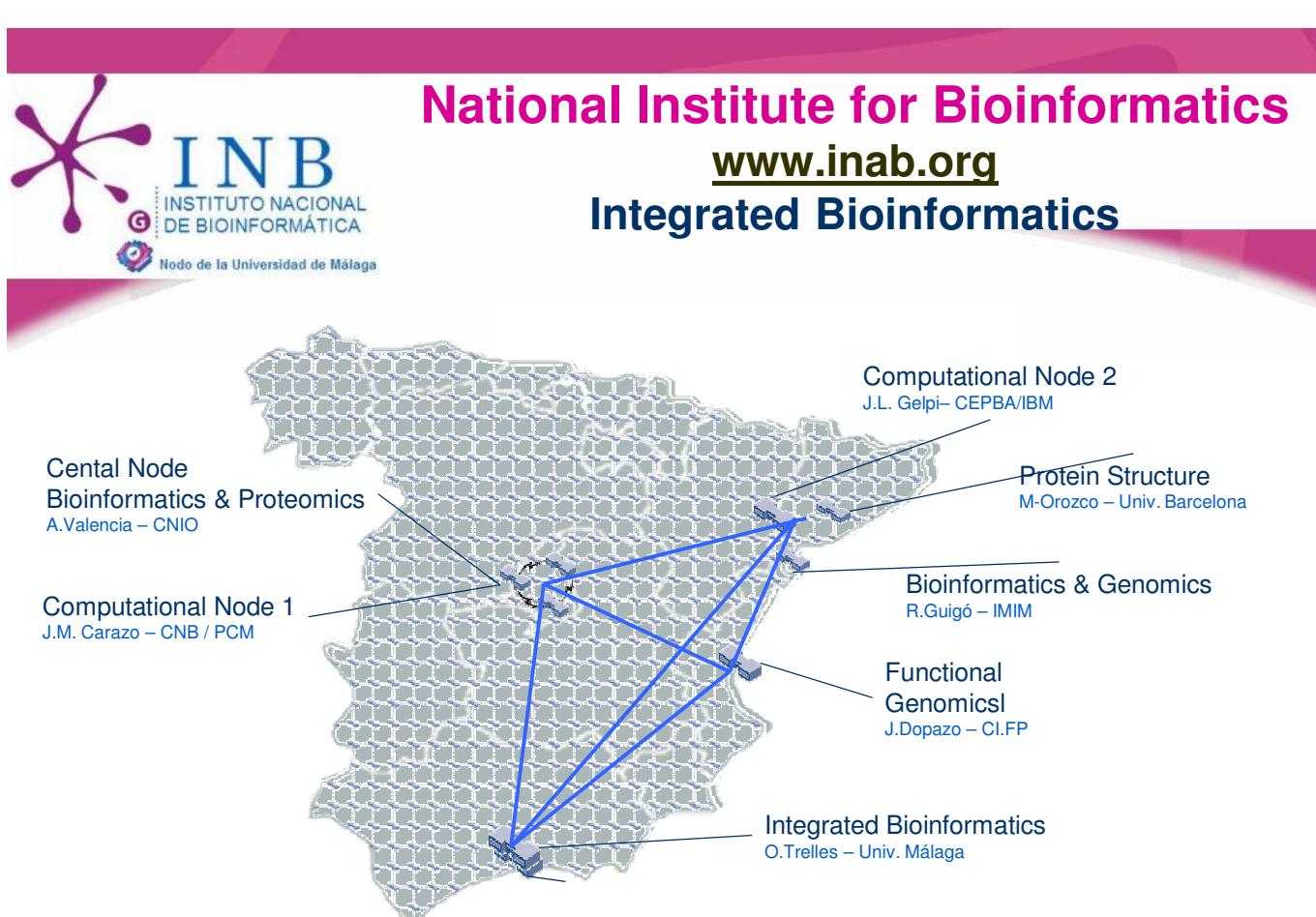


Data, Services and Computational Resources Integration at the INB



Oswaldo Trelles
University of Malaga, Spain
Integrated Bioinformatics - INB



Scalable & expandable “Virtual” Organization

The INB mission

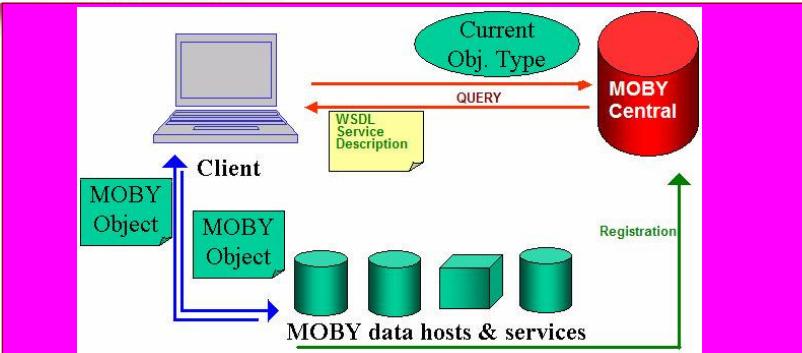
"produce and apply bioinformatic solutions in the development and workout of genomic and proteomics projects"

- 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

Integrated Bioinformatics, INB-UMA

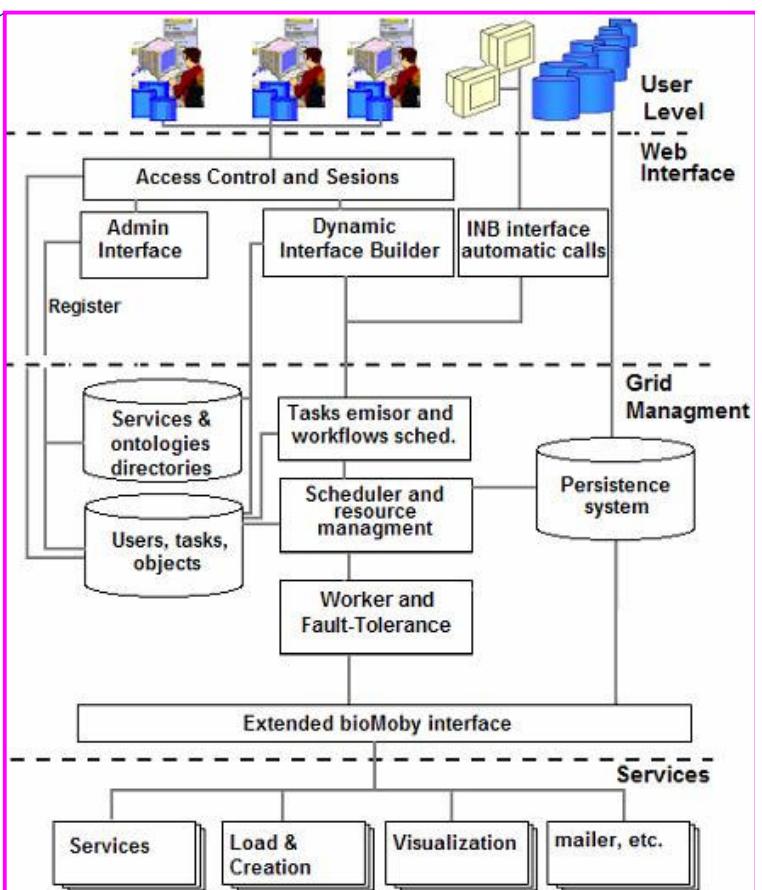
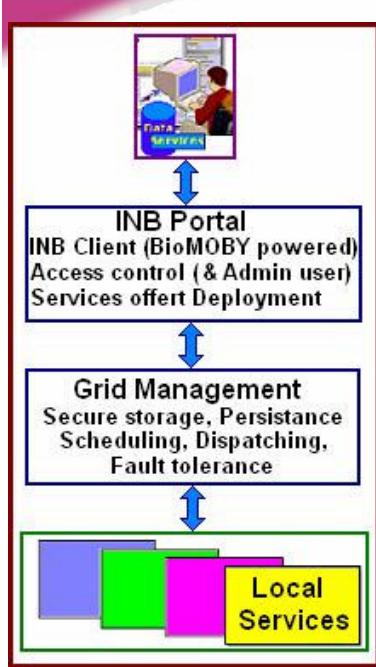
After careful analysis, the open source standard bioMOBY (Wilkinson and Links; 2002; Wilkinson, et al.; 2003) –included its pros and cons- was chosen as the underlying protocol to sustain inter-operational conjunction of nodes.

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.

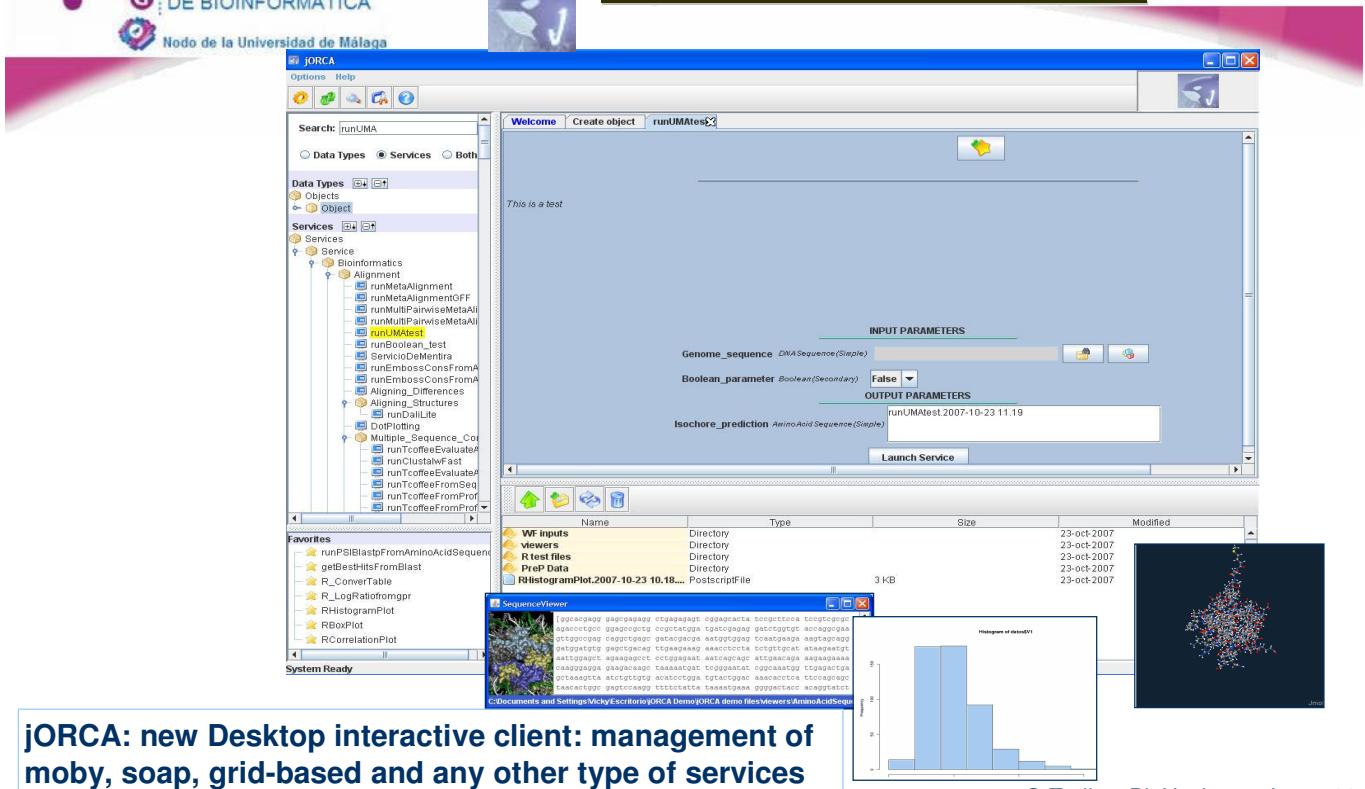
One of the important tasks from the INBG perspective was to extend the protocol to fulfil our objectives

O.Trelles, BioHackaton-Japan 08

The MOWServ-v1 Architecture



Integrated Bioinformatics, INB-UMA

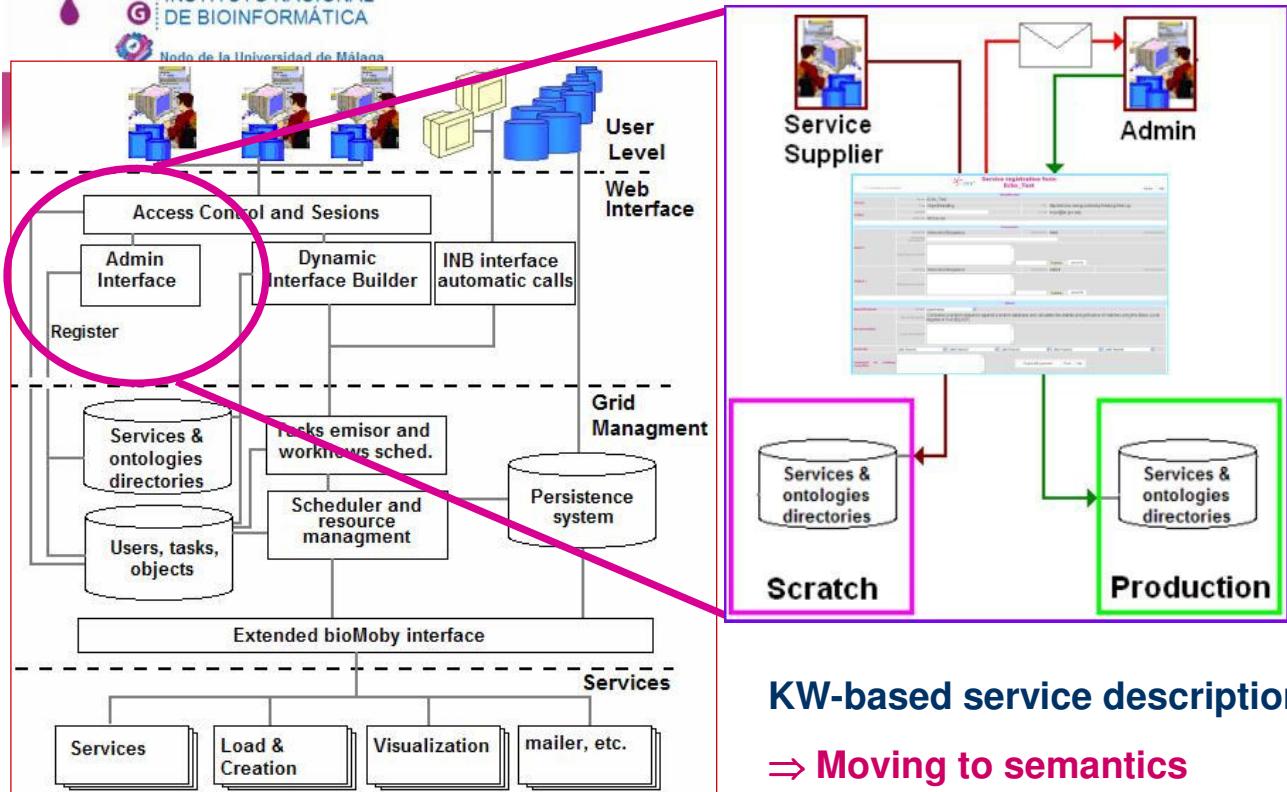


jORCA: new Desktop interactive client: management of moby, soap, grid-based and any other type of services
 Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

Controlled Registry System

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



Browsing Objects, Services & Namespaces

You are working as Anonymous User

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

Main menu:

- DataTypes
- Object
- AGI_LocusCode
- AlleleAssociation
- BasePairSequenceFeature
- Float
- GenericElement
- Gene
- GO_Term
- Integer
- Interaction
- InteractionMethod
- InteractionType
- Multi_Key_value_pair
- Sample
- SequenceWithBarcodeName
- String
- VIRTUALSEQUENCE
- AminoAcidSequence
- NucleotideSequence
- CommentedDNASequence
- DNASequenceWithGFFFeature
- RNASequence
- Services
- Namespaces

Help

Is that the INB benefits

Access to you can (data files)

If you a

DataTypes Services Namespaces

- Analysis
- Conversion
- Creation
- Parsing
- Resolution
- Retrieval
- getAnnotationsFromSwissProt
- getAminoAddSequencesFromUniprot
- getDescriptionsFromSwissProt
- getEntryFromSwissProt
- getFASTA
- getFASTAFromEMBL
- getFASTAFromGenBank
- getFASTAFromSwissProt
- getFASTAFromUniprot
- getGeneFromGenBank
- getGenericSequencesFromGenBank
- getGenericSequencesFromHGU
- getGenericSequencesFromGenBank
- getGenericSequencesFromSwissProt
- getGenericSequencesFromUniprot
- getHeaderFromPDB1
- getInteractingMethods
- getInteractionMethodDesc
- getInteractions
- getInteractorList
- getKeywordFromUniprot
- getKeywordsFromDBD
- getNBKsFromSwissProt
- getNBKsFromUniprot
- getNCBISideSequencesFromEMBL
- getNCBISideSequencesFromDB
- getNCBISideSequencesFromDBD
- getNCBISideSequencesFromPDB
- getNCBISideSequencesFromSwissProt
- getNCBISideSequencesFromUniprot
- getSWFromUniprot

the functionality of the system, and the frame) from which different name spaces

the 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

You are working as Anonymous User

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

Advanced Service Search

Advanced Service Search

ISS_Report LutefiskReport Matrix Pattern PepinfoReport Rubbish String text_plain TreeDet_Report VirtualSequence GenericSequence

Expand all Collapse all

Reset

Bioinformatics

- Alignment
- Annotations
- Biochemistry
- Clustering
- Database
- Distances
- GeneExpression
- Identification
- Proteomics
- SequenceAnalysis
- StructuralStudies
- Translating
- objectHandling

Object

- Annotation
- AnnotationHits
- Array
- BasicAnnotation
- Clustering
- DateTime
- Domain
- Element
- Float
- HierarchicalClustering
- HMMPfam_Report
- HMMSearch_Report
- Image_Encoded
- Integer
- Interaction

Matching Services

Input Object: **GenericSequence** Type of Service: **Alignment** Output Object: **HMMSearch_Report** Number of Services: 2

Workflow Help Where Submit

-runHMMSearchAgainstSeqs -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.

Creation Services

Creation services: generic
Up-down load objects

Plug-in specific
Transparent to users

Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

Main
User Objects

Object Generic Creation Service

OBJECT NAME

Id	<input type="text"/>
Namespace	<input type="text"/>
snp	<input type="text"/>
pValue	<input type="text"/>
validity	<input type="text"/>

[Back to the tree](#)

Object Types Tree:

- DataTypes
- Object
 - AlleleAssociation
 - BasicGFFSequence
 - Float
 - GFF
 - In
 - In
 - NOT validated against the GFF specification
- Interactor
- ISS_Output
- multi_key_value_Sample
- String
- VirtualSequence
- Services
- Namespaces

Service invocation

Automatic / Uniform
Objects: online creation
Up-download
Online help services

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

User ots2 Logged
[Logout](#)
your parameters

Analysis

- fromFastaUTToGFF
- runBlastFromSequence
- runBlastAminoAcidSequenceXML
- runBlastHtUte
- runBlastPfam
- runBlastMmSearch
- runISS
- runISSComplete
- runNCBIBlast
- runNCBIMlasMMI
- runNCut
- runNucleotideSequenceTo6ORFs
- runNucleotideSequence
- runPHDfromAminoAcidSequence
- runPHDfromFASTA
- runPSIBlastFromAminoAcidSequence
- runPSIBlastFromFASTA
- runPSIBlastFromSequence
- runReverseComplement
- runXNN

Main
User Objects
User Tasks
Workflows

runBlastAminoAcidSequence: Execute a blastall of proteins vs. proteins (blastp) with default parameters. If you want to put a different value than default, fill the secundary variable params (for example params = -e 0.5 -v 200 -b 100).

INPUT PARAMETERS

NAME	TYPE	VALUE
params	STRING (Secondary)	-e 10.0 -v 500 -b 250
database	STRING (Secondary)	nr
sequence	AminoAcidSequence	String2AA-2005-4-5-11:22:42

OUTPUT NAME

TYPE	NAME
------	------

Workflow Execution Monitor

ID Task: Workflow Name	Status	Input File	Output File
517: Homology search and Phylogenetic study	In Progress	in517.xml	

Workflow Execution Monitor (Task 517)

ServiceName	Input Objects	Output Objects	Status
Create_moby_data			Finished
runCreateTreeFromClustalw			Waiting
getBlastHitsFromBlast			Waiting
runNCBIBlast			In Progress
runClustalFast			Waiting
getAminoAcidSequenceCollection			Waiting
getAminoAcidSequence			Finished

[List of Workflows in Execution or Executed](#)

Tasks & Objects Monitoring

Main | User Objects | User Tasks

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

TID	Service	State
21	Generate Object	Finished
20		
19	getAminoAcidSequence	
18	getGenericSequence	

Error code: Parser XML

- 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
getGenericSequencefromGenBank
getGenericSequencefromEMBL
getFASTAfromTrEMBL
getFASTAfromGenBank
getFASTAfromEMBL
getFASTAfromSwissProt
getFASTAfromUniProt
```

Execute

Back to Service Tree

Reload

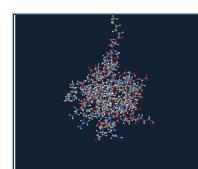
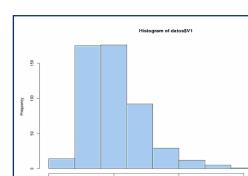
OID	TID	Object Type	Name	Date	View
58	21	Object	paraSergio	2005-04-07 08:25:37	
57	20	Object	objABCD_H	2005-04-06 19:13:38	
56	19	AminoAcidSequence	ABCDhumanSEQ	2005-04-06 19:30:13	
55	18	GenericSequence	ABCD	send object to compatible services	Waiting for task 18(Failed)
54	17	Object	ABCD_HUMAN	2005-04-06 19:28:29	
53	16	NCBI_BLAST_Text	runBlastAminoAcidSequence-2005-4-6-07:27:10		Waiting for task 16(Failed)
52	15	NCBI_BLAST_Text	string2AA-runNCBIBlastABR11		Waiting for task 15(Failed)
20	6	AminoAcidSequence	String2AA-2005-4-5-11:22:42	2005-04-05 11:23:00	

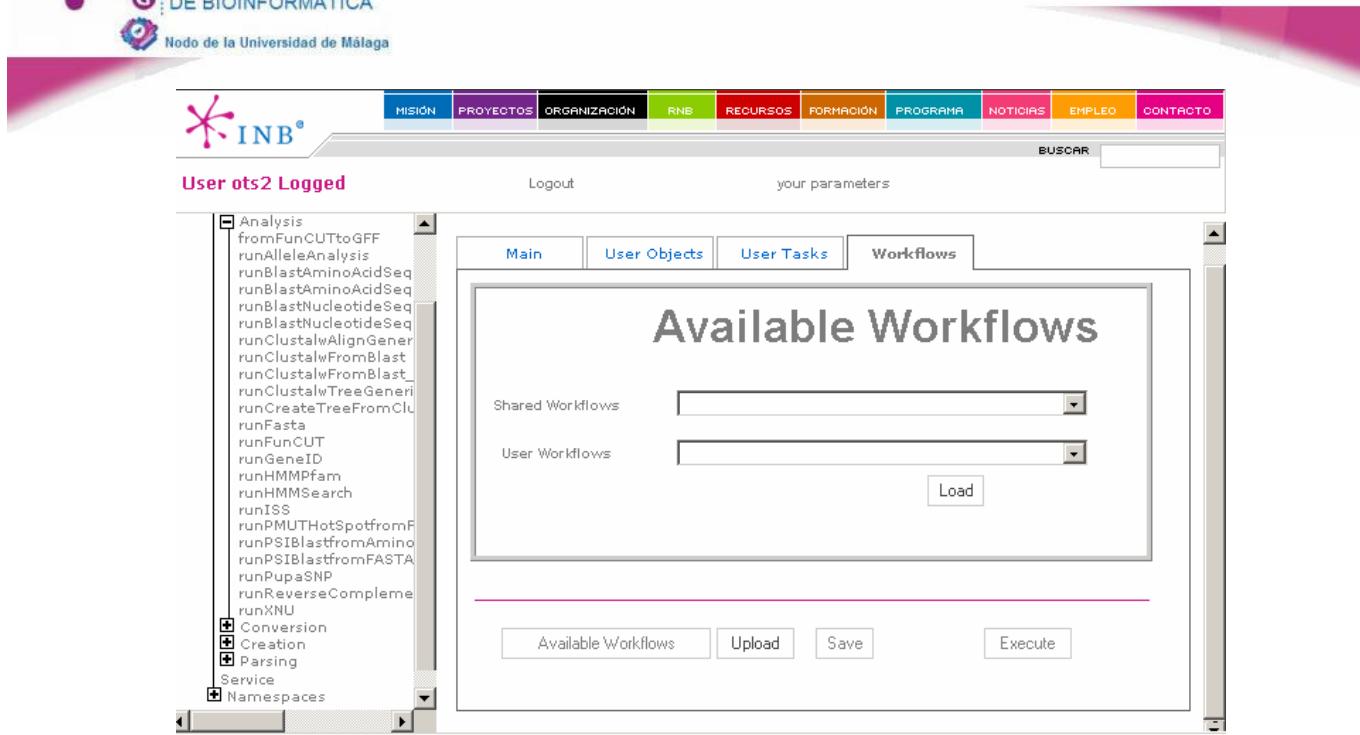
Integrated Bioinformatics, INB-UMA

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

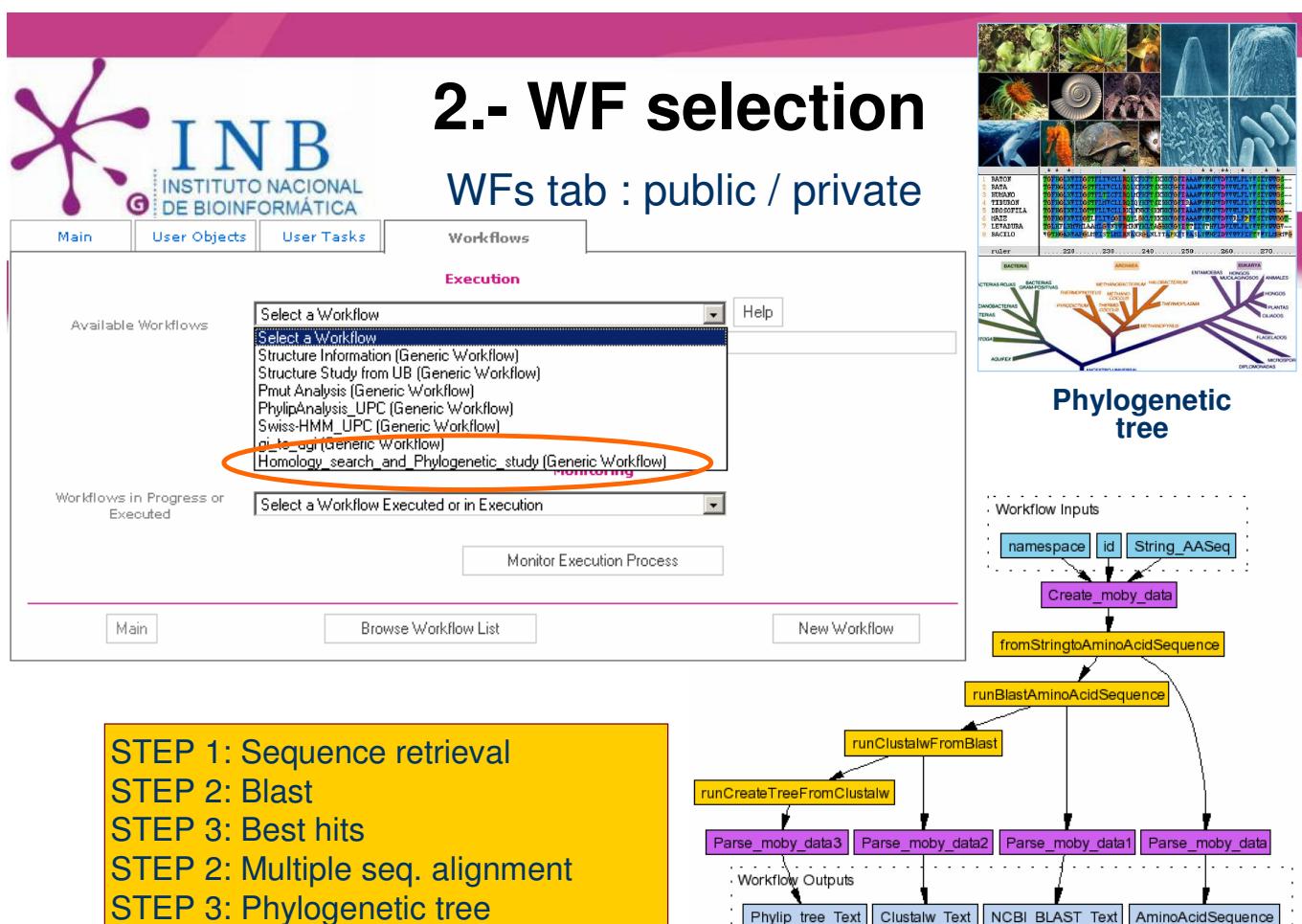




The screenshot shows the INB Workflows interface. At the top, there is a navigation bar with links: MISIÓN, PROYECTOS, ORGANIZACIÓN, RNB, RECURSOS, FORMACIÓN, PROGRAMA, NOTICIAS, EMPLEO, and CONTACTO. Below the navigation bar, it says "User ots2 Logged" and "Logout". There is also a search bar labeled "BUSCAR". The main content area is titled "Available Workflows". It has two dropdown menus: "Shared Workflows" and "User Workflows", both currently empty. Below these menus are four buttons: "Available Workflows", "Upload", "Save", and "Execute". To the left of the main content area, there is a sidebar with a tree view of available workflows, including Analysis, Conversion, Creation, Parsing, Service, and Namespaces.

Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08



The screenshot shows the INB Workflows interface with the "Execution" tab selected. On the left, there is a sidebar with "Available Workflows" and "Workflows in Progress or Executed". The "Available Workflows" dropdown is open, showing a list of workflows, with "Homology_search_and_Phylogenetic_study (Generic Workflow)" highlighted and circled in orange. Below the dropdown is a "Select a Workflow Executed or in Execution" dropdown. At the bottom of the sidebar are buttons for "Main", "Browse Workflow List", and "New Workflow". To the right of the sidebar, there is a large phylogenetic tree diagram with various organisms represented by icons and labels. Below the tree, there is a "Workflow Inputs" section with a flowchart showing the execution process. The flowchart starts with "Create_moby_data", which feeds into "fromStringtoAminoAcidSequence". This is followed by "runBlastAminoAcidSequence", "runClustalwFromBlast", "runCreateTreeFromClustalw", and "Parse_moby_data1", "Parse_moby_data2", "Parse_moby_data3". The final outputs are "Phylo_tree_Text", "Clustalw_Text", "NCBI_BLAST_Text", and "AminoAcidSequence".

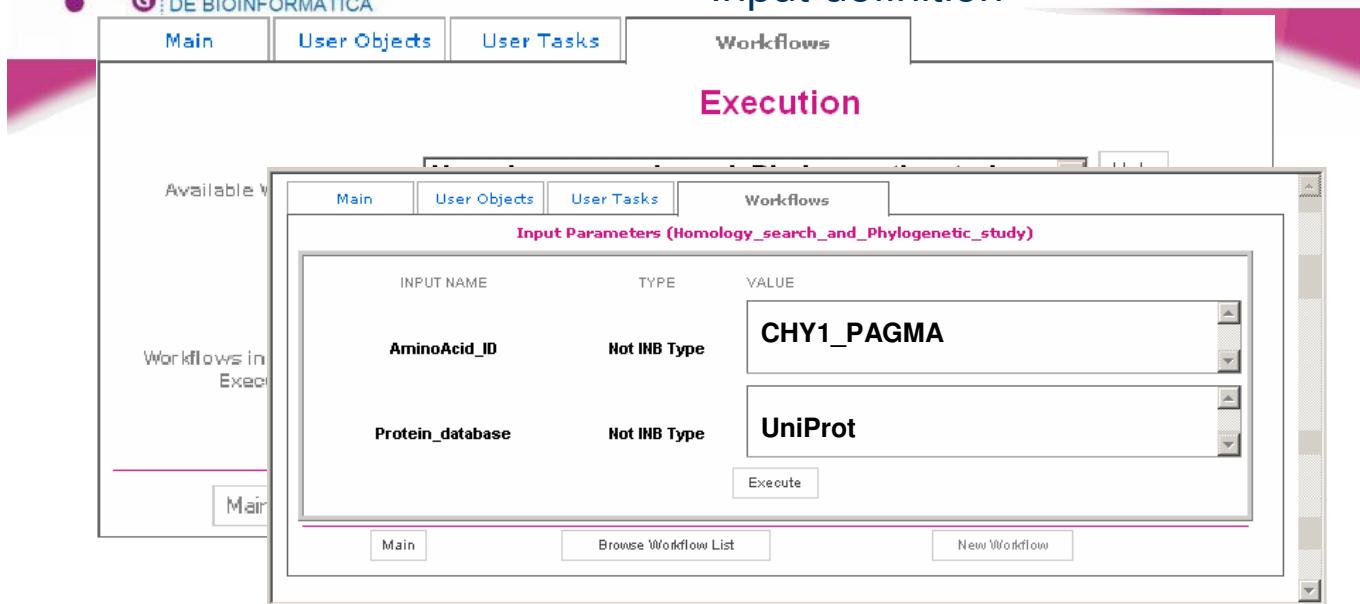
STEP 1: Sequence retrieval
STEP 2: Blast
STEP 3: Best hits
STEP 2: Multiple seq. alignment
STEP 3: Phylogenetic tree

Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

3.- specialise

Input definition



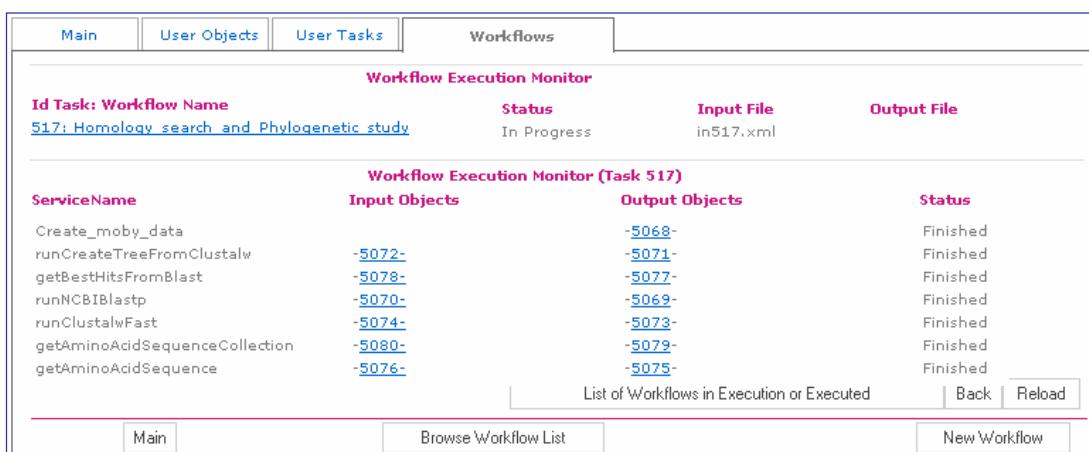
The screenshot shows the INB workflow execution interface. At the top, there are tabs: Main, User Objects, User Tasks, and Workflows. The Workflows tab is selected. Below the tabs, the word "Execution" is displayed. On the left, there are two vertical panels: "Available Workflows" and "Workflows in Execution". The main area is titled "Input Parameters (Homology_search_and_Phyllogenetic_study)". It contains two input fields: "AminoAcid_ID" with value "CHY1_PAGMA" and "Protein_database" with value "UniProt". A "Execute" button is at the bottom. At the very bottom of the interface, there are buttons for Main, Browse Workflow List, and New Workflow.

Launch execution

Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

4.- Monitoring

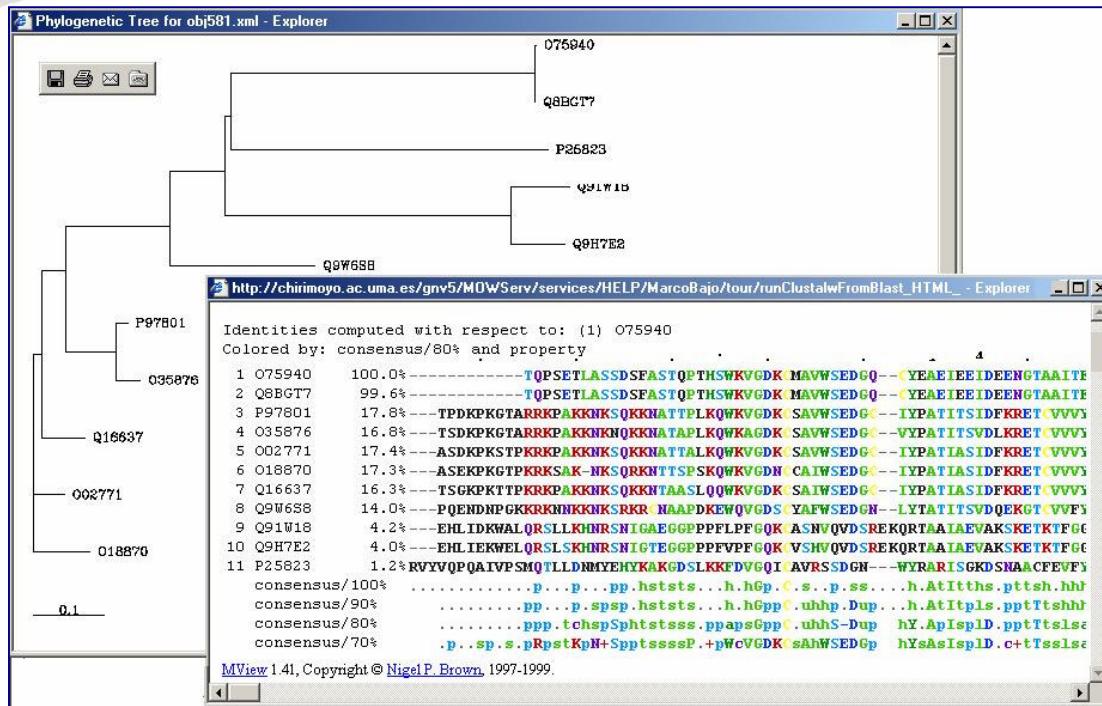


The screenshot shows the INB workflow monitoring interface. At the top, there are tabs: Main, User Objects, User Tasks, and Workflows. The Workflows tab is selected. Below the tabs, the title "Workflow Execution Monitor" is followed by a table for task 517. The table has columns: Id Task, Workflow Name, Status, Input File, and Output File. The entry is: "517: Homology_search_and_Phyllogenetic_study" with "In Progress" status, "in517.xml" as the input file, and no output file listed. Below this table is another table titled "Workflow Execution Monitor (Task 517)" with columns: ServiceName, Input Objects, Output Objects, and Status. It lists several service calls and their corresponding object IDs and statuses. At the bottom, there is a link "List of Workflows in Execution or Executed" and buttons for Back and Reload. At the very bottom, there are buttons for Main, Browse Workflow List, and New Workflow.

In progress

...done

5.- Results



Integrated Bioinformatics, INB-UMA

O.Trelles, BioHackaton-Japan 08

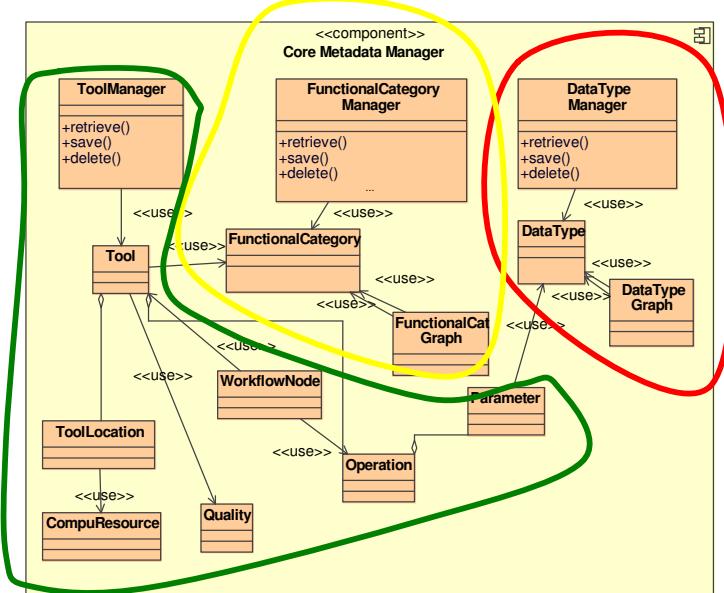
Open questions Short term

- **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



Bioinformatics and Information Technologies Laboratory

<http://chirimoyo.ac.uma.es/bitlab>



Arquitectura de Computadores



Lenguajes y Ciencias de la Computación



Genética



Biología Molecular y Bioquímica