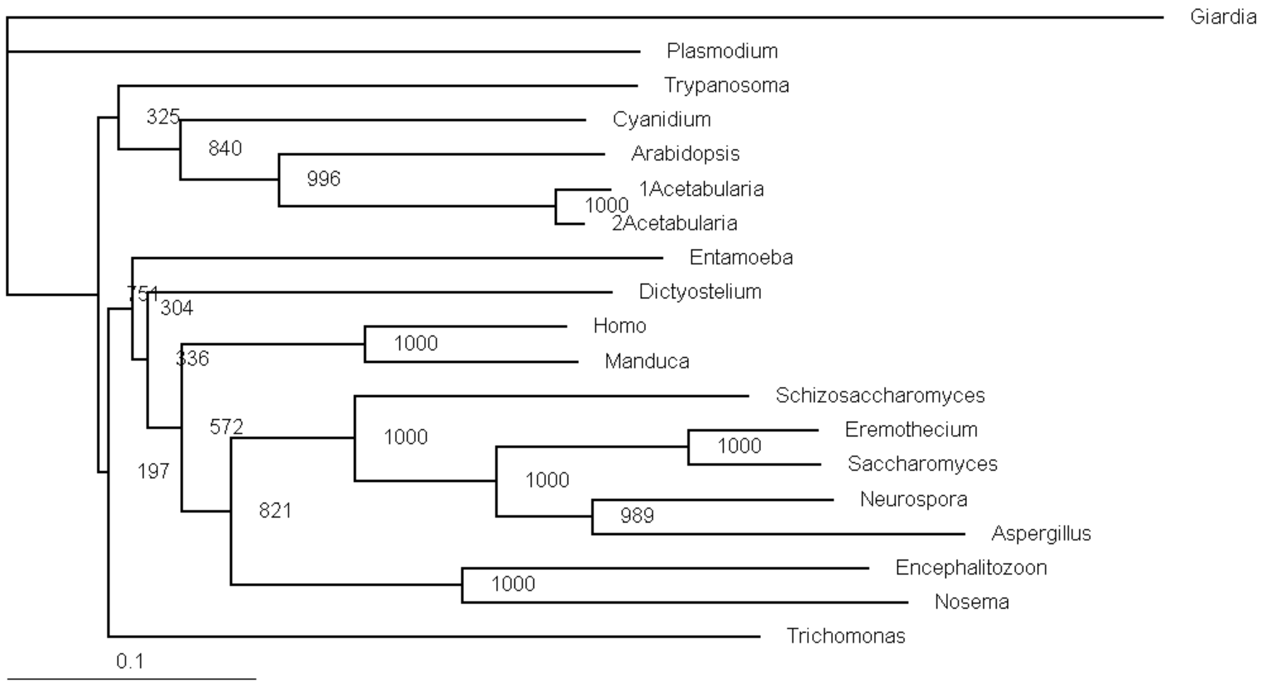
Sample output file



```
CLUSTAL X (1.8) multiple sequence alignment


Sulfolobus         ---------------------MVSEGRVVRVNGPLVIADGMREAQMFEVVYVSDLKLVGE
Thermococcus       ------------------------MGRIIRVTGPLVVADGMKGAKMYEVVRVGEMGLIGE
Acetabularia       ------------M----SKAKEGDYGSIKKVSGPVVVADNMGGSAMYELVRVGTGELIGE
Daucus             MPSVYGDRLTTFE----DSEKESEYGYVRKVSGPVVVADGMGGAAMYELVRVGHDNLIGE
Trypanosoma        -------MTSDKN----PYKTEQRMGAVKAVSGPVVIAENMGGSAMYELVQVGSFRLVGE
Drosophila         -----MSNLKRFD----DEERESKYGRVFAVSGPVVTAEAMSGSAMYELVRVGYYELVGE
Candida            MAGALENARKEIKRLSLDDTNESQYGQIYSVSGPVVIAENMIGCAMYELVKVGHDNLVGE
Neurospora         MAPQQNGA---------EVDG-IHTGKIYSVSGPVVVAEDMIGVAMYELVKVGHDQLVGE
Saccharomyces      MAGAIENARKEIKRISLEDHAESEYGAIYSVSGPVVIAENMIGCAMYELVKVGHDNLVGE
Borrelia           -------------------------------------------MNEVLFVKTAGRNLKAE
                                                                   .* ..   * .*

Sulfolobus         ITRIEGDRAFIQVYESTDGVKPGDKVYRSGAPLSVELGPGLIGKIYDGLQRPLDSIAKVS
Thermococcus       IIRLEGDKAVIQVYEETAGIRPGEPVEGTGSSLSVELGPGLLTSMYDGIQRPLDVLRQLS
Acetabularia       IIRLEGDTATIQVYEETSGLTVGDGVLRTKQPLSVDLGPGILGNIFDGIQRPLKAIADVS
Daucus             IIRLEGDSATIQVYEETAGLMVNDPVLRTHKPLSVELGPGILGNIFDGIQRPLKTIAKRS
Trypanosoma        IIRLEGDTATIQVYEETGGLTVGDPVYCTGKPLSLELGPGIMSEIFDGIQRPLDTIYRMV
Drosophila         IIRLEGDMATIQVYEETSGVTVGDPVLRTGKPLSVELGPGIMGSIFDGIQRPLKDINELT
Candida            VIRINGDKATIQVYEETAGVTVGDPVLRTGKPLSVELGPGLMETIYDGIQRPLKAIKDES
Neurospora         VIRINGDQATIQVYEETAGVMVGDPVLRTGKPLSVELGPGLLNNIYDGIQRPLEKIAEAS
Saccharomyces      VIRIDGDKATIQVYEETAGLTVGDPVLRTGKPLSVELGPGLMETIYDGIQRPLKAIKEES
Borrelia           VIRIRGNEVDAQVFELTKGISVGDLVEFTDKLLTVELGPGLLTQVYDGLQNPLPELAIQC
                   : *: *: .  **:* * *:  .: *  :   *:::****::  ::**:*.**  :

Sulfolobus         NSPFVARGVSIPALDRQTKWHFVP-KVKSGDKVGPGDIIGVVQETDLIE-HRILIPPNVH
Thermococcus       G-DFIARGLTAPALPRDKKWHFTP-KVKVGDKVVGGDILGVVPETSIIE-HKILVPPWVE
Acetabularia       GDVFIPRGVNVPSLDQTKQWEFRPSAFKVGDRVTGGDIIGIVPENSLLD-HKVMLLPQAK
Daucus             GDVYIPRGVSVPALDKDTLWEFQPKKIGEGDLLTGGDLYATVFENSLMQ-HHVALPPDAM
Trypanosoma        ENVFIPRGVQVKSLNDQKQWDFKP-CLKVGDLVSGGDIIGSVVENSLMYNHSIMIPPNVR
Drosophila         ESIYIPKGVNVPSLSRVASWEFNPLNVKVGSHITGGDLYGLVHENTLVK-HKMIVNPRAK
Candida            QSIYIPRGIDVPALSRTVQYDFTPGQLKVGDHITGGDIFGSIYENSLLDDHKILLPPRAR
Neurospora         NSIYIPRGIATPALDRKKKWEFTP-TMKVGDHIAGGDVWGTVYENSFISVHKILLPPRAR
Saccharomyces      QSIYIPRGIDTPALDRTIKWQFTPGKFQVGDHISGGDIYGSVFENSLISSHKILLPPRSR
Borrelia           G-FFLERGVYLRPLNKDKKWNFKK-TSKVGDIVIAGDFLGFVIEGTVHHQIMIPFYKRDS
                      :: :*:   .*     :.*       *. :  **. . : *  .     : .
```

example on bbcxsv1.biotech.uconn.edu

- •calculate multiple sequence alignment
- •go through options
- •do tree / tree options
  (positions with gaps, correct for multiple subs,
  support values to nodes, if you want to use treeview)



One way to draw trees on the road is phylodendron

Clustal also reads aligned sequences. If you input aligned sequences you can go directly to the tree section.

!! Be careful if you make a mistake, and the sequences are not aligned, your tree will look strange!!

!!! ALWAYS CHECK YOUR ALIGNMENT!!!

Also be careful when using the ignore positions with gaps option – there might not be many positions left.

Clustal is much better than its reputation. It is doing a great job in handling gaps, especially terminal gaps, and it makes good use of different substitution matrices, and the empirical correction for multiple substitutions is better than many other programs.

tcoffee

TCOFFEE extracts reliably aligned positions from several multiple or pairwise sequence alignments. It requires more thought and attention from the user than clustalw, but it helps to focus further analyses on those sites that are reliably aligned. A web interface is here.

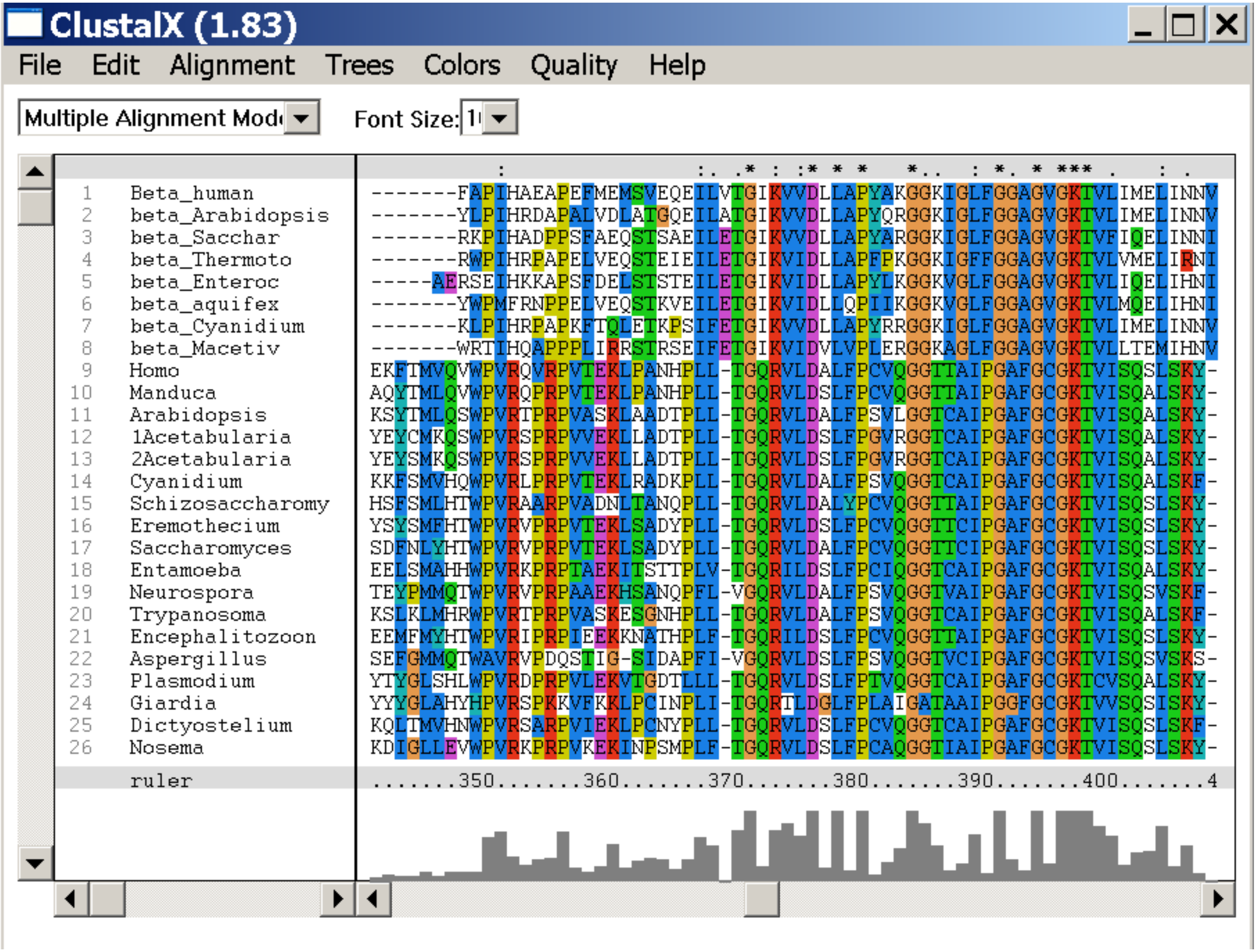
muscle

If you have very large datasets muscle is the way to go. It is fast, takes fasta formatted sequences as input file, and has a refinement option, that does an excellent job cleaning up around gaps.

The muscle home page is here , the manual is here
Muscle allows also allows profile alignments.

```
muscle -in VatpA.fa -out VatpA.afa
muscle -in VatpA.afa -out VatpA.rafa -refine
muscle -in beta.fa -out beta.afa
muscle -in beta.afa -out beta.rafa -refine
muscle -profile -in1 beta.rafa -in2
VatpA.rafa -out Abeta.afa
muscle –refine –in Abeta.afa –out Abeta.rafa
```

muscle alignment

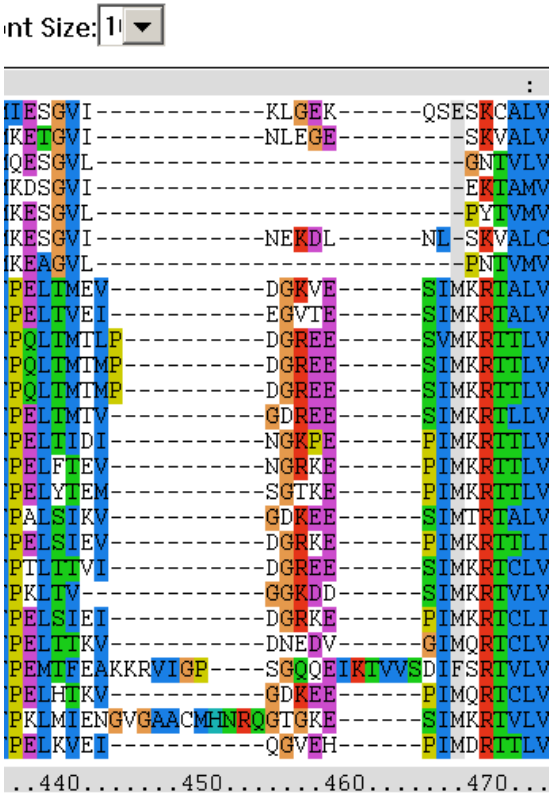


Multiple Alignment Mod

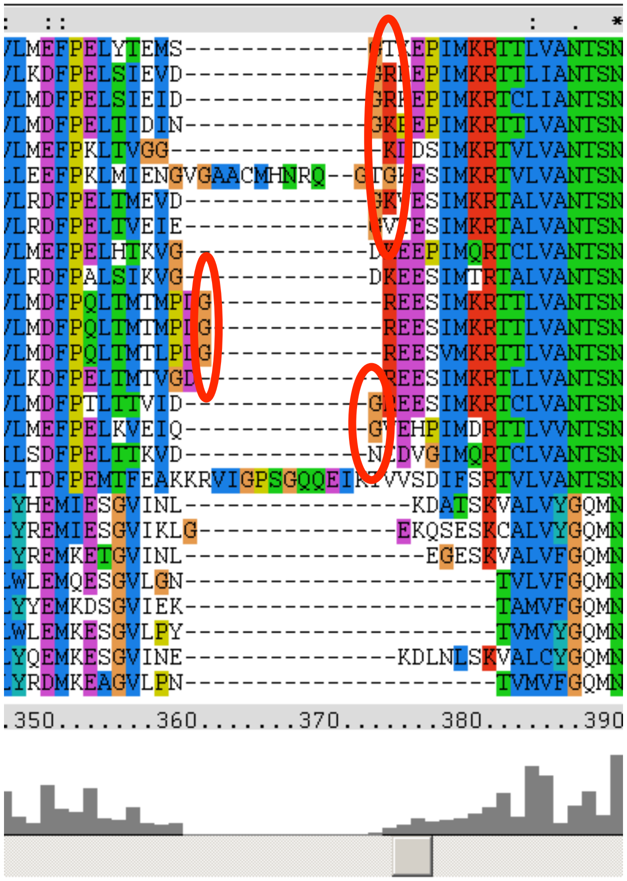
Font Size: 1(

```
                                    :               :. .* : :* * *   *..  : *. * *** .   : .
 1   Beta_human           -------FAPIHAEAPEFMEMSVEQEILVTGIKVVDLLAPYAKGGKIGLFGGAGVGKTVLIMELINNV
 2   beta_Arabidopsis     -------YLPIHRDAPALVDLATGQEILATGIKVVDLLAPYQRGGKIGLFGGAGVGKTVLIMELINNV
 3   beta_Sacchar         -------RKPIHADPPSFAEQSTSAEILETGIKVVDLLAPYARGGKIGLFGGAGVGKTVFIQELINNI
 4   beta_Thermoto        -------RWPIHRPAPELVEQSTEIEILETGIKVIDLLAPFPKGGKIGFFGGAGVGKTVLVMELIRNI
 5   beta_Enteroc         -----AERSEIHKKAPSFDELSTSTEILETGIKVIDLLAPYLKGGKVGLFGGAGVGKTVLIQELIHNI
 6   beta_aquifex         -------YWPMFRNPPELVEQSTKVEILETGIKVIDLLQPIIKGGKVGLFGGAGVGKTVLMQELIHNI
 7   beta_Cyanidium       -------KLPIHRPAPKFTQLETKPSIFETGIKVVDLLAPYRRGGKIGLFGGAGVGKTVLIMELINNV
 8   beta_Macetiv         -------WRTIHQAPPPLIRRSTRSEIFETGIKVIDVLVPLERGGKAGLFGGAGVGKTVLLTEMIHNV
 9   Homo                 EKFTMVQVWPVRQVRPVTEKLPANHPLL-TGQRVLDALFPCVQGGTTAIPGAFGCGKTVISQSLSKY-
10   Manduca              AQYTMLQVWPVRQPRPVTEKLPANHPLL-TGQRVLDSLFPCVQGGTTAIPGAFGCGKTVISQALSKY-
11   Arabidopsis          KSYTMLQSWPVRTPRPVASKLAADTPLL-TGQRVLDALFPSVLGGTCAIPGAFGCGKTVISQALSKY-
12   1Acetabularia        YEYCMKQSWPVRSPRPVVEKLLADTPLL-TGQRVLDSLFPGVRGGTCAIPGAFGCGKTVISQALSKY-
13   2Acetabularia        YEYSMKQSWPVRSPRPVVEKLLADTPLL-TGQRVLDSLFPGVRGGTCAIPGAFGCGKTVISQALSKY-
14   Cyanidium            KKFSMVHQWPVRLPRPVTEKLRADKPLL-TGQRVLDALFPSVQGGTCAIPGAFGCGKTVISQALSKF-
15   Schizosaccharomy     HSFSMLHTWPVRAARPVADNLTANQPLL-TGQRVLDALYPCVQGGTTAIPGAFGCGKTVISQSLSKY-
16   Eremothecium         YSYSMFHTWPVRVPRPVTEKLSADYPLL-TGQRVLDSLFPCVQGGTTCIPGAFGCGKTVISQSLSKY-
17   Saccharomyces        SDFNLYHTWPVRVPRPVTEKLSADYPLL-TGQRVLDALFPCVQGGTTCIPGAFGCGKTVISQSLSKY-
18   Entamoeba            EELSMAHHWPVRKPRPTAEKITSTTPLV-TGQRILDSLFPCIQGGTCAIPGAFGCGKTVISQALSKY-
19   Neurospora           TEYPMMQTWPVRVPRPAAEKHSANQPFL-VGQRVLDALFPSVQGGTVAIPGAFGCGKTVISQSVSKF-
20   Trypanosoma          KSLKLMHRWPVRTPRPVASKESGNHPLL-TGQRVLDALFPSVQGGTCAIPGAFGCGKTVISQALSKF-
21   Encephalitozoon      EEMFMYHTWPVRIPRPIEEKKNATHPLF-TGQRILDSLFPCVQGGTTAIPGAFGCGKTVISQSLSKY-
22   Aspergillus          SEFGMMQTWAVRVPDQSTIG-SIDAPFI-VGQRVLDSLFPSVQGGTVCIPGAFGCGKTVISQSVSKS-
23   Plasmodium           YTYGLSHLWPVRDPRPVLEKVTGDTLLL-TGQRVLDSLFPTVQGGTCAIPGAFGCGKTCVSQALSKY-
24   Giardia              YYYGLAHYHPVRSPKKVFKKLPCINPLI-TGQRTLDGLFPLAIGATAAIPGGFGCGKTVVSQSISKY-
25   Dictyostelium        KQLTMVHNWPVRSARPVIEKLPCNYPLL-TGQRVLDSLFPCVQGGTCAIPGAFGCGKTVISQSLSKF-
26   Nosema               KDIGLLEVWPVRKPRPVKEKINPSMPLF-TGQRVLDSLFPCAQGGTIAIPGAFGCGKTVISQSLSKY-

     ruler                .......350.......360.......370.......380.......390.......400.......4
```

muscle *vs* clustal



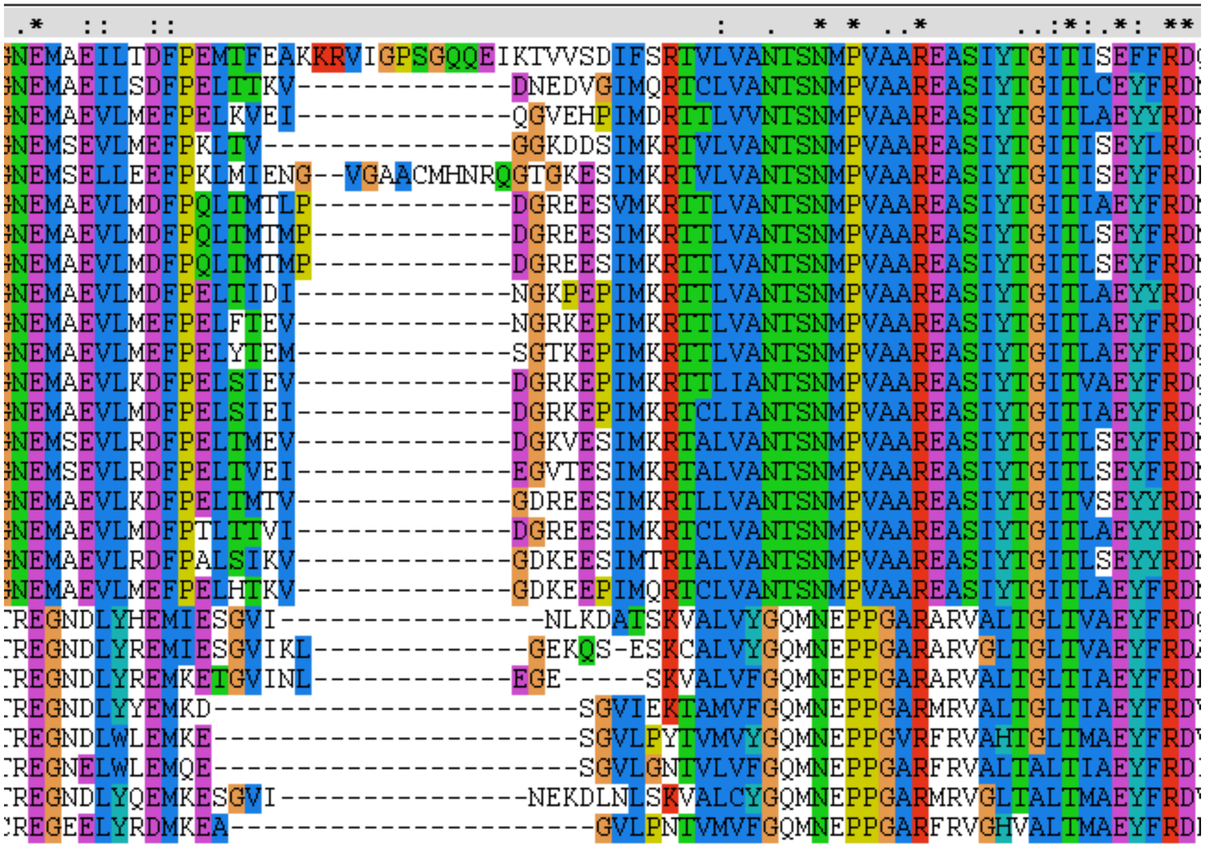
```
                                     :
IIESGVI-----------KLGEK------QSESKCALV
IKETGVI-----------NLEGE---------SKVALV
IQESGVL-------------------------GNTVLV
IKDSGVI-------------------------EKTAMV
IKESGVL-------------------------PYTVMV
IKESGVI-----------NEKDL------NL-SKVALC
IKEAGVL-------------------------PNTVMV
PELTMEV-----------DGKVE------SIMKRTALV
PELTVEI-----------EGVTE------SIMKRTALV
PQLTMTLP----------DGREE------SVMKRTTLV
PQLTMTMP----------DGREE------SIMKRTTLV
PQLTMTMP----------DGREE------SIMKRTTLV
PELTMTV-----------GDREE------SIMKRTLLV
PELTIDI-----------NGKPE------PIMKRTTLV
PELFTEV-----------NGRKE------PIMKRTTLV
PELYTEM-----------SGTKE------PIMKRTTLV
PALSIKV-----------GDKEE------SIMTRTALV
PELSIEV-----------DGRKE------PIMKRTTLI
PTLTTVI-----------DGREE------SIMKRTCLV
PKLTV-------------GGKDD------SIMKRTVLV
PELSIEI-----------DGRKE------PIMKRTCLI
PELTTKV-----------DNEDV------GIMQRTCLV
PEMTFEAKKRVIGP----SGQQEIKTVVSDIFSRTVLV
PELHTKV-----------GDKEE------PIMQRTCLV
PKLMIENGVGAACMHNRQGTGKE------SIMKRTVLV
PELKVEI-----------QGVEH------PIMDRTTLV
..440.......450.......460.......470...
```



```
:  ::                               :  .  *
VLMEFPELYTEMS-------------GTKEPIMKRTTLVANTSN
VLKDFPELSIEVD-------------GRKEPIMKRTTLIANTSN
VLMDFPELSIEID-------------GRKEPIMKRTCLIANTSN
VLMDFPELTIDIN-------------GKPEPIMKRTTLVANTSN
VLMEFPKLTVGG---------------KDDSIMKRTVLVANTSN
LLEEFPKLMIENGVGAACMHNRQ--GTGKESIMKRTVLVANTSN
VLRDFPELTMEVD-------------GKVESIMKRTALVANTSN
VLRDFPELTVEIE-------------GVTESIMKRTALVANTSN
VLMEFPELHTKVG-------------DKEEPIMQRTCLVANTSN
VLRDFPALSIKVG-------------DKEESIMTRTALVANTSN
VLMDFPQLTMTMPIG------------REESIMKRTTLVANTSN
VLMDFPQLTMTMPIG------------REESIMKRTTLVANTSN
VLMDFPQLTMTLPIG------------REESVMKRTTLVANTSN
VLKDFPELTMTVGD-------------REESIMKRTLLVANTSN
VLMDFPTLTTVID-------------GREESIMKRTCLVANTSN
VLMEFPELKVEIQ-------------GVEHPIMDRTTLVVNTSN
ILSDFPELTTKVD-------------NEDVGIMQRTCLVANTSN
ILTDFPEMTFEAKKRVIGPSGQQEIKTVVSDIFSRTVLVANTSN
LYHEMIESGVINL----------------KDATSKVALVYGQMN
LYREMIESGVIKLG--------------EKQSESKCALVYGQMN
LYREMKETGVINL-----------------EGESKVALVFGQMN
LWLEMQESGVLGN----------------------TVLVFGQMN
LYYEMKDSGVIEK----------------------TAMVFGQMN
LWLEMKESGVLPY----------------------TVMVYGQMN
LYQEMKESGVINE---------------KDLNLSKVALCYGQMN
LYRDMKEAGVLPN----------------------TVMVFGQMN
.350.......360.......370.......380.......390
```

more on alignment programs (statalign, pileup, SAM) here

the same region using tcoffee with default settings



```
 .*  ::  ::                                :  .  * * ..*     ..:*:.*: **
GNEMAEILTDFPEMTFEAKKRVIGPSGQQEIKTVVSDIFSRTVLVANTSNMPVAAREASIYTGITISEFFRD
GNEMAEILSDFPELTTKV-------------DNEDVGIMQRTCLVANTSNMPVAAREASIYTGITLCEYFRD
GNEMAEVLMEFPELKVEI-------------QGVEHPIMDRTTLVVNTSNMPVAAREASIYTGITLAEYYRD
GNEMSEVLMEFPKLTV---------------GGKDDSIMKRTVLVANTSNMPVAAREASIYTGITISEYLRD
GNEMSELLEEFPKLMIENG--VGAACMHNRQGTGKESIMKRTVLVANTSNMPVAAREASIYTGITISEYFRD
GNEMAEVLMDFPQLTMTLP------------DGREESVMKRTTLVANTSNMPVAAREASIYTGITIAEYFRD
GNEMAEVLMDFPQLTMTMP------------DGREESIMKRTTLVANTSNMPVAAREASIYTGITLSEYFRD
GNEMAEVLMDFPQLTMTMP------------DGREESIMKRTTLVANTSNMPVAAREASIYTGITLSEYFRD
GNEMAEVLMDFPELTIDI-------------NGKPEPIMKRTTLVANTSNMPVAAREASIYTGITLAEYYRD
GNEMAEVLMEFPELFTEV-------------NGRKEPIMKRTTLVANTSNMPVAAREASIYTGITLAEYFRD
GNEMAEVLMEFPELYTEM-------------SGTKEPIMKRTTLVANTSNMPVAAREASIYTGITLAEYFRD
GNEMAEVLKDFPELSIEV-------------DGRKEPIMKRTTLIANTSNMPVAAREASIYTGITVAEYFRD
GNEMAEVLMDFPELSIEI-------------DGRKEPIMKRTCLIANTSNMPVAAREASIYTGITIAEYFRD
GNEMSEVLRDFPELTMEV-------------DGKVESIMKRTALVANTSNMPVAAREASIYTGITLSEYFRD
GNEMSEVLRDFPELTVEI-------------EGVTESIMKRTALVANTSNMPVAAREASIYTGITLSEYFRD
GNEMAEVLKDFPELTMTV-------------GDREESIMKRTLLVANTSNMPVAAREASIYTGITVSEYYRD
GNEMAEVLMDFPTLTTVI-------------DGREESIMKRTCLVANTSNMPVAAREASIYTGITLAEYYRD
GNEMAEVLRDFPALSIKV-------------GDKEESIMTRTALVANTSNMPVAAREASIYTGITLSEYYRD
GNEMAEVLMEFPELHTKV-------------GDKEEPIMQRTCLVANTSNMPVAAREASIYTGITLAEYFRD
TREGNDLYHEMIESGVI----------------NLKDATSKVALVYGQMNEPPGARARVALTGLTVAEYFRD
TREGNDLYREMIESGVIKL-------------GEKQS-ESKCALVYGQMNEPPGARARVGLTGLTVAEYFRD
TREGNDLYREMKETGVINL------------EGE-----SKVALVFGQMNEPPGARARVALTGLTIAEYFRD
TREGNDLYYEMKD----------------------SGVIEKTAMVFGQMNEPPGARMRVALTGLTIAEYFRD
TREGNDLWLEMKE----------------------SGVLPYTVMVYGQMNEPPGVRFRVAHTGLTMAEYFRD
TREGNELWLEMQE----------------------SGVLGNTVLVFGQMNEPPGARFRVALTALTIAEYFRD
TREGNDLYQEMKESGVI---------------NEKDLNLSKVALCYGQMNEPPGARMRVGLTALTMAEYFRD
CREGEELYRDMKEA----------------------GVLPNTVMVFGQMNEPPGARFRVGHVALTMAEYFRD
```

more on alignment programs (statalign, pileup, SAM) here

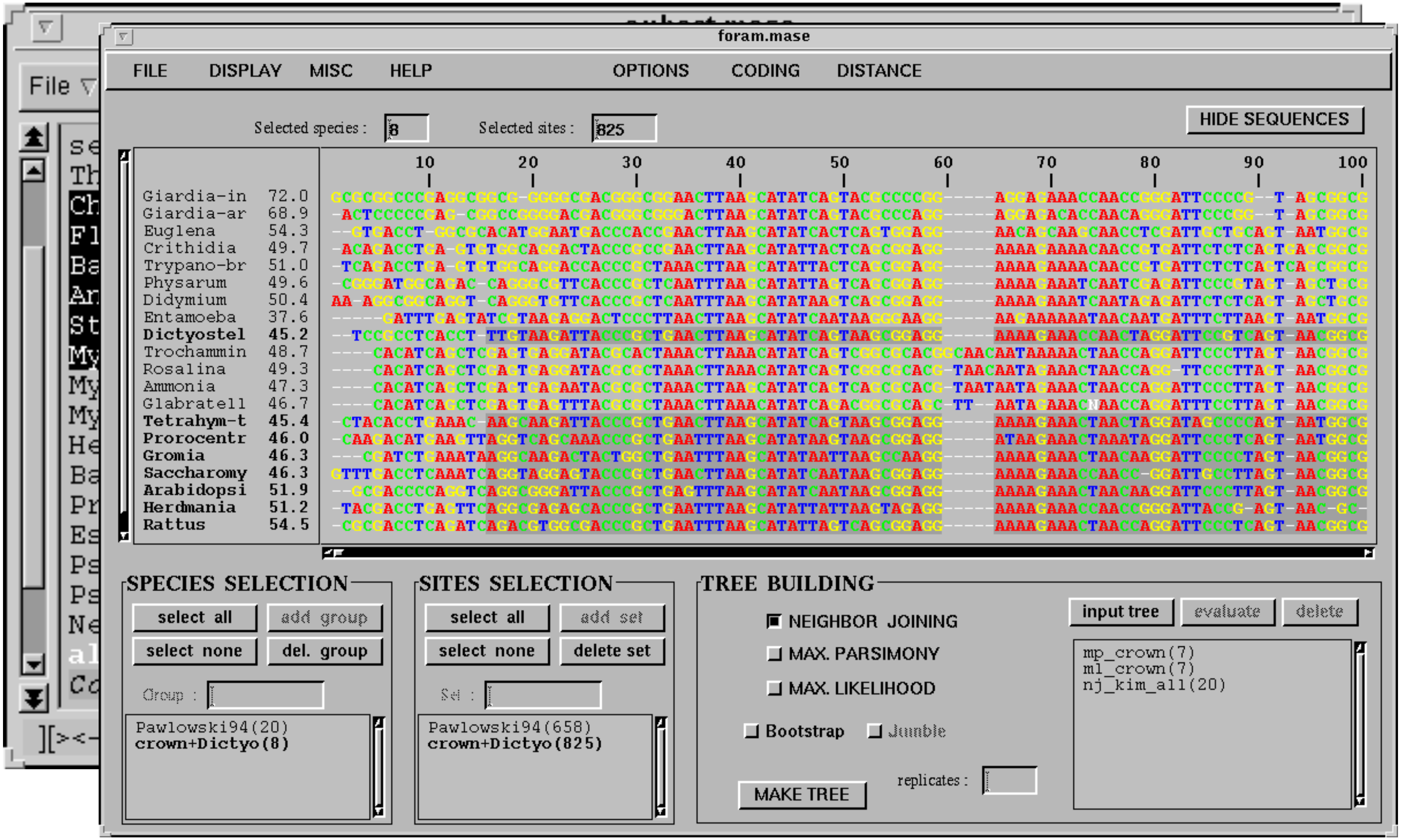
Sequence editors and viewers

Jalview Homepage, Description
Jalview as Java Web Start Application
(other JAVA applications are here)

Jalview is easy to install and run.
Test file is here (ATPase subunits)
(Intro to ATPases: 1bmf in spdbv)
(gif of rotation here, movies of the rotation are here and here)
(Load all.txt into Jalview,
colou**r** options,
mouse use,
PID tree,
Principle component analysis -> sequence space)
More on sequence space here

seaview – phylo_win

Another useful multiple alignment editor is **seaview, the companion sequence editor to phylo_win. It runs on PC and most unix flavors, and is the easiest way to get alignments into phylo_win.**



Selected species : 8 Selected sites : 825

```
                             10        20        30        40        50        60        70        80        90       100
                              |         |         |         |         |         |         |         |         |         |
Giardia-in   72.0  GCGCGGCCCGAGGCGGCG-GGGGCGACGGGCGGAACTTAAGCATATCAGTACGCCCCGG-----AGGAGAAACCAACCGGGATTCCCCG--T-AGCGGCG
Giardia-ar   68.9  -ACTCCCCCGAG-CGGCCGGGGACGACGGGCGGGACTTAAGCATATCAGTACGCCCAGG-----AGGAGACACCAACAGGGATTCCCGG--T-AGCGGCG
Euglena      54.3  --GTGACCT-GGCGCACATGGAATGACCCACCGAACTTAAGCATATCACTCAGTGGAGG-----AACAGCAAGCAACCTCGATTGCTGCAGT-AATGGCG
Crithidia    49.7  -ACAGACCTGA-GTGTGGCAGGACTACCCGCCGAACTTAAGCATATTACTCAGCGGAGG-----AAAAGAAAACAACCGTGATTCTCTCAGTGAGCGGCG
Trypano-br   51.0  -TCAGACCTGA-GTGTGGCAGGACCACCCGCTAAACTTAAGCATATTACTCAGCGGAGG-----AAAAGAAAACAACCGTGATTCTCTCAGTCAGCGGCG
Physarum     49.6  -CGGGATGGCAGAC-CAGGGCGTTCACCCGCTCAATTTAAGCATATTAGTCAGCGGAGG-----AAAAGAAATCAATCGAGATTCCCGTAGT-AGCTGCG
Didymium     50.4  AA-AGGCGGCAGGT-CAGGGTGTTCACCCGCTCAATTTAAGCATATAAGTCAGCGGAGG-----AAAAGAAATCAATAGAGATTCTCTCAGT-AGCTGCG
Entamoeba    37.6  -----GATTTGAGTATCGTAAGAGGACTCCCTTAACTTAAGCATATCAATAAGGGAAGG-----AAGAAAAAATAACAATGATTTCTTAAGT-AATGGCG
Dictyostel   45.2  --TCCGCCTCACCT-TTGTAAGATTACCCGCTGAACTTAAGCATATCAGTAAGCGGAGG-----AAAAGAAACCAACTAGGATTCCGTCAGT-AACGGCG
Trochammin   48.7  ----CACATCAGCTCGAGTGAGGATACGCACTAAACTTAAACATATCAGTCGGCGCACGGCAACAATAAAAACTAACCAGGATTCCCTTAGT-AACGGCG
Rosalina     49.3  ----CACATCAGCTCGAGTGAGGATACGCGCTAAACTTAAACATATCAGTCGGCGCACG-TAACAATAGAAACTAACCAGG-TTCCCTTAGT-AACGGCG
Ammonia      47.3  ----CACATCAGCTCGAGTGAGAATACGCGCTAAACTTAAGCATATCAGTCAGCGCACG-TAATAATAGAAACTAACCAGGATTCCCTTAGT-AACGGCG
Glabratell   46.7  ----CACATCAGCTCGAGTGAGTTTACGCGCTAAACTTAAACATATCAGACGGCGCAGC-TT--AATAGAAACNAACCAGGATTTCCTTAGT-AACGGCG
Tetrahym-t   45.4  -CTACACCTGAAAC-AAGCAAGATTACCCGCTGAACTTAAGCATATCAGTAAGCGGAGG-----AAAAGAAACTAACTAGGATAGCCCCAGT-AATGGCG
Prorocentr   46.0  -CAAGACATGAAGTTAGGTCAGCAAACCCGCTGAATTTAAGCATATAAGTAAGCGGAGG-----ATAAGAAACTAAATAGGATTCCCTCAGT-AATGGCG
Gromia       46.3  ---CGATCTGAAATAAGGCAAGACTACTGGCTGAATTTAAGCATATAATTAAGCCAAGG-----AAAAGAAACTAACAAGGATTCCCCTAGT-AACGGCG
Saccharomy   46.3  GTTTGACCTCAAATCAGGTAGGAGTACCCGCTGAACTTAAGCATATCAATAAGCGGAGG-----AAAAGAAACCAACC-GGATTGCCTTAGT-AACGGCG
Arabidopsi   51.9  --GCGACCCCAGGTCAGGCGGGATTACCCGCTGAGTTTAAGCATATCAATAAGCGGAGG-----AAAAGAAACTAACAAGGATTCCCTTAGT-AACGGCG
Herdmania    51.2  -TACGACCTGAGTTCAGGCGAGAGCACCCGCTGAATTTAAGCATATTATTAAGTAGAGG-----AAAAGAAACCAACCGGGATTACCG-AGT-AAC-GC-
Rattus       54.5  -CGCGACCTCAGATCAGACGTGGCGACCCGCTGAATTTAAGCATATTAGTCAGCGGAGG-----AAAAGAAACTAACCAGGATTCCCTCAGT-AACGGCG
```

select all
select none

Group :

Pawlowski94(20)
crown+Dictyo(8)

select all
select none

Set :

Pawlowski94(658)
crown+Dictyo(825)

☒ NEIGHBOR JOINING
☐ MAX. PARSIMONY
☐ MAX. LIKELIHOOD

☐ Bootstrap ☐ Jumble

replicates :

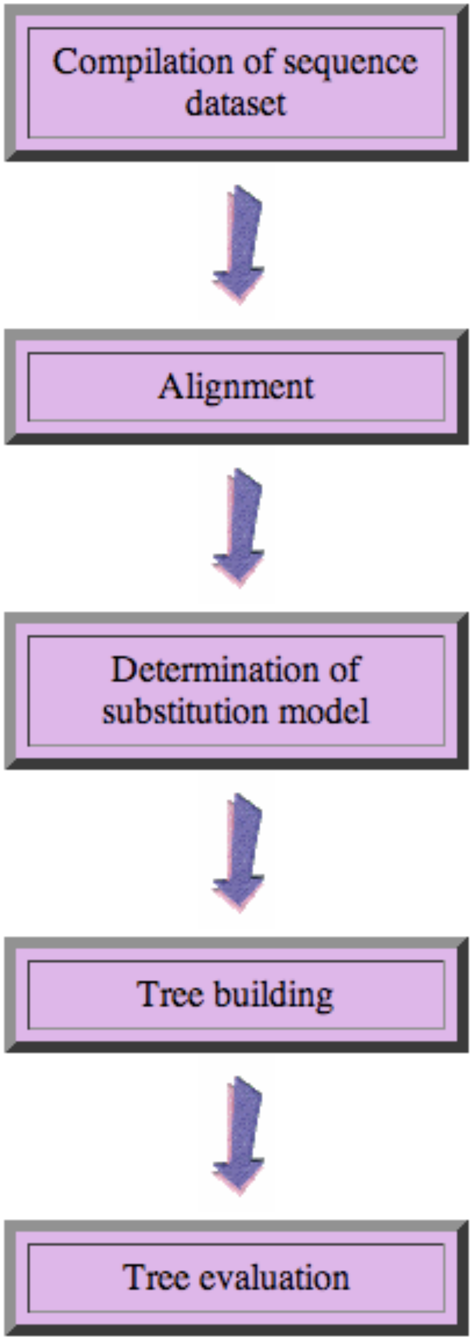
mp_crown(7)
ml_crown(7)
nj_kim_all(20)

Steps of the phylogenetic analysis

Phylogenetic analysis is an inference of evolutionary relationships between organisms. Phylogenetics tries to answer the question "How did groups of organisms come into existence?"

Those relationships are usually represented by tree-like diagrams.

Note: the assumption of a tree-like process of evolution is controversial!



Compilation of sequence dataset

Alignment

Determination of substitution model

Tree building

Tree evaluation

Phylogenetic reconstruction - **How**

Distance analyses

calculate pairwise distances
(different distance measures, correction for multiple hits, correction for codon bias)

make distance matrix (table of pairwise corrected distances)

calculate tree from distance matrix

i) using optimality criterion
(e.g.: smallest error between distance matrix
and distances in tree, or use
ii) algorithmic approaches (UPGMA or neighbor joining) B)

Phylogenetic reconstruction - How

Parsimony analyses

find that tree that explains sequence data with minimum number of substitutions
(tree includes hypothesis of sequence at each of the nodes)

Maximum Likelihood analyses

given a model for sequence evolution, find the tree that has the highest probability under this model.
This approach can also be used to successively refine the model.

Bayesian statistics use ML analyses to calculate posterior probabilities for trees, clades and evolutionary parameters. Especially MCMC approaches have become very popular in the last year, because they allow to estimate evolutionary parameters (e.g., which site in a virus protein is under positive selection), without assuming that one actually knows the "true" phylogeny.

more alignment programs: statalign

statalign from Jeff Thorne deserves more attention than it receives. Especially for divergent sequences the initial pairwise alignment usually determines the ultimate result of the phylogenetic reconstruction.

Statalign solves this problem by not calculating a multiple sequence alignment, rather it spends a lot of computational power to calculate pairwise alignments and it extract distances (and their potential error) from these pairwise alignments and then uses these in a distance pased reconstruction. The errors from the individual distances are used to generate bootstrap samples for the distance matrices.

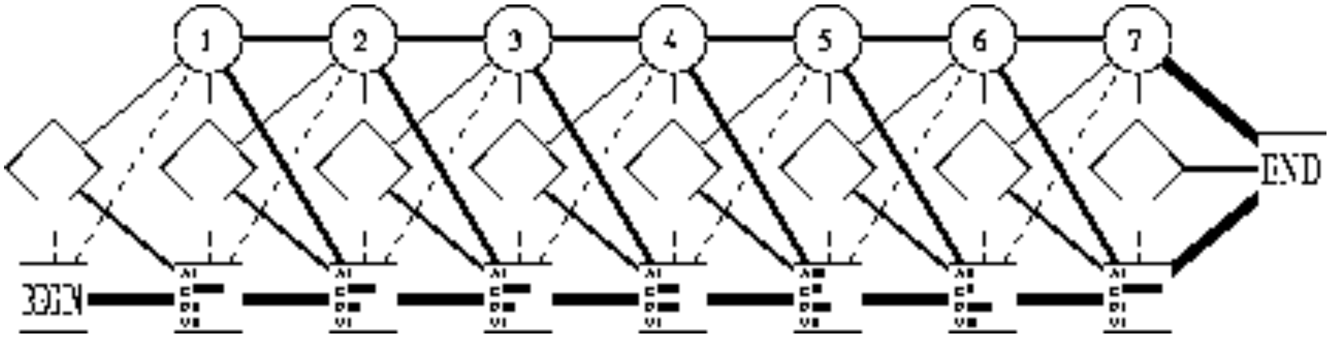
More at *Thorne JL, Kishino H (1992) Freeing phylogenies from artifacts of alignment. Mol Bio Evol 9:1148-1162*

statalign is available in several software archives (e.g. here), the readme file has plenty of information.

more alignment programs: SAM

SAM (sequence alignment and modeling system) by Richard Hughey, Anders Krogh, Christian Barrett, & Leslie Grate at UCSC.
http://www.cse.ucsc.edu/research/compbio/sam.html

The input consists of a multiple sequence file (aligned or not aligned) in FASTA format. The program uses secondary structure predictions, neighboring sites, etc. to place gaps. The program can be accessed through the www and run at UCSC



A linear hidden Markov model is a sequence of nodes, each corresponding to a column in a multiple alignment. In our HMMs, each node has a match state (square), insert state (diamond) and delete state (circle). Each sequence uses a series of these states to traverse the model from start to end. Using a match state indicates that the sequence has a character in that column, while using a delete state indicates that the sequence does not. Insert states allow sequences to have additional characters *between* columns. In many ways, these models correspond to profiles.

challenge:

Often one wants to build families of homologous proteins extracted from genomes. One way to do so is to find reciprocal best hits.
Tools:
The script *blastall.pl* takes the genomes indicted in the first line and calculates all possible genome against genome searches.

This script *simple_rbh_pairs.pl* takes two blastall searches (genome A versus genome B) in -m8 format and listing only the top scoring blast hit for each query) and writes the GI numbers of reciprocal best hits into a table.

The script *run_pairs.pl* runs all possible pairwise extractions of RBHs

Task: write a script that combines the pairwise tables keeping only those families that have a strict reciprocal best blast hit relationship in all genomes.