High throughput comparative genomics using a Chado backend

Mara Kim
mara.kim@vanderbilt.edu

GMOD 2014
Vanderbilt University

Jan 16, 2014
The Rokas Lab

- Computational Genomics laboratory
  - Fungal genetics
  - Human preterm birth
- RokasDB - Comparative genomics database
  - 200+ eukaryote genomes
  - Modified Chado schema
Measures of genomic similarity

- **Sequence similarity**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Alignment</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>P49789</td>
<td><img src="image" alt="Alignment Diagram" /></td>
<td>8e-27</td>
</tr>
<tr>
<td>P49779</td>
<td><img src="image" alt="Alignment Diagram" /></td>
<td>6e-18</td>
</tr>
<tr>
<td>P49775</td>
<td><img src="image" alt="Alignment Diagram" /></td>
<td>3e-07</td>
</tr>
<tr>
<td>Q11066</td>
<td><img src="image" alt="Alignment Diagram" /></td>
<td></td>
</tr>
<tr>
<td>Q09344</td>
<td><img src="image" alt="Alignment Diagram" /></td>
<td>4e-05</td>
</tr>
<tr>
<td>P49378</td>
<td><img src="image" alt="Alignment Diagram" /></td>
<td>0.001</td>
</tr>
<tr>
<td>P32084</td>
<td><img src="image" alt="Alignment Diagram" /></td>
<td>0.002</td>
</tr>
</tbody>
</table>

Measures of genomic similarity

- Ortholog estimation

Measures of genomic similarity

- Conservation of Synteny

Measures of genomic similarity

■ Functional similarity

Trees: with great power...

Trees: with great power . . .
comes great responsibility
Phylogeny vs. Similarity

- Phylogeny: shared ancestry (Darwinian homology)
- Similarity: common function (Owenian homology)
- Phylogenetic trees represent evolutionary history
- Relevance of model organisms depends on similarity
  - How do we quantify and visualize genomic similarity?

http://evolution.berkeley.edu/evolibrary/article/evo_08
Introducing...

The Genome Yardstick
What is a Yardstick?

**yard-stick** *n.* a rigid yard-measure; also *fig.*, **a standard of comparison**.


Our *standard of comparison* is the similarity between the human genome and other species which is used to gauge the similarity between two arbitrary genomes.
What is a Yardstick?
What is a Yardstick?
Goal

Quantify the molecular phenotype of the organisms on the yardstick
General Datapath
Blast Datapath

- **Distributed Filesystem [GPFS]**
  - $FASTFILE_PATH
    - *.fasta
    - *.fasta.phr
    - *.fasta.pin
    - *.fasta.psq

- **Cluster Node [TORQUE]**
  - **blastsqlclient**
  - Queries rokasdb for pending blast
  - Copies genome files
  - Calls blastp
  - Processes XML output and commits to rokasdb

- **Database Node**
  - rokasdb [PostgreSQL]
    - Chado
      - analysis
      - featureloc
      - feature
    - Audit
      - pendingblast
        - +Genome
        - +Gene
        - +probsize
        - +blast_id
        - +job_id
      - updatedfeature
        - +original_id
        - +feature_id
        - +type_id

- **Genome**
  - Rokaslab User
  - genomeImport
    - Parses genome files
    - Commits to rokasdb
  - blastsqlclient
    - Manages *.fasta files
    - Calls makeblastdb
    - Adds pending BLAST jobs
Score best hits

Score the best hit of every gene to every other genome

- Protein sequence identity
- Matrix adjusted protein identity (BLAST bitscore)
Score best hits

Score the best hit of every gene to every other genome

- Protein sequence identity
- Matrix adjusted protein identity (BLAST bitscore)
- Orthology estimation using reciprocal best hit
Calculate Aggregate Proteome Similarity

For each pair of genomes, find the average score by each measure
Aggregate Protein Sequence Identity Yardstick
Aggregate Bitscore Yardstick (BLOSUM62)
Specialized yardsticks

Generate yardstick using only a subset of the genome
- Metabolic Genome Yardstick
- Developmental Genome Yardstick

Weight genes by expression
- Placental Genome Yardstick
- Embryonic Genome Yardstick
Pathway centric yardsticks – Gene Set Enrichment Analysis

Fig. 1. A GSEA overview illustrating the method. (A) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the “gene tags,” i.e., location of genes from a set \( S \) within the sorted list. (B) Plot of the running sum for \( S \) in the data set, including the location of the maximum enrichment score \( ES(S) \) and the leading-edge subset.

Comparative module (featuregroups)

- Represent and annotate sets of features
- Avoids denormalization of ortholog annotation
- Allows access to sets via Foreign Keys
Proposed module structure
Complexity analysis

- featuregroup
  - Insert group: $O(\log n)$
  - Insert member: $O(\log n)$
  - Select members: $O(\log n)$
  - Annotate group: $O(\log n)$
  - Delete group: $O(\log n)$

- Tagging via featureprop
  - Insert group(!): $O(\log n)$
  - Insert member: $O(\log n)$
  - Select members: $O(n)$
  - Annotate group(!): $O(n)$
  - Delete group: $O(n)$
Summary

- Genome Yardsticks quantify similarity between organisms
- Specialized yardsticks test relevance of model systems
- Standard way to represent sets of genomic features is needed
Awknowledgements

- Dr. Antonis Rokas
- Dr. Kris McGary
- Dr. Jennifer Wisecaver

- GMOD Project
- March of Dimes
BLAST in Chado
Proposed SQL

--Groups
CREATE TABLE featuregroup (  
    featuregroup_id serial PRIMARY KEY,  
    name varchar(255),  
    uniquename text NOT NULL,  
    rank integer NOT NULL DEFAULT 0,  
    type_id integer NOT NULL REFERENCES cvterm  
    ON UPDATE CASCADE ON DELETE CASCADE  
    DEFERRABLE INITIALLY DEFERRED,  
    is_analysis boolean NOT NULL DEFAULT false,  
    UNIQUE(uniquename, rank, type_id)  
);  

--Group members
CREATE TABLE feature_featuregroup (  
    feature_featuregroup_id serial PRIMARY KEY,  
    featuregroup_id integer NOT NULL REFERENCES featuregroup  
    ON UPDATE CASCADE ON DELETE CASCADE  
    DEFERRABLE INITIALLY DEFERRED,  
    feature_id integer NOT NULL REFERENCES feature  
    ON UPDATE CASCADE ON DELETE CASCADE  
    DEFERRABLE INITIALLY DEFERRED,  
    UNIQUE(featuregroup_id, feature_id)  
);  

--Group annotation
CREATE TABLE featuregroupprop (  
    featuregroupprop_id serial PRIMARY KEY,  
    value text,  
    rank integer NOT NULL DEFAULT 0,  
    type_id integer NOT NULL REFERENCES cvterm  
    ON UPDATE CASCADE ON DELETE CASCADE  
    DEFERRABLE INITIALLY DEFERRED,  
    featuregroup_id integer NOT NULL REFERENCES featuregroup  
    ON UPDATE CASCADE ON DELETE CASCADE  
    DEFERRABLE INITIALLY DEFERRED,  
    UNIQUE(rank, type_id, featuregroup_id)  
);  

--Analysis featuregroups
CREATE TABLE analysisfeaturegroup (  
    analysisfeaturegroup_id serial PRIMARY KEY,  
    rawscore double precision,  
    normscore double precision,  
    significance double precision,  
    identity double precision,  
    analysis_id integer NOT NULL REFERENCES analysis  
    ON UPDATE CASCADE ON DELETE CASCADE  
    DEFERRABLE INITIALLY DEFERRED,  
    featuregroup_id integer NOT NULL REFERENCES featuregroup  
    ON UPDATE CASCADE ON DELETE CASCADE  
    DEFERRABLE INITIALLY DEFERRED,  
    UNIQUE(analysis_id, featuregroup_id)  
);  

--TODO? featuregroup_pub  
-- featuregroup_relationship, etc.
MaraKim_GMOD2014 1.0
e6d7383f6f2a69db45040ec596d254b1d5a460a8
## master