GeneSeqToFamily: a Galaxy workflow to find gene families based on the Ensembl Compara GeneTrees pipeline

NICOLA SORANZO

Data Infrastructure and Algorithms

@nicolasoranzo
Gene family

- A gene family (or orthogroup) is a set of similar genes descended from a single original gene.
- The study of homologous genes helps to understand the evolution of gene families.
- Various tools are available to identify gene families and provide an overview of syntenic regions evolution at the family level:
  - MSOAR, OrthoMCL, HomoloGene
  - but they don’t provide information about structural changes within a gene.

Genomicus

www.earlham.ac.uk
Ensembl GeneTree Pipeline

- Generates gene families based on **coding sequences**
- Provides details about **exon conservation**
- It uses various open source tools:
  - BLAST
  - hcluster_sg
  - T-Coffee
  - TreeBeST
- Issues:
  - Specific to certain **computing environments**
  - Only usable via **command line**
  - Require many **dependencies** to be fulfilled
  - Not flexible
GeneSeqToFamily
GeneSeqToFamily: preparation

- Removes coding sequences whose length is not a multiple of 3
- Keeps only the transcripts present in the gene feature information

Optionally:
- Keeps only the longest CDS per gene
- Changes the header of the FASTA sequences to the format required by TreeBest
GeneSeqToFamily: workflow overview
GeneSeqToFamily: Translation

- Transeq is part of the EMBOSS suite
GeneSeqToFamily: Pre-clustering alignment

- BLASTP is run all vs. all (using makeblastdb)
- BLAST parser calculates an edge weight as \(-\log_{10}(\text{BLAST e-value})/2\), capped at 100. It also removes the self-matching BLAST results and optionally filters out non-Reciprocal Best Hits.
GeneSeqToFamily: Cluster generation

- **hcluster_sg**: clustering tool for sparse graphs by Heng Li
- **hcluster_sg parser**: converts the hcluster_sg output into a collection of clusters, where each cluster is a list of IDs. It can also filter out clusters with a number of elements outside a specified range (TreeBeST requires at least 3 genes to generate a gene tree)
GeneSeqToFamily: Cluster alignment

- T-Coffee: Multiple Sequence Alignment tool, but can also be used to combine the output of other alignment methods (Clustal, MAFFT, MUSCLE...) into a single alignment
GeneSeqToFamily: Gene tree construction

- CDS for each cluster is generated using `filter_by_fasta_ids`
- Tranalign: also part of the EMBOSS suite
- TreeBeST (Tree Building guided by Species Tree) generates a consensus gene tree (using 5 different algorithms) based on a known species tree
- The final gene tree includes **phylogenetic annotations** (e.g. speciation or duplication)
GeneSeqToFamily: Gene Alignment and Family Aggregator (GAFA)

- The resulting database provides a reusable, persistent data store for visualization of synteny information with Aequatus
- Internally, GAFA converts each MSA from fasta_aln format to a simple CIGAR string (as in SAM format)
GeneSeqToFamilly: Visualisation

The GAFA SQLite database can be visualised with the Aequatus.js plugin
Aequatus.js
Aequatus.js

- New JavaScript library for visualisation of homologous genes
  - extracted from the standalone Aequatus software package http://aequatus.earlham.ac.uk
- Detailed view of gene structure across gene families
- Shared exons use the same colour in each representation
- Insertions (black bars) and deletions (red lines) w.r.t a guide gene

Aequatus: An open-source homology browser

bioRxiv 055632; doi: 10.1101/055632
Aequatus.js

- Genes nodes are coloured based on the potential **evolutionary event**
Additional tools

- Data generation:
  - **Ensembl REST API** tools: Get sequences by Ensembl ID, Get features by Ensembl ID
  - **ETE** tools: ETE taxa DB generator, ETE species tree generator
  - **ETE GeneTree splitter**: splits trees at duplication events
    - Useful when GSTF produces large families
  - **ETE homology classifier** and filter: determine homology type for each pair of genes in a tree (1-to-1, 1-to-many, many-to-many, paralogs)
Additional workflows

- Workflows to find **families for orphan genes**:
  - Retrieve the orphan genes, i.e. discarded after BLAST parser or by hcluster_s_g parser
  - Find homologous genes using SwissProt:
    - BLASTP orphan gene sequences against SwissProt
    - Map hit UniProt IDs to Ensembl IDs
    - Get sequences and feature info from Ensembl
  - Run GeneSeqToFamly on orphans and their homologues
Use cases
Use cases

- Species: 16 (15 rodents + rabbit)
- Genes: 302,383
- Gene families
  - With default params: 7247
  - After splitting: 22173
  - Families with just 1-to-1 orthologs: 2286
- For randomly-selected families, one-to-one orthologs identified by GSTF are consistent with Ensembl Compara
- Now testing with 19 species
- Similarly for a set of monocot and dicot plant genomes
Conclusion
Conclusion

- A user-friendly Galaxy workflow to identify gene families
  - Especially useful for unpublished genomes
- Scalable from one family to several genomes
  - We are pushing Galaxy to support 20K-element collections
- Flexibility to tailor the analysis by changing parameters and tools
- Interactive visualisation within Galaxy
- Easy to install (all tool dependencies available in Conda)
Future work

- Training material WIP PR: [https://github.com/galaxyproject/training-material/pull/778](https://github.com/galaxyproject/training-material/pull/778)
- Implement DB search in Aequatus.js viz
- Add gene positions to GSTF SQLite, and then show synteny in viz
- Search and visualisation of protein domains
Availability

  [https://doi.org/10.1093/gigascience/giy005](https://doi.org/10.1093/gigascience/giy005)
- ToolShed: [https://toolshed.g2.bx.psu.edu/view/earlhaminst](https://toolshed.g2.bx.psu.edu/view/earlhaminst)
- GitHub: [https://github.com/TGAC/earlham-galaxytools](https://github.com/TGAC/earlham-galaxytools)
- Aequatus.js visualisation plugin: PR to galaxy repo coming soon
- Tools and viz installed at [https://usegalaxy.eu/](https://usegalaxy.eu/)
- Contacts:
  - anil.thanki@earlham.ac.uk / nicola.soranzo@earlham.ac.uk
  - Twitter: @anilthanki / @nicolasoranzo
Acknowledgments

- Will Nash, Ben White, Wilfried Haerty
- Erin Baggs
- David Wright, David Thybert
- Marius van den Beek (Institut Curie), John Chilton and the Galaxy Team
- Matthias Bernt (UFZ)
- Björn Grüning (Uni Freiburg)
- Matthieu Muffato (EBI)
- Anil Thanki, Rob Davey