Contents for MACAW Help

MACAW is a program for locating, analyzing, and editing blocks of localized sequence similarity among multiple seqences and linking them into a composite multiple alignment.

Overviews

Abstracts
Terminology
Searching for Blocks
Statistical Significance of Blocks
Overview of MACAW Windows

How To...

Specify the Type of Sequence to Align (protein, DNA, or DNA)
Add Sequences to an Alignment
Select a Region of the Alignment
Extend a Block to More Sequences

Menu Commands

File Menu Commands
Edit Menu Commands
Sequence Menu Commands
Alignment Menu Commands
Window Menu Commands
Help Menu Commands

Miscellaneous Topics

MACAW Support Policy
MACAW System Requirements
MACAW Release Notes

Searching For Blocks

The question of how to find regions of local similarity among multiple sequences is not trivial, in large part due to the immense size of the space to be searched. Several attacks on the problem have been mounted, but each suffers from certain limitations, among which have been the necessity of specifying the length of the putative pattern beforehand, restrictions on the allowable length or position of such a pattern, and the requirement or preference that any pattern found appear in all the sequences tested. While we have made attempts to develop new algorithms that get around these problems (Schuler et al, 1991; Lawrence et al., 1993), we recognize that no ideal multiple local alignment algorithm exists today.

This realization guided our design of the MACAW program. We felt that it was important, first of all, to provide several search methods for finding local alignments. And secondly, we have attempted to involve the user as much as possible in the decision as to which blocks to accept or reject.

To search for blocks in MACAW:

1. Select a region of the alignment over which you would like to search
2. Choose the Search For Blocks command from the Alignment menu
3. In the dialog box, select a search method and set options as available

See Also

How To Select a Region of an Alignment

Statistical Significance of Blocks

Given a specific block B, an important question is whether it represents some real relationship, or whether it can be explained simply by chance. To get a handle on this question, we need to have a model of chance. If we assume that our sequences approximate a random sequence model, we can apply the statistical methods described by Karlin and Altschul (1990) to determine how surprising it is to observe B given the expected distribution of scores. The p-value calculated by this analysis tells us the probability that finding a block with a score or greater than or equal to that of B is due to chance.

Suppose we have found that the score if B is too high to be readily explainable by chance; what might this mean? It is of course possible that our random sequence model is at fault. Imagine, however, that the high score of B reflects some true relationship. Can we conclude that all the segments that comprise B are mutually related? The answer is no, for the null hypothesis of non-relatedness can be violated by just a subset of the segments in B. Even two closely related segments can boost the score for B to such an extent that the whole block can appear statistically significant. Any similarity the remaining sequences show may be readily explainable by chance. The case may of course be even more complicated, for a block may divide into two or more subsets of segments, each of which is internally related but whose mutual similarity is due to chance. The heuristic procedure used by MACAW for parsing diagonals seeks to avoid reporting blocks of this sort, but we have no direct way of deciding which among the segments of a significant block are mutually related. Nevertheless, this problem may be addressed by removing from the block any segments that seem only marginally similar to the rest. If the statistical significance of the block is unaffected or increases as a result, the similarity of the remaining segments to those that were removed may be ascribable to chance.

The concept of statistical significance is intimately tied to the idea of a null hypothesis, which here is tied to the idea of a search space. The p-values calculated by MACAW are best used simply as a rough guide to when a block has a surprisingly high score. A p-value should be reported formally only in conjunction with a description of the search space (i.e. the hypothesis) from which it draws its meaning.

Overview of MACAW Windows

Select the MACAW window for which you would like more information.

Windows

Schematic Window
Alignment Window
Search Results Window
Block Editor Window
Blocks Window
Color Map Window

Schematic Window

The Schematic Window shows the multiple alignment in schematic form. Each row corresponds to one sequence, with its name at the left and a thin bar representing the length of the sequence. Thickened portions of each bar indicate regions that have been linked and breaks indicate that gaps have been inserted as a result of linking.

See Also:

How To Select a Region of an Alignment

Alignment Window

Link (Alignment Menu)

Unlink (Alignment Menu)

Compress (Alignment Menu)

Shading (Alignment Menu)

Alignment Window

The Alignment Window shows the multiple alignment in text form, with upper case letters indicating the regions that have been linked and dashes indicating gaps that have been inserted as a result of linking.

See Also:

How To Select a Region of an Alignment

Schematic Window

Link (Alignment Menu)

Unlink (Alignment Menu)

Compress (Alignment Menu)

Shading (Alignment Menu)

Search Results Window

After using the Search For Blocks command to perform a search, the resulting blocks are displayed in the Search Results Window. Each row in the list represents a block and clicking on one with the mouse causes the selected region in both the Schematic and Alignment Windows to be changed reflect the block. The list is sorted in decreasing order of SP-score.

Dialog Box Options

View/Edit	Use this command to open a Block Editor Window for the selected block, allowing you to inspect the block and (optionally) edit its boundaries. Double-clicking the item in the listbox is a shortcut for this command.
Link	Use this command (after you have edited the block boundaries) to link the block into the alignment. Gaps will be inserted into the alignment as needed to bring the segments of the block into alignment. This performs exactly the same operation as the Link menu command.
Unlink	Use this command to remove links created by the Link command (see above). This performs exactly the same operation as the Unlink menu command.
Keep	Press this button to tell MACAW that you want to keep the selected block. When the window is closed, all blocks marked for keeping are transferred to the Blocks Window.
Keep when linked	When this option is enabled, blocks are automatically marked for keeping when they are linked.

See Also

Schematic Window

Alignment Window

How To Select a Region of an Alignment

Block Editor Window

After finding a block with the Search For Blocks command or creating one manually with the Define Block command, the Block Editor may be used to name the block, modify its boundaries, or exclude specific sequences from the block.

Since a block contains no gaps (by definition), it lies entirely on one m-diagonal. The block editor displays the block in this context and it is possible to scroll along the diagonal from beginning to end. The method used to shade the sequence characters is derived from the Segment Pair Overlap algorithm. Darker colors indicate a greater degree of overlap of the segment pairs, with the darkest shade (corresponding to 100% in the Color Map Window) indicating that a residue falls within a segment pair in all of the pairwise comparisons.

Several measures of block quality and significance -- such as SP-score, MP-score, and p-value -- are shown below the panel containing the diagonal. Based on these values, the shading patterns, or experimental evidence, you may decide to adjust the boundaries of the block. To do so, simply drag the edges of the box left or right with the mouse as desired. As you do this, the metrics are recalculated to reflect the new boundaries. For comparison, the original values are also shown.

With distantly related sequences, it is fairly common to find blocks in which only a subset of the sequences show a true relationship to one another. This will not be apparent from looking at the initial values of any of the measures, however the pattern of residue shading can be quite useful in detecting this situation. An additional test is to exclude the suspect sequence by removing the check mark from the checkbox to the left of the sequence name and see what effect this has on the block metrics. If removing a sequence has little impact on the SP-score and causes the MP-score to increase then the sequence is not contributing very much to the score and is unlikely to be related to the other sequences. Excluding the sequence should also cause the block to become more significant, i.e. have a smaller p-value (if you have chosen to display the p-values with the Summary Info command).

It is important to keep in mind that the statistical significance is only meaningful in the context of some specific search space. By default this is set to the product of the effective lengths of the sequences, but may be modified by pressing the "Search Space" button to open the Search Space Dialog Box.

Search Space Dialog Box

A key value in the calculation of the statistical significance of a block (p-value) is the search space N. It is calculated by multiplying together a series of sequence lengths, but you have a number of options on which lengths to use. In addition, you can enter your own value for N if you chose to calculate it in some other way.

Dialog Box Options

Use effective sequence lengths	Use this option to have MACAW calculate N using the effective sequence lengths, which may be set using the View Sequence command (by default they are the same as the actual lengths).
Use actual sequence lengths	Use this option to have MACAW calculate N using the actual lengths of the sequences loaded into the project.
Use space actually searched	Use this option to have MACAW calculate N using the lengths from the region that selected was prior to using the Search For Blocks command.
Use this value	Use this option to enter your own value for N.

Additional Values Shown

N	The search space (see above) calculated according to the options above.
K	A statistical constant derived from the distribution of block scores predicted by the random sequence model.
λ	A statistical constant derived from the distribution of block scores predicted by the random sequence model.
n	The number of sequences in the project.
m	The number of sequences spanned by the block.
S	The SP-score of the block.
p	The p-value.

See Also:
Statistical Significance of Blocks
How To Select a Region of an AlignmentRandom Sequence Model

Blocks Window

The Blocks Window contains a list of blocks that you have decided are meaningful and worth keeping. Usually this results from using the Search For Blocks command and pressing the "Keep" button on the Search Results Window (or pressing "Link" when the "Keep linked blocks" option is enabled). Alternatively, the block may have been created manually with the Define Block command. The order in which the blocks appear can be affected by selecting different sorting options from the drop-down lists at the top of the window.

Selecting a block from this list causes the selected region in the Schematic and Alignment windows to be set to reflect the sequence segments comprising the block. Double-clicking a block opens the Block Editor Window for that block. To remove a block from the list, use the Delete command or press the Del key when the block is selected.

Color Map Window

The Color Map Window shows the mappings of numerical values to shades that may be shown in the Schematic, Alignment, or Block Editor windows. The 100% level corresponds to the maximum value for whatever measure is being employed in the display. For example, in the Block Editor Window, where the shading reflects the degree of segment pair overlap, the most intense shade is used for residues that fell within segment pairs in all of the pairwise comparisons. In the Schematic and Alignment windows, various shading options are available. When either of the options involving score ("pair-score" or "mean-score") are used, the 100% shade corresponds to the largest score in the current score table and the 0% shade (usually just white) is used for all scores less than or equal to zero. Use the Summary Info command to select from any of the supplied color schemes.

See Also:
Schematic Window
Alignment Window
Block Editor Window

Terminology

Select the term for which you would like more information.

Terms

Block
Diagonal
Global Alignment
Information Per Parameter
Local Alignment
MP-score
P-value
Random Sequence Model
Score Table
Search Space
Segment Pair
SP-score

Block (definition)

A set of sequence segments of the same length taken from m sequences and aligned without gaps is called an m-block, or simply, a block. Since a block contains no gaps, it always lies entirely on one m-diagonal.

Diagonal (definition)

A specific register or offset imposed on a set of m sequences is called an m-diagonal, or simply, a diagonal.

Global Alignment (definition)

An alignment that is required to include all of characters from each sequences is called a global alignment. Global alignments are most useful for analyzing closely related sequences that are known ahead of time to be homologous. MACAW does not provide any tools for finding globally optimal alignments.

See Also:

Local Alignment (definition)

Information Per Parameter (definition)

The measure that one attempts to optimize using the Gibbs Sampler search method is effectively the information content of the pattern. However, the maximum value of the measure always increases with increasing pattern width. The quantity called "information per parameter" is independent of width, and may be used to compare blocks of different widths. MACAW reports the results of a Gibbs Sampler search sorted in decreasing order of information per parameter. This is helpful in selecting the optimal pattern width.

Local Alignment (definition)

An alignment that includes only the most similar regions of the sequences under consideration, even if they are fairly small, is called a local alignment. MACAW provides several methods for locating such regions that may span multiple sequences. Local alignments are especially useful for analyzing distantly related proteins or for finding short conserved domains.

See Also

Global Alignment (definition)

MP-score (definition)

Mean Pairwise Score. The MP-score is calculated by dividing the SP-score of a block by the number of possible pairwise combinations of component sequence segments.

P-value (definition)

Consider a set of n unrelated sequences. Any possible column aligning amino acids from the n sequences then has an associated probability, which is the product of the probabilities for the amino acids it contains. It also has an associated score, which in our case is the SP-score. What is the highest score for an n-block that can be expected to occur by chance? Recent statistical results answer just this question (Karlin & Altschul, 1990). In brief, the statistical significance of an n-block with score S is given by the formula

$$1 - \exp(-KN\, e^{-\lambda S})$$

where N is the search space for the analysis, and K and λ are constants determined by the possible scores for an n-column and their corresponding probabilities. This formula presupposes that a columns score is constant under permutation of the columns elements, that the expected score for a column is negative, and that the lengths of the n sequences being compared are not too dissimilar. All these are valid assumptions for the comparison of n unrelated protein sequences of typical size using SP-scores.

Random Sequence Model (definition)

To decide whether a block is statistically significant, one needs to have a model of chance. We can model a single protein by assuming that at any position each type of amino acid has a specific probability of occurring. We assume these probabilities are position independent and have no Markov dependence. Such a sequence we call a random protein sequence. While this is a simplified model for real proteins, it at least gives us a handle on the question of statistical significance.

Score Table

The score contribution for aligning any pair of sequence characters is given in a score table (also called a score matrix or substitution matrix). MACAW provides several score tables, which may be selected using the Summary Info command.

For aligning nucleotide sequences it has been customary to use the unitary matrix, which awards +1 for a match and -1 for a mismatch. However, the table supplied with MACAW for use with nucleotide sequences is slightly more stringent than the unitary matrix and awards +5 for a match and -6 for a mismatch. We have found that this reduces the number of background hits and improves running time significantly.

When aligning proteins, it has been customary to use one the "Dayhoff tables", such as PAM-120 or PAM-250, which are based on the point accepted mutation (PAM) model of evolution (Dayhoff, 1978. Atlas of Protein Sequence and Structure). These data are derived from global alignments of closely related sequences (less than 15% mismatches). More recently, the BLOSUM family of score tables has been developed (Henikoff & Henikoff, 1992, Proc.Natl.Acad.Sci.USA 89, 10915), which make use of substitution data from ungapped local alignments (blocks) instead of global alignments. The BLOSUM tables are more suitable for use with MACAW, which seeks blocks such as those from which the data are derived. Consequently, BLOSUM-62 is used as the default score table, with BLOSUM-45 and BLOSUM-80 also being supplied.

Obviously, using a different score table will have an impact on the expected distribution of block score. One consequence of this is that each table will have its own values of K and l for use in calculating p-values.

Search Space (definition)

It is important to understand that the statistical significance of a block is not some absolute quantity, independent of the context in which the block was found. The pvalue of a block represents the probability that a block with equal or greater score would appear among a set of random sequences of specified lengths. MACAW assumes as a default that this "search space is a set of n random sequences of the same lengths as the sequences loaded.

Under certain conditions, this is not the appropriate search space to use for estimating statistical significance. An example is provided by the situation in which two blocks representing clear homology among the sequences under consideration are already known. The question is whether any regions between these two blocks can be considered significantly related. The search space is now effectively reduced to the lengths of the protein segments between the blocks of already determined homology.

Finally, how do we assess the significance of an m-block found from a comparison of n sequences, where m<n ? For example, we may have input six protein sequences, and found a high-scoring 4-block B. Had we compared just the four sequences whose segments contribute to B, the calculation of N described above would apply, but we have in effect performed a much larger search. While rigorous results concerning the significance of B remain to be proved, we have used the sum, over all possible choices of m sequences, of the product of the sequence lengths: this is the effective size of the space searched. The parameters K and λ are of course those for m-sequence comparison. MACAW uses this procedure to estimate the statistical significance of any m-block. The numbers produced are based, of course, on a random protein model that can only approximate the sequence structure of real proteins, and therefore they must be used only as a rough guide.

Segment Pair (definition)

A pair of segments of the same length take from two sequences and aligned without gaps is called a segment pair. In order compare two segment pairs, we need to have some measure of quality. The typical method of calculating a segment pair score is to add up the score matrix values for all aligned pairs of letters. Normally when we speak of segment pairs, we mean ungapped pairwise alignments whose score is locally optimal, i.e. no adjustment to the boundaries of the segment pair will increase its score.

SP-score (definition)

Sum of the Pairs Score. A block is composed of several sequence segments. Taking any two of them, we can calculate a score as usual for a segment pair. If we do this for all possible pairs an sum the results we have the SP-score.

Abstracts

These abstracts are from papers describing the search algorithms and statistical methods used by MACAW.

Karlin & Altschul, 1990

Schuler, et al., 1991

Lawrence, et al., 1993

Proceedings of the National Academy of Sciences USA **87**, 2264-2268 (1990)

Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes

Samuel Karlin and Stephen F. Altschul

An unusual pattern in a nucleic acid or protein sequence or a region of strong similarity shared by two or more sequences may have biological significance. It is therefore desirable to know whether such a pattern can have arisen simply by chance. To identify interesting sequence patterns, appropriate scoring values can be assigned to the individual residues of a single sequence or to sets of residues when several sequences are compared. For single sequences, such scores and reflect biophysical properties such as charge, volume, hydrophobicity, or secondary structure potential; for multiple sequences, they can reflect nucleotide or amino acid similarity measured in a wide variety of ways. Using an appropriate random model, we present a theory that provides precise numerical formulas for assessing the statistical significance of any region with high aggregate score. A second class of results describes the composition of high-scoring segments. In certain contexts, these permit the choice of scoring systems which are "optimal" for distinguishing biologically relevant patterns. Examples are given of applications of the theory to a variety of protein features, highlighting segments with unusual biological features. These include distinctive charge regions in transcription factors and protooncogene products, pronounced hydrophobic segments in various receptor and transport proteins, and statistically significant subalignments involving the recently characterized cystic fibrosis gene.

Proteins: Structure, Function, and Genetics **9**, 180-190 (1991)

A Workbench for Multiple Alignment Construction and Analysis

Gregory D. Schuler, Stephen F. Altschul, and David J. Lipman

Multiple sequence alignment can be a useful technique for studying molecular evolution, as well as for analyzing relationships between structure or function and primary sequence. We have developed for this purpose an interactive program, MACAW (Multiple Alignment Construction and Analysis Workbench), that allows the user to construct multiple alignments by locating, analyzing, editing, and combining "blocks" of aligned sequence segments. MACAW incorporates several novel features. (1) Regions of local similarity are located by a new search algorithm that avoids many of the limitations of previous techniques. (2) The statistical significance of blocks of similarity is evaluated using a recently developed mathematical theory. (3) Candidate blocks may be evaluated for potential inclusion in a multiple alignment using a variety of visualization tools. (4) A user interface permits each block to be edited by moving its boundaries or by eliminating particular segments, and blocks may be linked to form a composite multiple alignment. No completely automatic program is likely to deal effectively with all the complexities of the multiple alignment problem; by combining a powerful similarity search algorithm with flexible editing, analysis and display tools, MACAW allows the alignment strategy to be tailored to the problem at hand.

Science **262**, 208-214 (1993)

Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment

Charles E. Lawrence, Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, and John C. Wootton

A wealth of protein and DNA sequence data is being generated by genome projects and other sequencing efforts. A crucial barrier to deciphering these sequences and understanding the relationships among them is the difficulty of detecting subtle residue patterns common to multiple sequences. Such patterns frequently reflect similar molecular structures and biological properties. A mathematical definition of this "local multiple alignment" problem suitable for full computer automation has been used to develop a new and sensitive algorithm, based on the statistical method of iterative sampling. This algorithm finds an optimized local alignment model for N sequences in Nlinear time, requiring only seconds on current workstations and allows the simultaneous detection and optimization of multiple patterns and pattern repeats. The method is illustrated as applied to helix-turn-helix proteins, lipocalins, and prenyltransferases.

File Menu Commands

Choose the command for which you need Help.

Commands

New Project
Open Project
Save Project
Save Project As
Export
Print
Exit

New Project (File Menu)

Use this command to begin a new alignment project. The Summary Info dialog box will be displayed, allowing you to modify settings as desired.

Note: Only one project may be open at any time, so the previous project will be automatically closed and you will be prompted to save it has been modified.

Open Project (File Menu)

Use this command to open an alignment project file that was previously saved by MACAW using the Save Project or Save Project As commands.

Note: Only one project may be open at any time, so the previous project will be automatically closed and you will be prompted to save it has been modified.

Save Project (File Menu)

Use this command to save the project to a disk file. When you save a project for the first time, MACAW displays the Save As dialog box so you can name your file. If you want to change the name or location of an existing project file, choose the Save Project As command.

Save Project As (File Menu)

Use this command to save the alignment project and assign a name and location to its disk file. To save a document with its current file name and location, use the Save Project command.

Note: If there is an existing file with the name you specified in the given directory, you will be asked if it is OK to replace it with the new project file you are saving.

Dialog Box Options

File Name	Type a new filename to save a document with a different name. Use the current name, or select a name in the list to save a document with an existing filename. **Note**: A filename can contain up to eight characters and an extension of up to three characters. MACAW adds the MCW extension if you do not specify one.
Drives	Select the drive in which you want to store the document.
Directories	Select the directory in which you want to store the document.

Export Alignment (File Menu)

Use this command to create a file containing a plain text representation of the alignment. Use the standard File Save As dialog box to specify the name of the file to be written.

Print (File menu)

Use this command to print the alignment schematic, alignment text, or blocks.

Dialog Box Options

Print	Select which items you would like to print.
Chars. Per Line	Enter the number of sequence character columns to be used in printing the alignment text or blocks.

After pressing the OK button, the standard Windows Print dialog box will then appear, allowing you to modify some common printing options. Press the Settings button for advanced printer setup options or to select a different printer.

Exit (File Menu)

Use the Exit command to close the current project and exit the MACAW application. If you have made any changes, you will be asked if you want to save your work.

Edit Menu Commands

Choose the command for which you need Help.

Commands

Cut
Copy
Paste
Delete

Cut (Edit Menu)

Use this command to put the selected item onto the clipboard and remove it from the project. The item may be either a sequence or a block. The clipboard data will be plain text and may be pasted into a variety of other programs, such as word processors and graphics applications.

Note: Sequences may be cut to the clipboard and inserted into another MACAW project using the Paste command. However, all residue links will be lost in the process.

Note: Blocks may not be transferred from one project to another using the clipboard because they are context-dependent.

Copy (Edit Menu)

Use this command to put the selected item onto the clipboard (without removing it from the project). The item may be either a sequence or a block. The clipboard data will be plain text and may be pasted into a variety of other programs, such as word processors and graphics applications.

Note: Sequences may be copied to the clipboard and inserted into another MACAW project using the Paste command. However, all residue links will be lost in the process.

Note: Blocks may not be transferred from one project to another using the clipboard because they are context-dependent.

Paste (Edit Menu)

Use this command to transfer sequence data from the clipboard into a MACAW alignment. The clipboard data must be the plain ASCII text representation of the sequence with either the simple header line of the FASTA format or no annotation at all.

Note: If there is no FASTA header line the name and description of the sequence will initially be empty. Use the View Sequence command to modify them as desired.

Note: If the FASTA style is used, multiple sequences may be copied and pasted in one step. Otherwise they must be transferred one at a time (because there is nothing to delimit the multiple sequences).

Delete (Edit Menu)

Use this command to remove the selected item (either a sequence or a block) from the MACAW project.

Note: If the selected item is a sequence, the Delete command is effectively the same as the Remove Sequence command.

Sequence Menu Commands

Choose the command for which you need Help.

Commands

Import Sequence
Remove Sequence
View Sequence

Import Sequence (Sequence Menu)

Use this command to read one or more sequences from a disk file for addition to the multiple alignment. Locate the file or enter its name in the standard File Open dialog box. The file must be plain text and in the FASTA format. Up to 32 sequences may be loaded at any time.

Note: It is important to select the sequence type (protein, DNA, or RNA) for the alignment using the Summary Info dialog box before importing any sequences. This determines which set of valid characters to assume when processing the file.

FASTA File Format

This sequence file format was originally used with FASTA and related programs, but is useful for other purposes because of its simplicity (an example is shown below). The file should be prepared with a text editor or word processor capable of saving files as plain ASCII text. A file may contain multiple sequence records, each of which begins with a ">" symbol as the first character on the line. The first word after this character is taken as the sequence name and the remainder of the line as the sequence description. The following lines (up to the beginning of the next sequence record) contain sequence characters. The line length and spacing style are not significant and MACAW additionally allows numbers within the sequence, which are ignored. Although length limits are not specified by FASTA, for MACAW's purposes, the sequence should not exceed 20,000 characters and the description should not exceed 2,000 characters.

Example

This example contains three sequences of the lipocalin family, named ICYA_MANSE, OBP_BOVIN, and LACB_BOVIN.

```
>ICYA_MANSE INSECTICYANIN (BLUE BILIPROTEIN).
GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLPLENENQGKCTIAEYKYDGKKASVYNS
FVSNGVKEYMEGDLEIAPDAKYTKQGKYVMTFKFGQRVVNLVPWVLATDYKNYAINYNCD
YHPDKKAHSIHAWILSKSKVLEGNTKEVVDNVLKTFSHLIDASKFISNDFSEAACQYSTT
YSLTGPDRH
>OBP_BOVIN ODORANT-BINDING PROTEIN (OBP) (OLFACTORY MUCOSA PYRAZINE-BINDING PROTEIN).
AQEEEAEQNLSELSGPWRTVYIGSTNPEKIQENGPFRTYFRELVFDDEKGTVDFYFSVKR
DGKWKNVHVKATKQDDGTYVADYEGQNVFKIVSLSRTHLVAHNINVDKHGQTTELTELFV
KLNVEDEDLEKFWKLTEDKGIDKKNVVNFLENEDHPHPE
>LACB_BOVIN BETA-LACTOGLOBULIN PRECURSOR (BETA-LG).
MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYVE
ELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLL
FCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI
```

Remove Sequence (Sequence Menu)

Use this command to remove a sequence and all links between its residues and those of other sequences in the multiple alignment.

Shortcut: Press the Del (Delete) key.

Note: Since it is easy to press the Del key by accident, you will be asked to confirm that you really want to remove the sequence before the command is actually carried out.

View Sequence (Sequence Menu)

Use this command to view the selected sequence or to edit its name, description, or effective length.

Shortcut: Double-click on the sequence name in either the Schematic or Alignment windows.

Dialog Box Options

Name	Enter a short mnemonic name for the sequence. Sequence names appear to the left of each row of a multiple alignment or block whenever it is displayed, printed or exported.
Description	Enter additional descriptive information about the sequence. This is for your own information and it appears only in this window.
Effective Length	The effective length of a sequence may be used in the calculation of the search space *N* that is used to determine the statistical significance of a block. For reference, the actual length of the sequence is also shown.

Note: The sequence is presented here for viewing purposes only and the field that contains it is not editable. Sequences may not be modified after they been imported into MACAW.

Alignment Menu Commands

Choose the command for which you need Help.

Commands

Search for Blocks
Define Block
Link
Unlink
Compress
Shading
Summary Info

Search for Blocks (Alignment Menu)

Use this command to search the selected region of the alignment for blocks of local similarity. Several search methods are provided.

Dialog Box Options

Search Method	Select the desired search method; additional dialog box options will be presented depending on the method chosen:
other options	Additional options apply to each method: Segment Pair Overlap Gibbs Sampler Regular Expression

Dialog Box Options for Segment Pair Overlap

If you used the Search For Blocks command and selected "Segment Pair Overlap" as the search method, the following options may be set:

Dialog Box Options	
Pairwise Score Cutoff	The minimum segment pair score needed to mark a diagonal.
Min. Seqs. per Block	The minimum number of sequence that a block must span in order to be reported.

Dialog Box Options for Gibbs Sampler

If you used the Search For Blocks command and selected "Gibbs Sampler" as the search method, the following options may be set:

Dialog Box Options	
Min. Pattern Width	Enter the smallest pattern width to look for.
Max. Pattern Width	Enter the largest pattern width to look for.
Random Seed	Enter some integer value that will be used to "seed" the pseudo random number generator.
Number of Trials	Each trial starts from a random model state and a random seed value and proceeds through many iterations of the Gibbs Sampler algorithm. Enter the desired number of trials.
Number of Iterations	Enter the maximum number of sampling iterations per trial.

Dialog Box Options for Regular Expression

If you used the Search For Blocks command and selected "Regular Expression" as the search method, the following options may be set:

Dialog Box Options	
Expression	Enter the regular expression describing the pattern you are looking for.

Define Block (Alignment Menu)

Use this command to create a block from the current selection, which may have been defined manually rather than as the result of a search. The block is presented in the Block Editor Window, where its score and statistical significance may be examined and its boundaries modified as desired. When you press the "OK" button, MACAW places the newly defined block into the Blocks Window . If you instead press "Cancel", the block is discarded.

Note: If the selected segments in the different sequences differ in width, the shortest width is used as the width of the block.

Link (Alignment Menu)

Use this command to align the currently selected sequence segments and create "links" between the residues appearing in each affected column. When residues are linked, they will remain in the same column despite any shifts that occur elsewhere in the alignment. Linked residues are shown in upper case in the Alignment window and with a thickened bar in the Schematic window. They may also be shaded for further emphasis with the appropriate Shading option.

Unlink (Alignment Menu)

Use this command to remove links that were created with the Link command. Note that, in some cases, this may result in some extra gaps that are not needed to properly align the blocks. If so, the unnecessary gaps may be removed with the Compress command.

Compress (Alignment Menu)

Use this command to remove unnecessary gaps from the alignment by shifting sequence segments as far as possible to the left, while still respecting whatever residue links may exist. Unnecessary gaps may result from use of the Unlink command.

Shading (Alignment Menu)

Use this group of commands to change the way the graphical and text alignments are shaded in the Schematic and Alignment windows or to remove the shading.

Setting	Effect
None	No shading is performed.
Pair-score	For each residue, a shade is selected based on the score matrix entry of that residue compared with the one in the same column of the selected row. In the selected row, this means that residues are compared with themselves and, even though all such comparisons will be exact matches, some matches are "worth more" than others. Therefore, you will observe some variation in the shading of the selected row, with darker shades indicating residues that are generally less mutable.
Mean-score	For each column, a shade is selected based on the score matrix values averaged over all possible pairs of residues in that column.
Links	For each residue, a shade is selected based on the number of residue links that it participates in. Linked blocks spanning all of the sequences in the data set will have the darkest shade, while blocks involving fewer sequences will be given progressively lighter shades.

Summary Info (Alignment Menu)

Use this command to change settings and preferences that affect the entire project.

Dialog Box Options

Title	Enter an (optional) title for the alignment. It will appear in the main window just below the menu bar as will as on printed output.
Sequence Type	Select the type of sequence (protein, DNA, or RNA) that the alignment will contain.
Scores	Choose from the list of score tables that are appropriate for the selected sequence type (see above). When aligning proteins, a variety of score tables from the PAM and BLOSUM series are available (default is BLOSUM62). For DNA and RNA, there is only one score matrix called "Nucleotide", which awards +5 for a match and -6 for a mismatch.
Colors	Choose from the list of color schemes to be used in shading blocks.
Significance	Typically, one is interested in only whether a particular block is significant and not its actual p-value. Select "number" if you would like to see p-values reported for the statistical significance of blocks. If you would prefer something more qualitative, try "yes/no/maybe" (see below).

What does "yes/no/maybe" mean ?

yes	$< 10^{-4}$ (significant)
maybe	10^{-4} to 10^{-1}
no	$> 10^{-1}$ (not significant)

Window Menu Commands

Choose the command for which you need Help.

Commands

Cascade
Tile Horizontal
Tile Vertical
Arrange Icons
Schematic
Alignment
Blocks
Color Map

Cascade (Window Menu)

Use this command to arrange two or more open windows in a stack, with a slight offset so that you can see the title bar of each.

Tile Horizontal (Window Menu)

Use this command to arrange two or more open windows side-by-side like tiles on a floor. Horizontal tiling favors wider, rather than taller windows.

Tile Vertical (Window Menu)

Use this command to arrange two or more open windows side-by-side like tiles on a floor. Vertical tiling favors taller, rather than wider windows.

Arrange Icons (Window Menu)

Use this command to arrange the icons for closed windows along the bottom of MACAW's main window.

Schematic (Window Menu)

Use this command to open the Schematic Window if it is an icon or bring it to the front if it is already open.

Alignment (Window Menu)

Use this command to open the Alignment Window if it is an icon or bring it to the front if it is already open.

Blocks (Window Menu)

Use this command to open the Blocks Window if it is an icon or bring it to the front if it is already open.

Color Map (Window Menu)

Use this command to open the Color Map Window if it is an icon or bring it to the front if it is already open.

Help Menu Commands

Choose the command for which you need Help.

Commands

Contents
Search For Help On
How to use Help
About MACAW

Contents (Help Menu)

Use this command to display the Contents for MACAW's Help file.

Search For Help On (Help Menu)

Use this command to start the Help system and display the "Search" dialog box, where you type or select keywords for the topic you are seeking

How To Use Help (Help Menu)

Use this command to get help on how to use the Microsoft Windows hypertext help system. Choosing this command is the same as pressing F1 when you are using Help.

About MACAW (Help Menu)

Use this command to display information about MACAW, including the author, version number and release date.

MACAW Support Policy

Scientists at the National Center for Biotechnology Information (NCBI) develop a variety of computational tools (MACAW is one example) as part of their own research and make them available to the scientific community. Descriptions of software tools and algorithms are published in appropriate journals, and both executables and source code are available upon request from the authors. However, the NCBI cannot provide user support for these tools.

If you have questions on the use of MACAW, first consult this on-line help document as well as the original published sources (see Abstracts). If you still have questions, you may send email to Greg Schuler (schuler@ncbi.nlm.nih.gov) or Stephen Altschul (altschul@ncbi.nlm.nih.gov), however, you should not expect a prompt reply. Please do not call NCBI expecting to get support for MACAW.

MACAW Release Notes

Release Notes for Version 2.0.0

What's New

- A new algorithm for finding ungapped local multiple alignments (blocks), the Gibbs Sampler (Lawrence, et al., 1993), has been incorporated.
- On-line help has been added (if you're reading this, you already know about it).
- Score tables and color definitions are now specified in a new Summary Info dialog box instead of requiring the use of the File Open command as before. In addition, the sequence type (protein, DNA, or RNA) is explicitly specified in the same dialog box (instead of being inferred from the selected score table).
- The choice between "yes/no/maybe" and numerical styles of statistical significance reporting is now a global setting that is selected in the Summary Info dialog box (previously it was set on a per-block basis).
- You can now use the Copy command to put the selected block (in the Blocks Window) on the clipboard (as plain ASCII text) for pasting into other applications.
- As a consequence of the above changes, the menus have been changed around quite a bit. Take a few moments to familiarize yourself with the new layout.

Known Bugs and Limitations

- Although the hard limit on the number of sequences you can load has been increased (to 32), there is a bug that causes the program to hang on 22nd (approx) sequence. It is recommended that you limit yourself to 20 sequences until a fix can be found.
- The program may crash when using the Segment Pair Overlap method and the score cutoff is too low. In general, it is a good idea to cancel a search if the progress monitor stalls at 25% for more than a few seconds.
- The program may crash if you attempt to link a block that is inconsistent with blocks that are already linked. If you want to link something that, in the Schematic Window appears to "cross over" a linked block, unlink the old block first.

Revision History

Date	Version	Platforms
01-Mar-91	1.00	Win16
16-Apr-91	1.01	Win16
10-May-91	1.02	Win16
01-Aug-92	1.03	Win16
20-Apr-92	1.04	Win16
15-Jun-92	1.05	Win16
12-Dec-92	1.06	Win16
19-Oct-93	2.0.0	Win16, Win32i, Win32a

See Also:
About MACAW Version Numbers
MACAW System Requirements

MACAW Support Policy

About MACAW Version Numbers

A new style of version numbering has been initiated beginning with the 2.0.0 release. The full version specifications consists of three numbers separated by decimal points and followed by a character string indicating the platform for which the executable was built (see MACAW System Requirements for a list of possible platforms), for example "2.0.0 Win32i". The least significant (rightmost) number is incremented for maintenance releases involving bug fixes or trivial additions. The middle number is incremented when a new feature is added but there is little overall change to the operation of the program. The most significant is incremented for major releases involving, for example, significant changes to the user interface, any change to the file format, or significant numbers of new features. In previous releases, "1.00" through "1.06", only two numbers were used and the 16-bit Windows platform was assumed in all cases.

System Requirements

Beginning with release 2.0.0, MACAW is available for multiple platforms, as shown in the following table, along with the system requirements for running MACAW on each platform.

Platform	Requirements
Win16	Microsoft® Windows or Microsoft® Windows for Workgroups, version 3.1 or greater. A personal computer that is based on the Intel® (or compatible) 80386 (or better) microprocessor and has at least 4 MB of memory (8 MB recommended), along with a Windows-supported graphics display and mouse.
Win32i	Microsoft® Windows NT for Intel, version 3.1 or greater. A personal computer that is based on the Intel® (or compatible) 80386 (or better) microprocessor and has at least 12 MB of memory (16 MB recommended), along with a Windows-supported graphics display and mouse.
Win32a	Microsoft® Windows NT for Alpha αXP, version 3.1 or greater. A personal computer that is based on the DEC® Alpha αXP microprocessor and has at least 12 MB of memory (16 MB recommended), along with a Windows-supported graphics display and mouse.

How To Specify the Sequence Type

MACAW can be configured to align either protein or nucleotide sequences. When you first install the program, the sequence type will be set to "protein", but if you will be aligning nucleotide sequences, use the Summary Info command to change the sequence type to either "DNA" or "RNA" (the choice between these two only affects whether "T"s or "U"s are shown on the display, but for all other purposes they are equivalent). When the sequence type is changed, the new type becomes the default for all new projects.
Note: Since the sequence type determines which sequence alphabet to use when importing sequences, it must be set to the desired state before importing any sequences.

See Also:

New Project (File Menu)

Import Sequence (Sequence Menu)

How To Add Sequences to an Alignment

When you launch MACAW or create a new alignment project, it will initially be empty and must be populated with the sequences to be aligned. Sequences may either be imported from disk files or transferred to MACAW from another program via the clipboard.

Use the Import Sequence command to import a sequence file that is in the FASTA format. FASTA files may contain multiple sequences, so it is often most convenient to prepare one file that contains all (or most) of the sequences that will go into one alignment. If you use the Entrez or NetEntrez programs to retrieve sequences from the databases supplied by CD-ROM or over the Internet, you can export them directly to FASTA files.

If you have sequence files in other formats, say GenBank flat file format, you can still incorporate the data into a MACAW alignment, albeit in a more roundabout way that makes use of the Windows clipboard.

- Run a text editor or word processor designed for Windows (e.g. notepad).
- In the other application, open the sequence file. Normally this is done using the Open command on the File menu.
- Select the sequence text, omitting any annotation that may be present.
- Use the other application's Copy command to put the sequence text on the clipboard.
- Use MACAW's Paste command to insert the sequence into the alignment.

Note: Since only the sequence is transferred, the name and description fields will initially be empty, but you can use the View Sequence command and edit them as desired. An exception to this rule is the case where the sequence data in the other application is in the FASTA format. In this case, you can include the header line in the selection to be put on the clipboard and MACAW will interpret the name and description exactly as if it had been imported from a file. This is useful in the case where you want to import just one sequence from a file that contains many.

Note: Be sure that the sequence type is correctly set so that MACAW will know whether to use the protein or nucleotide sequence alphabet when interpreting sequence data from either a file or the clipboard.

See Also:

How To Specify the Sequence Type

FASTA File Format

How To Select a Region of an Alignment

A number of menu commands expect a selected region of the alignment to be "selected" to serve as the target of whatever operation they perform. Search For Blocks and Define Block are good examples of this. The selection consists of no more than one contiguous segment on each of several sequences (but need not span all of the sequences). A region may be selected in either the Schematic Window or the Alignment Window by simply holding down the left mouse button while dragging the mouse to enclose the desired region in the selection marquee. Upon releasing the button, the selected region will be inverted.

Normally, the previous selection is cleared as soon as the mouse button is pressed, so the new selection replaces the existing one. However, there are two ways to extend the existing selection instead.

Using the right button instead of the left changes the selected segments of sequences within the marquee but preserves the previous selection for the other sequences. Think of this as extending across sequences.

In the other hand, pressing and holding the Shift key during the selection process extends the selection within sequences. For each sequence, the segment within the marquee is merged with the previous selection.

If the region you want to select corresponds to a block you have found or manually defined, simply clicking on its entry in either the Search Results Window or the Blocks Window will set the selected region to correspond to the block. This is most commonly done prior to using the Link command.

See Also:

- Schematic Window
- Alignment Window
- Blocks Window
- Search Results Window
- Search For Blocks (Alignment Menu)
- Define Block (Alignment Menu)
- Link (Alignment Menu)
- Unlink (Alignment Menu)

How To Extend a Block to More Sequences

You may find a block that involves only a subset of the sequence in the project, either because the score cutoff used with a segment pair overlap search was too high or because patterns in some of the sequences did not quite match a regular expression. The following strategy is often useful for extending that block to the remaining sequences.

- In the Search Results window, select the block that you want to extend.
- In the Schematic Window, extend the selection by holding the right mouse button and drawing a box around each of the sequences that is not represented in the original block.
- Repeat the search, but this time use a lower cutoff or a more inclusive regular expression or use a different search method, such as the Gibbs Sampler, which is more likely to find subtle patterns.

You can afford to relax the stringency of the search in this case since it will be performed on a reduced search space.

See Also:

How To Select a Region of an Alignment

Searching for Blocks