CBAT

Computational Biology Alignment Tools

Created By Caleb Sizemore

Version 3

User Assistance Guide
<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 Introduction</td>
<td>01</td>
</tr>
<tr>
<td>02 CBAT Edition</td>
<td>02</td>
</tr>
<tr>
<td>03 Menu Screen</td>
<td>03</td>
</tr>
<tr>
<td>04 Setup Screen</td>
<td>04</td>
</tr>
<tr>
<td>04.01 Number of sequences</td>
<td>04.01</td>
</tr>
<tr>
<td>04.02 Processes</td>
<td>04.02</td>
</tr>
<tr>
<td>04.02.01 Convert To DNA</td>
<td>04.02.01</td>
</tr>
<tr>
<td>04.02.02 Convert To mRNA</td>
<td>04.02.02</td>
</tr>
<tr>
<td>04.02.03 Convert To Protein</td>
<td>04.02.03</td>
</tr>
<tr>
<td>04.02.04 Find Double Strand</td>
<td>04.02.04</td>
</tr>
<tr>
<td>04.02.05 Global Alignment</td>
<td>04.02.05</td>
</tr>
<tr>
<td>04.02.06 Semiglobal Alignment</td>
<td>04.02.06</td>
</tr>
<tr>
<td>04.02.07 Local Alignment</td>
<td>04.02.07</td>
</tr>
<tr>
<td>04.02.08 Score Alignment</td>
<td>04.02.08</td>
</tr>
<tr>
<td>04.02.09 Longest Common Subsequence</td>
<td>04.02.09</td>
</tr>
<tr>
<td>04.02.10 Shortest Common Supersequence</td>
<td>04.02.10</td>
</tr>
<tr>
<td>04.02.11 Fragment Assembly</td>
<td>04.02.11</td>
</tr>
<tr>
<td>04.02.12 Star Alignment</td>
<td>04.02.12</td>
</tr>
<tr>
<td>04.02.13 Best Alignment</td>
<td>04.02.13</td>
</tr>
<tr>
<td>04.03 Input type</td>
<td>04.03</td>
</tr>
<tr>
<td>04.04 Scoring</td>
<td>04.04</td>
</tr>
<tr>
<td>05 Input Screen</td>
<td>05</td>
</tr>
<tr>
<td>06 Processing Screen</td>
<td>06</td>
</tr>
<tr>
<td>06.01 Error Screen</td>
<td>06.01</td>
</tr>
<tr>
<td>07 Output Screen</td>
<td>07</td>
</tr>
<tr>
<td>07.01 Matrix Viewer</td>
<td>07.01</td>
</tr>
<tr>
<td>08 Additional Tools</td>
<td>08</td>
</tr>
<tr>
<td>08.01 Protein / DNA Matrix</td>
<td>08.01</td>
</tr>
<tr>
<td>08.02 Blosum62 Scoring Matrix</td>
<td>08.02</td>
</tr>
<tr>
<td>08.03 PAM250 Scoring Matrix</td>
<td>08.03</td>
</tr>
<tr>
<td>09 Hot keys</td>
<td>09</td>
</tr>
<tr>
<td>10 Quick Questions</td>
<td>10</td>
</tr>
<tr>
<td>11 Disclaimer</td>
<td>11</td>
</tr>
</tbody>
</table>
Introduction

CBAT (pronounced c bat) stands for Computational Biology Alignment Tools. It provides beginning level tools to assist with the introductory study of bioinformatics. Developed over the course of two semesters, it started as a fulfillment of class assignments and became a complete program. Along with providing powerful and fast computational tools for sequence alignment and converting, it has an extremely user friendly environment. Whether you are using CBAT to run complex sequences or just to check some computations, it is the program for you.
CBAT Edition

CBAT has two editions, full and educational. The educational edition is for use by professors that want to turn features on after the students have learned it by hand. There are 12 different features in CBAT that can be turned on and off. When all 12 features have been turned on, you then have the full edition of CBAT.

The 12 features that can be turned on and off are:
- blosum & pam scoring
- output matrix tool
- best alignment
- converting tools
- fragment assembly
- global alignment
- l c subsequence
- local alignment
- score alignment
- s c supersequence
- semiglobal alignment
- star alignment
Menu Screen

The menu screen has four buttons which give you access to main parts of the program.

Start CBAT - takes you to the setup screen so that you can begin using the program.

Additional Tools - takes you to the tool menu.

Assistance - open the assistance menu.

About CBAT - displays information about the program.
Setup Screen

The setup screen is where the functions and options for all sequence manipulation are set. The on the setup screen there are four basic selection methods.

+ and -  The plus and minus are used to increase and decrease an amount

>>  The little double arrows indicates a dropdown menu.

[ ]  The check box is to turns a function on and off.

Then there are the buttons along the bottom of the screen as well.
Home — will return you to the main menu.
? — opens the assistance screen.
Input — will continue to the input screen where you enter your sequences.

Each item found on the setup screen is covered in the following sections.

Hot Keys:
Use the [~] key to view the assistance for the current selected process.
Use the [+ ] and [-] keys to increase and decrease the number of sequences.
04.01

Number of sequences

The number of sequences not only determines the total amount of sequences you will be imputing but also which processes are available to be used. There are three groups, processes that run a single sequence, double sequences, and three or more sequences.

Refer to processes to learn which processes are available in each group.
The main functionality of CBAT comes from the following 13 processes. The number of sequences handled, input types, scoring methods, options, and a description of the functions are listed for each process.

04.02.01
Process: Convert To DNA
Number of sequences: 1
Input type: mRNA & Protein
Scoring: Not used.
Options: None exist.
Function:
This process has two possible functions. When the input type is set to mRNA a double stranded DNA sequence will be produced and the output will display this. When the input type is set to Protein the total number of possible DNA sequences will be calculated. This is done because of the huge quantity of output that is possible when converting a DNA sequence to a protein sequences. However, you can use the Protein/DNA Matrix to find any of the possible sequences. The output will display the input with the number of possible DNA groups under it. The total possible DNA sequences, up to a million, will be listed below the output box.

04.02.02
Process: Convert To mRNA
Number of sequences: 1
Input type: DNA & Protein
Scoring: Not used.
Options: None exist.
Function:
This process does the same thing as Convert To DNA does, just for mRNA instead of DNA. Refer Convert To DNA for a description.

04.02.03
Process: Convert To Protein
Number of sequences: 1
Input type: DNA & mRNA
Scoring: Not used.
Options: None exist.
Function:
This process takes a single DNA or mRNA sequence and converts it to a double stranded protein sequence. Proteins are made from a group of three DNA characters. If the sequence that is input is not perfect divisible by three then a selector is provided on the output screen to view all possible protein sequences.
04.02.04
Process: Find Double Strand
Number of sequences: 1
Input type: DNA & mRNA
Scoring: Not used.
Options: None exist.
Function:
This process finds the double strand of a single DNA or mRNA sequence.

04.02.05
Process: Global Alignment
Number of sequences: 2
Input type: DNA, mRNA, & Protein
Scoring: User Defined, Blosum62, and PAM250.
Options: Number of Alignments and Matrix Viewer.
Function:
This process allows multiple input types and scoring methods to be used in performing global alignment on two sequences. Set the optimal alignments option to view up to 70 different alignments of sequences that get the same score. The output screen displays the aligned sequences with any matches between them. On this screen there is also the number of alignments found (if more then one), length of the sequences, alignment score, and the total number of matches, mismatches, and gaps found in the alignment. Use the matrix button at bottom right of the screen to run the matrix view.

04.02.06
Process: Semiglobal Alignment
Number of sequences: 2
Input type: DNA & mRNA
Scoring: User Defined only.
Options: Alignment method and Matrix Viewer.
Function:
This process performs three versions of basic semiglobal alignment which are Containment, Suffix-Prefix, and Prefix-Suffix. Refer to Global Alignment for a description of the output screen.

{Illustrations}
<table>
<thead>
<tr>
<th>Containment</th>
<th>Suffix-Prefix</th>
<th>Prefix-Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>xxxxxxxxxxx</td>
<td>xxxxxxxxx -</td>
<td>xxxxxxxxxx</td>
</tr>
<tr>
<td>-xxxxxx -</td>
<td>xxxxxxxxxx</td>
<td>xxxxxxxx</td>
</tr>
</tbody>
</table>
04.02.07
Process: Local Alignment
Number of sequences: 2
Input type: DNA & mRNA
Scoring: User Defined only.
Options: Matrix Viewer
Function:
This process finds the local alignment of two sequences if it exists. The sequences output are of the local alignment only. Refer to Global Alignment for a description of the output screen.

04.02.08
Process: Score Alignment
Number of sequences: 2
Input type: DNA, mRNA, & Protein
Scoring: User Defined, Blosum62, and PAM250.
Options: None exist.
Function:
This process scores two sequences without aligning them. It allows multiple input types and scoring methods, along with permitting gaps within the input sequences. When inputting sequences for this process use a negative - to represent a gap. The process will remove gaps over gaps and add gaps to the end of sequences to make them the same length. Refer to Global Alignment for a description of the output screen.

04.02.09
Process: Longest Common Subsequence
Number of sequences: 2
Input type: DNA & mRNA
Scoring: Not used.
Options: None exist.
Function:
This process finds the longest common subsequence (lcs for short) of two sequences. The lcs is found with the use of a modified Global Alignment function which scores the match a positive and the mismatch and gap as neutral. Then the matches of the aligned sequences create the lcs. The output screen displays the lcs with the two sequences aligned below it.

04.02.10
Process: Shortest Common Supersequence
Number of sequences: 2
Input type: DNA & mRNA
Scoring: Not used.
Options: None exist.
Function:
This process finds the shortest common supersequence (scs for short) of two sequences. The scs is found with the use of a modified Global Alignment function which scores the match a positive and the mismatch and gap as neutral. Then the two sequences now aligned are copied in to a single sequence called the scs. The output screen displays the scs with the two sequences aligned below it.
04.02.11
Process: Fragment Assembly
Number of sequences: 3 or more
Input type: DNA & mRNA
Scoring: User Defined only.
Options: None exist.
Function:
This process will produce a single sequence consensus from 3 or more fragment sequences with the use of the greedy algorithm. It not only finds the consensus of multiple fragments it also handles full and partial containment. This process is by far the most complex of all the functions along with requiring the longest amount of processing time. The output screen displays the fragment assembly consensus at the top of the output box with the all the input sequences aligned below it.

04.02.12
Process: Star Alignment
Number of sequences: 3 or more
Input type: DNA, mRNA, & Protein
Scoring: User defined, Blosum62, and PAM250.
Options: Set Star.
Function:
This process is a global alignment factory. It runs all the input sequences through the global alignment function to produce a pairwise matrix. From the pairwise matrix a single sequence (called the star), which scored the highest when aligned with all the other sequences, is selected. The global alignment is again run on all the sequences with the star to build the star alignment. The star can also be set to a cretin sequence prior to processing. The output screen displays the star followed by all the other sequences aligned below it. The pairwise score for the alignment is listed below the output box.

04.02.13
Process: Best Alignment
Number of sequences: 3 or more
Input type: DNA, mRNA, & Protein
Scoring: User Defined only.
Options: Number of Alignments.
Function:
This process, which is similar to the star alignment, finds the best alignment of two sequences from the group input. It is helpful when dealing with a handful of sequences which might all align but you need to know which score the highest. Refer to Global Alignment for a description of the output screen and the number of alignments option. The matrix view is not an option for this process since it deals with more then 2 sequences.
04.03

Input type

The input type selection varies for each process, but may contain any or all of the following types:

DNA sequence - uses A, C, G, & T
mRNA sequence - uses A, C, G, & U

Refer to processes to learn which input types are available for each process.
Scoring

Scoring allows you to set the values used in the alignment functions. Scoring is not available for all processes, but when it is available the options it has are based on the input type that has been chosen.

DNA (User Defined)
mRNA (User Defined)
Protein (User Defined, Blosum62, PAM250)

When User Defined scoring is selected you can set the values for match, mismatch, gap opening, and gap extension. A simple matrix is used to compute the alignment when gap opening is set to zero, otherwise a CAB matrix is used. However because of the special setup of Blosum62 and PAM250 you are only given the options of gap opening, and gap extension when they are selected and they always use a CAB matrix.

(example)
\[
\begin{array}{cccccc}
  c & c & c & c & g & t \\
  c & - & - & - & t & t
\end{array}
\]

Refer to processes to find out when scoring is available.
Input Screen

The input screen is where you enter the sequences. At the top of the screen is the process and input type that has been selected. The number of input boxes depends on the number of sequences you have chosen to use. Each box has the sequence number and a button that will only clear the information in it. Along the bottom are some important buttons.

Setup — return to the setup screen without losing any of your input.

? — opens the assistance screen.

Recall — loads the last input you used. You will lose any information currently in the input boxes.

Clear — clears every input box.

Process/Next — when this button reads process then when pressed it will begin to process the sequences you have input. When the button reads next it will display the next group of input boxes.

< (back arrow) — it is only displayed when there are previous input boxes to go back to.

Any information that has been entered remains unchanged while moving from setup to input and back, the next or previous buttons, or when an error is found with your input. The recall information is only updated with your last input when a process has been successfully completed.

Hot Keys:
Use the [~] key to view the assistance.
Use the [TAB] key to jump to the next input box.
Processing Screen

The processing screen is displayed while the sequences are being to run. The first function that is run on them is sequence cleanup. All of the sequences are striped of any characters that are not contained in the selected input type set. If any errors are found the processing is stopped and the error is displayed. Otherwise the processing continues and provides you with an up-to-date status. Each process displays its own individual status. Processing run time varies for each process and is highly effected by the quantity and length of sequences being used.

Refer to error screen to learn what causes an error and how to avoid them.

06.01 Error Screen
The error screen appears during processing when an error is found. There are three main problems that cause errors.

Nothing left — after the sequences were cleaned nothing remained. This is caused by leaving an input box empty then processing or by not entering the correct characters which are then cleaned out and nothing remains.

Not enough left — occurs when converting to a Protein. Three or more correct characters must be in the sequences.

Duplicate sequences — are not allowed.

A common mistake that causes most errors is forgetting which input type was selected. This causes an error because after the sequences have been cleaned nothing is left. In some cases however enough is left for the processing to continue even though you will no longer be returned the correct output. To help avoid this, the input type is always listed at the top of the screen.
Output Screen

The output screen displays the results of the sequences after processing. All the output screens are similar but have an individual setup that fit the needs of each process.

Refer to processes to learn more about each output screen.

Hot Keys:
Use the [~] key to view the assistance.
Use the [arrow] keys to manipulate the output box view.

07.01 Matrix Viewer
The matrix viewer provides an on screen printout of the alignment matrix with scores and arrows on it. There are two kinds of matrixes that can be viewed, basic which is a single matrix and a CAB matrix which is a complex three layer matrix. Along the bottom of the matrix are some important buttons.

Back — returns you to the main output screen.
? — opens the assistance screen.

The following controls only apply to matrixes that are larger then 8 by 8 in height or width.
*\ (box with line) — moves the matrix view window to (0,0 position) the top left.
\_ (line with box) — moves the matrix view window to the bottom right.
Arrows — moves the matrix view window one space left, right, up, and down.

CAB matrix controls are the same as the basic except for one addition.
C, A, B — select which of the three matrixes to view. The top left square displays the matrix that is currently being viewed.

Only some of the processes are able to take advantage of this feature. A matrix button will be located on the output screen at the bottom right when the matrix viewer is available for use.

Refer to processes to learn which once are able to use the matrix viewer.

Hot Keys:
Use the [~] key to view the assistance.
Use the [arrow] keys to move around the matrix.
Use the [Page Up] and [Page Down] keys to move to the top left or to the bottom right of the matrix.
Use the [c], [a], and [b] keys when viewing a CAB matrix to change between the matrixes.
Additional Tools

08.01 Protein / DNA Matrix
The protein / DNA matrix is a relationship chart for matching a protein with DNA or DNA with a protein. Their are two menus that are used to control the chart. Click on a character from the protein menu or select three, one from each column, from the DNA menu to active the chart. When the chart is active the screen will display the protein 1-letter code, protein 3-letter code, the protein full name, and a list of DNA groups. Use the buttons along the bottom of the screen to return to the tool menu, the main menu, or to receive assistance.
Refer to processes for full protein / DNA sequences converting.
Hot Keys:
Use the [~] key to view the assistance.
Use the [Lower Case a-z] keys to select a Protein.

08.02 Blosum62 Scoring Matrix
The Blosum62 scoring matrix is an interactive chart for finding the Blosum62 score of two protein characters with each other. Use the two protein menus on the left to select a protein A and a protein B to be scored. When the chart is active the screen will display the score along with highlighting the chart.
Refer to processes - score alignment for full sequence scoring.
Hot Keys:
Use the [~] key to view the assistance.
Use the [Lower Case a-z] keys to select a Protein A.
Use the [Upper Case A-Z] keys to select a Protein B.

08.03 PAM250 Scoring Matrix
The PAM250 scoring matrix is an interactive chart for finding the PAM250 score of two protein characters with each other. Use the two protein menus on the left to select a protein A and a protein B to be scored. When the chart is active the screen will display the score along with highlighting the chart.
Refer to processes - score alignment for full sequence scoring.
Hot Keys:
Use the [~] key to view the assistance.
Use the [Lower Case a-z] keys to select a Protein A.
Use the [Upper Case A-Z] keys to select a Protein B.
Hot keys

A hot key is when a shortcut to a button in the program has been mapped to a keyboard key. Quite a few hot keys can be found throughout CBAT. Learn to use them and navigate through the program faster. Each section of this assistance lists any hot keys that are related to it. You may also like to view the full list of all hot keys below.

Use the [~] key on any screen that has a ? button to view the assistance for that screen.

Use the [+] and [-] keys on the setup screen to increase and decrease the number of sequences.

Use the [TAB] key on the input screen to jump to the next input box.

Use the [arrow] keys on the output screen to manipulate the output box view.

Use the [arrow] keys on the matrix viewer screen to move around the matrix.

Use the [Page Up] and [Page Down] keys on the matrix viewer screen to move to the top left or to the bottom right of the matrix.

Use the [c], [a], and [b] keys on the matrix viewer screen when viewing a CAB matrix to change between the matrices.

Use the [Lower Case a-z] keys on the additional tools to select a Protein A.

Use the [Upper Case A-Z] keys on the additional tools to select a Protein B.
Quick Questions

Why was CBAT developed with Flash?
- CBAT was developed using Flash MX version 6. While Flash is not as powerful a development tool as C++ or Java, it was by far the best choice for this program. CBAT is a learning tool for people who are new to the world of bioinformatics, so it was important to have this program as user friendly as possible. It was also important to have it available for online use. Flash offers better visual aspects and usability than any other web tools currently available, as well as extremely powerful and complex action script that provided the string manipulation needed.

Why do some of the processes not allow me to select them?
- This is a feature of CBAT (Educational Version) which allows the professor to turn processes and options on and off. If you think that you should have access to the full version try restarting CBAT or refreshing the screen. If you are still having problems please email someone from the home page that linked you to this program.
Refer to CBAT Versions to see which version you are running and which, if any, functions are turned on for use.

What is the difference between Gap Opening and Gap Extension?
- Gap extension is the score given to ever gap encountered in an alignment. The gap opening score is added to the gap extension score for single gaps and the beginning of a gap group. Gap opening and extension are used to build a CAB Matrix and only the gap extension is used when building a basic matrix. Processes that only use the basic matrix have the gap opening set to zero and locked.

How come Convert To DNA does not work with a Protein input?
- It does work, it just does not give you a single DNA sequence. Only the total numbers of sequences are shown because of the huge quantity of output that is possible when converting a DNA (or mRNA) sequence to a protein sequences. However, you can use the Protein/DNA Matrix to find any of the possible sequences.

What is the best way to load sequences?
- You may be testing different groups of sequences and wish to use one that you have entered before. CBAT does not have a database of any kind. It only remembers the last input you used. The best way to save you sequences for use later is to create a text file that you can then copy and past them from and to.

How do I stop once I have started processing?
- Most processes do not take very long so just let them run. The only way to really end processing is to close the window and then restart the program.
What is the Do you want to abort the script? all about?
- The following message will pop up if you are trying to process extremely long sequences. A script in this movie is causing Macromedia Flash Player 6 to run slowly. If it continues to run, your computer may become unresponsive. Do you want to abort the script? (Yes) (No)  This is not an error and you should click No if you want to continue to process your sequences. This occurs when Flash runs for more the 15 seconds on any one frame. To avoid this use smaller sequences or stand by while processing and click No when you are prompted.

Why does it take so long for my sequences to be processed?
- Processing run time varies for each process and is highly effected by the quantity and length of sequences being used.

Why does the matrix button not work?
- This is a feature of CBAT (Educational Version) which allows the professor to turn processes and options on and off. If you think that you should have access to the full version try restarting CBAT or refreshing the screen. If you are still having problems please email someone from the home page that linked you to this program. Refer to CBAT Versions to see which version you are running and which, if any, functions are turned on for use.

How do I swap the x and y sequences on the matrix?
- To swap the x and y sequences of a matrix, simply return to the input screen and swap S1 with S2 then process again.

How do I select between the Basic Matrix and the CAB Matrix?
- Only the following three processes use the Matrix View: Global Alignment, Semiglobal Alignment, and Local Alignment. Of those three processes, Global Alignment is the only one that is able to view a CAB Matrix which is used when the scoring Gap Opening is not set to zero.

Why does CBAT give me a different answer then what is in my book?
- Many of the alignment functions produce numerous outcomes but only one is selected to be displayed on the output screen. Because of this, sometimes output is interpreted as wrong when in fact it is correct. It is just a different alignment which scored the same as another.

How can I save my output?
- There is really not a good way of saving output. CBAT is more like a bioinformatics calculator then anything else. Just type you results into a text file or write them down.
Disclaimer

While this program has been tested a great deal, there is no guarantee that it is free from all bugs. Also note that many of the alignment functions produce numerous outcomes but only one is selected to be displayed on the output screen. Because of this, sometimes output is interpreted as wrong when in fact it is correct. It is just a different alignment which scored the same as another. It is highly recommended that you double check any output that CBAT provides to be sure that the information is correct.