Tanveer Salim

MCDB 187AL

April 26, 2018

Methods

Constructing a Gene Model with Apollo

To determine gene models, the mRNA reference sequence was consulted using Apollo. Within the chr8 700,001..795,000 genomic region, the “Unique and MultiMapping FPM” was selected from the “Tracks” menu to display The Valley Oak genomic RNA reference sequence. Peaks on this reference map indicated the frequency at which the exon sequence was expressed in Valley Oak. To construct gene models, “IsoSeqTopTier” was selected from “Tracks” to display pieces of transcripts. Transcripts in the “IsoSeqTopTier” space that aligned with the RNA reference sequence in the “Unique and MultiMapping FPM” were dragged to the “User-Created Annotations” space and then merged together. Next, the merged gene model was right-clicked and the option for “Gene Sequence” was selected. If there were gaps in the constructed gene model, then first, the genomic sequence was selected, copied, and then pasted into the Augustus gene prediction tool. Augustus used an algorithm to predict its own gene model based on the genomic sequence. Augustus’s gene model was then compared to the constructed gene model for precision. If there were sequences in the gene model made by Augustus that patched the gaps in the Apollo constructed gene model, such edits were made to the gene model on Apollo accordingly. The Peptide sequence icon was selected, giving the peptide sequence of the gene model (Lewis et al 2002).

Searching for Homologous Sequences Using BLAST

This peptide sequence was pasted into NCBI’s Protein BLAST to search for conserved homologous sequences from other species. Photos of the Color Key, Putative Conserved Domains, and Top six Query sequence matches each from six unique species were taken. The Color Key is a tool to determine if the Apollo gene model is constructed correctly. The more complete the bandwith of a sequence is relative to subject homologous sequences, the better the indication that the gene model is complete. Secondly, the more red the bands of the Apollo gene model and the subject sequences are, the higher the quality of alignment. This is a second indication that the constructed Apollo gene model is complete. The Putative Conserved Domains page displays the conserved domain that is most likely the functional domain the Apollo constructed gene model belongs to. Query sequences were another means to check accuracy of constructed Apollo gene model. If the top six Query sequence results from six unique species do not match the input sequence, the Apollo gene model was most likely inaccurate (Castresana et al 2007).

Using Clustal Omega to Check Accuracy of Gene Model

The FASTA peptide sequences from these six unique species were pasted to Clustal Omega along with the peptide sequence of the constructed gene model. Clustal Omega generated a Multiple Sequence Alignment Page that was used to compare conservation of peptide sequence amongst the seven total species. Asterisks were drawn beneath the columns of sequences of the seven species when there was perfect conservation. The more asterisks available in the Multiple Sequence Alignment Page, the better the match. And this was a sign the constructed gene model was done correctly. A second tool Clustal Omega offered for checking the accuracy of the constructed gene model was the Phylogenetic Tree. If species of the same genus amongst the seven species were not drawn close to each other in the Phylogenetic Tree, this was a sign the gene model was designed incorrectly (Sievers et al 2011).

Using Biozentium’s SWISS To Determine Structure of Peptide Sequence

After the Apollo constructed gene model passed all these tests for accuracy, precision, and completeness, the peptide sequence of the Apollo constructed gene model was pasted into Biozentium’s SWISS MODEL search bar. SWISS then searched for a structure posted on the SWISS website that exactly matched the Apollo constructed gene model. If this was not available, SWISS instead displayed structure results for the topmost homologous peptide sequences. The results of the topmost homologous peptide sequence matches could then be used to infer the actual structure of the constructed Apollo Gene Model. Since the structure of a protein determines its function, establishing the structure of the constructed Apollo gene model is a major source of evidence that was used to infer the function of the gene model (Guex et al 2005).

Works Cited

Castresana, Jose. “On Homology Searches by Protein Blast and the Characterization of the Age of Genes.” *BMC Evolutionary Biology*, BioMed Central Ltd, 4 Apr. 2007, bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-7-53.

Guex, Nicolas, and Manuel C. Peitsch. “SWISS‐MODEL and the Swiss‐Pdb Viewer: An Environment for Comparative Protein Modeling.” *ELECTROPHORESIS*, Wiley-Blackwell, 14 Apr. 2005, onlinelibrary.wiley.com/doi/full/10.1002/elps.1150181505.

Lewis, et al. “Apollo: a Sequence Annotation Editor.” *Genome Biology*, BioMed Central, 23 Dec. 2002, genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-12-research0082.

Sievers, Fabian, et al. “Fast, Scalable Generation of High‐Quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology*, EMBO Press, 1 Jan. 2011, msb.embopress.org/content/7/1/539.print.