



Multiple Sequence Alignment

Hot Topic 5/24/06

Kim Walker



Outline

- Why are Multiple Sequence Alignments useful?
- What Tools are Available?
- Brief Introduction to ClustalX
- Tools to Edit and Add Features to your Alignment



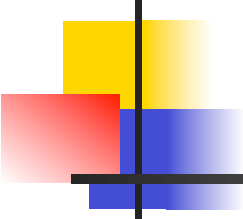
Why Do a Multiple Sequence Alignment?

- Multiple Sequence Alignment can reveal sequence patterns
 - Demonstration of homology between >2 sequences
 - Identification of functionally important sites
 - Protein function prediction
 - Structure prediction
 - Search for weak but significant similarities in database
 - Design PCR primers for related gene identification
 - Genome sequencing: contig assembly



Types of MSA

- Global Alignment - Attempts to align as many characters as possible from end to end
 - Optimal Global Alignments - Dynamic Programming
 - Global progressive - Match most common sequences together
 - Iterative multiple - Multiple re-building attempts to find best alignment
- Local Alignment - Attempts to align regions of sequences with the highest density of matches, creating >1 islands of sub-alignments



Multiple Sequence Alignment Programs

- Many resources exist either as web tools or free programs you can download and install
- Global Alignment Programs
 - ClustalX - window interface for ClustalW (EMBOSS) MSA tool, program can be download from here: <ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/>
 - Multialign - web tool using hierarchical clustering: <http://prodes.toulouse.inra.fr/multalin/multalin.html>
- Local Alignment Programs
 - Block Maker http://bioinfo.weizmann.ac.il/blocks/blockmkr/www/make_blocks.html



Getting Started with ClustalX

- Input sequences: Input files should be in FASTA format saved using a text based editor (not MS Word)

```
>Mouse_Oct4  
AGTCGATTGCA
```

Note: ClustalX doesn't like spaces in the header line

You can either have all your sequences in one file using File ->Load Sequences

Or add sequences one at a time using File->Append Sequences



Output Options

- Alignment -> Output Format Options
 - What do you plan to do with this alignment afterwards? ClustalX vs. MSF format of output file.
- Color - what colors do you want your nt or aa?
 - Defaults: DNA - A orange, C red, T blue, G green
 - Protein - GPST orange, HKR red, FWY blue, ILMV green

By changing the coldna.par or colprot.par files - you can highlight different AA properties or different nt regions (ex. gcgc region)



Alignment

- Alignment -> Alignment Parameters
 - There are many options here that advanced users can change to alter how ClustalX calculates the alignment
- Alignment -> Do Complete Alignment
 - ClustalX will ask where to save the Guide Tree File (which can be opened using TreeView) and the Alignment file itself
 - Note - the Tree file is Guide tree, not a phylogenetic Analysis. See the 'Trees' menu phylogenetic options



Alignment cont

- Symbols

- * indicates positions which have a single, fully conserved residue
- : indicates that one of the following 'strong' groups is fully conserved: STA, NEQK, NHQK, NDEQ, QHRK, MILV, MILF, HY, FYW
- . indicates that one of the following 'weaker' groups is fully conserved: CSA, ATV, SAG, STNK, STPA, SGND, SNDEQK, NDEQHK, NEQHRK, FVLIM, HFY
- An amino acid substitution matrix is used to define strong vs. weak groups



Profile Alignment

- Create a new alignment using an existing alignment or a profile as a guide.
- How to?
 - Load sequences of Profile 1
 - Do alignment
 - Switch to Profile Alignment Mode
 - Load sequences of Profile 2.
 - Now go to Alignment -> Align sequences to Profile 1, which will take 1 sequence at a time and align it to Profile 1
 - Alternatively, Align Profile 2 with Profile 1, in which case you need to have Profile 2 pre-aligned.



Are alignment programs 100% right?

- No alignment program can take the place of an educated human eye.
- Use programs like ClustalX as a guide, but don't be afraid to adjust the alignment as you see fit.
- Jalview is one program you can use to adjust your sequence alignment



What is Jalview?

- Jalview is an alignment editor that can be used for viewing, editing, analysis, annotation and publishing. Jalview can be download for free here: <http://www.jalview.org/>
- Documentation: <http://www.jalview.org/refcard.pdf>
- Since Jalview is Java based, users will have troubles installing and using Jalview without the newest version of Java. For instance, Mac users, The newest Jalview will only work with OS 10.4 (Tiger).



Jalview cont.

- Jalview can be used with nucleotide and protein sequences, but there are more features for protein sequence analysis



Jalview - Getting Started

- File -> Input Alignment -> From File
 - Choose which file type your alignment is
 - Will notify when sequences are successfully loaded
- Within alignment window
 - Color -> choose color scheme according to program
 - View -> Font - to change font and size



Basic Editing

- Inserting or deleting gaps
 - Single sequence - Press 'shift' and left mouse button, drag either left or right
 - Group of sequences - Press 'control' and left mouse button, drag either left or right
- Draw a Box, highlighting a sequence area
 - Click and drag your mouse over the sequence area.
 - Right click on the highlighted area and a menu will appear.
 - Groups -> Group color or Border color, You can change the color of the group and of the border of the box.



Jalview Advanced

- Jalview allows you to run:
- Jnet - a neural network secondary structure prediction method protein prediction
- ClustalW - the general multiple sequence alignment program in which ClustalX is based



Export

- File -> Export -> HTML, PNG, or EPS
 - Export file type depending on what you would like to do to your alignment next



References

- Bioinformatics: Sequence and Genome Analysis. David W. Mount. CSHL Press, 2001 and 2004.
- From Sequence to Multiple Alignment, <http://iona.wi.mit.edu/bio/education/how-to-ma.html>
- Spring Course Notes <http://jura.wi.mit.edu/bio/education/index.html>
- Questions? Need Help?
wibr-bioinformatics@wi.mit.edu