WEB-BASED BIOINFORMATICS PIPELINES FOR BIOLOGISTS

Integrative Services for Genomic Analysis (ISGA)
Chris Hemmerich
Center for Genomics and Bioinformatics
CONTACT: biohelp@cgb.indiana.edu
JUSTIFICATION AND HISTORY
ISGA BACKGROUND

- Provide a high-throughput microbial annotation service to local biologists
  - Reliable and pipelined execution
  - Efficient maintenance
  - Provide privacy and security for data
- High-quality (automated) annotation
  - Biologists able to customize parameters
  - Able to incorporate new programs and pipelines
ERGATIS
(ERGATIS.SOURCEFORGE.NET)

- Web-based analysis pipeline tool
- Wraps tools and utilities in “components”
- Ability to add new components
- Build new and customize existing pipelines
- In-depth monitoring of pipelines
- Underlying Workflow package supports SGE
- XML/BSML common data exchange format
- Includes prokaryotic annotation pipeline
ERGATIS WORKFLOW
A SLIGHT CORRECTION
WHY NOT EXPOSE ERGATIS?

- Insufficient accounts and permissions
- Shared interface for building and customizing pipelines
- Users must submit and retrieve results through filesystem
- Pipeline monitoring interface is slow and complex.
- Information of use to biologists is lost in “noise”
  - High number of components in a pipeline
  - Complexity of configuration interface
### Configuration

#### Parameters

- **training seq**
- **icm**
  
  ```
  ${OUTPUT_DIRECTORY}/${COMPONENT_NAME}.icm
  ```
  
  If self-training (using long-orfs), include long-orfs parameters/options here (Optional)

- **long orfs opts**
  
  ```
  -n -t 1.15
  ```
  
  If using a coordinates masking file, set this to at least include:
  ```
  -i
  ```

- **glimmer3 opts**
- **start codon usage**
  
  Used in id generation

- **project abbreviation**
  
  ```
  ${PROJECT}
  ```

#### Input

- **input file list**
- **input file**
- **input directory**
  
  The following is only used when iterating over an INPUT_DIRECTORY

- **input extension**
  
  `fsa`

#### Output

- **output token**
- **output directory**
  
  ```
  ${REPOSITORY_ROOT}/output_repository/${COMPONENT_NAME}/${PIPELINE}
  ```

- **bsml output list**
  
  ```
  ${OUTPUT_DIRECTORY}/${COMPONENT_NAME}.bsml.list
  ```

- **raw output list**
  
  ```
  ${OUTPUT_DIRECTORY}/${COMPONENT_NAME}.raw.list
  ```

---

Component: **glimmer3**
OUR SOLUTION

- Develop an alternative interface for biologists that uses the Ergatis backend
  - Administrators also use Ergatis

- New interface features
  - Accounts and permission system
  - File management
  - Simplify pipelines and component management by reducing functionality
  - Provide form validation, documentation and other features to improve usability
THE GOAL
ISGA: WHIRLWIND TOUR
PIPELINE CUSTOMIZATION

- Ability to toggle some clusters on/off.
- Some clusters contain parallel programs that can be independently toggled.
- Ability to edit component parameters
- Ability to save customizations to use with later data sets
Pipeline Building: Edit Parameters

Glimmer3 Refinement Pass

Required Parameters

- Long ORF Entropy Cutoff: 1.15
- Maximum Overlap: 50
- Minimum Gene Length: 110
- Threshold Score: 30
- Translation Table: (1) Bacterial Table

Optional Parameters

Pipeline Customization Tools

Overview
Return to the pipeline overview page, where you can view details and your current workflow.

Edit Pipeline Name and Description
Edit the name or description for the pipeline

View Inputs and Outputs
The files you will need to upload for your pipeline depend on the programs you have chosen to run. This Tool will allow you to view those inputs. Also view the files that your currently configuration will generate as output.

Finalize Pipeline
# Run Status

<table>
<thead>
<tr>
<th>Name</th>
<th>Prokaryotic Annotation Pipeline Run 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>7844057</td>
</tr>
<tr>
<td>Status</td>
<td>Running (Hide Detailed Status)</td>
</tr>
<tr>
<td>Started At</td>
<td>Jan 14, 2010 09:55 EDT</td>
</tr>
<tr>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>Input Files</td>
<td>sample_data.fna</td>
</tr>
</tbody>
</table>

## Detailed Status

### Jobs

<table>
<thead>
<tr>
<th>Job</th>
<th>State</th>
<th>Progress</th>
<th>Start (EDT)</th>
<th>End (EDT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pipeline</td>
<td>Running</td>
<td></td>
<td>Jan 14, 2010 09:55</td>
<td></td>
</tr>
<tr>
<td>TFBS Search</td>
<td>Running</td>
<td>11/13+</td>
<td>Jan 14, 2010 09:57</td>
<td></td>
</tr>
<tr>
<td>ORF Prediction</td>
<td>Complete</td>
<td>27/27</td>
<td>Jan 14, 2010 09:55</td>
<td>Jan 14, 2010 09:56</td>
</tr>
<tr>
<td>Additional Gene Analysis</td>
<td>Incomplete</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein Domain Search</td>
<td>Running</td>
<td>22/34+</td>
<td>Jan 14, 2010 09:57</td>
<td></td>
</tr>
<tr>
<td>Sequence Similarity Search</td>
<td>Running</td>
<td>20/46+</td>
<td>Jan 14, 2010 09:57</td>
<td></td>
</tr>
<tr>
<td>RNA Prediction</td>
<td>Complete</td>
<td>22/22</td>
<td>Jan 14, 2010 09:55</td>
<td>Jan 14, 2010 09:56</td>
</tr>
<tr>
<td>Alternate Start Site Analysis</td>
<td>Incomplete</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Output</td>
<td>Incomplete</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Close**
ISGA PIPELINE EXECUTION

- ISGA writes configuration and pipeline definition files to the Ergatis installation
- ISGA then triggers execution through Ergatis and receives the pipeline id in return
- Status is updated directly from Ergatis XML files
- Selected output is copied to ISGA, and the rest is available for download if needed
ISGA TOOLBOX

- Includes a GBrowse instance for visualizing annotation results
- BLAST support for pipeline results as query or database
- Text search against annotation results
- Tools can be executed over SGE and monitored
ADMINISTRATIVE TOOLS

- Lightly monitor status in ISGA w/ link to Ergatis page
- Notification when pipeline fails, ISGA will pick up a resumed pipeline
- Ability to redirect ISGA to a cloned Ergatis pipeline or cancel (w/ user notification)
- Disable new job submissions
UNDER THE HOOD

ISGA Web Interface
- pipeline builder
- genome browser
- monitor pipelines
- download results
- blast search

PostgreSQL Database
- pipeline specification
- user account
- annotation results

Ergatis
- XML configuration
- workflow engine

Shared Storage
- bioinformatics tools
- input and results

Sun Grid Engine
- computation nodes
- job scheduler

ISGA Backend
UNDER THE HOOD (CONTINUED)

- Perl & jQuery
- Persistence = PostgreSQL & YAML & XML
- Mason
- MasonX::WebApp
- Hacked up HTML::FormEngine
ADDING AN ERGATIS PIPELINE TO ISGA
64 Ergatis Components

Genome Sequence

ORF Prediction

RNA Prediction
  • tRNA
  • rRNA

Sequence Similarity Search*

Additional Gene Analysis
  • Metabolic Pathways
  • Metabolic Enzymes
  • Subcellular Localization
  • Ortholog/Paralog
  • Lipoprotein
  • Microarray Oligo Primer
  • Transcription Terminator

Protein Domain Search*

TFBS Search

Output
  • Combined Gene Function Prediction
  • Dynamic Analysis Tools (e.g. BLAST)
  • Visualization (e.g. Genome Browser)
  • Download Results
FIRST: UNDERSTAND THE PIPELINE

- ISGA takes a description of an Ergatis pipeline
  - YAML
  - Database Schema
  - Ergatis component .config files
- Document input and output of all components
- Which components are optional?
  - The user can upload previously generated data in their stead?
  - Alternative data from the pipeline can be used?
  - The pipeline is still useful without this functionality
SIMPLIFICATION

- Our microbial annotation pipeline is composed of 64 Ergatis components
  - Impossible to diagram for you on a slide or for a biologist on our web page
- Many of these components are file format conversions, program iterations, database preparation, etc...
  - They are not relevant to a high level view of the pipeline and offer no useful parameters for a biologist to customize
CLUSTERS OF ERGATIS COMPONENTS

- Break the pipeline into biologically meaningful clusters of one or more components
  - This is as much art as science, may depend on your audience
  - Example: ‘Alternative Start Site Analysis’

- overlap_analysis.default
- start_site_curation.default
- translate_sequence.translate_new_model
- parse_evidence.hypothetical
- hmmpfam.post_overlap_analysis
- parse_evidence.hmmpfam_post
- wu-blastp.post_overlap_analysis
- bsmI2fasta.post_overlap_analysis
- bsmI2featurerelationships.post_overlap
- xdformat.post_overlap_analysis
- ber.post_overlap_analysis
- parse_evidence.ber_post
- translate_sequence.final_polypeptides
- bsmI2fasta.final_cds
COMPONENT CUSTOMIZATION

- Scripts and XML files are unchanged
- ISGA stores the configuration template for each component
- Components with editable parameters have a YAML definition that is used to build the web form
- These values are incorporated into the configuration template
COMPONENT TEMPLATE

--- !perl/ISGA::ComponentBuilder
Name: RNAmmer
Description: ‘RNAmmer predicts 5s/8s, 16s/18s, and …’
Params:
RunBuilderParams:
- { templ: 'hidden', NAME: 'project_id_root', TITLE: 'Project Id Root', REQUIRED: 1, DESCRIPTION: 'The Id root used in bsml id generation', CONFIGLINE: '___project_id_root___' }
FUTURE ISGA WORK

- Incorporate additional pipelines
  - Small prokaryotic assembly pipeline
  - Comparative genomics
  - Functional genomics

- Add additional features
  - Make pipelines modular components of ISGA
  - Implement pipeline versioning
  - Pipeline and data sharing

- Ergatis Cloud Support?