Introduction to Molecular Biology and Bioinformatics Workshop

Biosciences eastern and central Africa - International Livestock Research Institute Hub (BecA-ILRI Hub), Nairobi, Kenya
May 5-16, 2014

Bioinformatics Manual

Sequence data analysis with CLC Main Workbench

Written and compiled by Joyce Njuguna, Mark Wamalwa, Rob Skilton
Getting started with CLC Main Workbench

A. Quality control of sequenced data
B. Assembling sequences
C. Nucleotide and protein sequence manipulation
D. BLAST sequence search
E. Aligning sequences
F. Building phylogenetic trees

Welcome to CLC Main Workbench -- a user-friendly sequence analysis software package for analysing Sanger sequencing data and for supporting your daily bioinformatics work.

Definition

Sequence: the order of nucleotide bases [or amino acids] in a DNA [or protein] molecule

DNA Sequencing: Biochemical methods used to determine the order of nucleotide bases, adenine(A), guanine(G), cytosine(C) and thymine(T) in a DNA strand

TIP

CLC Main Workbench includes an extensive Help function, which can be found in the Help menu of the program’s Menu bar. The Help can also be shown by pressing F1. The help topics are sorted in a table of contents and the topics can be searched. Also, it is recommended that you view the Online presentations where a product specialist from CLC bio demonstrates the software. This is a very easy way to get started using the program. Read more about online presentations here: http://clcbio.com/presentation.

A. Getting Started with CLC Main Workbench

1. Download the trial version of CLC Main Workbench 6.8.1 from the following URL: http://www.clcbio.com/products/clc-main-workbench-direct-download/

2. Double click on the CLC Main Workbench download file (.exe for windows) to install the program. Follow the on-screen instructions.

3. Your tutor will give you instructions for importing your license.

4. Open the CLC Main Workbench software

5. The user interface of CLC Main Workbench has:

   - A Toolbar at the top (Figure 1)
   - The Navigation Area is located on the left side of the screen. It is used for organizing and navigating through data. Its behavior is similar to the way files and folders are usually displayed on your computer.
• The **View Area** is the right-hand part of the screen, displaying your current work. The **View Area** may consist of one or more **Views**, represented by tabs at the top of the **View Area**

![Image of CLC Main Workbench Interface](image)

**Figure 1. The CLC Main Workbench Interface showing navigation area, view area, toolbar and the toolbox**

6. **Creating a folder.** It is best to organize data in folders in the navigation area.
   - To create a folder go to **File | New | Folder**, or click on the **New** icon on the tool bar and then select **Folder**.
   - Name the folder **Sequences** and click **OK**.

**Definition**

**Sequence trace file**: raw sequence data in the form of a chromatogram from a Sanger DNA sequencing reaction.

7. **Importing data**: This allows you to bring sequenced data into CLC from where it is stored on your computer. You can import many file types but in this tutorial we will look at importing DNA sequence trace files for further analysis.
• If your sequences are in an email, then save them first in a folder on your computer.
• Go to File | import, or click the Import icon on the tool bar.
• Navigate to where your sequences are stored on your computer. There should be two sequences for your own PCR product; one generated using a forward primer and another generated with a reverse primer. For each sequence there should be two files: a .seq file (a text file containing the sequences of bases), and an .ab1 file (the chromatogram of the sequence).
• Cntrl click to select the .ab1 files to import (See Figure 2, below). Then click Next.

![Figure 2. Importing data into CLC](image)

• Select your folder (Sequences) to save your sequences into it.
• Click Finish to save the sequences in CLC.
• Your sequence data has now been imported into CLC. If you expand your Sequences folder in the navigation area you should see your sequences.
• Double click on one of the sequences in the Navigation area to view it. Your imported sequence should look as in Figure 3, below.
B. Quality control of sequenced data

Definition

Trim sequence: removal of low quality data at the beginning and end of a sequence to improve the quality of your data.

Heterozygous: Having dissimilar alleles that code for the same gene

1. The first step in analysis of you sequence is to access the quality of the trace files and trim off poor quality reads so as to have good quality data for further analysis.
   • In most cases sequences do not start off very clearly about the first 20-30 bases (see Figure 4 below). This is due to non-fully activated taq polymerase at the start of the reaction. The bases close to the primer also do not get many terminations on them making their sequenced data unreliable.
   • The ends of sequences (base 500-800) also have low quality scores. This is due to diminishing nucleotides at this point of the reaction.
2. It is essential to know what the attributes of an ideal sequence are so as to aid in assessing the quality of your data

- **A good sequence** has smooth, evenly spaced peaks. Each peak has one color; the peaks may vary in height up to about 3-fold. The base line noise on a good sequence needs to be very minimal (see **Figure 5** below). An ideal sequence should not have overlapping peaks or two peaks at the same position.

![Figure 5. Example of an ideal/good sequence trace file](image)

- **Non-usable sequenced data**: when your data resembles (**Figure 6**) having non-distinct and overlapping peaks it is incomprehensive and unreliable for analysis. This presentation of sequenced data could be due to low concentration of DNA template also due to addition of the wrong or no primer to the sequencing reaction.

![Figure 6. Example of non-usable sequenced data](image)

- **Double peaks**: at times your data may have multiple peaks of same or different height at the same position (**Figure 7**). This could be due to clone contamination or a contaminated PCR reaction. However the double peaks could be due to a heterozygous position (two bases at the same position SNP).
  - In the case of a heterozygous position or a position in which the exact base is uncertain, we can use degenerate codes to represent the data.
3. **Trimming the sequences:** After carefully scrutinizing your sequence you can determine where your reliable sequence starts and ends. You then delete/trim the unreliable sequence from each end of your sequence file. Trimming eliminates base reads of poor quality *(and in the case of cloned inserts, the removal of vector contamination).*

   - To trim sequences go to **Toolbox menu | Sequencing Data Analysis | Trim Sequences.**
   - This gives the Trim Sequences dialogue box. Select all the sequences to be trimmed on the left panel and use the arrow to move them to the right panel, then click Next.
   - Specify the trimming parameters. The limit on trim quality scores can be increased or decreased to make trimming more or less stringent respectively.
     - Select **Ignore existing trim information**
     - Select **Trim using quality scores. Limit 0.05**
     - Select **Trim using ambiguous nucleotides. Residues 2**
     - Do not select vector trimming *(your PCR products were not cloned into a vector)*

   ![Figure 8. Trim sequence parameters](image)

   - Click ‘Next’ and choose to save the results.
• When the trimming is done the parts of the sequences that are trimmed are not actually removed but trim annotations are saved to the sequences. These annotated sections are ignored in further analysis.

![Trimming annotation](image)

Figure 9. Trimming annotations saved on sequences

C. Assembling sequences

**Definitions**

*Sequence assembly:* aligning and merging fragments of a much longer DNA sequence in order to reconstruct the original sequence.

*Assembly to a reference sequence:* assembly of the sequenced fragments using a similar sequence (a reference sequence) as a guide.

*De novo assembly:* assembly using only your sequenced fragments. This can be more difficult if the region of overlap is very small.

*Conflict:* a nucleotide position in the assembled sequence overlap which has different bases, or where there is a gap on one sequence and a base at the corresponding position on the other sequence.

*DNA Contig:* Contiguous sequence of DNA created by assembling overlapping DNA sequence reads. It consists of a multiple sequence alignment of sequence reads plus a consensus sequence.

*Consensus sequence:* A sequence of nucleotides [or amino acids] in common between regions of homology in related DNA [or protein] sequences.

1. **Assemble sequences:** In most cases forward and reverse primers are used to sequence a DNA fragment, such as a PCR product. Therefore, on sequencing you end up with a forward and reverse sequence. This was done for your PCR product in the Workshop.
In sequence assembly, the two sequences (forwards and reverse) are aligned where they overlap to get a contiguous sequence called a 'contig'.

- Assembly can be done to a reference sequence (e.g. a similar sequence obtained from a sequence database)
- Or de novo assembly can be done (we will do de novo assembly in the Workshop)

2. **De novo** assembly.
   - Go to Toolbox | Sequencing Data Analysis | Assemble sequences
   - Select the forward and reverse sequences you want to assemble and move them to the right panel (Figure 10 below).

![Figure 10. De novo assembly of sequences selecting forward and reverse sequences](image)

- Click **Next**. The dialogue box allows one to define assembly parameters (see Figure 11).
- You can increase or decrease the alignment options to change the stringency of the alignment. Select the following options for your assembly.
  - Select **Trimming**: Trim sequence ends before assembly
  - Select **Alignment options**: Minimum aligned read length: 25
  - Select **Alignment stringency**: Medium
  - Select **Conflicts**: Vote (A, C, G, T)
  - Select **Output options**: Create full contigs, including trace data
- Click **Next**.

9
The next dialogue box allows you to set trimming parameters if you had not trimmed the sequences prior to this. In any case, you can select the following parameters:

- Select **Ignore existing trim information**
- Select **Trim using quality scores. Limit 0.05**
- Select **Trim using ambiguous nucleotides. Residues 2**
- Do not select vector trimming (*your PCR products were not cloned into a vector*)

- Click **Next**.
- Select **Results handling: Save**. By selecting the **Save** option you can save your assembled contig in another folder to avoid confusion.
- Click **Next**.
- Click the **Folder Icon** and create a new folder (Figure 12) and label it **Contigs**. Highlight the new folder and click **Finish**.
- Your sequences will assemble and your contig, with the filename **Contig 1**, will be saved to the new **Contigs** folder. *An error message will appear on the screen if your sequences do not assemble.*
• Expand the Contigs folder in the navigation area you will see your assembled contig 1.
• Rename **Contig 1**: Highlight the file, go to **Edit menu select Rename**, then rename with **Sample_number**.
• The contig can be viewed by double clicking on it. The contig should appear as in **Figure 13**, having the alignment of **forward** and **reverse** sequence and a **consensus** sequence. You can also see the regions that were **trimmed** (greyed out), and any base **conflicts** that occur between the forwards and reverse sequences. **Coverage** indicates the regions of the consensus that were derived from one or both sequences.

Figure 12. Saving the assembled contig

Figure 13 The contig showing the forward, reverse and consensus sequences
3. **Resolving conflicts:** Conflicts arise during sequence assembly of the forward and reverse sequences as a result of a base mismatch in the overlapping regions of the forward and reverse sequences.

- To find any conflict in the contig click the ‘Find Conflict’ button on the side panel (Figure 14). You can zoom in to enhance visualization using the Zoom In tool on the toolbar.

![Figure 14. Finding and resolving conflicts](image)

- The above example shows a conflict in which the forward strand shows a base call of an "A" and the reverse strand shows a gap in that position.
- We assess the quality of reads at this position. In the Figure 14 example, examination of the reverse (lower) sequence shows us that the quality of chromatograph traces is low, which is often the case towards the ends of the sequence. However, the forward (upper) strand clearly has good quality peaks and can be trusted.
- To resolve the conflict, first select the erroneous position (in the example, the ‘gap’ in the reverse (lower) sequence): using the selection tool, drag the cursor over the position to select it. Put the cursor over the select position and right click, select **Edit selection** from the dialog box. Another dialog box may appear that says: **Warning: The sequence has quality scores. The sequence has trace data.** Click **OK**.
- Another dialog box appears that says **Replace Selection.** Replace the base (in the example it is a gap, written as ‘-’) with the appropriate base (in this example, the ‘-’ is replaced with an A).
- An editing warning pops-up: click **OK** to accept changes to the trace file.
- The red arrow that indicated a **conflict** now turns **green** and is labeled **Resolved conflict**.
- Click the **Find Conflict** button again to find the next conflict.
- Resolve all conflicts. Resolving conflicts is based on:
  - Quality of reads on each strand
  - Which of the peaks is of better quality
  - In some cases two bases may be picked on either sequence because it is genuinely a SNP position so judgment should be based on quality of reads but also background knowledge on the sequences being analyzed.
Degenerate codes – are used in the case that you cannot confidently resolve conflicts. See degenerate codes at http://www.bioinformatics.org/sms/iupac.html

4. Saving the consensus sequence: When all conflicts are resolved, you can extract the consensus sequence to be used for further analysis.

   • Right click on the name Consensus 1. Select Open copy of sequence.
   • A new window will appear with the consensus sequence
   • Put cursor on the tab of the new window.
   • Right click. Select Save as.
   • Select Folder icon, and make a new folder with the name Consensus_Sequences. Then enter name of the consensus sequence file: PCR_product_name_consensus. Highlight the new folder, and click OK.
   • Your consensus sequence will be saved in the new Consensus_Sequences folder, and is now ready for further analysis.
   • Close the file PCR_product_name_consensus by clicking on the X on the tab.
   • Then click on the X on the tab of the open window of your contig file, PCR_product_name_Contig. A dialog box will ask if you want to save changes to the file. Click Yes to save.

5. Remove PCR primer sequences from your consensus sequence:

   • Create a new folder. Name the folder: Primers
   • Import your PCR primer sequences into the Primers folder

   Plants: Rubisco primers with M13 tails
   rbcLa-forM13 TGTAAAACGACGGCCAGT ATGTCACCACAAACAGAGACTAAAGC
   rbcLa-revM13 CAGGAAACAGCTATGAC GTAAAATCAAGTCCACCRCG

   Animals:
   VF1d_t1 TCTCAACCAACCACAARGAYATYG
   VR1d_t1 TAGACTTCTGGGTTGCCRAAAYCA

   • Use Trim sequences option to remove the primer sequences from your consensus (by selecting the option Trim contamination from saved sequences) or manually remove the primer sequences.

Tip
You now have four folders:
 i) Your_Name_Sequences containing your trace files
 ii) Contigs containing your contig
 iii) Consensus_Sequences containing your consensus sequence
 iv) Primers containing your PCR primer sequences

Review each of the folders and files to make sure you understand how each file was derived.
Definition

**FASTA format**: this is a file format demarcated by the first line / title line starting with a greater than symbol “>” followed by the lines of sequence data (bases or amino acids). The tile line holds the sequence description. [http://en.wikipedia.org/wiki/FASTA_format](http://en.wikipedia.org/wiki/FASTA_format)

6. **Exporting sequences**: You can export your consensus sequence out of CLC and save it in another location on your computer.
   - Highlight the sequence and go to File | Export or click the export icon on the task bar
   - This brings a dialogue box that allows you to select the file format you want to save your sequence in and the location in your computer to save the sequence file.
   - You can select to export your sequence in the **FASTA (.fa / .fsa / .fasta) format** as it is a common file type for many sequence analysis programs.

   - **Plant Group**: export your consensus sequence and save it in another location on your computer. Email the sequence to all the members of the Plant Group. Conversely, save the sequences from your colleagues in Your_name folder.

   - **Animal Group**: export your consensus sequence and save it in another location on your computer. Email the sequence to all the members of the Animal Group. Conversely, save the sequences from your colleagues in Your_name folder.

**D. Nucleotide and protein sequence manipulation**

Definition

**Genes**: stretches of DNA that encode information for building proteins.

**Codon**: triplet of nucleotides that code for an amino acid (the building blocks of proteins)

**Start codon**: this is either of two codons AUG or GUG, that signal for the initial translation and the first amino acids in a polypeptide chain

**Stop codon**: this is any of three codons UAA, UAG or UGA that signal the termination of the synthesis of a protein.

**Restriction site**: location on a DNA molecule (with a palindromic sequence) that is recognized by restriction enzymes as a site to cleave.

**Translation**: this is the process by which RNA is converted to an amino acid sequence.
**Open reading frame:** Sequence of nucleotides in DNA that contain no stop codon and potentially translate to a peptide chain.

**Motif / domains:** this is a functional subunit of a protein

---

**I. Nucleotide sequence manipulation**

There several ways you can manipulate your nucleotide sequence to make more sense of the data within it. For this you will use your already generated consensus sequences.

1. **Convert DNA to RNA:** The difference between RNA and DNA is that RNA has the nucleotide uracil instead of thymine. The process of converting DNA to RNA is known as transcription. To convert your DNA sequences to RNA in CLC:

   - Go to Toolbox | Nucleotide analysis | Convert DNA to RNA.
   - This gives you a dialogue box that allows you to select the sequences you want to convert and move them to the right side panel.
   - Click next
   - Result handling select Save
   - Create a folder name it RNA-sequences to save in your RNA sequences.
   - From your navigation area you can see the folder containing the RNA sequences, double click on a RNA sequence to view it.

2. **Reverse Sequence:** In the case that you have a DNA sequence in its 5’ – 3’ direction and would like to view it in its 3’ – 5’ direction, you can reverse it using CLC to do this:

   - Go to Toolbox | Nucleotide analysis | Reverse Sequence.
   - This gives you a dialogue box that allows you to select the sequences you want to reverse and move them to the right side panel.
   - Click next
   - Result handling select Save
   - Create a folder name it Reverse-sequences to save in your reverse sequences.
   - From your navigation area you can see the folder containing the reverse sequences, double click to view the reverse sequence.

3. **Reverse complementary strand:** DNA occurs as a double helix strand. One of the strands occurs in the 5’ – 3’ direction while the complementary strand occurs in the 3’ – 5’ direction. Each base pair in a double helix structure must pair with its complementary strand:
   - Adenine pairs with thymine with a double hydrogen bond.
   - Guanine pairs with Cytosine with a double hydrogen bond.
In the case that you have one strand you may need to determine its complementary strand in its 5’ – 3’ directions. To do this you get the reverse compliment strand, on CLC this can be done by:

- Go to **Toolbox | Nucleotide analysis | Reverse compliment sequence**.
- This gives you a dialogue box that allows you to select the sequences you want to reverse compliment and move them to the right side panel.
- Click **next**
- Result handling select **Save**
- Create a folder to save your reverse compliment sequences.
- From your navigation area you can see the reverse compliment of your sequence double click on it to view it.

4. **Restriction site analysis**: this is key for finding the specific sequence on a DNA that restriction enzymes can cleave. It also aids one to find which enzymes can be used to cleave a given sequence. To do this on CLC

- Go to **Toolbox | Cloning and Restriction sites | Restriction Site Analysis**.
- This gives you a dialogue box that allows you to select the sequences that you want to use in performing restriction site analysis.
- Click **next**
- The next dialogue box allows you to select the enzymes to be considered. Select the enzyme(s) you wish to use for your restriction analysis and move them to the right side panel.
- **Click next**
- Number of cut sites: Select **one restriction site**
- **Click next**
- Result handling
  - Output Options: Select **Add restriction site annotations to the sequence**
  - Output Options: Select **Create restriction map**
  - Resulting handling: **Save**
- Create a folder to save your restriction enzyme analysis.
- From your navigation area you can see the result folder double click on it to view it.

5. **Translate nucleotide sequence to protein sequence**: this is the process by which a gene within a DNA sequence is converted to its corresponding amino acid sequence. Translation of DNA to Amino acids can be done in 6 possible reading frames, 3 in both strand directions see **Figure 15** below.
To translate a DNA sequence to amino acid sequence in CLC:

- Go to Toolbox | Nucleotide analysis | Translate to Protein.
- This gives you a dialogue box that allows you to select the sequence you want to translate and move it to the right side panel.
- Click next
- The next dialogue box allows you to set translation parameters, you can select the following parameters Figure 16:
  - Translation of whole sequence: select all the 6 reading frames
  - Do not select: Translation of coding regions
  - Genetic code translation table: select 1 Standard
  - Click next
- Result handling select Save
- Create a folder name it Protein-sequences to save in your translated sequences.
- From your navigation area you can see the folder containing the translated sequences, double click to view.
II. Protein sequence manipulation

Once you have translated your nucleotide sequence you now have amino acid sequences that you can manipulate variously.

1. Physico-chemical properties: it is possible to determine the physico-chemical properties of a protein sequence. These include the number of amino acids (sequence length), amino acid composition, molecular weight, aliphatic index, hydrophobicity plots and atomic composition. To determine this on CLC:

   • Go to Toolbox | Protein analysis | Create Protein report.
   • This gives you a dialogue box that allows you to select the protein sequences you want to analyse and move them to the right side panel.
   • Click next.
   • The next dialogue box allows you to choose the report content
     o Select Sequence statistics
     o Select Protein charge plot
     o Select hydrophobicity plot
     o Select Complexity plot
     o Select Secondary structure prediction
     o Click next
   • The next dialogue box allows you to set parameters for sequence statistics:
     o Layout: select individual statistics layout.
     o Do not select background distribution
   • The next dialogue box allows you to set parameters for hydrophobicity plot:
     o Hydrophobicity scale: select Kyte-Doolittle
• Window size: **select 11** (Increase window size, if your sequence is longer)

The next dialogue box allows you to set parameters for Complexity report

• Window size: **select 11** (Increase window size, if your sequence is longer)

• Result handling select **Save** and do not make a log

• Create a folder name it **Protein-properties** to save in your protein report.

• From your navigation area you can see the folder containing the protein report, double click to view.

2. **Motifs and Domains**

Motifs and domains are functional sub-units. Part of functional annotation of your protein sequence is finding motifs and conserved domains in your protein sequence. To search for domains on CLC:

• Go to **Toolbox | Protein analysis | Pfam Domain Search**

• This gives you a dialogue box that allows you to select the sequence you want to translate and move it to the right side panel.

• Click **next**

• The next dialogue box allows you to set parameters, you can select the following parameters:
  
  o Database and search type: **Search full domains and fragments.**
  o Database: **Select 100 most common domains**
  o Additional databases: **Open Plug-in and Resource manager**
  o Significance cutoff: E-value 1.0
  o **Click next**

• Result handling:
  
  o Output potions: **select add annotations to sequence and create table**
  o Result handling: **select save**
  o Log handling: Do not make log.
  o **Click next**

• Create a folder name it **Proteins Domains** to save in your domain search results.

• From your navigation area you can see the folder containing the domain results, double click to view.

• If there no domains found in your sequence an error message will appear on your screen.
E. BLAST sequence search

Definition

**Biological database:** computerized archive to store, organize and retrieve sequenced data.

**NCBI:** National center for biotechnology Information. A resource that provides analysis and retrieval of data from the GenBank and other public biological databases.

**Sequence similarity:** this is when two sequences are very alike in base pair or amino acid sequence.

**Homology:** Homologs diverge from a common ancestor and homology is inferred by the sequence, structural and functional similarity.

**Pairwise sequence alignment:** comparison of two sequences to establish conserved regions and residue-to-residue identity.

**Global alignment:** Finds optimal alignment over the entire length of the two compared sequences

**Local alignment:** aligns short regions of similarity between sequences.

**BLAST:** Basic Local Alignment Search Tool

- **Blastn:** Search a nucleotide database using a nucleotide query
- **Blastp:** Search a protein database using a protein query
- **Blastx:** Search a protein database using a translated nucleotide query
- **tblastn:** Search a translated nucleotide database using a protein query
- **tblastx:** Search a translated nucleotide database using a translated nucleotide query

BLAST is a tool used to calculate similarity for biological sequences, it find best local alignments. BLAST is used to search a sequence against a biological database. It returns a display of results as a list of sequences matches ordered by statistical significance.

**Use BLAST to find DNA sequences in databases**

1. Start by selecting a nucleotide sequence (one of your consensus sequences) to blast.
2. Then go to Toolbox | BLAST | BLAST at NCBI
3. Click Next
4. Choose program and database (select the right program see definitions section)
   - Program: blastn: (this program Blastn searches a DNA sequence against a DNA database, looking for DNA sequences with homologous regions to your nucleotide query sequence)
   - Database: Nucleotide collection (nr)
5. Click Next

6. Choose parameters
   - Limit by entrez query: All organisms
   - Choose filter: select Filter low complexity
   - Expect: 10
   - Word size: 11
   - Match/mismatch: Match 1, Mismatch -3
   - Gap costs: Existence 5, Extension 2
   - Max number of hit sequences: 50

7. Click Next

8. Results handling: select Save
   Log handling: do not select Make log

9. Click Next

10. Select/create folder for saving your Blast results

11. Click Finish to accept the parameter settings and begin the BLAST search.

12. Your computer now contacts NCBI and places your query in the BLAST search queue. After a short while the result should be received and opened in a new view.

13. A dialog box will appear informing you when the analysis has finished.

14. Double click on the Blast file in your folder to open it you should see results as below (Figure 17). Use the zoom in tool on the task bar to enable visibility of bases.

![BLAST result screenshot](image)

**Figure 17. BLAST results; zoom in to view the base pairs**
15. **Interpreting Blast results:** The figure shows the graphical display of blast results (also referred to as blast hits).

- The color is based on % similarity of your query sequence to the blast hit, green being the highest similarity down to black.
- The database hits are ordered by statistical significance.
- To view more details on the blast hit, **right click and select download and save** this will download its GenBank record.
- Click on the icons at the bottom of the page to view additional results (e.g. **show blast table** to view tabular results Figure 18)
- The tabular result below show a description of the hits in the 2nd column and on the third column is the E-value, which is a statistical measure of the significance of the hit.
- The icon **show text content** allows you to view the pairwise alignment between your Query sequence and the blast hit.

<table>
<thead>
<tr>
<th>Hit</th>
<th>Description</th>
<th>E-value</th>
<th>Score</th>
<th>%Ident</th>
</tr>
</thead>
<tbody>
<tr>
<td>F411949</td>
<td>Theileria parva clone Tp1 LS8 CD...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411941</td>
<td>Theileria parva clone Tp1 LS17 C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411942</td>
<td>Theileria parva clone Tp1 CDI C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411943</td>
<td>Theileria parva clone Tp1 CDI C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411944</td>
<td>Theileria parva clone Tp1 CD14 C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411945</td>
<td>Theileria parva clone Tp1 CD5 C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411946</td>
<td>Theileria parva clone Tp1 CD7 C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411947</td>
<td>Theileria parva clone Tp1 CD6 C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411948</td>
<td>Theileria parva clone Tp1 CD9 C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411949</td>
<td>Theileria parva clone Tp1 CD10...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411950</td>
<td>Theileria parva clone Tp1 CD15...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411951</td>
<td>Theileria parva clone Tp1 CD16...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411952</td>
<td>Theileria parva clone Tp1 CD18...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411953</td>
<td>Theileria parva clone Tp1 CD7...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411954</td>
<td>Theileria parva clone Tp1 BD2 C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411955</td>
<td>Theileria parva clone Tp1 BD4 C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411956</td>
<td>Theileria parva clone Tp1 BD5 C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411957</td>
<td>Theileria parva clone Tp1 BD7 C...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411958</td>
<td>Theileria parva clone Tp1 BD14...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
<tr>
<td>F411959</td>
<td>Theileria parva clone Tp1 BA19...</td>
<td>0.00</td>
<td>393.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Figure 18 Tabular Blast results**

16. **How to make conclusion on your Blast results:** the purpose of doing a BLAST search is to find other sequences in the publicly available biological databases that are similar to yours (finding homologs to your query sequence). To make conclusions on your blast results:

- Check the **E-value**-Expected value (this indicates the probability that your blast hit may have occurred by random chance.)
- The lower the E-value (the closer it is to 0) the more significant your result is. To be certain of homology you E-value must be below **10^-4 or 0.001.**
- **% Identity** the higher your identity the increasing likelihood of homology
- **Query coverage** – if a hit has high query coverage and similarity it increase the chances of homology
F. Aligning sequences

CLC can align two or more sequences to find regions of similarity. To align sequences you need to use the consensus sequences or input sequences in FASTA format:

**Definition**

**Sequence alignment:** comparison of two sequences to establish conserved regions and residue-to-residue identity.

**Multiple sequence alignment:** Is the alignment of more than two sequences

1. From the toolbox menu select | Alignment and Trees | Create Alignments.
   This brings a dialogue box allowing you to select the sequences to align. Move the sequences you want to align to the right panel and click next.
2. Set parameters for the alignment algorithm
   - You can choose to increase the gap settings to make the alignment more stringent.
     - Gap open cost: 10
     - Gap extension cost: 1
     - End gap cost: as any other
   - Select either fast (less accurate) or slow (very accurate) alignment. Fast alignment allows for use of an optimized fast algorithm, which is best for very long sequences. It is preferable that you use slow alignment in the Workshop.
3. In the result handling step select the save option and click next
4. Create a folder for saving alignment results and click finish
5. Double click the alignment in the navigation area to view it.
6. The alignment should look as in Figure 19. The consensus is shown, and the sequence logo displays the frequency of residues at each position. In this example the bases in the alignment are color-coded.

![Figure 18 Sequence alignment](image-url)
G. Building phylogenetic trees

It is possible to infer evolutionary relationships for aligned sequences. In CLC this can be done either by distance-based methods or by statistical methods (Maximum likelihood approach).

**Definition**

*Phylogenetics:* Study of evolutionary relatedness of organisms.

*Evolution:* change in distribution of allele frequencies from one generation to the next

*Phylogenetic tree:* branching diagram that infers evolutionary relationship of various species based on their physical or genetic traits

*Evolutionary relatedness:* is indicated by sequence similarity

*Bootstrapping:* test for the reliability of inferred tree topology

1. From the **toolbox menu** select | **Alignment and Trees** | **Create Trees**
   This allows one to create distance based phylogenetic trees.
   For Maximum likelihood trees **toolbox menu** select | **Alignment and Trees** | **Maximum Likelihood Phylogeny**.

2. This options open up a dialogue box that requires you to select the alignment on which your tree will be built. Select the alignment on the left panel and move it to the right panel and click next.

3. The **set parameters dialogue** box allows one to set the tree parameters.
   - For the distance based methods the tree building parameters include:
     a. Selecting the **algorithm**. Either **UPGMA** or **Neighbor Joining**.
        - **UPGMA** method assumes the evolution has occurred at a constant rate in the different lineages hence the root of the tree is estimated.
        - **Neighbor joining:** building a tree where the evolutionary rates are free to differ in the different lineages
     b. **Bootstrap** analysis is also done to evaluate the reliability of the inferred trees. The wizard allows one to adjust the number of replicates of bootstrap analysis. Set the bootstrap value to 1000 and click **next**.
     c. In result select the save option and create a folder where you can save your phylogenetic tree.
For the Maximum Likelihood tree a number of parameters to be set are different (Figure 20).

a. **Set starting tree**: specify a starting tree for the tree reconstruction:
   - Neighbor joining
   - UPGMA or
   - A tree from a file

b. **Select substitution models.** Maximum likelihood is estimated under the assumption of one of four substitution models
   - Jukes Cantor (Jukes and Cantor, 1969)
   - Kimura 80 (Kimura, 1980)
   - HKY (Hasegawa et al., 1985)
   - GTR (Yang, 1994a)

c. **Rate Variation**: Substitution rates can be allowed to differ among the individual nucleotides by selecting the Include rate variation option.

d. **Estimation**: Selecting the **estimate substitution rate** option maximum likelihood values of the free parameters in the rate matrix describing the assumed substitution model are found. If the **Estimate topology** option is selected, a search in the space of tree topologies for that which best explains the alignment is performed. The **Estimate Gamma distribution parameter** allows estimation of the Gamma distribution parameter to be switched on or off.

e. Click next and choose either to save or open the tree and click next to finish the tree is generated and displayed. Once the tree is generated various tree view preferences can be selected, change the Tree layout or labels as desired.
Online Bioinformatic tools

1. Perform nucleotide sequence analysis:
   a) Go to the URL http://www.ebi.ac.uk/Tools/emboss/
   b) Select the Transeq sequence translation option
   c) Get a sequence of your choice (have a fasta nucleotide sequence in a file)
   d) Paste the sequence in the box provided and select under frame options (all six frames)
   e) Click Submit to translate sequence
   f) This return a 6-frame translation of your sequence you can then choose the correct reading frame.
   g) To get complementary and reverse complementary sequences go to the URL http://emboss.bioinformatics.nl/cgi-bin/emboss/revseq
   h) Paste your nucleotide sequence and select options for the reverse complement and complementary sequence.

2. Perform Protein sequence analysis:
   a) Use Protparam at the url http://web.expasy.org/protparam/ to get the Physio-chemical properties of the sequence.
   b) To search for Motifs and domains on Interpro go to the url http://www.ebi.ac.uk/Tools/pfa/iprscan/
   c) Structure prediction on HHpred find program at URL http://toolkit.tuebingen.mpg.de/hhpred
   d) Protein visualization with PyMOL