Phylogenomic Inference of Protein Molecular Function

One of the fundamental paradigms in computational biology is function prediction by homology. In this framework, a gene or protein is compared against other genes or proteins in a database, and, if a sequence can be detected whose similarity is statistically significant, the function of the unknown gene or protein is inferred based on the known (or presumed) function of the homolog.

Homology-based predictions are used to gain a first-order approximation of the molecular function of the proteins encoded in a genome, and to prioritize experimental investigation. While computationally efficient methods for pairwise sequence comparison—notably BLAST (Altschul et al., 1990)—have been developed to make this approach feasible in high-throughput, homology-based function prediction is not without its dangers. Systematic errors associated with this paradigm have become increasingly apparent (Bork and Koonin, 1998; Eisen, 1998; Galperin and Koonin, 1998). Biological processes such as gene duplication (Fitch, 1970), domain shuffling (Doolittle and Bork, 1993; Doolittle, 1995), and speciation (Galperin and Koonin, 1998; Gerlt and Babbitt, 2001) produce families of related genes whose gene products can have vastly different molecular functions. Finally, existing database errors can be propagated through function prediction by homology (Brenner, 1999; Devos and Valencia, 2001; Gilks et al., 2002).

This unit presents a workflow (Fig. 6.9.1) for phylogenomic inference of protein molecular function (Eisen, 1998; Sjölander, 2004) for a sequence of interest. The first step involves identification of homologs and construction of a multiple sequence alignment. For this task, the FlowerPower Web server is presented in Basic Protocol 1.

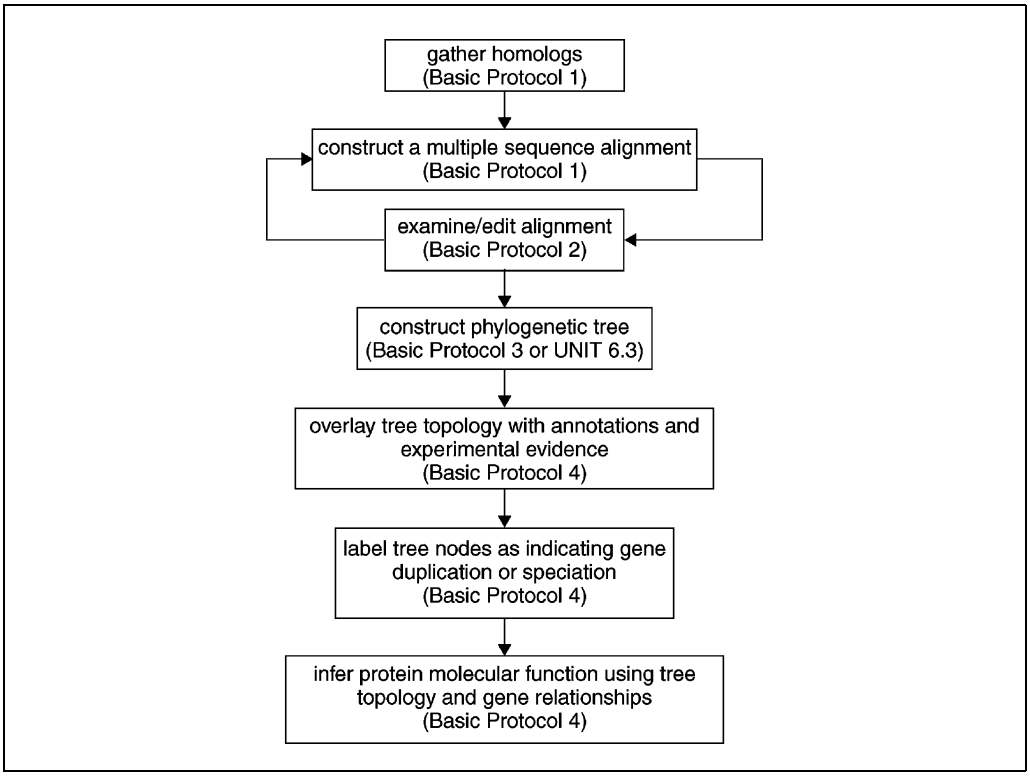

Figure 6.9.1 Workflow for phylogenomic analysis. For every step in the workflow, the basic protocol in this unit that describes it is given in parentheses.

Contributed by Nandini Krishnamurthy and Kimmen Sjölander
Current Protocols in Bioinformatics (2005) 6.9.1-6.9.19

Next, the sequence alignment is analyzed and edited using the Belvu software (Basic Protocol 2). A phylogenetic tree can then be constructed using Neighbor Joining or the BETE software (Basic Protocol 3). The phylogenetic tree is displayed with annotations culled for the members of the family using the annotation software TreeNotator (Basic Protocol 4). Finally, the tree topology is analyzed to label branch points as indicative of either speciation or gene-duplication events, enabling the discrimination of orthologs from paralogs. Changes in biochemical function (and sometimes structure) can be traced along the evolutionary tree. For proteins whose functions are unknown, consistency of database annotations within subtrees containing the protein can be used as the basis for function prediction (a process termed "subtree neighbors"; Zmasek and Eddy, 2002).

The reliability of a phylogenomic inference of protein molecular function obviously depends on the accuracy of the phylogenetic tree, which depends in turn on the accuracy of the multiple sequence alignment and on a representative and thorough sampling of the protein family under analysis. Accuracy of annotation data is an additional critical dependency.

BASIC PROTOCOL 1

IDENTIFYING HOMOLOGS AND CONSTRUCTING A MULTIPLE SEQUENCE ALIGNMENT USING FlowerPower AND MUSCLE

The first step in phylogenetic inference of protein function involves identifying homologs to a given protein and constructing a multiple sequence alignment as input to tree construction. The primary protocol for this stage, as detailed below, presents the use of the Berkeley Phylogenomics Group FlowerPower Web server for clustering and aligning protein sequence homologs to a user-supplied seed sequence. FlowerPower constructs a multiple sequence alignment during the homolog clustering process; users can download both the FlowerPower alignment and the MUSCLE (Edgar, 2004)

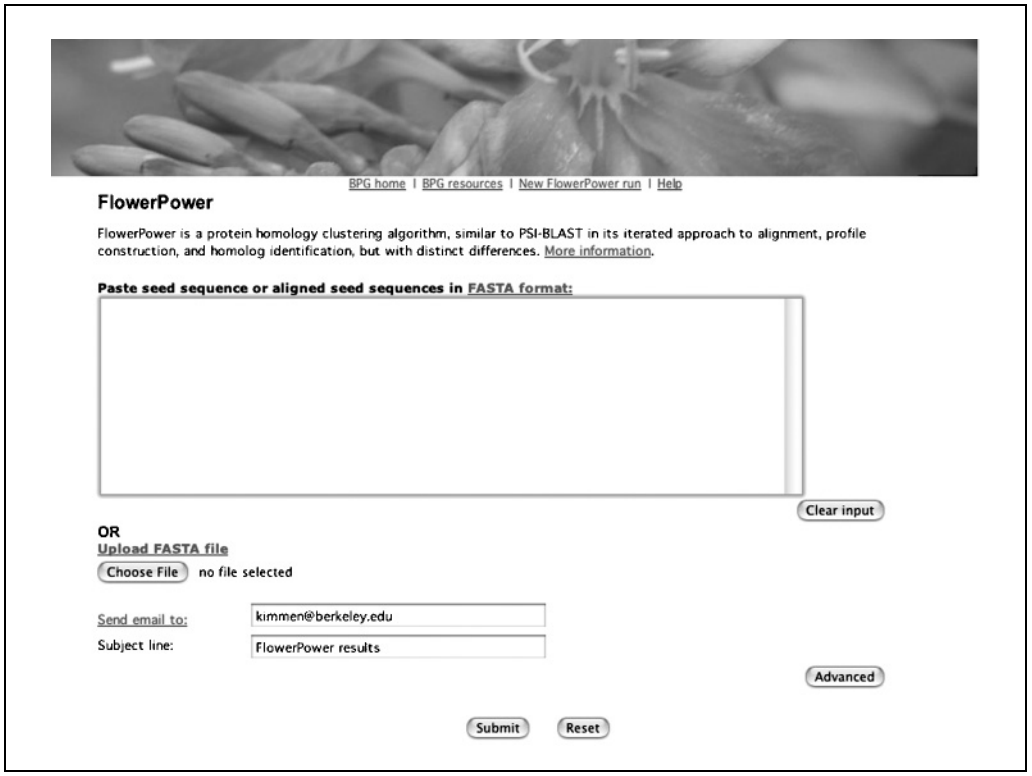

Figure 6.9.2 FlowerPower submission page. Paste the seed sequence in the box provided, in FASTA format. Type a valid E-mail address in the box provided. Results will be sent to that address.

realignment of the same sequences, as well as the raw (unaligned) sequences. Users who prefer to use a different method for multiple alignment would naturally download the raw sequences.

Necessary Resources

Hardware

Any computer with an Internet connection

Software

Web browser

Files

Protein sequence: The sequence to be used as a seed for FlowerPower should be in FASTA format (see *APPENDIX 1B*). An example FASTA format file is also shown in Figure 6.9.4.

1. Point the browser at the FlowerPower Web server (*http://phylogenomics.berkeley.edu/flowerpower/*).
2. Paste the seed sequence in FASTA format into the box provided (see Fig. 6.9.2).
3. Enter a functioning E-mail address as directed. Results will be sent to this address.

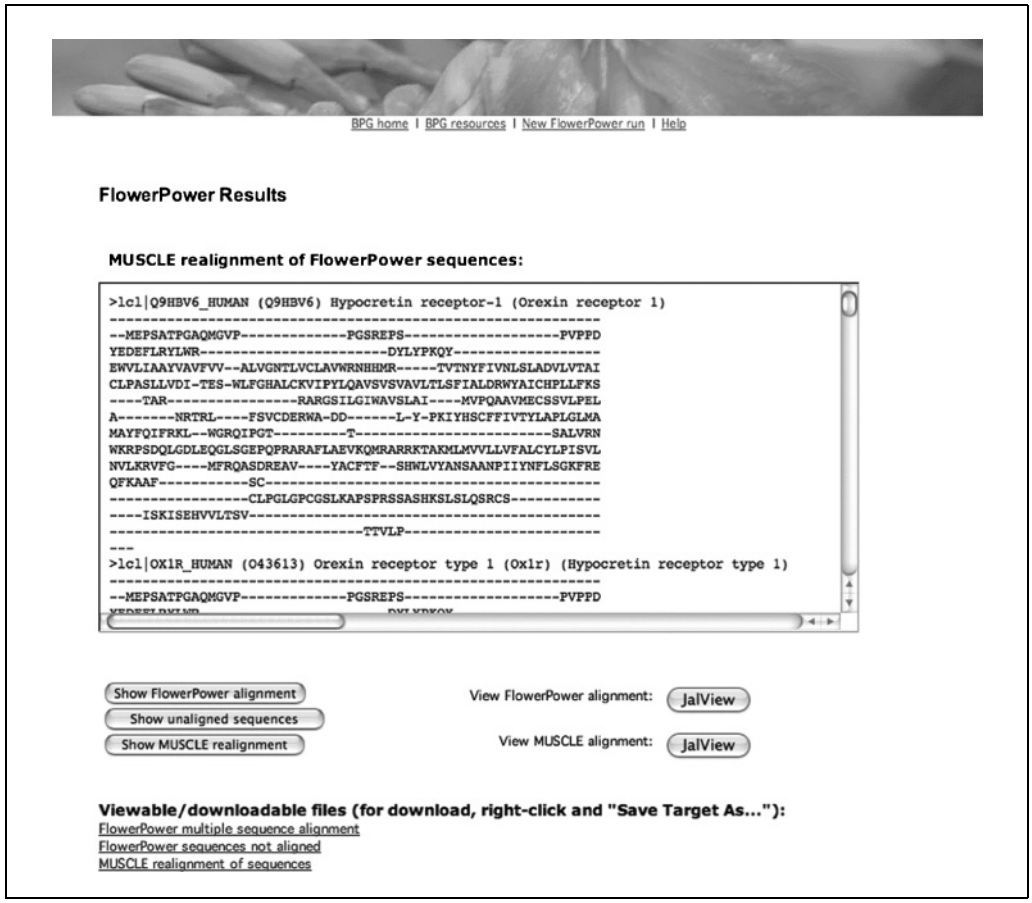

Figure 6.9.3 FlowerPower results page. The FlowerPower and MUSCLE alignments can be downloaded by clicking on the hyperlinks at bottom left. The FlowerPower cluster (unaligned sequences) can be downloaded using the second hyperlink from the bottom. The MUSCLE and FlowerPower alignments can be viewed directly on the Web page by clicking on the JalView buttons at right.

4. Use the default parameters to gather global homologs to the seed from the UniProt database.

   Users can override default parameters by clicking on the Advanced button. Advanced settings make it possible to select a different sequence database and to adjust the inclusion cutoff parameters (e.g., minimum length coverage and pairwise identity). Note that overriding default inclusion cutoff parameters can result in nonglobal homologs being included in the final result, producing inaccuracies in the phylogenomic analysis.

5. Click Submit.

6. Retrieve results sent by E-mail.

   The Web page includes two alignments of the sequences retrieved by FlowerPower (Fig. 6.9.3); the first alignment is constructed by FlowerPower and the second is a realignment of these sequences using the MUSCLE software. If a different multiple alignment program is desired, download the unaligned sequences to the local computer and construct the alignment separately.

7. Download the MUSCLE alignment to the local computer.

BASIC PROTOCOL 2

MULTIPLE SEQUENCE ALIGNMENT ANALYSIS AND EDITING USING Belvu

The accuracy of a phylogenetic tree is dependent on the accuracy of the input multiple sequence alignment (MSA). Alignment masking is often used to increase the phylogenetic signal in the MSA (see Commentary). In practice, masking is accomplished by editing the alignment to be used as the basis for phylogenetic inference. For this task, an alignment viewer/editor is critical. The authors' method for accomplishing this task employs the Belvu alignment editing software from the Karolinska Institute (Stockholm, Sweden) to remove selected columns from the alignment. Other alignment viewer/editors may provide the same functionality, but with a different interface.

Necessary Resources

Hardware

Unix system with X Windows

Software

Alignment editor (e.g., Belvu; see the Support Protocol)

Files

Protein sequence: The multiple sequence alignment in aligned FASTA format (Fig. 6.9.4).

Open the multiple sequence alignment file in Belvu

1. Download and install Belvu (see Support Protocol).

2. To use Belvu on a Unix/Linux machine, type `belvu <MSA>` (where `<MSA>` is the name of the multiple sequence alignment file) at the prompt. This will start up the program and make it possible to view and edit the alignment.

Edit the alignment

The FlowerPower algorithm is designed to restrict clusters to sequences that are globally alignable. Steps 3 to 5 may therefore not be necessary when FlowerPower is used to gather homologs (but should still be confirmed). All manipulations explained below can

```
>gi|2494987|sp|P79292|OPRX_PIG/1-370
----MESLFPAPFWEVLYGSPLQGNLSLLSPNHSLLPPHLLLNASHG--------------------AFLPLGLK-VTIV
GLYLAVCVGGLLGNCLVMYVILRHTKMKTATNIYIFNLALADTAVLLTLPFQGTDVLLGFWPFGNALCKAVIAIDYYNMF
TSAFTLTAMSVDRYVAICHPIRALDVRTSSKAQAVNVAIWALASIVGVPVAIMGSAQVEDEE--IECLVEIPAPQDYWGP
VFA-VCIFLFSFVIPVLIISVCYSLMVRRLRGVRLLS-----------GSREKDRNLRRITRLVLVVVAVFVGCWTPVQV
FVLVQGLGVQP-GSETAVAVLRFCTALGYVNSCLNPILYAFLDENFKACFRKFCCAPT----------------------
------------RRREMQVSDRVRSIA-KDVALACKTSETVPRPA--------------------
>gi|730228|sp|P41144|OPRK_CAVPO/1-380
--MGRRRQGPAQPASELPARN----ACLLPNGSAWLPGWAEPDGNGS----------AGPQDEQLEPAHISPAIP-VIIT
AVYSVVFVVGLVGNSLVMFVIIRYTKMKTATNIYIFNLALADALVTTTMPFQSTVYLMNSWPFGDVLCKIVISIDYYNMF
TSIFTLTMMSVDRYIAVCHPVKALDFRTPLKAKIINICIWLLSSSVGISAIILGGTKVREDVDIIECSLQFPDDDYSWWD
LFMKICVFVFAFVIPVLIIIVCYTLMILRLKSVRLLS-----------GSREKDRNLRRITRLVLVVVAVFIICWTPIHI
FILVEALGSTS-HSTAALSSYYFCIALGYTNSSLNPILYAFLDENFKRCFRDFCFPIK----------------------
------------MRMERQSTSRVRNTV--QDPAYMRNVDGVNKPV--------------------
```

Figure 6.9.4 Aligned FASTA format. The aligned FASTA format displays aligned residues in uppercase and gaps as dashes.

be done using the pull-down menu options at the top of the window, using right-click in Windows or Apple-click on the Macintosh.

3. *Remove any sequences appearing not to be globally alignable with the seed:* Ideally, one wants to exclude any sequence having significant inserts or deletions relative to the seed and to obtain a bidirectional overlap between the seed and included sequences of ≥70% (i.e., at least 70% of the amino acids in each database hit should align to corresponding residues in the seed, and vice versa). Identify sequences having large contiguous inserts (e.g., >50 amino acids) or deletions relative to the seed; these may represent structural domains conferring changes in function relative to the seed. Delete selected sequences using the Edit pull-down menu option "Remove highlighted line." Alternatively, one can select the "Remove many sequences" option and then double-click on each sequence to be removed from the alignment.

4. *Remove any sequences having very low pairwise identity with the seed:* To accomplish this using the Belvu viewer, first select the seed sequence by clicking on its identifier, then select the Sort pull-down menu option "Sort by identity to highlighted sequence." This will sort the sequences in the alignment, placing the seed (or sequences identical to the seed) at the top. A general rule of thumb in predicting homology based on amino acid identity is to require ≥30% identity with the seed over at least 80 amino acids, with lower levels of identity being permitted for longer sequences (Sander and Schneider, 1991). Experienced users may choose to retain more remotely related sequences with lower percent identities and alignment coverage (but caution is advised in these cases). Delete sequences using the Edit pull-down menu as described in step 1.

5. *Remove sequences with no close homologs in the alignment:* Choose the Edit pull-down menu option "Remove outliers." This option removes any sequence having less than a minimum percent identity to at least one other sequence in the MSA. The default minimum percent identity required to retain a sequence in the alignment is 20%, but can be reset by the user (the authors recommend using the default).

6. *Alignment masking—remove highly variable columns and columns containing a significant number of gap characters:* This step is designed to increase the signal-to-noise ratio in the alignment.

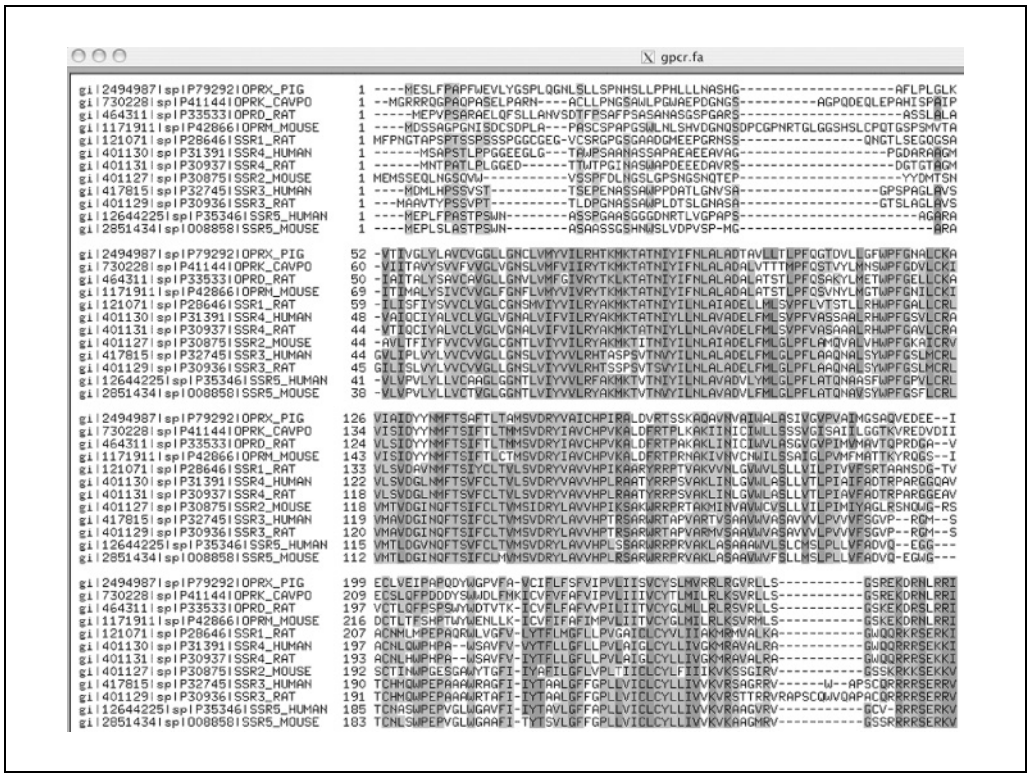

Figure 6.9.5 Belvu display of an MSA. The alignment is displayed using the Belvu wrap-around mode option. This displays a region of variability at the amino-terminus, and another region (at bottom) where only one of the sequences aligns, causing the remaining sequences to have gap characters at these positions. The variable amino-terminus and gappy columns can be deleted during alignment masking, prior to use of the alignment as the basis for phylogenetic tree construction.

a. Remove variable N- and C-terminal regions (Fig. 6.9.5). Variable N- and C-termini can be deleted by selecting an amino acid in the alignment at the edge of a variable region and then using the Edit menu options "Remove columns to the left of cursor (inclusive)" and "Remove columns to the right of cursor (inclusive)."

b. Remove gappy columns. Columns with many gap characters can be removed using the Edit pull-down menu option "Remove gappy columns." The authors recommend setting the threshold to 80%, although lower values (down to 50%) can be used.

c. Remove regions with low sequence similarity. These columns can be identified by visual inspection of the alignment (see step 7 below for coloring scheme), and deleted using the Edit menu option "Remove columns" (setting the start and end indices manually). Alternatively, the user can use the Edit menu option "Remove columns according to conservation," setting the maximum conservation to 0.2 or less.

7. *Remove sequences not matching at family-defining motifs:* The default Belvu alignment coloring uses the Blosum62 scores to color columns; cyan for columns having high Blosum62 scores (i.e., very similar amino acids), dark blue for somewhat less conserved, gray for still less conserved, and no coloring for residues with poor Blosum62 scores (Fig. 6.9.5). Use this coloring to identify key motifs and delete sequences not matching at these motifs. Information from experimental investigation about key residues should be included at this stage.

8. *Save the alignment.* When finished, save the alignment using the File pull-down menu option "Save alignment as..." and select Aligned Fasta. Be sure to give the saved alignment a new name, so that the original alignment will not be overwritten.

DOWNLOADING AND INSTALLING THE Belvu SOFTWARE

SUPPORT PROTOCOL

This protocol describes how to obtain and install the Belvu software used for multiple sequence alignment analysis and editing in Basic Protocol 2.

Necessary Resources

Hardware

Unix system with X Windows

Files

Belvu executables for Linux/Unix platforms Sun, SGI, and Dec Alpha are available via anonymous FTP at *ftp.cgb.ki.se* in the directory `/pub/prog/belvu/`. Documentation for the Belvu software is available at *http://www.cgb.ki.se/cgb/groups/sonnhammer/Belvu.html*.

1. Go to *ftp://ftp.cgb.ki.se/pub/prog/belvu/*.
2. Download file `belvu.LIN_2.28`.
3. Rename file `belvu.LIN_2.28` to `belvu`.
4. Move the belvu file to a location on the local computer's path where executables are stored (e.g., `/usr/bin or /usr/local/bin`).
5. Make sure that the file is an executable by typing:

```
chmod 755 belvu
```

CONSTRUCTING A PHYLOGENETIC TREE USING BETE

BASIC PROTOCOL 3

The authors' recommended approach for this task involves constructing a Neighbor-Joining tree (Saitou and Nei, 1987) and performing bootstrap analysis, as described in *UNIT 6.3*. As an alternative, the following protocol describes the use of the BETE Web server to construct a phylogenetic tree. It is assumed that a large number of proteins (>50) are included in the alignment, so computationally efficient methods for tree construction such as BETE (Sjölander, 1998) or Neighbor-Joining must be employed. For further confidence in an estimated phylogeny, the authors recommend the use of more than one tree method, followed by derivation of a consensus tree from the set of estimated phylogenetic trees (using the Consense software from the PHYLIP suite).

Necessary Resources

Hardware

Any computer with an Internet connection

Software

Web browser

Files

Multiple Sequence Alignment: The alignment produced in Basic Protocol 2 can be used for tree construction. The alignment should be in aligned FASTA format or in SAM A2M format. Examples of alignments in aligned FASTA format and A2M format are presented in Figures 6.9.4 and 6.9.6 respectively.

```
>gi|1083836|pir||A55259
mgrrrqgpaqpas------------ELPARN.ACLLPNGSAWLPGWAEPDGNGSagpqdeqlepAHISPAIPVIITAVYS
VVFVVGLVGNSLVMFVIIRYTKMKTATNIYIFNLALADALVTTTMPFQSTVYLMNSWPFGDVLCKIVISIDYYNMFTSIF
TLTMMSVDRYIAVCHPVKALDFRTPLKAKIINICIWLLSSSVGISAIILGGTKVREDVdiIECSLQFPDDDYSWWDLFMk
ICVFVFAFVIPVLIIIVCYTLMILRLKSVRLLSGSREKDRNLRRITRLVLVVVAVFIICWTPIHIFILVEALGSTSHSTA
ALSSYYFCIALGYTNSSLNPILYAFLDENFKRCFRDFCFPIKMRMERQSTSRVRNTVQDPAYMRNVDGVNKPv
>gi|20379020|gb|AAM21070.1|
.............MESPIQIFRGEPGPTCAPsACLPPNSSAWFPGWAEPDSNGSagsedaqlepAHISPAIPVIITAVYS
VVFVVGLVGNSLVMFVIIRYTKMKTATNIYIFNLALADALVTTTMPFQSTVYLMNSWPFGDVLCKIVISIDYYNMFTSIF
TLTMMSVDRYIAVCHPVKALDFRTPLKAKIINICIWLLSSSVGISAIVLGGTKVREDVdvIECSLQFPDDDYSWWDLFMk
ICVFIFAFVIPVLIIIVCYTLMILRLKSVRLLSGSREKDRNLRRITRLVLVVVAVFVVCWTPIHIFILVEALGSTSHSTA
ALSSYYFCIALGYTNSSLNPILYAFLDENFKRCFRDFCFPLKMRMERQSTSRVRNTVQDPAYLRDIDGMNKPv
```

Figure 6.9.6 UCSC A2M format. The A2M (for "align2model", i.e., the alignment of a sequence to a hidden Markov model) format is designed to indicate the states used by an HMM to generate a sequence. HMM states include Match states (representing the consensus structure for a family), Delete states (used to skip over a consensus position), and Insert states (used to generate additional amino acids not contained in the consensus). The A2M format displays aligned residues in uppercase and "unaligned" characters (emitted in HMM insert states) in lowercase. Gaps in the aligned regions are indicated as dashes (-) and are emitted in HMM skip/delete states. By contrast dots (.) are inserted post hoc in columns containing lowercase letters emitted in Insert state, so that all sequences have the same length.

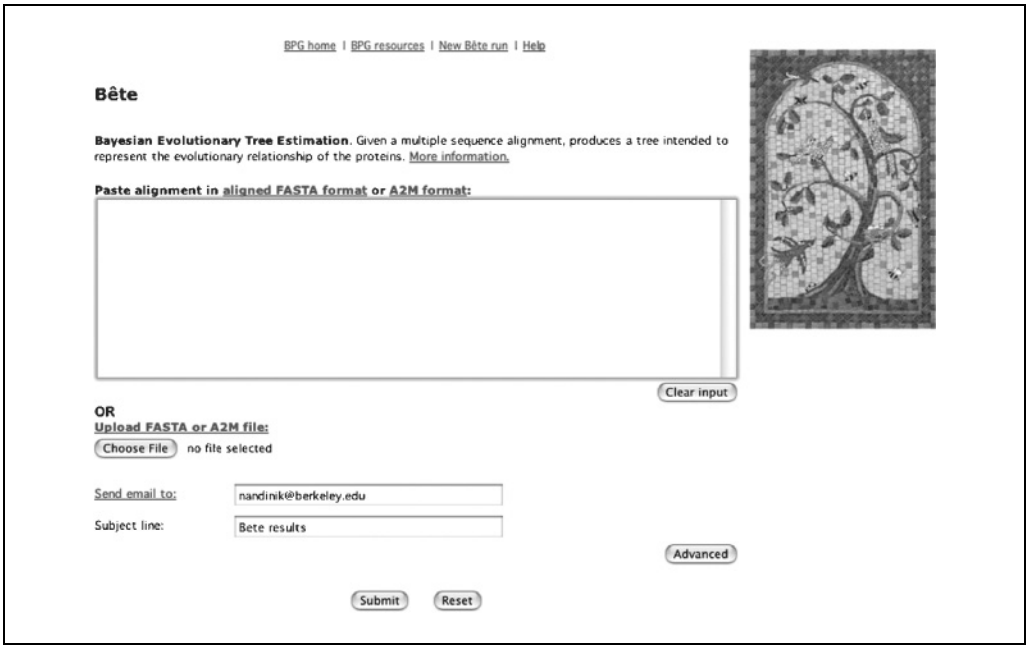

Figure 6.9.7 BETE Web server submission page. The multiple sequence alignment is pasted into the box provided. Results are sent by E-mail.

1. Point the browser to *http://phylogenomics.berkeley.edu/bete* (Fig. 6.9.7).
2. Paste the multiple sequence alignment in the window provided or upload an alignment file (in FASTA or A2M format) by clicking the Browse button.
3. Enter a working E-mail address and click Submit.
4. Retrieve results sent by E-mail. The E-mail body will include a hyperlink for retrieval of results.
5. Download subfamily identification and tree files.

   The results page (Fig. 6.9.8) contains the alignment file separated by subfamilies and a tree file in Newick format (UNIT 6.2), which can be viewed using any tree viewer, or annotated with the TreeNotator utility.

Figure 6.9.8 BETE results page. Download the Newick-format tree for annotation by the TreeNotator software. To view the BETE subfamily decomposition, download the "Alignment organized by subfamily" file.

PHYLOGENOMIC INFERENCE OF MOLECULAR FUNCTION USING TreeNotator

BASIC PROTOCOL 4

Phylogenomic inference of protein function requires a cluster of related proteins with associated experimental and annotation data, along with a phylogenetic tree displaying their predicted evolutionary relationships. To facilitate the tree analysis and prediction of molecular function in a phylogenetic context, the use of the Berkeley Phylogenomics Group TreeNotator Web server is presented. TreeNotator enables phylogenetic tree annotation for proteins drawn from the GenBank and UniProt databases. Annotations are retrieved for sequences in the tree and used to label the leaves; the annotated tree is displayed online using the Java-based ATV software (Zmasek and Eddy, 2001), and can also be downloaded for analysis and display using other tree viewer software tools. Function prediction of a sequence (or sequences) of interest can then be performed by integrating experimental data and annotations in an evolutionary context.

Necessary Resources

Hardware

Any computer with an Internet connection

Software

Web browser

Files

Phylogenetic tree in Newick format (*UNIT 6.2*), generated in Basic Protocol 3. To enable annotation retrieval from sequence databases such as UniProt and GenBank, sequence identifiers must be in a prespecified Newick format. An example Newick file format is shown in Figure 6.9.9.

NOTE: This protocol assumes that sequences in the tree are drawn from either the UniProt database (Apweiler et al., 2004) or the Genbank (Benson et al., 2004) database. If other sequence databases are used as a source for sequences, TreeNotator will not be able to retrieve annotations. In these cases, the authors recommend the use of TreeView (*UNIT 6.2*) to display and print the tree(s), and manual annotation of the trees to predict molecular function.

Display the phylogenetic tree overlaid with annotations

1. Point the browser at the TreeNotator site (*http://phylogenomics.berkeley.edu/treenotator*).

```
(((((((401124:100.0,401125:100.0):78.0,121071:100.0):100.0,((401131:100.0,1351119:100.0):100.0,401130:100.0):100.0):100.
0,((((401128:100.0,401129:100.0):100.0,417815:100.0):100.0,((730838:100.0,2851434:100.0):100.0,12644225:100.0):100.0):89
.0,(((401126:100.0,464813:100.0):85.0,464812:100.0):81.0,(401127:100.0,267008:100.0):100.0):100.0):96.0):100.0,(((((464311:
100.0,417418:100.0):94.0,32363499:100.0):100.0,2494985:100.0):100.0,(((2494986:100.0,6093615:100.0):71.0,(2851402:100.
0,20139232:100.0):100.0):74.0,(1171911:100.0,464314:100.0):100.0):100.0):82.0,(((464313:100.0,464312:100.0):99.0,730229:
100.0):87.0,730228:100.0):100.0):88.0):100.0,((548427:100.0,548426:100.0):94.0,1352647:100.0):48.0):54.0,2494987:100.0):1
00.0,730230:100.0);
```

Figure 6.9.9 Newick file format. The Newick format is a standard tree file format readable by most tree viewers. The format uses nested parentheses to indicate the join order of the nodes. In the above example, the sequences in the tree are represented by their GenBank identifier, and the bootstrap values are indicated at each node. For example `((401124:100.0,401125:100.0):78.0,121071:100.0)` means that the sequences 401124 and 401125 are joined at a node with bootstrap value of 78.0, and this node is joined to sequence 121071 with a bootstrap value of 100.

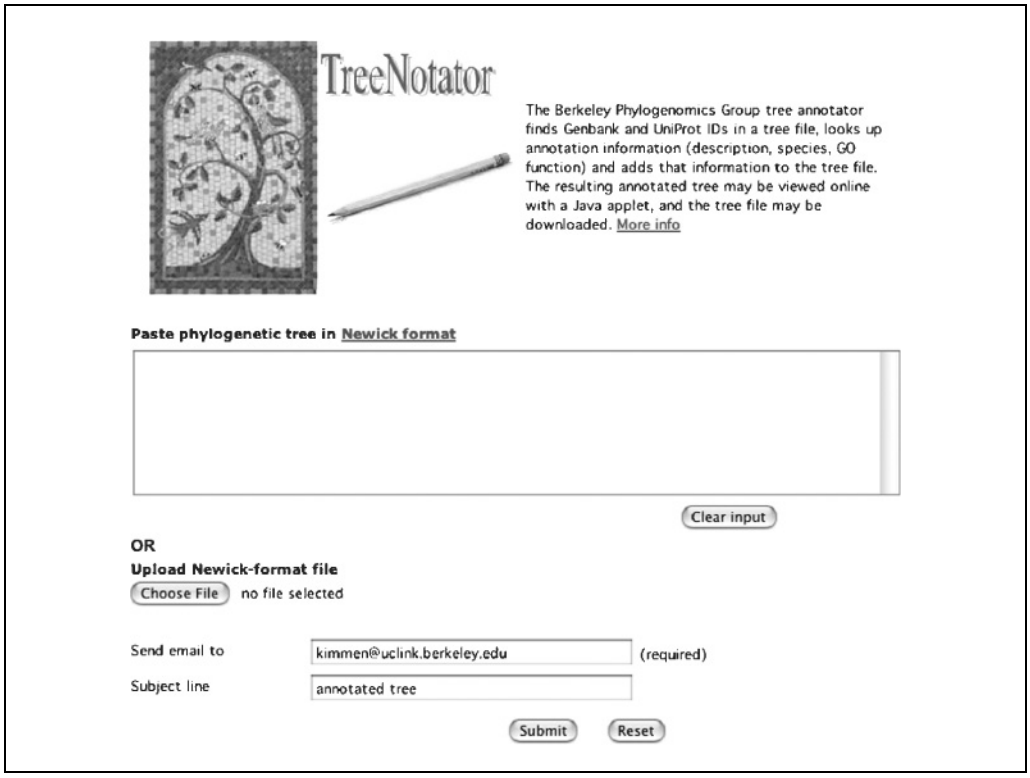

Figure 6.9.10 TreeNotator submission page. The tree is pasted in the input box in Newick format. Results are returned by E-mail to the address provided.

2. Paste the tree file in Newick format in the box provided or upload a tree file by clicking the "Choose file" button (Fig. 6.9.10).
3. Enter a working E-mail address as directed. Results will be sent to this address.
4. Click Submit.
5. Retrieve results by clicking on the hyperlink provided in the E-mail. The TreeNotator results Web page (Fig. 6.9.11) provides a tool to view the tree with annotations using the ATV software. It is also possible to download annotated tree files for use with other tree-viewing software applications such as TreeView (*UNIT 6.2*).

The TreeNotator display

6. Click on View Tree in the results page. This will open a window that displays the tree and sequence annotations using the ATV software (Fig. 6.9.12). It may be necessary

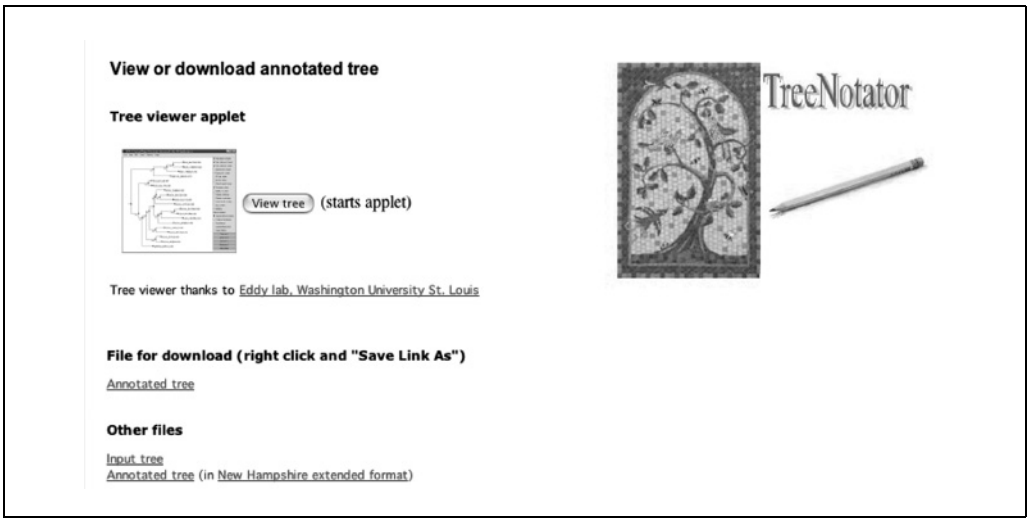

Figure 6.9.11 TreeNotator results page. Click on the "View tree" button to display the annotated tree using the ATV software, or download the annotated tree (immediately under "File for download") for display using other phylogenetic tree visualization software.

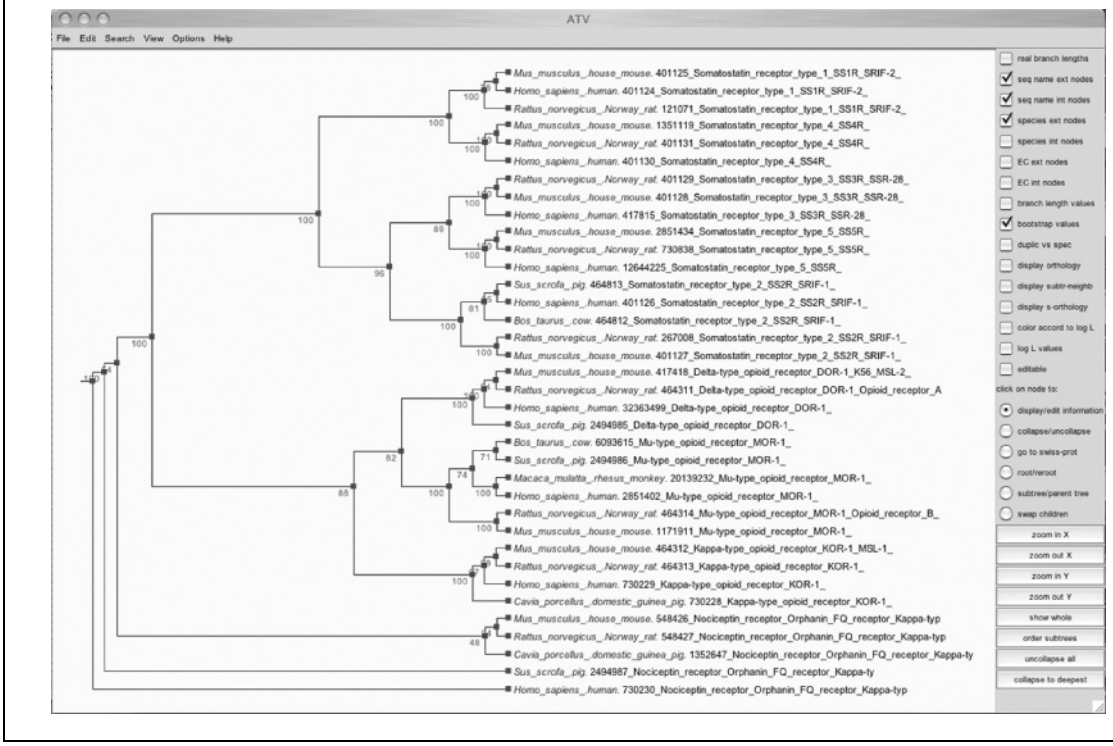

Figure 6.9.12 Annotated tree displayed with ATV viewer. For each sequence, the following data are retrieved from the corresponding database (UniProt or GenBank) and displayed at the leaves: Species, GenBank, or UniProt idenitifier and the annotation. The bootstrap values are displayed in green at internal nodes.

to resize the window to display the tree topology more accurately, as the initial display tends to compress the tree.

7. Options for manipulating the tree display are provided in the panel on the right. Although there are several options, only those relevant for this unit are discussed. The user is encouraged to explore the other options.

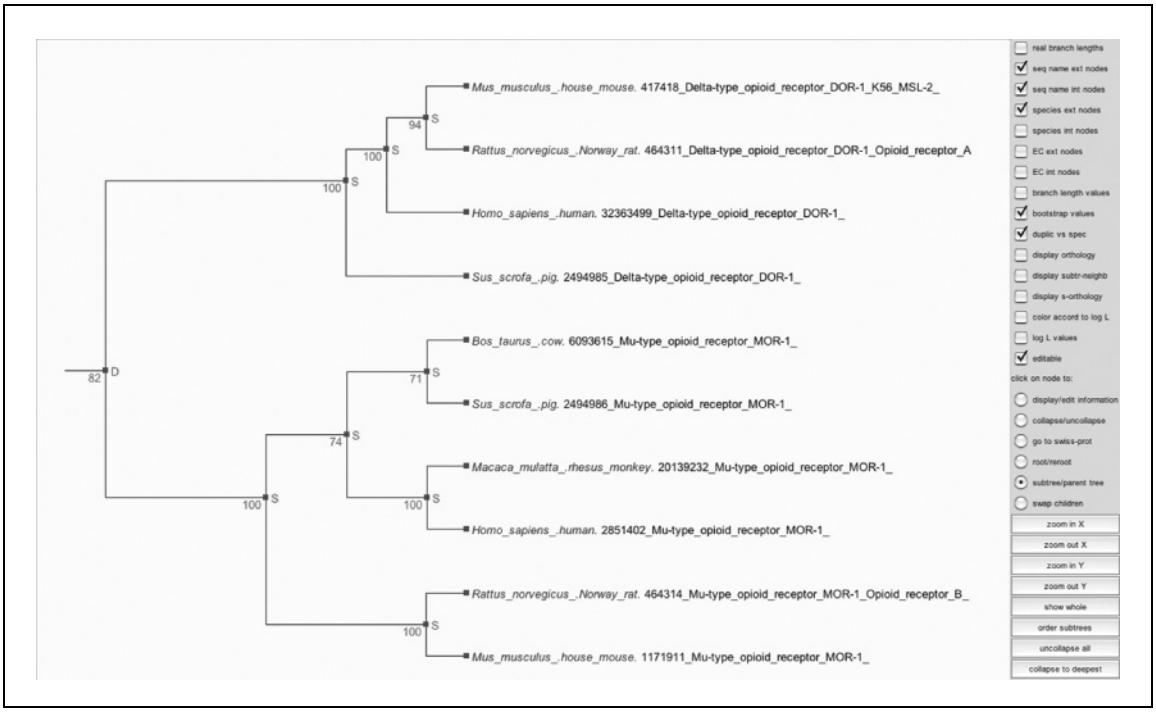

Figure 6.9.13 In the tree shown above, the root is labeled D to indicate a duplication event, joining two subtrees, each of which contains a group of putative orthologs. All other internal nodes are labeled S, to indicate speciation events. Nodes for subtrees containing only one representative of a species can safely be labeled with an S. While labeling nodes as indicative of speciation is straightforward, labeling nodes as indicative of duplication is less so, since multiple protein sequences encoded by the same gene can be present in a dataset (e.g., due to splice variants or simple database duplicates).

8. All the following steps are to be performed using the options panel on the right.

   a. If a large number of sequences are used, the initial display may be compressed vertically, so that sequence identifiers on the leaves are not legible and subtree structure is not obvious. To view subtrees clearly, click on "zoom in Y." To show the full tree, click "show whole."

   b. To display the organism name, select "species ext nodes."

   c. To display the sequence ID and annotation, select "seq name ext nodes."

   d. To display bootstrap values, select "bootstrap values." To hide bootstrap values, deselect this box. Bootstrapping is a popular resampling method and bootstrap values indicate statistical support for subtrees in the MSA used as input. For further explanation of bootstrapping, see *UNIT 6.1*.

   e. To edit tree information, select "editable" and click on the node to edit. This will open a separate window with options for annotating the node with Name, Taxonomy ID, EC number, etc.

   f. To restrict the display to a subtree, select "subtree/parent tree" and click on the node of interest. This will display only the selected node and the subtree descending from that node.

Label tree nodes as indicating gene duplication or speciation

Label subtree nodes as indicating either speciation or duplication events, based on analysis of the tree topology (Fig. 6.9.13). To label nodes as indicating speciation or duplication events, select "editable," click on the node of interest, and, in the window that opens, select "duplication" or "speciation" or "not assigned," click Write to Tree, and close the

window. To display this on the tree, select "duplic vs spec." This will display duplication at a node with a "D" (in red) and speciation with an "S" (also shown in red).

9. To revert to the original tree and discard any changes to the tree display, click File in the menu bar on the upper left corner and select Reload.

Identifying functionally consistent subtrees

10. Identify three sets of "basis" subtrees:

    a. Inspect the phylogenetic tree to identify subtrees containing orthologous proteins. Nodes whose subtrees contain only single representatives from individual species can be labeled as indicative of speciation; the sequences in these subtrees are putative orthologs. Subtrees containing multiple (nonidentical) sequences from the same species should not be considered orthologous. (See Figs. 6.9.13 and 6.9.14).

    b. Inspect the phylogenetic tree to identify subtrees with consistent annotation. Many sequences may have uninformative annotations (e.g., hypothetical, unknown), while others may have simply been annotated by a hit to a PFAM HMM or some other domain analysis (e.g., a match to a Conserved Domain in the NCBI CDD, or to a COG). Attempt to identify any sequence with experimental support for the assigned function.

    c. If bootstrap values have been computed, identify subtrees with significant bootstrap support. A good rule of thumb for this is 70%.

11. The three "basis" subtrees identified in step 10 (groups a, b and c) should now be examined. Sequences in these "basis" subtrees can be assumed to share a similar function, with orthologous groups (group a) having the highest likelihood of functional similarity, followed by consistently labeled subtrees having high bootstrap support (the intersection of groups b and c).

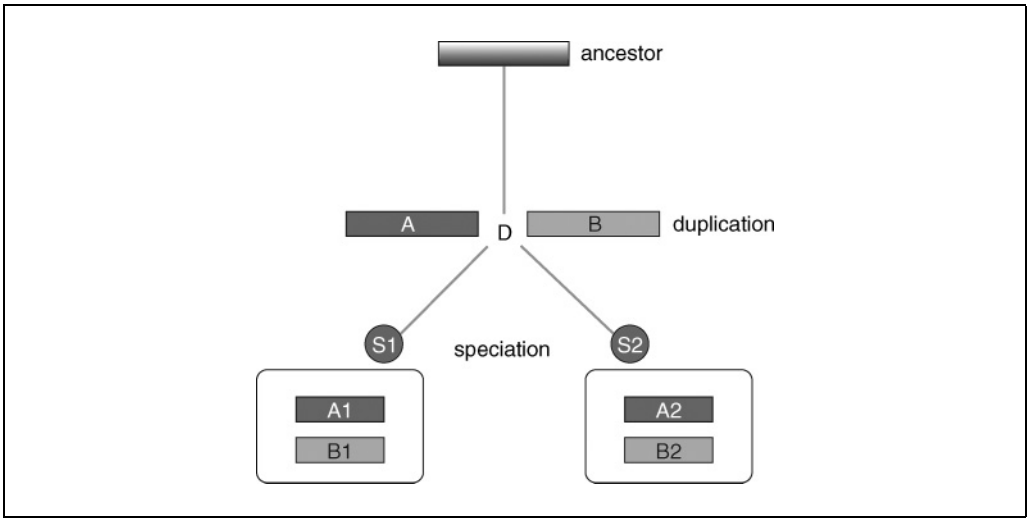

Figure 6.9.14 Discriminating orthologs and paralogs. This figure represents the evolution of a protein family through the joint processes of gene duplication and speciation. The ancestral gene, shown at top, undergoes duplication in the ancestral genome (at the node labeled D), producing two paralogous genes, A and B. A subsequent speciation event produces two species, S1 and S2, each having a copy of the A and B genes. A1 and B1 are clearly paralogs, as are A2 and B2. If the definition of ortholog used is "the same gene in different species," then according to this tree, A1 and A2 are clearly orthologs, as are B1 and B2. By contrast, it cannot be asserted that A1 and B2 are orthologs, nor can it be asserted that B1 and A2 are orthologs. The A and B genes are related by a duplication event in the ancestral genome where they are paralogous. This figure is adapted from Koonin (2001).

Infer molecular function

12. Label these basis subtrees with the molecular function assigned to sequences in the subtree (assuming credible experimental evidence supports the assigned function) using the "editable" feature in the ATV software as explained above (in step 8e, above). Sequences in these subtrees can then be assigned as potentially having the molecular function of their subtree neighbors.

COMMENTARY

Background Information

Clustering homologs for phylogenomic inference

Several factors are critical when clustering homologs for phylogenomic inference. The most important point concerns whether the phylogenomic inference is being performed on the whole-chain level (i.e., including all domains in all proteins gathered) or only along a single conserved domain found in proteins with otherwise divergent folds. This distinction is critical for accurate prediction of molecular function, due to the impact of domain fusion and fission processes in the evolution of protein families.

Domain shuffling in protein families requires special attention to avoid errors in prediction by homology. This is particularly problematic, as standard methods of homolog detection typically ignore whether two proteins align globally or only locally. For example, plant receptor-like kinases (RLKs) are composed of extracellular leucine-rich repeat (LRR) regions followed by a transmembrane domain and a cytoplasmic kinase domain. Receptor-like proteins (RLPs) are very similar to RLKs, but lack the cytoplasmic kinase domain. Polygalacturonase-inhibiting proteins (PGIPs) are similar to both RLPs and RLKs in containing an extracellular LRR region, but lack the transmembrane and cytoplasmic domains. PGIPs, RLPs, and RLKs have distinct molecular functions and overall architectures, but each has significant sequence similarity to the others along their LRR region, which can be hundreds of amino acids long.

The protocols presented here assume that a whole-chain analysis is being performed, and FlowerPower default parameters are designed to match this assumption. If one prefers to perform a phylogenomic analysis on the domain level, one can either adjust the FlowerPower options (using the Advanced settings page) or use another method, such as PSI-BLAST, to retrieve homologs.

Ideally, the set of proteins gathered should include all identifiable homologs sharing the same overall architecture. Practically speaking, this can be difficult. First, the overall architecture (i.e., sequence of structural domains) is known for only a small fraction of proteins; differentiating globally homologous proteins from those sharing only a local similarity is not always straightforward. Second, most bioinformatics methods for identifying proteins homologous to a query (or "seed") protein will identify all proteins sharing any significant *local* similarity to the query. This is true for the pairwise methods such as BLAST (Altschul et al., 1990), as well as the iterative profile methods such as PSI-BLAST (Altschul et al., 1997). This stage therefore requires the investigator to combine homolog identification and analysis in order to produce a final set of reliable homologs.

The FlowerPower algorithm is designed explicitly for phylogenomic analysis of protein molecular function. Like PSI-BLAST, FlowerPower employs an iterative approach to clustering, but instead of using a single HMM or profile to expand the cluster, it identifies subfamilies using the BETE algorithm and then selects and aligns new homologs using HMMs constructed for each subfamily (Brown et al., 2005). Subfamily HMMs compete for new sequences, enabling improved alignment quality (particularly in regions of structural variability across the family as a whole), and prevent drifting of the profile away from the seed sequence (profile drift). This is accomplished by the persistent representation of the seed sequence (and close homologs) by subfamily HMMs. FlowerPower employs alignment analysis following each iteration to reduce the intrusion of non-homologs.

Multiple sequence alignment construction and analysis

The accuracy of a phylogenetic tree depends on the accuracy of the multiple sequence alignment on which it is based. Studies have shown that multiple sequence alignment methods are sensitive to attributes characterizing many protein superfamilies—i.e., large numbers (in the hundreds and thousands) of

sequences, high sequence variability, and length differences (McClure et al., 1994; Thompson et al., 1999). To the degree that phylogenomic analysis is restricted to reasonable numbers of closely related sequences, alignment accuracy can be expected to be high, with corresponding increased likelihood of accuracy in the resulting tree topology. However, when large numbers of divergent sequences are included in a cluster, alignment and tree topology accuracy can be expected to decrease. The protocol presented here includes the use of the MUSCLE multiple sequence alignment program to align the proteins retrieved by FlowerPower.

Of the available multiple sequence alignment software tools, the MUSCLE algorithm is computationally efficient and has solid performance on a number of benchmark datasets (Edgar, 2004). Users who have elected to use a different program to align their sequences should download the unaligned sequences instead of the MUSCLE realignment, and then use the program of choice to construct the alignment.

Masking the multiple sequence alignment

One of the fundamental assumptions of phylogenetic inference is positional homology, i.e., all the residues in a column of an MSA descend from a residue in the ancestral protein. However, protein sequence and structural variability can cause this assumption to be violated. Alignment *masking* is designed to maximize the phylogenetic signal in the multiple sequence alignment by deleting (or downweighting) columns corresponding to either misaligned or structurally divergent regions in the molecules. These high-structural-divergence regions are more often found at the amino- and carboxy-terminus of the proteins, but may also be found at the exposed surface of the proteins (in particular, in loop regions) between conserved blocks. Two basic approaches to alignment masking have been proposed. The first approach (presented here) involves deleting columns that appear unreliable or that contain large numbers of gap characters. The multiple sequence alignment created is analyzed to discriminate the conserved structural features of the family as a whole. This analysis is used to determine which regions of the alignment should be used for phylogenetic tree construction and to identify and remove potentially spurious database hits. Both alignment rows and columns can be edited independently to produce a final multiple sequence alignment containing only the conserved structural features of highly credible homologs. An alternative approach (Wheeler et al., 1995) involves the construction of a concatenated superalignment of several separate multiple alignments, varying parameters and using the concatenated alignment as input to phylogenetic tree construction.

Tree construction and analysis

As noted earlier, the accuracy of a phylogenomic analysis of protein function depends on the accuracy of phylogenetic tree topology. Unfortunately, ensuring phylogenetic tree accuracy is far from straightforward. Detailed phylogenomic analyses of protein families using different tree-estimation tools have shown a lack of consistency across methods (see, for example, Citerne et al., 2003). The lack of robustness of phylogenetic tree estimation under different conditions is supported by simulation studies (Felsenstein, 1988; Hasegawa and Fujiwara, 1993; Kuhner and Felsenstein, 1994), which show that most phylogenetic inference methods are sensitive to conditions that are commonly observed in protein superfamilies, e.g., large numbers of sequences, high levels of divergence, lack of positional homology, and nonuniformity of conservation across character states (columns) in the alignment. Importantly, while some methods are more robust with respect to these issues than others, no method is robust under all circumstances.

The potential uncertainty in a phylogenetic tree topology can be identified in several ways. Bootstrap analysis is traditionally performed to identify subtrees with good statistical support. This analysis can be supplemented through the use of different phylogenetic tree-estimation software tools and/or different multiple sequence alignments as input. The different topologies produced by these different methods and alignments will support a different set of hypotheses about the relationships between proteins and their molecular functions. Because of the inherent inconsistencies between tree-construction methods and their dependencies on the input multiple sequence alignment, the authors of this unit recommend that biologists use a combination of different alignment and tree-construction methods, followed by the identification of a consensus tree topology across all the derived trees.

It is assumed that a large number (>50) of proteins are included in the alignment, so that computationally efficient methods for tree construction such as BETE (Sjölander, 1998) or Neighbor-Joining (Saitou and Nei, 1987)

must be employed. If fewer sequences are being analyzed, users can employ more computationally intensive approaches. Primary sources of phylogenetic tree construction software include the PHYLIP Web site (*http://evolution.genetics.washington.edu/phylip.html*), MrBayes (Huelsenbeck and Ronquist, 2001), and PAUP (Swofford, 2002; *UNIT 6.4*). Additional information on these programs can be found in Chapter 6 of this manual (Inferring Evolutionary Relationships).

This unit recommends the use of Neighbor-Joining (NJ) and bootstrap analysis to construct a phylogenetic tree and identify subtrees with statistical support. Neighbor-Joining is one of the most commonly used methods for phylogenetic tree construction. Its computational efficiency enables NJ to be used to estimate phylogenies for large datasets and for estimation of bootstrap support. The alternative to this is the use of the Bayesian Evolutionary Tree (BETE) algorithm (Basic Protocol 3). The main advantage of BETE for phylogenomic analysis lies in its automatic prediction of functional subfamilies. BETE estimates a phylogenetic tree and automatically identifies functional subfamilies (Sjölander, 1998) given an input multiple sequence alignment. BETE uses relative entropy between profiles constructed using Dirichlet mixture densities (Sjölander et al., 1996) to build a tree, and minimum encoding cost principles to cut the tree into subtrees to define the subfamily decomposition. BETE subfamilies often correspond to orthologous groups, but may instead contain ultra-paralogs from the same species, or, simply, very similar proteins (including both paralogs and orthologs).

Using phylogenetic trees to predict function

As described earlier, gene-duplication events produce families of related proteins having different functional specificities. Phylogenomic inference is designed to enable users to differentiate orthologs (related by speciation events) from paralogs (genes in the same species related by duplication events). Orthologs are expected to share a common function (although this is not always the case), while paralogs are expected to be somewhat more divergent in function.

Two new methods for differentiating orthologs and paralogs have been developed very recently: Resampled Inference of Orthologs (RIO; Zmasek and Eddy, 2002) and Orthostrapper (Storm and Sonnhammer, 2002). Both methods involve analysis of a phylogenetic tree to label nodes as indicative of either speciation or duplication events.

Complicating matters, orthology does not always suffice for correct inference of molecular function (Eisen and Wu, 2002; Zmasek and Eddy, 2002). Paralogs can occasionally have greater functional similarity than orthologs, particularly in cases where the gene duplications are recent, or in the case of subfunctionalization (e.g., paralogs may perform the same function but specialize for different tissue types). Similarly, orthologous genes in distantly related species might perform different functions. The authors look for consistent annotation within subtrees (an approach called *subtree neighbors*; Zmasek and Eddy, 2002). When available, relevant biological information should be incorporated along with annotations. The use of the term "ortholog" differs among investigators in the field; some use this term to refer to homologous genes in different species, whereas others restrict the term ortholog to indicate only those genes or proteins in different organisms that are clearly "the same gene in different species" and that can thus be assumed to share an identical molecular function (Koonin, 2001). See Figure 6.9.14 for an illustration of these issues.

In practice, it can be very difficult to unambiguously label subtree nodes as corresponding to speciation or duplication events. The presence of multiple copies of the same sequence in the database, as well as allelic or splice variants of the same gene, can be interpreted incorrectly as entirely different genes and would consequently appear to be paralogs in the tree. If it is not possible to remove any such sequences from the alignment used to estimate the tree, detailed scrutiny of the tree will be necessary to avoid errors. For this reason, it may be simpler to use a "subtree neighbors" approach when predicting molecular function. An automatic tool for identifying putative orthologs is the Orthostrapper program, available at *http://orthostrapper.cgb.ki.se/* (Hollich et al., 2002).

The accuracy of sequence annotations should be confirmed. A large fraction of sequence annotations are based on homology and may be suspect. The UniProt database includes information about the annotation source, using Evidence Codes (e.g., Traceable Author Statement (TAS), for annotations supported by scientific publications). The authors strongly recommend checking biological literature for descriptions of experiments

performed for the original assignment of molecular function, as not all experiments are equally well designed.

Critical Parameters and Troubleshooting

Clustering and alignment

The sequences included by FlowerPower are selected according to the user-specified parameter settings. The default settings (particularly for fractional coverage of the seed sequence and of database hits) have been determined empirically and are designed to maximize inclusion of homologous proteins with identical architectures while reducing the number of sequences with different architectures. Occasionally, these default parameters may be overly conservative, resulting in the retrieval of very few sequences. There are two ways to overcome this problem. To include additional homologs, one can either change the default database for homolog retrieval from, the smaller but higher-quality UniProt database to NR, or adjust the inclusion parameters to make them less restrictive (e.g., by reducing the query and hit coverage parameters). In the latter case, it will be necessary to pay special attention to the alignment-editing step, to ensure that any non-homologs are rejected from the alignment used as the basis for tree construction.

As an alternative to FlowerPower, biologists can use BLAST or PSI-BLAST to retrieve sequences and then manually remove sequences with significant length differences. A multiple sequence alignment for the remaining sequences can be derived using the alignment program of choice. This alignment will need to be examined to confirm that all sequences included in the cluster can be aligned along their entire lengths (i.e., that all sequences appear to have the same domain architecture and are globally alignable).

Alignment masking

Protein superfamilies can have regions of structural divergence that introduce noise in a phylogenetic estimation. To improve the expected accuracy of the phylogenetic tree, these regions should be identified and removed (or masked) prior to phylogenetic tree estimation. A detailed discussion of the various masking methods used is beyond the scope of this unit. The authors recommend that users construct and compare several phylogenetic trees using different masking protocols. Recommended masking protocols include: (i) removing columns with many gaps (e.g., where $<30\%$ of sequences align) and (ii) removing columns with average pairwise Blosum62 scores below zero. Trees estimated using masked alignments as input should be compared with trees based on unmasked alignments.

Tree construction and analysis

One of the seldom-noted issues in phylogenetic reconstruction involves the uncertainty of a phylogenetic analysis. If the same alignment is used as input to three different tree-estimation programs, three (or even more) distinct trees can be produced. Small changes to the alignment algorithm, or different methods for masking a multiple sequence alignment, can also produce unexpected changes in the tree topology.

For these reasons, it is strongly encouraged that users construct phylogenetic trees via two or more distinct methods, rather than rely on a single method. For even greater robustness in phylogenetic reconstruction, the authors of this unit recommend constructing several multiple alignments using different alignment methods and using each (following judicious alignment masking) as input to the phylogenetic tree methods selected. To avoid the inherent bias of the guide tree used by most progressive methods, it is recommended that the alignment methods selected include at least one iterative (i.e., nonprogressive) approach. The final set of trees can be used to derive a consensus tree using the PHYLIP Consense software (see *UNIT 6.3*).

The authors strongly encourage users to retrieve experimentally confirmed data on molecular function, interactions, biological processes, cellular localization, and other aspects of a protein's overall function and role. These data are often difficult to separate from annotations supplied entirely by homology, which are likely to contain errors.

Users of the PHYLIP software should take note of the following caution. Sequence IDs are often truncated by PHYLIP to the first 10 characters following the > sign in the FASTA sequence. Truncated IDs will cause errors in TreeNotator. To avoid these errors, the authors recommend that the tree file be manually edited to repair any truncated IDs, or that a different tree-reconstruction program be used.

Acknowledgments

This work was supported in part by Grant #0238311 from the National Science Foundation and by Grant #R01 HG002769-01 from the National Institutes of Health.

Literature Cited

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl. Acids Res.* 25:3389-3402.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., and Yeh, L.S. 2004. UniProt: The Universal Protein knowledgebase. *Nucl. Acids Res.* 32:D115-D119.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2004. GenBank: Update. *Nucl. Acids Res.* 32:D23-D26.

Bork, P. and Koonin, E.V. 1998. Predicting functions from protein sequences—where are the bottlenecks? *Nature Genet.* 18:313-318.

Brenner, S.E. 1999. Errors in genome annotation. *Trends Genet.* 15:132-133.

Brown, D., Krishnamurthy, N., Dale, J.M., Christopher, W., and Sjölander, K. 2005. Subfamily HMMs in functional genomics. *Pac. Symp. Biocomput.* 322-333.

Citerne, H.L., Luo, D., Pennington, R.T., Coen, E., and Cronk, Q.C. 2003. A phylogenomic investigation of CYCLOIDEA-like TCP genes in the Leguminosae. *Plant Physiol.* 131:1042-1053.

Devos, D. and Valencia, A. 2001. Intrinsic errors in genome annotation. *Trends Genet.* 17:429-431.

Doolittle, R.F. 1995. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* 64:287-314.

Doolittle, R.F. and Bork, P. 1993. Evolutionarily mobile modules in proteins. *Sci. Am.* 269:50-56.

Edgar, R.C. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.

Eisen, J.A. 1998. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8:163-167.

Eisen, J.A. and Wu, M. 2002. Phylogenetic analysis and gene functional predictions: Phylogenomics in action. *Theor. Popul. Biol.* 61:481-487.

Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. *Annu. Rev. Genet.* 22:521-565.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19:99-113.

Galperin, M.Y. and Koonin, E.V. 1998. Sources of systematic error in functional annotation of genomes: Domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol.* 1:55-67.

Gerlt, J.A. and Babbitt, P.C. 2001. Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.* 70:209-246.

Gilks, W.R., Audit, B., De Angelis, D., Tsoka, S., and Ouzounis, C.A. 2002. Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* 18:1641-1649.

Hasegawa, M. and Fujiwara, M. 1993. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol. Phylogenet. Evol.* 2:1-5.

Hollich, V., Storm, C.E., and Sonnhammer, E.L. 2002. OrthoGUI: Graphical presentation of Orthostrapper results. *Bioinformatics* 18:1272-1273.

Huelsenbeck, J.P. and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754-755.

Koonin, E.V. 2001. An apology for orthologs—or brave new memes. *Genome Biol.* 2:COMMENT1005.

Kuhner, M.K. and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11:459-468.

McClure, M.A., Vasi, T.K., and Fitch, W.M. 1994. Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.* 11:571-592.

Sander, C. and Schneider, R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9:56-68.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-425.

Sjölander, K. 1998. Phylogenetic inference in protein superfamilies: Analysis of SH2 domains. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6:165-174.

Sjölander, K. 2004. Phylogenomic inference of protein molecular function: Advances and challenges. *Bioinformatics* 20:170-179.

Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S., and Haussler, D. 1996. Dirichlet mixtures: A method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.* 12:327-345.

Storm, C.E. and Sonnhammer, E.L. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 18:92-99.

Swofford, D. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Mass.

Thompson, J.D., Plewniak, F., and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucl. Acids Res.* 27:2682-2690.

Wheeler, W.C., Gatesy, J., and DeSalle, R. 1995. Elision: A method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. *Mol. Phylogenet. Evol.* 4:1-9.

Zmasek, C.M. and Eddy, S.R. 2001. ATV: Display and manipulation of annotated phylogenetic trees. *Bioinformatics* 17:383-384.

Zmasek, C.M. and Eddy, S.R. 2002. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 16:3(1):14.

Key References

Bork and Koonin, 1998. See above.

The authors of this paper identify common problems associated with function prediction by homology and present ways to avoid these errors.

Eisen, 1998. See above.

Jonathan Eisen's cogent presentation of the raison d'etre behind phylogenomic analysis for improving prediction of gene function.

Sjölander, 2004. See above.

A detailed view of the challenges in phylogenomic analysis, with a description of new methods for key tasks in a phylogenomic pipeline.

Internet Resources

http://phylogenomics.berkeley.edu/resources

The BPG resources Web site includes a variety of user-friendly resources for phylogenomic inference of protein molecular function. A description of all the available tools can also be found on the Web site.

Contributed by Nandini Krishnamurthy and Kimmen Sjölander
University of California
Berkeley, California