# Project DNA1: Project DNA1 Assembly and Analysis of a cloned DNA fragment using Staden Software

Objective: To learn to use PC‑based, DNA analysis software to:

(The project can be one, several or all of these objectives)

1. Assemble DNA traces to form one contiguous DNA sequence.
2. Search this sequence for restriction sites.
3. Search this sequence for possible genes (open reading sequences)
4. Produce a formatted DNA sequence, showing RE sites and other features for display or printing
5. Use BLAST to search Genbank for similar sequences
6. Download sequence data from public databases
7. Align and compare DNA sequences
8. Use phylogenetic software to infer relationships among sequences
9. Display the analyzed data as diagrams in Microsoft Word, Powerpoint, pDRAW or other presentation programs

This project uses the H600 dataset as an example. Any of the B393, H600 or H637 datasets can be used in this project.

Storage of Data: Although in a real project some data will be copied and placed in the lab book, the large volume of paper generated by DNA analysis is usually stored in a binder or archived as digital data. In our lab all files are retained on the computer and archived onto CD-R (now DVD-R).

Please save data in directories organized by project. There will be no hard copy files to be saved for this class but you may print them out for yourself if you wish. Instead of printing the files onto paper, you will create Adobe Acrobat files. Other specific computer files will need to be saved as well. The required data is listed at the end of each project.

A tip on saving DNA and Protein sequences: I save individual DNA sequences as GenBank (\*.gb) format, DNA assemblies are Staden databases (\*.aux) and all alignment project files as multiple sequence alignment (\*.msf) format. That way you can tell them apart on the computer. For instance, in this project I would have GenBank files H600‑T3.gb and H600‑T7.gb each containing an individual file and H600.0.aux containing the Staden DNA assembly. H600 con.gb or H600con.fa would be the consensus sequence.

**The pH600 Fragment**

The pH600 clone is a small, HindIII fragment of viral DNA (BaCMV OCOM4-37) cloned into the vector pBluescript II SK(+). The sequencing facility has sent two sequences to you. One sequence was obtained using the T3 primer and the other sequence was generated using the T7 primer.

Objective: To derive the complete sequence of both strands of the fragment, construct a restriction map identifying the sites of all the unique enzymes in the fragment, search GenBank for similar sequences, download any 4 of these sequences to BioEdit.

Procedure: There are several different ways to accomplish these tasks. We will assemble the DNA using manual methods and GAP4 program of Staden. There is a 72 page mini-manual for Staden. The complete, >650 page Staden manual is in the Staden Package folder as a Adobe Acrobat (\*.pdf) file. The consensus sequence will be annoted in Spin and sequences we download from GenBank will be entered into Clustal for alignment.

The sequence and trace files needed for your assignments can be located from the Datasets Page. You will see a H600.zip file containg two trace files and a text file containing the plasmid vector sequence. .

# Opening Trace files using Trev (For more info, see page 53 of the Staden Mini-Manual)

1. Using Staden Trev to open the 600-T3.abi trace file. You will have to click File > Open then select Files of Type as Staden considers \*.ab1 to be trace files not \*.abi. A viewable chromatogram with uneditable (above) and editable text (below) will appear.

*If the lower editable text is not visible, click View>Display edits*

1. You can clip away the undesired MCS sequence at the 5’ end of the sequence by clicking Edit > Left Quality then scrolling to the end of the DNA that you want to clip off and clicking

*For this sequence it is nt 85 just before the HindIII site (AAGCTT)*

*You are actually clipping away vector but use Left Quality since the program will export the sequence data without clipping out the Vector*

1. Similarly you can clip away the undesired MCS sequence (or low quality trace data) at the 3’ end of the sequence by clicking Edit > Right Quality then scrolling to point in the DNA that you want the clip to start and clicking

*Cut off the data after the HindIII site (AAGCTT) at ~270*

1. To edit or make changes in the editable sequence you can also click Edit > sequence and a cursor will appear in the editable DNA sequence. You can type in bases and to remove bases use delete as backspace (This is a unix program ported to windows so some key commands are unusual).
2. To save the sequence in an editable file click File > Save as > Plain text

*-You will have to type the .txt extension on the file so programs will recognize it as a text file.*

-*almost all DNA analysis programs will accept DNA or protein sequences as plain text*

1. You can search for any sequence (e.g. RE sites) by clicking View > Search and entering the sequence.

*Try searching for* AAGCT*, the HindIII site*

1. Save the edited sequence as 600-T3.txt
2. When you close the file do not save the trace! This will alter the raw data. If you wish to save the trace with the altered sequence use a name that indicates the sequence has been edited.

*The program no longer allows you to save (and change) the trace*

1. If you have a postscript capable printer you can print out a color picture of the trace by clicking File > Print > Page Options then selecting paper (A4 to US letter) and landscape versus portrait
2. Then Click File > Print > Trace Options to select the range of bases to print. Usually we only print the first 600 bp or so.
3. It is easier to print out a trace using FinchTV or ChromasLite.

Assembling DNA Traces to form a Consensus Sequence

(part 1 preGAP4)

1. Create a Staden H600 folder within the Project DNA1 folder

*-Staden programs create many databases and auxiliary files during assembly.*

1. Copy the two trace files into this folder
2. Start the program preGAP4
3. Click the Add files button and navigate to DNA Data folder > DNA Project 1 > Staden H600
4. In order to see the files you will need to change the Files of Type to Any
5. Click on the two \*.abi files to select them, then click open

-*the paths of the two files should now be visible in the List of Files to process*

1. Now click the configure Modules tab

-*modules are different programs preGAP can use to edit and evaluate DNA traces*

*-we will first perform the simplest possible manual assembly*

1. Select Modules > Deselect all modules
2. Check Estimate Base Accuracies and configure module by checking S/N original
3. Check Initialize Experiment Files and you do not have to configure this module
4. Check Interactive clipping and you do not have to configure this module
5. Check Gap4 Shotgun assembly and put H600 in the Gap4 database name and select Create new database.

-*this is where you can change the parameters for assembly. How much match is needed etc.*

1. Click run at the lower left of the screen

-*the program will now process our data using the modules in the order we selected*

1. Since we are Interactive Clipping (or manual editing) the program brings up the first file in the Trev program.

 -*remember, the sequencing primer binds within the MCS so 3'-end (the first part) of this sequence is the vector. You also must trim the 5'-end because the trace will degrade at some point and the DNA sequence will not be reliable. Since this is a small fragment, the sequencing trace spans the viral DNA fragment and continues through the other side of the MCS. This must be removed.*

 -*it is very helpful to view a map of your vector with the sequence of the MCS while editing*

1. Search for the HindIII sites by clicking View > Search and entering the HindIII recognition sequence AAGCTT then clicking Next

-*You should be at nucleotide (nt) 65 (if this is 600-T7)*

1. To remove or clip away the 5’ MCS sequence select Edit > Left Vector and click the sequence immediately to the left of the AAGCTT site (the background should turn pink)

-*you are not actually removing any data, the program will just “hide it”*

1. Click View > Search > next to go to the next HindIII site

-*You should be at nucleotide (nt) 257*

1. Now click Edit > Right Quality and click the sequence immediately to the right of the AAGCTT site.

*-This should be right vector but I wanted you to see the different colour*

1. Now you are finished so click on Next file and click yes to save your sequence.
2. The 600-T3 trace should appear in Trev. Now process it using the same procedure. When you are done, close the window by clicking the upper right x and click yes to save your sequence.
3. The main pregagp4 window should appear with the Textual Output displayed and the following text at the end

……

- Report from 'Gap4 shotgun assembly' -

SEQ D:/….your actual file location…/Staden H600/600-T3.exp: assembled

SEQ D:/ ….your actual file location…/DNA1 H600/Staden H600/600-T7.exp: assembled

 \*\*\* Processing finished \*\*\*

1. Your assembly is done. Close the Pregap4 and open GAP4 program to view the results and edit the consensus sequence

### Assembling DNA Traces to form a Consensus Sequence

Part 2 GAP4 (Genome Assembly program)

1. Open GAP4 and click File > open then navigate to the Staden H600 folder, select H660.aux and open
2. The main GAP4 Window and the Contig Selector Windows will open. Since the 2 traces assembled into one contig or contiguous sequence there isn’t much to see in the Contig Selector window (yet).

You should get the following in the logfile (or very similar)

Mon 24 Sep 14:46:45 2012: check database

------------------------------------------------------------

Database is logically consistent

Mon 24 Sep 14:46:45 2012: Database information

------------------------------------------------------------

Database size 8000 Max reading length 30000

No. Readings 2 No. Contigs 1

No. Annotations 2 No. Templates 2

No. Clones 1 No. Vectors 1

Total contig length 192 Average length 192.0

Total characters in readings 378

Average reading characters per consensus character 1.97

Average used length of reading 189.00

Current maximum consensus length is 100000

1. Go to the main GAP4 Window and click View > Template Display and OK
2. The two traces that make up the Contig are shown
3. Click the +50% to zoom in. The DNAfragment is 192 bp. Since there are only two traces there isn’t much information displayed. This window will tell you a lot about your assembly.

# Reading Colour Meaning

Red = Primer unknown, Cyan = Forwards primer, Orange = Reverse primer, Dark cyan = Custom forward primer, Dark orange = Custom reverse primer

# Template Colour Meaning

Blue Template contains only readings from one end

Pink Template contains both forward and reverse readings within this contig

Green Template contains both forward and reverse readings, but they are in separate contigs

Black Readings on the template are within the same contig but are in contradictory orientations or are an unexpected distance apart

Yellow Readings on the template are within different contigs and are consistent

Dark yellow Readings on the template are within different contigs and are inconsistent

1. Double click on the red lines to bring up the Contig Editor. This is where the consensus sequence is examined and edited
2. Double click on the T7 then the T3 sequence to open the Trace Display.
3. If you go back to the Template display window try clicking in different areas. The corresponding sequences are shown in the Contig Editor and Trace display.
4. There are no disagreements in this assembly (not the normal situation) so you cannot see that areas of conflict are indicated by colors (can be configured).
5. Save the consensus by clicking File > Save Consensus > normal then select output format as Fasta and name the file H600con.fa and click OK

-*Fasta format files are text files contain a title indicated by “> name” followed by a hard return then the body of the file is the DNA sequence in simple text format*

1. Close GAP4
2. Open the H600con.fa file in BioEdit and annote the sequence
3. Save it as a Genbank sequence.

GAP4 can be used to construct restriction maps and analyze the assembly but we will use these functions on a more complex assembly.

Creating a Diagram of plasmid pH600 using pDraw

This is a simple DNA mapping and display program but is simple to use. This program allows you to select specific enzymes. It also allows you to build plasmids by combining different DNA sequences. Finally, it will allow you to choose restriction enzymes, cut your construct and show what it would look like if electrophoresed on an agarose gel.

1. You can enter DNA sequences into pDRAW several ways
2. To paste data in, start Notepad, open the H600con.fa fasta file.
3. Select the entire DNA sequence and copy to the clipboard
4. Start pDraw and click File>New>enter new sequence

 -*this opens a text editing window*

1. Paste your DNA sequence from the clipboard into the text window then press OK.

-*the DNA sequence is mapped using the default settings which usually include the methylation sensitive residues and the RE recognition sequence*

*-If you open a sequence in pDRAW it is a very good idea to check the sequence by clicking Edit > Edit Sequence and inspecting the DNA. Sometimes the header is imported and you get bad data in the sequence*

1. Save your sequence in the pDraw format (\*.pdw).

*-Save and save often*

 -*if you wish, you may close BioEdit to increase system resources*

1. The default display setting is most of the RE’s and it also shows the RE recognition sites. A little too much clutter. Let’s fix it. Click Settings>plot> then check the explicitly selected enzymes only box and click apply.
2. To add RE Sites to the diagram, cClick Settings>plot then click the explicitly selected tab. Pick specific enzymes from the menu on the left. Note that the number of times each enzyme cuts the DNA is displayed. To select more than one enzyme hold down the Ctrl key. Click apply to show the desired sites. There aren’t many in the short 192 bp H600 sequence. (try *Hin*dIII)
3. This program will display a lot of information about a sequence. Click View>Restriction Analysis for the RE data as a text file. This can be selected and pasted into a word processor or just printed.Hit View>Sequence and the DNA sequence is displayed.

Searching GenBank

Once you have a DNA sequence some of the first questions you ask about it are “what does it code for, or what does it do?” Probably one of the first things a researcher will do is search GenBank and look for similar sequences whose functions may be characterized.

1. Start BioEdit and open the H600 file.

2. Double-click on the title to bring up the DNA editing window.

3. Copy your DNA sequence to the clipboard (in FASTA format) by Clicking Edit>Copy

4. Open Internet Explorer by double-clicking on the mol biology bookmarks.htm in your Utilities folder.

5. Go to “www.ncbi.nlm.nih.gov “.This will open the home page of the National Center for Biotechnology Information.

6. Open Blast and click on the BLAST (**B**asic **L**ocal **A**rea **S**earch **T**heorem) program.

 *-this program is on the right side of the screen, under Popular Resources*

 *Under Basic Blast is Nucleotide-nucleotide and there is BLAST (blastn)*

 *-this program allows you to compare you DNA sequences against all the sequences in the entire GenBank database*

 -*A second program (BLASTX) will translate you DNA sequence in the 3 forward and 3 complementary reading frames and compare these 6 possible proteins to all the protein sequences in GenBank*. *BLASTX is more likely to detect homology as considerable DNA variation can occur and thus reduce homology without changing the amino acid composition.*

7. Paste your sequence into the window.

 -*the sequence needs to be in FASTA format*

A sequence in FASTA format begins with a signal-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a great-than (>) symbol in the first column. It is recommended that all lines of text be shorter than 80 characters in length. An example sequence in FASTA format is:

>600‑T3

AAGCTTTGCCCGCTGCACCAAGACCGCTCATCAAGTCGTCCAGGCCCCTGAGATATGTAGGGACGGTGTCGAACACCTTTCCTTCCACGTGAACGACCCGTTGCTTGTAAGCGTTGAACTCTTGCATAATCTCTTCCAGATTGAAAACATTGCTGGCACGCAGTTCATCCTCGGAGTACAGTTCCAAAGC

8. Choose Search parameters

*-under choose search set pick others*

*-under program selection choose somewhat similar sequences*

9. Click search. A new screen will come up telling you when your results will be ready and formatted.

 -*you are searching the non-redundant database of DNA sequences at GenBank*

 -*NOTE if you scroll farther down the screen you will see an option to have the search results*

10. Click the large Blast button and wait...and wait...and wait...

 -*having the results emailed to me is my method of choice but that’s not allowed anymore*

1. Eventually you will get a screen with a graphical interface that looks somewhat like this. BLAST is always updating so it will not look exactly like this.



With a lot of textual information below (See next page)



This (gi|12620808|gb|AF324835.1|AF324835) should be the first sequence you get. The gb means this sequence was originally submitted to GenBank, emb would mean it was first in the European molecular biology database, dbj is a Japanese database. The big databases share sequences and update often.

AF324835 is the accession number. This is the identifying code for the sequence. Then comes the gene or DNA name. Finally is the score. If you used the T7 you probably got a score of 377, if you used the consensus you would get a score of 381. This is as good as it gets for DNA. It is identical.

12. Click on the accession number and the DNA sequence will be displayed in GenBank format.

13. Click on Send (near the top right) select the clipboard and click add to clipboard and the sequence is added to the clipboard. You can add many sequences to the clipboard. This is very useful when collecting sequences from GenBank to align and analyze.

 -*if you machine has been set to refuse cookies the clipboard will not work.*

 *-use send to file instead. I have been having problems with OSU security myself*

14. You can save individual sequences by clicking Send>File>create file

You can select complete record and the type of format. It will ask you if you want to open it or save it to file. t **You must save it.**

If you choose Save, it will open the Save As dialog box. Pick a folder, give the sequence a name that makes sense and save it. I always put the accession number in the sequence name.

15. Open the sequence in a text editor such as Notepad. Note all the annotation.

16. Go back to the search results and save the first 4 sequences to a file giving it the name Project H600 DNA homology.

 -*the H600 fragment is part of the BaCMV gB gene so the RhCMV homology is to be expected*

17. To save the BLAST search results return to the search results page and select all text (Control A).

18. Open a word processing program (Microsoft Word or Wordperfect)

19. Click Paste>Special and select unformated text.

 -*Change the left and right margins to 0.25 and the font to Courier New, size 10. You will also need to go to Home>Paragraph> and click don’t add space between paragraphs of the same style and the results will align perfectly. Scroll down a few pages to see the actual alignments.*

20. Print the first 3 pages of the actual alignments as an acrobat file

21. Save the BLAST search.

22. Now repeat the search but on the first page select Program>BLASTX instead of the default BLASTN.

23. When you rerun the search notice how many more significant homology sequences are recovered.

24. Repeat steps 12-24 with the BLASTX results.

Data to be produced:

Computer files Acrobat documents

H600-T3.abi H600-T3 DNA Sequence

H600-T7.abi H600-T3 Trace file (chromatogram)

H600-T3.gb H600-T7 DNA Sequence

H600-T7.gb H600-T7 Trace file (chromatogram)

H600 consensus.gb H600 consensus DNA Sequence

H600 BLASTN search (2-3 pages) H600 Restriction map (any format)

H600 BLASTX search (2-3 pages)

H600.aux GAP4 database

H600 BLASTN search

H600 BLASTX search