DOGMA

Tutorial: chloroplast genomes

Stacia Wyman
Department of Computer Sciences
Williams College
Williamstown, MA 01267

March 10, 2005

ASSUMPTIONS:

- You are using Internet Explorer under OS X on the Mac.
- You have set the font size in the Internet Explorer preferences to 14.
- The resolution of your monitor is set to at least 1152 x 768.

This tutorial will walk you through some of the steps for annotating annotating a chloroplast genome. It will describe all of the basic features of the software as well as preparing files for input to `DOGMA`. Please email staciacs.williams.edu with any comments or questions.

1 Preparing the input file

The first thing you should do is prepare the input file. For this tutorial, we will annotate the *Nicotiana* chloroplast genome. If you choose to do the tutorial on your own sequence, directions for preparing the input file are in section 1.2. The directions for downloading *Nicotiana* for the tutorial follow.

1.1 Downloading the *Nicotiana tabacum* chloroplast genome from NCBI

1. Using your browser, go to: `http://www.ncbi.nlm.nih.gov:80/entrez/viewer.fcgi?val=NC_001879`

2. Change the second button from the left in the top button bar to “FASTA.”
3. Hit the “Display” button.
4. Click on the “Send to” button.
5. When the window appears saying the browser doesn’t know what to do with the file, hit “Save File As...”.
6. Save the file as “nicotiana.fa” (*not* as batchseq.cgi) on your desktop.

The file nicotiana.fa will be in the proper format for uploading.

1.2 Preparing your own data file for input

The input file should be in FASTA format. The genome should be in one contig, containing only the nucleotides `A,C,G,T`, in uppercase. It should be a plain text document. To prepare your file for uploading to `DOGMA`, several steps may be required depending on what program you saved your file from.

- **From `Sequencher`:**
  - Save the file as Pearson/FASTA format from the export menu.
  - Then you have two options:
    1. **Using a terminal window:**
       - Open a terminal window. (The terminal app is in the Utilities folder under applications in OS X on the Mac.)
       - `cd` to the directory with the file in it (if the file is on the Desktop, open the terminal window and then type `"cd Desktop"` at the prompt).
       - At the prompt, type:

         $$native2ascii < filename >< filename >$$

       - Then the file should be ready for uploading to `DOGMA`.
    2. **Using `Microsoft Word`:**
       - Open the file in `Microsoft Word.`
       - Save the document as MS-DOS Text. **NOT TEXT ONLY**
- **From `Microsoft Word`:**
  - Save the file as MS-DOS text (as in number 2 above).
- **Other:**
  - If you aren’t sure what program the file came from and it isn’t uploading correctly to `DOGMA`, try running `native2ascii` on it (as in number 1 above, doing this to a file will do no harm if it’s already in ASCII format). The symptom of not uploading correctly is that `DOGMA` finds no genes in input sequence.

2 Getting a DOGMA Userid

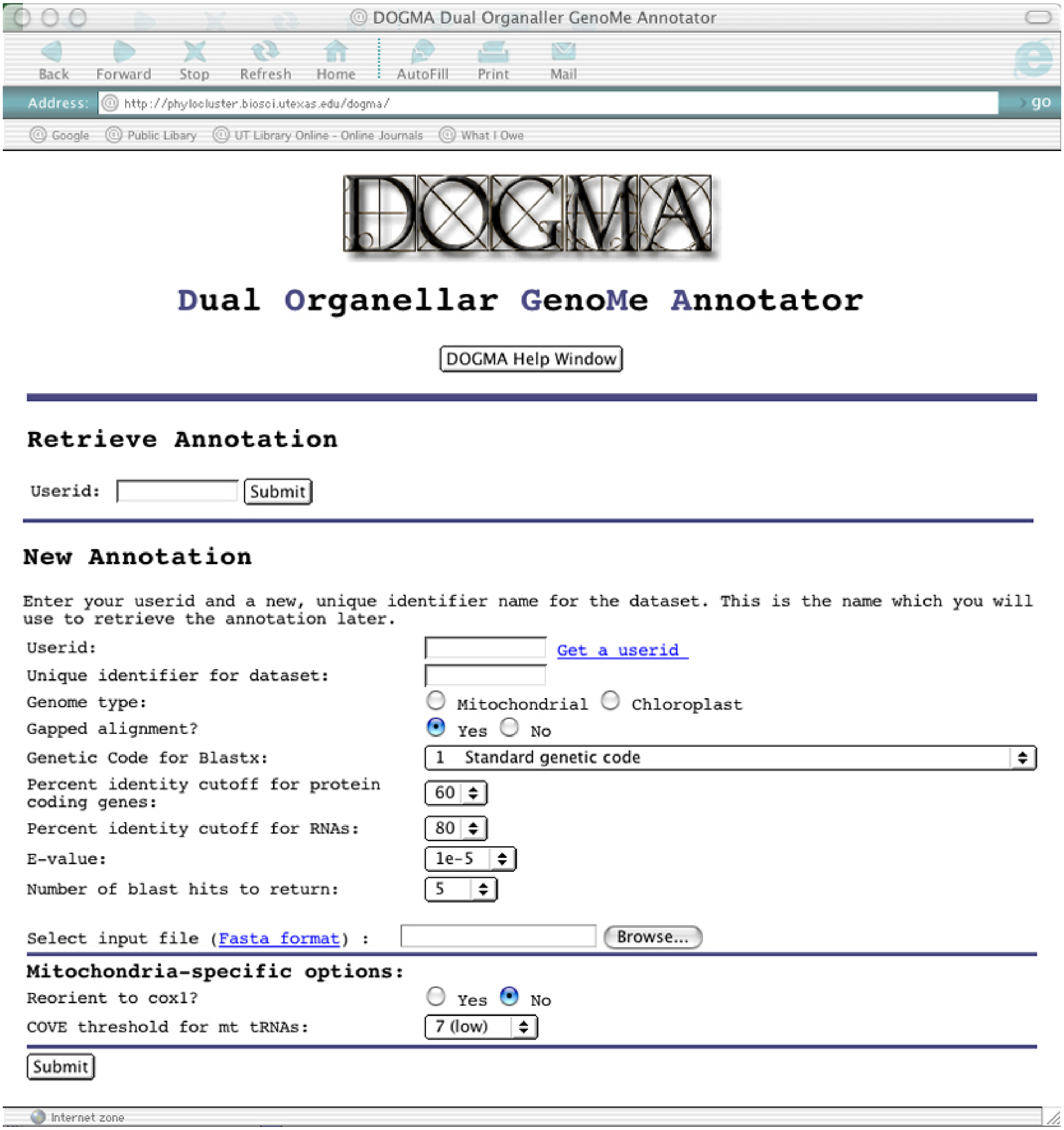
The DOGMA website is password protected. You can create your own userid and password for DOGMA using a link off the DOGMA home page. You will use the new userid from that point on to log on to the DOMGA website. The userid/password is to keep your data private. Other users of the site (and the rest of the world) cannot see your data unless you give them your userid and password. This is the userid you will use to retrieve annotations later.

1. Go to `http://bugmaster.jgi-psf.org/dogma`
2. For now, ignore everything except click on the "Get a userid" link to the right of the Userid input box about half way down the page (see Figure 1).
3. Create a userid. DOGMA is case-sensitive, and userids should not have any spaces or punctuation in them.
4. Create a password. DOGMA is case-sensitive, and password should not have any spaces or punctuation in them. This site does not use encryption so do not use an important password.
5. Enter your email address. This is so I can track users and contact you if there's problems with the website or your data.
6. Hit submit, then, if the creation was successful, go back to the main form page.

3 Filling out the form

Details on the different items in the DOGMA form can be found in the Help Window. The main form for DOGMA is shown in Figure 1.

- Once you've created a userid, put that userid in the userid field.
- Then fill in "nicotiana" for the unique id. This is how you'll identify the annotation for retrieval later.
- Choose chloroplast for the genome type.
- Initially, choose a gapped alignment (for cp genomes, you should run it both ways).
- Leave the default for the genetic code. **Note:** for new annotations, you can choose Genetic Code 11, which has is identical to the Standard Genetic Code, but has additional start codons.
- For protein coding genes, 60 percent is a good cutoff for *Nicotiana*. This may need to be changed for genomes which are divergent from existing genomes in the DOGMA database (DOG-base). For genomes not too diverged from other in the database, you can set this value to a higher number.



http://phylocluster.biosci.utexas.edu/dogma/

Dual Organellar GenoMe Annotator

Retrieve Annotation

Userid: [ ] Submit

New Annotation

Enter your userid and a new, unique identifier name for the dataset. This is the name which you will use to retrieve the annotation later.

Userid:	[ ] Get a userid
Unique identifier for dataset:	[ ]
Genome type:	○ Mitochondrial ○ Chloroplast
Gapped alignment?	● Yes ○ No
Genetic Code for Blastx:	1 Standard genetic code
Percent identity cutoff for protein coding genes:	60
Percent identity cutoff for RNAs:	80
E-value:	1e-5
Number of blast hits to return:	5

Select input file (Fasta format) : [ ] Browse...

Mitochondria-specific options:

Reorient to cox1?	○ Yes ● No
COVE threshold for mt tRNAs:	7 (low)

Submit

Figure 1: The DOGMA form.

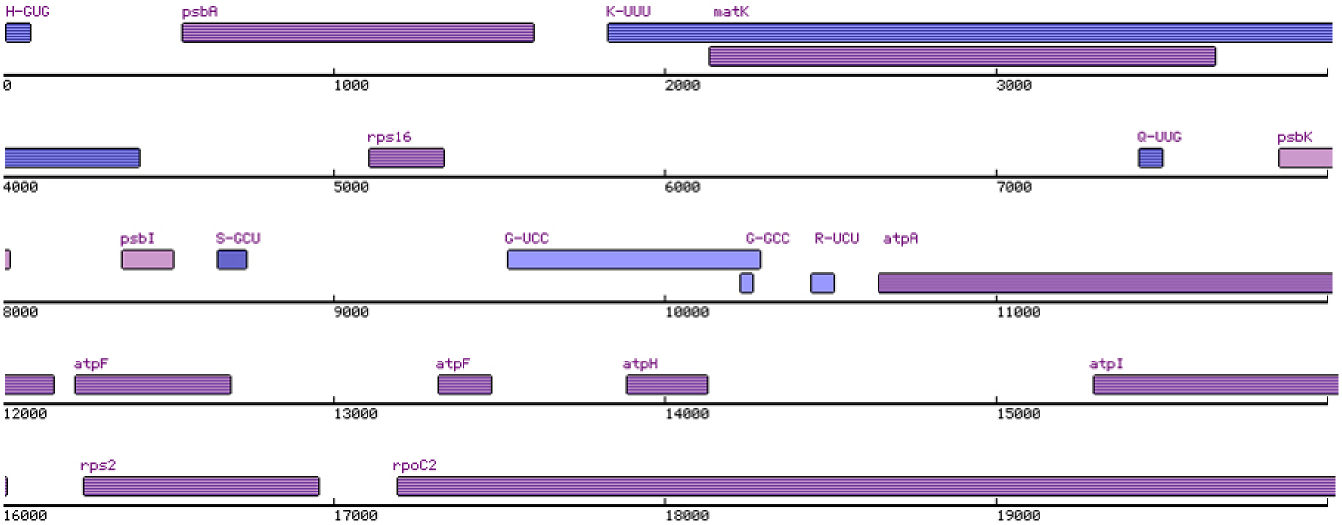


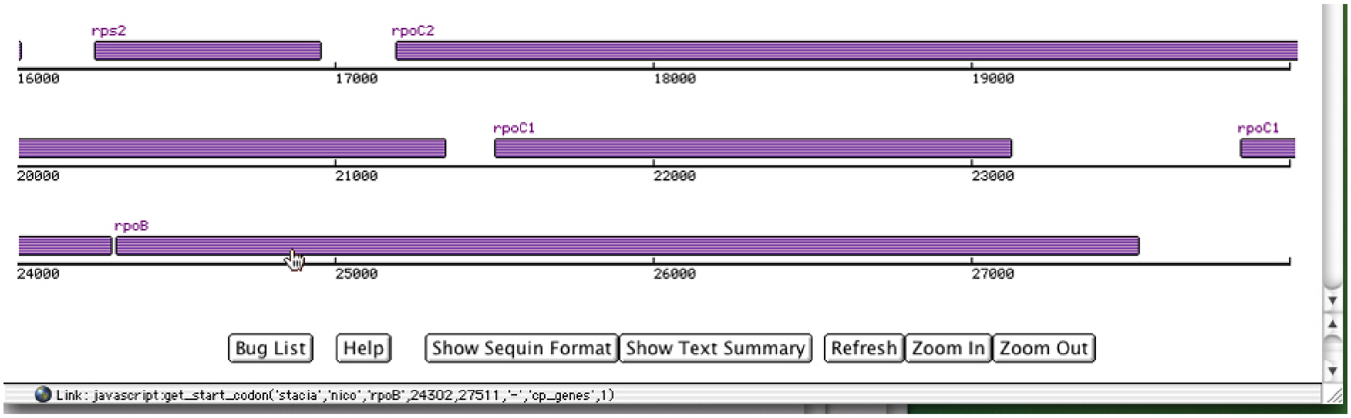
Figure 2: The Number Line Panel for *Nicotiana* from 0 to 20,000.

- For RNAs in *Nicotiana*, the percent identity cutoff should be high to eliminate spurious hits. Again, for more diverged genomes, this may need to be set lower, but for *Nicotiana*, set it to 95%.
- It's unclear if e-value is an appropriate way to judge the quality of a BLAST hit in DOGMA. For *Nicotiana*, leave it at 1e-5, a less stringent cutoff.
- The number of BLAST hits returned determines how many BLAST hits are showing the annotation. For this tutorial, leave it at 5. (This is a good size with respect to the size of the Blast Hit Panel.)
- Click browse and find the "nicotiana.fa" file on your desktop.
- Click the "Submit" button.

DOGMA takes about 10 minutes to annotate a chloroplast genome. Before the annotation returns, you'll get a list of genes that it is working on and then you'll be redirected to the main annotation page for *Nicotiana*.

4 The Number Line Panel

The middle panel (of three) of the DOGMA Annotation Window contains the Number Line Panel with the genes laid out on a number line (Figure 4). For chloroplast genes, it initially shows four 1,000 nucleotide intervals per row.



```
Link: javascript:get_start_codon('stacia','nico','rpoB',24302,27511,'-','cp_genes',1)
```

Figure 3: Mousing over the gene block displays that gene's information in the bottom of the browser window.

You can scroll down to see all of the genome. The protein coding genes are purple, the tRNAs are blue and the rRNAs are green. The genes on the forward strand are light-colored and genes on the reverse strand are dark-colored (striped, actually). If you can't see the whole number line, you'll need to widen your browser window.

To find a specific gene in an unknown location, you'll need to look it up in the Text Summary. Say we're looking for the

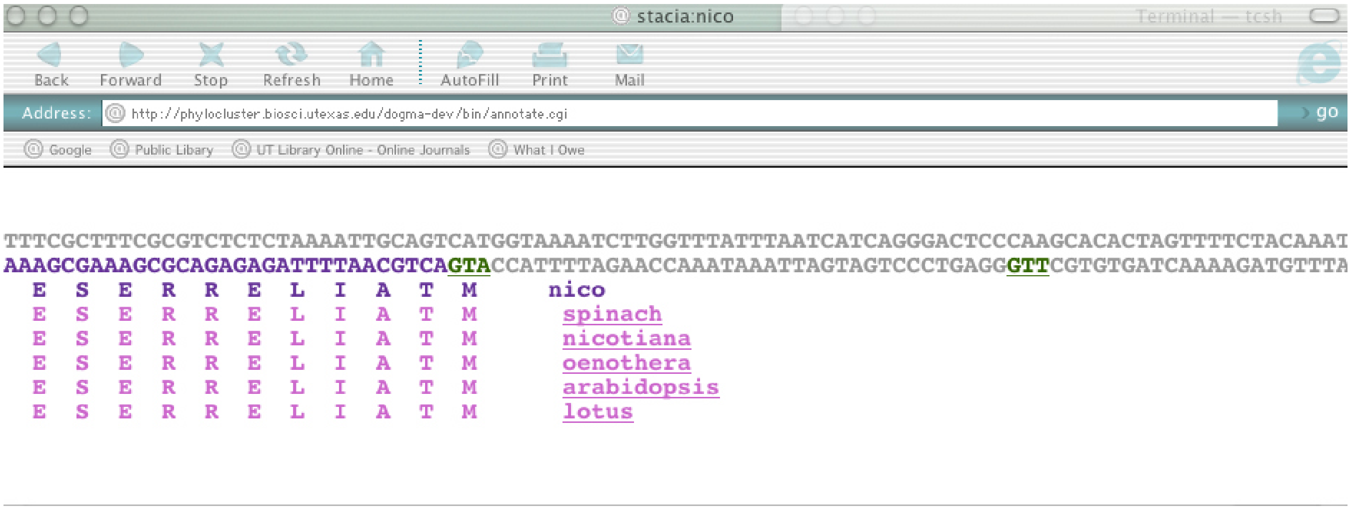
If you mouse over the colored block for a gene, information about that gene will appear in the bottom of the browser window. In Figure 4, the mouse (shown as a hand) is shown over gene `rpaoB`, and the information for that gene is shown in the bottom of the window. It lists: your userid, the unique id of this annotation, the name, the start, the end, and the strand of this gene.

This is convenient if the gene names overlap, you can mouse over each colored gene block to see what the genes are.

5 Bottom Panel

Selecting buttons in the bottom panel will perform various tasks.

- **Bug List** The bug list is list of things I need to do, am working on, or bugs that need to be fixed. The completed items with date that it was completed are listed at the bottom.
- **Help** This button opend the Help Window (if it's not already open).
- **Show Sequin Format** After each gene is annotated, an entry for it will be stored in the Sequin file. Click on "Show Sequin Format" will display the contents of this file. When preparing to submit a genome, this file can be saved for input to Sequin. The data in this



Address: http://phylocluster.biosci.utexas.edu/dogma-dev/bin/annotate.cgi

```
TTTCGCTTTCGCGTCTCTCTAAAATTGCAGTCATGGTAAAATCTTGGTTTATTTAATCATCAGGGACTCCCAAGCACACTAGTTTTCTACAAAT
AAAGCGAAAGCGCAGAGAGATTTTAACGTCAGTACCATTTTAGAACCAAATAAATTAGTAGTCCCTGAGGGTTCGTGTGATCAAAAGATGTTTA
  E  S  E  R  R  E  L  I  A  T  M     nico
  E  S  E  R  R  E  L  I  A  T  M      spinach
  E  S  E  R  R  E  L  I  A  T  M      nicotiana
  E  S  E  R  R  E  L  I  A  T  M      oenothera
  E  S  E  R  R  E  L  I  A  T  M      arabidopsis
  E  S  E  R  R  E  L  I  A  T  M      lotus
```

Figure 4: top panel.jpg.

file is tab-delimited and the format must be preserved for Sequin to read it, so it cannot be cut-and-pasted into a document.

- **Show Text Summary** Clicking on the "Show Text Summary" button will bring you to a summary of all the initial BLAST hits for all the genes. Because the genes have not been annotated, they do not include stop codons and may have incorrect start codons.
- **Refresh** This restores the Number Line Panel to its original proportions after zooming in or out.
- **Zooming** You can zoom in or zoom out of the Number Line Panel with these buttons.

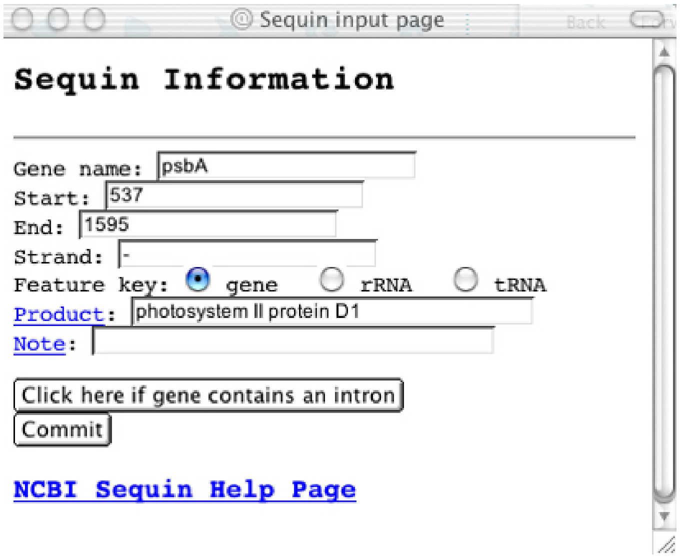
6 The Blast Hit Panel

6.1 A gene on the reverse strand: psbA

Click on the colored block for the psbA gene. In the top panel will appear the nucleotide sequence of both strands of the genome as well as the gene and it's 5 closest BLAST hits (Figure 6.1).

Additionally, the Sequin Window (Figure 6.1) opened and went behind this browser window. The initial values for the start and end of the gene are filled out in the Sequin Window.

As illustrated in Figure 6.1, the nucleotide sequence of the forward strand is on top with the nucleotide sequence of the reverse strand below. For a gene, the nucleotide sequence of the whole



Sequin Information

Gene name: psbA
Start: 537
End: 1595
Strand: -
Feature key: ● gene ○ rRNA ○ tRNA
Product: photosystem II protein D1
Note:

Figure 5: The Sequin Window pops up when you click on a gene in the Number Line Panel. Selecting stop and start codons in the Blast Hit Panel updates the values in the Sequin Window.

gene is displayed, plus the 60 upstream and downstream nucleotides of the gene. Because `psbA` is a protein coding gene, the nucleotide sequence is displayed purple.

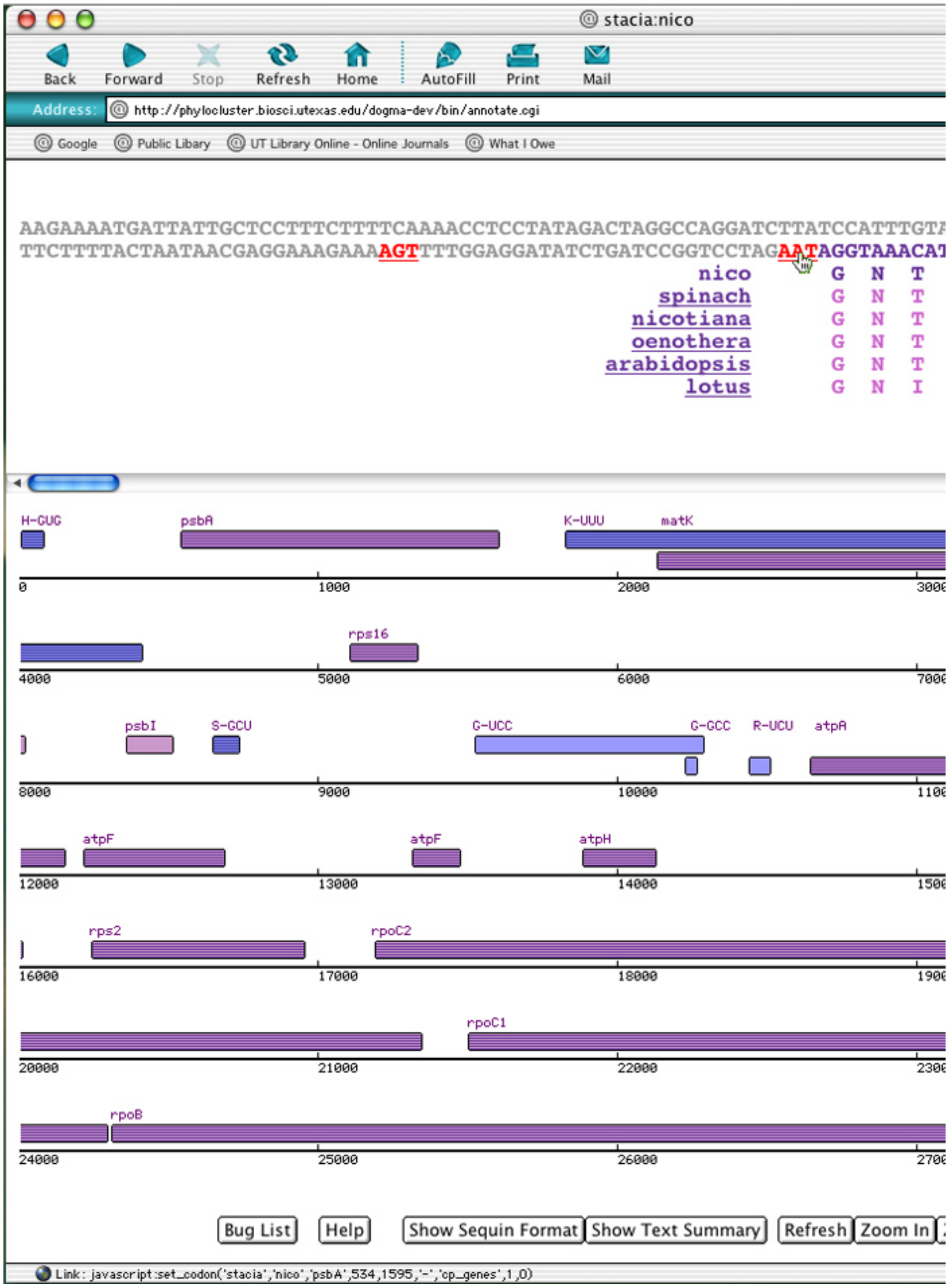
The previous and next genes for `psbA` are not within 60 nucleotides of either end of the gene, so they are not displayed and the nucleotides are displayed grey. The amino acid sequences from the `BLAST` hits are displayed below the nucleotide sequence (for genes on the reverse strand).

If you click on the gene name at the top of the panel, the `BLAST` results for that gene appear in a separate window (Figure 7).

If you click on the taxon name "spinach" next to the amino acid sequence, the database entry for `psbA` in spinach will appear in a separate window (Figure 6.1).

The potential stop codons are displayed as red links (see Figure 6.1). For `psbA`, there is only one potential stop codon in frame with the gene. Bring the Sequin Window into view next to or just behind the Annotation Window. Click on the red nucleotides `AAT`, and you will see the start of the gene for `psbA` is changed to include the stop codon (it changes the start for genes on the reverse strand).

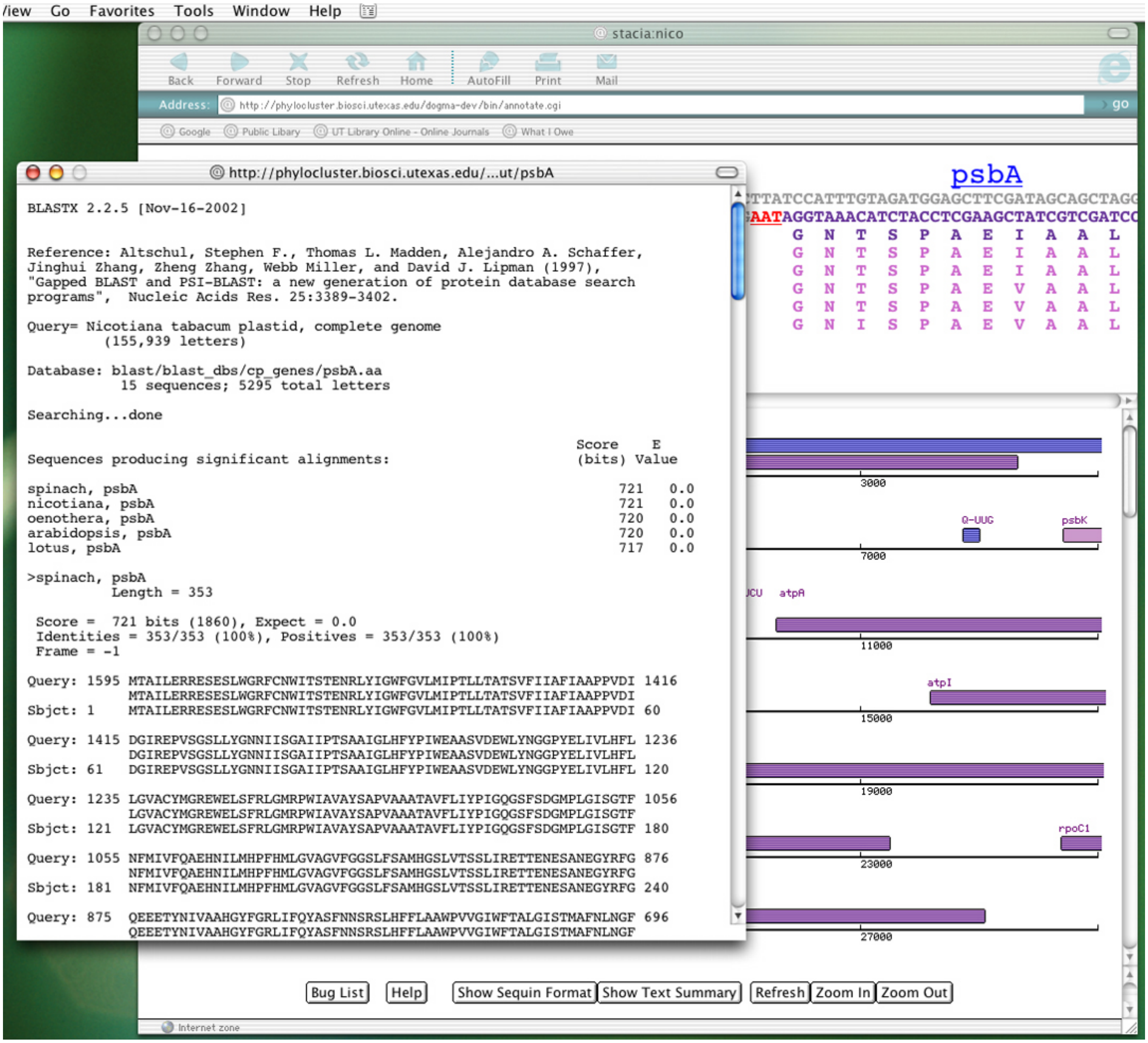
If you scroll the top panel all the way to the right, you will see the potential in-frame start codons for `psbA` highlighted as green links in the downstream sequence (because `psbA` is on the reverse



```
AAGAAAATGATTATTGCTCCTTTCTTTTCAAAACCTCCTATAGACTAGGCCAGGATCTTATCCATTTGTA
TTCTTTTACTAATAACGAGGAAAGAAAAGTTTTGGAGGATATCTGATCCGGTCCTAGAATAGGTAAACAT
                                                 nico     G  N  T
                                              spinach     G  N  T
                                            nicotiana     G  N  T
                                            oenothera     G  N  T
                                          arabidopsis     G  N  T
                                                lotus     G  N  I
```

Link: javascript:set_codon('stacia','nico','psbA',534,1595,'-','cp_genes',1,0)

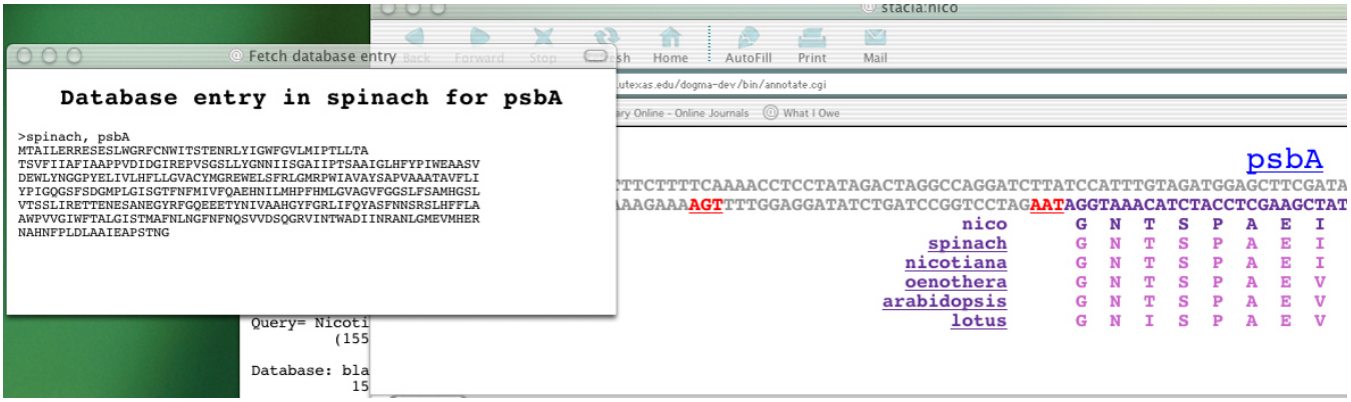
Figure 6: stop codon.jpg.



View Go Favorites Tools Window Help
stacia:nico
Back Forward Stop Refresh Home AutoFill Print Mail
Address: http://phylocluster.biosci.utexas.edu/dogma-dev/bin/annotate.cgi
go
Google Public Libary UT Library Online - Online Journals What I Owe
http://phylocluster.biosci.utexas.edu/...ut/psbA
psbA
TTATCCATTTGTAGATGGAGCTTCGATAGCAGCTAGG
AATAGGTAAACATCTACCTCGAAGCTATCGTCGATCC
G N T S P A E I A A L
G N T S P A E I A A L
G N T S P A E I A A L
G N T S P A E V A A L
G N T S P A E V A A L
G N I S P A E V A A L
3000
Q-UUG
psbK
7000
JCU
atpA
11000
atpI
15000
19000
rpoC1
23000
27000
Bug List
Help
Show Sequin Format
Show Text Summary
Refresh
Zoom In
Zoom Out
Internet zone

```
BLASTX 2.2.5 [Nov-16-2002]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= Nicotiana tabacum plastid, complete genome
         (155,939 letters)

Database: blast/blast_dbs/cp_genes/psbA.aa
           15 sequences; 5295 total letters

Searching...done


                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

spinach, psbA                                                       721   0.0
nicotiana, psbA                                                     721   0.0
oenothera, psbA                                                     720   0.0
arabidopsis, psbA                                                   720   0.0
lotus, psbA                                                         717   0.0

>spinach, psbA
          Length = 353

 Score =  721 bits (1860), Expect = 0.0
 Identities = 353/353 (100%), Positives = 353/353 (100%)
 Frame = -1

Query: 1595 MTAILERRESESLWGRFCNWITSTENRLYIGWFGVLMIPTLLTATSVFIIAFIAAPPVDI 1416
            MTAILERRESESLWGRFCNWITSTENRLYIGWFGVLMIPTLLTATSVFIIAFIAAPPVDI
Sbjct: 1    MTAILERRESESLWGRFCNWITSTENRLYIGWFGVLMIPTLLTATSVFIIAFIAAPPVDI 60

Query: 1415 DGIREPVSGSLLYGNNIISGAIIPTSAAIGLHFYPIWEAASVDEWLYNGGPYELIVLHFL 1236
            DGIREPVSGSLLYGNNIISGAIIPTSAAIGLHFYPIWEAASVDEWLYNGGPYELIVLHFL
Sbjct: 61   DGIREPVSGSLLYGNNIISGAIIPTSAAIGLHFYPIWEAASVDEWLYNGGPYELIVLHFL 120

Query: 1235 LGVACYMGREWELSFRLGMRPWIAVAYSAPVAAATAVFLIYPIGQGSFSDGMPLGISGTF 1056
            LGVACYMGREWELSFRLGMRPWIAVAYSAPVAAATAVFLIYPIGQGSFSDGMPLGISGTF
Sbjct: 121  LGVACYMGREWELSFRLGMRPWIAVAYSAPVAAATAVFLIYPIGQGSFSDGMPLGISGTF 180

Query: 1055 NFMIVFQAEHNILMHPFHMLGVAGVFGGSLFSAMHGSLVTSSLIRETTENESANEGYRFG 876
            NFMIVFQAEHNILMHPFHMLGVAGVFGGSLFSAMHGSLVTSSLIRETTENESANEGYRFG
Sbjct: 181  NFMIVFQAEHNILMHPFHMLGVAGVFGGSLFSAMHGSLVTSSLIRETTENESANEGYRFG 240

Query: 875  QEEETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVVGIWFTALGISTMAFNLNGF 696
            QEEETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVVGIWFTALGISTMAFNLNGF
```

Figure 7: Clicking on the gene name in the Blast Hit Panel brings up a window with the raw BLAST output.



Database entry in spinach for psbA

```
>spinach, psbA
MTAILERRESESLWGRFCNWITSTENRLYIGWFGVLMIPTLLTA
TSVFIIAFIAAPPVDIDGIREPVSGSLLYGNNIISGAIIPTSAAIGLHFYPIWEAASV
DEWLYNGGPYELIVLHFLLGVACYMGREWELSFRLGMRPWIAVAYSAPVAAATAVFLI
YPIGQGSFSDGMPLGISGTFNFMIVFQAEHNILMHPFHMLGVAGVFGGSLFSAMHGSL
VTSSLIRETTENESANEGYRFGQEEETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLA
AWPVVGIWFTALGISTMAFNLNGFNFNQSVVDSQGRVINTWADIINRANLGMEVMHER
NAHNFPLDLAAIEAPSTNG
```

psbA

```
TTCTTTTCAAAACCTCCTATAGACTAGGCCAGGATCTTATCCATTTGTAGATGGAGCTTCGATA
AAGAAAAGTTTTGGAGGATATCTGATCCGGTCCTAGAATAGGTAAACATCTACCTCGAAGCTAT
                                nico     G  N  T  S  P  A  E  I
                             spinach     G  N  T  S  P  A  E  I
                           nicotiana     G  N  T  S  P  A  E  I
                           oenothera     G  N  T  S  P  A  E  V
                         arabidopsis     G  N  T  S  P  A  E  V
                               lotus     G  N  I  S  P  A  E  V
```

Figure 8: Clicking on the taxon name in the Blast Hit Window brings up a window with the database entry for the current gene for that taxon.

strand) as well as within the sequence. To choose a different start codon than the end of the BLAST hit, click on the link and it will change the end of the gene. The start codon for psbA is correct, so we don't need to change it.

You may also change the start and end of the gene manually by typing in the fields in the Sequin Window. If you mouse over the green start or red stop codon, the new start or stop information will appear at the bottom of the browser window (see Figure 8).

Fill in any other Sequin information for this gene (the product will eventually show up in the window so you won't have to enter it). Then click on the "Commit" button. The Sequin Window will close and the new information for psbA has been recorded.

Click on the "Show Sequin Format" button in the bottom panel, and the Sequin information for psbA will be seen. Hit the back button, then hit the "Show Text Summary" button. The text summary is the initial BLAST hit for each gene. You will see that psbA has now been updated to include the stop codon.

6.2 A gene on the forward strand: psbD

Protein coding genes on the forward strand are similar to the reverse strand. Scroll down to the number line with psbD on it and click the colored box for that gene (at approximately 34,500). A new Sequin Window will appear, and in the Blast Hit Panel, the psbD gene appears with the BLAST hits above the purple nucleotides representing that gene.

There is not a gene within the 60 upstream nucleotides of psbD, so the nucleotide sequence is grey. The potential start codons are highlighted as green links. The original start codon is correct so scroll all the way to the right in the Blast Hit Panel.

The downstream gene from psbD, psbC, overlaps. Because it overlaps and is on the same strand, it is truncated where it overlaps. The potential stop codons are show in the nucleotide sequence of psbC. Click on the first highlighted stop codon, then click the "Commit" button in the Sequin Window.

6.3 A gene with an intron: ndhB in irB

For protein coding genes that have an intron, both pieces must be recorded together in the Sequin input file. Scroll down to ndhB in irB (at approximately 142,500) and click on the first ndhB. You'll notice the start codon is good, but there is no stop codon. Click on the next ndhB. You'll notice there is no start codon, but there is a stop codon. Click on the stop codon. Then go to the Sequin Window and click on the "Click here if gene contains an intron" button. All of the ndhB genes appear (in both copies of the inverted repeat). The two exons we'd like to join are the last two. Click the two bottom check boxes. You'll notice that the start of the gene changed to include both exons. Click on the "Commit" button. Now if you view the Sequin format, you'll see the two coding sequences listed for the ndhB gene and if you click the "Refresh" button, that gene will include the two exons and the intron.

6.4 A tRNA: trnG-GCC

Next we'll look at annotating a tRNA. Scroll down to trnG-GCC (at approx. 38,000) and click on the colored box. The gene appears in the Blast Hit Window with the nucleotide sequence BLAST hits (Figure 10). You'll notice there are no highlighted start or stop codons. The change the start or the end of the tRNAs, you must change it manually in the Sequin Window. For this tRNA, the BLAST match for the start and end are accurate, so click on "Commit" in the Sequin Window. The window will disappear, and you can check the entry for the tRNA in the Sequin file by clicking on "Show Sequin Format" in the bottom panel.

6.5 A tRNA with an intron: trnA-UGC

An example of a tRNA with an intron is trnA-UGC. Scroll down to 105,000 and click on the box for trnA-UGC. Because the intron is well-conserved, the BLAST hit contains the intron in the tRNA. The BLAST hits for wheat and rice are split up into two separate hits because there is a larger gap in the database sequence than in the query sequence.

6.6 An rRNA:

For rRNAs, there is a button in the Sequin Window which allows the user to join two pieces of rRNA which BLAST did not think (BLAST thinks?) was contiguous.