

Data, Services and Computational Resources Integration at the INB

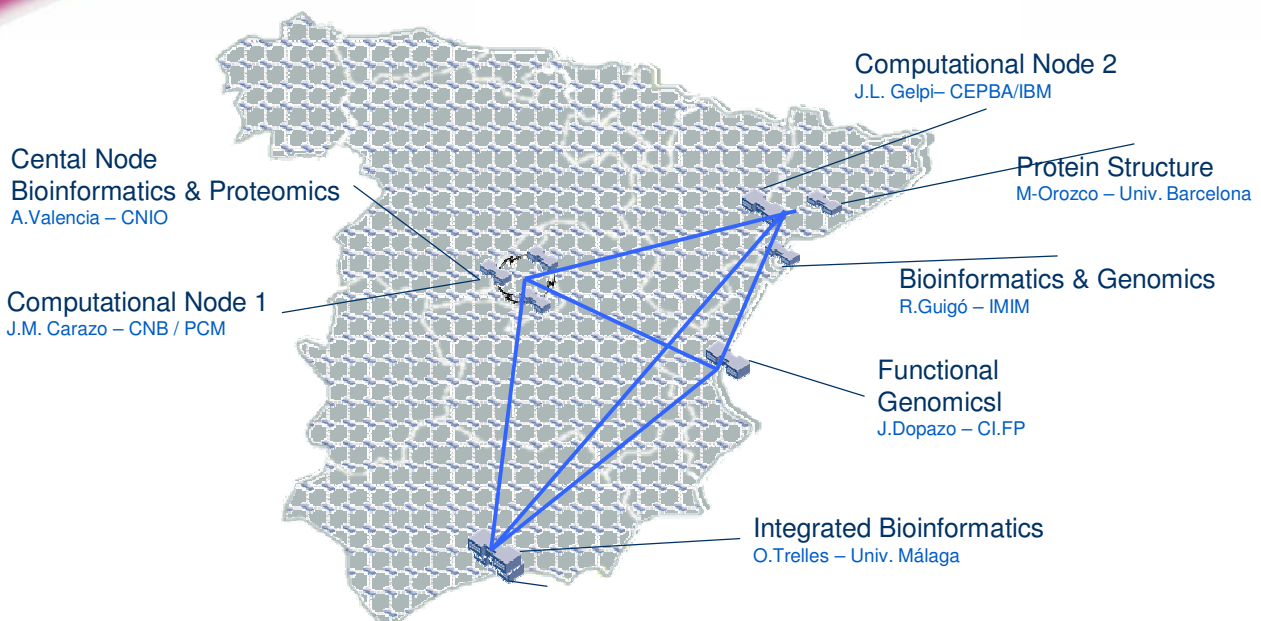


Oswaldo Trelles
University of Malaga, Spain
Integrated Bioinformatics - INB

National Institute for Bioinformatics

www.inab.org

Integrated Bioinformatics



Scalable & expandable “Virtual” Organization

The INB mission

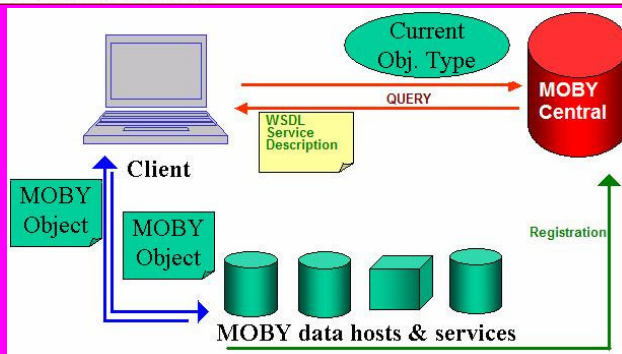
“produce and apply bioinformatic solutions in the development and workout of genomic and proteomics projects”

- Promote the **Bioinformatics dissemination** in Spain.
- Collaborate and provide **technical & scientific support** in G&P projects.
- Contribute in the **emergence and consolidation of local groups** (research or service's bioinformatics).
- Provide basic **Training in bioinformatics**.
- **Activate bioinformatics projects** from the INB.
- Promote the development and competitiveness of the **companies** in this field.
- **Internationalisation** of the INB activities.

Main lines of work at GNV5

- Integration architecture
 - MOWServ client - connecting and presenting the service offering of INB
- BioMOBY proposals
 - Protocol improvements (error handling, async, mirroring)
- Identification and restructuring of MOWServ internals
 - MOWServ is a generic client (uniform interface) but more specific clients are needed (jORCA)
- Development of software libraries

Design specifications



- **Persistence** of user data files
- **Controlled** Ontology and Services Registry
- Task **scheduling**, Robustness (long services)
- Objects & tasks **monitoring**
- **Generic Client** (**extended set of facilities**)
- **Creation** & visualization functionality
- **Help & training** support

After careful analysis, the open source standard bioMOBY (Wilkinson and Links; 2002; Wilkinson, et al.; 2003) –included its pros and cons- was chosen as the underlying protocol to sustain inter-operational conjunction of nodes.

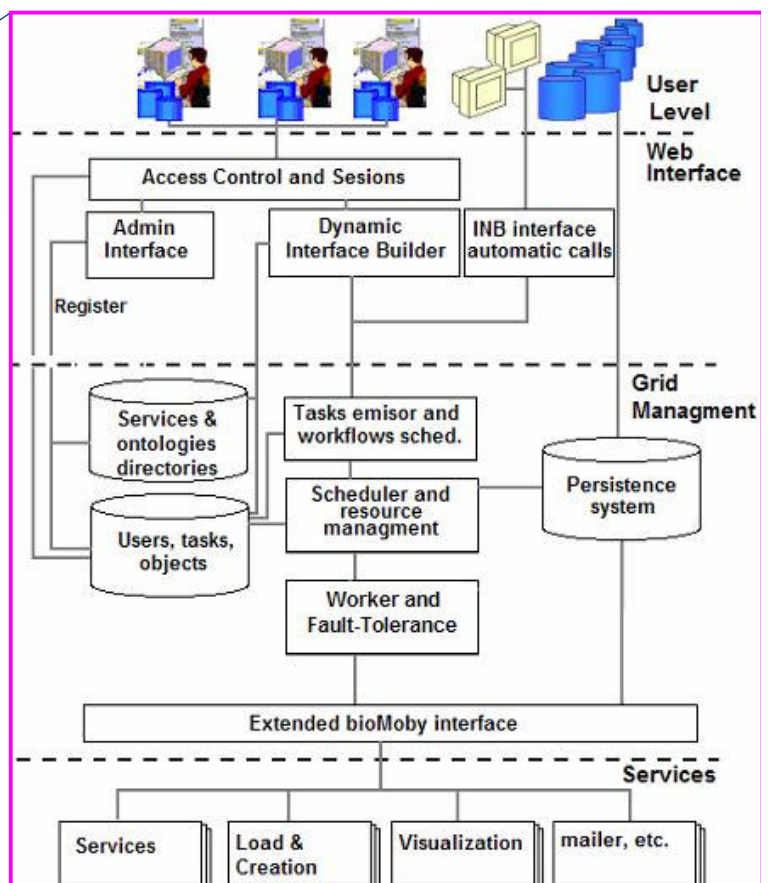
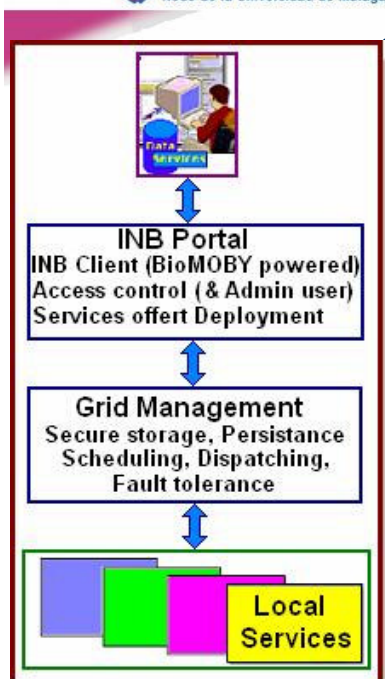
Although from the beginning we realized of some limitations in the protocol, the advantages for a rapid deployment of results as well as the increasing number of services becoming available under this platform were on top of benefits.

One of the important tasks from the INBG perspective was to extend the protocol to fulfil our objectives

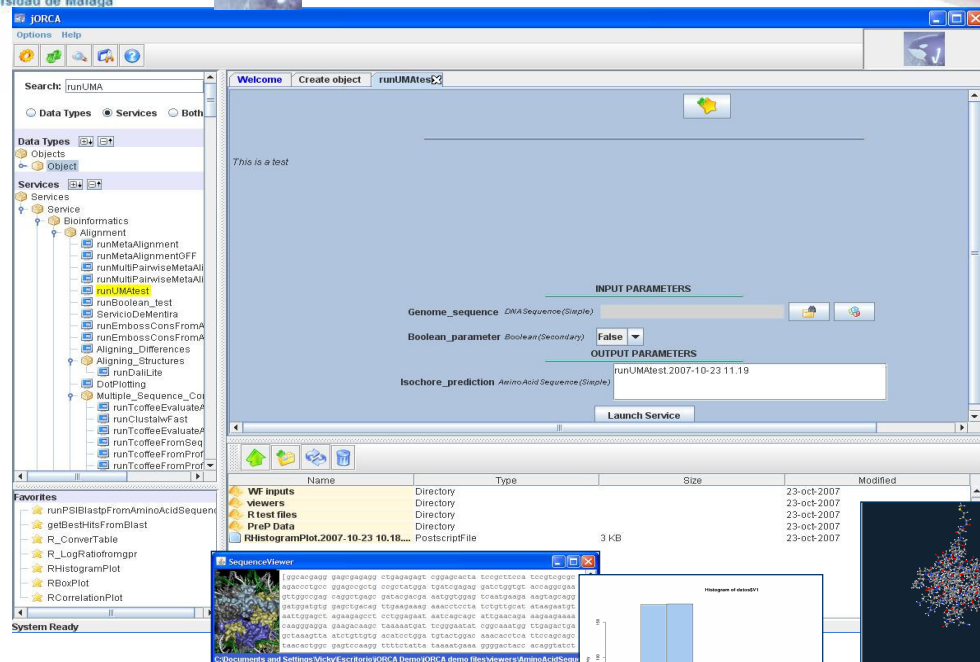
O.Trelles, BioHackaton-Japan 08

Integrated Bioinformatics, INB-UMA

The MOWServ-v1 Architecture



Integrated Bioinformatics, INB-UMA



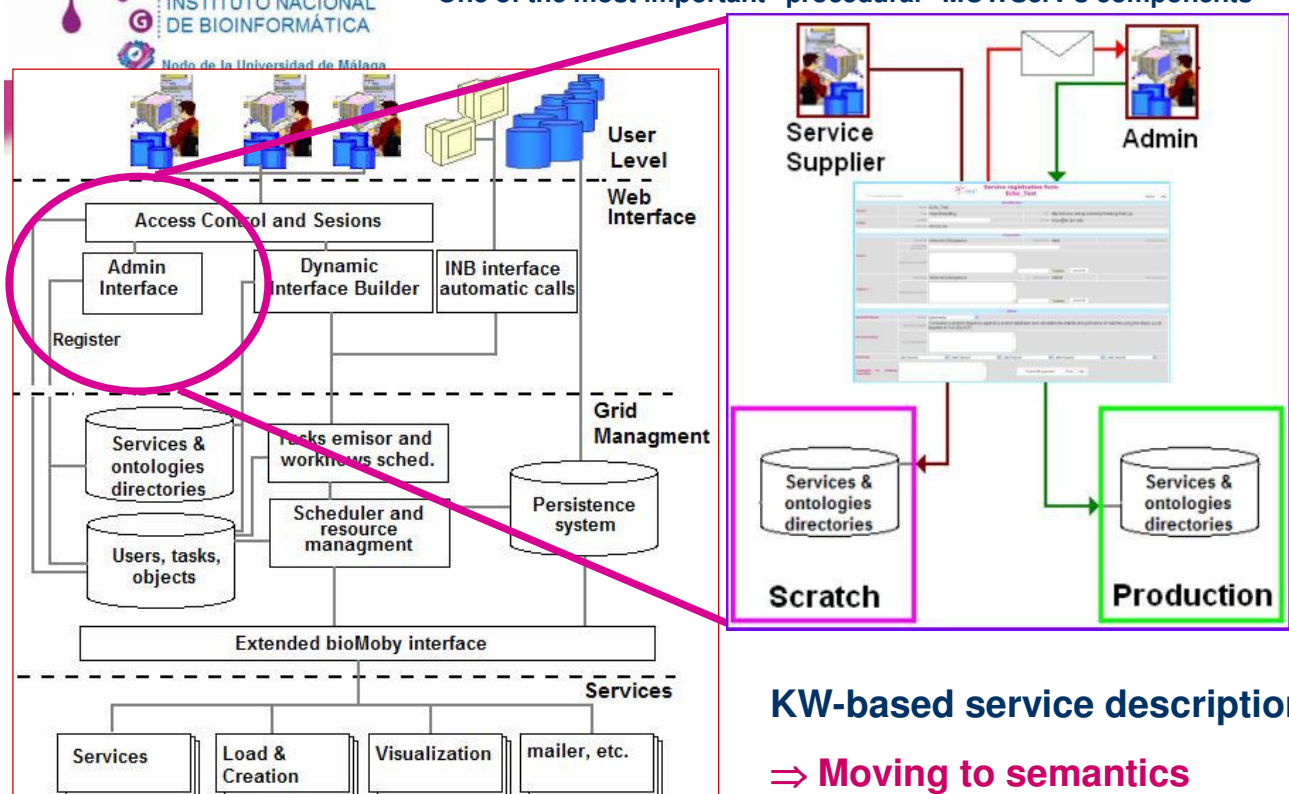
JORCA: new Desktop interactive client: management of moby, soap, grid-based and any other type of services

Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

Controlled Registry System

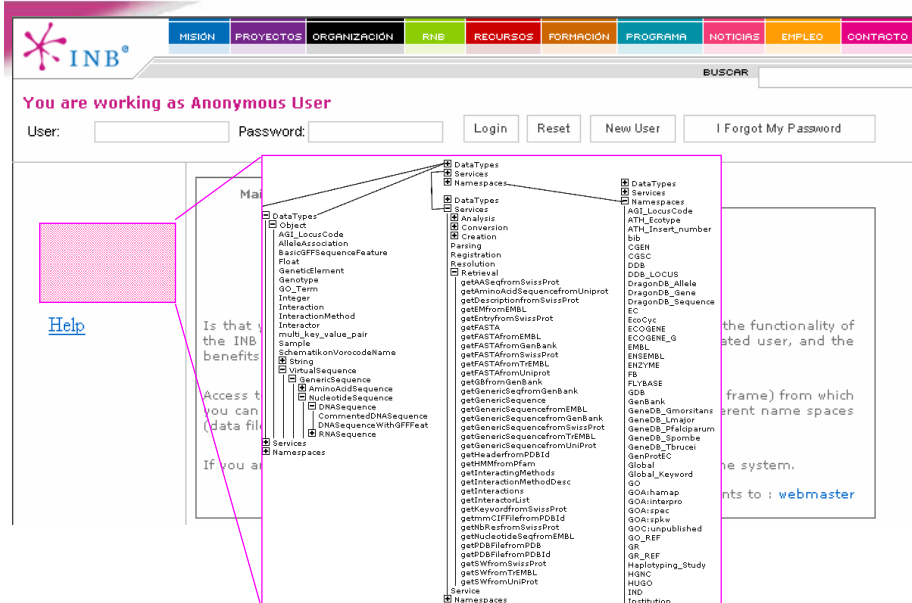
One of the most important “procedural” MOWServ’s components



KW-based service description

⇒ **Moving to semantics**

Browsing Objects, Services & Namespaces



You are working as Anonymous User

User: Password: Login Reset New User I Forgot My Password

[Help](#)

Is that v...
the INB
benefits

Access t...
you can
(data fil...
If you a...

the functionality of
ated user, and the

frame) from which
arent name spaces

ne system.

nts to : [webmaster](#)

Quality control:
registering / daily
availability /
documentation

Favourites section

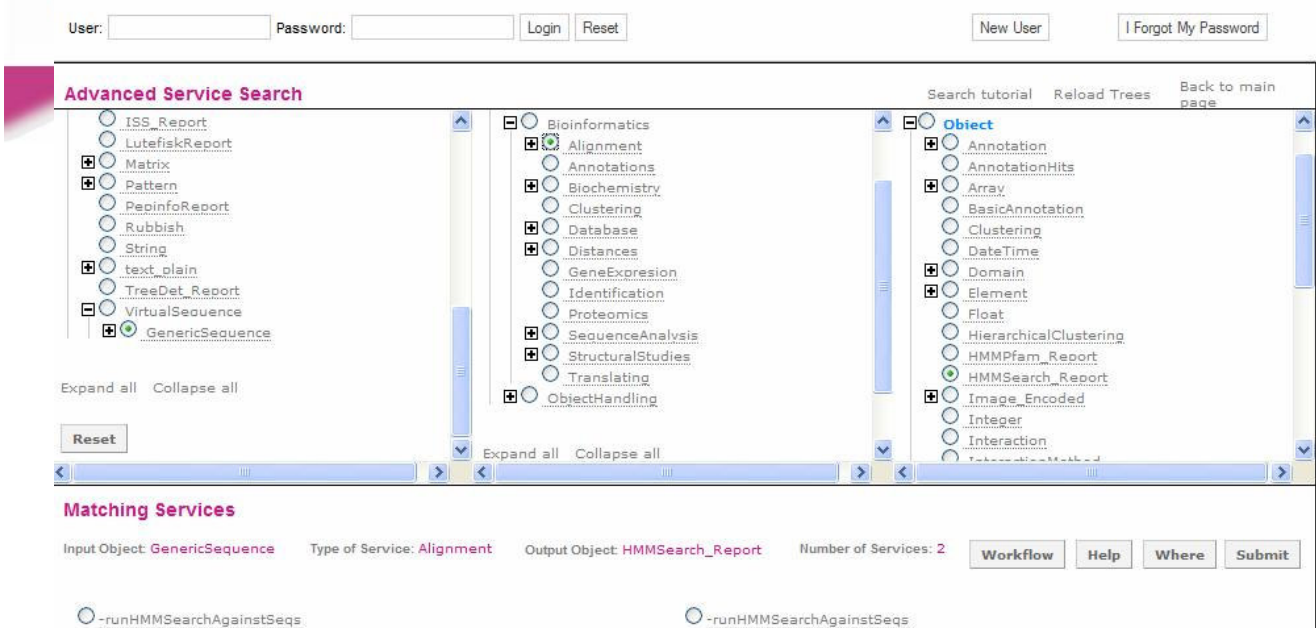
Dynamic & automatic offer
deployment in browseable tree
organization

Intelligent client for integrating bioinformatics services”, Navas-Delgado, et al.
Bioinformatics, vol.22 no.1 2006 pages 106-111

Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

Advanced search: Example



Advanced Service Search

Search tutorial Reload Trees Back to main page

Input Object: **GenericSequence** Type of Service: **Alignment** Output Object: **HMMSearch_Report** Number of Services: 2 Workflow Help Where Submit

☐ -runHMMSearchAgainstSeqs

Example: a user has a new sequence and wants to obtain a quality alignment using HMM in order to get the family sequence where the query sequence belongs. Select GenericSequence as Input Object, and next HMMSearch_Report as output object. In this way, 4 services are suggested. Then select Alignment in Service window, and only 2 services (the same with 2 mirrors): runHMMSearchAgainstSeqs, are identified which is useful for the final target.

Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

Plug-in specific

Transparent to users

O.Trelles, BioHackaton-Japan 08

Service invocation

- Different protocols (soap / Moby / grid / ...)
- JORCA: plug-in workers
- Security issues (anonymous data)
- Controlled access to resources

O.Trelles, BioHackaton-Japan 08

Tasks & Objects Monitoring

- Task monitoring
- Objects persistence
- Pipelining capabilities
- The object object problem

INB available services for ABCDhumanSEQ

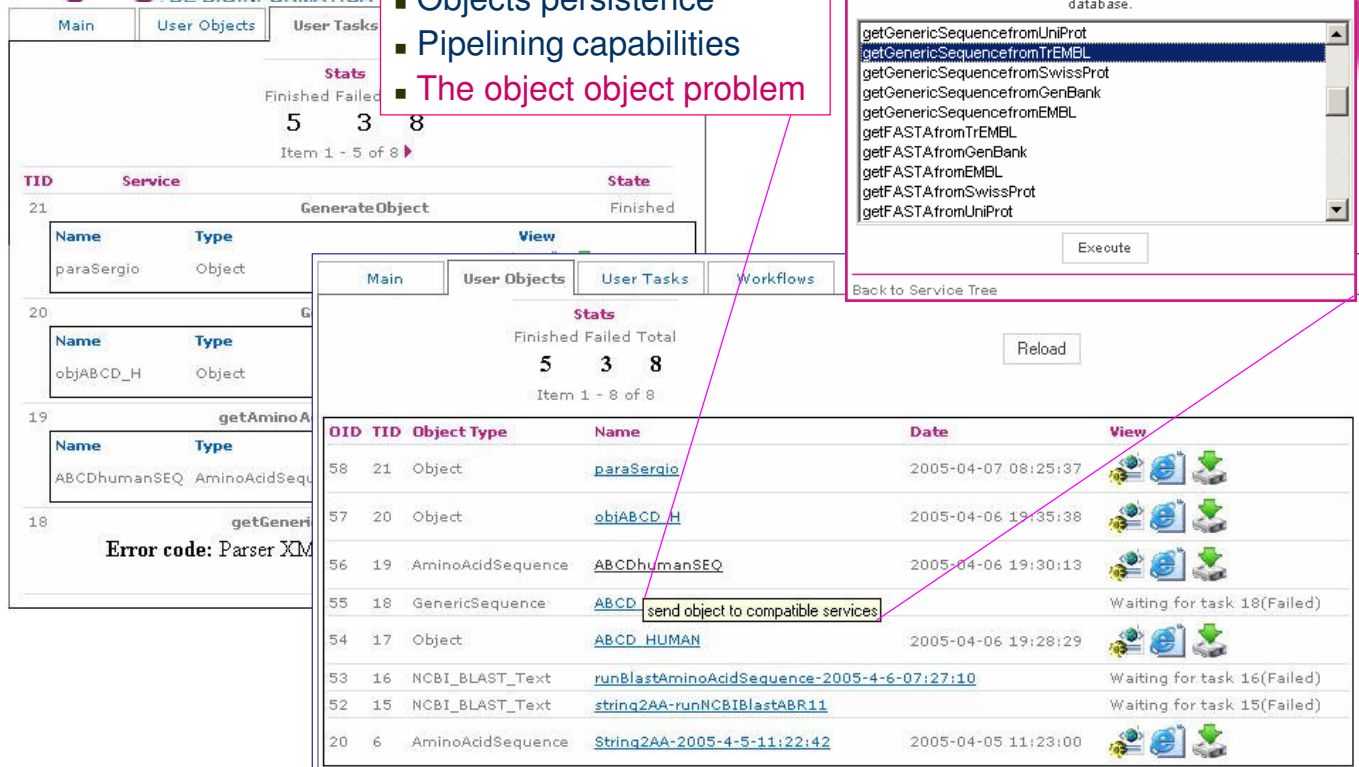
Retrieve a sequence as a GenericSequence object from UniProt-TrEMBL database.

```
getGenericSequencefromUniProt
getGenericSequencefromTrEMBL
getGenericSequencefromSwissProt
getGenericSequencefromEMBL
getFASTAfromTrEMBL
getFASTAfromGenBank
getFASTAfromEMBL
getFASTAfromSwissProt
getFASTAfromUniProt
```

Execute

Back to Service Tree

Reload



Stats
Finished Failed Total
5 3 8
Item 1 - 5 of 8

TID	Service	State
21	GenerateObject	Finished

Name **Type**
paraSergio Object

Name **Type**
objABCD_H Object

Name **Type**
ABCDhumanSEQ AminoAcidSequence

Name **Type**
ABCD HUMAN Object

Name **Type**
runBlastAminoAcidSequence-2005-4-6-07:27:10 NCBI_BLAST_Text

Name **Type**
string2AA-runNCBIBlastABR11 NCBI_BLAST_Text

Name **Type**
String2AA-2005-4-5-11:22:42 AminoAcidSequence

Error code: Parser XM

OID **TID** **Object Type** **Name** **Date** **View**

58	21	Object	paraSergio	2005-04-07 08:25:37	
57	20	Object	objABCD_H	2005-04-06 19:35:38	
56	19	AminoAcidSequence	ABCDhumanSEQ	2005-04-06 19:30:13	
55	18	GenericSequence	ABCD	2005-04-06 19:28:29	
54	17	Object	ABCD HUMAN	2005-04-06 19:28:29	
53	16	NCBI_BLAST_Text	runBlastAminoAcidSequence-2005-4-6-07:27:10	2005-04-06 19:28:29	
52	15	NCBI_BLAST_Text	string2AA-runNCBIBlastABR11	2005-04-06 19:28:29	
20	6	AminoAcidSequence	String2AA-2005-4-5-11:22:42	2005-04-05 11:23:00	

Waiting for task 18(Failed)

Waiting for task 16(Failed)

Waiting for task 15(Failed)

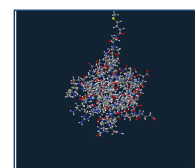
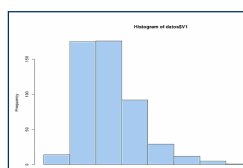
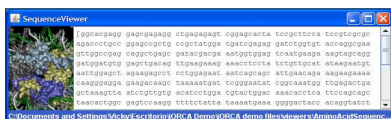
send object to compatible services

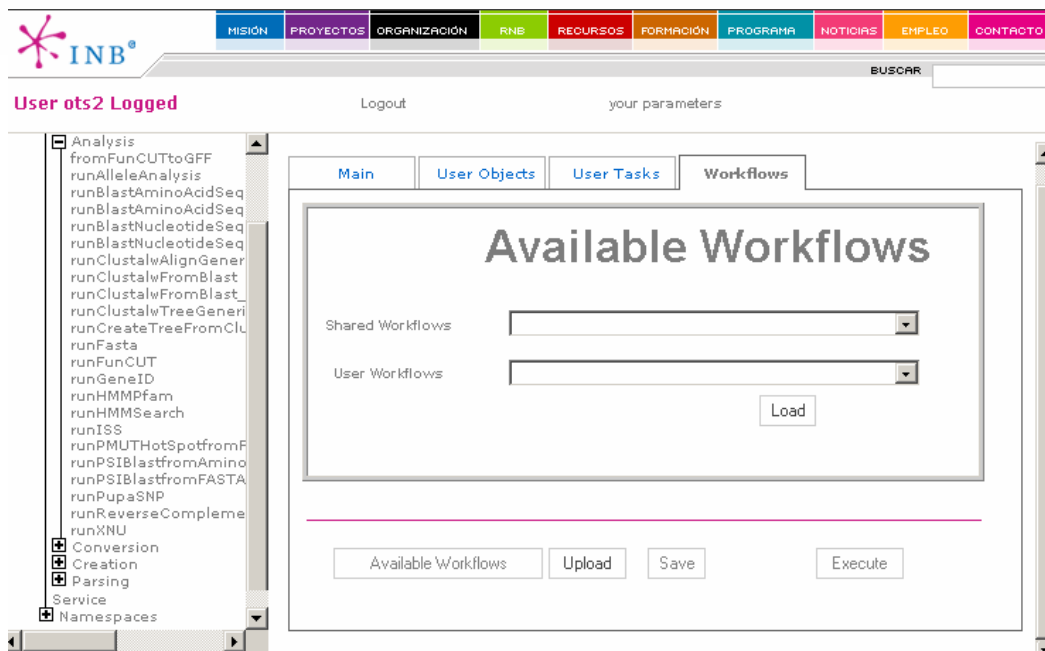
Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

Viewing results

- Plug-in based architecture:
 - Each object can be associated with a XSLT file
- Service providers can submit specialized viewers for their data (e.g. jMOL)
- Results always viewable as XML or HTML.
- Improvements: "Interactive" visualization / registering



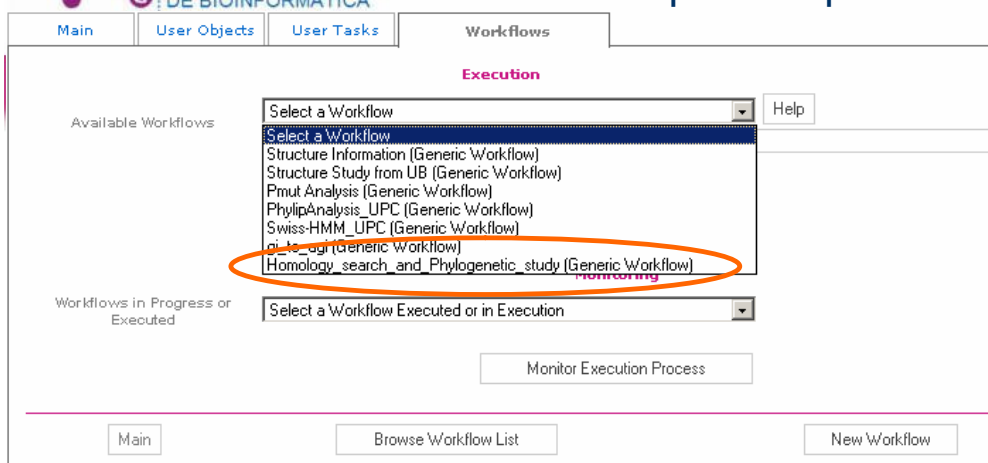


Integrated Bioinformatics, INB-UMA

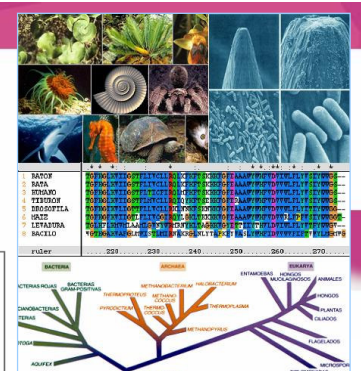
O.Trelles, BioHackaton-Japan 08

2.- WF selection

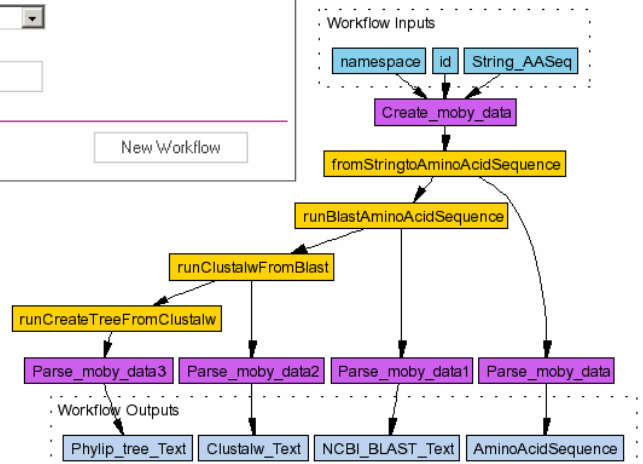
WFs tab : public / private



STEP 1: Sequence retrieval
STEP 2: Blast
STEP 3: Best hits
STEP 2: Multiple seq. alignment
STEP 3: Phylogenetic tree



Phylogenetic tree

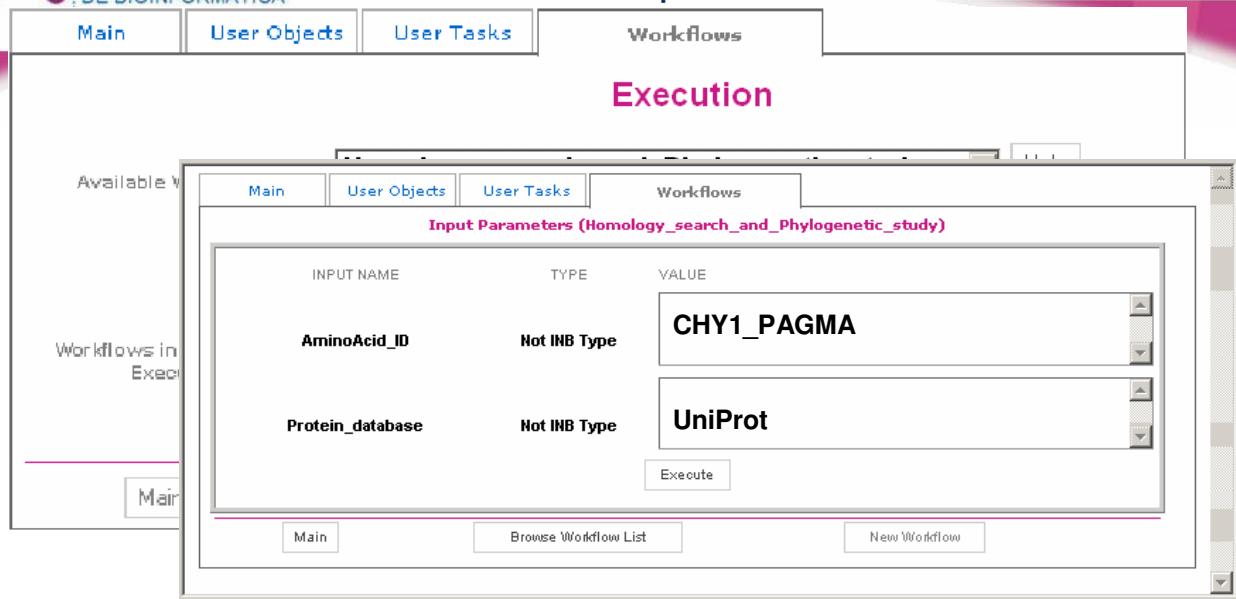


Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

3.- specialise

Input definition



Execution

Available Workflows

Workflows in Execution

Main | User Objects | User Tasks | Workflows

Input Parameters (Homology_search_and_Phylogenetic_study)

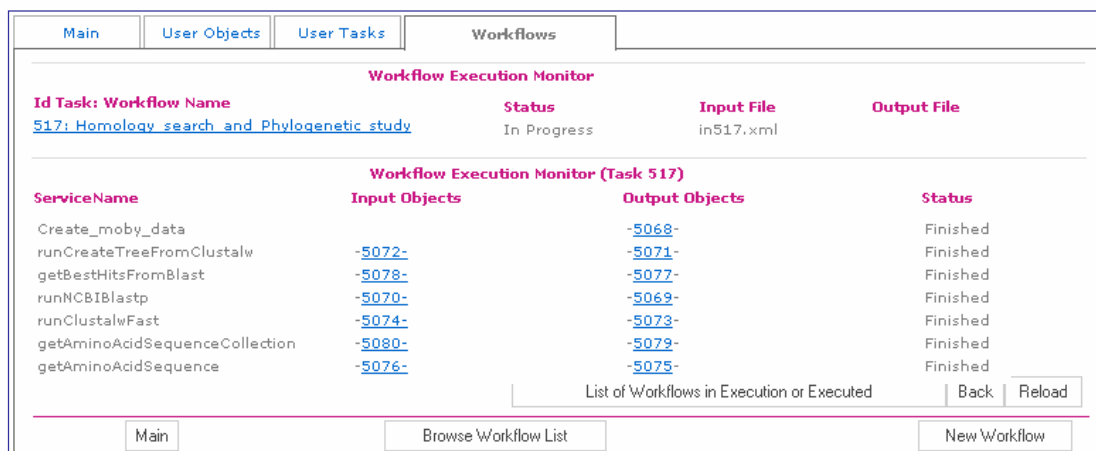
INPUT NAME	TYPE	VALUE
AminoAcid_ID	Not INB Type	CHY1_PAGMA
Protein_database	Not INB Type	UniProt

Execute

Main | Browse Workflow List | New Workflow

Launch execution

4.- Monitoring



Workflow Execution Monitor

Id Task: Workflow Name
[517: Homology_search_and_Phylogenetic_study](#)

Status
In Progress

Input File
in517.xml

Output File

ServiceName	Input Objects	Output Objects	Status
Create_moby_data		-5068-	Finished
runCreateTreeFromClustalw	-5072-	-5071-	Finished
getBestHitsFromBlast	-5078-	-5077-	Finished
runNCBIBlastp	-5070-	-5069-	Finished
runClustalwFast	-5074-	-5073-	Finished
getAminoAcidSequenceCollection	-5080-	-5079-	Finished
getAminoAcidSequence	-5076-	-5075-	Finished

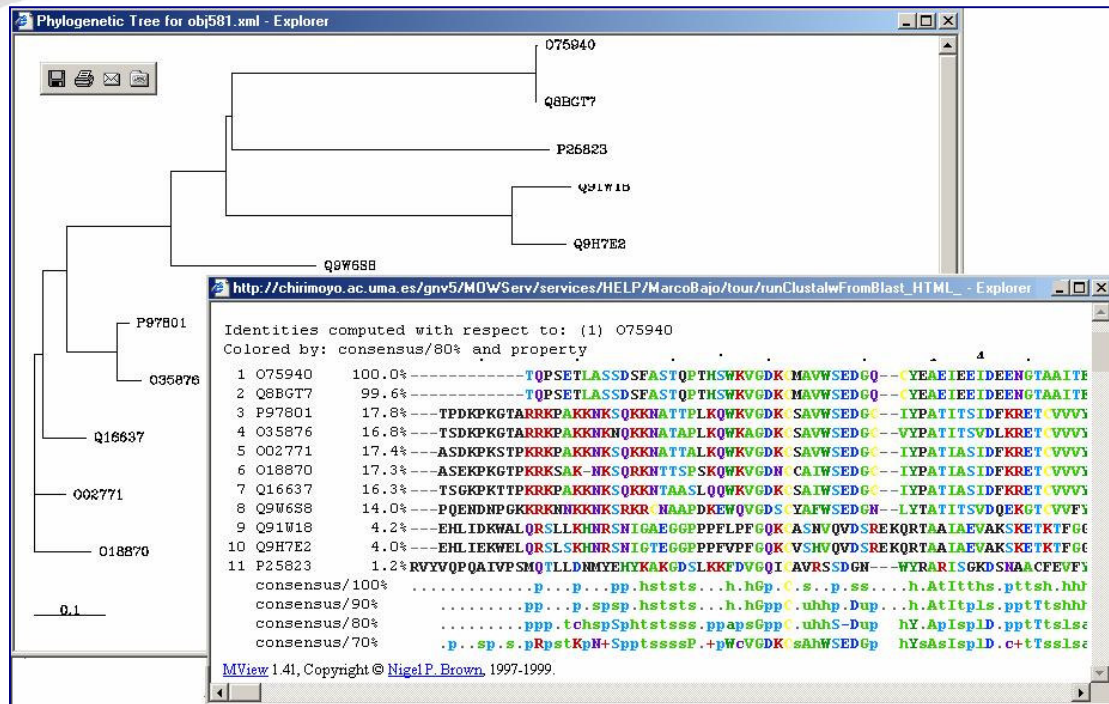
List of Workflows in Execution or Executed | Back | Reload

Main | Browse Workflow List | New Workflow

In progress

...done

5.- Results



Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

Open questions

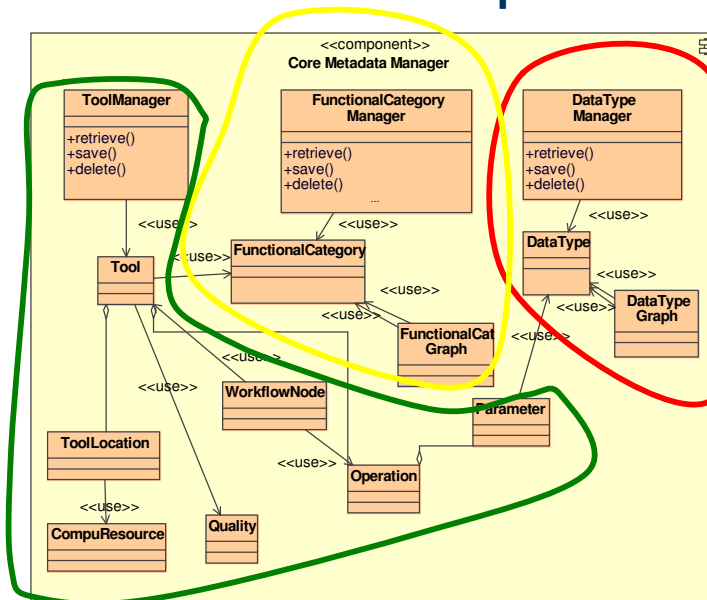
Short term

- **Asynchronous services (& progress status notification)**
- **Error handling**
- **Replication (mirroring) system**
- **Daily Service quality control**
- **Service's help system**
- **Advanced service discovering**
- **Collections**
- **Large data sets (network overloading)**
- **Indirect (by reference) pass of data (locality of D & S)**
- **Workflows (storage in the repository, WWE-jmf)**
- **Semantic annotations (datatypes & Services)**
- **Define services categories**
- **New datatypes: Gene expression; images; ...**

Open questions

long term

- **Review / redefine the repository data model**
 - **modular / distributed / users / persistence / security...**



Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

Integrated Bioinformatics Node



www.bitlab-es.com

