Peter Rice

EMBRACE

Building Inter-Operable Services for Biology

Peter Rice, EBI (pmr@ebi.ac.uk)

23 April 2009
EMBRACE

• European Union "Network of Excellence" 5-year project
• Coordinated by Graham Cameron, Director of Services at EBI
• Test cases requiring integration of data content and tools
• Application interface standards for data content:
  • DNA and protein sequence data
  • Structure and image data
  • Gene and protein expression
  • Literature and text mining
• Analysis tools using data content standards
  • Sequence analysis tools (EMBOSS etc.)
  • Structure analysis tools
  • ... and tools for all the other data types
• Taverna as an example user interface.
EMBRACE - as seen by bioinformatics developers

Sequence /Genomics

Expression /Function

Structure /Images

Application interface
Application
User interface
EMBRACE: Technology Choice

• Initial survey of webservice and grid technologies
• Web Services
  • Web Service Description Language (WSDL)
  • "Document/literal wrapped" style
    • Complicated WSDL, but everything fully defined
  • WS-I Web services interoperability standard
  • SAWSDL web service annotation standard
• Grid Services:
  • EGEE Virtual Organization
  • Test case implementations
    • Demonstrate technology
    • Show improved performance
    • Highlight technical issues
  • Grid service standards monitoring
EMBRACE: GeneFinder Test Case

• Start with a known characteristic in one species
• Map to species with well-characterized genomes
• Identify potential genes and their functions
EMBRACE: GeneFinder Test Case

• Define data resources and analysis tools required
EMBRACE: GeneFinder Test Case

• Taverna workflow example:
  • Human and dog genome comparison
  • Using EMBRACE services
EMBRACE: Grid Test Cases

• Grid implementation on EGEE
  • Implement service on Grid Bioinformatics Platform
  • Interface follows WSDL web service interface
  • Taverna (or other client) to process workflow
EMBRACE: Data content services

- Sequence and sequence related data
- EBI providing:
  - Core bioinformatics data
    - EMBL DNA sequence (GenBank/DDBJ)
    - UniProt Protein sequence
    - Ensembl Genomics
    - ArrayExpress Gene expression
    - PDBe 3D structures
- Other partners providing
  - Motifs, Domains, structure families, Alignments,
  - Transcription factors, Image data,
  - SNPs, Untranslated regions
  - Genomics data
EMBRACE: EB-eye

- EB-eye web search engine and interface
  - Standard interface to EBI core public data
- Also implemented as SOAP web services
- Web search automation by web services
- Choice of clients
  - Workflow systems (Taverna, ...)
  - Programming languages (Java, ...)
  - Simple scripts (Perl, ...)
EB-eye Search Engine

<table>
<thead>
<tr>
<th>Database Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomes</td>
<td>1</td>
</tr>
<tr>
<td>Nucleotide Sequences</td>
<td>135</td>
</tr>
<tr>
<td>Protein Sequences</td>
<td>16</td>
</tr>
<tr>
<td>Macromolecular Structures</td>
<td>1</td>
</tr>
<tr>
<td>Small molecules</td>
<td>1</td>
</tr>
<tr>
<td>Gene Expression</td>
<td>1</td>
</tr>
<tr>
<td>Molecular Interactions</td>
<td>1</td>
</tr>
<tr>
<td>Reactions &amp; Pathways</td>
<td>0</td>
</tr>
<tr>
<td>Protein Families</td>
<td>0</td>
</tr>
<tr>
<td>Enzymes</td>
<td>0</td>
</tr>
<tr>
<td>Literature</td>
<td>195</td>
</tr>
<tr>
<td>Ontologies</td>
<td>0</td>
</tr>
<tr>
<td>EBI Web Site</td>
<td>0</td>
</tr>
</tbody>
</table>

Refine your search:

Search for LFC in *All the EBI*

with the following keywords

Refine
EB-eye Search Engine

Search for LFC in Protein Sequences

13 results found in UniProtKB

ARMS2_MOUSE
Q60875, Q89115, O3T84, Q3TJ6, Q8CHE1, Q923E0, Q9E8G7
Rho guanine nucleotide exchange factor 2
View: in UniProt format, in SRS, in UniSave, in Interpro Matches
References: Ensemble, InterPro, Medline, Taxonomy, EMBL-Bank (Coding Sequence), GO, ArrayExpress, EMBL-Bank

Q9W736_XENLA
Q9W736
Submitted as: Guanine nucleotide exchange factor LFC
View: in UniProt format, in SRS, in UniSave, in Interpro Matches
References: InterPro, Medline, EMBL-Bank (Coding Sequence), Taxonomy, EMBL-Bank

Q4J72_BRAFL
Q4J72
Submitted as: Aromatic amine oxidase CYP19
View: in UniProt format, in SRS, in UniSave, in Interpro Matches
References: InterPro, Taxonomy, EMBL-Bank (Coding Sequence), GO, EMBL-Bank

1 result found in UniRef100

UniRef100_Q9W736
Cluster: Guanine nucleotide exchange factor LFC (Fragment)
EMBRACE: EB-eye Web Services

- Same functionality as the web interface

```java
listDomains()
getDomainsHierarchy()
listFields(domain)
listFieldsInformation(domain)
listAdditionalReferenceFields(domain)
getDomainsReferencedInDomain(domain)
getNumberOfResults(domain, query)
getAllResultsIds(domain, query)
getDetailedNumberOfResult(domain, query, flat)
getResultsIds(domain, query, start, size)
getResults(domain, query, fields, start, size)
getDomainsReferencedInEntry(domain, entry)
getEntry(domain, entry, fields)
getEntries(domain, entries, fields)
getEntryFieldUrls(domain, entry, fields)
getEntriesFieldUrls(domain, entries, fields)
getReferencedEntries(domain, entry, referencedDomain)
getReferencedEntriesSet(domain, entries, referencedDomain, fields)
getReferencedEntriesFlatSet(domain, entries, referencedDomain, fields)
```
EB-eye Web service clients

Web Services at the EBI

Introduction

The EBI provides programmatic access to various data resources and analysis tools via Web Services.

Web Services is an integration technology. To ensure software from various sources work well together, this technology is built on open standards such as:

- Simple Object Access Protocol (SOAP), a messaging protocol for transporting information.
- Web Service Description Language (WSDL), a standard method of describing Web Services and their capabilities.
- Representational state transfer (REST), a software architecture style.

For the transport layer, Web Services utilise most of the commonly available network protocols, especially the Hypertext Transfer Protocol (HTTP).

An introductory tutorial is available, which covers most common programming languages.

Important Note

We kindly ask users to submit NO MORE THAN 25 JOBS AT THE TIME AND NOT TO SUBMIT MORE JOBS UNTIL YOU HAVE OBTAINED RESULTS FOR THE LAST 25. There are many people using these services at present and a fair share policy has been implemented that allows us to block users that submit jobs in a manner that prevents others from using the service. This block may affect access to the EBI Web Services for an entire organisation or a class B or C subnet. Also make sure you USE A REAL EMAIL ADDRESS in your submissions. Using a fake email means we cannot contact you and will very likely result in your jobs being killed and your IP, Organisation or entire domain being black-listed. We do apologise for any inconvenience this may cause.
EMBRACE: BioMart

- Originally developed as an interface to ENSEMBL
- Federated database structure easy to query and maintain
- Web-based MartView for users to browse data
- MartShell for command line queries using SQL
- MartLib for Java applications

- EMBRACE emphasises web services
- MartService provides SOAP document/literal interface
- Also REST style web services for large scale results

- Common interface for all
EMBRACE: Analysis tools

• Originally planned sequence and sequence related tools
• EBI providing:
  – EMBOSS package
  – EMBASSY (EMBOSS-associated) third-party tools
  – SoapLab web services
• Other partners providing
  – Pattern and motif analysis/prediction
  – Protein structure modelling and binding site prediction
  – BioConductor services
  – GUI interfaces (Utopia/Cinema)
  – 3D image analysis
  – Systems biology modelling tools
EMBRACE: EMBOSS

• European Molecular Biology Open Software Suite
  – Open source sequence analysis
  – Written in C
  – 200+ applications
  – Additional 100+ ported applications (PHYLIP, HMMER, ViennaRNA)
  – Interface definition files (ACD)
  – Internal use of ontologies (e.g. Sequence Ontology)
  – Command line and 100+ other interfaces/wrappers

• Future plans
  – Next-Generation Sequence data analysis
  – Sequence metadata management
  – Multiple external resource integration
BioCatalogue Registry

<< Go to www.biocatalogue.org

You are here: Home » public

Show pagesource  Old revisions

Recent changes  Index  Login

BioCatalogue

"The Life Science Web Services Registry"

BioCatalogue will provide a curated and comprehensive catalogue of biological web services, thereby enabling users (people and programs) to discover and use these services easily. It also aims to provide a platform with several (standardized) interfaces and a suite of tools for registration of services by the community of users as well empower the community to extend and enhance the system. BioCatalogue will provide a centralized biological web services market place which will be accessible to the world as it will be searchable and indexable to search engines. BioCatalogue will provide a quality of service standard for biological web services thereby enabling services to be classified and checked for availability, reliability and other quality measures.

BioCatalogue is a BBSRC funded project and has been running since 1st June 2008. The project is a joint venture between the EMBL-EBI (led by Rodrigo Lopez) and the myGrid project at the University of Manchester (led by Carole Goble).
Ontologies

• **EMBRACE needs ontologies**
  - Definitions of data types
  - Definitions of syntax (e.g. Sequence formats)
  - Definitions of tool functionality / algorithms

• **Existing ontologies and controlled vocabularies**
  - MyGrid
  - BioMOBY
  - EMBOSS application definition files

• **None of these can define all EMBRACE services**
Ontologies

• Building the ontologies at EBI
  – Data types
  – Data formats
  – Tool functions
  – Terms, definitions and hierarchy
  – Other sources:
    • BioMOBY
    • MyGrid

• First release (EDAM) planned May 2009

• OBO Open BioMedical Ontologies
  – www.obofoundry.org
  – Common OBO standards
  – Cross-references with other relevant ontologies
Defining Service Metadata

- **Self-describing services:**
  - EB-eye services search any EBI data resource
    - Each has its own output data types, formats, and query terms
    - Output format can be selected by the user
  - Simplest to define by example
    - More difficult for automatic workflow building
    - Very easy for human workflow editors
Defining Service Metadata

• **Automatic input detection**
  – EMBOSS tools can accept any standard sequence format
    • 40+ formats (and many sub-formats e.g. FASTA id/description lines)
  – Some tools accept protein or DNA sequences
  – Can be defined as “sequence data”
  – ... But this confuses automatic workflow building

• **Multiple output formats**
  – One format by default
  – 40+ alternative formats for sequence data
  – Formats for alignments, features, phylogenetic trees, ...
EMBRACE Partners

www.embracegrid.info

Graham Cameron, Peter Rice, Alan Bleasby — EBI, Cambridge, GB
Toby Gibson — EMBL, Heidelberg, DE
Andreas Gisel — Institute of Biomedical Technologies, Section Bari, CNR, IT
Teresa Attwood — University of Manchester, GB
Marco Pagni—Swiss Institute of Bioinformatics, CH
Erik Bongcam-Rudloff — LCB/BMC, Uppsala, SE
Vincent Breton — CNRS, Clermont Ferrand, FR
Søren Brunak — CBS, Lyngby, DK
José-María Carazo — CNB, Madrid, ES
Arne Elofsson — DBB, Stockholm, SE
Daniel Kahn — INRA/CNRS, Toulouse, FR
Ralf Herwig — MPI für Molekulare Genetik, Berlin, DE
Eija Korpelainen — CSC, Espoo, FI
Christine Orengo — University College London, GB
Yitzhak Pilpel — Weizmann Institute of Science, IL
Gert Vriend — CMBI, Nijmegen, NL
Alfonso Valencia — INTA-CAB, Madrid, ES
Christian Bryne — University of Bergen, NO
Acknowledgments

Taverna:
  Tom Oinn ... and his many collaborators
SoapLab,
  Martin Senger, Mahmut Uludag
EMBOSS:
  Alan Bleasby, Jon Ison, Lisa Mullan, Ian Longden, Bijay Jassal, Thomas Laurent (EBI),
    Gary Williams, Claude Beezley, Hugh Morgan (RFCGR)
    Tim Carver, Jaina Mistry (Sanger)
EMBRACE:
  Alan Bleasby, Jon Ison, Mahmut Uludag, Shaun McGlinchey, Syed Haider,
    Graham Cameron, Kerstin Nyberg (EBI), Toby Gibson (EMBL)
    Vincent Breton (CNRS France), Erik Bongcam-Rudloff (LCB Sweden),
    Gert Vriend (CMBI, Netherlands), Steve Pettifer (Manchester)
BioMart:
  Arek Kasprcyk
BioCatalogue:
  Rodrigo Lopez, Carole Goble (Manchester)