

MCB 371/372

Sequence alignment

Sequence space

4/4/05

Peter Gogarten

Office: *BSP 404*

phone: *860 486-4061,*

Email: *gogarten@uconn.edu*

Questions on the Needleman Wunsch Algorithm?

a step by step illustration is [here](#)

a more realistic example is [here](#)

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

assignment:

On Monday we used the perl script [extract_lines.pl](#).

Modify the script so that it prints out an additional column that for each blasthit gives the bitscore of the alignment divided by the alignment lengths.

Hints:

`chomp ($line);` removes the `\n` character at the end of `$line`

`$parts[i]` contains the $(i+1)$ s entry of the line.

If `$a[11]` and `$a[3]` are variables in an array that contain numbers you can assign a new variable to the ratio using the command

```
$ratio_of_values=$a[11]/$a[3];
```

To print things in a line in the output separated by tabs and a new line symbol at the end you could say for example:

```
print OUT "$line\t$parts[12]\t$ratio\n";
```

Review Dotlet and Blast (chapter 11).

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.

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.

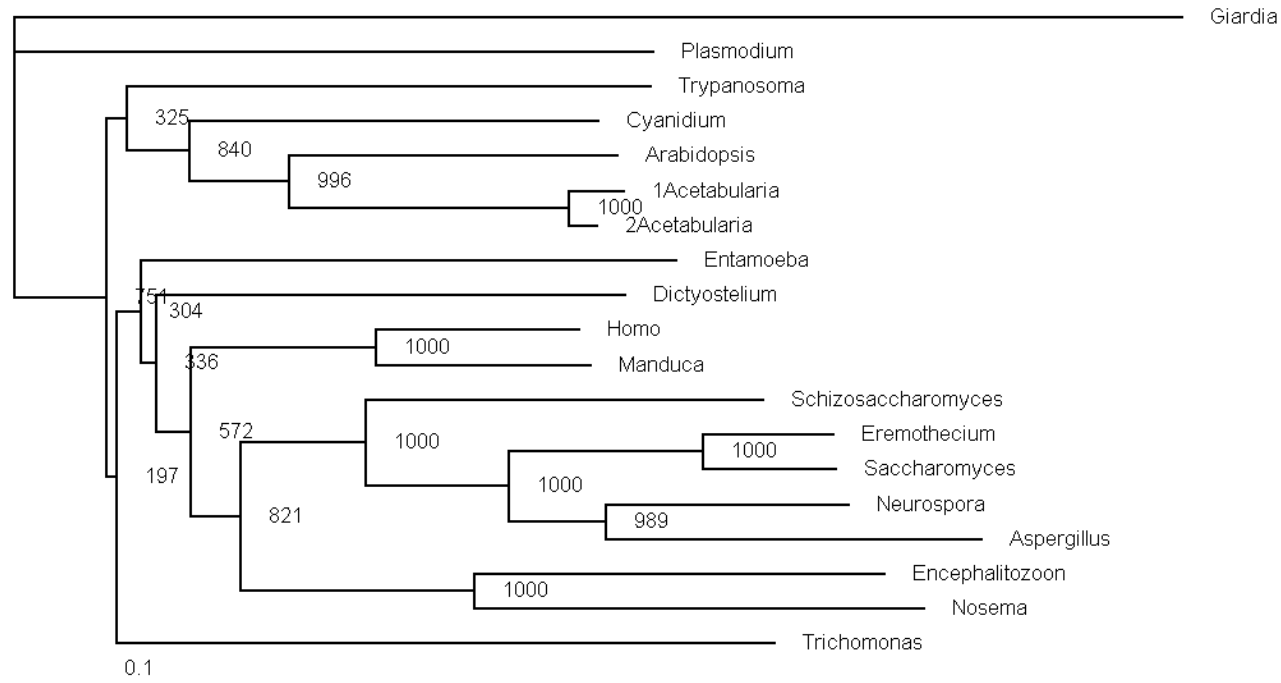
It is important to realize that the second step is the most important. The relationships found here will create a serious bias in the final alignment. The better your guide tree, the better your final alignment.

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](#)

example on bbcxs.v1.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

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 description is [here](#), 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](#)

Muscle also allows profile alignments.

```
muscle -in VatpA.fa -out VatpA.afa
```

```
muscle -in VatpA.afa -out VatpA.rafa -refine
```

```
muscle -in beta.afa -out beta.rafa -refine
```

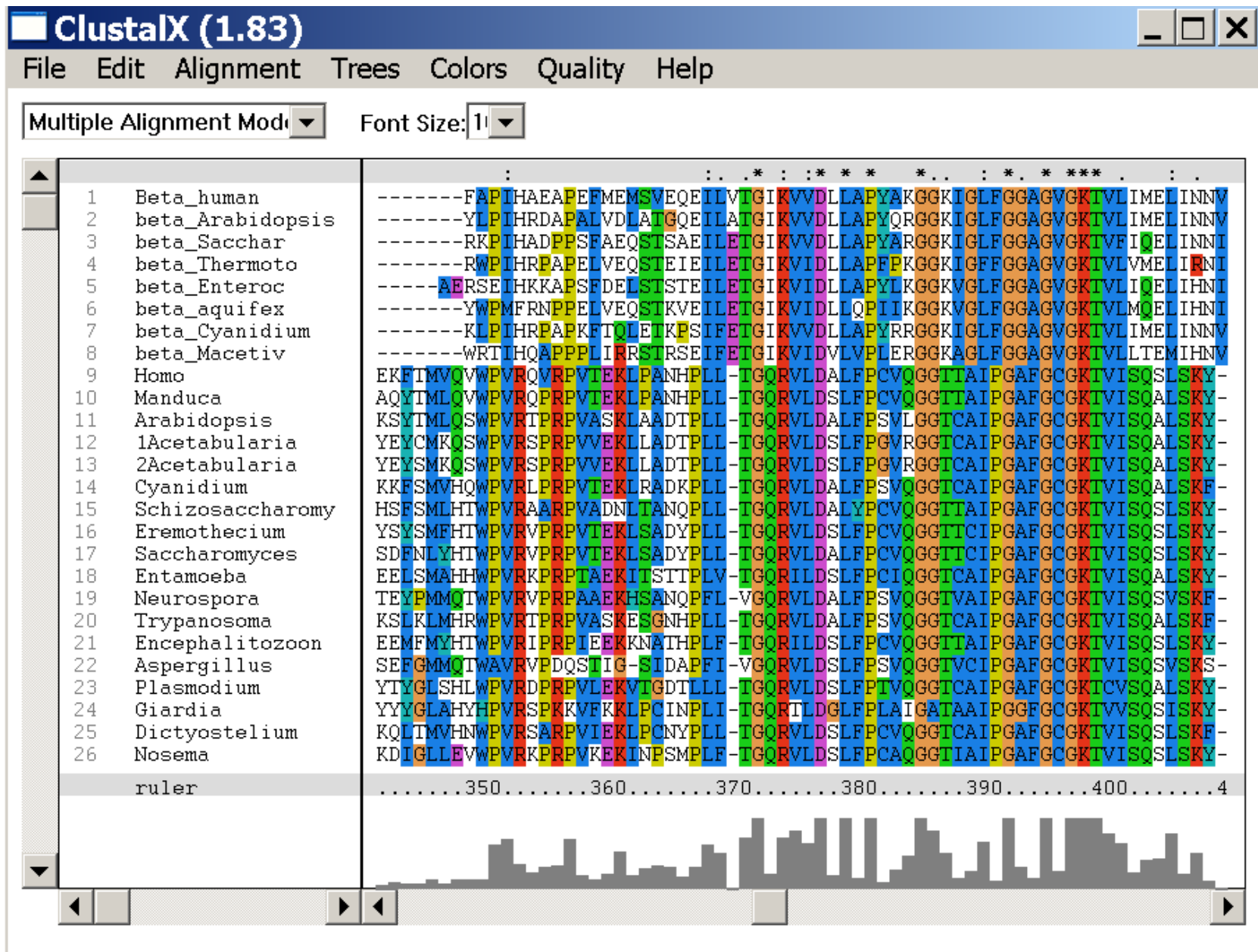
```
muscle -in beta.afa -out beta.rafa -refine
```

```
muscle -profile -in1 beta.rafa -in2
```

```
VatpA.rafa -out Abeta.afa
```

```
muscle -refine -in Abeta.afa -out Abeta.rafa
```


muscle alignment



the same region using tcoffee with default settings

```
.*  ::  ::  :  .  *  *  .*  .  .  :  :  :  *  *  :  **
#NEMAEILTDFFPEMTFEAKKRVIGPSGQQEIKTVVSDIFSRTVLVANTSNNMPVAAAREASITYTGITISEFFRDQ
#NEMAEILSDFPELTIKVI-----DNEDVGIHQRTCLVANTSNNMPVAAAREASITYGITLCEYFRDI
#NEMAEVLMDFPELKVET-----QGVHEPIMDRITLVVNTSNMPVAAAREASITYGITLAEYYRDI
#NEMSEVLMDFPKLIV-----GGKDDSIMKRTVLVANTSNNMPVAAAREASITYGITISEYLRDI
#NEMSELLEEFPKLMIENG--VGAACMHNRTGTGKESIMKRTVLVANTSNNMPVAAAREASITYGITISEYFRDI
#NEMAEVLMDFPQLTMTLP-----DGREESVMKRTTLVANTSNNMPVAAAREASITYGITIAEYFRDI
#NEMAEVLMDFPQLTMTMP-----DGREESIMKRTTLVANTSNNMPVAAAREASITYGITLSEYFRDI
#NEMAEVLMDFPQLTMTMP-----DGREESIMKRTTLVANTSNNMPVAAAREASITYGITLSEYFRDI
#NEMAEVLMDFPELTIIDI-----NGKPEPIMKRTTLVANTSNNMPVAAAREASITYGITLAEYYRDI
#NEMAEVLMDFPELFTIEV-----NGRKEPIMKRTTLVANTSNNMPVAAAREASITYGITLAEYFRDI
#NEMAEVLMDFPELYIEM-----SGTKEPIMKRTTLVANTSNNMPVAAAREASITYGITLAEYFRDI
#NEMAEVLKDFPELSIEV-----DGRKEPIMKRTTLVANTSNNMPVAAAREASITYGITVAEYFRDI
#NEMAEVLMDFPELSIEI-----DGRKEPIMKRTCLVANTSNNMPVAAAREASITYGITIAEYFRDI
#NEMSEVLRDFPELTMEV-----DGKVESIMKRTALVANTSNNMPVAAAREASITYGITLSEYFRDI
#NEMSEVLRDFPELTVEI-----EGVTEPIMKRTALVANTSNNMPVAAAREASITYGITLSEYFRDI
#NEMAEVLKDFPELTMIV-----GDREESIMKRTLLVANTSNNMPVAAAREASITYGITVSEYFRDI
#NEMAEVLRDFPALSIV-----GDREESIMKRTCLVANTSNNMPVAAAREASITYGITLAEYYRDI
#NEMAEVLRDFPALSIV-----GDREESIMKRTCLVANTSNNMPVAAAREASITYGITLAEYYRDI
#NEMAEVLMDFPELTIKVI-----GDKEEPIMQRTCLVANTSNNMPVAAAREASITYGITLAEYFRDI
#NEMAEVLMDFPELTIKVI-----GDKEEPIMQRTCLVANTSNNMPVAAAREASITYGITLAEYFRDI
#REGNDLYHEMIESGVI-----NLKDATSKVALVYQGMNEPPGARARVALTGLTVAEYFRDI
#REGNDLYREMIESGVIKVI-----GEKQS-ESKCALVYQGMNEPPGARARVGLTGLTVAEYFRDI
#REGNDLYREMKEITGVINL-----EGE-----SKVALVFGQMNEPPGARARVALTGLTIAEYFRDI
#REGNDLYYEMKD-----SGVIEKIAMVFGQMNEPPGARMRVALTGLTIAEYFRDI
#REGNDLWLEMKE-----SGVLPYITVMVYQGMNEPPGVFRVVAHTGLTMAEYFRDI
#REGNELWLEMQE-----SGVLGNTVLVFGQMNEPPGARFRVALTALTIAEYFRDI
#REGNDLYQEMKESGVI-----NEKDLNLSKVALCYQGMNEPPGARMRVGLTALTMAEYFRDI
#REGELYRDMKEA-----GVLNNTVMVFGQMNEPPGARFRVGHVALTMAEYFRDI
```

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](#))

Easy to install and run. Test on all.txt (ATPase subunits)

(do 1bmf in spdbv)

(gif of rotation [here](#), movies of the rotation are [here](#) and [here](#))

(Load all.txt into Jalview,

colour options,

mouse use,

PID tree,

Principle component analysis -> sequence space)

Info 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.

foram.mase

FILE DISPLAY MISC HELP OPTIONS CODING DISTANCE

Selected species : 8 Selected sites : 825 HIDE SEQUENCES

Species	10	20	30	40	50	60	70	80	90	100
Giardia-in	CCGCGCCCGAGGCGCG	CGGCGGCGGCGGAACTTAA	CATATCACTACCCCGG	AGGA	AAACCAACCGGATT	CCCCG	T	AGCGCG		
Giardia-ar	ACTCCCGCGAC	CGCGCGGACGACGCGGACTTAA	CATATCACTACCCCGG	ACTA	ACACCAACCGGATT	CCCCG	T	AGCGCG		
Euglena	GTACCT	GCGCACATGAAAT	ACCCACCGAACTTAA	CATATCACTCACT	CACT	CACT	CACT	CACT	CACT	CACT
Crithidia	ACACACCTAA	GTGTGACAGACTACCCCGG	AACTTAA	CATATTA	CACTCACT	CACT	CACT	CACT	CACT	CACT
Trypano-br	TCACACCTAA	GTGTGACAGACTACCCCGG	AACTTAA	CATATTA	CACTCACT	CACT	CACT	CACT	CACT	CACT
Physarum	CGGAT	CGCACAC	CAAGCC	TTACCCCGCTCAATTTAA	CATATTA	CACTCACT	CACT	CACT	CACT	CACT
Didymium	AA	AGCGCGGAGT	CAAGCT	TTACCCCGCTCAATTTAA	CATATTA	CACTCACT	CACT	CACT	CACT	CACT
Entamoeba	GATTTCACTATC	CTAA	CACTCCCTTAACTTAA	CATATCAATA	CACTCACT	CACT	CACT	CACT	CACT	CACT
Dictyostel	TCCCGCTCACCT	TTGTAAGATTACCCCGCT	CAACTTAA	CATATCACTAAGCGGAGG	AAAA	AAACCAACCTAGGATT	CCGTCAL	T	AACGGCG	
Trochammin	CACATCACTCA	ATTAAGATACGACTAA	AACTTAA	CATATCACTCACTCACT	CACT	CACT	CACT	CACT	CACT	CACT
Rosalina	CACATCACTCA	ATTAAGATACGACTAA	AACTTAA	CATATCACTCACTCACT	TAACAATA	AAACTAACCAAG	TTCCCTTA	T	AACGGCG	
Ammonia	CACATCACTCA	ATTAAGATACGACTAA	AACTTAA	CATATCACTCACTCACT	TAATAATA	AAACTAACCAAGATT	CCCTTA	T	AACGGCG	
Glabratell	CACATCACTCA	ATTAAGATACGACTAA	AACTTAA	CATATCACTCACTCACT	TT	AATA	AAACNAA	CACTTTCCCTTA	T	AACGGCG
Tetrahym-t	CTACACCT	AAAC	AAACCAAGATTACCCCGCT	CAACTTAA	CATATCACTAAGCGGAGG	AAAA	AAACTAACCTAGGATA	CCCCGAGT	AATGGCG	
Prorocentr	CAATCAT	AAATTAAGCTCAG	CAACCCCGCTCAATTTAA	CATATCACTAAGCGGAGG	ATAAC	AAACTAAATAGGATT	CCCTCACT	AATGGCG		
Gromia	CCATCT	AAATAAGGCAAGACTACT	CGCTCAATTTAA	CATATCAATAAGCGGAGG	AAAA	AAACTAACCAAGGATT	CCCTTA	T	AACGGCG	
Saccharomy	TTTTACCTCA	AAATCAGCTAGGACTACCCCGCT	CAACTTAA	CATATCAATAAGCGGAGG	AAAA	AAACCAACCGGATT	CCCTTA	T	AACGGCG	
Arabidopsi	GCGACCCCACT	CAGGCGGATTACCCCGCT	CAGTTTAA	CATATCAATAAGCGGAGG	AAAA	AAACTAACCAAGGATT	CCCTTA	T	AACGGCG	
Herdmania	TACACCTCA	ATTCAGGCGGAGCACCCCGCT	CAATTTAA	CATATTAATAAGCGGAGG	AAAA	AAACCAACCGGATT	ACCG	ACT	AAC	GC
Rattus	GCGACCTCA	SATCAGACCTGGCGACCCCGCT	CAATTTAA	CATATTAATAAGCGGAGG	AAAA	AAACTAACCAAGGATT	CCCTCACT	AACGGCG		

SPECIES SELECTION: select all, add group, select none, del. group, Group: Pawlowski94(20), crown+Dictyo(8)

SITES SELECTION: select all, add set, select none, delete set, Set: Pawlowski94(658), crown+Dictyo(825)

TREE BUILDING: NEIGHBOR JOINING, MAX. PARSIMONY, MAX. LIKELIHOOD, Bootstrap, Jumble, input tree, evaluate, delete, MAKE TREE, replicates: []

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