How to use Fgenesh in Galaxy Australia

Anna Syme, November 2024

What is Fgenesh?

- Fgenesh is a tool for genome annotation
- Link to more information: https://www.biocommons.org.au/fgenesh-plus-plus

Can I use Fgenesh in Galaxy Australia to annotate a genome?

- Yes, but we recommend a basic understanding of Galaxy, repeat masking, and genome annotation.
  - See the Galaxy Training Network for tutorials: https://training.galaxyproject.org/
- To request access to Fgenesh:
  - https://site.usegalaxy.org.au/request/access/fgenesh

Preparing to use Fgenesh

- Log in to Galaxy Australia
- Upload your data
  - genome assembly in fasta format, e.g. `assembly.fasta`
  - the same genome assembly, but masked, e.g. `masked_assembly.fasta` (Note: The developers of Fgenesh recommend to use a hard-masked rather than soft-masked genome)
- Sample data for this tutorial:
  - Genome assembly of fungal plant pathogen *Mucor mucedo*
  - Import this history to get test data: https://usegalaxy.org.au/u/anna/h/input-data-for-fgenesh-tutorial-new-copy
  - The history information describes where this data is from.

Split input files (to speed up the next step)

- to split `assembly.fasta`:
  - find tool: **FGENESH split**
  - for fasta file input: `assembly.fasta`
  - output filename format: use sequence header
  - output file extension: fa (non-masked)
  - run tool
  - output = a collection of assembly fasta files: one file per contig.
- to split `masked_assembly.fasta`:
  - find tool: **FGENESH split**
  - fasta file input: `masked_assembly.fasta`

    - output filename format: use sequence header
    - output file extension: fa.masked (masked)
    - run tool
    - output = a collection of fasta files: one file per contig

Annotate the assembly

- find tool: **FGENESH annotate**
- input type:
    - choose: assembled genome, in multiple contigs
- multiple sequence:
    - collection of assembly fasta files (output from fgenesh split)
- use repeat masking sequence:
    - choose: repeat masked genome in multiple contigs
- repeat masked sequence:
    - collection of repeat masked assembly files (output from fgenesh split)
- select matrix type:
    - use a built-in index
- select a species matrix:
    - choose approximately nearest related species (e.g., for the sample data, type in "Mucor")
- select db type:
    - use a built-in index
- select a reference database:
    - choose mammal or non-mammal (e.g., for the sample data, choose non-mammal)
- select nr db type:
    - this only applies if you are using a set of known protein sequences, and have selected the option further down: USE_PROTEINS = yes
    - if you aren't using proteins, disregard this and leave as default setting
- all other settings: use defaults or change as needed.
- run tool
- outputs: a collection of gff3 files and a collection of txt files

Merge outputs

- find tool: **FGENESH merge**
- input file type: gff
- collection: the gff3 output from the annotation step
- run tool
- output: a a single gff3 file of annotations

Summarise results

- find tool: **Genome annotation statistics**
- Annotation to analyse: the merged FGENESH gff3 file
- Reference genome: use a genome from history
- Corresponding genome sequence: `masked_assembly.fasta`
- run tool
- outputs: Genome annotation statistics, see the summary file