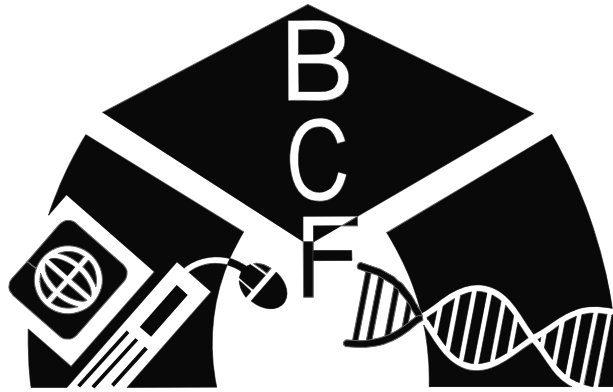


THE UNIVERSITY OF  
**ARIZONA**<sup>®</sup>  
TUCSON ARIZONA



**Biotechnology Computing Facility**

## **Introduction to GCG/SeqLab**

### **Part B**

**Electropherograms**  
**Multiple Alignments (ClustalW)**  
**NetBlast & NetFetch**  
**Editing Techniques & Features**

For assistance please contact:

<http://bcf.arl.arizona.edu/>

Gavin Nelson – 621.3206 – [gmn@email.arizona.edu](mailto:gmn@email.arizona.edu)  
Susan Miller – 626.2597 – [sjmillar@email.arizona.edu](mailto:sjmillar@email.arizona.edu)  
Nirav Merchant – 621.8379 – [nirav@arl.arizona.edu](mailto:nirav@arl.arizona.edu)



# Contents

---

Overview .....	3
<b>Preparations .....</b>	<b>4</b>
What you need .....	4
<b>Getting Started.....</b>	<b>5</b>
Before You Begin.....	5
Retrieving your Data from the Sequencing Facility .....	5
Launch SeqLab .....	6
<b>Opening Your</b>	
<b>Sequences\Electropherograms .....</b>	<b>7</b>
Importing Data .....	7
Viewing an Electropherogram .....	8
<b>Using your Electropherogram</b>	
<b>To Troubleshoot Sequences.....</b>	<b>9</b>
Not Enough DNA .....	9
Too much DNA .....	10
Mixed Sequence .....	10
Clean Vector, Noisy Insert .....	11
Sequence Dies Off .....	11
Homopolymer Region .....	12
No Sequence .....	13
Cleaning your Sequence .....	14
<b>Multiple Alignment using ClustalW.....</b>	<b>15</b>
Selecting Sequences .....	15
Grouping & Sorting .....	17
<b>Database Search &amp; Retrieval.....</b>	<b>19</b>
NetBlast & NetFetch .....	19
Viewing the New Sequences .....	22
Closing Up .....	24
<b>Summary.....</b>	<b>25</b>

## *Overview*

### **In this tutorial you will learn how to:**

- ❖ Download Electropherograms from the DNA Sequencing Facility.
- ❖ View Traces in SeqLab and analyze the quality of the sequences.
- ❖ Remove unwanted bases at the ends of sequences.
- ❖ Do a multiple alignment using ClustalW.
- ❖ Group & Sort sequences.
- ❖ Use NetBlast & NetFetch to search databases.
- ❖ View/Edit Features in sequences.

# Preparations

---

This is the second half of a two-part tutorial. Please complete 'Introduction to GCG/SeqLab Part A' prior to attempting this tutorial.

## *What you need*

To use this tutorial you will need an account on amadeus to access the GCG package. Use the online form at:

**<http://bcf.arl.arizona.edu/online-tools/new-account-form.html>**

You will then be sent a login and password to access the GCG server.

Programs in the GCG package can be run in two ways:

- Command line; UNIX prompt in a terminal window
- Graphical User Interface; SeqLab via BioDesk.

Command-line GCG consists of initializing the GCG package and then typing the commands.

**This tutorial does not cover command line procedures.**

The preferred method of using the SeqLab graphical interface, requires a familiarity with BioDesk\*.



**Please complete our 'Introduction to BioDesk' tutorial before continuing.**

You may need a File Transfer program to allow you to transfer files to and from your GCG account folder. We recommend SSH (Secure Shell). SSH Clients generally include both a command line option and a file transfer feature.

Please see the campus software repository for a "no cost" SSH client:

**<https://sitelicense.arizona.edu/ssh/ssh.shtml>**



BioDesk is a complete graphical interface to a variety of UNIX based genetic analysis software. Please see our web site for more information:  
<http://bcf.arl.arizona.edu/biodesk/>

# Getting Started

## Before You Begin

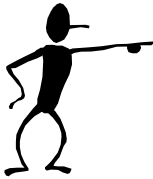
Before you begin working with this tutorial, it is suggested that you complete Part A – Introduction to the SeqLab Interface.

## Retrieving your Data from the Sequencing Facility

1. **Start your web browser and go to the following web site to pickup your sequenced DNA:**

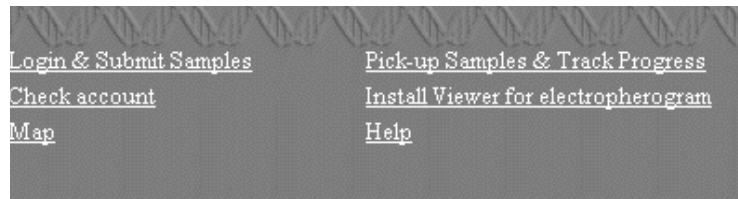
<http://uofadna.arl.arizona.edu/>

This will take you to the “System Login” page; note the toolbar on the top of the page.



2. **Select “Pick-up Samples & Track Progress” from the toolbar.**

This will take you to another login page, enter the following information where requested:



**Login:** demo  
**Password:** user

3. **Click on OK; then click on Submit for Tracking number.**

The Information Requested page appears.

### Information Requested:

Filename	Date Submitted	Sample Received Date	Date Queued	Status
<input checked="" type="checkbox"/> 12_1az_tmp_17303.CRO <input type="checkbox"/> 12_1az_tmp_17303.Seq	14-Apr-1999	14-Apr-1999	Not Processed	Ready to download
<input checked="" type="checkbox"/> 13_2az_tmp_17304.CRO <input type="checkbox"/> 13_2az_tmp_17304.Seq	14-Apr-1999	14-Apr-1999	Not Processed	Ready to download
<input checked="" type="checkbox"/> 14_3az_tmp_17305.CRO <input type="checkbox"/> 14_3az_tmp_17305.Seq	14-Apr-1999	14-Apr-1999	Not Processed	Ready to download

Make sure this box is checked if you wish to download all the sequence files (selected above) as a single multiple sequence FASTA format file

**4. Click on the Check-Boxes for the three .CRO files**

At the bottom of this page you will see a partial form.

Server	amadeus.biosci.arizona.edu
Login	demo2
Passwd	*****
Destination directory	tutorial
<input type="button" value="Send file(s)"/> <input type="button" value="Clear Selection"/>	

**5. Select 'amadeus.biosci.arizona.edu' for Server, enter your account login, password, and a destination directory of 'tutorial'.**

**6. Click on the 'Send file(s)' button.**

You will get a confirmation page that looks something like this:

Connected to host aretha.biosci.arizona.edu with login name demo25...

Changing path to tutorial ...  
Uploading /usr/home/ftp/pub/seq/CACHE2/12\_1az\_tmp\_17303.CRO -> sent  
Uploading /usr/home/ftp/pub/seq/CACHE2/13\_2az\_tmp\_17304.CRO -> sent  
Uploading /usr/home/ftp/pub/seq/CACHE2/14\_3az\_tmp\_17305.CRO -> sent

**7. Close web browser.**



Select the Check Box below the file list table to retrieve your sequences in FastA format. Trace information is lost, so it will be more difficult to clean and verify the validity of the sequence. However if you're only interested in a small segment, some time can be saved.

*Launch SeqLab*

The basic steps are as follows:

- 1. Launch BioDesk viewer.**
- 2. Connect to your session, and enter your password.**
- 3. Open the BioDesk menu (click on desktop) and select GCG/SeqLab**



# Opening Your Sequences\Electropherograms

## Importing Data

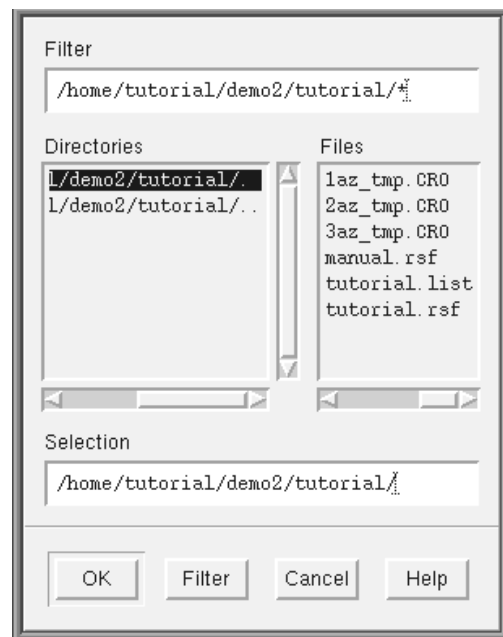
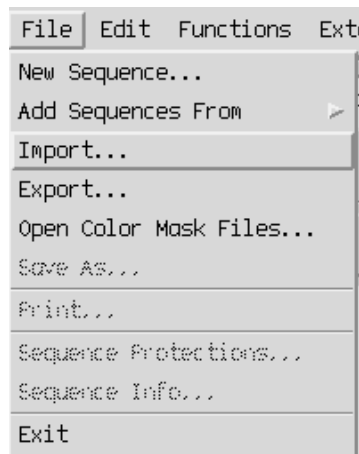
1. Make sure that you are in Editor mode.



a) Click on the “Mode:” option menu and select Editor.

2. Click on “File” from the menu bar, and select “Import...”

The Import Sequences dialog box appears.



You will see the 3 Electropherograms (.CRO files); possibly in addition to files created in Part A of this tutorial.

3. Double-Click on each of the Electropherograms to import them into the Editor window.

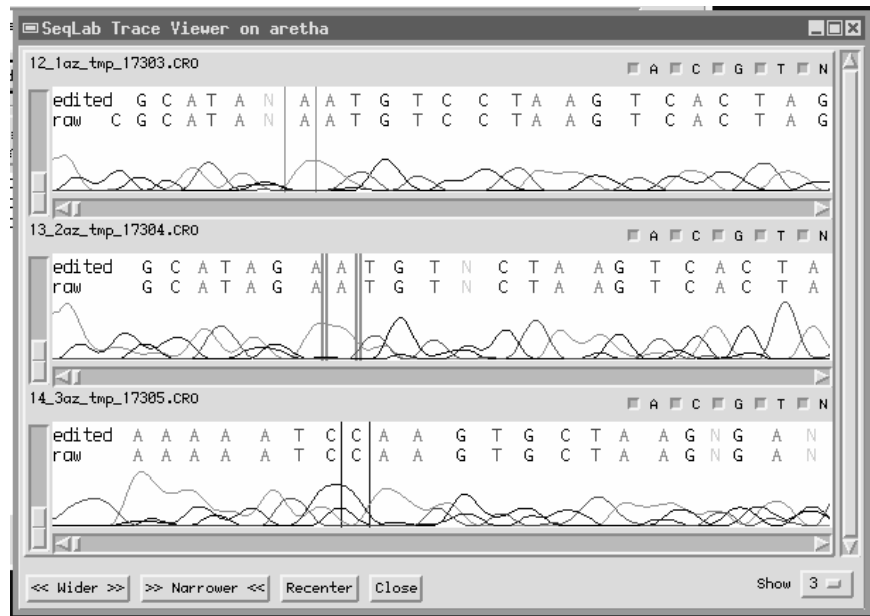
You should now see the three sequences in the Editor.

4. Click on Cancel, when you’re done importing.

## Viewing an Electropherogram

1. Click on “Windows” from the menu bar and select “Traces”.

The Trace Viewer window will appear.



2. Click on a base in the Trace Viewer window and you will see that the corresponding base in the Editor is highlighted with white.

You may have re-size/re-position your windows to see both at once.

You can use the slider bars and buttons to navigate.



A common use of the Trace Viewer is to examine ‘no calls’ to determine what the base should have been; one can then easily edit the sequence and place the correct base over the ‘no call’. (Set your keyboard action to Overstrike).

3. Click on Close.

The Trace Viewer disappears and you are left with the Editor window.

# Using your Electropherogram To Troubleshoot Sequences

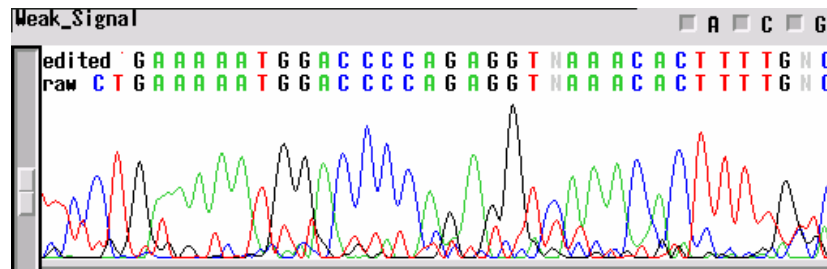
There are a number of potential problems that can effect the quality and reliability of a sequence produced by the Automated Sequencing Facility.

## Not Enough DNA

The most common problem associated with poor sequencing results is a low concentration of DNA. The easiest way to troubleshoot this using the Electropherogram is by looking at the signal strength.

When you receive an e-mail prompting you to pick up the sequences, there will be four values (one for each base) next to the file names. This is the sum of the area under the peaks of that given base, or the signal strength. A normal signal strength of a fragment 500 bp or longer should have values between 100 and 500 for each of these numbers. If you see that your signals are lower than this, you can assume that one of three problems could have caused this.

1. Template DNA concentration was overestimated.  
(This is usually the case)
2. Primer concentration was overestimated
3. Mismatch in primer sequence.  
(A single mismatch can cause the signal to decrease by ten-fold)

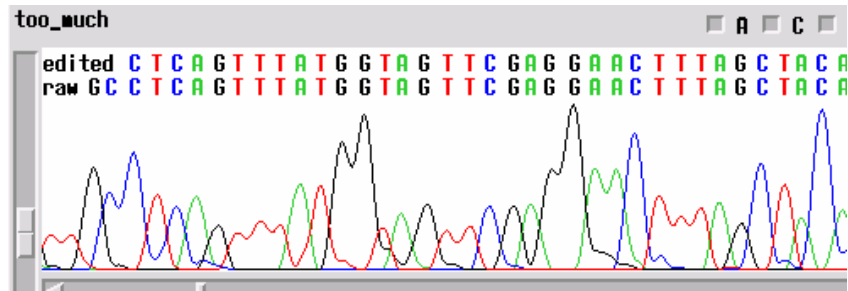


### Results of this on Sequence:

Often, when you get results with low signals, you will also get amplified background noise under the real peaks. This can cause no-calls or miscalls throughout the sequence.

## Too much DNA

Much like agarose gels, when a lane is overloaded on automated sequencers, the bands can become fuzzy and undelimited. You can tell that too much DNA was loaded if the signal strengths are over 1000 for four base values.



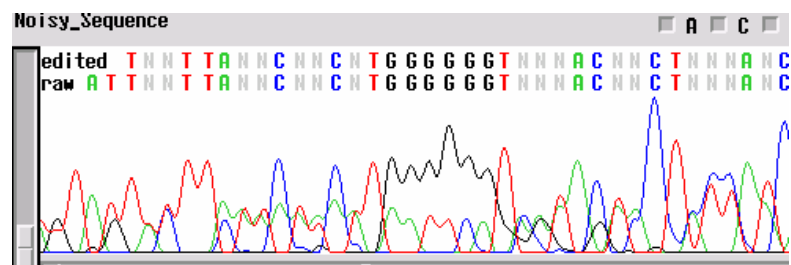
### Results of this on Sequence:

If too much DNA is loaded into the wells, the basepeaks will have rounded tips resulting in a decrease in the length of read.

## Mixed Sequence

This is commonly found with PCR products created directly from genomic DNA. If several products are amplified from the template, and a single product is not isolated using either gel extraction or another technique, the primers will sequence multiple products. You can tell that this is the case when you have a normal signal strength, however, there are several peaks in a single location. The possible causes of this problem are:

1. Multiple templates being sequenced by a single primer.
2. Multiple primers sequencing a template  
-Usually caused by a PCR reaction that did not have excess primers sufficiently removed.
3. Multiple binding sites on a single template  
-This is common when an insert is cut out of or amplified from a vector and subcloned into another. Often, individuals do not realize that they inserted a fragment that contains a promoter site into a vector with the same site. If we use the primer for this site, we will get overlapping sequence.

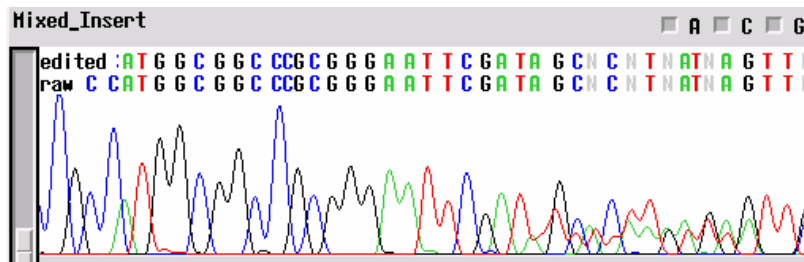


### Results of this on Sequence:

This problem results in several miscalls and no-calls. For most of these sequences, it is recommended that the data not be used.

### *Clean Vector, Noisy Insert*

With the popularity of TA cloning on the rise, this problem has become more prominent. When a subcloning procedure allows the insert to be oriented in either direction in the vector, positive colonies can have different sequences. If two positive colonies grow extremely close together and have different orientations, the Electropherogram will have clean peaks until the site of insertion, where sequences get mixed.



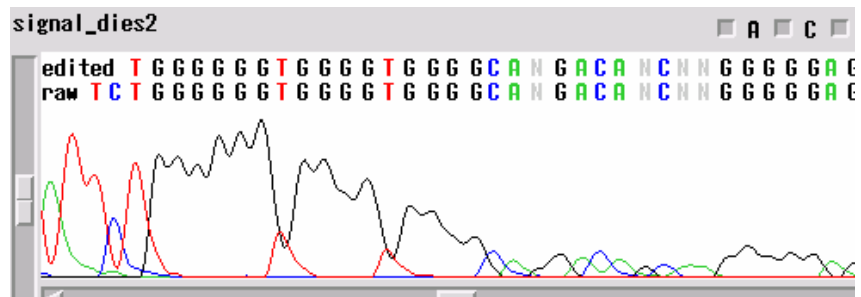
### Results on Sequence:

Unfortunately, most researchers do not care about the vector sequence, so these results are usually not readable

### *Sequence Dies Off*

A sequence can either gradually die off, or it can have a sharp drop-off where the peaks stop suddenly. The gradual die off is usually caused by a contaminant in the prep, causing enzyme efficiency to drop. Some contaminants that could cause this are EDTA, ethanol, PEG, Tris (high amounts), NaOAc, phenol/CHCl<sub>3</sub>, or CsCl. These could also result in no sequence at all if present in high enough concentrations.

A sudden drop-off is usually caused by a hairpin loop or a region with high G-C content. DMSO can be added to the sequencing reaction if the presence of a region such as this is expected.



### Results on Sequence:

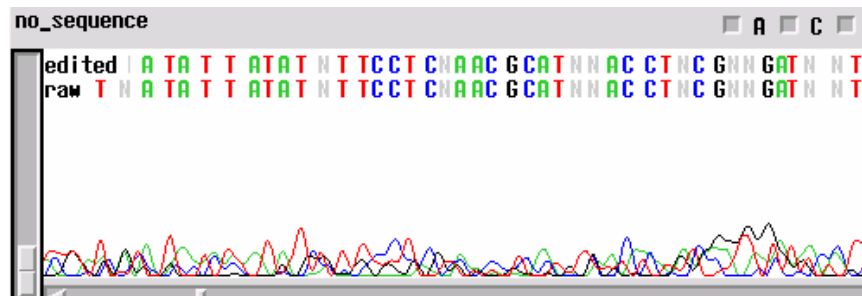
This will result in a decrease in length of read



## No Sequence

If there were no results generated by the reaction, you will be informed that the lane in the gel corresponding to that sample was blank. This could be caused by a number of different problems. Some of the more common causes are:

1. Not enough template DNA to generate visible results.
2. Not enough primer to generate visible results.
3. No primer site on the template DNA corresponding to the oligo in the reaction.
4. High enough concentration of contaminant to completely cease enzyme activity.



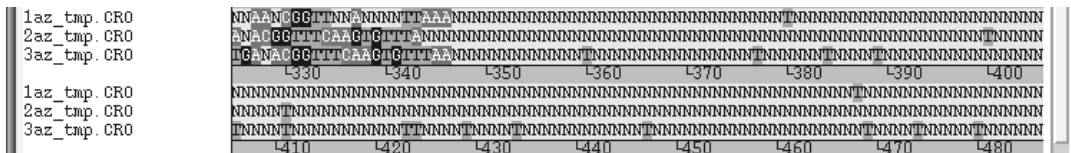
### Results on Sequence:

Although a sequence will still be available to download, the basecalls will have been generated from peaks found only in background noise. Please do not attempt to use this data if we inform you that the results were blank.

## Cleaning your Sequence



If your Editor is set to “Wrap”(or you scroll to the end of your sequences), you will notice that there are some “NNN” or “no calls” at the very end of the sequence, let’s remove those.



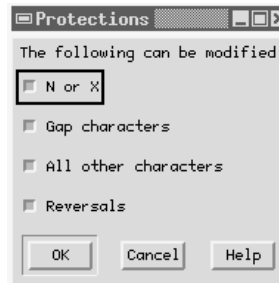
### 1. Select all three sequences.

You can use your mouse to Drag select, or use Control-Click.



### 2. Unprotect the sequences.

Select all options.



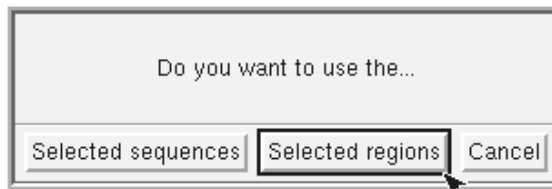
### 3. Use the mouse (click-drag) to select the bases you want to delete.

Since these are fairly long sequences, you may need to zoom out (~16:1) to select the entire length.



### 4. Click on the CUT button.

A “Which Selection” dialog box appears.



### 5. Click on “Selected Regions”.

# Multiple Alignment using ClustalW

## Selecting Sequences

### 1. Select a Region

- a) Using your mouse, drag a box across the region you want to align.

The bottom-right corner of the Editor window shows the start and end bases for your selection.

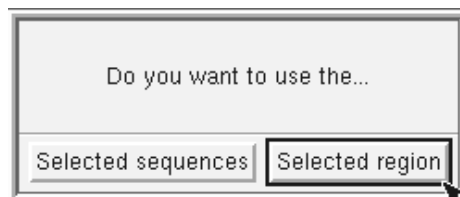
For this tutorial we will select from 1 to approximately 300.



The bases that are selected will turn white, indicating that they are selected.

- 2. Click on “Extensions” from the menu bar, and select “CulstalW...”

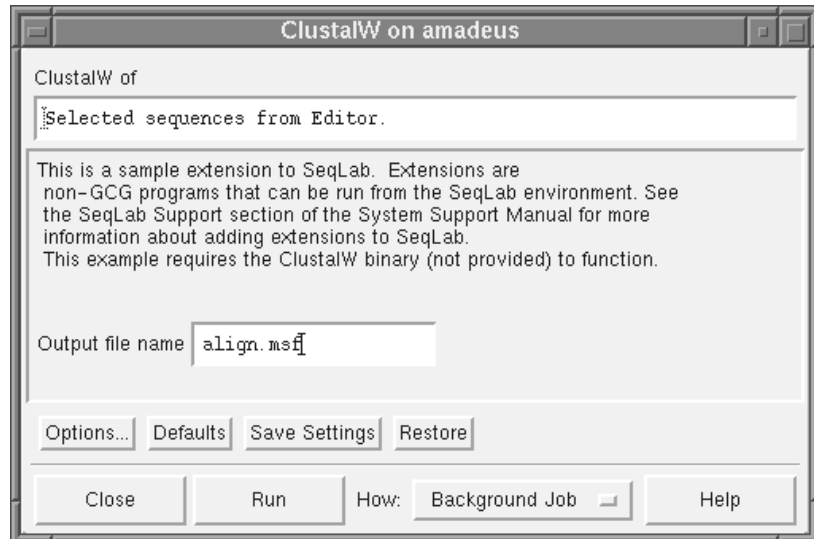
The “Which Selection?” dialog box appears.



- 3. Click on “Selected Regions”.

The ClustalW window appears.

## Running ClustalW



#### 4. Rename the output file from clustal.msf to align.msf

Make sure you end your file name with .msf and it doesn't have '-' in the file name.

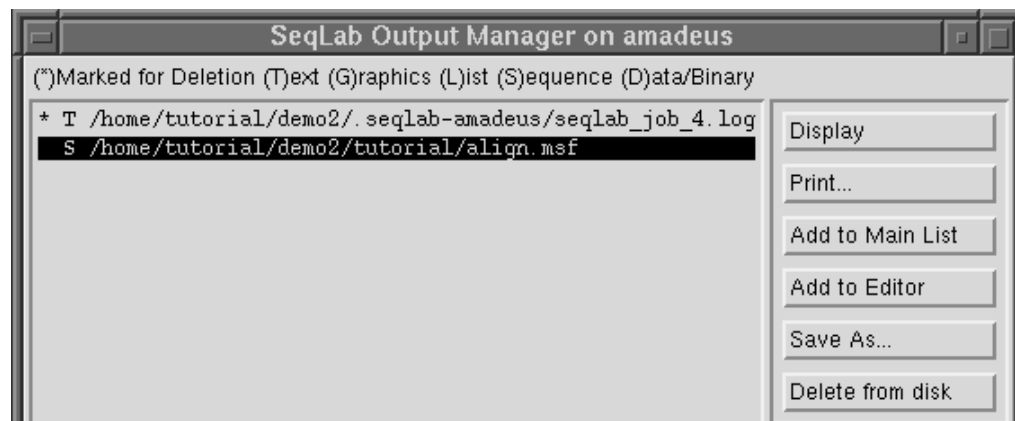
#### 5. Click on Options...

Take a moment to examine the ClustalW options available; we will use only the defaults for this tutorial. However, it is from here that one determines the speed and accuracy of the comparison (i.e. gap penalty).


#### 6. Click on Close to remove the Options window

#### 7. Click on Run.

Once the job is complete, three windows will appear; one is a log of the job, one is a display of the results and the other is the Output Manager. Close all but the Output Manager.



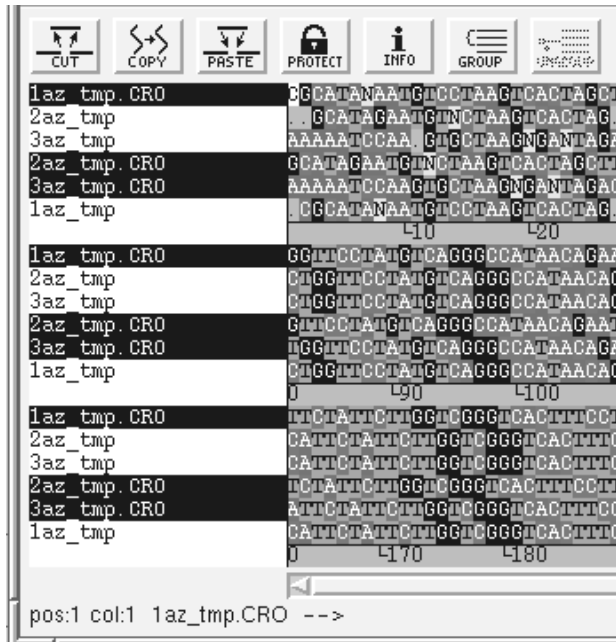
8. Select 'align.rsf' from the Output Manger and click on "Add to Editor"



If the aligned sequences have the same name as the original sequences, SeqLab will prompt you to overwrite it or load a second copy.

Grouping & Sorting

The aligned sequences may be mixed with the original sequences, let's organize them

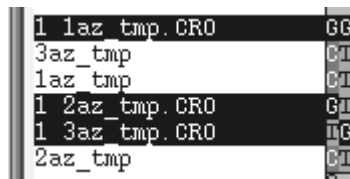


9. Hold-down the Control key and select all the sequences ending in .CRO by clicking on them.

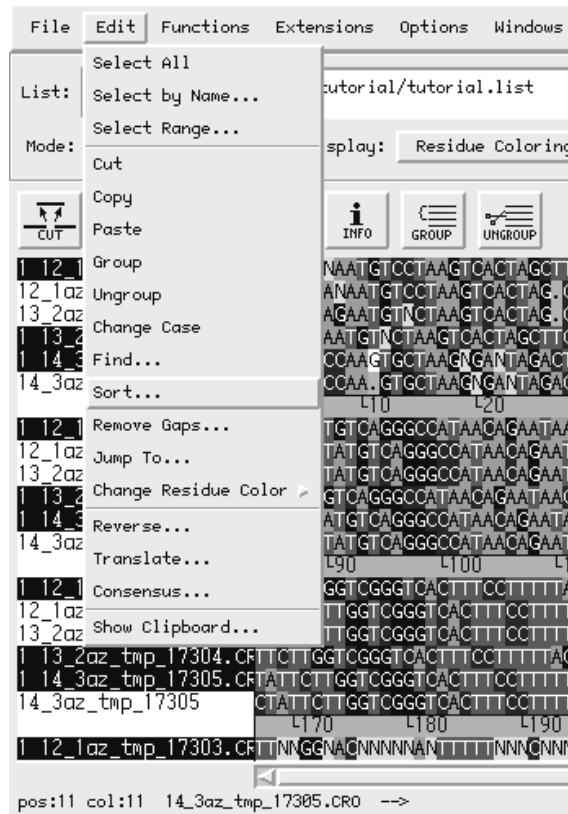


10. Release the Control key and click on the "Group" button.

You will see a "1" before the names signifying that they belong to group #1.

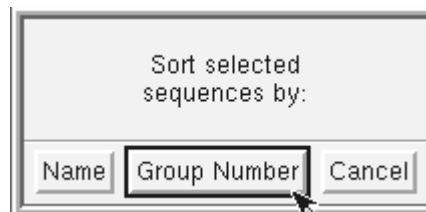


## 11. Click on “Edit” from the menu bar and select “Sort...”



The Sort dialog box appears.

## 12. Click on “Group Number”.



Now your sequences are nicely arranged!

# Database Search & Retrieval

## NetBlast & NetFetch

“NetBlast” is a direct interface to the NCBI web server. You can send multiple blast searches without waiting for prior searches to complete. The unique feature of NetBlast is that it uses “NetFetch” to retrieve the matching sequences directly into SeqLab, you don’t have to download each of them!

### 1. Select a region from a single sequence.

We will take approximately 200 bp from the last of our aligned sequences.

#### a) Select a sequence currently in the editor.



```
1 1az_tmp.CRO      GGTTCC
1 2az_tmp.CRO      GTTCCG
1 3az_tmp.CRO      TGGTTC
3az_tmp           CTGGTT
1az_tmp           CTGGTT
2az_tmp           CTGGTT
```

#### b) Click on Edit from the menu bar and select “Select Range...”

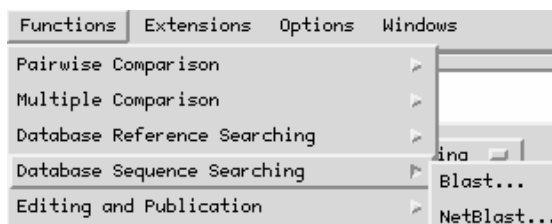
The Select Range dialog box appears.



#### c) Enter 200 in the “End:” text box and click on “Select” then “Close”.

The selected range will turn white.

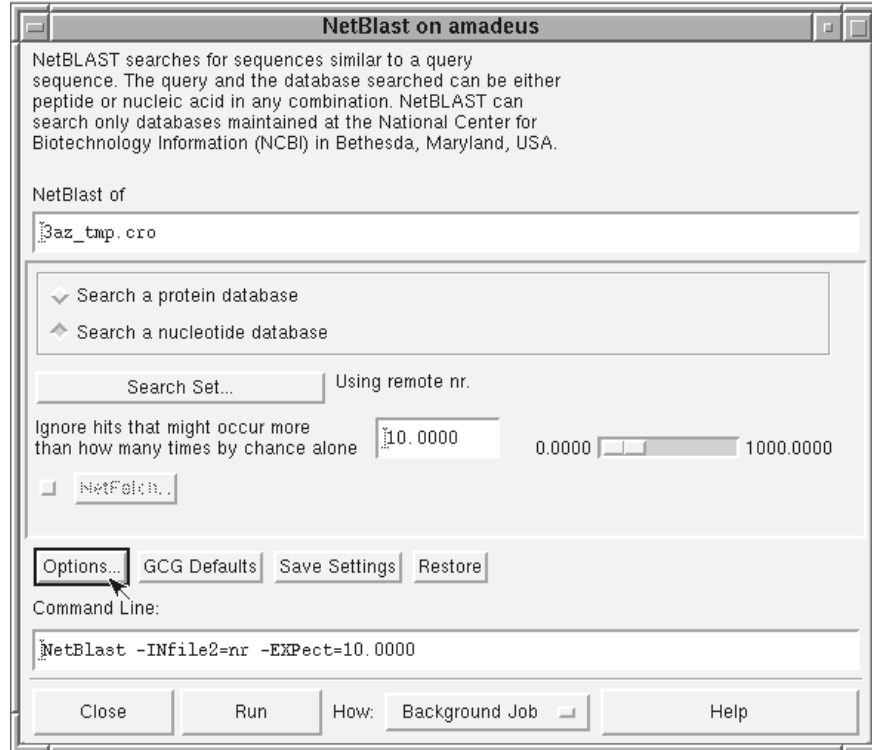
### 2. Click on “Functions” from the menu bar and select “Database Sequence Searching” → “NetBlast...”



The “Which Selection” dialog box appears.

**3. Click on “Selected Region”.**

The NetBlast window appears.

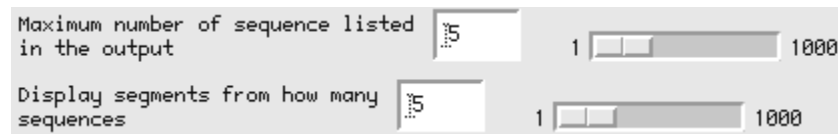


**4. Click on “Options...” in the NetBlast window.**

The “NetBlast Options” window appears.

You will see 2 slider bars with default values of 250 and 100.

**5. Click on the text box next to the slider bars and change the values to 5.**



Since we will be doing a NetFetch, we don’t want to get 250 sequences back!

**This is a very important point! 250 sequences can result in over 9 Megabytes of Data!**



**6. Click on “Close”.**

The Options window disappears and you are left with the NetBlast window.

**7. Click on “Run”.**

The NetBlast search has begun, you can see the job progress in the Job Manager window.

**9. Click on “Windows” from the menu bar and select “Job Manager”.**

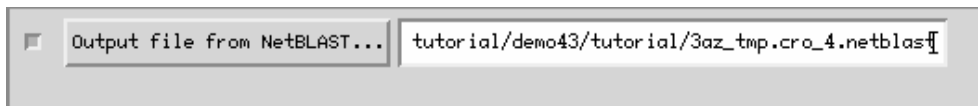
Once the NetBlast completes, 3 windows will be displayed. Job Manager, Output Manager, and the NetBlast job report, .

**10. Click on “Close” for all windows (except the editor).**

Now we will use the NetBlast report to start a NetFetch.

**11. Click on “Functions” from the menu bar and select “Database Sequence Searching” → “NetFetch...”**

The NetFetch dialog box will appear.



**12. Click on the ‘Output file from NetBlast button and select the NetBlast report from your tutorial folder.**

**13. Click on Run.**

When the job finishes the output manger will appear along with the NetFetch report.

**14. Close the report window.**



**It’s a good idea to always check the Job Manager to see if there are errors running the job. You don’t want to be waiting for a job that never got processed.**

## Viewing the New Sequences

1. **From the Output Manager, select the .rsf file.**

The file name is given by the Output Manager and depends on previous outputs from the current session.  
Your NetBlast output will have the same number in its file name.

2. **Click on “Add to Editor”**

The new sequences will appear in the Editor window.

3. **Click on “Close” to remove the Output Manager.**

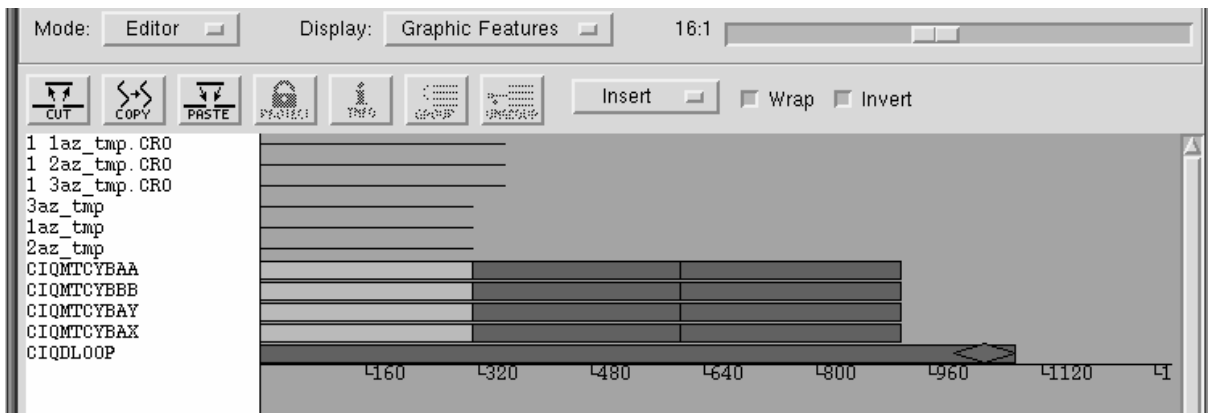
4. **Click on “Display:” (in the Editor) and select “Graphic Features”.**



Your newly added sequences become bars of different colors. The original sequences have no features defined, you will see only a line indicating length.

5. **“Drag” the zoom slider to the right until the entire length of the sequences is revealed (Approx. 16:1)**

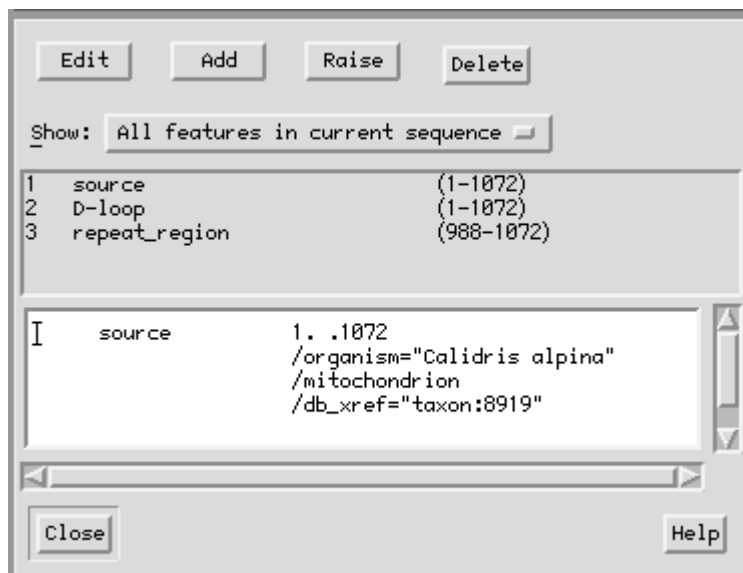
You will see geometric shapes formed at those sites where defined features are located.



**6. Place the Editing cursor on the first of the new sequences (CIQDLOOP).**

Click anywhere on the sequence

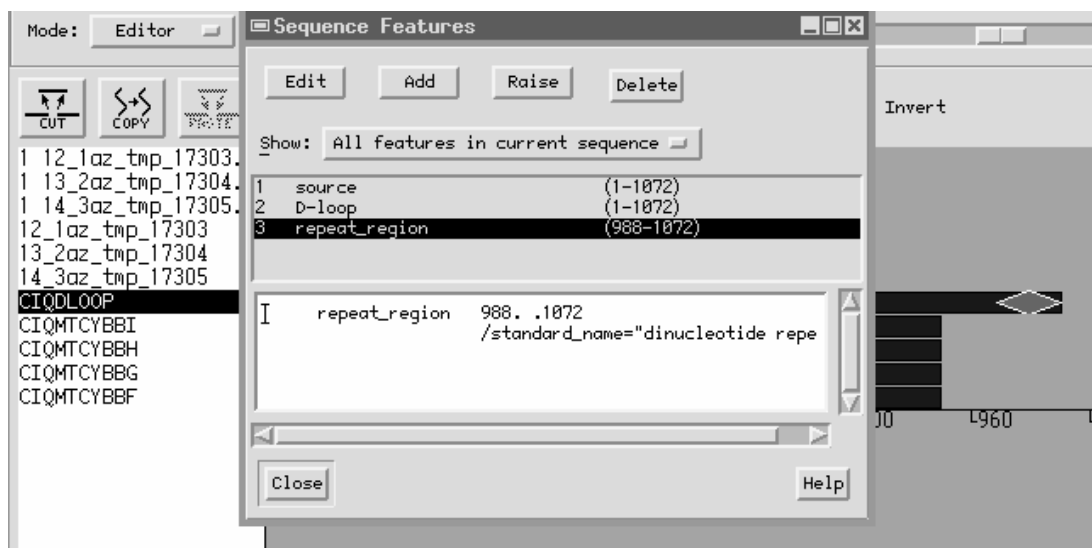
**7. Click on “Windows” from the menu bar and select “Features”.**



The SequenceFeatures window appears.

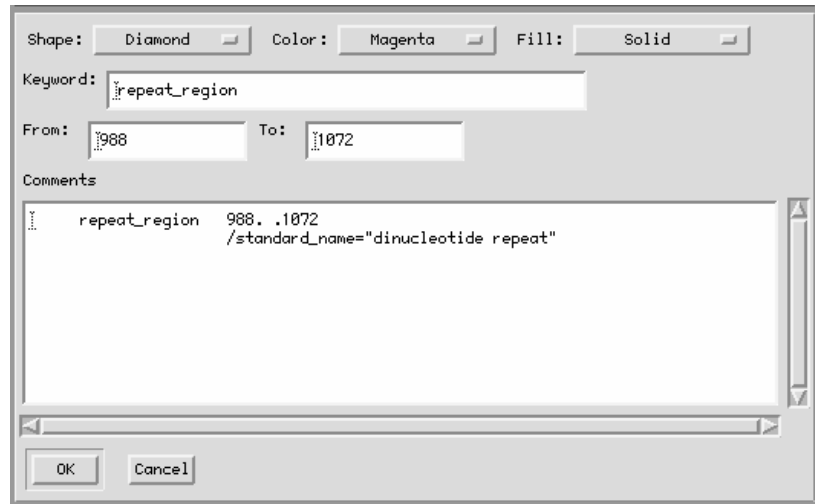
**8. Make sure that “Show:” is set to “All features...”**

You can click on the feature number in the “Sequence Features” window and the corresponding region will be highlighted in the editor.



In this case, there are two features found on this sequence; the D-loop and the repeat\_region.

### 9. Select “Edit” to view/change the Features characteristics



From here you can change the shape, color, fill, and location of the features.

You can also use this to create and/or annotate features of your own sequences.

### *Closing Up*



### 10. Click on the Mode: button to switch to Main List.

You will be prompted to save your sequences, this includes all the sequences visible in the editor, output from NetBlast and other GCG programs is saved by default.

### 11. Click on “File” from the menu bar, and select “Save List”.

### 12. Click on “File” from the menu bar and select “Exit”.

### 13. Close your BioDesk viewer, unless you are using one of our demo accounts, then you should use the EXIT button to kill the session.

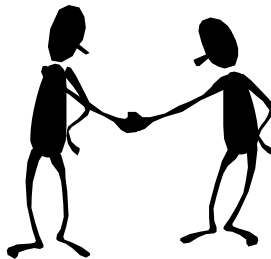
# Summary

---



## In this tutorial you learned how to:

- ❖ Download Electropherograms from the DNA Sequencing Facility.
- ❖ View Electropherograms in SeqLab and extract the sequences.
- ❖ Remove unwanted bases at the ends of sequences.
- ❖ Do a multiple alignment using ClustalW.
- ❖ Group & Sort sequences.
- ❖ Use NetBlast & NetFetch to search databases.
- ❖ View/Edit Features in sequences.



# Introduction to GCG/SeqLab

## Part B

### Feedback

Date \_\_\_\_\_

**How would you rate this workshop in the following areas (1=poor, 10=excellent)?**

Quality of Content \_\_\_\_\_

Handout \_\_\_\_\_

Information Covered \_\_\_\_\_

Presentation \_\_\_\_\_

Overall Evaluation \_\_\_\_\_

Would you recommend this workshop to your students, colleagues, etc... \_\_\_\_\_

What additional areas in bioinformatics/computing would you like training/workshops in?

Any other comments?

Thank you for your Feedback!