LAB 3 — MULTIPLE SEQUENCE ALIGNMENT

[Software needed: MEGA X – see page 11 for download link, and web access]

In this lab we will learn how to align sets of nucleotide sequences from different organisms to identify their similarities and differences. Multiple sequence alignment (MSA) is an extremely powerful methodology underlying many bioinformatic analyses. It can be used in conjunction with a wide range of evolutionary and functional analyses to identify everything from evolutionary relationships, to functional motifs and interaction domains. Reliable MSAs identify homologous residues among related sequences, data which are absolutely essential for phylogenetic analyses and motif identification.

There are many methods available to produce MSAs. Fortunately, producing a MSA is very easy. Unfortunately, it is often just as easy to produce a poor MSA as a good MSA, and a bad MSA can weaken or completely compromise your analysis. In this lab, you will become familiar with a few of the MSA applications available. Additionally, you should also come to understand that producing a good MSA is a learned skill, and very often requires more than a simple plug-and-play approach. While all good MSA applications will give you their 'best' solution, often, some manual adjustment and biological intuition will produce an even better alignment.

There are two major classes of MSA: global alignments and local alignments. You need to be familiar with the difference between these two, and use the application that is most appropriate for your particular needs. The single biggest mistake made with MSAs is assuming there is one tool that works equally well for all jobs. For example, most people simply use Clustal for their MSAs. While Clustal is a powerful application, it is completely inappropriate for a large number of alignment problems.

Box 1. Global and Local Alignments

Global alignments are those in which the set of sequences to be aligned are similar across their entire length. It is important to recognize that while global alignments assume you can align the whole sequence, they do permit gaps and differences in lengths (i.e. when some sequence are much longer or shorter than others).

Local alignments are those in which there is similarity only at particular sub-regions of the sequence. Use a local alignment, for example, when part of the sequence is conserved, yet the rest has diverged so much that no similarity remains.

The figures below (adapted from the MAFFT site) nicely illustrate the differences between global and local alignment. In each figure, "X"s indicate alignable residues, "o"s indicate unalignable residues and "-"s indicate gaps,

A global alignment would be appropriate in this case since the sequences are similar across their entire length (from one end to the other):

```
                                XXXXXXXXXXX-XXXXXXXXXXXXXXX
                                XX-XXXXXXXXXXXXXX-XXXXXXXX
                                XXXXX----XXXXXXXX---XXXXXXX
                                XXXXX-XXXXXXXXX----XXXXXXX
                                XXXXXXXXXXXXXXX----XXXXXXX
```

A global alignment will also work in the following case when using most good global alignment applications, even though the sequences are of very different lengths:

```
        ooooooooooooooooooooooooooooooXXXXXXXXXX-XXXXXXXXXXXXXX----------------
        --------------------------------XX-XXXXXXXXXXXXX-XXXXXXXoooooooooo-------
        -----------------oooooooooooooXXXXX----XXXXXXXX---XXXXXXXoooooooooo-------
        --------ooooooooooooooooooooooXXXXX-XXXXXXXXX----XXXXXXXoooooooooooooooo
        --------------------------------XXXXXXXXXXXXXXX----XXXXXXX----------------
```

The following sequences would require a local MSA approach. Note that there are sub-regions within each sequence that can be aligned, but the overall sequences cannot. Additionally, these conserved regions are not found in all sequences:

```
ooooooooXXX------XXXX--------------------------------XXXXXXXXXX-XXXXXXXXXXXXXXXoooooooooooo
---------XXXXXXXXXXXXXooo----------------------------XXXXXXXXXXXXXXXXX-XXXXXXXX-------------
-----oooXXXXXX---XXXXoooooooooo---------------------XXXXX----XXXXXXXXXXXXXXXXXXoooooooooooo
---------XXXXX----XXXXoooooooooooooooooooooooooooooXXXXX-XXXXXXXXXXX--XXXXXXX-------------
-------------------------------------------------XXXXX---XXXXXXXXX--XXXXXXXooooo--------
```

Clustal

Clustal is by far the most popular MSA application. It is a global alignment tool that can be run on nearly any computer platform. It is also available through a number of web servers, and is integrated into a wide range of bioinformatic packages. We will use the Clustal implementation in MEGA X, which is a very powerful and easy to use phylogenetics package.

1. Download a FASTA file containing unaligned nucleotide sequences from the Coursera website (download it from the Bioinformatic Methods I "Lab 3 – Multiple Sequence Alignment" tab, the same tab where you retrieved this lab document, using the "**these sequences**" link…note that if you're a Mac user you will need to remove a .txt extension that automatically gets added such that the file name ends with just "…Labs3,4_sequences.fas", otherwise Mega won't recognize the file. This sometimes happens with certain browsers, too.). You can always view any text-based file with simple text editor program. We recommend doing this for any file you are working with, just to confirm that it contains the information you require in the correct format. Crimson or Sublime Editors or others (see **Appendix 1**) are very handy for this kind of thing, and offer some powerful features not found in Notepad.

2. The sequences have been given new headers from the rather lengthy headers they were given by the NCBI download system that we explored in earlier labs to the following format, so that the header information is compact enough to be completely displayed in MEGA's alignment applications:

   >Genus_species_GI
   sequence…

For example, we changed:

```
>gi|42567417:17-484 Arabidopsis thaliana ribosomal protein L11 family…
ACTGTCTA…
```

to:

```
>A_thaliana_42567417
ACTGTCTA…
```

This seemingly time-consuming step illustrates the importance of having some basic programming abilities. If you intend to do any serious bioinformatics you would be well served to spend a little bit of time learning a good scripting language, such as Python. A few minutes of simple scripting could automate this task and finish the job in fractions of a second. Alternately, sometimes such tasks can be accomplished with regular expressions using the search-and-replace functions of a good text editor: see **Appendix 1** for some tips!

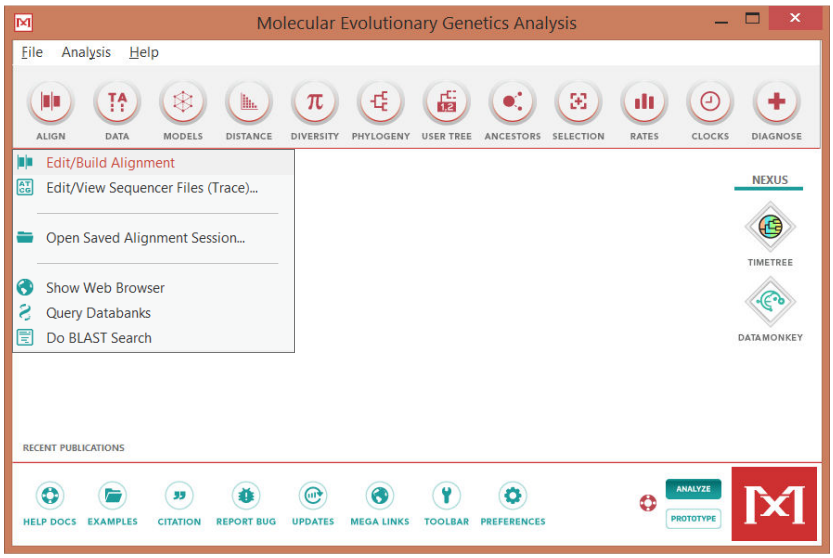


Figure 1. The MEGA X interface with the "Align" menu options shown.

3. Open up the alignment editor of MEGA X (install MEGA X from http://megasoftware.net)
   - Click the **Align** tab (see Figure 1; *don't* use the File menu item!)
   - Select **Edit/Build Alignment**, then **Retrieve sequences from a file** to load your file from Step 1 (if you can't see it in the menu, then make sure a ".txt" extension hasn't been added).
   - Explore your sequences by scrolling to the end. Note that you can make the name field (on the left side of the window) larger in the same way you would increase the size of a table column in Excel or other spreadsheet program.
     a. *What do you notice about the sequence lengths?*
     b. *What do you notice about the present state of the "alignment"?*
   - Click on **Translated Protein Sequence** tab just above the sequences (if MEGA asks if you want to use the current standard genetic code, check "yes")
     c. *What are you looking at now, and how does it differ from what you viewed previously?*
     d. *Save your **unaligned** protein sequences under **Data / Export Alignment > Fasta format** (add something like "prot" to the name before the .fas extension) for use with **MAFFT** later on in the lab.*

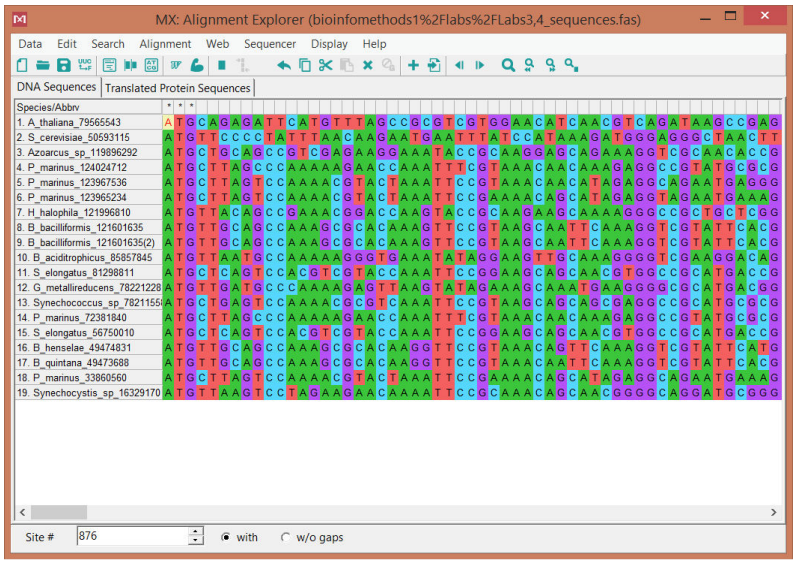


Figure 2. MEGA's alignment viewer.

4. Go back to the **DNA sequences** and start to perform a Clustal alignment
    - You will need to select the sequences to align. Select all by doing **Edit / Select All.**
    - Click on **Alignment/Align by ClustalW**
    - You are looking at the DNA alignment parameters. Review these and try to interpret what they mean.
        a. *What do you think the **Gap Opening Penalty** and **Gap Extension Penalty** mean?*
        b. *Why do you want to control these separately? (We touched on this last lab.)*
    - Cancel out of this menu for the moment.

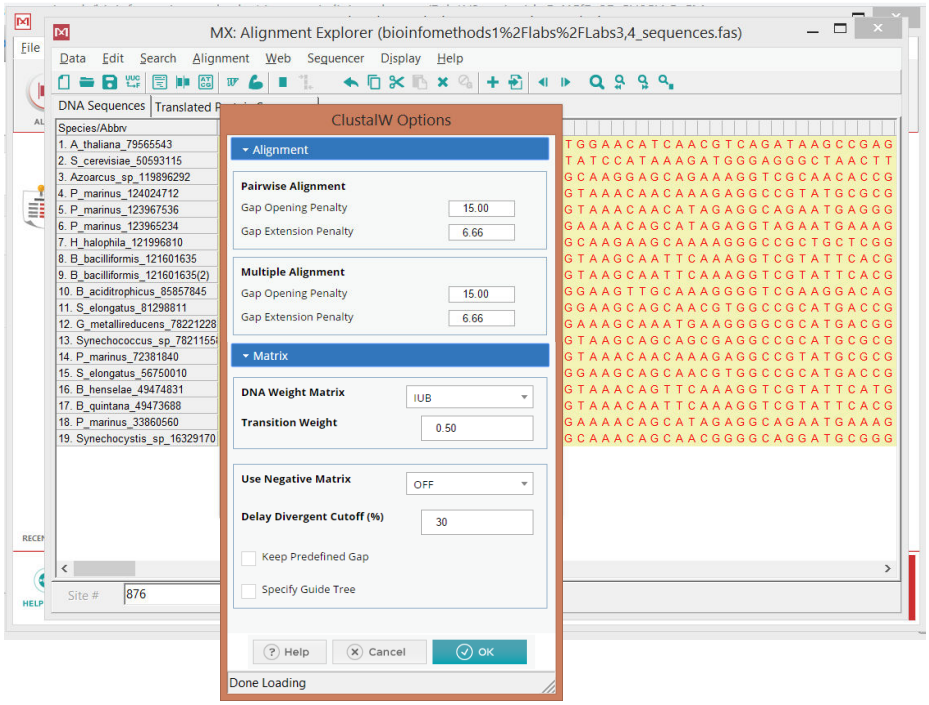


Figure 3. Clustal DNA alignment parameters (both Alignment and Matrix parameter sections are expanded in this figure)

5. Let's return to **Translated Protein Sequences** to do the actual alignment, and click on **Alignment/Align by ClustalW**
    - Select all sequences via **Edit / Select All.**
    - Click on **Alignment/Align by ClustalW**
        a. *How do the protein alignment parameters differ from the DNA alignment parameters?*
    - Change 'gap opening penalty' to 20 in both multiple and pairwise alignment categories, and ensure the 'gap extension penalty' is 0.1 (change if necessary). *Note: Gonnet is another, slightly less common "flavour" of substitution matrix.*
    - Run the alignment by clicking **OK.**
    - You may have to click on the aligned sequences again to see the colour coding
    - You can see more of the alignment by **Display / Font**, and reducing the font size.
    - Note that once you align the translated protein sequences, the corresponding DNA sequences also are aligned, and you can switch back and forth between these views by clicking the tabs above the sequences.

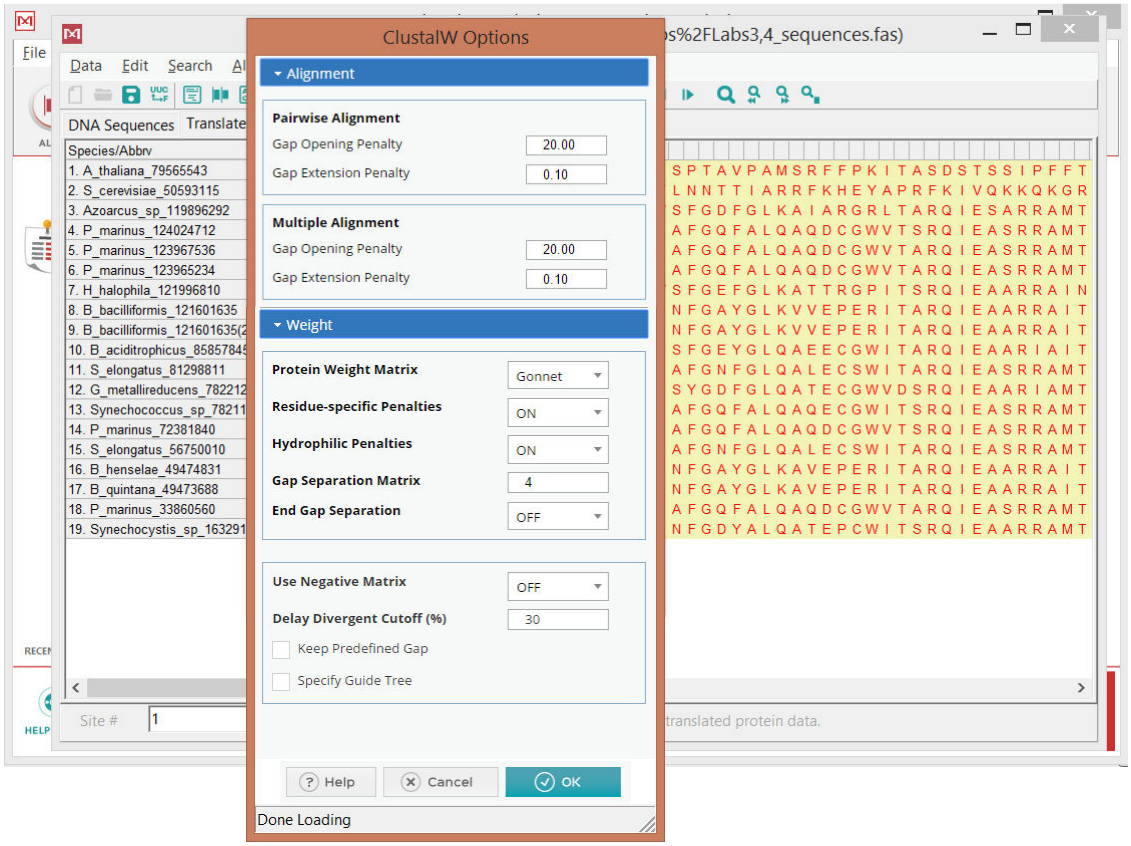


Figure 4. Clustal protein alignment parameters, with Gap Opening parameters set to 20 from their default values of 10 (see Appendix 1 for alignment defaults!). Again, both Alignment and Matrix parameter sections are expanded in this figure).

Scroll through both DNA and translated protein alignments, note consistently coloured bars, which denote conserved residues

- If you look through the aligned translated protein sequences you will notice some alignment columns that are all of a single colour, yet have a number of different amino acids.
  a. *Why is this? Hint: think about the basis of substitution matrices!*
- Typically, Clustal identifies the conservation of the alignment column using the following notation: *****, **:**, and **.** which indicate perfect conservation, strong conservation and weak conservation respectively. You should be aware that not all versions of MEGA follow this convention.
  c. *What general region(s) in the alignment show the strongest conservation? What might this mean?*
  d. *Examine the gaps generated in the alignments. Do they seem "appropriate"?*

Figure 5. Clustal aligned protein sequences (with Display / Font set to size 4)

6. Let's align the translated protein sequences again. Go back to the **Alignment / Align by ClustalW** menu and set the gap open and extension penalties to 100 and 0.1 respectively. Rerun the alignment.
   a. *What changed in the alignment?*
   b. *Are there any sequences in the alignment that seem to be very badly matched to the rest? Which ones? What might this mean biologically?*

Lab Quiz Question 1

7. Now we need to save the alignments. Note that you will save either the DNA or protein alignment depending on which tab is currently being viewed. You can export the sequences in either FASTA or MEGA format. MEGA format is very close to FASTA format for simple data, but permits further information about the sequences to be stored.
   - Export your best nucleotide alignment in FASTA format using a file name along the lines of 'Lab3_dna_clustal.fas'
   - Export your best protein alignment in FASTA format using a file name along the lines of 'Lab3_pro_clustal.fas'
   - You can also do the same in MEGA format using the .meg file extension. You will be prompted for a title when saving in MEGA format. This can be any description of the data in the file.

8. Translating DNA sequences to protein sequences can be a very good way to increase the power of your alignment. Commonly, if you want to continue to work with the DNA sequence you simply translate to protein, align, then back-translate to DNA again. MEGA does this seamlessly, which makes it very powerful. But IMPORTANTLY, you cannot always do this.
    a. *What is an obvious time that you would not want to translate your DNA sequence to protein prior to alignment? Hint – are all DNA sequences protein-coding?*
    b. *Assuming it is legitimate to do the translation, what extremely important issue do you need to pay attention to during the translation? Hint – can you translate a sequence in any reading frame?*

Lab Quiz Question 2

9. Exit MEGA.

DIALIGN

DIALIGN is a very powerful local alignment MSA that is particularly good at identifying short conserved motifs embedded in long unalignable regions. This program has a number of nice features, including automatic translation and back-translation of DNA to protein sequences. DIALIGN comes in a number of different versions, and, in some, the program will present a normalized score below each alignment column representing the quality of that alignment column, and explicitly show you which residues are actually aligned and which are too divergent for a reliable alignment. You can access DIALIGN in a number of ways. You can load and run it locally, or access it through a number of web interfaces.

1. Go to http://dialign.gobics.de/ and select **CHAOS-DIALIGN.**

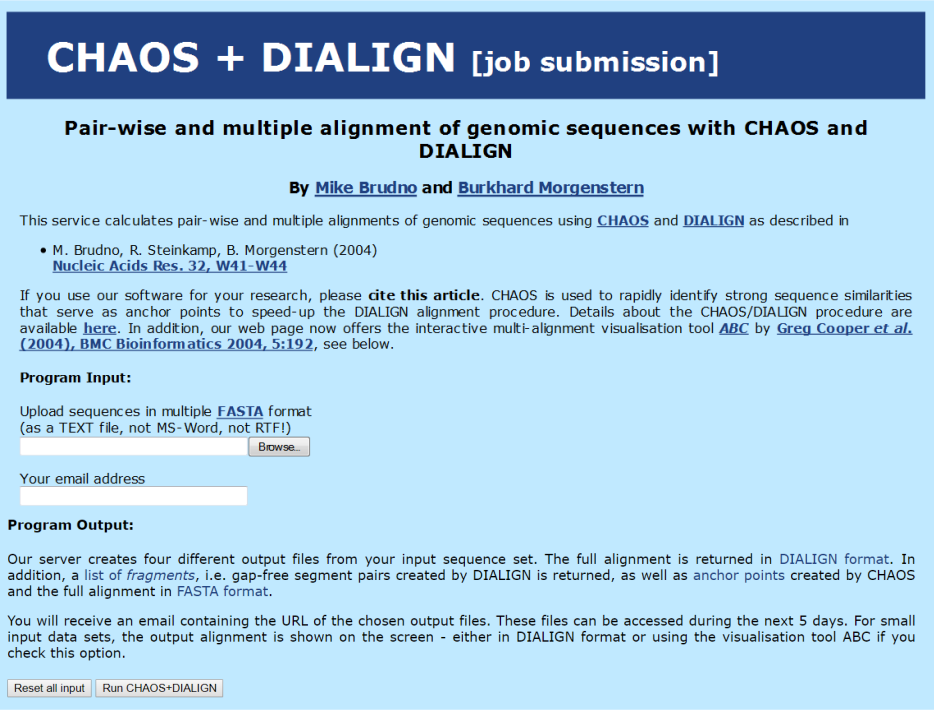


Figure 6. The CHAOS+DIALIGN home page

2. Upload your unaligned **nucleic** acid sequences in FASTA format from above, enter your email address, and click **Run CHAOS+DIALIGN**. *(Note that while Dialign can deal with protein sequences, this particular server is designed to deal with long DNA sequences efficiently).*

3. Click on the "additional output files" link on the DIALIGN output page, and then on the "Full Alignment" in DIALIGN format link, see **Figure 7**.

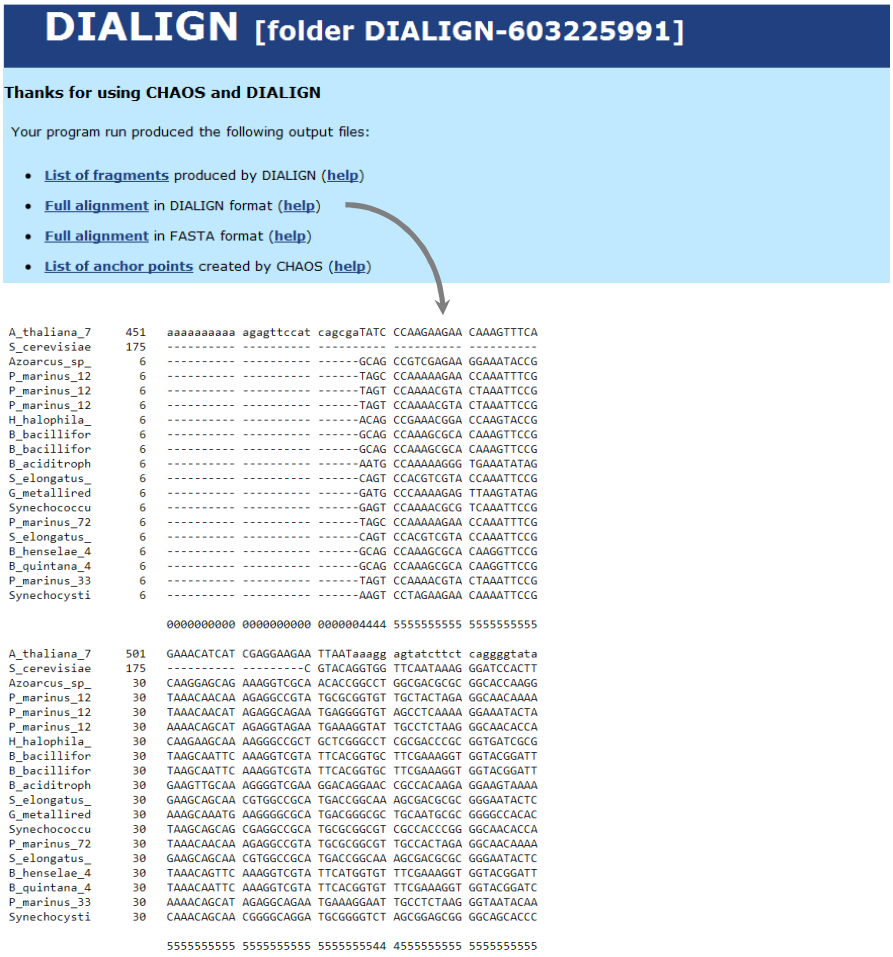


Figure 7. CHAOS+DIALIGN output, in DIALIGN format

4. You can download the alignment file in a number of formats, including fasta. In **DIALIGN format**, there are two important and handy features:
   - Capital letters denote aligned residues, i.e. residues involved in at least one of the "fragments" the alignment consists of. Lower-case letters denote residues not belonging to any of these selected "fragments". These are not considered to be aligned by DIALIGN. Thus, if a lower-case letter is standing in the same column with other letters, this is pure chance; these residues are not considered to be homologous.
   - Numbers below the alignment roughly reflect the relative degree of local similarity among the sequences. The numbers are normalized such that every position gets a value between 0 and 9, 9 for the region of maximum similarity in the alignment

   a. *Are there any obvious differences between the Clustal and Dialign MSAs?*

MAFFT

MAFFT is a remarkably powerful, versatile, and fast application that seems to work quite well on nearly all MSA problems. It provides a number of very significant advantages over other methods; most notably, it is able to automatically adjust its particular choice of algorithms to provide the most appropriate MSA for each dataset. As stated on the MAFFT website, MAFFT offers various multiple alignment strategies, which can be classified into three types, [1] progressive (similar to Clustal), [2] iterative refinement using an objective scoring function to assess the quality of the MSA, and [3] iterative refinement using an even more sophisticated scoring function that is particularly good at dealing with sequence gaps (insertions and deletions). In general, there is a tradeoff between speed and accuracy. The order of speed is 1 > 2 > 3, whereas the order of accuracy is the reverse.

MAFFT is versatile enough to be able to align datasets with very long sequences (~1,000,000bp) as well as datasets with very large numbers of sequences (>50,000 sequences). Furthermore, it performs these extreme alignments generally much faster than any other algorithm.

MAFFT also has the option of incorporating homologous sequences in order to increase the accuracy of the alignment. This is called Mafft-homologs, and is nicely described in this figure from the MAFFT website.

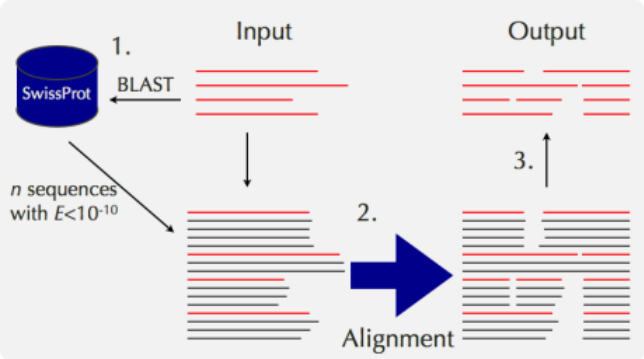


Figure 8. MAFFT-homologs. (1) Collect a number (50 by default) of close homologs (*E*=1e-10 by default) of the input sequences. (2) Align the input sequences and homologs all together using the L-INS-i strategy. (3) Remove the homologs.
(from http://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html)

Finally, MAFFT provides an excellent website that very clearly describes the advantages and applications of the various different methods. The website also provides a web server for the application, which not only does the MSA, but also performs a basic phylogenetic analysis.

1. Go to the MAFFT website http://mafft.cbrc.jp/alignment/server/, and either upload your unaligned **protein** sequence file from the first part of the lab (Step 3), or copy and paste the fasta sequences directly into the window.

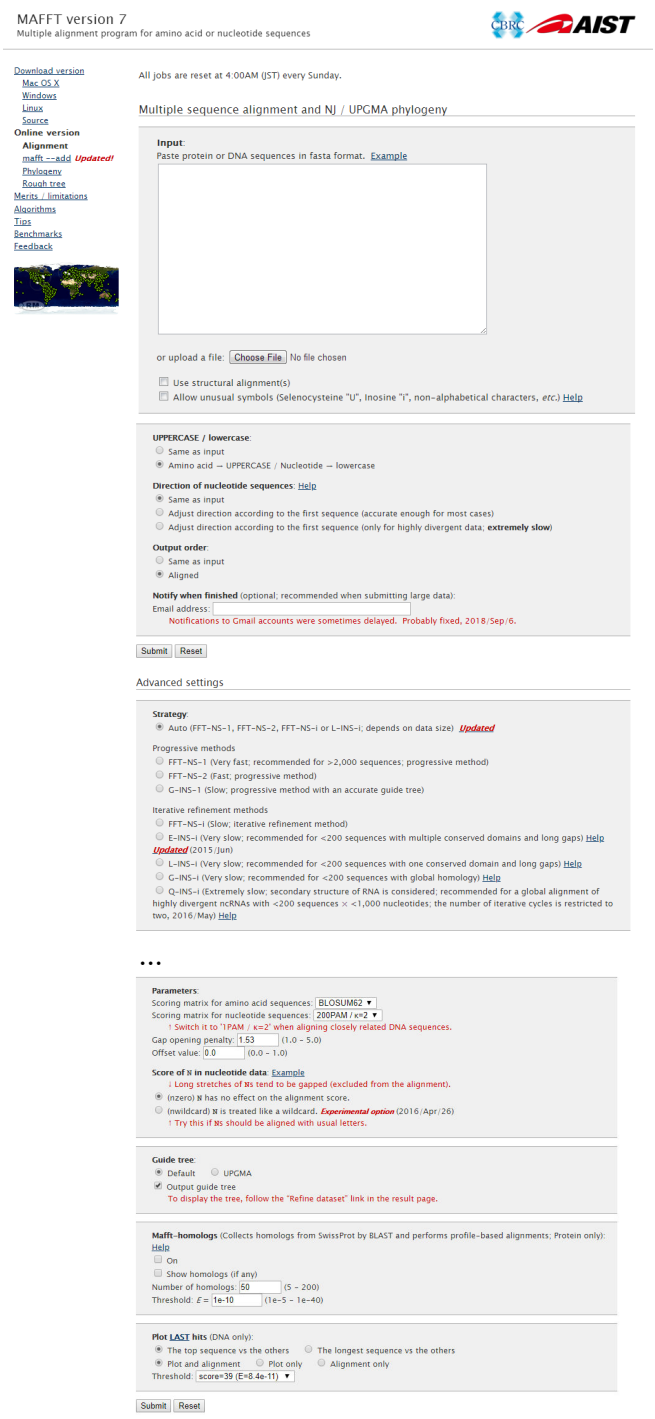


Figure 9. The MAFFT online alignment interface.

- Choose **Auto** under **Strategy**.
- Up the **gap opening** penalty to 3.0 and the **offset value** (gap extension penalty) to 0.2.
- Do not select anything in the Mafft-homologs section, and leave other settings at default.
- Submit the unaligned **protein** file.

2. Notice that MAFFT gives you a Clustal-like output along with the column conservation indicators along the bottom. The aligned FASTA format is presented below, and then below

that is the specific method used. You can also save the data in a number of formats from links at the top of the page, as well as build the phylogenetic tree. MSAViewer and Jalview are nice viewing tools for multiple sequence alignments. To be able to see the MAFFT quality scores check out the Jalview Desktop view – you may need to add http://mafft.cbrc.jp to the Exception Site List using the Configure Java tool). We will be building trees next module.

3. Without closing the previous window, open a new window or tab and return to the MAFFT input page as before.
   - Repeat the same steps, but this time click **On** for **Mafft-homologs** (section at the bottom of the page)
   - Select **show homologs**
   - Set the **number of homologs** to 100, and leave other settings at their default values.
   - Run the alignment.

4. Again, don't close the window, but open a new window or tab and return to the MAFFT input page as before.
   - Repeat the same steps, but this time select **FFT-NS-1** in the **Strategy (manual) section.**
   - Set the **gap open penalty** to 1, and the **offset value** to 0.
   - Turn off **show homologs**
   - Run the alignment.

5. Play with the parameter settings for a few more alignments to get a feel for how they influence the alignment. Remember to open a new window or tab each time.

6. Examine the different MAFFT outputs.
   - Click the **View** button near the top of the page to open links to MSA viewer options. The MSAViewer works reliably without futzing around with Java settings.
     a. *Does one look better (i.e.: "better" gaps, less random) than the other? Which one?*
     b. *Why does searching for homologs make a better alignment?*
     c. *Describe how the alignments respond to modifying the parameters.*

Lab Quiz Question 3

7. You can save your alignments in a variety of sequence formats either through the MAFFT output page (options near the top of the page) or through the File menu of Jalview. When you save your alignment it is a good idea to make the name informative so that you know what algorithm was used – e.g. save the output under a name such as 'Lab3_pro_MAFFT- FFT-NS-i.fas'.

End of Lab!

Where to get it:
MEGA X http://www.megasoftware.net/

Lab 3 Objectives

By the end of Lab 3 (comprising the lab including its boxes, and the lecture), you should:

- be familiar with how to use Clustal, Dialign, and MAFFT for performing multiple sequence alignments, and know when to use which algorithm;
- know why and when it would be advantageous to align protein sequences over DNA sequences;
- be aware of the difference between progressive and iterative alignment methods;
- know how a multiple sequence alignment is scored using sum of pairs scoring;
- be able to identify conserved residues in an alignment;
- know what the effect of changing gap open and extension penalties will be on an alignment;
- understand why adding homologs might improve alignments.

Do not hesitate to ask the instructor or TA if you do not understand any of the above after reading the relevant material.

Further Reading

Chapters 6 “Blast and Multiple Sequence Alignment” in *Concepts in Bioinformatics and Genomics* by Jamil Momand and Alison McCurdy, Oxford University Press, 2017. pp. 118-125 (ClustalW section).

F Jeanmougin, JD Thompson, M Gouy, DG Higgins, TJ Gibson (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci. 23(10):403-5.

K Tamura, J Dudley, M Nei, S Kumar (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 24(8):1596-9.

M Brudno, R Steinkamp, B Morgenstern (2004) The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences. Nucleic Acids Res. 32:W41-4.

K Katoh, K Misawa, K Kuma, T Miyata (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059-66.

N Essoussi, K Boujenfa, M Limam (2008) A comparison of MSA tools. Bioinformation 2(10):452-5.

Appendix 1: Regular Expressions

In computing, a ***regular expression***, or regex, is a powerful and concise way for matching certain strings of text, such as particular patterns of characters, words etc. Regular expressions are available in many search and replace functions in such programming languages as Perl, Ruby and Python, but are also available in many good text editors, such as Crimson (available for free at http://www.crimsoneditor.com/), Sublime (free trial version at http://sublimetext.com), or Atom (open source at http://atom.io – slightly trickier to install).

As an example, to use a regular expression in order to automatically change the string:

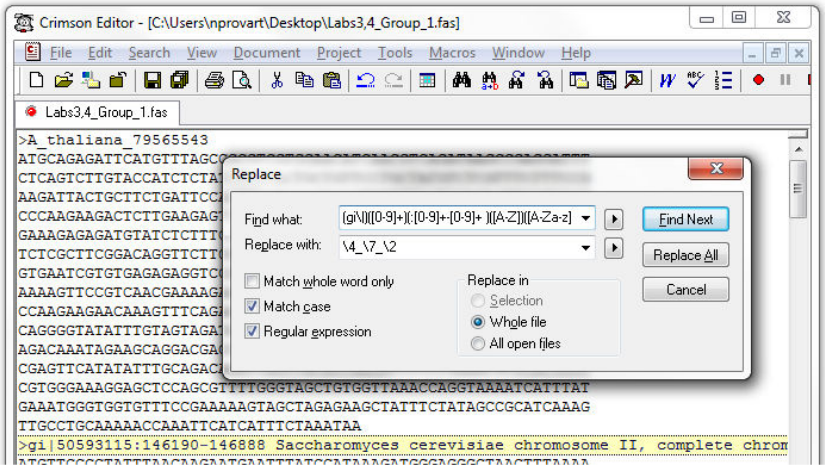
```
>gi|42567417:17-484 Arabidopsis thaliana ribosomal protein L11 family protein
(AT4G35490) mRNA, complete cds
```

to the string:

```
>A_thaliana_42567417
```

use the Replace function, check the "Regular expression" and "Match case" options and enter:

Find what: `(gi\|)([0-9]+)(:[0-9]+-[0-9]+ )([A-Z])([A-Za-z]*)( )([A-Za-z]+)(.*)`
Replace with: `\4_\7_\2` (or `$4_$7_$2` in Atom)



Explanation

The idea is to group the regular expression for matches with parentheses, and then use "backreferences" to these groups to denote the ones you want to keep using the replace function:

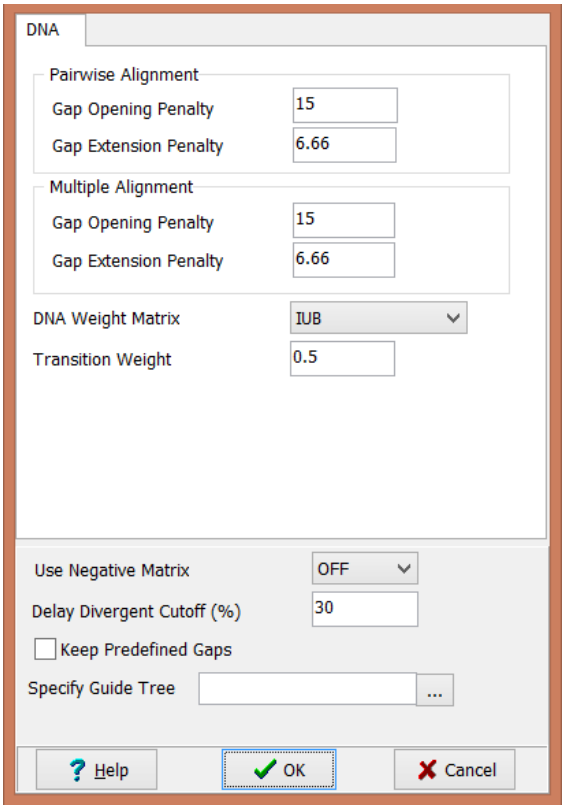
```
 1      2         3                    4        5           6   7           8
(gi\|)([0-9]+)(:[0-9]+-[0-9]+ )([A-Z])([A-Za-z]*)( )([A-Za-z]+)(.*)
```

`\4_\7_\2` *(Crimson or Sublime) or* `$4_$7_$2` *(Atom). Interpretation: when replacing the found text, use the matched text in the 4th group, followed by "_", then the matched text from the 7th group, then "_", finishing with the matched text from the 2nd group.*

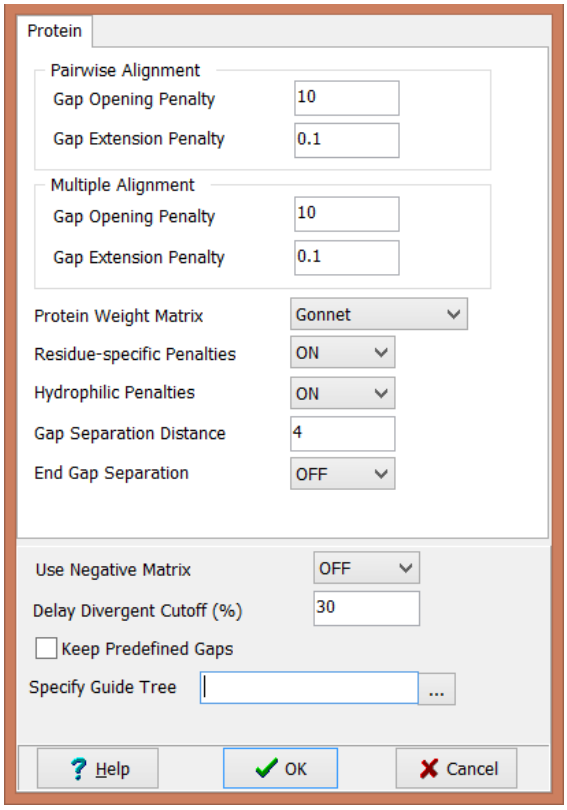
There's a ton of stuff you can do with regexes, even in a text editor! For more information – on what [A-Za-z]+ matches for example – see http://www.regular-expressions.info/tutorial.html.

Appendix 2: Clustal Alignment Defaults in MEGA

MEGA doesn't have a "use defaults" button, so if you change some parameters and want to go back to the default parameters, here they are:



Default Settings for DNA Alignments



Default Settings for Protein Alignments