Protein Sequence Alignment and Phylogenetic Analysis

Overview:

Gene sequence comparison is a powerful tool for molecular biologists for both the isolation of specific sequences and the characterization of newly cloned sequences.

In this course, we have already compared conserved regions of homologous proteins from different species to design degenerate RT-PCR primers so that we can clone cDNA sequences encoding the same protein from additional sources.

In today's lab exercise, we will examine some computer tools that are useful for characterizing the evolutionary relationships between protein sequences. This process can be important for both analyzing the evolution of known gene sequences and for placing newly isolated sequences into an evolutionary context. Sequence similarities between different proteins are often an indicator of similar structures and functions.

Our analyses will have two types of outputs:

1) Indications of amino acid conservation between different protein sequences, including percent identical amino acids and sequence alignments, and

2) Phylogenetic trees--diagrams indicating patterns of common ancestry and evolutionary divergence between different sequences.

Because we have not yet cloned any sequences of our own, we will obtain sequence information for different proteins from Genbank, the public database of protein and DNA sequences maintained by the National Center for Biotechnology Information (NCBI), part of the National Institutes of Health (NIH). All biological journals require that newly-reported sequences must be deposited in a public database prior to publication of a paper. Because the Genbank database and databases maintained in Europe (Swissprot) share information freely, either source of information represents a comprehensive and current collection of sequence information.

In the lab exercise, we will analyze CYP1A sequences. A homework assignment will consist of a similar analysis of a different gene family that will be assigned.

1. Obtaining sequence information from GenBank:

A) key word search.

»»Using a web browser, access the NCBI home page [http://www.ncbi.nlm.nih.gov/]. Near the top of the page, you will see a horizontal search bar. A choice menu allows you to choose GenBank, PubMed, OMIM, Protein, Nucleotide, etc. Set this choice to PROTEIN. In the white box at the right, type in CYP1A. Database entries containing this key word will be listed on a new page. How many entries are returned? _____

[NOTE: PubMed is an especially useful tool that allows you to search the biomedical literature using keywords or authors. This is THE place to find biomedical journal articles.]

»»Now refer back to the table of aligned sequences that we used last week to design primers. How many sequences had I placed in the alignment? _____ Are they the same sequences? All of these sequences were derived from GenBank. Why the difference in number and identity? Are all your GenBank results CYP1A homologs, or are other types of sequences identified as well? Why?

»»Try entering "CYP1A and *Xenopus*" [Boolean operators can be used in these searches].
Try "CYP1A6 and *Xenopus*." Did either of these searches yield an entry?
Try "cytochrome and *Xenopus*." How many entries did this search retrieve? Which of these are you interested in?

This exercise illustrates the random and ambiguous nature of *key word* searches. If the key word was placed in the GenBank entry then the entry will likely be included on the hit list. But if the key word is not there, then relevant sequences may be omitted from the hit list. The approach is very useful if you know exactly what you're looking for and what it's called, but for more general data-mining applications, the limitations are severe. Furthermore, this type of search yields a set of sequences based on preconceived notions about which sequences will be important in your analysis. There may be proteins with sequence similarities that you are unaware of and can't discover based simply on the name of the protein. The keyword search may also identify *functionally* relevant proteins that share no sequence conservation, e.g. a transcription factor that regulates expression of the enzyme that was the subject of your search. To get a comprehensive list of truly homologous protein sequences, a BLAST search is a better way to go.

[Note: To view and/or copy a single sequence, click on the Accession Number for that sequence (the first thing in the entry; consists of letters and numbers). This links to the entire entry, including the sequence, journal article, and other annotations. *Click on several links within the entry to discover what information is available*.]

B) BLASTP search

NOTE:
The on-line BLAST guide is a good resource for general information on how to use these web pages [http://www.ncbi.nlm.nih.gov/BLAST/about/].
The HELP tab provides access to more in-depth tutorials about BLAST searching [http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs] I urge you to try these out and explore the possibilities. Following is a limited example of how you might use a BLAST search to identify sequences of interest for a subsequent phylogenetic analysis.

An alternative to searching for sequences based on key words is to search the sequence database using a sequence that you know is relevant. A BLAST search takes this sequence and compares it with all the sequences in the database. Results are returned as a ranked list of homologous sequences, an alignment of the amino acids in your query sequence and each returned entry, and an expression of the degree of homology (i.e., % identities). BLAST searches can be conducted using amino acid sequences (BLASTP) or nucleotide sequences (BLASTN). Some other NCBI BLAST searches include[1]:

BLASTX: compares a nucleotide query sequence translated in all reading frames against a protein sequence database
TBLASTN: compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
TBLASTX: compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database (Please note that tblastx program cannot be used with the nr database on the BLAST Web page).

»»Conduct a BLASTP search using one of the CYP1A sequences identified in the first keyword search of GenBank (repeat the search, if necessary). Highlight and copy (or write down) the accession number for one of the sequences.

A convenient **BLAST** link is located in the blue bar at the left of your GenBank results page. It will direct you to the BLAST home page [http://www.ncbi.nlm.nih.gov/BLAST/]. Under the heading Protein BLAST, click on **blastp (protein-protein BLAST)**.

»»Now enter the accession number into the query box at the top of the page and click on the BLAST button. [NOTE: this box can also be used for entering actual amino acid sequences.]

A new page will appear with important information about the progress of your BLAST search. It will verify the accession number and number of amino acids in the sequence you entered. It will assign an ID number to your search and estimate the time required to complete it. This is dependent upon traffic and varies from a few seconds to several minutes.

The results appear in a new web page. They consist of three types of information.

The first part of the result is a graphical overview of homologous sequences. This is a color-coded indication of the degree of homology in the top-ranked matches. A key to the colors is found above. When you mouse over a lines, the name and accession number of the matched protein will appear in the box above the graphic. If you click on a line, you will be taken to an amino acid alignment of your query protein and that particular result protein located farther down the results page.

[1] descriptions taken from the NCBI web page: http://www.ncbi.nlm.nih.gov/BLAST/blast_program.html

Below the graphical result is a ranked list of all the BLAST matches. The best matches are listed first. You can click on the accession number at left to access the GenBank entry page for that sequence.

Below the ranked list and GenBank accession numbers is a list of pairwise alignments. These are the same alignments you accessed by clicking lines in the graphical overview section, above. Here you see the actual positions of amino acid identities shared between your query sequence and the matched sequence ("subject"). Indicated above each alignment is the number and percenage of identical positions.

BLAST vs. Key Word Search: The BLAST search identified more proteins than they keyword search of GenBank. It also ranked them according to degree of identity from your query protein and provided detailed information about exactly which amino acid positions are identical. A link at the top of the list will even construct a phylogenetic tree of the results. This type of information is very helpful in choosing which proteins from the database should be used in a focused phylogenetic analysis to test specific hypotheses. If you conduct the search with a newly-cloned sequence, BLAST searching provides an important clue about the identity and function of that sequence based on its similarities to known proteins.

2. Amino acid alignment and construction of phylogenetic trees.

There are at least three basic types of computer algorithms for the construction of phylogenetic trees. Consensus is lacking on which method most reliably constructs the evolutionary history of gene sequences. In this exercise, we will learn how to construct phylogenetic trees using the "neighbor-joining method," which is a type of "distance method." Other methods include the "maximum parsimony method" and the "maximum likelihood method." Although many contemporary experts in molecular evolution criticize neighbor-joining and other distance methods, it is said to be valid for relatively small data sets (such as those we will work with), and it is nonetheless widely used because it requires relatively little computing power and because it can be accomplished using free, publicly available software.

Regardless of the chosen algorithm, the first step in phylogenetic tree construction is the *alignment* of all sequences included in the analysis. We will use a program called CLUSTAL X to perform an amino acid alignment of CYP1A protein sequences.

»»To access CLUSTAL X on Kenyon Biology public computers, use this shortcut path:
Start Menu > Programs > Bioinformatics > clustalX.exe.

[NOTE: you can copy this software to your own computer from P:class/biology/BIOL 264/ Bioinformatics-Powell/clustalx1_81_msw. Copy the entire folder to the C: drive on your personal computer.
Publicly available at http://bips.u-strasbg.fr/en/Documentation/ClustalX/
Mac users: see http://www.embl-heidelberg.de/~chenna/clustal/darwin/.]

A blank window will appear, and we want to fill the window with the sequences to be analyzed.

**NOTE: Sequences cannot be entered directly into this window. They must be obtained from a separate text file. I have already prepared such a file for you [P:class\biology\BIOL264\CYP1_sequences.txt].

When you prepare your own text file of sequences for analysis (e.g. in your next lab report), you will obtain them from GenBank via BLAST and/or keyword search. While you can copy and paste sequence data from the GenBank entry pages, be sure to copy the proper format called FASTA. Here's how:

a) From a GenPept report page of an individual sequence, look for a selection menu near the top entitled Display.
b) Select FASTA. In the FASTA format, the first character should be ">". That symbol is followed by a brief descriptor to indicate the species and gene. There should be an open line between the end of a sequence and the FASTA title for the next sequence. There should not be an open line between titles and corresponding sequences.
c) this file is created in a word processing program (I use MSWord for Mac; you probably like something else). Regardless of the word processing program used to create the file, it must be saved as a TEXT ONLY file (.txt), or CLUSTALX will not read it.

NOTE: previous classes have discovered that if you open this file in a program like NOTEPAD, it will not work with the alignment software. Stay away from NOTEPAD, SIMPLETEXT, TEXTEDIT etc. *This often requires opening the file from within an application, not merely double-clicking an icon.*

»»Open [P:class\biology\BIOL 264\CYP1_sequences.txt] and observe the sequences in FASTA format. IMPORTANT: copy this file to [H:] now, and access the copy on [H:] henceforth. DO NOT use the copy on [P:] directly!

»»Within the CLUSTALX program, goto the FILE menu and choose LOAD SEQUENCES. A new window will appear from which you can select CYP1_sequences.txt *FROM YOUR H-drive*! The sequences should appear in the CLUSTALX window.

»»Under the ALIGNMENT window, select DO COMPLETE ALIGNMENT. A window will appear asking about the direction for the output files.

IMPORTANT: make sure the output files are directed to [H:], not [P:class]. . . ! If you input the sequences from the text file on [H:], this should happen automatically.

The alignment process takes a few minutes. It happens faster if there are not other applications running on the computer (including webmail, Facebook). When the process is finished, you can notice how the positions of the amino acids have shifted. The colors correspond to different types of amino acids (e.g. purples are negative (E, D), oranges are positive (R, K), etc.) A ruler above the sequences indicates identities with a ":" symbol. Other symbols indicate "similarities," i.e. not an exact residue but one which shares chemical properties (e.g. D instead of an E; I instead of a V).

NOTE: This is not a great format for the graphical presentation of alignments. We will discuss some options for that later in the course. This is a good method for calculating alignment for use in drawing trees, however, and that's what we will proceed to do.

With the completed alignment in hand, it is now time to draw the tree.

Under the TREE menu, there are two choices: "draw N-J tree" and "bootstrap N-J tree." When you "draw" a tree, the computer does the operation one time and generates an output file. When you "bootstrap" the tree, the computer does the operation many times (usually 1000), introducing small misalignments in each iteration. Sometimes it gets a different result. The NUMBERS that appear within a bootstrap tree indicate how many of the 1000 trees contained an identical branch point. It's a democratic process, and branch points with less than 1/2 the vote are discarded. Overall, large bootstrap values indicate strong statistical support for a particular branch.

DRAW YOUR TREE:

1. Under the TREES menu, choose OUTPUT FORMAT OPTIONS. In the dialog box select NODE in the Bootstrap labels on popup menu.

2. Under the TREES menu, choose BOOTSTRAP N-J TREE. A window will appear asking for three pieces of information. The first two are a RANDOM NUMBER SEED and a NUMBER OF BOOTSTRAP REPLICATES. The default values on these are fine. The last box asks for the direction of the output file. Again, if you started from the text file in your H-drive, the output will also go to your H-drive.

A tree output file is not a visible tree. We need another program to convert the output file to a graphical tree format. This program is called *TreeView*.

To launch the TreeView application, follow this path:
Start Menu > Programs > Bioinformatics > Treeview.

[Note: you can have your very own copy of TreeView by copying in from the P: drive.
P:class/biology/biol364/treev32.exe
Public access at http://taxonomy.zoology.gla.ac.uk/rod/treeview.html]

BUG ALERT: A small introductory window will appear in the center of your screen. To make it go away (as it ought to do on its own), right click the treeview tab at the bottom of the screen. Select Close. Repeat this process. Now you're ready to work. Remember that this is free software. Don't curse the developers. Definitely don't curse the course instructor.

Under the FILE menu, select OPEN. Use the window that appears to select [CYP1_sequences.phb] on your H-drive.

[NOTE: the bootstrap tree file will always take the name of the text file originally aligned in the .phb format.]

The tree will appear immediately in the window.

Changing the Tree's Appearance:

The default appearance may be the “slanted cladogram.” Use a menu button (Or Trees Menu) to change this to a “phylogram.” Other styles of trees are also available, but the rectangular form is standard in the manuscripts we have read.

The OUTGROUP of a tree should be the least related sequence(s). In this CYP1A tree, CYP1B sequences define the outgroup. To make sure these sequences are in the outgroup, choose DEFINE OUTGROUP from the TREE menu. Select appropriate sequences in the left window and hsift them to the right using the arrow button. Note that you cange change your mind and

move a sequence back with the other arrow button. Click OK when you're done. You may also need to ROOT WITH OUTGROUP, a command in the same window. Not how the arrangement of the tree changes with each of these adjustments.

[NOTE: If your bootstrap values don't appear on the tree, see step 1 on page 7.]

You may need to resize windows to get a good look at your tree.

To print the tree in even nicer form, COPY it under the EDIT menu and PASTE into a Powerpoint slide. Ungroup the pasted object. Now all the text and all the lines can be edited, resized, etc. These are skills you likely have learned in other contexts.

QUESTIONS to PONDER and DISCUSS:

1. Which CYP1 enzymes are the most distantly related to the others (hint: the earliest branch point).

2, Evolutionary Relationships and Nomenclature:

i) There are multiple forms of CYP1A for several mammals (CYP1A1 and CYP1A2). There are duplicate CYP1A sequences in *Xenopus* (CYP1A6 and CYP1A7) and chicken (CYP1A4 and CYP1A5) as well. Why not simply call the frog and bird sequences 1A1 and 1A2, just like the mammals?

ii) What about the fishes, which have only a single CYP1A gene? Should these be designated 1A1?

3. In general, does the evolution of the CYP1A gene match what you know about the overall relationships between vertebrate groups? What does your conclusion say about the utility of inferring relationships between SPECIES based on the sequence of a single gene?

Assignment: LAB REPORT #2:

In class, each of you will be assigned a gene or gene family to analyze. You will need to identify homologous protein sequences in GenBank and construct a comprehensive amino acid alignment and a phylogenetic tree using CLUSTALX and Treeview.

Your report should include:

- a descriptive title
- A brief introduction giving some background about the function of the protein/protein family in question. Useful references are available as links within the genbank entry. I recommend you read and cite some of these.
- A materials and methods section that gives the accession numbers of the sequences you analyzed, how they were identified, and names the computer programs used in the analysis.
- A results section that includes figures of the translated cDNA sequence and your phylogenetic tree. Model your figures and figure legends after the papers issued in the first class period (Morrison et al. 1995, 1998).
- A discussion section that comments on the relationship of the tree to overall evolution of the taxa, any gene multiplicity observed, and anything you think might be important about the functions of the proteins in the different groups. Cite references discovered in your sequence search as appropriate.

It should be about 4-5 pages long (double-spaced), including the figure and references. We will discuss the details of the report at the end of class.

DUE DATE:

Background reading assigned for class discussion associated with this module:

Robinson-Rechavi et al. (2001) Euteleost fish genomes are characterized by expansion of gene families. *Genome Research* 11:781-788.

Lynch, M (2002) Gene Duplication and Evolution. Science 297:945-947.

Unda, Faride (2007) Introduction to phylogenetics. *Science Creative Quarterly*: http://www.scq.ubc.ca/?p=140

Relevant background reading from previous weeks:

Morrison, Hilary G., E. Jennifer Weil, Sibel I. Karchner, Mitchell L. Sogin, and John J. Stegeman, 1998. Molecular cloning of CYP1A from the estuarine fish *Fundulus heteroclitus* and phylogenetic analysis of CYP1 genes: update with new sequences. *Comparative Biochemistry and Physiology Part C* 121: 231-240.

Morrison H.G., M.F. Oleksiak, N.W. Cornell, M.L. Sogin, J.J. Stegeman. 1995. Identification of cytochrome P-450 1A (CYP1A) from two teleost fish, toadfish (*Opsanus tau*) and scup (*Stenotomus chrysops*), and phylogenetic analysis of CYP1A genes. *Biochem. J.* 308: 97-104.

Berdtson, A., and Chen, T. 1994. Two unique CYP1 genes are expressed in response to 3-methylcholanthrene treatment in rainbow trout. *Archives of Biochemistry and Biophysics* 310: 187-195.