Workflows in bioinformatics:
meta-analysis and prototype implementation of a workflow generator

Alexander Garcia Castro, Samuel Thoraval, Leyla J Garcia and Mark A Ragan
The Institute for Molecular Bioscience

Comparative and Computational Genomics

- Automated inference of vertical and lateral gene transmission in microbial genomes
- Novel algorithmic methods in comparative genomics
- Relational database structures for genomic data
1. introduction
2. meta analysis
   • Syntactic components
   • Algebraic components
3. implementation
   • Architecture
   • Examples
   • Screenshots
4. Future works
1. Introduction

Bioinformatic experiments often interleaves *information access* and *algorithm execution* in a problem-specific workflow.

```
Input

Process: sequence of activities

Output
```
These activities need to be controlled by a workflow management system that will:

- Perform data transformations (to ensure syntactic interoperability)
- Interact with the sequence of activities and coordinate them

We focus our analysis only on those closely related to workflow design issues:

- Piping of data
- Conditional statements
- Iteration
- Recursion
- Suspension/Resumption
2. meta analysis

Our workflow follows a **task-flow** model
Our workflow model closely follows those of Lei and Singh[1] and Stevens et al.[2]

1. R. Stevens, C Goble, P Baker, A Brass: A classification of tasks in bioinformatics. Bioinformatics 2001
2. Lei K, Singh M: A Comparison of Workflow Metamodels.
A **stage** holds one or more independent tasks, annotations and results.

A protocol with task concurrency

*in silico* experiments are captured as **protocols**, enabling sharing and replication
Algebraic Operators

- Iteration (I)
- Recursion (R)
- Condition (C)
- Suspension/Resumption (S)
We have also used formal concept analysis to define valid objects and their behaviours:

We have defined operators in terms of pre and post-conditions:

Example:
$I[\text{Transformer}, (\text{CC}1, \text{CC}2, \ldots, \text{CC}n)]: (\text{CC}1', \text{CC}2', \ldots, \text{CC}n')$

Pre-condition: $T = \text{Transformer} \land T \neq \text{blank}$
$C = \{\text{CC}1, \text{CC}2, \ldots, \text{CC}n\}$ such that $\text{CC}i \in \{\text{Biological data types}\}$

Post-condition:
$C' = \{\text{CC}'1, \text{CC}'2, \ldots, \text{CC}'n\}$ such that $\text{CC}'i = T(\text{CC}i) \land 1 \leq i \leq n$

We are currently working on a complete FCA for biological data types and operations (database and analytical)
3. Implementation

Gpipe, a GUI workflow generator for PISE[3]

Features include:
- Pise Method browser with piping capabilities
- Definition of a complete bioinformatics experiment
- Jobs monitoring and execution
- Protocol sharing
- Grid computing

Gpipe substantially mitigate the syntactic complexity behind the process of building a workflow

3. Letondal C: A Web interface generator for molecular biology programs on Unix. Bioinformatics 2001, 17:73-82
GPipe Architecture

Grid computing

Coordination

Share/Reproduce

Method definition

Flexibility

Persistence

Hide syntactic complexity

Protocol
Clustalw: Multiple Alignments ([Des Higgins](http://if-web1.imb.uq.edu.au/))

Add clustalw | Add clustalw to a new workflow

Reset | Run clustalw | samuel.thoraval@librophyt.com | your e-mail

Select a server for clustalw

Sequences File (or Alignment File for Bootstrap and Tree actions) (-infile) (format)

Phylip alignment output format (-output)

Toggle Slow/Fast pairwise alignments (-quicktree) ? Speed: Slow, Fast

Advanced clustalw form
Clustalw: Multiple Alignments (Des Higgins)

Apply parameters modification

Pipe clustalw | Pipe clustalw to another program | Remove clustalw | Remove this method (piped methods will be removed)

samuel.thoraval@librophyt.com | your e-mail

http://if-web1.imb.uq.edu.au/cgi-bin/Pise/5.a/clustalw.pl | Select a server for clustalw

Sequences File (or Alignment File for Bootstrap and Tree actions) (-infile) (format)

Phylip alignment output format (-output)
Toggle Slow/Fast pairwise alignments (-quicktree) ? | Slow ☐ Fast
Clustalw

Choose a protpars to pipe clustalw (results of type *.aln, *.gde, *.phy)

Choose a drawgram to pipe clustalw (results of type *.dnd, *.ph, *.phb)
**Phylib**: protpars - Protein Sequence Parcimony Method (**Felsenstein**)

Add protpars to the workflow

- samuel.thoraval@librophyt.com your e-mail

Select a **server** for protpars

**Advanced protpars form**

References:

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Felsenstein, J. 1989. PHYLIP -- Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.

**Pise** form generator version: (04 Oct 2005 08:48)
Clustalw: Multiple Alignments (Des Higgins)

Apply parameters modification

Pipe clustalw  |  Pipe clustalw to another program  |  Remove clustalw  |  Remove this method (piped methods will be removed)

samuel.thoraval@librophyt.com  your e-mail

Select a server for clustalw

Sequences File (or Alignment File for Bootstrap and Tree actions) (-infile) (format)

Phylip alignment output format (-output)
Toggle Slow/Fast pairwise alignments (-quicktree)  @ Slow  @ Fast
Phylip: protdist - Program to compute distance matrix from protein sequences (Felsenstein)

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Felsenstein, J. 1989. PHYLIP -- Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.
Run the workflow

Run the workflow

Click on "Jobs" in the workflow explorer if not redirected...

Save the workflow

This is a perl script to be edited and executed (for programmers).

This is an XML file to save, edit and share the workflow.

Your methods settings

Parameters for clustalw (1)

| Parameter                                                                 | Value |
|--------------------------------------------------------------------------|-------|
| Hydrophilic residues list (-hgapresidues)                                |       |
| Residue specific penalties (Pascarella gaps) (-nopgap)                   | on    |
| Helix terminal positions: number of residues inside helix to be treated as terminal (-helixendin) | 3     |
| Loop gap penalty (-loopgap)                                              | 1     |
| Protein weight matrix (-matrix)                                          | gonnet|
| Gap extension penalty (-pgapext)                                          | 0.10  |
| Toggle Slow/Fast pairwise alignments (-quicktree)                        | slow  |
## Jobs Informations

Click on 'Jobs' or **bookmark** to display this page later again...

| METHOD      | JOB               | STATUS | From METHOD | RESULTS               |
|-------------|-------------------|--------|-------------|-----------------------|
| clustalw    | A27083112838017   | DONE   | -           | results for clustalw  |
| protpars    |                  | SUBMITTED | clustalw (1) |                       |
| protdist    |                  | SUBMITTED | clustalw (1) |                       |

[Refresh Jobs Informations]
Jobs Informations

Click on 'Jobs' or bookmark to display this page later again...

| METHOD   | JOB              | STATUS | From METHOD | RESULTS                  |
|----------|------------------|--------|-------------|--------------------------|
| clustalw | A27083112838017  | DONE   | -           | results for clustalw     |
| protpars | A27099112838017  | DONE   | clustalw (1)| results for protpars     |
| protdist | A27100112838017  | DONE   | clustalw (1)| results for protdist     |

Refresh Jobs Informations
Clustalw

Results:

infi.le.aln (6.76 Ko)

filtersites Run the selected program on infi.le.aln

infi.le.dnd

drawgram Run the selected program on infi.le.dnd

clustalw.out

standard error file

From now, this files will remain accessible for 10 days at:
http://if-web1.imb.uq.edu.au/seqanal/tmp/clustalw/A27083112838017/
You can save them individually by the Save file function if needed.

Unix exact command:
clustalw -infi.le=infi.le.data -align -gapopen=10 -gapext=0.2 -pwgapopen=10 -pwgapext=0.1

Your input data:
infi.le.data

Help

References:
Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22:4673-4680.
Annotate the workflow:

Click on a stage node to annotate a **stage** and its **methods**.

**Workflow annotation:**

A workflow for a rodent phylogeny. The exon 28 of gene eVWF (von Willebrand Factor) from different rodents are used for the analysis.
Annotation of stage 1

Annotate stage 1:

clustalw

Annotate the methods of stage 1:

clustalw (1):

multiple alignment
#!/usr/bin/perl
#
# macro to be used with Pise/bioperl module
# generated from GPipe
# See http://www-alt.pasteur.fr/~letondal/Pise/#pisebioperl
# (requires bioperl 1.2 with bioperl-run, or else a Pise distribution
# and PiseWorkflow
#
#
use lib "/usr/local/src/Pise/5.a";
use Bio::Tools::Run::AnalysisFactory::Pise;
use Bio::Tools::Run::PiseWorkflow;

my $factory = new Bio::Tools::Run::AnalysisFactory::Pise();

#--------------------------- clustalw ---------------------------

$clustalw1 = $factory->program('clustalw',
                               -remote => "http://if-web1.imb.uq.edu.au/cgi-bin/Pise/5.a/clustalw.pl",
                               -email => "samuel.thoraval@librophyt.com");

#--------------------------- Hydrophilic residues list (-hgapresidues)
$clustalw1->hgapresidues([qw(G P S N D Q E K R)]);

#--------------------------- Loop gap penalty (-loopgap)
$clustalw1->loopgap(1);

#--------------------------- Helix terminal positions: number of residues inside helix to be treated as terminal (-helixendin)
$clustalw1->helixendin(3);

#--------------------------- Protein weight matrix (matrix)
<xml version="1.0" encoding="UTF-8">
<!-- Protocole definition file generated by GPipe -->
<protocol>
  <annotation></annotation>
  <stage>
    <annotation></annotation>
    <task id="1" email="samuel.thoraval@librophyt.com"/>
    <annotation></annotation>
    <transformer name="clustalw" version="1.82" server="http://if-web1.imb.uq.edu.au/cgi-bin/Pise/5.a/cgi-bin/Pise/5.a/clusterw"/>
    <pipe_component>
      <destination ref="2" pipetype="readseq_ok_alig"/>
      <destination ref="3" pipetype="readseq_ok_alig"/>
    </pipe_component>
    <parameter name="hgapresidues">
      <value>G</value>
      <value>P</value>
      <value>S</value>
      <value>N</value>
      <value>D</value>
      <value>Q</value>
      <value>E</value>
      <value>K</value>
      <value>R</value>
    </parameter>
    <parameter name="pgap">
      <value>on</value>
    </parameter>
    <parameter name="helixendin">
      <value>3</value>
    </parameter>
    <parameter name="loopgap">
      <value>1</value>
    </parameter>
    <parameter name="matrix">
      <value>gonnet</value>
    </parameter>
    <parameter name="pwgapext">
      <value>-10</value>
    </parameter>
  </stage>
</protocol>
Open a protocol

Open to load the specified protocol into GPipe

Specify a protocol definition file (XML file):

/home/slf/tmp/rodent_philogeny.xml
3. Future Work

- Define a more wide-ranging algebra including query operations
- Add a semantic layer (method functions and operations) to attempt to separate the domain knowledge from the operational knowledge

The biologist must be able to describe, with a controlled vocabulary, the type of experiment and the system guide him to build that experiment.
Condition: sequences are aligned or not. (i) Sequences are not aligned then use ClustalW. (ii) Next stage.

if avg_distance < 0.1
    dist_method = OPTION
elseif avg_distance >= 0.1 and avg_distance < 0.3
    dist_method = OPTION
else
    dist_method = OPTION

J (Blanck_Transformer)(CC1, CC2...CCn)

Puzzle

Maximum likelihood tree

Parsimony

Ednadist Eneigbo Econsense

Distance tree