Phylogenetically-Informed Annotation (PIA)

Introducting PIA

We developed a new method for rapidly identifying light-interacting genes in transcriptomes. These genes include those involved in phototransduction, eye development, pigment synthesis, circadian cycles, and other light-interacting pathways. Our methods are optimized for detecting genes in metazoans, but could be adapted for annotating transcriptomes from non-metazoans.

For annotation, we use BLAST to search transcriptomes or gene models from full genomes for 109 separate genes known to be involved with light interactions in metazoans. We then use a likelihood algorithm to place the BLAST hits for each gene on to a corresponding, pre-calculated gene tree (we have separate trees for each of our 109 genes). As a method for assigning identities to transcripts, we find that PIA is more objective than methods based on similarity alone (*e.g.* Blast2Go) and more efficient than past methods based on constructing gene trees.

Getting started

1) Visit: http://galaxy-dev.cnsi.ucsb.edu/pia/

2) Go to the toolbar along the top of the window and click on "User". Then select "Register" from the pull-down menu. Enter your email address, a password, and a username.

3) Go to the "History" toolbar on the right side of the window and click on the gear-shaped icon. Then, from the pull-down menu, select "Create New". This will get you started on a new history in Galaxy.

Running PIA

1) To upload a transcriptome, go to the "Tools" toolbar on the left side of the window. Select "Get Data" and then "Upload File". You will then have the option of uploading a data set from either a file on your computer or from a specific URL.

2) PIA does not annotate transcriptomes directly, but instead annotates protein sequences predicted from transcriptomes. To generate predicted protein sequences from your transcriptome, run the "Get open reading frames (ORFs)" under the Analyze Data option in the Tools menu.

3) To search for genes to annotate within your data set, select the "pia" tool under Analyze data. You have the option of searching transcriptomes for single genes (*e.g.* opsins) or functional gene sets (*e.g.* a dozen genes involved in rhabdomeric-type phototransduction).You can also adjust the e-value cut-off for the BLAST search that will return hits from your transcriptome. Lower e-values will return fewer possible genes than higher ones. Finally, you can adjust the maximum number of hits returned by each BLAST search. Generally, we suggest starting with an e-value cut-off of 1e-20 and retain the top 3 hits. These parameters return quite similar genes and we encourage users to try out a variety of search settings to see how this affects results.

4) The "pia" tool will give several different output files in your history. The "all genes hit" file contains the sequences of all genes recovered by the initial BLAST searches. The "name tab tree" file has all of your hits placed on to their corresponding gene trees. You can view these trees as a PDF by running the "tab2trees" tool under Analyze Data and selecting "name tab tree" as the input.

5) You also have the option of running workflows. Published workflows may be found under the "Shared Data" menu at the top of the Galaxy window. Click on "Published Workflows", then select the workflow of your choice and import it into your current history. A helpful feature provided by these workflows is that they will return hits from your searches in both amino acid and nucleotide formats.

Interpreting results

The output of PIA is a PDF document in which hits for a selected set of light-interacting genes have been placed on to pre-calculated gene trees using a likelihood-based algorithm. The sequences that comprise each pre-calculated tree come from predicted protein databases associated with the complete genomes of about 30 different taxa. The trees also include "Landmarks", which are genes that have been well-characterized functionally. Sequences marked "LANDMARK1" – highlighted with red squares – represent the genes whose orthologs we are seeking. The genes marked "LANDMARK2" are also well-characterized, but are not orthologous to the ones in which we are interested. The trees also include "Queries" – marked by yellow circles – which are the BLAST hits pulled from your input data set.

A promising hit from your BLAST search is one that represents an ortholog of a light-interacting gene whose function has been established previously (see sequences labeled "LANDMARK1"). Promising hits tend to fall on short branches in phylogenetic positions that make sense given established relationships between species. It is important to remember that PIA will give you hits that represent genes that are not orthologs of those for which you are searching. PIA works well as a filter, but there is some subjectivity in deciding which BLAST hits are orthologs of particular genes of interest. As noted above, branch length and phylogenetic placement are important considerations. We also suggest BLASTing all promising hits back to GenBank and runny other analyses – contamination is always a possibility.

If you are interested in searching for genes that are not included in our initial set of 109 trees, please contact the authors. We plan to expand our list of available gene trees in the near future.