Multiple Sequence Alignment
Modern MSA algorithms, principles and performance

By: Nirvana Nursimulu

Questions

(1) *MAFFT: Multiple Alignment using Fast Fourier Transform*

Recall the MAFFT algorithm. During the presentation, we have focussed on the alignment for 2 sequences only. Describe how alignment of more than 2 sequences can be obtained. (Refer to the appropriate paper in the pre-reading.)

(2) *Questioning alignment algorithms*

Following the presentation on Multiple Sequence Alignment, you have decided to use the "best algorithm out there". Thus, you have aligned your sequences using PROBSCONS. Without so much as giving it a second thought, you have simply copied and pasted your alignment, left it as *final*, and gone off to play Skyrim (http://en.wikipedia.org/wiki/The_Elder_Scrolls_V:_Skyrim). Why should this practice be frowned upon?

(3) *Inferring homology from an MSA*

From a multiple sequence alignment, one can infer homology. You have used PROBSCONS to align a number of sequences, but the homology implied by the program you are using seems unlikely. What are some ways to troubleshoot this?

(4) *Practical use*

Please ccess the multi-fasta file at the following URL: http://www.embl.de/~seqanal/courses/commonCourseContent/sequences/a2msFullLength.fasta We shall be using these sequences (of alpha-2-macroglobulins protein sequences: http://en.wikipedia.org/wiki/Alpha-2-Macroglobulin) to compare T-coffee to Clustal, and evaluate our initial results with MAFFT and MUSCLE.

(a) *Clustal versus T-coffee*

Access the CLUSTAL page at the EBI (http://www.ebi.ac.uk/Tools/msa/clustalw2/) and enter the sequences into the input box. Use the default parameters. Do the same with T-coffee: http://www.ebi.ac.uk/Tools/msa/tcoffee/.

Look at the alignments generated by the different programs. Are they the same? Is there a significant difference if any? If there is any difference, which alignment would you trust more?

Under Result Summary, start Jalview for both alignments to view the conservation patterns. What is a clear difference? What is a reason for this?

Basing ourselves on only one alignment algorithm (whichever of Clustal and T-coffee seems better) demonstrates poor judgement. Let's look at MAFFT and MUSCLE.

(b) *Evaluation of results using MAFFT and MUSCLE*

Access the MAFFT page at the EBI (http://www.ebi.ac.uk/Tools/msa/mafft/) and enter the sequences into the input box. Set "Output format" to "ClustalW".

Access the MUSCLE page at the EBI (http://www.ebi.ac.uk/Tools/msa/muscle/) and enter the sequences into the input box. Use output format "CLUSTALW", and under "More options", set the "Output tree" to have value "from second iteration". Have "Output order" be "aligned".

View the trees for each. What do you notice? Compare with what was output for T-coffee.

Go under "Result Summary" and view the alignments for both MAFFT and MUSCLE using Jalview. Report what you observe.

(c) *In a nutshell…*

…well, what can we say?

Answers

(1) There are 3 methods:

(i)	FFT-NS-1	(very fast, but very approximate and unreliable)
(ii)	FFT-NS-2	(fast)
(iii)	FFT-NS-i	(slow)

Really, the first two are progressive methods, and the last is an iterative refinement method.

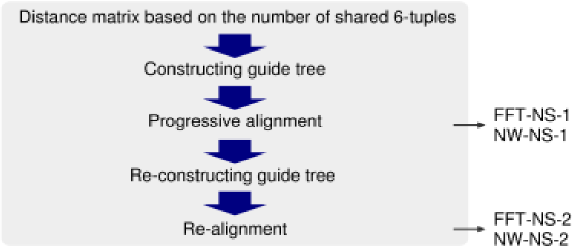
In FFT-NS-1, a guide tree based on the all-pairwise comparison is built. Then, input sequences are progressively aligned following the branching order of sequences in the guide tree. The guide tree is constructed from a distance matrix using the UPGMA method, where the distance matrix is constructed by taking into consideration 6 physico-chemical groups in which amino acids can be grouped.

In FFT-NS-2, the guide tree is re-computed from the FFT-NS-1 alignment. A second progressive alignment is then carried out.

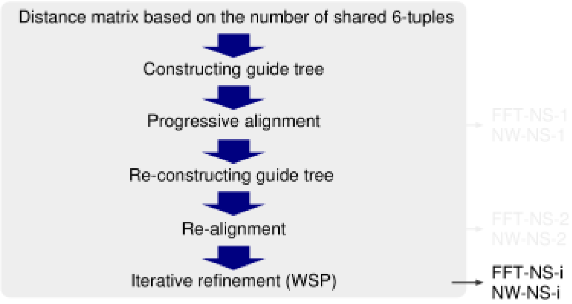
In FFT-NS-i, an initial alignment from FFT-NS-2 is utilized, and subjected to the iterative refinement method as described during the presentation. This may run for as many cycles as desired.

To illustrate (taken from http://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html):

- *Method used for FFT-NS-1 and FFT-NS-2*



- *Method used for FFT-NS-i*


Distance matrix based on the number of shared 6-tuples
Constructing guide tree
Progressive alignment
FFT-NS-1
NW-NS-1
Re-constructing guide tree
Re-alignment
FFT-NS-2
NW-NS-2
Iterative refinement (WSP)
FFT-NS-i
NW-NS-i

---

(2) It is not necessary that the alignment calculated by the algorithm (however superior it is) makes sense biologically. Thus comes the need for manual editing of the multiple sequence alignment. (Note that this is different from the editing described in the answer for question 3.) Here are some of the ways one may proceed:

(i) Reducing the number of indels.
Indels have lower probability of happening than substitutions. Also, in the case of nucleotide sequences, for instance, an indel shifts the reading frame, and this may be less favoured as this may cause any resulting protein to be non-functional.

(ii) Moving indels to more plausible regions (amino acid sequences).
Indels may have been placed in less plausible regions simply to maximize sequence similarities that are actually not that meaningful. By “plausible”, we refer to the fact that it is more likely that in a related sequence, it is more likely that at the same position there will be an amino acid with a similar property (ie, acidic, basic, hydrophobic, polar).

(iii) Conserving motifs (amino acid sequences).
We would prefer secondary structure to not be disrupted since it often entails function. Therefore, say a glycine is involved at a crucial point, we would expect it to

be conserved amongst the different sequences. Then, we would try to align the glycine's of the different sequences in the same column.

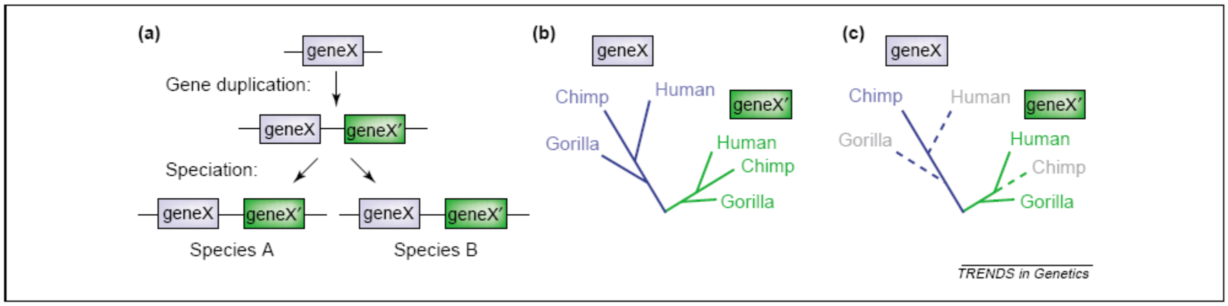
---

(3) Ways of troubleshooting:

(i) Include more sequences.

(ii) Most importantly: did you edit the alignment before sending it out for phylogenetic reconstruction?

(i) If sequences are missing, what are truly paralogues may be misinterpreted as homologues. A solution is to have as many sequences as we can when constructing the multiple sequence alignment. Consider the following for an example (taken from "Phylogeny for the faint of heart: a tutorial" by Sandra L. Baldauf: TRENDS in Genetics, Vol. 19, No. 6, June 2003)



TRENDS in Genetics

Fig. 3. The problem with paralogues. (a) Paralogous genes are created by gene duplication events. Gene X is duplicated in a common ancestor to species A and B resulting in two paralogous genes, X and X′. All subsequent species inherit both copies of the gene (unless one or the other is lost somewhere along the way). (b) Phylogenetic analysis of the X/X′ gene family gives two parallel phylogenies. All sequences of gene X are orthologues of each other, and all the sequences of gene X′ are orthologues of each other. However, X and X′ are paralogues. Both the X and X′ subtrees show the true relationships among the three species. The subtrees are also each other's natural outgroup, and as a result each subtree is rooted with the other (reciprocally rooting). (c) A tree of the X/X′ gene family can be misleading if not all the sequences are included (because of incomplete sampling or gene loss). If the broken branches are missing, then the true species relationships are misrepresented.

(ii) Before sending out the alignment for phylogenetic reconstruction, manually edit the alignment. To get the phylogenetic tree of interest, we want to compare sequences that are different but not too different at the same time. If there is too great a difference, any resemblance will not be adequately captured by the program at use; if the similarity is too great, we are not gaining any information by performing phylogenetic reconstruction. Thus, our goal is to edit the alignment such that we keep only parts of sequences that have similarity, while being careful to "conserve" the differences we may be observing.

Hence, uncertain alignments should be removed, as well as large areas of dissimilarity (this may include frayed C- and N-termini).

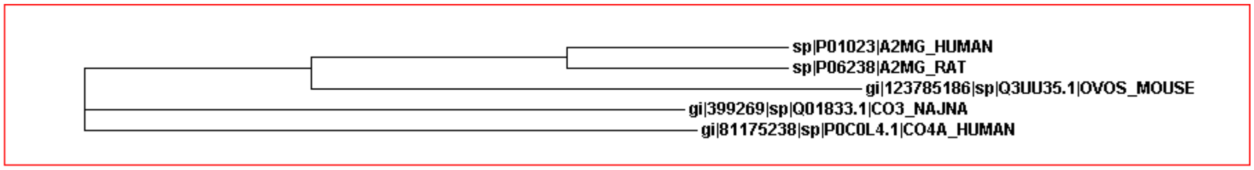
---

(4)

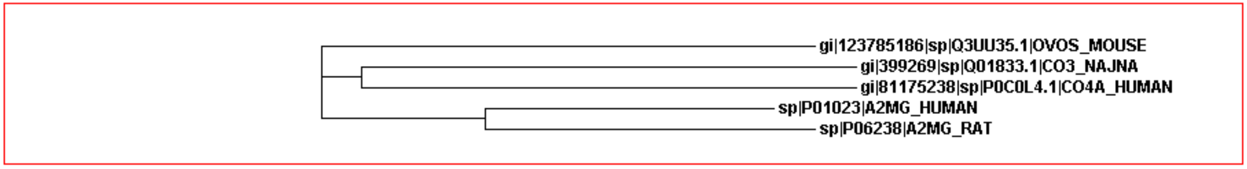
(a) *Clustal versus T-coffee*

Why yes, of course the alignments are different!!! This is obvious from the alignment itself: in Clustal, the first few residues are not aligned unlike with T-coffee. When viewing the guide tree, there are some clear differences.

- *CLUSTAL*


sp|P01023|A2MG_HUMAN
sp|P06238|A2MG_RAT
gi|123785186|sp|Q3UU35.1|OVOS_MOUSE
gi|399269|sp|Q01833.1|CO3_NAJNA
gi|81175238|sp|P0C0L4.1|CO4A_HUMAN

- *T-coffee*


gi|123785186|sp|Q3UU35.1|OVOS_MOUSE
gi|399269|sp|Q01833.1|CO3_NAJNA
gi|81175238|sp|P0C0L4.1|CO4A_HUMAN
sp|P01023|A2MG_HUMAN
sp|P06238|A2MG_RAT

While some of the pairs of sequences that have been marked to be most similar are the same in both alignments, the distances showing relatedness are clearly different.

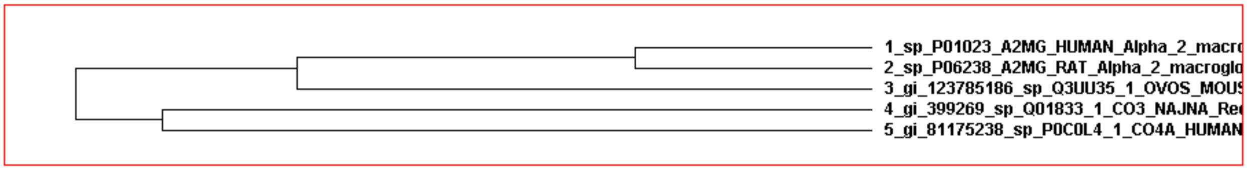
The second alignment would be trusted more than the first, because, as discussed during the presentation, Clustal does not really consider the information from all the sequences when constructing the guide tree.

Once Jalview has been started, a clear difference is in the consensus sequence displayed. The consensus sequence picked out by T-coffee is longer (1914 versus 1823). However, a common pattern is that at the end, there is an area of low conservation, followed by high conservation, but then by low conservation. Furthermore, while the conservation patterns are slightly different, one sees the match at position 1536 in the Clustal alignment and at position 1626 in the T-coffee alignment. A visual inspection reveals that this difference in the position of the match is because of a great number of insertions, for example at position 712 in the T-coffee alignment.

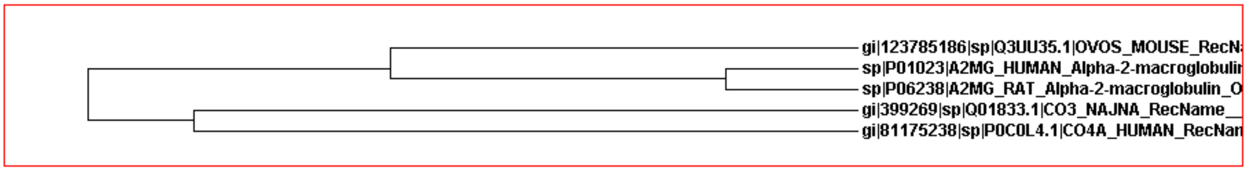
(b) *Evaluation of results using MAFFT and MUSCLE*

The trees demonstrate the same kinds of relationships with the same relationships between distances (even if the distances are different)! The same kinds of relationships are also seen as those reported by T-coffee, though more clusters are defined through MAFFT and MUSCLE.

- *MAFFT*


1_sp_P01023_A2MG_HUMAN_Alpha_2_macr
2_sp_P06238_A2MG_RAT_Alpha_2_macroglo
3_gi_123785186_sp_Q3UU35_1_OVOS_MOUS
4_gi_399269_sp_Q01833_1_CO3_NAJNA_Re
5_gi_81175238_sp_P0C0L4_1_CO4A_HUMAN

- *MUSCLE*


gi|123785186|sp|Q3UU35.1|OVOS_MOUSE_RecN
sp|P01023|A2MG_HUMAN_Alpha-2-macroglobuli
sp|P06238|A2MG_RAT_Alpha-2-macroglobulin_O
gi|399269|sp|Q01833.1|CO3_NAJNA_RecName_
gi|81175238|sp|P0C0L4.1|CO4A_HUMAN_RecNa

The consensus sequence lengths are smaller than those reported by T-coffee (1860 for MUSCLE and 1869 for MAFFT). The conservation pattern at the end is seen again: with a region of low conservation, followed by high conservation, and finally low conservation. A visual inspection reveals that MUSCLE and MAFFT show similar results at this end (compare especially at the beginning of the high conservation region).

(c) *In a nutshell…*

The different algorithms perform differently in a few different ways. However, the results we obtain are very similar in nature. Maybe this is because the protein in question (alpha-2-macroglobulins) is highly conserved as its function is crucial. To assess the quality of the alignments as such, one would have to factor in biological context, and as was described previously, motifs should be sought to be conserved. This goes slightly beyond the scope of this exercise as more information would need to be sought regarding the protein of interest, but the key points should have been grasped.

Furthermore, a point that may or may not be important is that, at the end, T-coffee and Clustal have highly similar trees, and that MUSCLE and MAFFT have highly similar trees too. This may be because, while superior, T-coffee still uses the technique of progressive alignment as Clustal. While fundamentally different, MUSCLE and MAFFT both use progressive alignment followed by iterative refinement. This may be why both MUSCLE and MAFFT reveal more clusters than T-coffee or Clustal.

The end! ♪♫