

WebLab: a data-centric, knowledge-sharing bioinformatic platform

Wang Jun 王珺

Center for Bioinformatics

Peking University

wangjun@mail.cbi.pku.edu.cn

Contents

1

Introduction

2

Data Management

3

Analysis Service

4

Team Work

5

Others

Contents

1

Introduction

2

Data Management

3

Analysis Service

4

Team Work

5

Others

Introduction

Why great demands are proposed for an integrative system?

1

Large biological data are produced

2

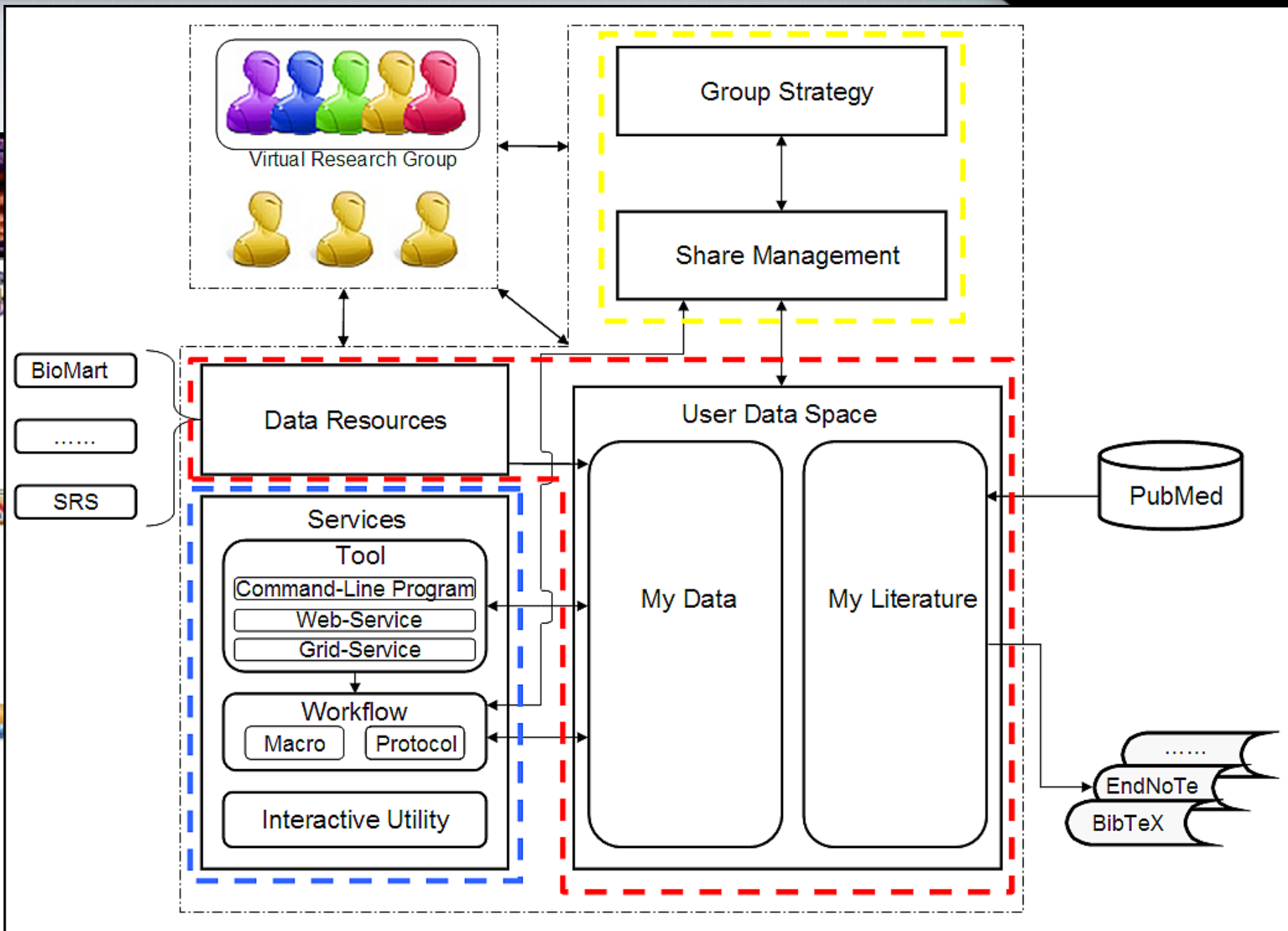
Thousands of literatures are pouring

3

Plentiful complex bioinformatic analysis tools

4

Collaboration of researchers from various fields



Contents

1

Introduction

2

Data Management

3

Analysis Service

4

Team Work

5

Others

Data Management: "My Data"

(A)

bar-headed goose greylag goose

"Tag" View

- greylag goose 1fawa.gp bio:seq:genbank
- greylag goose 1faw.pdb unknown

delete tag greylag goose >>My Data

(B)

Edit Comment Delete Comment Edit Tag Attach Tag Edit Format

+Expand all -Collapse all

- My Data my data help
- Hemoglobin
 - data
 - bar-headed goose 1a4f.pdb unknown
 - bar-headed goose 1a4fa.gp bio:seq:genbank
 - greylag goose 1fawa.gp bio:seq:genbank
 - greylag goose 1faw.pdb unknown
 - 8hba.fasta unknown
 - results
 - 8 vertebrate hemoglobin alpha chain
 - bar-headed_graylag.needle bio:align:emboss:pair
 - bar-headed_graylag.diffseq bio:report:emboss:diffseq
 - bar-headed_graylag.gff bio:feature:gff

Data can be uploaded locally or retrieved from remote database.

"Comment" knowledge

New Directory New File Delete Move Rename

Share Cancel Share Import Cancel Import Refresh

Data Management: "My Literature"

+Expand all -Collapse all

(A)

- My Literature [my literature help](#)
 - 2D
 - Secondary Chou-PY_Biochem_1974_2D.pdf [HTML View](#) | [Fetch Citation](#)
 - Secondary Kabat-EA_PNAS_1973_2D.pdf [HTML View](#) | [Fetch Citation](#)
 - Secondary Fasman-GD_Biophys-J_1976_Histone.txt [HTML View](#) | [Fetch Citation](#)
 - 3D
 - Tertiary Anfinsen-CB_Science_1973_Fold.pdf [HTML View](#) | [View Citation](#) | [Delete Citation](#) | [Export Citation](#)
 - Tertiary Sternberg-MJ_Nature_1978_Review.pdf [HTML View](#) | [View Citation](#) | [Delete Citation](#) | [Export Citation](#)
 - Tertiary Kabat-EA_PNAS_1972_3D.doc [HTML View](#) | [Fetch Citation](#)
 - blast
 - Altschul-SF_JMB_1990_BLAST.pdf [HTML View](#) | [Fetch Citation](#)
 - Lipman-DJ_NAR_1984_Statistics.pdf [HTML View](#) | [Fetch Citation](#)
 - Pearson-W_MethEnz_1990_FASTA.pdf [HTML View](#) | [Fetch Citation](#)
 - Pearson-W_PNAS_1988_Search.pdf [HTML View](#) | [Fetch Citation](#)
 - Smith-TF_JMB_1981.pdf [HTML View](#) | [Fetch Citation](#)
 - Wilbur-WJ_PNAS_1983_Search.pdf [HTML View](#) | [Fetch Citation](#)
 - Needleman-S_JMB_1970_Global.pdf [HTML View](#) | [Fetch Citation](#)

HTML View

WebLab will fetch and then index the citation information for **Kabat-EA_PNAS_1973_2D.pdf**

through **pubmed ID**

Pubmed ID (like 11748933):

through **title**

Title:

Fetch citation information

Citation information for **Kabat-EA_PNAS_1973_2D.pdf** **(D)**

Proc Natl Acad Sci U S A.1973-May;70(5):1473-1477.

The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in protein comparison of predicted and experimental determination of β -sheets in concanavalin A.

A Kabat, T Wu.

A 20 x 20 table of tripeptides has been compiled that may be used to locate beta-sheet breaking and alpha-breaking residues in proteins. It is based on the definition of an alpha-helical and a beta-sheet domain (Psi) map based on the occurrences of alpha-helices and beta-sheets in 12 known proteins whose sequence three-dimensional structure have been determined. Each entry in the 20 x 20 table lists three numbers, the occurrences of the middle amino acid (n) in relation to its nearest neighbors (n - 1) and (n + 1) in the domain, the beta-sheet domain and outside these regions. The regions between two beta-sheet-breaks would be permissively beta-sheet regions. The sequence of concanavalin A has been examined in this manner. The 13 beta-strands defined by x-ray crystallography, 10 were in agreement with the permissively beta-strand, and in the remaining three, beta-sheet-breaking residues were the third in one, and the third, fourth, a residues in another, and the sixth residue in the third from the beginning of the beta-strands. The finding strong support for the role of nearest-neighboring amino acids in determining secondary structure of protein.

PMID: 4514316

Export citation information for **(E)**

Kabat-EA_PNAS_1972_3D.pdf
Anfinsen-CB_Science_1973_Fold.pdf

to

- Endnote
- BibTeX
- ADS reference format
- ISI format
- RIS format
- Word 2007 bibliography format

Export citation

Full Text Search **(F)**

Full Text:

Refine Search

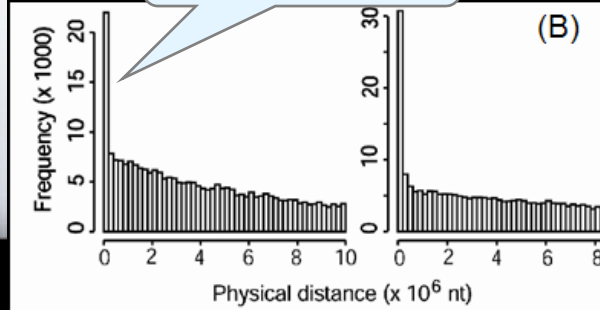
Title:

Authors:

Journal:

Publish Date: Specify a date range from To

Search against own literature space



Under the directory **Literature Home**, using search expression **-body:(structure prediction)+title:(protein)+publishdate:[1970-01 TO 1995-01]**, you get the following search results:

total count: 2 >>Advanced Search PageSize: **(G)**

- Anfinsen-CB_Science_1973_Fold.pdf score:1.0
*Fulltext: ... The studies on the renaturation of fully denatured ribonuclease required many supporting investigations (6-8) to establish, finally, the generality which we have occasionally called (9) the "thermodynamic hypothesis." This hypothesis states that the three-dimensional structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other components...
*Title: Principles that govern the folding of protein chains.
*Publish Date: 1973-07-20
- Jessen_t_pnas_1991_hb.pdf score:0.7513678
*Fulltext: ... so that a hydrophilic side chain would protrude into the surrounding water (9). None of these substitutions is likely to affect the...
*Title: Adaptation of bird hemoglobins to high altitudes: demonstration of molecular mechanism by protein engineering.
*Publish Date: 1991-08-01

Search results

Data Management: "My MetaPackage"

+Expand all -Collapse all

- My MetaPackage my metapackage help
 - reElectronicPCRdb
 - reverseEPCR weblab:metadata:reElectronicPCRdb
 - dnaDB weblab:metadata:reElectronicPCRdb
 - blastdb
 - hba7-formatdb weblab:metadata:blastdb:protein

Define your own databases

blastp(v2.2.15) - Search protein database using a protein query

+add to ToolBox

Compares an amino acid query sequence against a protein sequence database

Input

* Query file:

Output

* save result in directory:

* Output file name: (bio:align:blast)

Basic Options

* Choose database:

* Filter low complexity regions:

* Expect:

* Matrix:

- Swissprot protein sequences (swissprot)
- built-in databases**
 - Swissprot protein sequences (swissprot)
 - Non-redundant protein sequences (nr)
 - Protein Data Bank proteins (pdb)
- user-created databases**
 - hba7-formatdb**

Blast against your own databases

Contents

1

Introduction

2

Data Management

3

Analysis Service

4

Team Work

5

Others

Analysis Service: "Program"

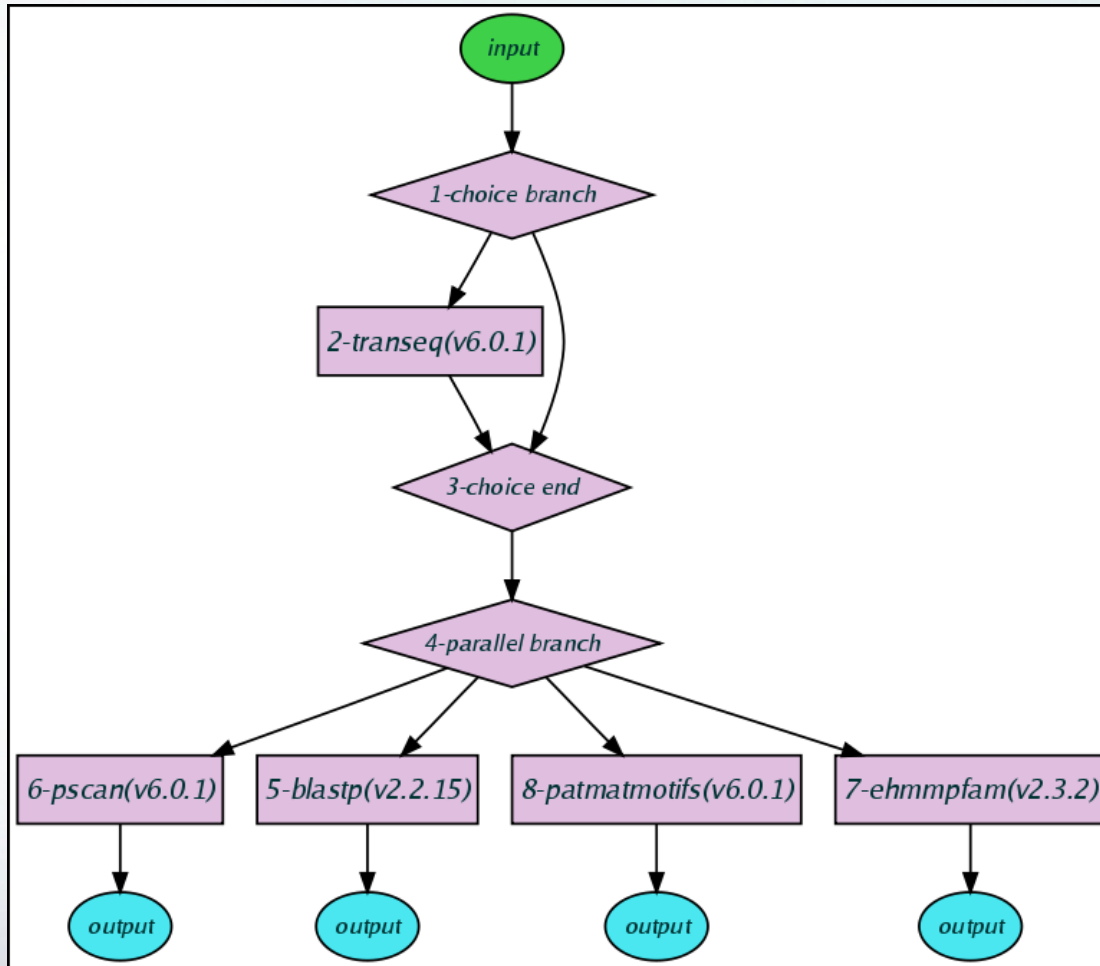
Package View | **Function View** | Alphabet View

+ Expand all - Collapse all

- [-] Protein
 - [+] Structure
 - [+] Similarity Detecting
 - [+] Sequence Feature Identifying
 - [+] Conceptual Back Translation
 - [+] Enzyme
 - [+] Phylogeny Reconstruction
 - [+] Mass Spectrometry Analysis
 - [+] Sequence Feature Display
 - [+] Mutation
- [-] Data Manipulation
 - [+] Database Retrieval
 - [+] Data Display
 - [+] Alignment Editing and Visualization
 - [+] Sequence Edit
 - [+] Format Converting
 - [+] Sequence Information Display
 - listor(v6.0.1) - Write a list file of the logical OR of two sets of sequences
 - [+] Random Sequence Generation
- [+] Literature
- [+] Ontology
- [+] DNA/RNA
- [+] Pathway
- [-] Sequence Comparison
 - [+] Multiple Sequence Alignment
 - [+] Database Searching
 - [+] Analysis of Aligned Sequences
 - [+] Pairwise Sequence Alignment
 - [+] Alignment Editing and Visualization
- [+] Expression
- [-] Database Construction
 - [+] Blast
 - [+] Electronic PCR

- **264 tools in total**
- **Uniform interface**
- **Three types of tools:**
 - (1) Command-line programs**
 - (2) Web-services**
 - (3) Grid-services**


Analysis Service: "Workflow"



- **Workflow**
Complex analysis tasks can be run automatically
- **Two types:**
Macro
Protocol
- **Customized**


Analysis Service: "Resource"

BioMart



BioMart is a query-oriented data management system developed jointly by the Ontario Institute for Cancer Research (OICR) and the European Bioinformatics Institute (EBI). Please click [here](#) to enter the query page for the BioMart system integrated in WebLab.

PlantTFDB



Plant Transcription Factor Databases

A comprehensive database which contains information on transcription factors including five plant genomes, gene sequences, and gene expression data. Please click [here](#) to enter the query page for the PlantTFDB system integrated in WebLab.

SRS ID/AC search

Choose database:

Through:

FASTA format:

Each line of FASTA format:

SRS keywords search

Choose database:

FASTA format:

One or several keywords:

If the number of result entries is no more than 100, all of them are provided; otherwise, only 100 entries are provided.

Dataset:

Attribute:

- Features
 - GENE:
 - Ensembl Gene ID
 - Description
 - Gene End (bp)
 - Transcript End (bp)
 - External Transcript ID
 - Ensembl cDNA length
 - % GC content
 - Ensembl Transcript ID
 - Chromosome Name
 - Strand
 - Gene name
 - External Transcript DB
 - Ensembl Peptide length
 - Biotype
 - Ensembl Peptide ID
 - Gene Start (bp)
 - Transcript Start (bp)
 - Gene DB
 - Ensembl CDS length
 - Transcript count
 - Source
 - EXTERNAL:
 - PROTEIN:
 - GENOMIC REGION:
 - Homologs

Filter:

gene_id:

transcript_id:

mouse_gene_id:

Chromosome name:

Gene Start (bp):

Gene End (bp):

Strand:

Chromosome Regions:

id_list_filters:

only excluded

We provide uniform interface for different data sources.

Search

Analysis Service: "Utility"

Utility includes some mini popular bioinformatics analysis tools written in javascript or applet. If you know some other applications which are pretty useful, please give us some advice at your free time. [utility help](#)

+ Expand all - Collapse all

- Bioverse
- Dotlet
 - Dotlet(applet) Dotlet is a program for com
- GlycanBuilder
- Vimida
- Sequence Manipulation Suite
 - Sequence Analysis
 - Sequence Figures
 - Format Conversion
 - Miscellaneous
 - Random Sequences
- jalview
 - jalview(applet) Jalview is a multiple alignm
- WebMol

The screenshot shows a Jalview applet window displaying a multiple sequence alignment. The window title is "/utility/jalview/uniref50.fa". The menu bar includes File, Edit, Select, View, Format, Colour, Calculate, and Help. The alignment is shown with columns numbered 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. A red box highlights the region from position 10 to 50. The sequences are color-coded by conservation, with yellow indicating high conservation and blue indicating low conservation. The sequences listed on the left are FER_CAPAA, FER_CAPAN, FER1_SOLLC, Q93XJ9_SOLTU, FER1_PEA, Q7XA98_TRIPR, FER1_MESCR, FER1_SPIOL, FER3_RAPSA, FER1_ARATH, FER_BRANA, FER2_ARATH, Q93260_ARATH, FER1_MAIZE, and O80429_MAIZE. The status bar at the bottom shows "Sequence 9 ID: FER3_RAPSA" and "Java Applet Window".

Contents

1

Introduction

2

Data Management

3

Analysis Service

4

Team Work

5

Others

Team Work: "Group Strategy"

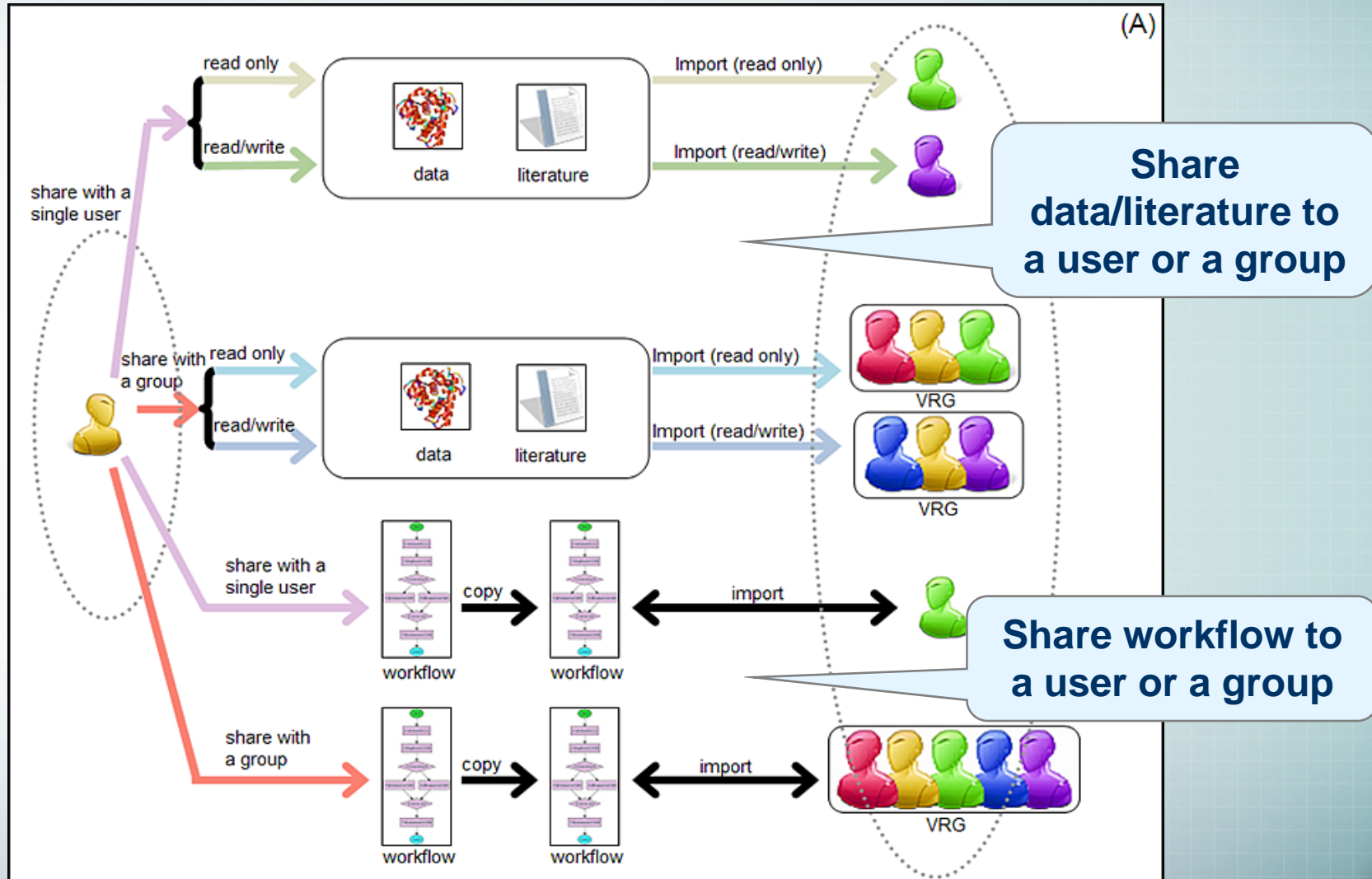
User Created Group				>>Create Your Group
NAME	DESCRIPTION	MEMBERSHIP		ACTION
text-mining	BioNLP	wangjun	gssbc07b_f5	delete edit
Microarray analysis	BioConductor	luojc	wangjun gaoge	

User Attended Group				
NAME	DESCRIPTION	CREATOR	MEMBERSHIP	ACTION
weblab-dev		gssbc07b_f5	kongl wujm wangjun zhaosq liuxq gaoge alygunia liuxc	quit
text-mining	BioNLP	wangjun	wangjun gssbc07b_f5	quit
Microarray analysis	BioConductor	wangjun	luojc wangjun gaoge alygunia	quit

User can create a Virtual Research Group (VRG)

User attended VGRs

Team Work: "Sharing Mechanism"



Contents

1

Introduction

2

Data Management

3

Analysis Service

4

Team Work


5

Others

Others: "Document & License"



WebLab User Guide

- I. Introduction
- II. User Registration 😊
- III. Organizing Your Data: User Space 
 1. My Data 😊
 2. My Literature 😊
 3. My MetaPackage
- IV. Doing Your Job in WebLab
 1. Utility - Handful Tools
 2. Program - the Build Blocks 😊
 3. Workflow (Protocol & Macro) - A Nice Bundle 😊
 4. Job controller - You're the Master of Your Jobs!
 5. My Toolbox - My Little Nice Toolkit
 6. Resource - Get Original Data 😊
- V. Working with Your Collaborators - Group Strategy and Sharing Mechanism
 1. Overview
 2. Group - You're Not Alone 😊
 3. Sharing - Do Not Re-invent the Wheel 😊

**Detailed document
and video**

**Open Source
(GPLv3)**

WebLab's whole source code is released as "Free Software" under the GNU General Public License version 3 (GPLv3), and freely available to anyone for downloading.



WebLab 1.0 Source Code

Others: “EMBOSS Link”

<http://emboss.sourceforge.net>

Web interfaces

Most web interfaces to EMBOSS provide a form for setting options and running applications. Difficulties can arise because of 'dependencies' in the application input. These include where the value of one parameter depends on the value or property of another, for example sequence length, or where the available options depend on the input, for example whether a sequence is DNA or Protein. Some web interfaces handle dependencies better than others.

To use a web interface, all you need are a web browser and internet access. Some of these interfaces run 'anonymously' so any user can run the applications. In most cases however you will be expected to register with the site before you can use their services.

The interfaces we have seen so far are:

WebLab

WebLab collect provide easy web access to the most popular bioinformatics programs. It is task-oriented instead of tool-oriented. It provides pre-defined protocols to guide biologists to work step by step through the entire analysis procedure.

Pise

PISE (french pronunciation like the english word 'peas') was developed by Catherine Letondal at the Institute Pasteur. Using an XML definition for each application, PISE generates a Web interface and other common interfaces automatically.

For EMBOSS, the ACD files were converted with scripts to generate the XML definitions. Where ACD files caused problems, we simply fixed the ACD files.

EMBOSS Explorer.

An extremely popular web interface to EMBOSS from Luke McCarthy at Plant Biotechnology Institute, Canada

Others: "BioMart Link"

<http://www.biomart.org>



HOME

MARTVIEW

MARTSERVICE

DOCS

CONTACT

NEWS

CREDITS

BioMart Project

BioMart is a query-oriented data management system developed jointly by the [Ontario Institute for Cancer Research \(OICR\)](#) and the [European Bioinformatics Institute \(EBI\)](#).

The system can be used with any type of data and is particularly suited for providing 'data mining' like searches of complex descriptive data. BioMart comes with an 'out of the box' website that can be installed, configured and customised according to user requirements. Further access is provided by graphical and text based applications or programmatically using web services or API written in Perl and Java. BioMart has built-in support for query optimisation and data federation and in addition can be configured to work as a DAS 1.5 Annotation server. The process of converting a data source into BioMart format is fully automated by the tools included in the package. Currently supported RDBMS platforms are MySQL, Oracle and Postgres.

BioMart is completely Open Source, licensed under the LGPL, and freely available to anyone without restrictions.

Powered by BioMart software:

- [BioMart Central Portal](#)
- [Ensembl](#)
- [HapMap](#)
- [HTGT](#)
- [HGNC](#)
- [Dictybase](#)
- [Wormbase](#)
- [Gramene](#)
- [Europhenome](#)
- [UniProt](#)
- [Rat Genome Database](#)
- [DroSpeGe](#)
- [ArrayExpress DW](#)
- [Eurexpress](#)
- [GermOnLine](#)
- [PRIDE](#)
- [PepSeeker](#)
- [VectorBase](#)
- [Pancreatic Expression Database](#)
- [Reactome](#)
- [EU Rat Mart](#)
- [Paramecium DB](#)

Third party software with BioMart Plugin:

[Bioclipse](#) [biomaRt-BioConductor](#) [Cytoscape](#) [Galaxy](#) [Taverna](#) [WebLab](#)