User-designed web services to support heterogeneous biological data retrieval

Marie-Dominique DEVIGNES, Hervé de PALMA, Laurent PIERRON, Lionel DOMENJOUDE and Malika SMAIL-TABBONE

Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)
CNRS – INRIA - Université Henri Poincaré
Proliférateurs des Peroxysomes (EA3446) Université Henri Poincaré
Nancy, France
One biological question
« Is this DNA fragment located upstream some interesting gene? "

Querying
- Formulate a query
- Describe what you want
- Track the origin of the result
- Repeat the process

versus

Browsing
- Activate hyperlinks
- Discover what you cannot describe
- Disorientation
- Cognitive overload

Find a compromise between efficiency and exploration

2005/10/06  Devignes - NETTAB 2005
OUTLINE

1. Heterogeneous biological data retrieval
   A. User constraints
   B. Case study

2. The Xcollect application
   A. Principles
   B. Architecture

3. Example - Demo

4. Encapsulation into applications
1. Heterogeneous biological data retrieval

A. User constraints

- ✔ Unlimited choice of databases or treatments
- ✔ Minimal maintenance costs
- ✔ Easy adjustment to databases changes
- ✔ Chaining of elementary steps (workflow)
- ✔ Quality metadata retrieval
- ✔ Generic solution
- ✔ Automatic treatment of list of entries
- ✔ Transformable session document
B. Case studies (1/2)

"Is this DNA fragment located upstream some interesting gene?"

**Xprom: From sequence to genomic context**

1. Heterogeneous biological data retrieval
2. Xcollect application
3. Example – Demo
4. Encapsulation into applications

WET LAB → DNA.fasta → AGGTCA?

PPRE-Finder → REPEATS?

Repeat Masker

Genomic localization

BLAT on Genome Browser

2005/10/06
Devignes - NETTAB 2005
B. Case studies (2/2)

**XfirstExonMRNA-Hs**

- **AF422928**
- **AF422927**
- **BC010390**
- **BC017031**
- **U60520**

List of mRNA accession numbers

- **Description**
- **Associated citations**

GenBank

**Genomic alignment**

First exon genomic coordinates

BLAT on Genome Browser

Genome Browser

2005/10/06  Devignes - NETTAB 2005
2. The « Xcollect » application

A. Principles:
   1. Interaction with resources via HTTP protocol (URL queries, screen-scraping)
   2. Distinguish scenario design and scenario enactment
   2. Process any scenario (generic)

B. Models:
   - XML scenario model
   - Generic application that executes any scenario represented under this model
   - XML session model
   - Generic XSL sheet to display the results
Schematic functioning

1. Scenario description in generic model
   - User-designed scenario
   - Xcollect-config
   - .xml
   - Xcollect-execute
   - Input data
   - XML session document with generic structure

Java application written by:
Yvan Norsa and Herve de Palma
- User-friendly interface to enter the scenario
Xcollect interface

Xcollect configuration module is a graphic interface used to create or modify scenarios that can be executed by Xcollect.

Xcollect launcher is used to execute a scenario on an entry parameter.

Scenario: `collec\XCOLLECT-2005\scenarios\Xprom-scenario.xml`

Comment: `test`

Entry: `\XCOLLECT-2005\Entries\TestSetXprom\SeqCPT1.txt`
Configuration module

1. Heterogeneous biological data retrieval
2. Xcollect application
3. Example-Demo
4. Encapsulation into applications

Properties:
- Name: GenomeBrowser_update
- Type: URL
- URL: http://genome.ucsc.edu/cgi-bin/hgGateway
- Current URL: http://genome.ucsc.edu/cgi-bin/hgGateway?org=Human
- Multiple Outputs

Help:
- Name: This is the name of this elementary step.
- Type: Choose between an URL or a Class as a type of this elementary step. If an URL is chosen, fill the URL field with the address of the source, otherwise, if a Class file is selected, choose this file by clicking the "Try" button.
- The Class files are placed in the "classes/processing" folder.
- Multiple Outputs: check this box if you want to get several outputs with the same characteristics.

Commands:
- Add a New Input
- Add a New Output
- Save Properties
1. Heterogeneous biological data retrieval

Enactment module

2. Xcollect application

Xcollect configuration module is a graphic interface used to create or modify scenarios that can be executed by Xcollect.

Launcher Module

Xcollect launcher is used to execute a scenario on an entry parameter.

Scenario: C:\Xcollect\XCOLLECT-2005\scenarios\Xprom-6.xml

Comment: test Xprom october 3

Entry: |Excised_Fasta_Sequences|1.253consensus.fasta.txt

Launch Xcollect

Step genomicContext

ElementaryStep BlatSearch
http://genome.ucsc.edu/cgi-bin/Blat?userSeq=%201034-002+1%20+-%200x20+comment%0aGTATCCACGGGCAAACACCACCGATCGTCTGCTCAAGGACCATGTCATACACCAAGAGGACG

Inputs:
- userSeq = 1034-002+1
- so consensus:
  - GATCCACGGGCAAACACCACCGATCGTCTGCTCAAGGACCATGTCATACACCAAGAGGACG

2005/10/06

Devignes - NETTAB 2005 12
2. Example – Demo

→ Go to DEMO
4. Encapsulation into applications

Web services

Goals:
✓ to facilitate re-use of scenarios
✓ to incorporate into higher-order workflows

Method

Xprom_ws.properties
- display-name = getGenomic Context
- scenario = Xprom-scenario
- output = XML file
- input = sequence fasta file
- description = Retrieves genomic localisation from Genome Browser and annotates by PpreFinder and RepeatMasker

Xcollect_ws deploy

ws registry

Xcollect_ws factory

Client: ws browser

Sun JWSDP
JAX-RPC
JAXR

Apache Tomcat

Herve de PALMA
Laurent PIERRON
LORIA

2005/10/06

Devignes - NETTAB 2005
CONCLUSION - PERSPECTIVES

1. **Xcollect**: a generic application to collect and integrate heterogeneous data from biological web sources:
   - Along a « user-defined » scenario
   - Adaptable to any data source

2. **Make Xcollect scenarios available as web services**

3. **Semantic description of web services to facilitate their discovery**
   - Semantic « match-maker service »: e.g. Feta – myGrid
   - Analogous situation with biological databases when designing scenarios
     → **BioRegistry project at LORIA and INIST (CNRS)**
Thank you for your attention!