MCB 5472

Sequence alignment

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

Asignments from last week

Geneplot

In a perfect world you do not want to plot gi numbers but positions in a genome. The script addnumnuc.pl adds the nucleotide position of the ORF (the central one) to the beginning of the annotation line.

.ptt files

Available on the ftp server at NCBI or each chromosome. E.g.

Fervidobacterium nodosum Rt17-B1, complete genome - 1..1948941
1750 proteins

Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product
43..1377	+	444	154248706	-	Fnod_0001	-	-	chromosomal replicatio
1453..1635	+	60	154248707	-	Fnod_0002	-	-	4Fe-4S ferredoxin iron-
1976..3829	+	617	154248708	-	Fnod_0003	-	-	hypothetical protein
3826..4926	+	366	154248709	-	Fnod_0004	-	-	basic membrane lipopr
5136..6701	+	521	154248710	-	Fnod_0005	-	-	ABC transporter relate
6698..7732	+	344	154248711	-	Fnod_0006	-	-	inner-membrane transl
7729..8688	+	319	154248712	-	Fnod_0007	-	-	inner-membrane transl
8734..9132	+	132	154248713	-	Fnod_0008	-	-	protein of unknown fu
9261..9617	+	118	154248714	-	Fnod_0009	-	-	hypothetical protein
9745..10020	+	91	154248715	-	Fnod_0010	-	-	histone family protein
10098..11342	-	414	154248716	-	Fnod_0011	-	-	metal depe
11361..13514	-	717	154248717	-	Fnod_0012	-	-	hypothetic
13511..14161	-	216	154248718	-	Fnod_0013	-	-	hypothetic
14158..15102	-	314	154248719	-	Fnod_0014	-	-	putative m
15115..15969	-	284	154248720	-	Fnod_0015	-	-	putative ad
16022..17008	-	328	154248721	-	Fnod_0016	-	-	putative C
17156..17566	+	136	154248722	-	Fnod_0017	-	-	protein of u
17594..19282	+	562	154248723	-	Fnod_0018	-	-	sigma54 sp
19623..19859	+	78	154248724	-	Fnod_0019	-	-	hypothetic
19856..20074	+	72	154248725	-	Fnod_0020	-	-	hypothetic
20095..20289	+	64	154248726	-	Fnod_0021	-	-	hypothetic

.....

Addnumnuc.pl part 1

```
#!/usr/bin/perl -w
#decided to have input file entered in command line
#call program followed by genome name.
#the program assumes that a file with the extensions ptt and faa exist in the same dirctory.
#####INPUT Name of multiple seq file containing ORF of genome, open file and assign IN filehandle #
unless(@ARGV==1) {die "please provide genome name in command line \n
file should contain multiple sequences in fasta format \n
a file with the ptt table should be in the same directory\n\n";}
#$num=0;
$filename=$ARGV[0];
@nameparts=split(/\./, $filename);
#print $parts[0];
$orfs="$nameparts[0]"."\.faa";
$ptt="$nameparts[0]"."\.ptt";

open(IN, "< $ptt") or die "cannot open $ptt:$!";
$line=<IN>; # read t1st line
    if ($line=~/complete genome/) {  #look forheader
        print "$line\n";};
$line=<IN>; # read 2nd line
    print "$line\n";
$line=<IN>; # read 3rd line
if  ($line=~/Location   Strand/) {  #look for beginning of table

    while (defined ($line=<IN>)){ # read through rest of table line by line

        @parts=split/\t/,$line;
        @fromto=split/\.\./,$parts[0];
        $middle = (($fromto[1]+$fromto[0])/2);
        print "$fromto[1]\t$fromto[0]\t$middle\t$parts[3]\t";
 $gi_hash{$parts[3]}=$middle;
        print "\n";
    }
}
@gi_names = sort(keys(%gi_hash));
$total=scalar(@gi_names);

print "total number of GIs= $total\n";
foreach (@gi_names) {
    print "gi number $_ is located at $gi_hash{$_}\n";
    }

close(IN);
```

Addnumnuc.pl part 2

```
# read in and process faa file

open(IN, "< $orfs") or die "cannot open $filename:$!";
$outfilename = "$nameparts[0]"."\.num\.faa";
open(OUT, "> $outfilename")||die "cannot open: $!";

##############################################


while (defined ($line=<IN>)){ # read through file line by line

    if ($line=~/^>/) {  #look for beginning of line starting with > (^ is an anchor for the beginning of the line)
        $line =~ m/gi\|(\d+)\|/; #match gi|number capture number in $1
        $num=$gi_hash{$1};
        $line =~ s/^>//;
        # print "$1 $num \n";
        $line= ">"."$num\t"." $line";
        };
print "$line"; #print to screen
print OUT "$line"; #print to OUT
    }

close(IN);
close(OUT);
```

Results in a multiple fasta file where each annotation line starts with the nucleotide position in the chromosome:

Tmar.num.faa:

```
>385.5     gi|15642776|ref|NP_227817.1| hypothetical protein TM0001 [Thermotoga maritima MSB8]
MVYGKEGYGRSKNILLSECVCGIISLELNGFQYFLRGMETL
>545.5     gi|15642777|ref|NP_227818.1| hypothetical protein TM0002 [Thermotoga maritima MSB8]
MSPEDWKRLICFHTSKEVLKQTLDDAQQNISDSVSIPLRKY
>1828      gi|15642778|ref|NP_227819.1| hypothetical protein TM0003 [Thermotoga maritima MSB8]
METVKAYEVEDIPAIGFNNSLEVWKLFPASSSRSTSSSFQ
>1974.5    gi|15642779|ref|NP_227820.1| hypothetical protein TM0004 [Thermotoga maritima MSB8]
MKDLYERFNNSLEVWKLVELFGTSIRIHLFQ
>4131      gi|15642780|ref|NP_227821.1| DNA helicase, putative [Thermotoga maritima MSB8]
MTVQQFIKKLVRLVELERNAEINAMLDEMKRLSGEEREKKGRAVLGLTGKFIGEELGYFLVRFGRRKKID
TEIGVGDLVLISKGNPLKSDYTGTVVEKGERFITVAVDRLPSWKLKNVRIDLFASDITFRRQIENLMTLS
SEGKKALEFLLGKRKPEESFEEEFTPFDEGLNESQREAVSLALGSSDFFLIHGPFGTGKTRTLVEYIRQE
VARGKKILVTAESNLAVDNLVERLWGKVSLVRIGHPSRVSSHLKESTLAHQIETSSEYEKVKKMKEELAK
LIKKRDSFTKPSPQWRRGLSDKKILEYAEKNWSARGVSKEKIKEMAEWIKLNSQIQDIRDLIERKEEIIA
SRIVREAQVVLSTNSSAALEILSGIVFDVVVVDEASQATIPSILIPISKGKKFVLAGDHKQLPPTILSED
AKDLSRTLFEELITRYPEKSSLLDTQYRMNELLMEFPSEEFYDGKLKAAEKVRNITLFDLGVEIPNFGKF
WDVVLSPKNVLVFIDTKNRSDRFERQRKDSPSRENPLEAQIVKEVVEKLLSMGVKEDWIGIITPYDDQVN
LIRELIEAKVEVHSVDGFQGREKEVIIISFVRSNKNGEIGFLEDLRRLNVSLTRAKRKLIATGDSSTLSV
HPTYRRFVEFVKKKGTYVIF
```

.....

Geneplot using EXCEL part 1

Format databank using Tpet.num.faa

```
>formatdb -i Tpet.num.faa -p T -o T
```

Search databank using Tmar.num.faa using blastall with -m8

```
> blastall -p blastp -d Tlet.num.faa -i
Tmar.num.faa -o Tlet_Tmar.tab -F F -m 8 -W 2 -a
2 -e 0.001
```

You could use different E values

Load output (in this case Tlet_Tmar.tab) into Excel;
(note the script addnumnuc added an extra tab)

Geneplot using EXCEL part 2

TpetTmar.xls

	A	B	C	D	E	F	G	H	I	J	K	L
1	4131	939008	99.38	650	4	0	1	650	1	650	0	1182
2	5622.5	937516.5	99.71	345	1	0	1	345	1	345	0	638
3	6803	936336	99.32	438	3	0	1	438	1	438	0	833
4	7832	935307	99.55	224	1	0	1	224	1	224	######	435
5	8419	934720	99.3	142	1	0	1	142	1	142	1.00E-60	224
6	8419	1278886.5	37.32	142	78	3	6	141	5	141	1.00E-13	67.8
7	8419	843801	32.89	152	91	4	1	141	1	152	2.00E-11	60.5
8	9541.5	933597.5	100	607	0	0	1	607	1	607	0	1206
9	9541.5	1360664	52.49	602	272	5	7	596	27	626	0	626
10	9541.5	700745.5	52.64	530	247	1	7	536	5	530	######	561
11	9541.5	932477	35.79	95	57	2	8	99	33	126	2.00E-10	59.7
12	9541.5	908506.5	46.3	54	26	2	541	594	189	239	1.00E-08	54.3
13	9541.5	1361839.5	36.78	87	43	3	2	81	69	150	3.00E-06	46.2
14	9541.5	932040	30.34	89	57	2	2	86	75	162	4.00E-06	45.8
15	9541.5	684306.5	23.15	324	200	12	89	412	66	340	4.00E-05	42.4
16	9541.5	896459.5	40.74	54	29	1	543	596	207	257	5.00E-05	42
17	9541.5	1364003.5	41.46	41	23	1	548	588	12	51	4.00E-04	39.3
18	10665	932477	100	126	0	0	3	128	1	126	3.00E-56	209
19	10665	933597.5	35.79	95	57	2	35	128	8	99	3.00E-11	59.7
20	10665	700745.5	29.79	94	62	2	35	127	6	96	5.00E-04	35.8
21	10665	1360664	26.5	117	67	3	16	121	14	122	5.00E-04	35.4
22	11120	932040	99.38	162	1	0	15	176	1	162	1.00E-90	324
23	11120	1361839.5	43.59	156	85	2	20	175	2	154	1.00E-32	131
24	11120	701797.5	41.73	139	79	2	38	175	18	155	4.00E-29	119
25	11120	1358741.5	38.1	84	46	3	89	171	565	643	4.00E-08	50.1
26	11120	1360664	33.71	89	52	2	95	176	28	116	9.00E-08	48.9
27	11120	933597.5	30.34	89	57	2	89	176	2	86	8.00E-07	45.8
28	11120	1364883.5	29.11	79	45	3	99	176	1	69	4.00E-05	40

Plotting column B against A ->

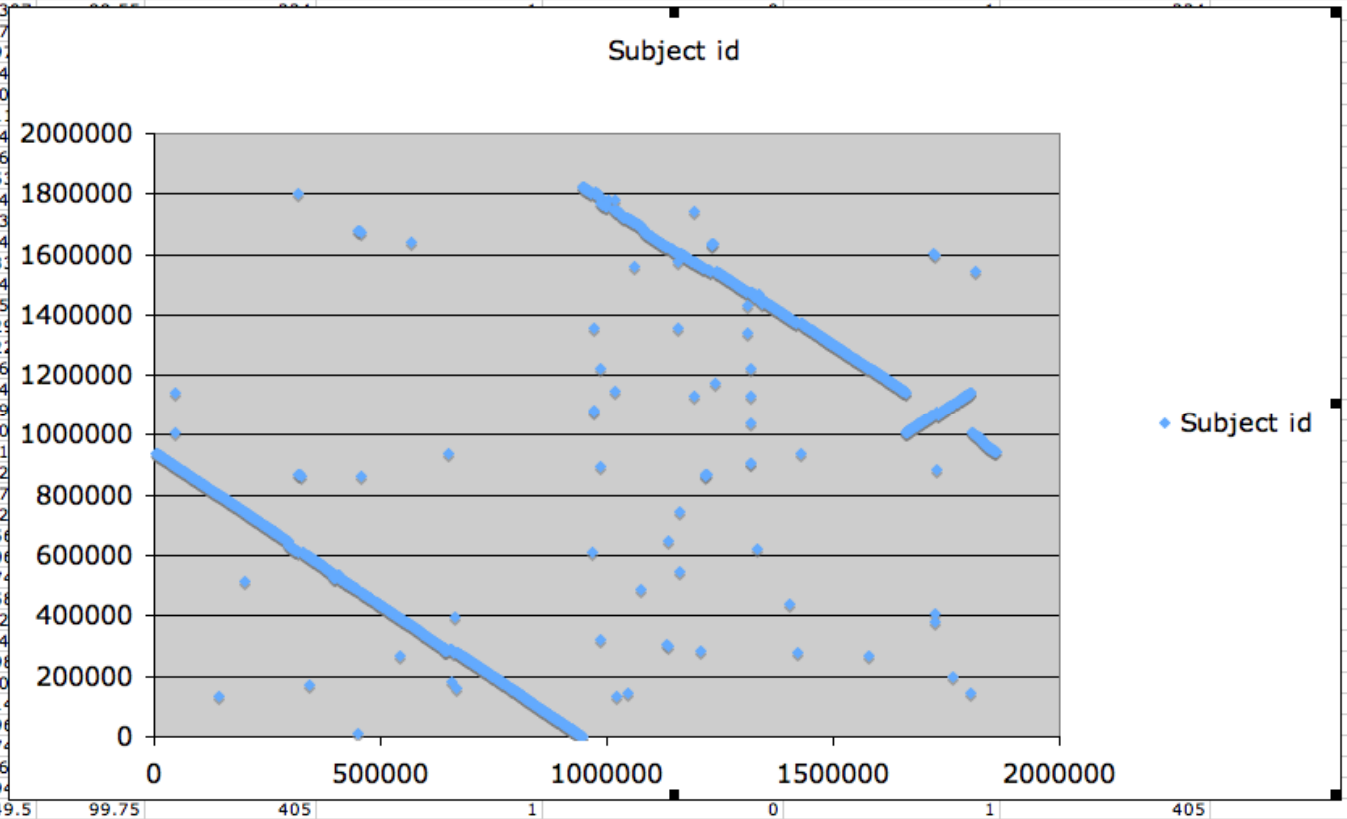
Geneplot using EXCEL part 3

To only plot the top scoring hits use extract_lines.pl -->

Geneplot using EXCEL part 4

TpetTmar.xls

	B	C	D	E	F	G	H	I
1	Subject id	% identity	alignment length	number of mismatches	number of gap openings	position of query start	position of query end	position of subject start
2	939008	99.38	650	4	0	1	650	1
3	937516.5	99.71	345	1	0	1	345	1
4	936336	99.32	438	3	0	1	438	1
5	9353							1
6	9347							1
7	93359							1
8	9324							1
9	9320							1
10	93121							1
11	9304							1
12	9296							9
13	92915							1
14	9284							1
15	9273							1
16	9264							1
17	92583							1
18	9254							1
19	9245							1
20	922729							1
21	920722							1
22	9186							2
23	9174							1
24	9169							1
25	9160							1
26	9151							1
27	9142							1
28	9127							1
29	9112							26
30	909756							24
31	908506							1
32	907174							18
33	906158							1
34	9052							1
35	9044							1
36	903898							1
37	9030							1
38	902414							1
39	901696							1
40	900274							1
41	8986							9
42	1142494							1
43	897749.5	99.75	405	1	0	1	405	1
44	1007494.5	79.75	79	16	0	30	108	51
45	896459.5	100	261	0	0	1	261	1
46	895734	100	154	0	0	1	154	1
47	894525.5	99.85	667	1	0	3	669	1
48	893361	100	110	0	0	1	110	1



TpetTmarTOP | TpetTmarALLE10-3 | Sheet3

Plotting Tpet_Tmar.tab.top

PSIBlast to find transposase homologs

- **Download transposase sequence transposase.fa**
- **Download genome as nucleotide sequence**
- **Format genome**

- `formatdb -i Tpet.fna -p F -o T`
- `blastpgp -i transposase.fa -d nr -I T -h 0.00001 -j 6 -C transposase.chk -a2`
- `blastall -i transposase.fa -d Tpet.fna -p psitblastn -R transposase.chk -o transposase_Tpet.tab -a2 -m8 -F F`

transposase_Tpet.tab:

```
ass5 jpgogarten$ more transposase_Tlet.tab
1|      gi|157362870|ref|NC_009828.1|   18.54   426     334     9       11      423     463614  462364  3e-101   361
1|      gi|157362870|ref|NC_009828.1|   15.46   194     158     5       10      197     462999  462448  4e-14   72.7
1|      gi|157362870|ref|NC_009828.1|   12.21   434     335     17      5       392     1945857 1947041 2e-08   53.4
1|      gi|157362870|ref|NC_009828.1|   19.20   125     92      5       249     364     1079762 1080133 6e-08   51.9
1|      gi|157362870|ref|NC_009828.1|   12.29   293     247     12      13      295     669830  669084  2e-07   50.0
1|      gi|157362870|ref|NC_009828.1|   14.61   178     132     8       160     317     1375657 1375151 1e-05   44.6
1|      gi|157362870|ref|NC_009828.1|   10.29   175     145     6       144     306     336563  337063  5e-05   42.3
1|      gi|157362870|ref|NC_009828.1|   16.12   273     199     14      149     391     1314911 1315603 7e-05   41.9
1|      gi|157362870|ref|NC_009828.1|   12.93   348     291     8       8       343     2023445 2022588 0.001   38.0
1|      gi|157362870|ref|NC_009828.1|   11.25   160     125     7       257     399     1943255 1942806 0.001   38.0
```

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 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**
Muxcle homepage is at http://www.drive5.com/muscle/docs.htm
Sequence files are also in script folder.

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).

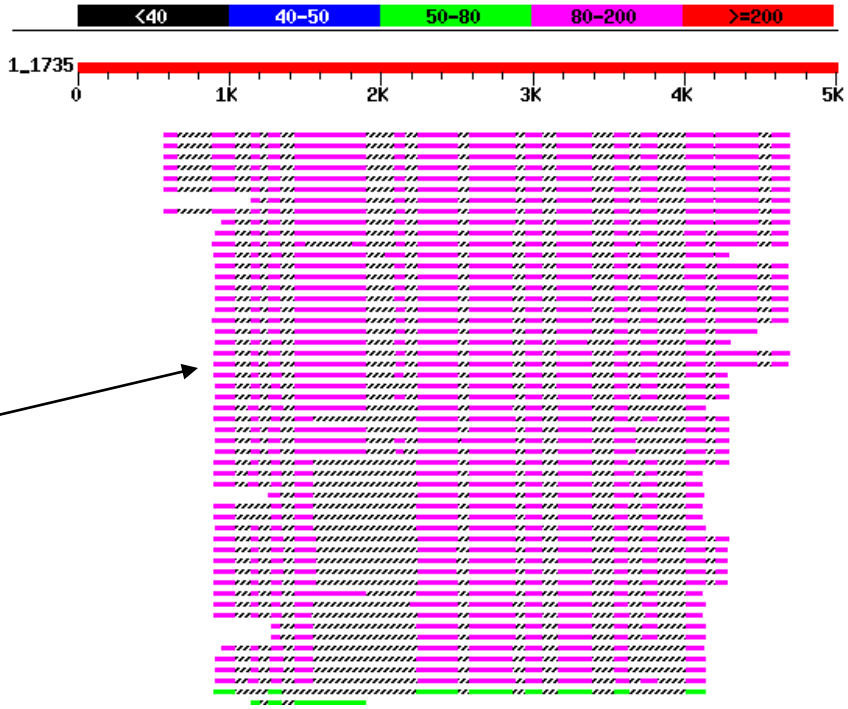
dotlet

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 VTVGDPVLRTGKPLSVELGPG

>Saccharomyces cerevisiae gi|137464|sp|P17255|VATA_YEAST VACUOLAR ATP SYNTHASE CATALYTIC SUB
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
                       :: :*:   .*     :.*       *. :  **. . : *  .     : .
```

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.

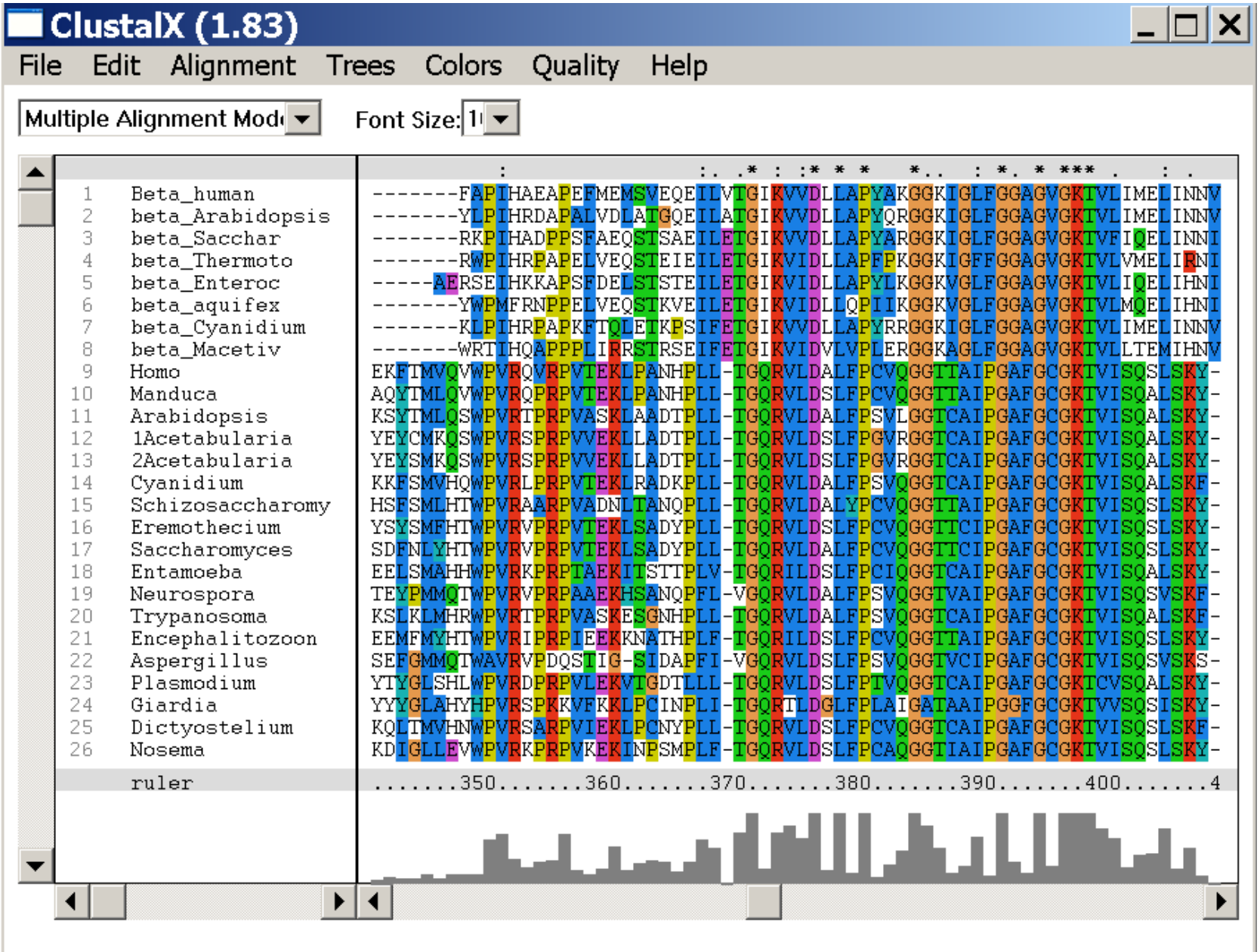
muscle

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.

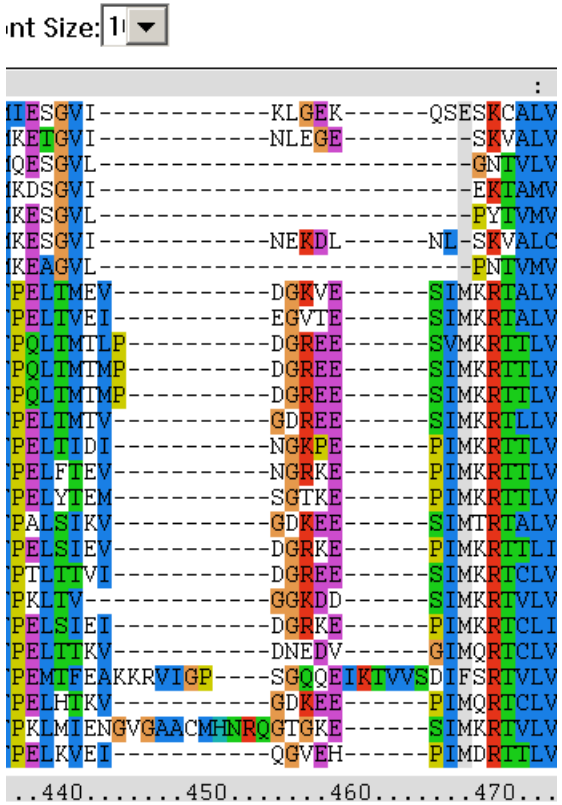
The muscle home page is here , the manual is here
Muscle allows also allows profile alignments.

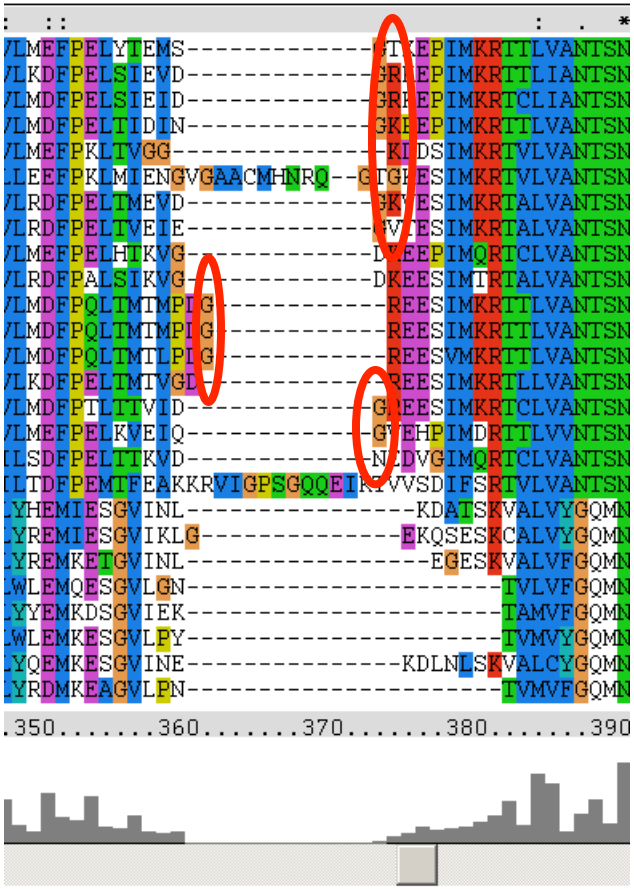
```
muscle -in VatpA.fa -out VatpA.afa
muscle -in VatpA.afa -out VatpA.rafa –refine
muscle -in beta.fa -out beta.afa
muscle -in beta.afa -out beta.rafa -refine
muscle -profile -in1 beta.rafa -in2
VatpA.rafa -out Abeta.afa
muscle –refine –in Abeta.afa –out Abeta.rafa
```

muscle alignment



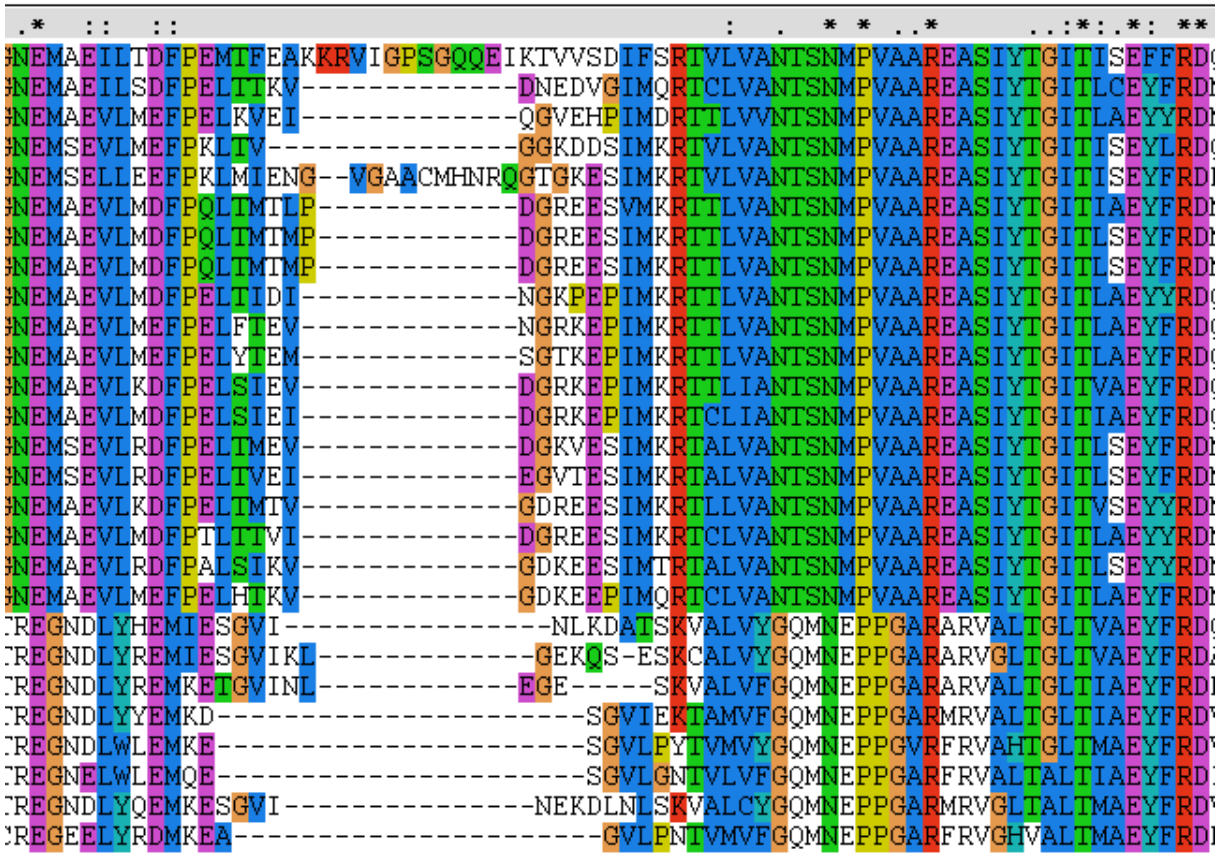
muscle *vs* clustal





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

the same region using tcoffee with default settings



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

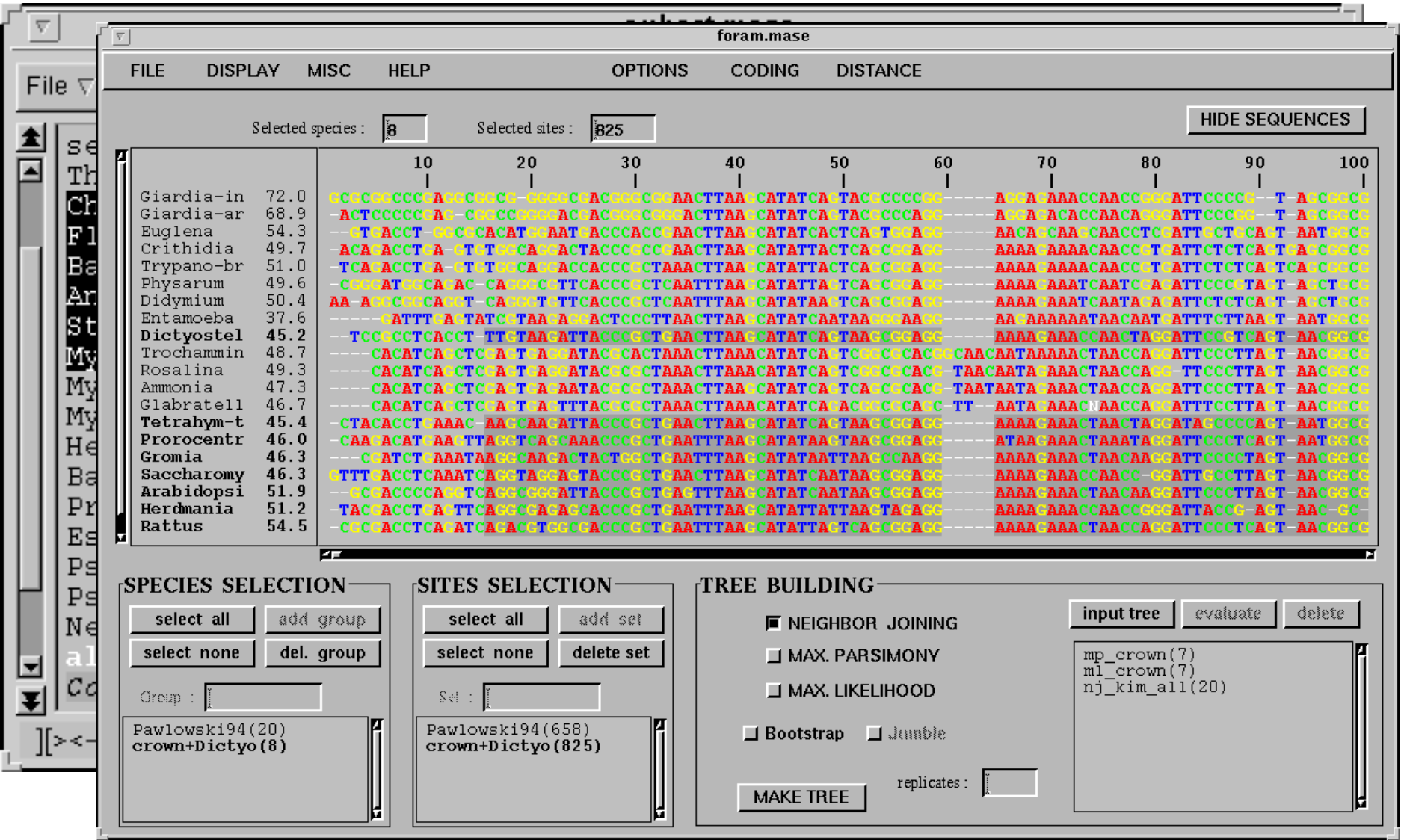
Sequence editors and viewers

Jalview Homepage, Description

Jalview is easy to install and run.
Test file is here (ATPase subunits)
(Intro to ATPases: 1bmf in spdbv)
(gif of rotation here)
(Load all.txt into Jalview,
colo**u**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**, it runs on most platforms, uses either **clustal** or **muscle** alignments, andhas simple parsimony, distance and ml programs.



Clustalw (and other progressive alignment programs):

Good alignment programs, alignments match regions that have same structures.

Not very useful for phylogenetic reconstruction: Alignment is strongly biased towards guide tree. Also, quality of alignment presumably depends on guide tree.

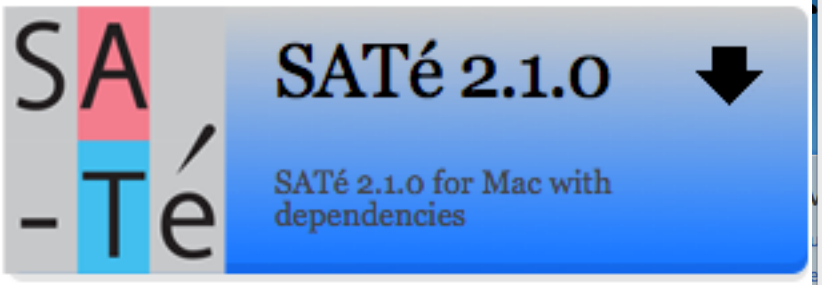
Solutions:

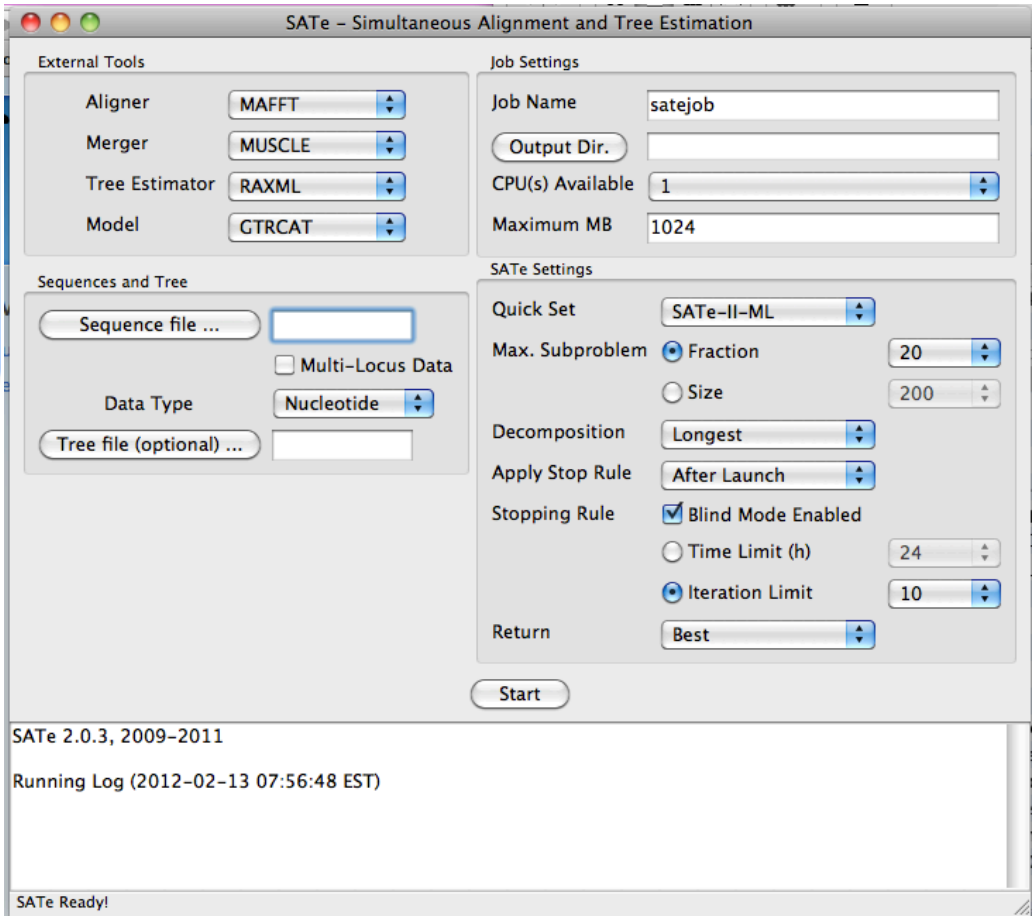
- **Use different guide trees, especially if you want to test different phylogenetic hypotheses**
- **Use an alignment program that creates less bias (muscle)**
- **Use a program that optimizes tree and alignment simultaneously.**

SATé

Simultaneous Alignment and Tree Estimation

http://phylo.bio.ku.edu/software/sate/sate.html





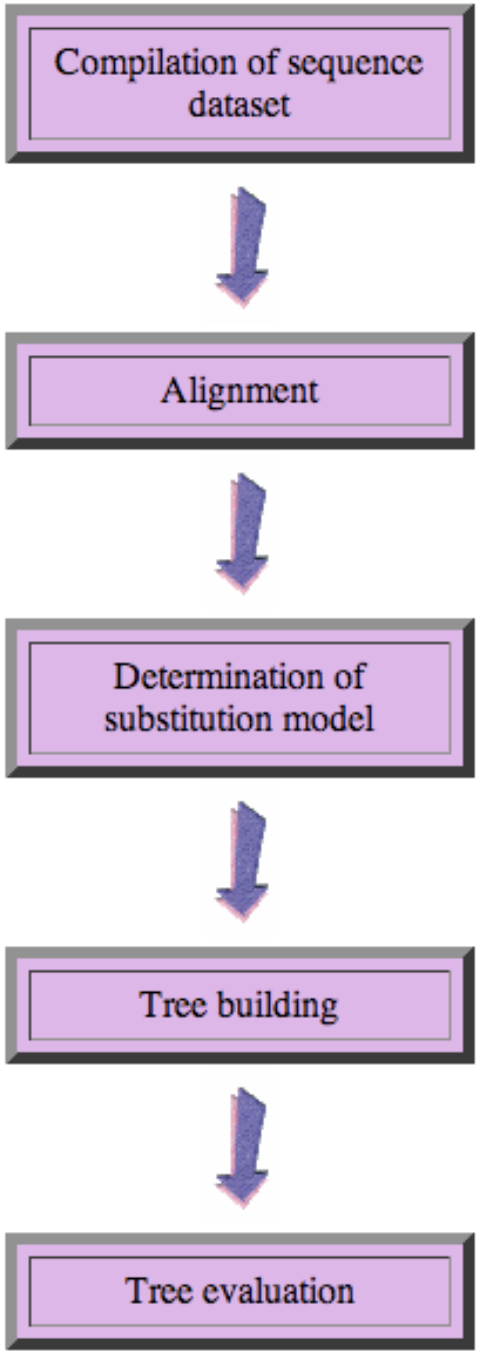
GUI works well on iMacs, but uses only local processors.

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!



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.

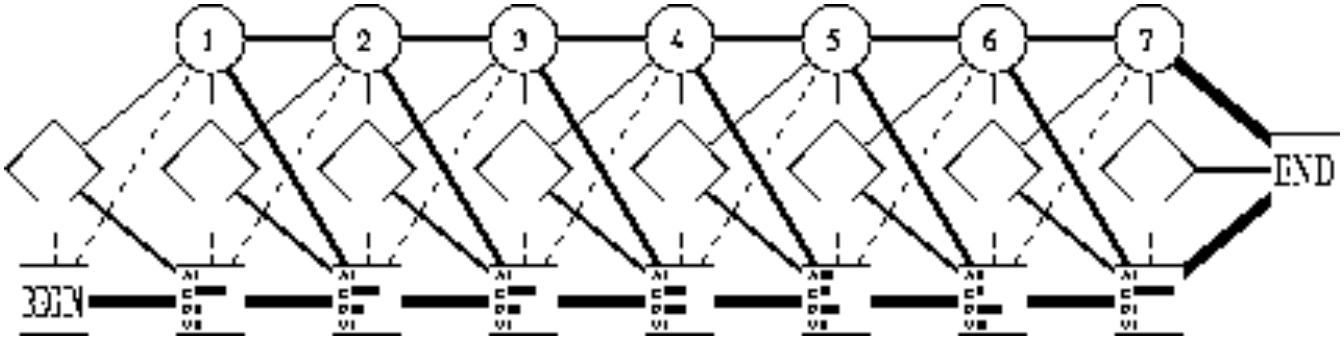
More at *Thorne JL, Kishino H (1992) Freeing phylogenies from artifacts of alignment. Mol Bio Evol 9:1148-1162*

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

more alignment programs: SAM

<u>SAM</u> (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.