ClustalX Help

You can get the latest version of the ClustalX program here:

ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/

For full details of usage and algorithms, please read the *ClustalW.Doc* file.

```
Toby  Gibson                         EMBL, Heidelberg, Germany.
Des   Higgins                        UCC, Cork, Ireland.
Julie Thompson/Francois Jeanmougin   IGBMC, Strasbourg, France.
```

Index

General help for CLUSTAL X (1.8)

Clustal X is a windows interface for the ClustalW multiple sequence alignment program. It provides an integrated environment for performing multiple sequence and profile alignments and analysing the results. The sequence alignment is displayed in a window on the screen. A versatile coloring scheme has been incorporated allowing you to highlight conserved features in the alignment. The pull-down menus at the top of the window allow you to select all the options required for traditional multiple sequence and profile alignment.

You can cut-and-paste sequences to change the order of the alignment; you can select a subset of sequences to be aligned; you can select a sub-range of the alignment to be realigned and inserted back into the original alignment.

Alignment quality analysis can be performed and low-scoring segments or exceptional residues can be highlighted.

ClustalX is available for a number of different platforms including: SUN Solaris, IRIX5.3 on Silicon Graphics, Digital UNIX on DECStations, Microsoft Windows (32 bit) for PC's, Linux ELF for x86 PC's and Macintosh PowerMac. (See the README file for Installation instructions.)

SEQUENCE INPUT

Sequences and profiles (a term for pre-existing alignments) are input using the FILE menu. Invalid options will be disabled. All sequences must be included into 1 file. 7 formats are automatically recognised: NBRF/PIR, EMBL/SWISSPROT, Pearson (Fasta), Clustal (*.aln), GCG/MSF (Pileup), GCG9 RSF and GDE flat file. All non-alphabetic characters (spaces, digits, punctuation marks) are ignored except "-" which is used to indicate a GAP ("." in MSF/RSF).

SEQUENCE / PROFILE ALIGNMENTS

Clustal X has two modes which can be selected using the switch directly above the sequence display: MULTIPLE ALIGNMENT MODE and PROFILE ALIGNMENT MODE.

To do a MULTIPLE ALIGNMENT on a set of sequences, make sure MULTIPLE ALIGNMENT MODE is selected. A single sequence data area is then displayed. The ALIGNMENT menu then allows you to either produce a guide tree for the alignment, or to do a multiple alignment following the guide tree, or to do a full multiple alignment.

In PROFILE ALIGNMENT MODE, two sequence data areas are displayed, allowing you to align 2 alignments (termed profiles). Profiles are also used to add a new sequence to an old alignment, or to use secondary structure to guide the alignment process. GAPS in the old alignments are indicated using the "-" character. PROFILES can be input in ANY of the allowed formats; just use "-" (or "." for MSF/RSF) for each gap position. In Profile Alignment Mode, a button "Lock Scroll" is displayed which allows you to scroll the two profiles together using a single scroll bar. When the Lock Scroll is turned off, the two profiles can be scrolled independently.

PHYLOGENETIC TREES

Phylogenetic trees can be calculated from old alignments (read in with "-" characters to indicate gaps) OR after a multiple alignment while the alignment is still displayed.

ALIGNMENT DISPLAY

The alignment is displayed on the screen with the sequence names on the left hand side. The sequence alignment is for display only, it cannot be edited here (except for changing the sequence order by cutting-and-pasting on the sequence names).

A ruler is displayed below the sequences, starting at 1 for the first residue position (residue numbers in the sequence input file are ignored).

A line above the alignment is used to mark strongly conserved positions. Three characters ('*', ':' and '.') are used:

'*' indicates positions which have a single, fully conserved residue

':' indicates that one of the following 'strong' groups is fully conserved:-

```
                STA
                NEQK
                NHQK
                NDEQ
                QHRK
                MILV
                MILF
                HY
                FYW
'.' indicates that one of the following 'weaker' groups is fully conserved:-

                CSA
                ATV
                SAG
                STNK
                STPA
                SGND
                SNDEQK
                NDEQHK
                NEQHRK
                FVLIM
                HFY
```

These are all the positively scoring groups that occur in the Gonnet Pam250 matrix. The strong and weak groups are defined as strong score 0.5 and weak score =<0.5 respectively.

For profile alignments, secondary structure and gap penalty masks are displayed above the sequences, if any data is found in the profile input file.

Back to Index

Input / Output Files

LOAD SEQUENCES reads sequences from one of 7 file formats, replacing any sequences that are already loaded. All sequences must be in 1 file. The formats that are automatically recognised are: NBRF/PIR, EMBL/SWISSPROT, Pearson (Fasta), Clustal (*.aln), GCG/MSF (Pileup), GCG9/RSF and GDE flat file. All non-alphabetic characters (spaces, digits, punctuation marks) are ignored except "-" which is used to indicate a GAP ("." in MSF/RSF).

The program tries to automatically recognise the different file formats used and to guess whether the sequences are amino acid or nucleotide. This is not always foolproof.

FASTA and NBRF/PIR formats are recognised by having a "" as the first character in the file.

EMBL/Swiss Prot formats are recognised by the letters "ID" at the start of the file (the token for the entry name field).

CLUSTAL format is recognised by the word CLUSTAL at the beginning of the file.

GCG/MSF format is recognised by one of the following:

- - the word PileUp at the start of the file.
- - the word !!AA_MULTIPLE_ALIGNMENT or !!NA_MULTIPLE_ALIGNMENT at the start of the file.
- - the word MSF on the first line of the file, and the characters .. at the end of this line.

GCG/RSF format is recognised by the word !!RICH_SEQUENCE at the beginning of the file.

If 85% or more of the characters in the sequence are from A,C,G,T,U or N, the sequence will be assumed to be nucleotide. This works in 97.3% of cases but watch out!

APPEND SEQUENCES is only valid in MULTIPLE ALIGNMENT MODE. The input sequences do not replace those already loaded, but are appended at the end of the alignment.

SAVE SEQUENCES AS... offers the user a choice of one of six output formats: CLUSTAL, NBRF/PIR, GCG/MSF, PHYLIP, NEXUS or GDE. All sequences are written to a single file. Options are available to save a range of the alignment, switch between UPPER/LOWER case for GDE files, and to output SEQUENCE NUMBERING for CLUSTAL files.

LOAD PROFILE 1 reads sequences in the same 7 file formats, replacing any sequences already loaded as Profile 1. This option will also remove any sequences which are loaded in Profile 2.

LOAD PROFILE 2 reads sequences in the same 7 file formats, replacing any sequences already loaded as Profile 2.

SAVE PROFILE 1 AS... is similar to the Save Sequences option except that only those sequences in Profile 1 will be written to the output file.

SAVE PROFILE 2 AS... is similar to the Save Sequences option except that only those sequences in Profile 2 will be written to the output file.

WRITE ALIGNMENT AS POSTSCRIPT will write the sequence display to a postscript format file. This will include any secondary structure / gap penalty mask information and the consensus and ruler lines which are displayed on the screen. The Alignment

Quality curve can be optionally included in the output file.

WRITE PROFILE 1 AS POSTSCRIPT is similar to WRITE ALIGNMENT AS POSTSCRIPT except that only the profile 1 display will be printed.

WRITE PROFILE 2 AS POSTSCRIPT is similar to WRITE ALIGNMENT AS POSTSCRIPT except that only the profile 2 display will be printed.

POSTSCRIPT PARAMETERS

A number of options are available to allow you to configure your postscript output file.

PS COLORS FILE:

The exact RGB values required to reproduce the colors used in the alignment window will vary from printer to printer. A PS colors file can be specified that contains the RGB values for all the colors required by each of your postscript printers.

By default, Clustal X looks for a file called 'colprint.par' in the current directory (if your running under UNIX, it then looks in your home directory, and finally in the directories in your PATH environment variable). If no PS colors file is found or a color used on the screen is not defined here, the screen RGB values (from the Color Parameter File) are used.

The PS colors file consists of one line for each color to be defined, with the color name followed by the RGB values (on a scale of 0 to 1). For example,

RED 0.9 0.1 0.1

Blank lines and comments (lines beginning with a '#' character) are ignored.

PAGE SIZE: The alignment can be displayed on either A4, A3 or US Letter size pages.

ORIENTATION: The alignment can be displayed on either a landscape or portrait page.

PRINT HEADER: An optional header including the postscript filename, and creation date can be printed at the top of each page.

PRINT QUALITY CURVE: The Alignment Quality curve which is displayed underneath the alignment on the screen can be included in the postscript output.

PRINT RULER: The ruler which is displayed underneath the alignment on the screen can be included in the postscript output.

PRINT RESIDUE NUMBERS: Sequence residue numbers can be printed at the right hand side of the alignment.

RESIZE TO FIT PAGE: By default, the alignment is scaled to fit the page size selected. This option can be turned off, in which case a font size of 10 will be used for the sequences.

PRINT FROM POSITION / TO: A range of the alignment can be printed. The default is to print the full alignment. The first and last residues to be printed are specified here.

USE BLOCK LENGTH: The alignment can be divided into blocks of residues. The number of residues in a block is specified here. More than one block may then be printed on a single page. This is useful for long alignments of a small number of sequences. If the block length is set to 0, The alignment will not be divided into blocks, but printed across a number of pages.

Back to Index

Editing Alignments

Clustal X allows you to change the order of the sequences in the alignment, by cutting-and-pasting the sequence names.

To select a group of sequences to be moved, click on a sequence name and drag the cursor until all the required sequences are highlighted. Holding down the Shift key when clicking on the first name will add new sequences to those already selected.

(Options are provided to Select All Sequences, Select Profile 1 or Select Profile 2.)

The selected sequences can be removed from the alignment by using the EDIT menu, CUT option.

To add the cut sequences back into an alignment, select a sequence by clicking on the sequence name. The cut sequences will be added to the alignment, immediately following the selected sequence, by the EDIT menu, PASTE option.

To add the cut sequences to an empty alignment (eg. when cutting sequences from Profile 1 and pasting them to Profile 2), click on the empty sequence name display area, and select the EDIT menu, PASTE option as before.

The sequence selection and sequence range selection can be cleared using the EDIT menu, CLEAR SEQUENCE SELECTION and CLEAR RANGE SELECTION options respectively.

To search for a string of residues in the sequences, select the sequences to be searched by clicking on the sequence names. You can then enter the string to search for by selecting the SEARCH FOR STRING option. If the string is found in any of the sequences selected, the sequence name and column number is printed below the sequence display.

In PROFILE ALIGNMENT MODE, the two profiles can be merged (normally done after alignment) by selecting ADD PROFILE 2 TO PROFILE 1. The sequences currently displayed as Profile 2 will be appended to Profile 1.

The REMOVE ALL GAPS option will remove all gaps from the sequences currently selected. WARNING: This option removes ALL gaps, not only those introduced by

ClustalX, but also those that were read from the input alignment file. Any secondary structure information associated with the alignment will NOT be automatically realigned.

The REMOVE GAP-ONLY COLUMNS will remove those positions in the alignment which contain gaps in all sequences. This can occur as a result of removing divergent sequences from an alignment, or if an alignment has been realigned.

Back to Index

Multiple Alignments

Make sure MULTIPLE ALIGNMENT MODE is selected, using the switch directly above the sequence display area. Then, use the ALIGNMENT menu to do multiple alignments.

Multiple alignments are carried out in 3 stages:

1) all sequences are compared to each other (pairwise alignments);

2) a dendrogram (like a phylogenetic tree) is constructed, describing the approximate groupings of the sequences by similarity (stored in a file).

3) the final multiple alignment is carried out, using the dendrogram as a guide.

The 3 stages are carried out automatically by the DO COMPLETE ALIGNMENT option. You can skip the first stages (pairwise alignments; guide tree) by using an old guide tree file (DO ALIGNMENT FROM GUIDE TREE); or you can just produce the guide tree with no final multiple alignment (PRODUCE GUIDE TREE ONLY).

REALIGN SELECTED SEQUENCES is used to realign badly aligned sequences in the alignment. Sequences can be selected by clicking on the sequence names - see Editing Alignments for more details. The unselected sequences are then 'fixed' and a profile is made including only the unselected sequences. Each of the selected sequences in turn is then realigned to this profile. The realigned sequences will be displayed as a group at the end the alignment.

REALIGN SELECTED SEQUENCE RANGE is used to realign a small region of the alignment. A residue range can be selected by clicking on the sequence display area. A multiple alignment is then performed, following the 3 stages described above, but only using the selected residue range. Finally the new alignment of the range is pasted back into the full sequence alignment.

By default, gap penalties are used at each end of the subrange in order to penalise terminal gaps. If the REALIGN SEGMENT END GAP PENALTIES option is switched off, gaps can be introduced at the ends of the residue range at no cost.

ALIGNMENT PARAMETERS displays a sub-menu with the following options:

RESET NEW GAPS BEFORE ALIGNMENT will remove any new gaps introduced into the sequences during multiple alignment if you wish to change the parameters and try again. This only takes effect just before you do a second multiple alignment. You can make phylogenetic trees after alignment whether or not this is ON. If you turn this OFF, the new gaps are kept even if you do a second multiple alignment. This allows you to iterate the alignment gradually. Sometimes, the alignment is improved by a second or third pass.

RESET ALL GAPS BEFORE ALIGNMENT will remove all gaps in the sequences including gaps which were read in from the sequence input file. This only takes effect just before you do a second multiple alignment. You can make phylogenetic trees after alignment whether or not this is ON. If you turn this OFF, all gaps are kept even if you do a second multiple alignment. This allows you to iterate the alignment gradually. Sometimes, the alignment is improved by a second or third pass.

PAIRWISE ALIGNMENT PARAMETERS control the speed/sensitivity of the initial alignments.

MULTIPLE ALIGNMENT PARAMETERS control the gaps in the final multiple alignments.

PROTEIN GAP PARAMETERS displays a temporary window which allows you to set various parameters only used in the alignment of protein sequences.

(SECONDARY STRUCTURE PARAMETERS, for use with the Profile Alignment Mode only, allows you to set various parameters only used with gap penalty masks.)

SAVE LOG FILE will write the alignment calculation scores to a file. The log filename is the same as the input sequence filename, with an extension .log appended.

OUTPUT FORMAT OPTIONS

You can choose from 6 different alignment formats (CLUSTAL, GCG, NBRF/PIR, PHYLIP, GDE and NEXUS). You can choose more than one (or all 6 if you wish).

CLUSTAL format output is a self explanatory alignment format. It shows the sequences aligned in blocks. It can be read in again at a later date to (for example) calculate a phylogenetic tree or add in new sequences by profile alignment.

GCG output can be used by any of the GCG programs that can work on multiple alignments (e.g. PRETTY, PROFILEMAKE, PLOTALIGN). It is the same as the GCG .msf format files (multiple sequence file); new in version 7 of GCG.

NEXUS format is used by several phylogeny programs, including PAUP and MacClade.

PHYLIP format output can be used for input to the PHYLIP package of Joe Felsenstein. This is a very widely used package for doing every imaginable form of phylogenetic analysis (MUCH more than the the modest introduction offered by this program).

NBRF/PIR: this is the same as the standard PIR format with ONE ADDITION. Gap characters "-" are used to indicate the positions of gaps in the multiple alignment. These files can be re-used as input in any part of clustal that allows sequences (or alignments or profiles) to be read in.

GDE: this format is used by the GDE package of Steven Smith and is understood by SEQLAB in GCG 9 or later.

GDE OUTPUT CASE: sequences in GDE format may be written in either upper or lower case.

CLUSTALW SEQUENCE NUMBERS: residue numbers may be added to the end of the alignment lines in clustalw format.

OUTPUT ORDER is used to control the order of the sequences in the output alignments. By default, it uses the order in which the sequences were aligned (from the guide tree/dendrogram), thus automatically grouping closely related sequences. It can be switched to be the same as the original input order.

PARAMETER OUTPUT: This option will save all your parameter settings in a parameter file (suffix .par) during alignment. The file can be subsequently used to rerun ClustalW using the same parameters.

ALIGNMENT PARAMETERS

PAIRWISE ALIGNMENT PARAMETERS

A distance is calculated between every pair of sequences and these are used to construct the phylogenetic tree which guides the final multiple alignment. The scores are calculated from separate pairwise alignments. These can be calculated using 2 methods: dynamic programming (slow but accurate) or by the method of Wilbur and Lipman (extremely fast but approximate).

You can choose between the 2 alignment methods using the PAIRWISE ALIGNMENTS option. The slow/accurate method is fast enough for short sequences but will be VERY SLOW for many (e.g. 100) long (e.g. 1000 residue) sequences.

SLOW-ACCURATE alignment parameters:

These parameters do not have any affect on the speed of the alignments. They are used to give initial alignments which are then rescored to give percent identity scores. These % scores are the ones which are displayed on the screen. The scores are converted to distances for the trees.

Gap Open Penalty: the penalty for opening a gap in the alignment.

Gap Extension Penalty: the penalty for extending a gap by 1 residue.

Protein Weight Matrix: the scoring table which describes the similarity of each amino acid to each other.

Load protein matrix: allows you to read in a comparison table from a file.

DNA weight matrix: the scores assigned to matches and mismatches (including IUB ambiguity codes).

Load DNA matrix: allows you to read in a comparison table from a file.

See the Multiple alignment parameters, MATRIX option below for details of the matrix input format.

FAST-APPROXIMATE alignment parameters:

These similarity scores are calculated from fast, approximate, global align- ments, which are controlled by 4 parameters. 2 techniques are used to make these alignments very fast: 1) only exactly matching fragments (k-tuples) are considered; 2) only the 'best' diagonals (the ones with most k-tuple matches) are used.

GAP PENALTY: This is a penalty for each gap in the fast alignments. It has little effect on the speed or sensitivity except for extreme values.

K-TUPLE SIZE: This is the size of exactly matching fragment that is used. INCREASE for speed (max= 2 for proteins; 4 for DNA), DECREASE for sensitivity. For longer sequences (e.g. 1000 residues) you may wish to increase the default.

TOP DIAGONALS: The number of k-tuple matches on each diagonal (in an imaginary dot-matrix plot) is calculated. Only the best ones (with most matches) are used in the alignment. This parameter specifies how many. Decrease for speed; increase for sensitivity.

WINDOW SIZE: This is the number of diagonals around each of the 'best' diagonals that will be used. Decrease for speed; increase for sensitivity.

MULTIPLE ALIGNMENT PARAMETERS

These parameters control the final multiple alignment. This is the core of the program and the details are complicated. To fully understand the use of the parameters and the scoring system, you will have to refer to the documentation.

Each step in the final multiple alignment consists of aligning two alignments or sequences. This is done progressively, following the branching order in the GUIDE TREE. The basic parameters to control this are two gap penalties and the scores for various identical/non-indentical residues.

The GAP OPENING and EXTENSION PENALTIES can be set here. These control the cost of opening up every new gap and the cost of every item in a gap. Increasing the gap opening penalty will make gaps less frequent. Increasing the gap extension penalty will make gaps shorter. Terminal gaps are not penalised.

The DELAY DIVERGENT SEQUENCES switch delays the alignment of the most distantly related sequences until after the most closely related sequences have been aligned. The setting shows the percent identity level required to delay the addition of a sequence; sequences that are less identical than this level to any other sequences will be aligned later.

The TRANSITION WEIGHT gives transitions (AG or CT i.e. purine-purine or pyrimidine-pyrimidine substitutions) a weight between 0 and 1; a weight of zero means that the transitions are scored as mismatches, while a weight of 1 gives the transitions the match score. 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. The default is set to 0.5.

The PROTEIN WEIGHT MATRIX option allows you to choose a series of weight matrices. For protein alignments, you use a weight matrix to determine the similarity of non-identical amino acids. For example, Tyr aligned with Phe is usually judged to be 'better' than Tyr aligned with Pro.

There are three 'in-built' series of weight matrices offered. Each consists of several matrices which work differently at different evolutionary distances. To see the exact details, read the documentation. Crudely, we store several matrices in memory, spanning the full range of amino acid distance (from almost identical sequences to highly divergent ones). For very similar sequences, it is best to use a strict weight matrix which only gives a high score to identities and the most favoured conservative substitutions. For more divergent sequences, it is appropriate to use "softer" matrices which give a high score to many other frequent substitutions.

1) BLOSUM (Henikoff). These matrices appear to be the best available for carrying out data base similarity (homology searches). The matrices currently used are: Blosum 80, 62, 45 and 30. BLOSUM was the default in earlier Clustal X versions.

2) PAM (Dayhoff). These have been extremely widely used since the late '70s. We currently use the PAM 20, 60, 120, 350 matrices.

3) GONNET. These matrices were derived using almost the same procedure as the Dayhoff one (above) but are much more up to date and are based on a far larger data set. They appear to be more sensitive than the Dayhoff series. We currently use the GONNET 80, 120, 160, 250 and 350 matrices. This series is the default for Clustal X version 1.8.

We also supply an identity matrix which gives a score of 10 to two identical amino acids and a score of zero otherwise. This matrix is not very useful.

Load protein matrix: allows you to read in a comparison matrix from a file. This can be either a single matrix or a series of matrices (see below for format).

DNA WEIGHT MATRIX option allows you to select a single matrix (not a series) used for aligning nucleic acid sequences. Two hard-coded matrices are available:

1) IUB. This is the default scoring matrix used by BESTFIT for the comparison of nucleic acid sequences. X's and N's are treated as matches to any IUB ambiguity symbol. All matches score 1.9; all mismatches for IUB symbols score 0.

2) CLUSTALW(1.6). A previous system used by ClustalW, in which matches score 1.0 and mismatches score 0. All matches for IUB symbols also score 0.

Load DNA matrix: allows you to read in a nucleic acid comparison matrix from a file (just one matrix, not a series).

SINGLE MATRIX INPUT FORMAT The format used for a single matrix is the same as the BLAST program. The scores in the new weight matrix should be similarities. You can use negative as well as positive values if you wish, although the matrix will be automatically adjusted to all positive scores, unless the NEGATIVE MATRIX option is selected. Any lines beginning with a # character are assumed to be comments. The first non-comment line should contain a list of amino acids in any order, using the 1 letter code, followed by a * character. This should be followed by a square matrix of scores, with one row and one column for each amino acid. The last row and column of the matrix (corresponding to the * character) contain the minimum score over the whole matrix.

MATRIX SERIES INPUT FORMAT ClustalX uses different matrices depending on the mean percent identity of the sequences to be aligned. You can specify a series of matrices and the range of the percent identity for each matrix in a matrix series file. The file is automatically recognised by the word CLUSTAL_SERIES at the beginning of the file. Each matrix in the series is then specified on one line which should start with the word MATRIX. This is followed by the lower and upper limits of the sequence percent identities for which you want to apply the matrix. The final entry on the matrix line is the filename of a Blast format matrix file (see above for details of the single matrix file format).

Example.

CLUSTAL_SERIES

MATRIX 81 100 /us1/user/julie/matrices/blosum80 MATRIX 61 80 /us1/user/julie/matrices/blosum62 MATRIX 31 60 /us1/user/julie/matrices/blosum45 MATRIX 0 30 /us1/user/julie/matrices/blosum30

PROTEIN GAP PARAMETERS

RESIDUE SPECIFIC PENALTIES are amino acid specific gap penalties that reduce or increase the gap opening penalties at each position in the alignment or sequence. See the documentation for details. As an example, positions that are rich in glycine are more likely to have an adjacent gap than positions that are rich in valine.

HYDROPHILIC GAP PENALTIES are used to increase the chances of a gap within a run (5 or more residues) of hydrophilic amino acids; these are likely to be loop or random coil regions where gaps are more common. The residues that are "considered" to be hydrophilic can be entered in HYDROPHILIC RESIDUES.

GAP SEPARATION DISTANCE tries to decrease the chances of gaps being too close to each other. Gaps that are less than this distance apart are penalised more than other gaps. This does not prevent close gaps; it makes them less frequent, promoting a block-like appearance of the alignment.

END GAP SEPARATION treats end gaps just like internal gaps for the purposes of avoiding gaps that are too close (set by GAP SEPARATION DISTANCE above). If you turn this off, end gaps will be ignored for this purpose. This is useful when you wish to align fragments where the end gaps are not biologically meaningful.

Back to Index

Profile and Structure Alignments

By PROFILE ALIGNMENT, we mean alignment using existing alignments. Profile alignments allow you to store alignments of your favourite sequences and add new sequences to them in small bunches at a time. A profile is simply an alignment of one or more sequences (e.g. an alignment output file from Clustal X). Each input can be a single sequence. One or both sets of input sequences may include secondary structure assignments or gap penalty masks to guide the alignment.

Make sure PROFILE ALIGNMENT MODE is selected, using the switch directly above the sequence display area. Then, use the ALIGNMENT menu to do profile and secondary structure alignments.

The profiles can be in any of the allowed input formats with "-" characters used to specify gaps (except for GCG/MSF where "." is used).

You have to load the 2 profiles by choosing FILE, LOAD PROFILE 1 and LOAD PROFILE 2. Then ALIGNMENT, ALIGN PROFILE 2 TO PROFILE 1 will align the 2 profiles to each other. Secondary structure masks in either profile can be used to guide the alignment. This option compares all the sequences in profile 1 with all the sequences in profile 2 in order to build guide trees which will be used to calculate sequence weights, and select appropriate alignment parameters for the final profile alignment.

You can skip the first stage (pairwise alignments; guide trees) by using old guide tree files (ALIGN PROFILES FROM GUIDE TREES).

The ALIGN SEQUENCES TO PROFILE 1 option will take the sequences in the second profile and align them to the first profile, 1 at a time. This is useful to add some new sequences to an existing alignment, or to align a set of sequences to a known structure. In this case, the second profile set need not be pre-aligned.

You can skip the first stage (pairwise alignments; guide tree) by using an old guide tree file (ALIGN SEQUENCES TO PROFILE 1 FROM TREE).

SAVE LOG FILE will write the alignment calculation scores to a file. The log filename is the same as the input sequence filename, with an extension .log appended.

The alignment parameters can be set using the ALIGNMENT PARAMETERS menu, Pairwise Parameters, Multiple Parameters and Protein Gap Parameters options. These are EXACTLY the same parameters as used by the general, automatic multiple alignment procedure. The general multiple alignment procedure is simply a series of profile alignments. Carrying out a series of profile alignments on larger and larger groups of sequences, allows you to manually build up a complete alignment, if necessary editing intermediate alignments.

SECONDARY STRUCTURE PARAMETERS

Use this menu to set secondary structure options. If a solved structure is known, it can be used to guide the alignment by raising gap penalties within secondary structure elements, so that gaps will preferentially be inserted into unstructured surface loop regions. Alternatively, a user-specified gap penalty mask can be supplied for a similar purpose.

A gap penalty mask is a series of numbers between 1 and 9, one per position in the alignment. Each number specifies how much the gap opening penalty is to be raised at that position (raised by multiplying the basic gap opening penalty by the number) i.e. a mask figure of 1 at a position means no change in gap opening penalty; a figure of 4 means that the gap opening penalty is four times greater at that position, making gaps 4 times harder to open.

The format for gap penalty masks and secondary structure masks is explained in a separate help section.

Back to Index

Secondary Structure / Gap Penalty Masks

The use of secondary structure-based penalties has been shown to improve the accuracy of sequence alignment. Clustal X now allows secondary structure/ gap penalty masks to be supplied with the input sequences used during profile alignment. (NB. The secondary structure information is NOT used during multiple sequence alignment). The masks work by raising gap penalties in specified regions (typically secondary structure elements) so that gaps are preferentially opened in the less well conserved regions (typically surface loops).

The USE PROFILE 1(2) SECONDARY STRUCTURE / GAP PENALTY MASK options control whether the input 2D-structure information or gap penalty masks will be used during the profile alignment.

The OUTPUT options control whether the secondary structure and gap penalty masks should be included in the Clustal X output alignments. Showing both is useful for understanding how the masks work. The 2D-structure information is itself useful in judging the alignment quality and in seeing how residue conservation patterns vary with secondary structure.

The HELIX and STRAND GAP PENALTY options provide the value for raising the gap penalty at core Alpha Helical (A) and Beta Strand (B) residues. In CLUSTAL format, capital residues denote the A and B core structure notation. Basic gap penalties are multiplied by the amount specified.

The LOOP GAP PENALTY option provides the value for the gap penalty in Loops. By default this penalty is not raised. In CLUSTAL format, loops are specified by "." in the secondary structure notation.

The SECONDARY STRUCTURE TERMINAL PENALTY provides the value for setting the gap penalty at the ends of secondary structures. Ends of secondary structures are known to grow or shrink, comparing related structures. Therefore by default these are given intermediate values, lower than the core penalties. All secondary structure read in as lower case in CLUSTAL format gets the reduced terminal penalty.

The HELIX and STRAND TERMINAL POSITIONS options specify the range of structure termini for the intermediate penalties. In the alignment output, these are indicated as lower case. For Alpha Helices, by default, the range spans the end-helical turn (3 residues). For Beta Strands, the default range spans the end residue and the adjacent loop residue, since sequence conservation often extends beyond the actual H-bonded Beta Strand.

Clustal X can read the masks from SWISS-PROT, CLUSTAL or GDE format input files. For many 3-D protein structures, secondary structure information is recorded in the feature tables of SWISS-PROT database entries. You should always check that the assignments are correct - some are quite inaccurate. Clustal X looks for SWISS-PROT HELIX and STRAND assignments e.g.

```
FT   HELIX       100    115
FT   STRAND      118    119
```

The structure and penalty masks can also be read from CLUSTAL alignment format as comment lines beginning "!SS_" or "!GM_" e.g.

```
!SS_HBA_HUMA    ..aaaAAAAAAAAAAaaa.aaaAAAAAAAAAAaaaaaaAaaa.........aaaAAAAAA
!GM_HBA_HUMA    112224444444444222122244444444442222224222111111111222444444
HBA_HUMA        VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
```

Note that the mask itself is a set of numbers between 1 and 9 each of which is assigned to the residue(s) in the same column below.

In GDE flat file format, the masks are specified as text and the names must begin with "SS_ or "GM_.

Either a structure or penalty mask or both may be used. If both are included in an alignment, the user will be asked which is to be used.

Back to Index

Phylogenetic Trees

Before calculating a tree, you must have an ALIGNMENT in memory. This can be input using the FILE menu, LOAD SEQUENCES option or you should have just carried out a full multiple alignment and the alignment is still in memory. Remember YOU MUST ALIGN THE SEQUENCES FIRST!!!!

The method used is the NJ (Neighbour Joining) method of Saitou and Nei. First you calculate distances (percent divergence) between all pairs of sequence from a multiple alignment; second you apply the NJ method to the distance matrix.

To calculate a tree, use the DRAW N-J TREE option. This gives an UNROOTED tree and all branch lengths. The root of the tree can only be inferred by using an outgroup (a sequence that you are certain branches at the outside of the tree .... certain on biological grounds) OR if you assume a degree of constancy in the 'molecular clock', you can place the root in the 'middle' of the tree (roughly equidistant from all tips).

BOOTSTRAP N-J TREE uses a method for deriving confidence values for the groupings in a tree (first adapted for trees by Joe Felsenstein). It involves making N random samples of sites from the alignment (N should be LARGE, e.g. 500 - 1000); drawing N trees (1 from each sample) and counting how many times each grouping from the original tree occurs in the sample trees. You can set N using the NUMBER OF BOOTSTRAP TRIALS option in the BOOTSTRAP TREE window. In practice, you should use a large number of bootstrap replicates (1000 is recommended, even if it means running the program for an hour on a slow computer). You can also supply a seed number for the random number generator here. Different runs with the same seed will give the same answer. See the documentation for more details.

EXCLUDE POSITIONS WITH GAPS? With this option, any alignment positions where ANY of the sequences have a gap will be ignored. This means that 'like' will be compared to 'like' in all distances, which is highly desirable. It also automatically throws away the most ambiguous parts of the alignment, which are concentrated around gaps (usually). The disadvantage is that you may throw away much of the data if there are many gaps (which is why it is difficult for us to make it the default).

CORRECT FOR MULTIPLE SUBSTITUTIONS? For small divergence (say <10%) this option makes no difference. For greater divergence, this option corrects for the fact that observed distances underestimate actual evolutionary distances. This is because, as sequences diverge, more than one substitution will happen at many sites. However, you only see one difference when you look at the present day sequences. Therefore, this option has the effect of stretching branch lengths in trees (especially long branches). The corrections used here (for DNA or proteins) are both due to Motoo Kimura. See the documentation for details.

Where possible, this option should be used. However, for VERY divergent sequences, the distances cannot be reliably corrected. You will be warned if this happens. Even if none of the distances in a data set exceed the reliable threshold, if you bootstrap the data, some of the bootstrap distances may randomly exceed the safe limit.

SAVE LOG FILE will write the tree calculation scores to a file. The log filename is the same as the input sequence filename, with an extension .log appended.

OUTPUT FORMAT OPTIONS

Three different formats are allowed. None of these displays the tree visually. You can display the tree using the NJPLOT program distributed with Clustal X OR get the PHYLIP package and use the tree drawing facilities there.

1) CLUSTAL FORMAT TREE. This format is verbose and lists all of the distances between the sequences and the number of alignment positions used for each. The tree is described at the end of the file. It lists the sequences that are joined at each alignment step and the branch lengths. After two sequences are joined, it is referred to later as a NODE. The number of a NODE is the number of the lowest sequence in that NODE.

2) PHYLIP FORMAT TREE. This format is the New Hampshire format, used by many phylogenetic analysis packages. It consists of a series of nested parentheses, describing the branching order, with the sequence names and branch lengths. It can be read by the NJPLOT program distributed with ClustalX. It can also be used by the RETREE, DRAWGRAM and DRAWTREE programs of the PHYLIP package to see the trees graphically. This is the same format used during multiple alignment for the guide trees. Some other packages that can read and display New Hampshire format are TreeTool, TreeView, and Phylowin.

3) PHYLIP DISTANCE MATRIX. This format just outputs a matrix of all the pairwise distances in a format that can be used by the PHYLIP package. It used to be useful when one could not produce distances from protein sequences in the Phylip package but is now redundant (PROTDIST of Phylip 3.5 now does this).

4) NEXUS FORMAT TREE. This format is used by several popular phylogeny programs, including PAUP and MacClade. The format is described fully in: Maddison, D. R., D. L. Swofford and W. P. Maddison. 1997. NEXUS: an extensible file format for systematic information. Systematic Biology 46:590-621.

BOOTSTRAP LABELS ON: By default, the bootstrap values are correctly placed on the tree branches of the phylip format output tree. The toggle allows them to be placed on the nodes, which is incorrect, but some display packages (e.g. TreeTool, TreeView and Phylowin) only support node labelling but not branch labelling. Care should be taken to note which branches and labels go together.

Back to Index

Colors

Clustal X provides a versatile coloring scheme for the sequence alignment display. The sequences (or profiles) are colored automatically, when they are loaded. Sequences can be colored either by assigning a color to specific residues, or on the basis of an alignment consensus. In the latter case, the alignment consensus is calculated automatically, and the residues in each column are colored according to the consensus character assigned to that column. In this way, you can choose to highlight, for example, conserved hydrophylic or hydrophobic positions in the alignment.

The 'rules' used to color the alignment are specified in a COLOR PARAMETER FILE. Clustal X automatically looks for a file called 'colprot.par' for protein sequences or 'coldna.par' for DNA, in the current directory. (If your running under UNIX, it then looks in your home directory, and finally in the directories in your PATH environment variable).

By default, if no color parameter file is found, protein sequences are colored by residue as follows:

Color	Residue Code
ORANGE	GPST
RED	HKR
BLUE	FWY
GREEN	ILMV

In the case of DNA sequences, the default colors are as follows:

Color	Residue Code
ORANGE	A
RED	C
BLUE	T
GREEN	G

The default BACKGROUND COLORING option shows the sequence residues using a black character on a colored background. It can be switched off to show residues as a colored character on a white background.

Either BLACK AND WHITE or DEFAULT COLOR options can be selected. The Color option looks first for the color parameter file (as described above) and, if no file is found, uses the default residue-specific colors.

You can specify your own coloring scheme by using the LOAD COLOR PARAMETER FILE option. The format of the color parameter file is described below.

COLOR PARAMETER FILE

This file is divided into 3 sections:

1) the names and rgb values of the colors 2) the rules for calculating the consensus 3) the rules for assigning colors to the residues

An example file is given here.

```
 ------------------------------------------------------------------
@rgbindex
```

```
RED          0.9 0.1 0.1
BLUE         0.1 0.1 0.9
GREEN        0.1 0.9 0.1
YELLOW       0.9 0.9 0.0




@consensus
% = 60% w:l:v:i:m:a:f:c:y:h:p
# = 80% w:l:v:i:m:a:f:c:y:h:p
- = 50% e:d
+ = 60% k:r
q = 50% q:e
p = 50% p
n = 50% n
t = 50% t:s




@color
g = RED
p = YELLOW
t = GREEN if t:%:#
n = GREEN if n
w = BLUE if %:#:p
k = RED if +
 --------------------------------------------------------------------
```

The first section is optional and is identified by the header @rgbindex. If this section exists, each color used in the file must be named and the rgb values specified (on a scale from 0 to 1). If the rgb index section is not found, the following set of hard-coded colors will be used.

```
RED          0.9 0.1 0.1
BLUE         0.1 0.1 0.9
GREEN        0.1 0.9 0.1
ORANGE       0.9 0.7 0.3
CYAN         0.1 0.9 0.9
PINK         0.9 0.5 0.5
MAGENTA      0.9 0.1 0.9
YELLOW       0.9 0.9 0.0
```

The second section is optional and is identified by the header @consensus. It defines how the consensus is calculated.

The format of each consensus parameter is:-

```
c = n% residue_list
```

```
        where
```

```
              c             is a character used to identify the parameter.
              n             is an integer value used as the percentage cutoff
                            point.
              residue_list  is a list of residues denoted by a single
                            character, delimited by a colon (:).
```

For example: # = 60% w:l:v:i

will assign a consensus character # to any column in the alignment which contains more than 60% of the residues w,l,v and i.

The third section is identified by the header @color, and defines how colors are assigned to each residue in the alignment.

The color parameters can take one of two formats:

```
1) r = color
2) r = color if consensus_list




        where
              r             is a character used to denote a residue.
              color         is one of the colors in the GDE color lookup
table.
              residue_list  is a list of residues denoted by a single
                            character, delimited by a colon (:).
```

Examples: 1) g = ORANGE

will color all glycines ORANGE, regardless of the consensus.

2) w = BLUE if w:%:#

will color BLUE any tryptophan which is found in a column with a consensus of w, % or #.

Back to Index

Alignment Quality Analysis

QUALITY SCORES

Clustal X provides an indication of the quality of an alignment by plotting a 'conservation score' for each column of the alignment. A high score indicates a well-conserved column; a low score indicates low conservation. The quality curve is drawn below the alignment.

Two methods are also provided to indicate single residues or sequence segments which score badly in the alignment.

Low-scoring residues are expected to occur at a moderate frequency in all the sequences because of their steady divergence due to the natural processes of evolution. The most divergent sequences are likely to have the most outliers. However, the highlighted residues are especially useful in pointing to sequence misalignments. Note that clustering of highlighted residues is a strong indication of misalignment. This can arise due to various reasons, for example:

1. Partial or total misalignments caused by a failure in the alignment algorithm. Usually only in difficult alignment cases.

2. Partial or total misalignments because at least one of the sequences in the given set is partly or completely unrelated to the other sequences. It is up to the user to check that the set of sequences are alignable.

3. Frameshift translation errors in a protein sequence causing local mismatched regions to be heavily highlighted. These are surprisingly common in database entries. If suspected, a 3-frame translation of the source DNA needs to be examined.

Occasionally, highlighted residues may point to regions of some biological significance. This might happen for example if a protein alignment contains a sequence which has acquired new functions relative to the main sequence set. It is important to exclude other explanations, such as error or the natural divergence of sequences, before invoking a biological explanation.

LOW-SCORING SEGMENTS

Unreliable regions in the alignment can be highlighted using the Low-Scoring Segments option. A sequence-weighted profile is used to indicate any segments in the sequences which score badly. Because the profile calculation may take some time, an option is provided to calculate LOW-SCORING SEGMENTS. The segment display can then be toggled on or off without having to repeat the time-consuming calculations.

For details of the low-scoring segment calculation, see the CALCULATION section below.

LOW-SCORING SEGMENT PARAMETERS

MINIMUM LENGTH OF SEGMENTS: short segments (or even single residues) can be hidden by increasing the minimum length of segments which will be displayed.

DNA MARKING SCALE is used to remove less significant segments from the highlighted display. Increase the scale to display more segments; decrease the scale to remove the least significant.

PROTEIN WEIGHT MATRIX: the scoring table which describes the similarity of each amino acid to each other. The matrix is used to calculate the sequence- weighted profile scores. There are four 'in-built' Log-Odds matrices offered: the Gonnet PAM 80, 120, 250, 350 matrices. A more stringent matrix which only gives a high score to identities and the most favoured conservative substitutions, may be more suitable when the sequences are closely related. For more divergent sequences, it is appropriate to use "softer" matrices which give a high score to many other frequent substitutions. This option automatically recalculates the low-scoring segments.

DNA WEIGHT MATRIX: Two hard-coded matrices are available:

1) IUB. This is the default scoring matrix used by BESTFIT for the comparison of nucleic acid sequences. X's and N's are treated as matches to any IUB ambiguity symbol. All matches score 1.0; all mismatches for IUB symbols score 0.9.

2) CLUSTALW(1.6). The previous system used by ClustalW, in which matches score 1.0 and mismatches score 0. All matches for IUB symbols also score 0.

A new matrix can be read from a file on disk, if the filename consists only of lower case characters. The values in the new weight matrix should be similarities and should be NEGATIVE for infrequent substitutions.

INPUT FORMAT. The format used for a new matrix is the same as the BLAST program. Any lines beginning with a # character are assumed to be comments. The first non-comment line should contain a list of amino acids in any order, using the 1 letter code, followed by a * character. This should be followed by a square matrix of scores, with one row and one column for each amino acid. The last row and column of the matrix (corresponding to the * character) contain the minimum score over the whole matrix.

QUALITY SCORE PARAMETERS

You can customise the column 'quality scores' plotted underneath the alignment display using the following options.

SCORE PLOT SCALE: this is a scalar value from 1 to 10, which can be used to change the scale of the quality score plot.

RESIDUE EXCEPTION CUTOFF: this is a scalar value from 1 to 10, which can be used to change the number of residue exceptions which are highlighted in the alignment display. (For an explanation of this cutoff, see the CALCULATION OF RESIDUE EXCEPTIONS section below.)

PROTEIN WEIGHT MATRIX: the scoring table which describes the similarity of each amino acid to each other.

DNA WEIGHT MATRIX: two hard-coded matrices are available: IUB and CLUSTALW(1.6).

For more information about the weight matrices, see the help above for the Low-scoring Segments Weight Matrix.

For details of the quality score calculations, see the CALCULATION section below.

SHOW LOW-SCORING SEGMENTS

The low-scoring segment display can be toggled on or off. This option does not recalculate the profile scores.

SHOW EXCEPTIONAL RESIDUES

This option highlights individual residues which score badly in the alignment quality calculations. Residues which score exceptionally low are highlighted by using a white character on a grey background.

SAVE QUALITY SCORES TO FILE

The quality scores that are plotted underneath the alignment display can also be saved in a text file. Each column in the alignment is written on one line in the output file, with the value of the quality score at the end of the line. Only the sequences currently selected in the display are written to the file. One use for quality scores is to color residues in a protein structure by sequence conservation. In this way conserved surface residues can be highlighted to locate functional regions such as ligand-binding sites.

CALCULATION OF QUALITY SCORES

Suppose we have an alignment of m sequences of length n. Then, the alignment can be written as:

```
A11 A12 A13 .......... A1n
A21 A22 A23 .......... A2n
.
.
Am1 Am2 Am3 .......... Amn
```

We also have a residue comparison matrix of size R where C(i,j) is the score for aligning residue i with residue j.

We want to calculate a score for the conservation of the jth position in the alignment.

To do this, we define an R-dimensional sequence space. For the jth position in the alignment, each sequence consists of a single residue which is assigned a point S in the space. S has R dimensions, and for sequence i, the rth dimension is defined as:

```
Sr =    C(r,Aij)
```

We then calculate a consensus value for the jth position in the alignment. This value X also has R dimensions, and the rth dimension is defined as:

```
Xr = (    SUM    (Fij * C(i,r)) ) / m
       1<=i<=R
```

where Fij is the count of residues i at position j in the alignment.

Now we can calculate the distance Di between each sequence i and the consensus position X in the R-dimensional space.

```
Di = SQRT   (    SUM    (Xr - Sr)(Xr - Sr) )
```

```
                    1<=i<=R
```

The quality score for the jth position in the alignment is defined as the mean of the sequence distances Di.

The score is normalised by multiplying by the percentage of sequences which have residues (and not gaps) at this position.

CALCULATION OF RESIDUE EXCEPTIONS

The jth residue of the ith sequence is considered as an exception if the distance Di of the sequence from the consensus value P is greater than (Upper Quartile + Inter Quartile Range * Cutoff). The value used as a cutoff for displaying exceptions can be set from the SCORE PARAMETERS menu. A high cutoff value will only display very significant exceptions; a low value will allow more, less significant, exceptions to be highlighted.

(NB. Sequences which contain gaps at this position are not included in the exception calculation.)

CALCULATION OF LOW-SCORING SEGMENTS

Suppose we have an alignment of m sequences of length n. Then, the alignment can be written as:

```
A11 A12 A13 .......... A1n
A21 A22 A23 .......... A2n
.
.
Am1 Am2 Am3 .......... Amn
```

We also have a residue comparison matrix of size R where C(i,j) is the score for aligning residue i with residue j.

We calculate sequence weights by building a neighbour-joining tree, in which branch lengths are proportional to divergence. Summing the branches by branch ownership provides the weights. See (Thompson et al., CABIOS, 10, 19 (1994) and Henikoff et al.,JMB, 243, 574 1994).

To find the low-scoring segments in a sequence Si, we build a weighted profile of the remaining sequences in the alignment. Suppose we find residue r at position j in the sequence; then the score for the jth position in the sequence is defined as

```
Score(Si,j) = Profile(j,r)   where Profile(j,r) is the profile score
                               for residue r at position j in the
                               alignment.
```

These residue scores are summed along the sequence in both forward and backward directions. If the sum of the scores is positive, then it is reset to zero. Segments which score negatively in both directions are considered as 'low-scoring' and will be highlighted in the alignment display.

Back to Index

Command Line Parameters

DATA (sequences)

Parameter	***Description***
-PROFILE1=file.ext and -PROFILE2=file.ext	*profiles (aligned sequences)*

VERBS (do things)

Parameter	***Description***
-HELP or -CHECK	*outline the command line parameters*
-ALIGN	*do full multiple alignment*
-TREE	*calculate NJ tree*
-BOOTSTRAP(=n)	*bootstrap a NJ tree (n= number of bootstraps; def. = 1000)*
-CONVERT	*output the input sequences in a different file format*

PARAMETERS (set things)

General settings:*

Parameter	***Description***
-INTERACTIVE	*read command line, then enter normal interactive menus*
-QUICKTREE	*use FAST algorithm for the alignment guide tree*
-TYPE=	*PROTEIN or DNA sequences*
-NEGATIVE	*protein alignment with negative values in matrix*
-OUTFILE=	*sequence alignment file name*
-OUTPUT=	*GCG, GDE, PHYLIP, PIR or NEXUS*
-OUTORDER=	*INPUT or ALIGNED*

-CASE=	*LOWER or UPPER (for GDE output only)*
-SEQNOS=	*OFF or ON (for Clustal output only)*

Fast Pairwise Alignments:

Parameter	***Description***
-TOPDIAGS=n	*number of best diags.*
-WINDOW=n	*window around best diags.*
-PAIRGAP=n	*gap penalty*
-SCORE=	*PERCENT or ABSOLUTE*

Slow Pairwise Alignments:

Parameter	***Description***
-PWDNAMATRIX=	*DNA weight matrix=IUB, CLUSTALW or filename*
-PWGAPOPEN=f	*gap opening penalty*
-PWGAPEXT=f	*gap opening penalty*

Multiple Alignments:

Parameter	***Description***
-USETREE=	*file for old guide tree*
-MATRIX=	*Protein weight matrix=BLOSUM, PAM, GONNET, ID or filename*
-DNAMATRIX=	*DNA weight matrix=IUB, CLUSTALW or filename*
-GAPOPEN=f	*gap opening penalty*
-GAPEXT=f	*gap extension penalty*
-ENDGAPS	*no end gap separation pen.*
-GAPDIST=n	*gap separation pen. range*
-NOPGAP	*residue-specific gaps off*
-NOHGAP	*hydrophilic gaps off*

-HGAPRESIDUES=	*list hydrophilic res.*
-MAXDIV=n	*% ident. for delay*
-TYPE=	*PROTEIN or DNA*
-TRANSWEIGHT=f	*transitions weighting*

Profile Alignments:

Parameter	***Description***
-NEWTREE1=	*file for new guide tree for profile1*
-NEWTREE2=	*file for new guide tree for profile2*
-USETREE1=	*file for old guide tree for profile1*
-USETREE2=	*file for old guide tree for profile2*

Sequence to Profile Alignments:

Parameter	***Description***
-NEWTREE=	*file for new guide tree*
-USETREE=	*file for old guide tree*

Structure Alignments:

Parameter	***Description***
-NOSECSTR2	*do not use secondary structure/gap penalty mask for profile 2*
-SECSTROUT=STRUCTURE or MASK or BOTH or NONE	*output in alignment file*
-HELIXGAP=n	*gap penalty for helix core residues*
-STRANDGAP=n	*gap penalty for strand core residues*
-LOOPGAP=n	*gap penalty for loop regions*

-TERMINALGAP=n	*gap penalty for structure termini*
-HELIXENDIN=n	*number of residues inside helix to be treated as terminal*
-HELIXENDOUT=n	*number of residues outside helix to be treated as terminal*
-STRANDENDIN=n	*number of residues inside strand to be treated as terminal*
-STRANDENDOUT=n	*number of residues outside strand to be treated as terminal*

Trees:

Parameter	***Description***
-SEED=n	*seed number for bootstraps*
-KIMURA	*use Kimura's correction*
-TOSSGAPS	*ignore positions with gaps*
-BOOTLABELS=node OR branch	*position of bootstrap values in tree display*

Back to Index

References

The ClustalX program is described in the manuscript:

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 24:4876-4882.

The ClustalW program is described in the manuscript:

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22:4673-4680.

The ClustalV program is described in the manuscript:

Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) CLUSTAL V: improved software for multiple sequence alignment. CABIOS 8,189-191.

The original Clustal program is described in the manuscripts:

Higgins,D.G. and Sharp,P.M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS 5,151-153.

Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73,237-244.

Some tips on using Clustal X:

Jeannmougin,F., Thompson,J.D., Gouy,M., Higgins,D.G. and Gibson,T.J. (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci, 23, 403-5.

Some tips on using Clustal W:

Higgins, D. G., Thompson, J. D. and Gibson, T. J. (1996) Using CLUSTAL for multiple sequence alignments. Methods Enzymol., 266, 383-402.

You can get the latest version of the ClustalX program by anonymous ftp to:

ftp-igbmc.u-strasbg.fr ftp.embl-heidelberg.de ftp.ebi.ac.uk

Or, have a look at the following WWW site:

http://www-igbmc.u-strasbg.fr/BioInfo/

Back to Index