

Multiple Sequence Alignment

BME 110: CompBio Tools

Todd Lowe

April 28, 2009

Original Slides: Carol Rohl

Multiple Sequence Alignment

- Multiple sequence alignment (MSA) is one of the most important bioinformatics tools
 - Many applications require accurate MSAs
 - PSI-BLAST
 - Family and domain classification
 - Pattern identification
 - Structure prediction
 - secondary structure
 - fold recognition
 - Phylogeny
 - Full-genome alignments in browsers
-

In Proteins, Common Conservation Patterns

- Cys pairs -disulfide bonds
 - His, Ser -catalytic sites
 - Cys, His -metal binding sites
 - Gly, Pro -ends of 2° structure elements, turns
 - Lys, Arg, Asp, Glu - ligand binding
 - Lys/Arg-Asp/Glu pairs - salt bridges
 - Leu -coiled coils, leucine zippers
 - Motifs, secondary structure, indels
-

Methods

- Dynamic Programming
 - Gives the optimal solution, but prohibitively slow for >6-8 sequences
 - MSA program is an example
 - Progressive Alignment
 - ClustalW
 - <http://www.ebi.ac.uk/clustalw/index.html>
(most commonly used)
 - Tcoffee
 - <http://igs-server.cnrs-mrs.fr/Tcoffee/>
(a little better, but slower)
 - Iterative
 - better than progressive methods, but slower
 - Dialign
 - HMMs
-

Progressive Alignment

1. Calculate global pair-wise alignments for all pairs
 - Needleman and Wunsch
 2. Use pairwise alignment scores to calculate a guide tree describing the distance between all pairs of sequences
 3. Align the sequences progressively
 - Start with the two most closely related sequences
 - Add in sequences in order of increasing distance
- ClustalW uses this method
-

ClustalW Example

- Input: 5 sequences detected by BLASTp using human SNAP-25 as a query
- Default parameters, output order: input

```
>sp_P13795
MAEDADMRNELEEMQRRADQLADESLESTRRMLQLVEESKDAGIRTLVMLDEQGEQLERIEEGMDQINKD
MKEAEKNLTDLGKFCGLCVCPCNKLKSSDAYKKAAGNNQDGVVASQPARVVDEREQMAISGGFIRRVTND
ARENEMDENLEQVSGIIGNLRHMALDMGNEIDTQNRQIDRIMEKADSNKTRIDEANQRATKMLGSG
```

```
>gi_31242623
MPAAAPPAENGAAPKTELQELQMKQQQVVDES�DSTRRMLALCEESTEVGMRTIVMLDEQGEQLDRIEE
GMQINADMREAENLNGMEKCCGICVLPNKSASFKEDDGTWKGNDGKVVNNQPQRVMDDRNGLGPQA
GYIGRITNDAREDEMEENMGQVNTMIGNLRNMAIDMGSELENQNRQIDRINRKGDSNATRIAAANERAH
LLK
```

```
>gi_3822409
MPTTAEPAQENGAPRSELQELQLKAGQVTDFTLESTRRMLALCEESKEAGIRTLVALDDQGEQLERIEEN
MDQINADMKEAEKNLTGMEKFCGLCVLPWNKSAPFKENEDAWKGNDDGKVVNNQPQRVMDGSGGLPQGG
YIGRITNDAREDEMEENVGQVNTMIGNLRNMAIDMGSELENQNRQIDRIKKAEM
```

```
>gi_39593308
MSARRGAPGGQRHPRPYAVEPTVDINGLVLPADELDKGLNVGIDEKTIESLESTRRMLALCEESKEAG
IKTLVMLDDQGEQLERCEGALDTINQDMKEAEDHLKGMKCCGLCVLPWNKTDDFEKNSEYAKAWKDD
GGVISDQPRITVGDPTMGPQGGYITKITNDAREDEMDENIQQVSTMVGNLRNMAIDMSTEVSNQNRQLDR
IHDKAQSNEVRVESANKRAKNLITK
```

```
>gi_32567202
MSGDDDIPEGLEAINLKMNATDDSLSTRRMLALCEESKEAGIKTLVMLDDQGEQLERCEGALDTINQD
MKEAEDHLKGMKCCGLCVLPWNKTDDFEKTEFAKAWKDDGGVISDQPRITVGDSSMGPQGGYITKIT
NDAREDEMDENVQVSTMVGNLRNMAIDMSTEVSNQNRQLDRIHDKAQSNEVRVESANKRAKNLITK
```



Input Formats

- FASTA format
 - Download from NCBI, ExPASy, EBI, Pfam ...
 - Sequence names should be
 - Unique
 - 15 characters or less
 - Comprised of only A-Z,a-z,0-9 and _
(Do not use #,\$%@|*!:,., or spaces)
-

ClustalW Output

CLUSTAL W (1.82) Multiple Sequence Alignments

Sequence format is Pearson

Sequence 1: sp_P13795 206 aa
Sequence 2: gi_31242623 213 aa
Sequence 3: gi_3822409 195 aa
Sequence 4: gi_39593308 235 aa
Sequence 5: gi_32567202 207 aa

Start of Pairwise alignments

Aligning...

Sequences (1:2) Aligned. Score: 57
Sequences (1:3) Aligned. Score: 59
Sequences (1:4) Aligned. Score: 52
Sequences (1:5) Aligned. Score: 51
Sequences (2:3) Aligned. Score: 77
Sequences (2:4) Aligned. Score: 53
Sequences (2:5) Aligned. Score: 54
Sequences (3:4) Aligned. Score: 60
Sequences (3:5) Aligned. Score: 61
Sequences (4:5) Aligned. Score: 87

Guide tree file created: [/ebi/extserv/old-work/clustalw-20040206-01234219.dnd]

Start of Multiple Alignment

There are 4 groups

Aligning...

Group 1: Sequences: 2 Score:3818
Group 2: Sequences: 3 Score:3429
Group 3: Sequences: 2 Score:4233
Group 4: Sequences: 5 Score:3386

Alignment Score 7423

CLUSTAL-Alignment file created [/ebi/extserv/old-work/clustalw-20040206-01234219.aln]

ClustalW Guide Tree

- The guide tree shows the distances between sequences obtained from the initial pairwise alignments.
- This is the order that sequences were added into the MSA
- Guide tree is *not* a phylogenetic tree (it's just a rough estimate of similarity), however a true phylogenetic tree can be generated after making an alignment



Progressive Alignment

- Greedy algorithm
 - Breaks problem up into smaller problems
 - Finds best solution to each small problem
 - Combine solutions to get answer to whole problem
- Not necessarily the global answer
 - Doesn't use all information in solving sub-problems
 - Suboptimal answers for small problems may combine to give a better overall answer
- Gaps: once created, they stay as part of alignment for rest of alignment iterations



ClustalW Alignment

CLUSTAL W (1.82) multiple sequence alignment

```
sp_P13795      ---MAEDAD-----MRNELEEMQRRADQLADESLESTRRML 33
gi_31242623   MPAAAPPAENG-----AAVPKTELQELQMKQQQVVDES LDSTRRML 41
gi_3822409    MPTTAEPAQE-----NGAPRSELQELQLKAGQVTDETLSTRRML 40
gi_39593308   MSARRGAPGGQRHPRPYAVEPTVDINGLVLVPADMSDELKGLNVGIDEKTIESTRSTRRML 60
gi_32567202   MSGDDDIPEG-----LEAINLKMNATDDSLSTRRML 33
               .                               *:  ::      .  ::*:*****

sp_P13795      QLVVESKDAGIRTLVMLDEQGEQLERIEEGMDQINKDMKEAEKNLTDLGKFCGLCVCPN 93
gi_31242623   ALCEESTEVMRTIVMLDEQGEQLDRIEEGMDQINADMREAEKNLSGMEKCCGICVLPN 101
gi_3822409    ALCEESKEAGIRTLVALDDQGEQLERIEENMDQINADMKEAEKNLTGMEKFCGLCVLPWN 100
gi_39593308   ALCEESKEAGIKTLVMLDDQGEQLERCEGALDTINQDMKEAEDHLKGMEKCCGICVLPWN 120
gi_32567202   ALCEESKEAGIKTLVMLDDQGEQLERCEGALDTINQDMKEAEDHLKGMEKCCGICVLPWN 93
               *  *** . : * : * : * * * : * * * : * * * : * * * : * * * : * * *

sp_P13795      KLKSSDA---YKKA WGNNDG-VVASQPARVVDEREQMAISGGFIRRVTNDARENEMDEN 149
gi_31242623   KSASFKE---DDGTWKGNDDGKVVNNQPQRMDDRNGLGPQAGYIGRITNDAREDEMEEN 158
gi_3822409    KSAPFKE---NEDAWKGNDDGKVVNNQPQRMDDGSGLGPQGGYIGRITNDAREDEMEEN 157
gi_39593308   KTDDFEKNSEYAKAWKKDDGGVISDQPRITVGDPT-MGPQGGYITKITNDAREDEMEND 179
gi_32567202   KTDDFEK-TEFAKAWKKDDGGVISDQPRITVGDSS-MGPQGGYITKITNDAREDEMEND 151
               *      .           : *  : : * * * : . * *  . : : :  : . . * : * : : * * * * * : * * *

sp_P13795      LEQVSGIIGNLRHMALDMGNEIDTQNRQIDRIMEKADSNKTRIDEANQRATKMLGSG 206
gi_31242623   MGQVNTMIGNLRNMAIDMGSELENQNRQIDRINRKGDSNATRIAAANERAHDLLK-- 213
gi_3822409    VQVNTMIGNLRNMAIDMGSELENQNRQIDRIKKAEM----- 195
gi_39593308   IQQVSTMVGNLRNMAIDMSTEVSNQNRQLDRIHDKAQSNEVRVESANKRAKNLITK- 235
gi_32567202   VQQVSTMVGNLRNMAIDMSTEVSNQNRQLDRIHDKAQSNEVRVESANKRAKNLITK- 207
               :  ** .  : : * * * : * * : * * . * : . . * * * : * * * * * * * : * . :
```

Interleaved Formats

- Most common output formats for MSAs are interleaved:
 - MSF, ASN, BLAST query-anchored formats
 - All sequences are stacked up, and chopped into blocks of ~60 residues
 - Easy for humans to read, but difficult to edit
 - Tools for converting formats are available on the web
 - EMBOSS tool for conversion (`squizz_convert`)
-

Aligned FASTA (A2M) Format

```
>SN29_RAT/142-196
PSSRLKEAINTSKDQESKYQASHPNLRRLHDAE---LDSVPASTV-----NTEVY-----P
KNSSL---R-----A
>SN29_HUMAN/142-197
PNNRLKEAISTSKEQEAKYQASHPNLR-----KLDDTDPVPRGA---GSAMSTDA-YP
KNPHL---R-----A
>SN25_TORMA/95-148
PCNK----LKNFEAGGAYKKVWGNNQD-----G-VVASQP-ARVMD-DREQMA-----M
SGGYI--RRI-TDDA
>O93578/11-59
PCNK----MKS-----GASKAWGNNQD-----G-VVASQP-ARVVD-EREQMA-----I
SGGFI--RRV-TDDA
>SN25_DROME/98-149
PCNK----SQSFK---EDDGTWKGND-----GKVNNQP-QRVMD-DRNGM-----MA
QAGYI--GRI-TNDA
```

- Uppercase and '-' characters are alignment columns. There must be the same number of aligned characters in all sequences.
 - Insertions that are not part of the alignment, are indicated with lower case and '.' characters. These are not read (i.e. they're for humans only)
 - Benefits
 - Easily machine readable
 - Readable by most programs that read FASTA format
-

Graphical - Jalview

- Postscript, PDF, HTML
- Looks pretty and very visually informative
- Completely useless for further computational analysis.

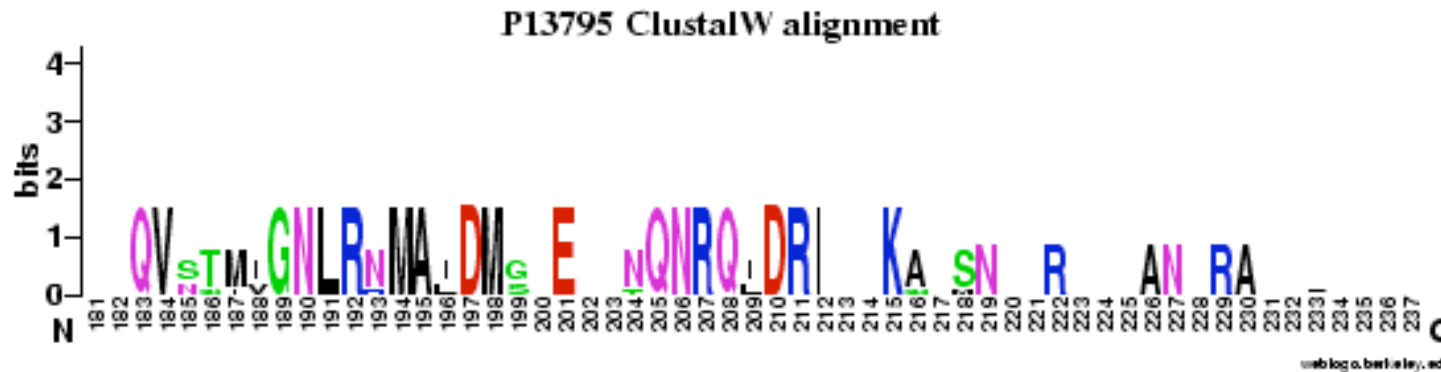
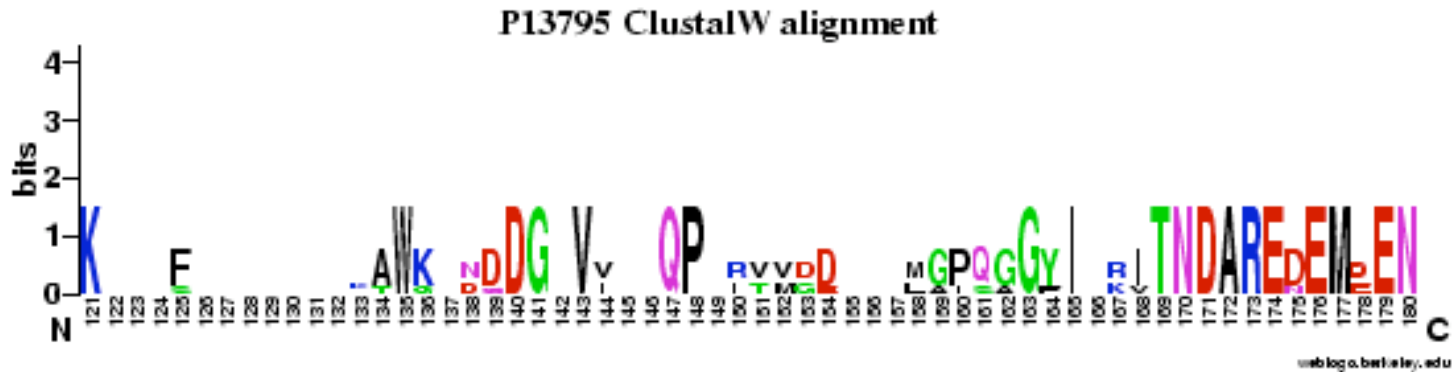
DO NOT SAVE GRAPHICS AS YOUR ONLY OUTPUT

- Jalview -- Java alignment editor (<http://www.jalview.org>)
 - Available as an online applet or as an application
 - Makes nice pictures and allow interactive editing



Sequence Logos

- Logos are another useful visualization of alignments that allow conserved positions to be easily picked out.
- Multiple tools available on the web or can be downloaded:
- <http://weblogo.berkeley.edu>



Tcoffee

- Makes a library of pair-wise global and several local alignments
- Tries to find a multiple alignment that has best consensus with all alignments in the library.
- Still a progressive algorithm
- Slower, but usually a bit better than ClustalW



Other Uses of MSA Servers

- ClustalW can refine an alignment
 - If sequences are aligned when submitted, this info is used.
- Tcoffee can
 - Combine alignments
 - Evaluate alignment quality
 - Use structural information if available



Criteria for a Good MSA

- Most methods align proteins on the basis of sequence similarity, but what we really want to know is:
 - Evolutionary similarity
 - Functional similarity
 - Structural similarity
 - If the sequences are closely related, these similarities are all equivalent. As sequences become more divergent, these similarities may not be equivalent.
 - There isn't necessarily one 'correct' alignment for a family. MSA doesn't *necessarily* reflect a true structural or functional alignment.
-

Which Sequences?

- Don't include too many
 - Problems are VERY slow for many sequences
 - Start with 10-15 or so.
- Closely related sequences are easy to align, but less informative. The converse is true for more distantly related sequences
 - No identical sequences
 - Each sequence 30-70% identical with at least half of the other sequences.



Strategies

- Visually inspect alignment and try eliminating sequences that seem problematic.
- Avoid sequences with long insertions and/or terminal extensions
- “Orphan” sequences (highly divergent members of a family) usually don’t disrupt alignment because they’re the last to be aligned.



Collections of MSAs

- Domain and family collection databases not only have sequences grouped by domain/family, but also have MSAs that were used for classification.
- Example: Pfam <http://pfam.janelia.org/>

