MCB 371/372

Sequence alignment
3/30/05

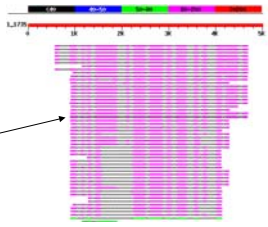
Peter Gogarten
Office: *BSP 404*
phone: *860 486-4061,*
Email: *gogarten@uconn.edu*

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). Currently, the applet is reported to only work in Internet Explorer ONLY.
One application of this program is to find internal duplications and to locate exons.

Example: this sequence against itself
genomic sequence against Protein

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



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 and 6319974**

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

The Needlemann 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**

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, 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.

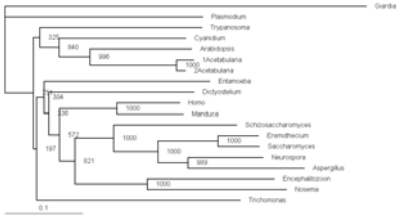
It is important to realize that the second step is the most important. The relationships found here will create a serious bias in the final alignment. The better your guide tree, the better your final alignment.

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

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.

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 description is here, a web interface is here.

muscle

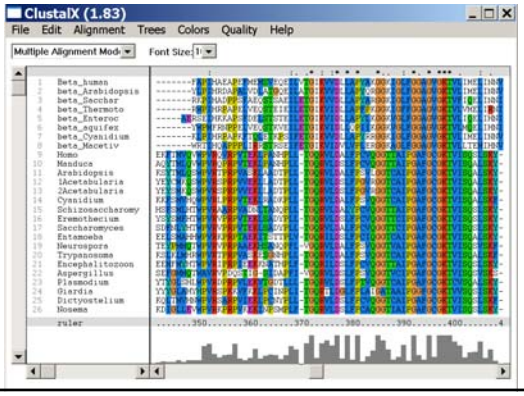
If you have very large datasets muscle is the way to go. It is fast, takes fasta formated sequences as input file, and has a refinement option, that does an excellent job cleaning up around gaps.

The muscle home page is here

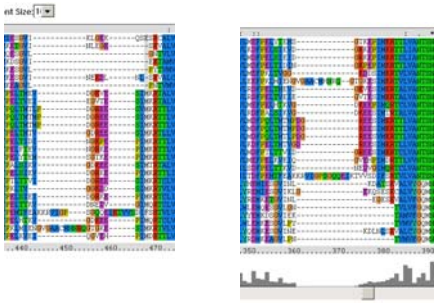
Muscle allows also allows profile alignments.

```
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



muscle *vs* clustal



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

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";
```

Demo using **putty** to bbcxsrv.biotech.uconn.edu
- maybe

follow instructions of exercise one task 6 – these are the commands

```
formatdb -i p_abyssi.faa -o T -p T

blastall -i t_maritima.faa -d p_abyssi.faa -o
blast.out -p blastp -e 10 -m 8 -a2

./extract_lines.pl blast.out
```

sftp results
load into spreadsheet
sort data, do histogram …
the extract_lines.pl script is here (you can sftp it into your account, you'll need to chmod 755 extr*.pl afterwards)

Command Line

The favored operating system flavor in computational biology is UNIX/LINUX.
The command line is similar to DOS.
Some of the frequently used commands are here

`pwd`	`ps`
`ls`	`ps aux`
`ls -l`	`rm`
`chmod`	`more`
`chmod a+x blastall.sh`	`cat`
`chmod 755 *.sh`	`vi (text editor)`
`cd`	`ps`
`cd ..`	`ps aux`
`cd $HOME`	`ssh`
`passwd`	`sftp`

For windows a good ssh program is putty.
UConn also has a site license for the ssh program from ssh.com