Using ClustalX for multiple sequence alignment

Jarno Tuimala

December 2004

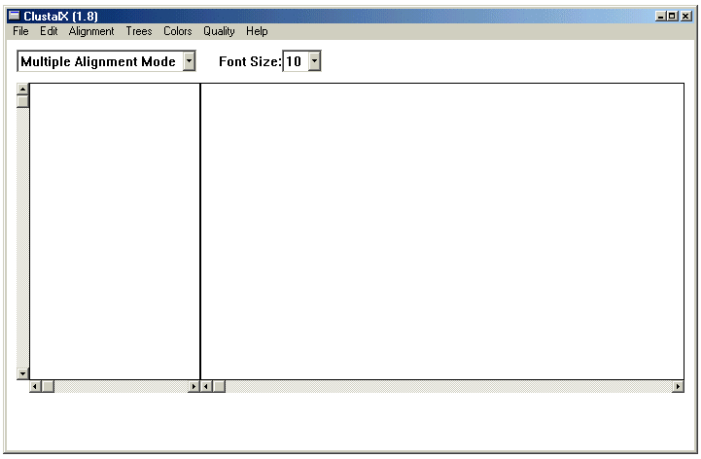
Index

Quick Start

In order to make a multiple sequence alignment using ClustalX, you should have your sequences in FastA format. If you do not know haw to do this, check the chapter "Creating the input file for multiple sequence alignment".

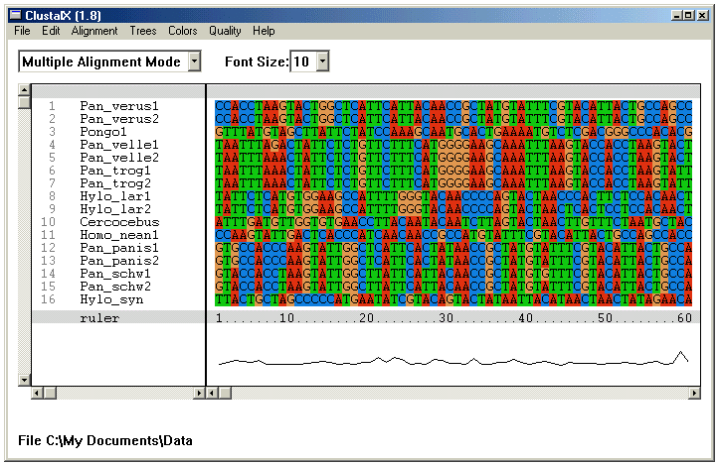
1. Open ClustalX

After starting ClustalX, and you will see a window that looks something like the one below.



2. Read in the FastA-formatted sequences

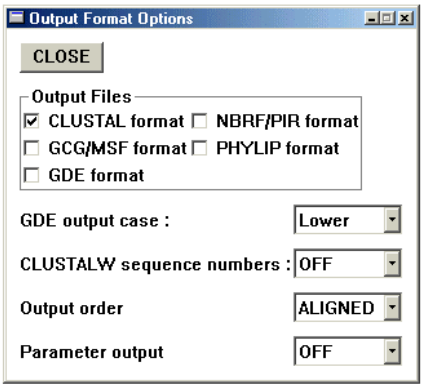
Pull down the File-menu, and choose Load Sequences menu item. Navigate to the folder (subdirectory) that contains the input file (text-file containing the sequences in FastA-format), and choose that file. Sequences should appear in the ClustalX window.



The left pane (in the figure above) lists the sequences according to the name that follows ">" symbol in the input file. The right pane shows the beginning of each sequence. You can scroll to the right to see the rest of each sequence by using the scroll bar at the bottom of the pane.

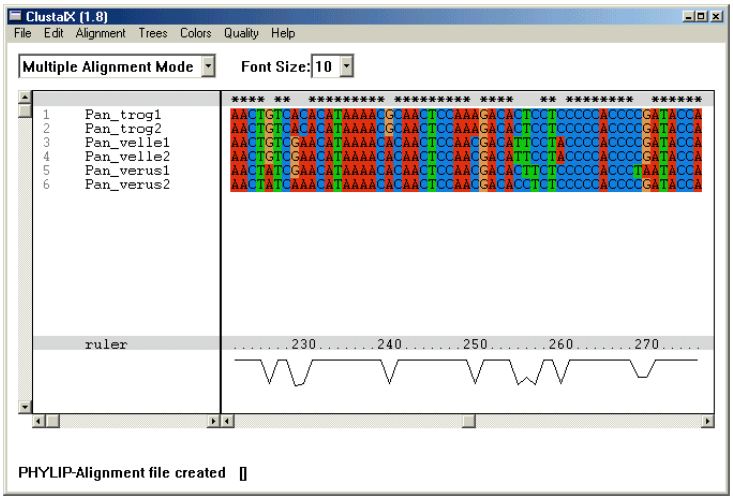
3. Modify the output format option, if necessary

Before aligning the sequences, you should make sure the output format options (from menu Alignment -> output format options) are set correctly. If you'd like to continue with phylogenetic analysis using Phylip package, you should select PHYLIP format. Note, that you should always save the Clustal formatted sequence alignment, also. Here's an example of the output format option settings:



4. Create an alignment

In order to make the actual alignment, select "Do complete alignment" from the menu Alignment. At that point ClustalX asks for output file names. Your sequence alignment is automatically saved in those files once the alignment is ready. After the alignment has been successfully calculated, a new view will appear, and it might look something like that:



Now that the alignment has been created, you can close ClustalX, and use the generated alignment files in other programs.

Creating the input file for multiple sequence alignment

Here, ClustalX is going to be used for sequence alignment. It, like any other computer program requires the data it manipulates (the input file) to be in a format it can recognize. You can use your favourite word processor to create the input file, but I use Notepad.

In the previous chapters, we pasted the found sequences into the text editor. Often the unedited files look like this:

```
gi|15146064|gb|AY040893.1| Homo sapiens individual VP37 mitochondrial control region
GGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCA
CGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTC
ATCCTGTTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACCTACTAAAGTGTGTTAATTAATT
AATGCTTGTAGGACATAATAATAACAATTG  gi|15146065|gb|AY040894.1|  Homo  sapiens  individual  VP5
mitochondrial control region
GGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTATGCA
CGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTC
ATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTAATTAATT
```

ClustalX can recognize several formats for the sequences, but we will use the FastA format, because that's the one most easily downloaded from the databanks. The FastA format can be recognized, because the first line begins with the ">" character. This line contains the title of the sequence. The sequence will start from the next line. The ">" character should be followed by one word (only letters or numbers and _), which ClustalX will use as the name for the sequence in the multiple alignment that it creates.

Clustal treats everything between ">" and the first space as the sequence name. I suggest you to save the original title, and just enter the new name (up to 10 characters but not more) for the sequence and one space after that. This way you can still save the Genbank accession number in the same file as the sequences. You will need the accession number, if you're ever going to publish your results, so save them!

For the sake of clarity, we will put a blank line after the first sequence before the second sequence. The order of the sequences in the file is not important.

After these formatting procedures, the aforementioned sequences should look like this.

```
>hs_vp37 gi|15146064|gb|AY040893.1| Homo sapiens individual VP37 mitochondrial control
GGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCA
CGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTC
ATCCTGTTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACCTACTAAAGTGTGTTAATTAATT
AATGCTTGTAGGACATAATAATAACAATTG

>hs_vp5 gi|15146065|gb|AY040894.1| Homo sapiens individual VP5 mitochondrial control
GGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTATGCA
CGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTC
ATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTAATTAATT
```

Save the sequences in this FastA-format as a plain text file (also known as ASCII). In the current Word version XP, the sequences should be saved as plain text, and then the encoding should be changed to MS-DOS in order to make sure that the file works in every system.

Multiple alignment theory

Dynamic programming can be used to align multiple sequences also. It creates an optimal alignment, but cannot be used for more than five or so sequences because of the calculation time. Therefore, progressive method of multiple sequence alignment is often applied.

Clustal performs a global-multiple sequence alignment by the progressive method. The steps include:

a) Perform pair-wise alignment of all the sequences by dynamic programming
b) Use the alignment scores to produce a phylogenetic tree by neighbor-joining
c) Align the multiple sequences sequentially, guided by the phylogenetic tree

Thus, the most closely related sequences are aligned first, and then additional sequences and groups of sequences are added, guided by the initial alignments to produce a multiple sequence alignment showing in each column the sequence variations among the sequences.

Sequence contributions to the multiple sequence alignment are weighted according to their relationships on the predicted evolutionary tree. Weights are based on the distance of each sequence from the root. The alignment scores between two positions of the multiple sequence alignment are then calculated using the resulting weights as multiplication factors.

As more sequences are added to the profile, gaps accumulate and influence the alignment of further sequences. Clustal calculates gaps in a novel way designed to place them between conserved domains. Gaps found in the initial alignments remain fixed. New gaps are then introduced into the multiple alignment when more sequences are added, but gaps can never be deleted, only added. Clustal also implements methods, which try to compensate for the scoring matrix (e.g., PAM), expected number of gaps, and differencies in sequence length.

Clustal has advanced options:

a) Add sequences with weight
b) Add weights to different sequence positions
c) Add a sequence or alignment to an alignment
d) Use user-defined tree for alignment

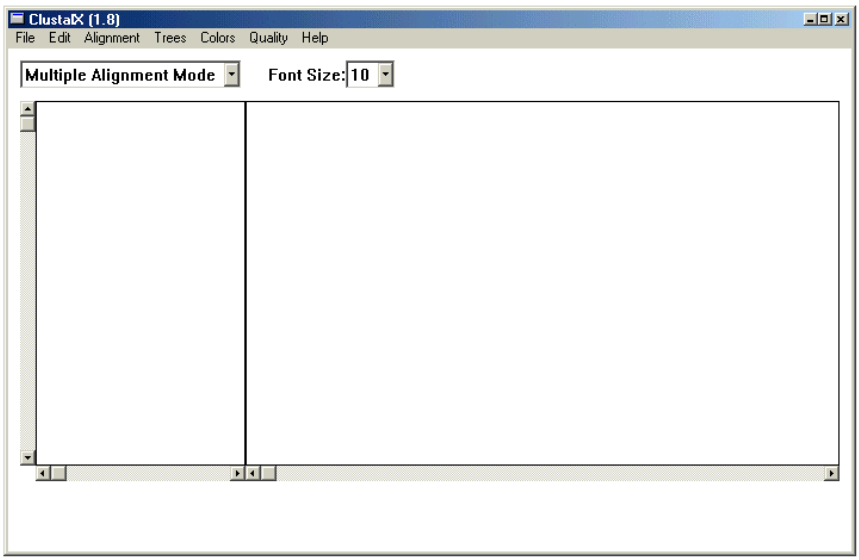
Some of these will be discussed in the next chapters.

The problem with progressive alignment is the dependence of the ultimate multiple sequence alignment on the initial pair-wise alignments. The very first sequences to be aligned are the most closely related on the sequence tree. If these sequences align very well, there will be few errors in the initial alignments. However, the more distantly related these sequences, the more errors will be made, and these errors will be propagated to the multiple sequence alignment. A second problem with the progressive alignment

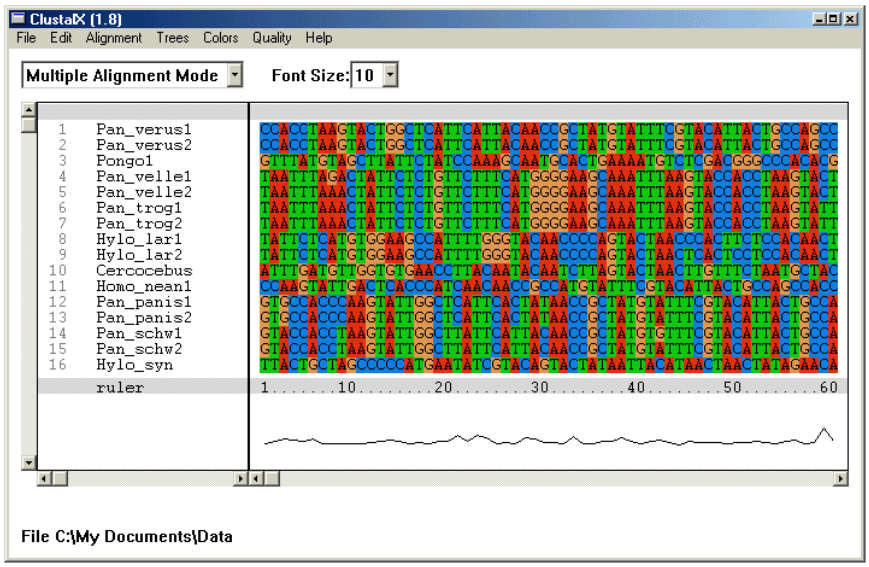
method is the choise of suitable scoring matrices and gap penalties that apply to the set of sequences.

Getting the data into ClustalX

Start ClustalX, and you will see a window that looks something like the one below.



Pull down the File-menu, and choose Load Sequences menu item. Navigate to the folder (subdirectory) that contains the input file (text-file containing the sequences in FastA-format), and choose that file.



The left pane (in the figure above) lists the sequences according to the name that follows ">" symbol in the input file. The right pane shows the beginning of each sequence. You can scroll to the right to see the rest of each sequence by using the scroll bar at the bottom of the pane.

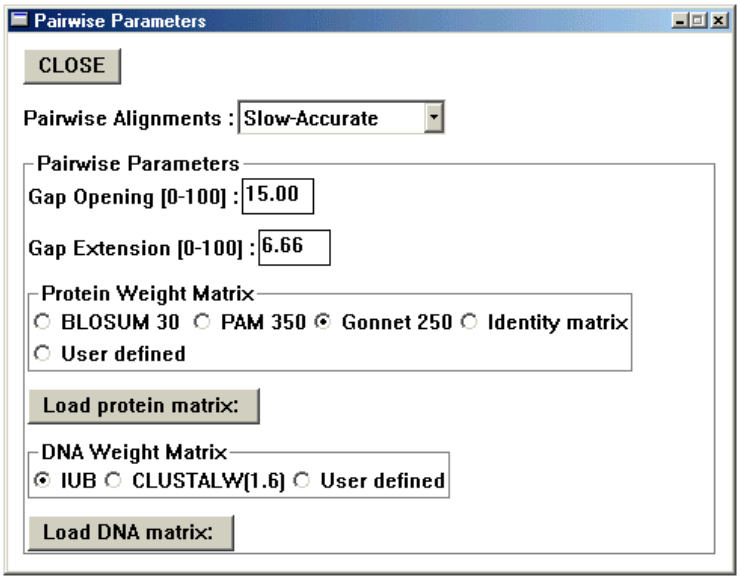
Setting up the alignment parameters

The alignment is done is several succeeding steps: (from Clustal documentation)

1. Reset All Gaps (Alignment->Alignment parameters, Edit->Remove all Gaps)
2. Refine Pairwise Alignment Parameters (Alignment->Alignment parameters)
3. Refine Multiple Alignment Parameters (Alignment->Alignment parameters)
4. Refine Output Format Options (Alignment->Output Format Options)
5. Write Alignment as Postscript (File->Write Alignment as Postscript)
6. Assess the quality of the alignment
    a. Not satisfied -> Go to step 1.
    b. Satisfied -> Refine the alignment by hand

Pairwise alignment parameters

In order to create the pairwise alignment, ClustalX needs to know what penalties to assign for the creation of a gap and for the extension of that gap. Choose Pairwise Alignment Parameters from the Alignment-menu. You will see a dialog box like the one below.



The first choise, Pairwise Alignments, allows you to choose between Slow-Accurate and Fast-Approximate methods. The Slow-method is preferred, but if you are aligning so many sequences or the sequences are so long that the program takes a long time to run,

you may want to use the Fast-method. The Fast-method uses a k-tuple method for pairwise alignment, whereas Slow-method uses a full dynamic programming algorith.

The box shows the default values for Gap Opening and Gap Extension. Decreasing the gap penalties will allow the introduction of more gaps, and less mismatches. This may result is matches that do not reflect homology (identity by descent). Increasing gap penalties wil have an opposite effect, but may result in missing matches that actually are homologies.

Weight matrix parameters can be changed too. The IUB DNA weight matrix scores matches as 1.9 and mismatches as 0, except that it scores all X's and N's as matches to any IUB ambiguity symbols. The Protein Weight Matrices are equivalent to the same matrices used as evolutionary models in the production of the dendrogram. All the matrices have their strenghts and down-sides: PAM has been used for years, but is now somewhat outdated, and the Gonnet can be more appropriate for your purposes. BLOSUM seems to be best for searching databases.

You can create and load in your own matrix into ClustalX. For the description of the file format, take a look at the file matrices.h in the ClustalX-folder:

For amino acids:

```
char *amino_acid_order = "ABCDEFGHIKLMNPQRSTVWXYZ";

short blosum30mt[]={
  4,
  0,  5,
 -3, -2, 17,
  0,  5, -3,  9,
  0,  0,  1,  1,  6,
 -2, -3, -3, -5, -4, 10,
  0,  0, -4, -1, -2, -3,  8,
 -2, -2, -5, -2,  0, -3, -3, 14,
  0, -2, -2, -4, -3,  0, -1, -2,  6,
  0,  0, -3,  0,  2, -1, -1, -2, -2,  4,
 -1, -1,  0, -1, -1,  2, -2, -1,  2, -2,  4,
  1, -2, -2, -3, -1, -2, -2,  2,  1,  2,  2,  6,
  0,  4, -1,  1, -1, -1,  0, -1,  0,  0, -2,  0,  8,
 -1, -2, -3, -1,  1, -4, -1,  1, -3,  1, -3, -4, -3, 11,
  1, -1, -2, -1,  2, -3, -2,  0, -2,  0, -2, -1, -1,  0,  8,
 -1, -2, -2, -1, -1, -1, -2, -1, -3,  1, -2,  0, -2, -1,  3,  8,
  1,  0, -2,  0,  0, -1,  0, -1, -1,  0, -2, -2,  0, -1, -1, -1,  4,
  1,  0, -2, -1, -2, -2, -2, -2,  0, -1,  0,  0,  1,  0,  0, -3,  2,  5,
  1, -2, -2, -2, -3,  1, -3, -3,  4, -2,  1,  0, -2, -4, -3, -1, -1,  1,  5,
 -5, -5, -2, -4, -1,  1,  1, -5, -3, -2, -2, -3, -7, -3, -1,  0, -3, -5, -3, 20,
  0, -1, -2, -1, -1, -1, -1, -1,  0,  0,  0,  0,  0, -1,  0, -1,  0,  0,  0, -2, -1,
 -4, -3, -6, -1, -2,  3, -3,  0, -1, -1,  3, -1, -4, -2, -1,  0, -2, -1,  1,  5, -1,  9,
  0,  0,  0,  0,  5, -4, -2,  0, -3,  1, -1, -1, -1,  0,  4,  0, -1, -1, -3, -1,  0,2,4};

/*
```

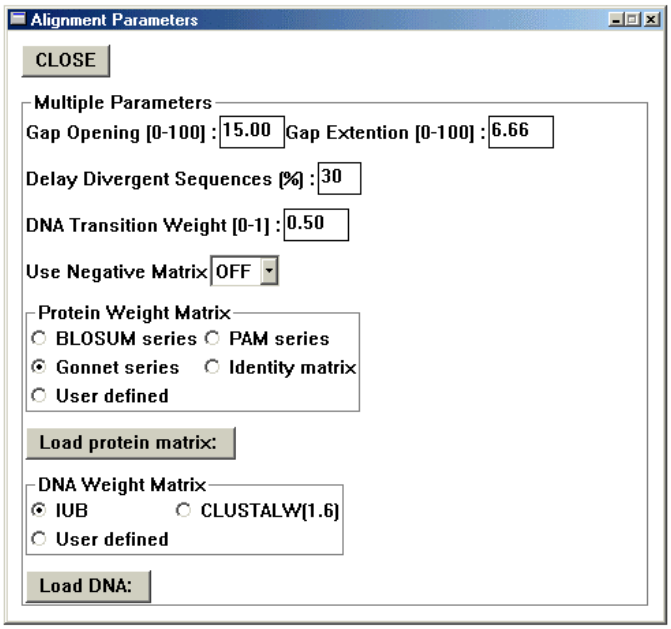
For the DNA:

```
char *nucleic_acid_order = "ABCDGHKMNRSTUVWXY";

short clustalvdnamt[]={
 10,
  0,  0,
  0,  0, 10,
  0,  0,  0,  0,
  0,  0,  0,  0, 10,
  0,  0,  0,  0,  0,  0,
  0,  0,  0,  0,  0,  0,  0,
  0,  0,  0,  0,  0,  0,  0,  0,
  0,  0,  0,  0,  0,  0,  0,  0,  0,
  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, 10,
  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, 10,
  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0};
```

Multiple alignment parameters

Choose Multiple Alignment Parameters from the Alignment-menu. A new dialog box will appear.



The pairwise and multiple alignment options are set independently, because Clustal needs to know both. As have been already discussed, a matrix of pairwise alignments is first calculated, and on the basis of those distances, a multiple alignment is formed. Using different settings for these alignment steps gives us more flexibility to affect how the alignment is done.

Compared to pairwise alignment there are a couple of new settings. Delay Divergent Sequences determines how different two sequences must be in order for their alignment to be delayed. This tries to compensate for the bias introduced by progressive alignment method.

DNA transition weight can be modified. Weight 0 means that transitions are scored as mismatches, and weight 1 mean that transitions are given the same weight as transversions. For distantly related DNA sequences, the weight should be near to zero; for closely related sequences it can be useful to assign a higher score.

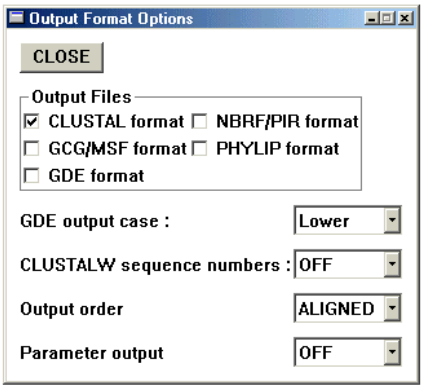
You don't have to choose the individual Weight Matrix any more, because we know how dissimilar the sequences are by now. ClustalX will automatically choose the most appropriate matrix inside one matrix series. So, if you changed the matrix series to be used in the pairwise alignment options, be sure to change it here too.

Alignment output-format

The last thing to be modified before performing the alignment is the output-format. When ClustalX creates an alignment it writes the aligned sequences into a file. There are different multiple alignment formats, which might be needed depending on what program you have planned to use for further analyses.

For the phylogenetic construction using Phylip-package, choose the Phylip as an output format. You should also always remember to write the aligned sequences in the Clustal-format, because it is convenient for both publishing the alignment and refining the alignment in the Notepad. If you plan to publish the sequences, turn the sequence numbering on.

You can modify the output-format by selecting the Output Format Options menu item form the Aligment-menu.



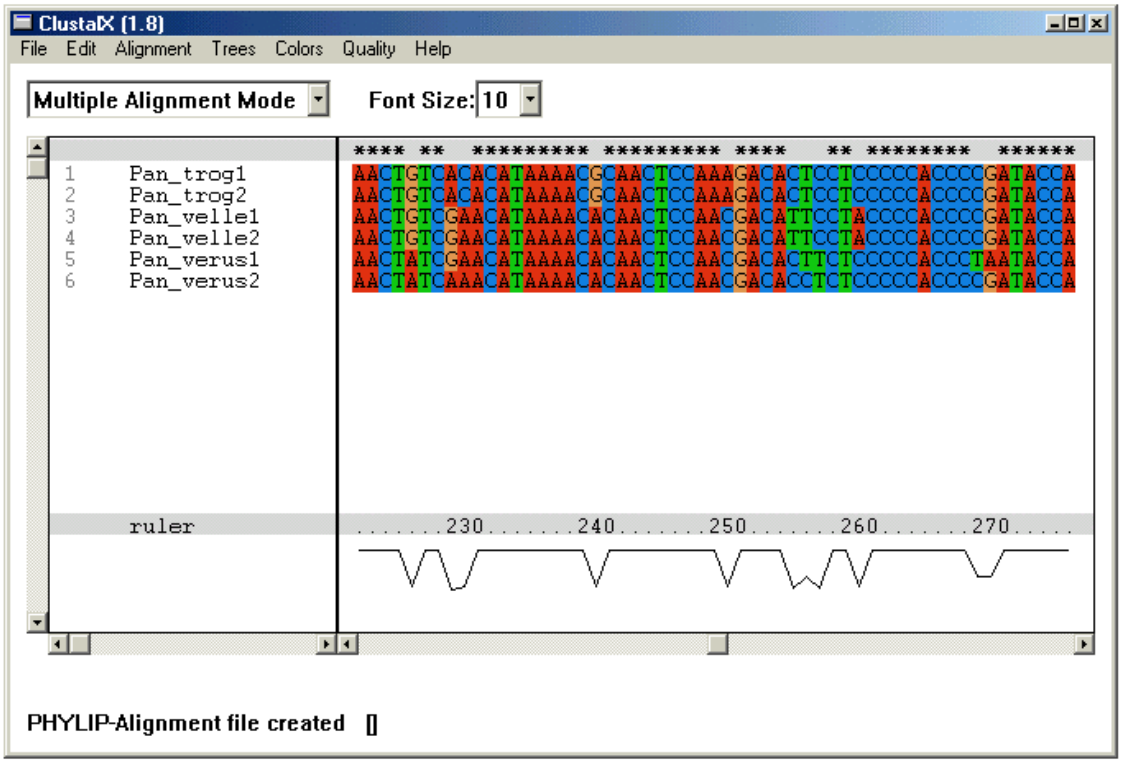
Creating the alignment

Finally, it is time to create the alignment. Choose the Do Complete Alignment under the Alignment menu.

ClustalX tells you what it is doing, but it usually locks the computer, so do not plan to use the computer for any other purposes while the alignment is still under construction.

One word of warning, though. ClustalX does not understand spaces in the names of the folders. Therefore, for example, My Documents, cannot be used in the path.

After the alignment is done, the main window will be updated with the aligned sequences.



The bases are colored, which makes the assessment of alignment that much easier. The histogram (or line) below the ruler indicates the degree of similarity. Peaks indicate positions of high similarity, and valleys positions of low similarity.

The grey line just above the sequences is used to mark strongly conserved positions. The "*" character indicates positions that have been fully conserved.

Writing alignment as Postscript

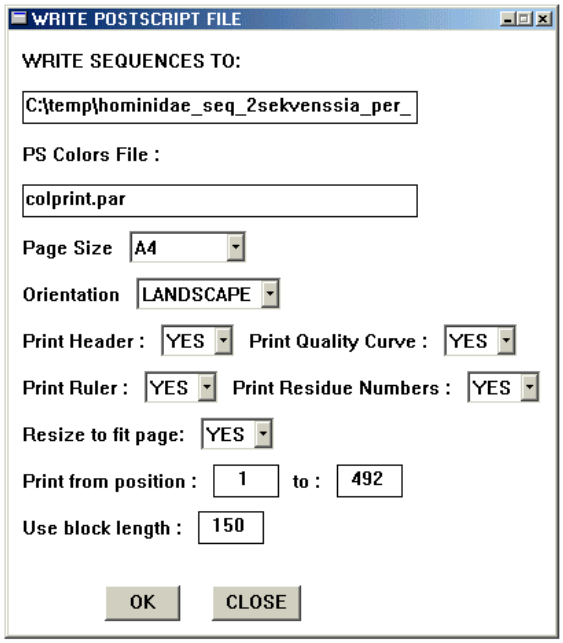
It is possible to use the files we have just created to construct a phylogenetic tree, but the quality and the value of that tree will be no better than the quality of the alignment, and we have not yet considered quality.

It is also important to understand that no matter how dissimilar the sequences are, ClustalX will always produce an alignment. The mere existence of the alignment does not mean that the sequences are related. It is up to the user to ensure that the sequences in the data set are actually homologous and therefore alignable.

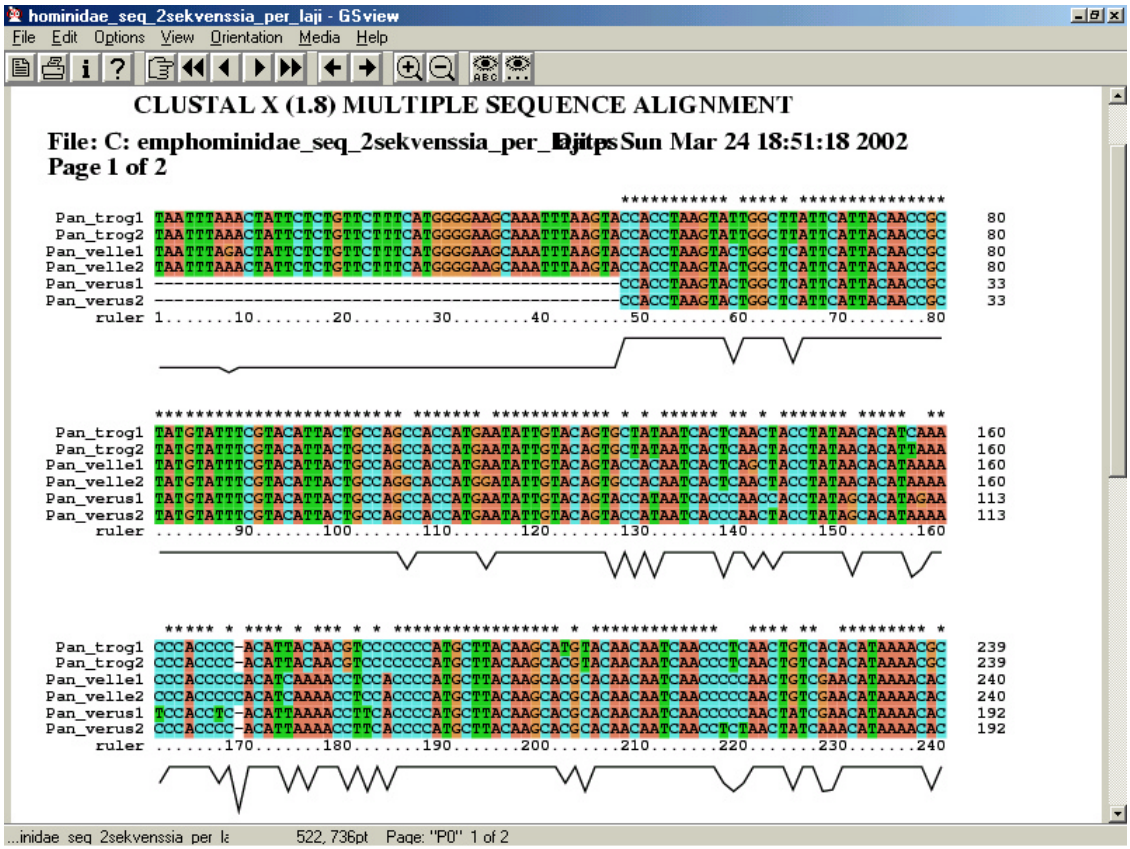
The quality of the alignment is easier to check if there exists hard copies of the sequences. You have two possibilities. You might print the ClustalX-format alignment, which exists in the text-format, and use that, but then you'll lose the information of the colors.

The other possibility is to install Ghostscript and Ghostview programs, which allow the user to manipulate Postscript documents, and print the alignment as postscript, which includes the colors.

Go to File-emu and select Write Alignment as Postscript. A dialog box opens.



You have some options to change, but after you are satisfied with them, click OK. Open the postscript-file in Ghostview. The Ghostview-program might give you some warnings about the incorrect type of the postscript-file, but ignore them by clicking on the Cancel-button.



Now you can print this alignment by selecting the printer-button in the left and top corner of the window.

Assessing the quality of the alignment

In practice, many alignments are produced, and they are compared together. You can start by using the default ClustalX options for pairwise and multiple alignment options. In the next alignments try lowering and highering the gap options by 50%. Produce, say 10-20 different alignments, and then compare those together. Often you need to try more than 10 different settings in order to find the best alignment. What you also need to keep in mind is that alignment is not an absolute thing. It is a best guess according to some algorithm used by a computer program or by an experienced human eye. It is necessary for the user to carefully examine each alignment to see if it makes biological sense.

After printing the alignments you need to examine them to see if most of the gaps make sense. If many of the gaps seem to be arbitrary (i.e., you think you could have done better by eye), then you need to improve the alignment. If there are large regions that are present in only one or two sequences, you may need to delete those regions from the alignment. In practice, most of the programs used for phylogenetic tree reconstruction, will delete all the column containing any gaps from the analysis, so you don't need to bother.

There are some guidelines on how to assess the quality. First remember, that we want to maximize similarity, and minimize dissimilarity. Therefore, the number of gaps is one parameter you should pay attention to. Use the histogram in the bottom of the alignment and the "*" characters above it to assess the conservation of different areas of the alignment.

It is also biologically unsound to assume that there are many gaps with equal spaces between them. Usually gaps are clustered, and are more common in certain areas than in other areas. So, look for the number and length of the conserved blocks of columns. If the pattern of the gaps looks like they have been randomly inserted, choose another alignment. This is, of course, assuming that the sequences are relatively closely related.

If you are aligning an area where there are no functional genes, the above's all you can do. If you have some knowledge on the functional regions of the sequence you are aligning, you can use this information when assessing the quality of the alignment.

The functional regions are often more or less conserved between the relatively closely related sequences. Therefore, quite a few gaps should be inserted into those areas, and most of the gaps should be inserted into less well conserved areas, for example, in the spacer regions between alfa-helixes.

Although it is very time consuming, attempting to improve the alignment through this process of examination and modification of penalties is probably the single most important thing you can do to ensure a high-quality alignment and make a high-quality phylogeny estimation possible.

Advanced alignment strategies

If you have very difficult sequences to align, you can try iterative alignment procedure in order to get a better estimate of the real alignment. First, produce an initial alignment with some quite closely fitting parameter values. Then, produce a new alignment from this initial alignment without removing gaps before this second alignment. You can iterate the alignment using the same settings, but doing the alignment based on the previous alignment multiple times. Sooner or later the alignment will stabilize, and will not change anymore with the same parameters. This is the best, "iterated", alignment for those settings.

Normally, it is important to reset the gaps before producing the alignment with new settings, but this is not done with iterative alignment.

You can also realign only a part of the sequences. Hold down the left mouse button, and paint a selection. Then, from the Alignment menu select Realign Selected Residue Range. ClustalX will then do the alignment using the current setting only for the selected residues. You can also produce a new alignment for selected sequences (Realign Selected Sequences). Sequences can be selected by holding down the left mouse-button, and then dragging downwards in the sequence name list.

Advanced options

Do alignment from the guide tree
If you have some data on the relationships of the sequences, you can construct a tree in the Newick-format, and use that for producing a multiple sequence alignment.

Produce a tree like the one below:

```
(
(
Pan_verus1:0.02428,
Pan_verus2:0.01474)
:0.03203,
(
Pan_velle1:0.00437,
Pan_velle2:0.00579)
:0.01402,
(
Pan_trog1:0.00306,
Pan_trog2:0.00306)
:0.05015);
```

Save the tree in a file in text-format.

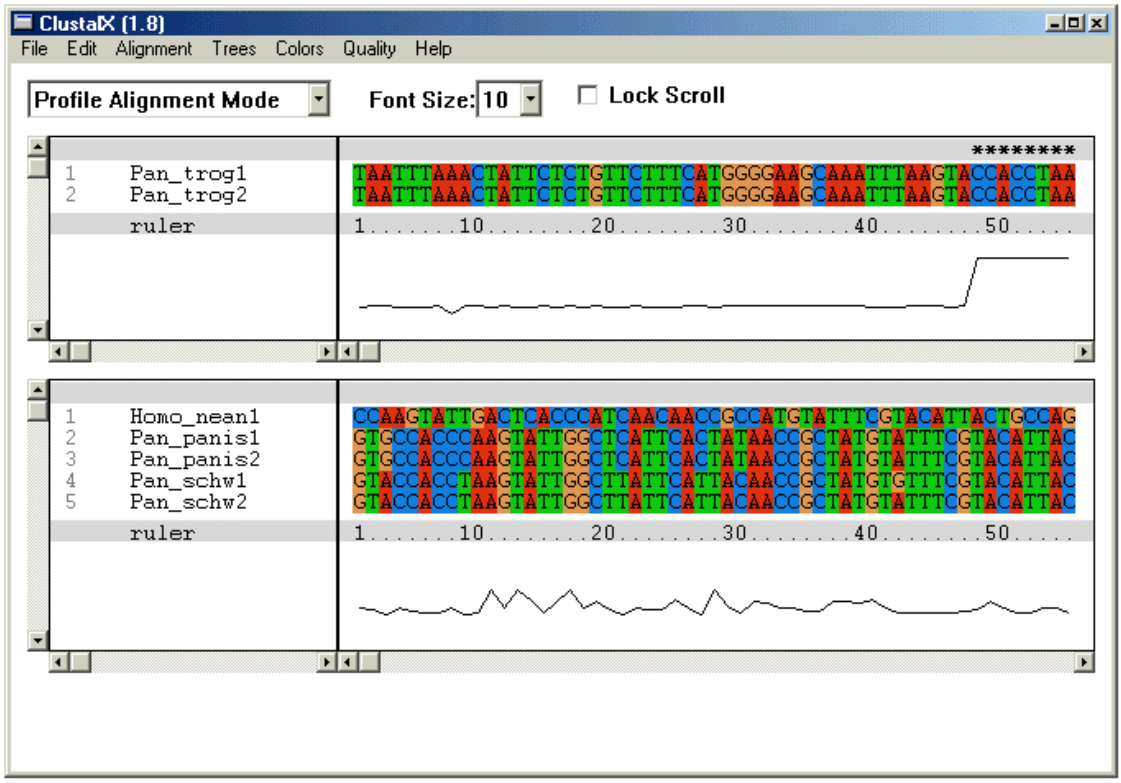
In Clustal, select the Do Alignment From Guide Tree from the Alignment menu. This way you can try to get around the difficulties of the progressive alignment, which might create wrong alignments, if the sequences are very divergent. This can also be used to combine morphological data into the sequence alignment step: the knowledge of the relationships of the taxons can be used to guide the alignment process.

Profile alignment

Profile alignment is used for a couple of purposes, but we first discuss how to align a new sequence into an existing alignment.

There is a pull-down menu on the main window top left-hand corner. From the menu change to Profile Alignment Mode. The Alignment view window is now split into two parts. The upper part contains the alignment we just created, and lower is empty. The upper part is called “profile 1” and the lower part is “profile 2”.

In the File-menu you have the options to load profiles 1 and 2. Now we already have the profile one available, so we only need to load the profile 2. Let’s do that. The main window is updated, and look something like the one below.



From the alignment-menu, select first the option “Align Sequences to Profile 1”. After that select “Align Profile 2 to Profile 1”. This will create an alignment file, where all the sequences are together. Note that you should always first align the sequences in the profile 1 before aligning those (already aligned) sequences to the profile 2!

This is a very handy way to add new sequences into an existing alignment. Otherwise you would have needed to calculate the initial alignment again, which could have been very laborous in the case of many or very long sequences.

Using secondary structure information in the profile alignment

As has been shortly discussed above the sequence alignment can be guided by the secondary structure of the protein, if such information is available. Often such alignments are more biologically plausible than the alignments done "randomly".

In ClustalX the gap penalties are raised at core alpha helix (A) or beta strand (B) residues. The structure information can be used only in the Profile Alignment Mode. These gap penalties cannot be used in the multiple alignment mode. There are two ways to include structure information in Clustal, but here we present only the easier one, which describes the domain areas of the protein. Then the penalties are adjusted in the ClustalX dialog box.

First we need to create the input files. In the first input file, which the first sequence of all the sequences to aligned, a descriptions of the domains (helix or strand) is included. The second input file contains the rest of the sequences in the FastA-format.

The information about the domains is most easily acquired from the SWISS-PROT descriptions. Find the relevant information from http://www.ebi.ac.uk/swissprot/, and after you have acquired the SRS results, click on the Accession Number link on the top of the page. This will take you to the plain text description.

From the description find the lines starting with two capital letters: ID, FT, SQ, and the sequence. Copy those lines into a text file (i.e., into Notepad) and save the file. It should now something like the one below.

```
ID   XRC1_HUMAN     STANDARD;      PRT;   633 AA.
FT   HELIX       315    403       BRCT 1.
FT   HELIX       538    629       BRCT 2.
SQ   SEQUENCE   633 AA;  69525 MW;  30CC2421345ABFC2 CRC64;
MPEIRLRHVV SCSSQDSTHC AENLLKADTY RKWRAAKAGE KTISVVLQLE KEEQIHSVDI
GNDGSAFVEV LVGSSAGGAG EQDYEVLLVT SSFMSPSESR SGSNPNRVRM FGPDKLVRAA
AEKRWDRVKI VCSQPYSKDS PFGLSFVRFH SPPDKDEAEA PSQKVTVTKL GQFRVKEEDE
SANSLRPGAL FFSRINKTSP VTASDPAGPS YAAATLQASS AASSASPVSR AIGSTSKPQE
SPKGKRKLDL NQEEKKTPSK PPAQLSPSVP KRPKLPAPTR TPATAPVPAR AQGAVTGKPR
GEGTEPRRPR AGPEELGKIL QGVVVVLSGF QNPFRSELRD KALELGAKYR PDWTRDSTHL
ICAFANTPKY SQVLGLGGRI VRKEWVLDCH RMRRRLPSRR YLMAGPGSSS EEDEASHSGG
SGDEAPKLPQ KQPQTKTKPT QAAGPSSPQK PPTPEETKAA SPVLQEDIDI EGVQSEGQDN
GAEDSGDTED ELRRVAEQKE HRLPPGQEEN GEDPYAGSTD ENTDSEEHQE PPDLPVPELP
DFFQGKHFFL YGEFPGDERR KLIRYVTAFN GELEDYMSDR VQFVITAQEW DPSFEEALMD
NPSLAFVRPR WIYSCNEKQK LLPHQLYGVV PQA
```

The ID line gives description of the sequence and what it codes. The FT lines describe what kind of domains are present in the protein. Those should say either HELIX or STRAND (always double-check those; they might be inaccurate). The SQ line gives molecular weight and some other information of the succeeding amino acid sequence.

For checking the secondary structures, go to http://www.embl-heidelberg.de/predictprotein/submit_def.html and paste in the first protein sequence. In a short while the results will be emailed to you. From the results, you'll find a description:

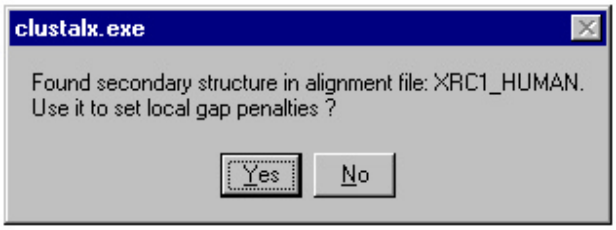
```
         AA      |MPEIRLRHVVSCSSQDSTHCAENLLKADTYRKWRAAKAGEKTISVVLQLEKEEQIHSVDI|
         PHD sec |  EEEEEEEEEE     HHHHHHHHH HHHHHHHHHH    EEEEEE      EEEEE |
         Rel sec |9936789977725787649999986511368789999468836998852354533544 51|
detail:
         prH sec |0000000000000001168999887644678888999621000000000011110000 00|
         prE sec |0067788988752100000000000000000000000000037998864312125666 64|
         prL sec |9932100011146888720000012254321110000278862000125576653222 25|
subset: SUB sec |LL.EEEEEEEE.LLLLL.HHHHHHHH...HHHHHHHH.LLL.EEEEEE..L.L..E..E.|
accessibility
3st:     P_3 acc | eebebebbbbbbeeebeebbeebbebee  ebebbeeeeebbbbbbebeeeeebbbbeb|
10st:    PHD acc |3970606000000997076007700707745706007777700000060677760000 60|
         Rel acc |0023250368462011200583144331001211644123426398503253204045 04|
subset: SUB acc |.....b..bbbb.......bb..bb.........bbe...e.b.bbb...e...b.bb.b|
```

The line marked with AA gives the original sequence, and the the next line, starting with PHD sec gives the predicted secondary structures. Rel sec gives the reliability of the secondary strusture prediction. H mean helix, and E mean sheet. You can use this information for the description of the structures for ClustalX input.

You have just created the first input file. The second file should include the other sequences you're interested in in FastA-format (see above).

After preparing these input files, go to ClustalX, and switch to Profile Alignment Mode.

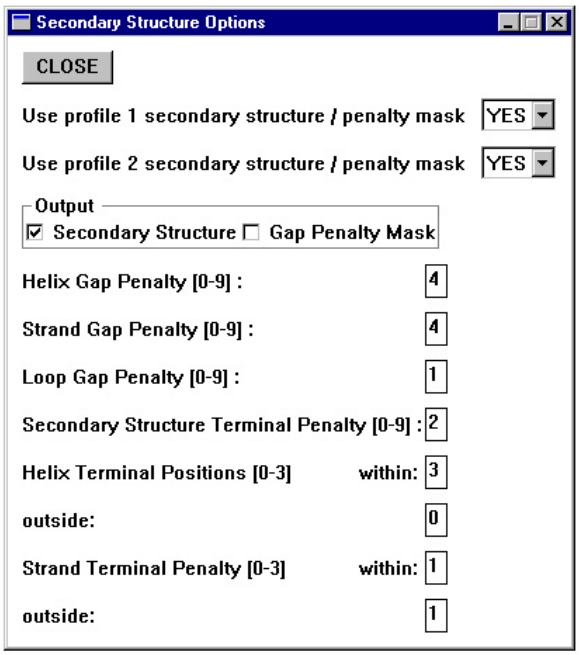
From the File menu select Load Profile 1, and search for the first input file. If ClustalX recognizes the file to contain the weights for the gaps it asks you whether to use the penalties or not.



If you have formatted the files as described above, but ClustalX does not recognize the weights, go to Alignment->Alignment Parameters->Secondary Structure parameters. From the opening dialog box turn the Use Profile Secondary Structure option on (yes).

After that try to load the first input file again.

If the loading was successful, go to Alignment->Alignment Parameters->Secondary Structure parameters. A dialog box opens.



In the first input file you have set up the areas of protein consisting of helixes and strands. In the box you can set the gap penalties for these areas. The penalties are multiples of the normal penalty given by the Multiple Sequence Alignment Parameters.

After setting up the parameters, load in the second Profile (File->Load Profile 2). The alignment is then done in two phases as previously described: First align sequences to the profile 1, and then align profile 2 to profile 1.

A new multiple alignment is created, and the gaps are more often inserted into the areas outside the described secondary structures than within them, depending on the parameters.