Analysis of Mystery *Agrobacterium* Sequence

Part I: Initial analysis of open reading frames

I. Overview

You should all now have received a DNA sequence taken from the bacterium *Agrobacterium tumefaciens*. See Lab 6 in your manual for reasons why this is a remarkable organism worthy of your profound attention. Besides Brad, you are the only person on earth to consider the sequence you have been given. Your task for the remainder of the semester is to plumb the depths of this sequence and find out what secrets you can tear from Nature's grasp.

The first step (as described further on p.3 of Lab 6), is to find one or more large open reading frames and describe them. You have two useful tools at your disposal. *EditBase* will help you find open reading frames and open them to your inspection. The THIRD appendix to Lab 2 (cleverly named Appendix II) tells you how to find open reading frames both on the strand you were given and its complement. The web program DAS will help you analyze the regions of predicted proteins for hydrophobicity.

II. How to get sequence into *EditBase*

Depending on how your sequence file was sent, it might be possible to import it directly into *EditBase*. Otherwise, you'll have to run it through a word processor first.

II.A. Direct import into *EditBase*

1. Make sure that your sequence file is in the same directory as *EditBase*
2. Enter *EditBase*
3. Import file: Click on **F**ile, then **I**mport. Press F1 to get a list of files on your disk or subdirectory. Select the sequence file. Select **D**NAsis as the format and **N**ucleotide bases as the type. You should see the sequence appear on the screen.
4. Check carefully for spaces. Check particularly bases 51, 102, etc. If there are spaces within the sequence, then abort the process by quitting *EditBase* (answer **N**o when asked whether you want to save the changes to a file). In this case you'll need to use a word processor (see Section II.B or II.C).

5. If there are no spaces, then save the sequence to a file (click **F**ile, then **S**ave), using a filename other than the original. I suggest something with the extension ".SEQ" to mark it as an *EditBase*-readable file.

II.B. Cleaning up file in a word processor

1. Read the file into your favorite word processor.
2. Do a global replace, eliminating all spaces. In MS Word, click **E**dit, then **R**eplace. Then enter a single space in the "Find what" box. Click on the "Replace all" button.
3. Close the dialog box and save the file as text. In MS Word, this is done automatically unless you accessed the file in a strange way.
4. Import the file as described in Section II.C.

II.C. Cut and paste file into *EditBase*

1. Read the file into your favorite word processor.
2. Block out the text. In MS Word, click **E**dit, then Select a**l**l.
3. Copy text to clipboard. In MS Word, click **E**dit, then **C**opy.
4. Open *EditBase* (get past the copyright screen)
5. Press Alt-Space (first the Alt key and then, while the Alt key is still depressed, tap the space bar). This will bring up a Windows menu.
6. Click on **E**dit, then **P**aste. You should see the sequence appear on the screen.
7. Make sure all of the sequence is there.
8. If all is well, then save the sequence to a file (click **F**ile, then **S**ave), using a filename other than the original. I suggest something with the extension ".SEQ" to mark it as an *EditBase*-readable file.

III. How to get sequence into DAS

To use DAS, you will need to supply it with the amino acid sequence of open reading frames you find. To do this, you need to generate the amino acid sequence as one-letter abbreviations. *EditBase* will help you part of the way, then you have to seek outside help, either through a word processor or *AASeq*, a program that was sent to you.

III.A. Generate amino acid sequence (part 1)

1. Find open reading frame (described in lab manual, third appendix of Lab 2)
2. Block out the open reading frame in question by clicking on the beginning of the open reading frame and then clicking on the end.
3. Turn on amino acid translation for the block if it isn't on already, by clicking on **B**lock, then coding o**N**.

4. Toggle to the one-letter amino acid abbreviations by clicking on **T**oggle, then clicking on **L**ength of AA abbrevs to 1 (if it says 3, then press Esc). You should now see one letter sitting over every triplet in the open reading frame. Figure 2 from the Sept 1 notes shows all 20 amino acids with their one letter codes.
5. Output open reading frame to a file (with the open reading frame still highlighted). To do this:
    a. Click on **A**nalyze
    b. Click on **L**isting of sequence
    c. Press return to accept 60 bases per line
    d. Press **N**o to exclude spaces between groups
    e. Press **F**ile to send output to a file
    f. Type in some name you'll remember

Unfortunately, *EditBase* outputs not only the amino acids but the bases as well, not to mention all sorts of other garbage. You have two ways to clean up the output so that DAS can handle it. Section III.B. describes a method that is easy but foreign to your experience. Section III.C. describes what might be a familiar, if tedious, method.

III.B. Clean up *EditBase* output using handy utility to extract amino acid sequence

1. In Windows, click on **Start** button and click on **R**un.
2. Browse to find the program named *AASeq*, attached to this message.
3. Run that program. You'll need to have in your mind the filename where you put the raw output from *EditBase*. It will probably be on a floppy disk (Drive A:), in C:\Eudora\attach, or wherever your mail program puts attachments.
4. After running the program, bring up the resulting file in your favorite word processor.
5. Go to Section III.D.

III.C. Tidy up sequence manually in word processor

1. Bring up the crude output from *EditBase* into your favorite word processor
2. Get rid of all header lines, all lines with DNA sequence, anything except lines with amino acid letters
3. Remove all spaces (you can do this with a global REPLACE)
4. Go to Section III.D.

III.D. Run DAS

1. Go to DAS web site (http://www.sbc.su.se/~miklos/DAS/)
2. Cut and paste amino acid sequence from word processor into DAS

IV. Interpretation of results from DAS

This program analyzes the hydrophobicity of segments of an amino acid sequence and tells you if it thinks it is hydrophobic enough to be in a membrane. You should note the segments higher than the thresholds given in the graphical output and record the start and end points in your amino acid sequence. The sequence numbers given by DAS correspond to amino acids. The numbers in *EditBase*, of course, refer to nucleotide position. You'll have to translate from the first to the second to locate the regions. Check on the program – look up a few amino acids. Are the regions identified as hydrophobic truly hydrophobic, in your judgment?

How can you tell whether a region is likely to span a membrane. DAS doesn't give you an answer, it merely provides you with information about hydrophobicity. To give you a feeling how to interpret the information, I've supplied you with the sequences of two known proteins: hexokinase (a soluble globular protein; recall notes from Sept 1) and rhodopsin (a membrane-spanning protein; recall Problem 1.14, although that protein wasn't exactly rhodopsin). Run these protein through DAS.

You'll note that it isn't enough that a region pierces the hydrophobicity threshold. The **length** of the region is also important. It has to be long enough to span a membrane! Let rhodopsin teach you how long that length must be.