

Editing and Publishing Alignments



Modified by Dietlind Gerloff
for use in BME110/BIOL181
Winter 2009

© Wiley Publishing, 2007. All Rights Reserved.



Outline

- ✓ A quick recap: uses of MSAs
(see also Lecture from Feb 3)
- ✓ Get a working idea of the most common
sequence formats
- ✓ Know how to edit an MSA with Jalview

4 Ways of Using MSAs . . .

<i>Application</i>	<i>Procedure</i>
Extrapolation	Determine the function of your protein
Phylogenetic analysis	Build a Phylogenetic tree
Pattern identification	Discover important positions
Domain identification	Turn your alignment into a domain profile

4 More Ways of Using MSAs

DNA regulatory elements identification	Use your alignment to discover promoters
Structure prediction	Predict the secondary structure of proteins and RNA molecules
nsSNP analysis	Discover important allelic variations in human and other animals (nsSNP: non-synonymous single-nucleotide polymorphisms)
PCR analysis	Select your PCR primers

Some Guidelines for Choosing the Right Sequences

<i>Problem</i>	<i>Diagnostic</i>
Proteins or DNA	Use proteins whenever possible.
Many sequences	Start with 10~15 sequences, 50 at max
Very different sequences	Avoid sequences very different from the rest of the set
Identical sequences	Avoid identical sequences; they never help
Partial sequences	Can trigger errors. Avoid them
Repeated domains	Can trigger errors. Extract the domains

Selecting a Method

- ✓ Many alternative methods exist for MSAs
- ✓ Most of them use the **progressive algorithm**
- ✓ They all are approximate methods
- ✓ None is guaranteed to deliver the best alignments
- ✓ All existing methods have pros and cons
 - **ClustalW** is the most popular (21,000 citations)
 - **T-Coffee** and **ProbCons** are more accurate but slower
 - **MUSCLE** is very fast, ideal for very large datasets

Alignments and Formats

- ✓ Many alternative formats exist for MSAs
- ✓ One format does not always have a clear advantage over another
 - That said, some are very good (e.g. .a2m) but not very widely used
 - Changing formats is possible - however, best is to plan ahead because...
 - Annotation information can sometimes be lost in a format change
- ✓ Not all formats contain the same information

The Most Common Sequence Formats

<i>Name</i>	<i>Type</i>	<i>Usage</i>
post-script, PDF, HTML	Graphic	Terminal formats suitable for printing only
FASTA	Text Non-interleaved	Easy to manipulate Supported by most programs
PIR	Text Non-interleaved	Similar to FASTA but with an extra line Can incorporate limited extra annotation
MSF	Text Interleaved	Official MSA format Supported by most programs
Selex	Text Interleaved	Extended version of MSF Can include extra annotation Supported by few programs
ALN	Text Interleaved	Simplified MSF Output of ClustalW Supported by many programs
Phylip	Text Interleaved	Variant of ALN Useful for doing phylogenetic analysis

Interleaved and Non-interleaved

✓ The MSF Format

- Interleaved

```

Fileup
MSF: 79 Type: P Check: 9672 ..
Name: hmg1_trybr oo Len: 79 Check: 6600 Weight: 1.000
Name: hmg1_mouse oo Len: 79 Check: 9313 Weight: 1.000
Name: hmg1_chite oo Len: 79 Check: 217 Weight: 1.000
Name: hmg1_wheat oo Len: 79 Check: 3534 Weight: 1.000
//
hmg1_trybr KKDSNAPKRA MTSFMPFSSD
hmg1_mouse KPKRPRSAYM IVUSESPOEA
hmg1_chite ADKPKRFLSA VHLWLSAIRE
hmg1_wheat DPNKPKRAPS AFPVPMGEFR
hmg1_trybr FRSKHSDLSI UEMSKARQAA
hmg1_mouse KDDSAGCKLK LUNEAUKNLS
hmg1_chite SJKREMPDK VTEUANKGSE
hmg1_wheat EEFKQKPKM KSUARUGKA
hmg1_trybr VKELGPEERK UVEEMAEKDK
hmg1_mouse PEEKQV IQL AKDDRI RYDN
hmg1_chite LARGLKQSE UEMKATYKQ
hmg1_wheat GERMSLSES EKAPYUAKAN
hmg1_trybr ERYKREM...
hmg1_mouse EHSUVEEQMA E...
hmg1_chite WYIKOLESE RAGG...
hmg1_wheat KLGVEYNKAI ARYKNGESA

```

✓ The FASTA Format

- Non-interleaved

```

>hmg1_trybr
kkdsnapkrantsf m f s s d f r s k h
s d l s i v e m s k a a g a a w k e l g p e e r k
v y e e m a e k d k e r y k r e m -----
>hmg1_mouse
kpkprpsayniyvsesf qeakddsa
ggklklvneawkn l s p e e k q a y i q l
akddrirydnemksw e e q m a e -----
>hmg1_chite
adkpkprlsaymlwlnsares ikre
npdf kutevakkgge l w r g l k d k s e
w e a k a a t a k q n y i r a l q e y e r n g g -----
>hmg1_wheat
d p n k p k r a p s a f f v f m g e f r e e f k q
k n p k n k s v a a v g k a a g e r w k s l s e s
e k a p y v a k a n k l k g e y n k a i a a y n k
g e s a

```

Choosing Your Format

✓ When choosing a format, ask yourself four questions:

- Is it supported by the programs I need to use ?
- Can my collaborators use it?
- Can it support all of my annotation ?
- Is it easy to read and manipulate ?

Potential Information Loss When Converting MSAs

<i>Information Type</i>	<i>Nature of the Loss</i>
Sequence name	Names can be modified Long names can be truncated Special symbol can be replaced
Upper/lowercase	Casing can change Official FASTA only supports uppercase
Gap type	Gaps use the '-' symbol This can change with some formats (MSF: . or ~)
Annotation	Most annotation gets lost
Special amino acid or nucleotides	All formats and programs do not support the same alphabet

Editing your MSA

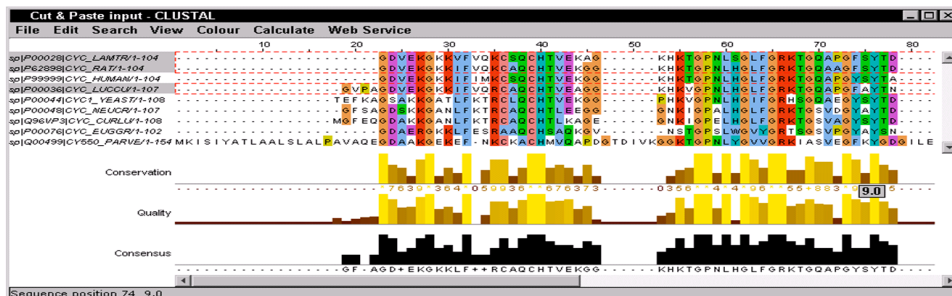
- ✓ If your MSA looks bad . . .
 - Don't torture the online server
 - Edit the MSA yourself locally
- ✓ Only use a standard word processor if you . . .
 - Know the tricks (e.g. how to cut/copy/paste vertical selections (text columns))
 - Know exactly what change you would like to make (no coloring scheme to help you)
 - Are well-rested (no protection from overwriting/deleting)
- ✓ Generally it is best to learn, and use, a dedicated MSA editor
- ✓ The most popular online tool is Jalview
 - You can get it at www.jalview.org

With Jalview You Can . . .

- ✓ Modify your MSA
- ✓ Remove some of the redundant sequences
- ✓ Insert/remove gaps
- ✓ Shift portions of the MSA
- ✓ Modify the alignment of a sub-group of sequences
- ✓ Recompute some portions of your alignment

Some Special Features of Jalview

- ✓ Computation of a consensus sequence
- ✓ Applying any color scheme to your MSA
- ✓ Removal of the redundancy
- ✓ The new Jalview also allows you to view structures etc
- ✓ Make sure you have the documentation handy when working with it!



Reminder: You can view your MSA as a LOGO Graph

- ✓ A LOGO graph summarizes an MSA
- ✓ Tall letters indicate highly conserved positions
- ✓ Short letters indicate poorly conserved positions
- ✓ LOGO graphs are ideal for identifying conserved patterns
- ✓ For example: <http://weblogo.berkeley.edu>

