

Bioinformatics Ontology: Towards the Automatics Generation of Bioinformatics Workflow for Web Services

Konagaya Akihiko

Project Director
Advanced Genome Information Technology
Research Group
RIKEN Genomic Sciences Center

Contents

- Introduction of Ontology
- Web Services for Bioinformatics
- Automatics Workflow Generation
- Lessons from our First Experience

Introduction of Ontology

Tacit and Explicit Knowledge

We should start from the fact that
'we can know more than we can tell'.

Michael Polanyi, "The Tacit Dimension" 1967



Michael Polanyi (1891-1976)

Rainbow Color

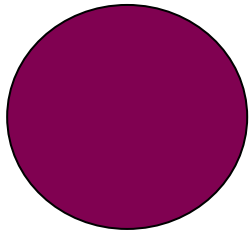
How many colors can you see in rainbow?



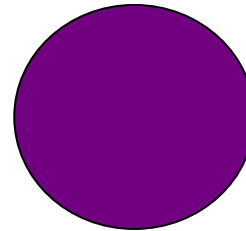
Ontology for Rainbow Colors



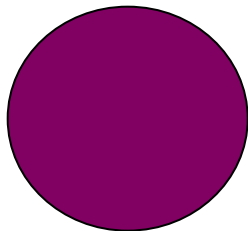
Which are Purple?



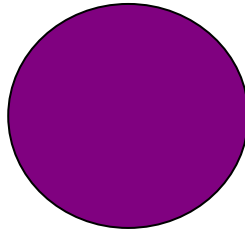
#800050



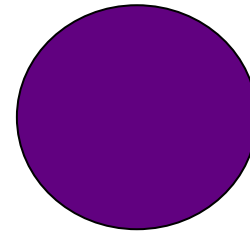
#700080



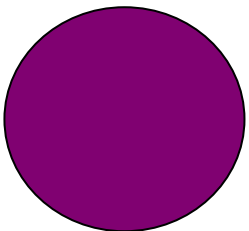
#800060



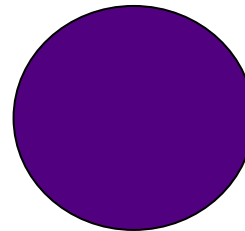
#800080



#600080

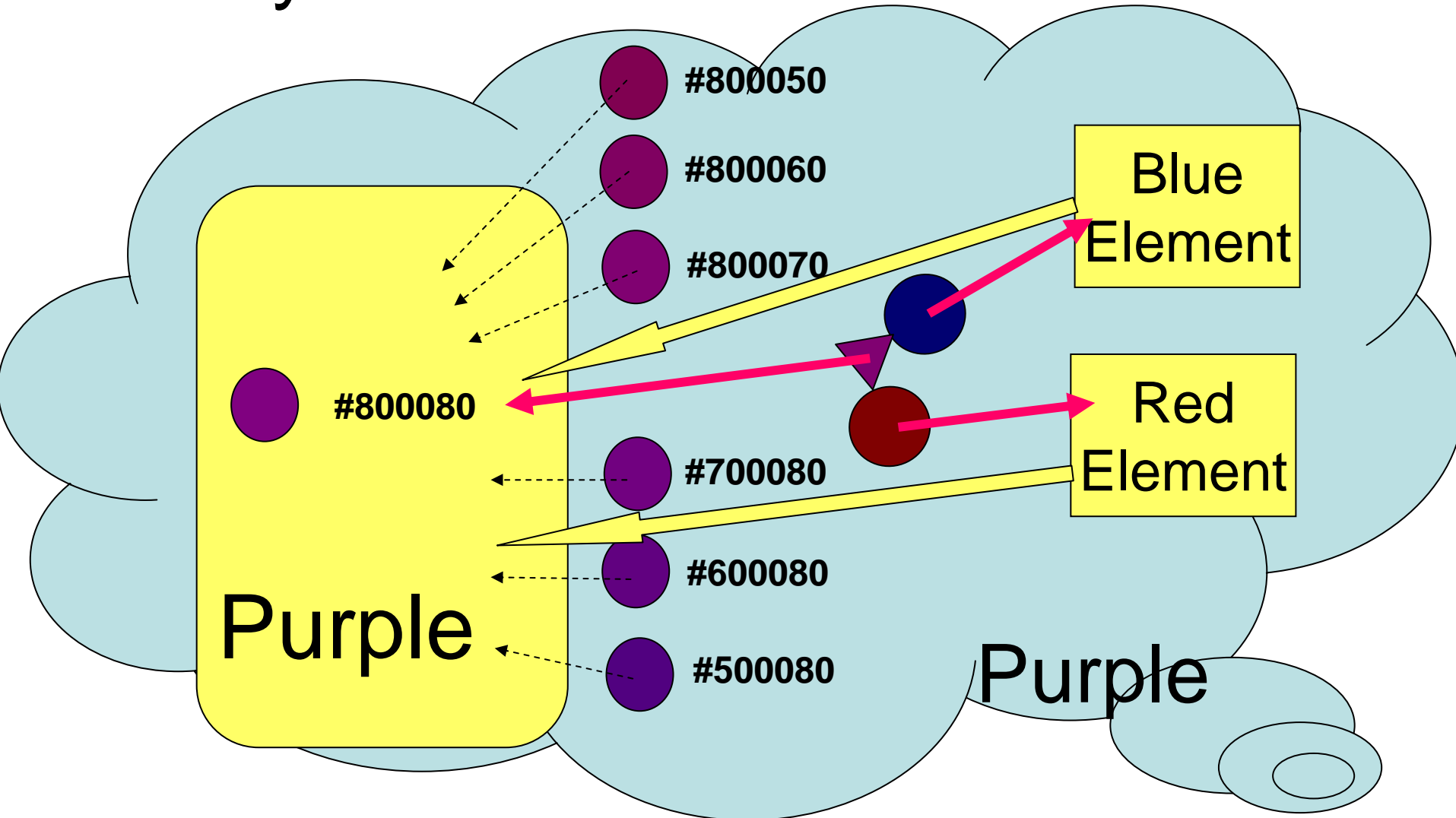


#800070



#500080

Representation by Elements and Constructor

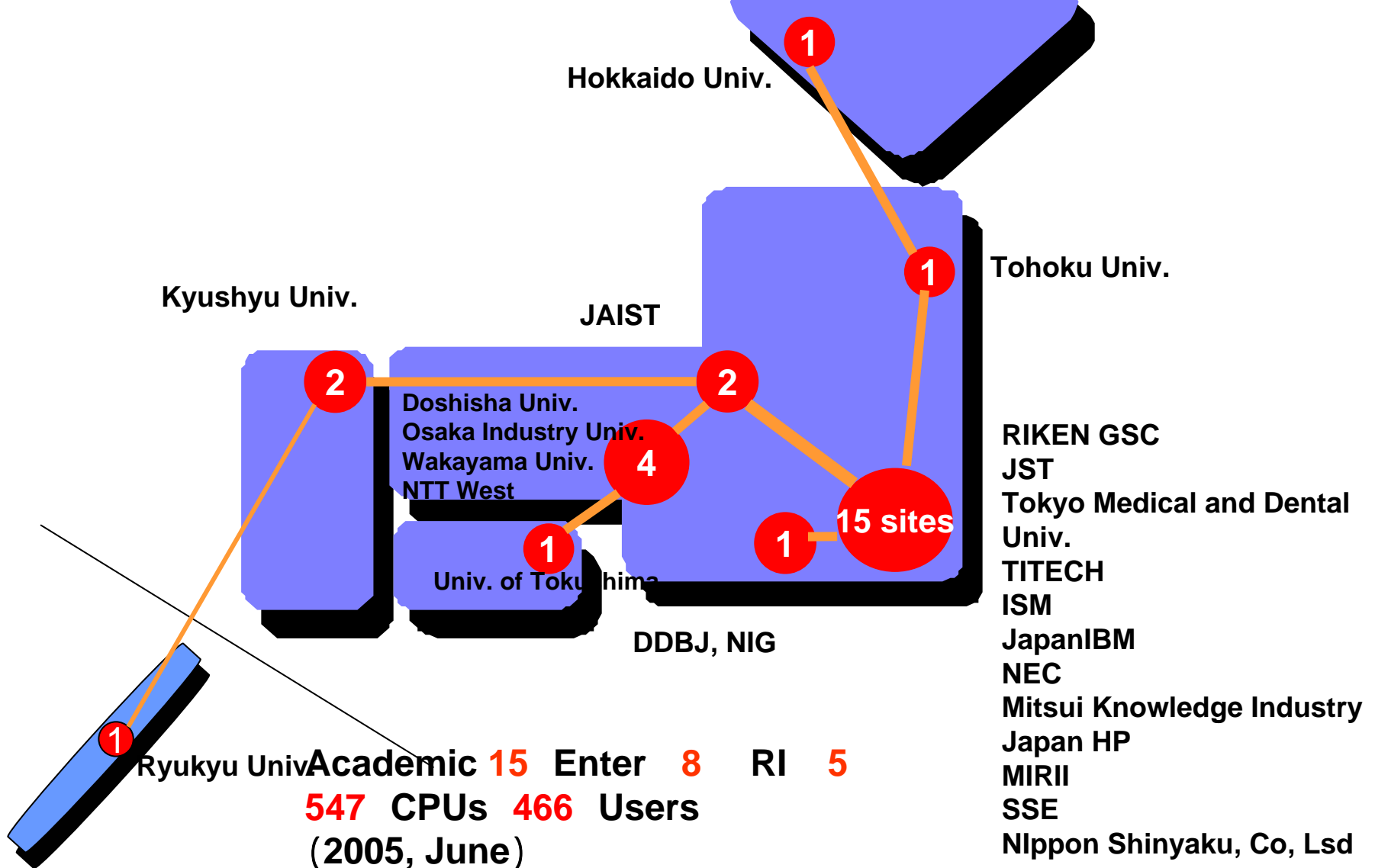


Web Services for Bioinformatics

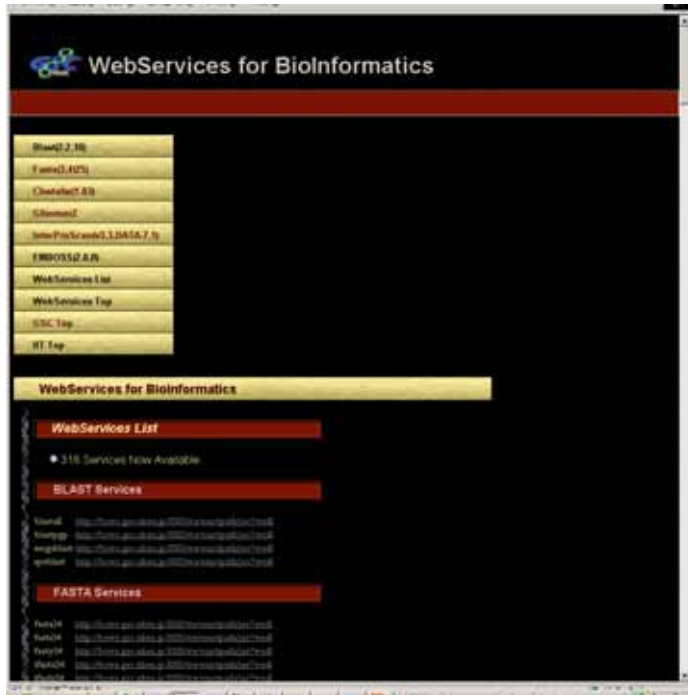
Formulation of Community



OBIGrid VPN



Bioinformatics Web Services on Grid



GRIDIFIED

BLAST, FASTA, ClustalW,
Glimmer2, InterProScan,

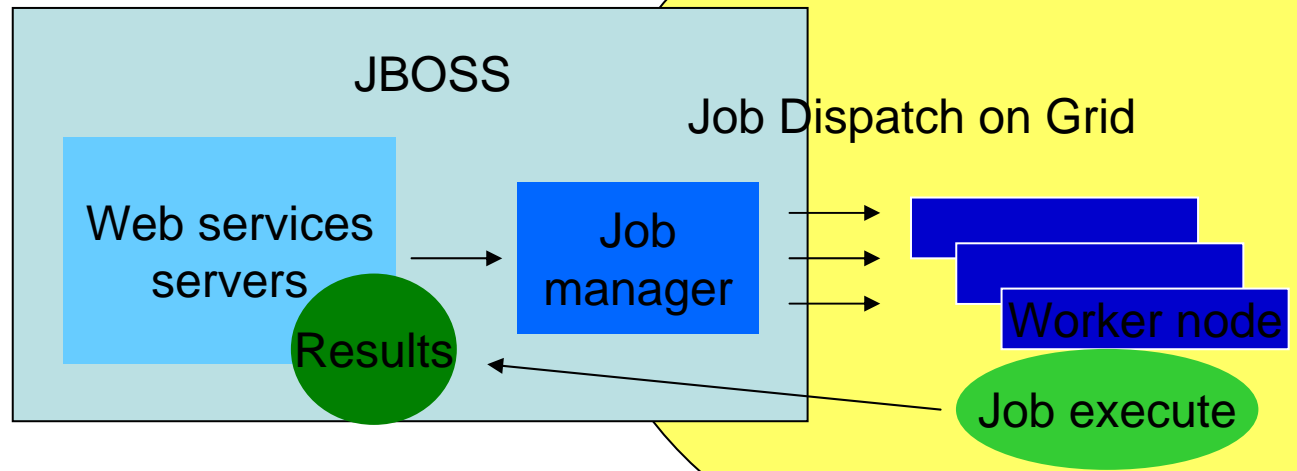
<http://jkt.gsc.riken.jp/sp/spbio/wslist.jsf>

Client



call web
services

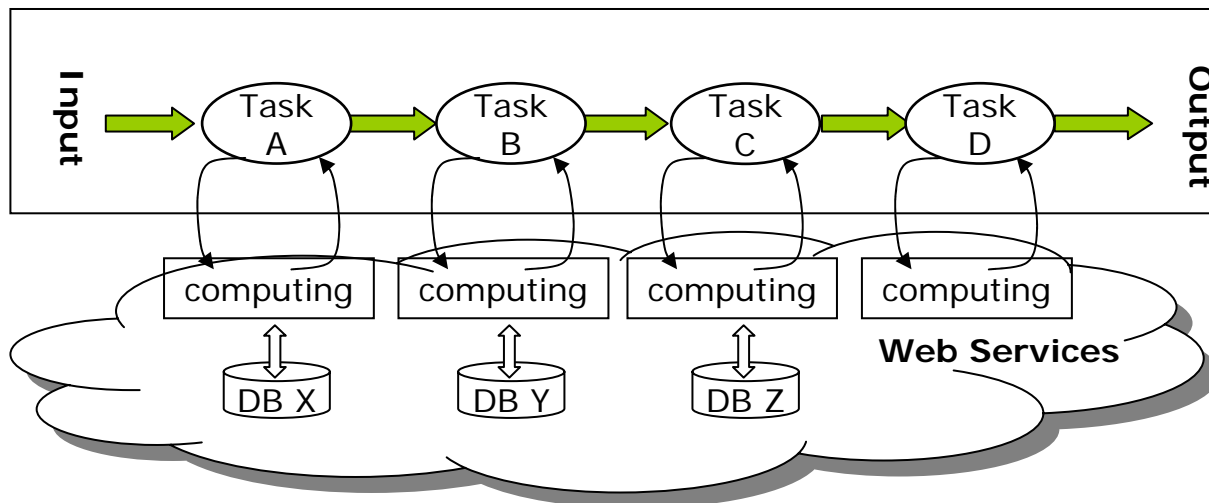
Return



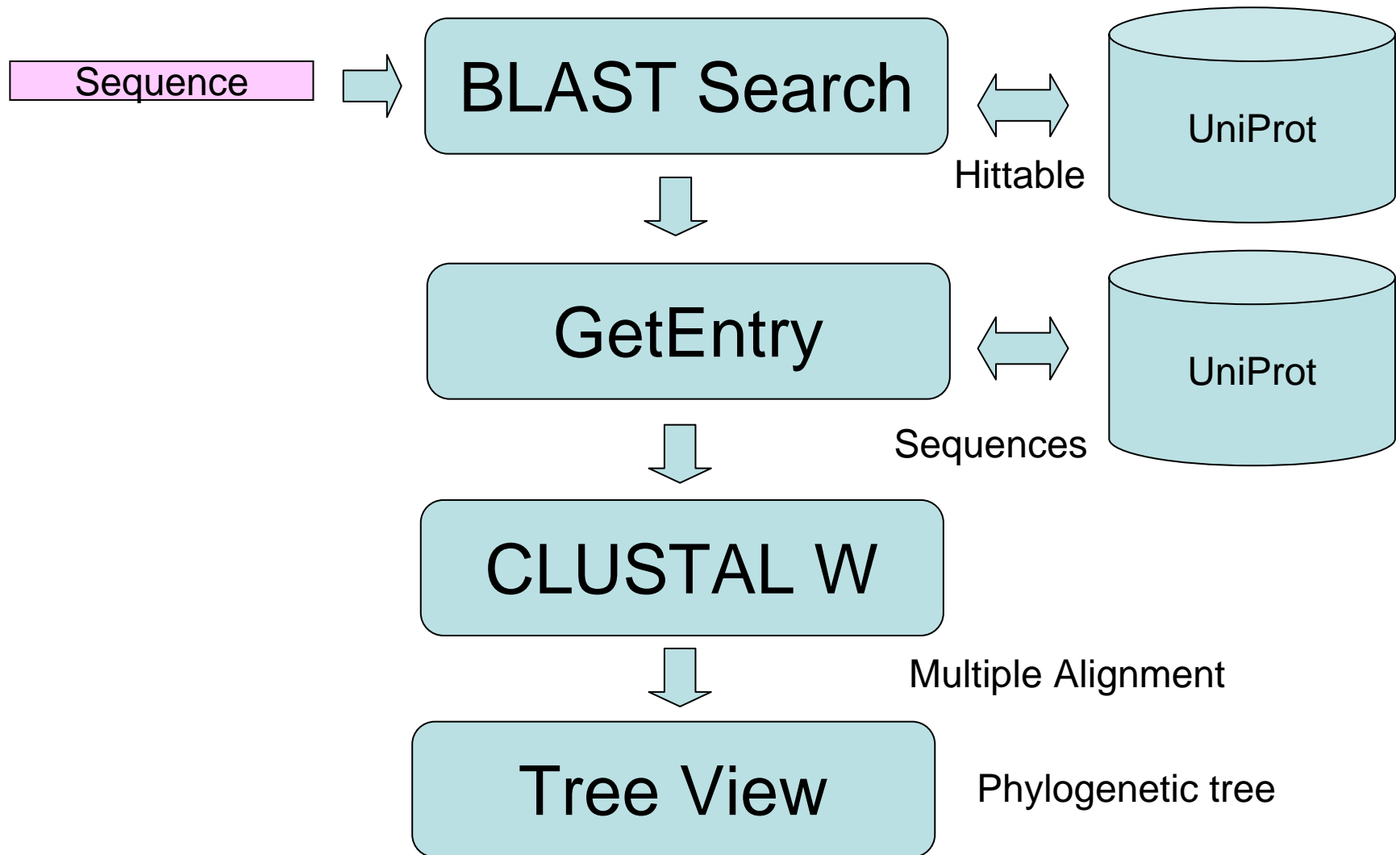
Advantages of Web Services

- Liberating from the maintenance of biological databases and tools
- Scalability of computational resources
- High-level application programming interface

Web Services



Very Simple Work Flow



Manual Workflow on Web Apps

The image displays a manual workflow on web applications for sequence analysis, shown within a Microsoft Internet Explorer browser window. The workflow involves several steps:

- BLAST Search:** The first screenshot shows the BLAST search interface. The "PROGRAM" dropdown is set to "blastn" (DNA query vs. DNA database). The "DATABASE" dropdown is set to "DDBJ ALL (DDBJ, EMBL, GenBank)". The "ANALYSIS" section shows "CLUSTALW" selected.
- getentry:** The second screenshot shows the "getentry" web application, which is used to retrieve sequence data from the DDBJ database. It includes fields for "Accession" and "E-mail Address".
- ClustalW:** The third screenshot shows the ClustalW alignment interface. The "ALIGN" section is active, and the "DOTSINOUT" option is selected. The "TYPE" dropdown is set to "AUTO", and the "OUTPUT" dropdown is set to "clustal".
- PHYLIP:** The fourth screenshot shows a PHYLIP phylogenetic tree. The tree is rooted and shows the relationships between several species: *Rattus norvegicus*, *Mus musculus*, *Homo sapiens*, *Takifugu rubripes*, *Xenopus laevis*, *Gallus gallus*, *Anopheles gambiae*, and *Danio rerio*.

The browser window also shows a "DDBJ services emergency suspension" notice, indicating a system issue on May 10, 2006.

Web Service Programming

```
#!/usr/bin/perl

use SOAP::Lite;

# SOAP API
# specify WSDL
my $service = SOAP::Lite-> service('http://xml.nig.ac.jp/wsdl/GetEntry.wsdl');

# call web service
$result = $service->getXML_DDBJEntry("AB000003");

# print result
print $result;
```

<http://www.xml.nig.ac.jp/perl.txt>

Why don't we use workflow tools?

The screenshot displays the Taverna Workbench interface, which is used for managing and executing workflows. The main window is titled "Enactor invocation" and shows a table of processor status.

Type	Name	Last event	Event timestamp	Event detail
	Blast2_program	ProcessComplete	28-Jul-2004 11:37...	
	comparer	ProcessComplete	28-Jul-2004 11:39...	
	Fasta_to_numbered	ProcessComplete	28-Jul-2004 11:39...	
	simplifier	ProcessComplete	28-Jul-2004 11:39...	
	ncbiblast	ProcessComplete	28-Jul-2004 11:39...	
	repeatmasker	ProcessComplete	28-Jul-2004 11:38...	
	retrieve	ProcessComplete	28-Jul-2004 11:39...	
	copyright	ProcessComplete	28-Jul-2004 11:37...	
	blast2	ProcessComplete	28-Jul-2004 11:39...	
	lister	ProcessComplete	28-Jul-2004 11:39...	

Below the table, there are sections for "Intermediate inputs" and "Intermediate outputs".

On the right side, there is an "Advanced model explorer" window showing a hierarchical view of workflow objects, including "Workflow inputs", "Workflow outputs", "Processors", and "Run Workflow".

At the bottom, there is a "Run Workflow" window with a list of input documents and a "Run Workflow" button.

The background of the main window shows a workflow diagram with various processors and data flows.

Needs Automatic Workflow Generate Tool from Very High Level Specification

apply **Blastp** to **UniProt**

GetEntry from **UniProt**

apply **CLUSTALW**

apply **TreeView**

Automatics
Generation

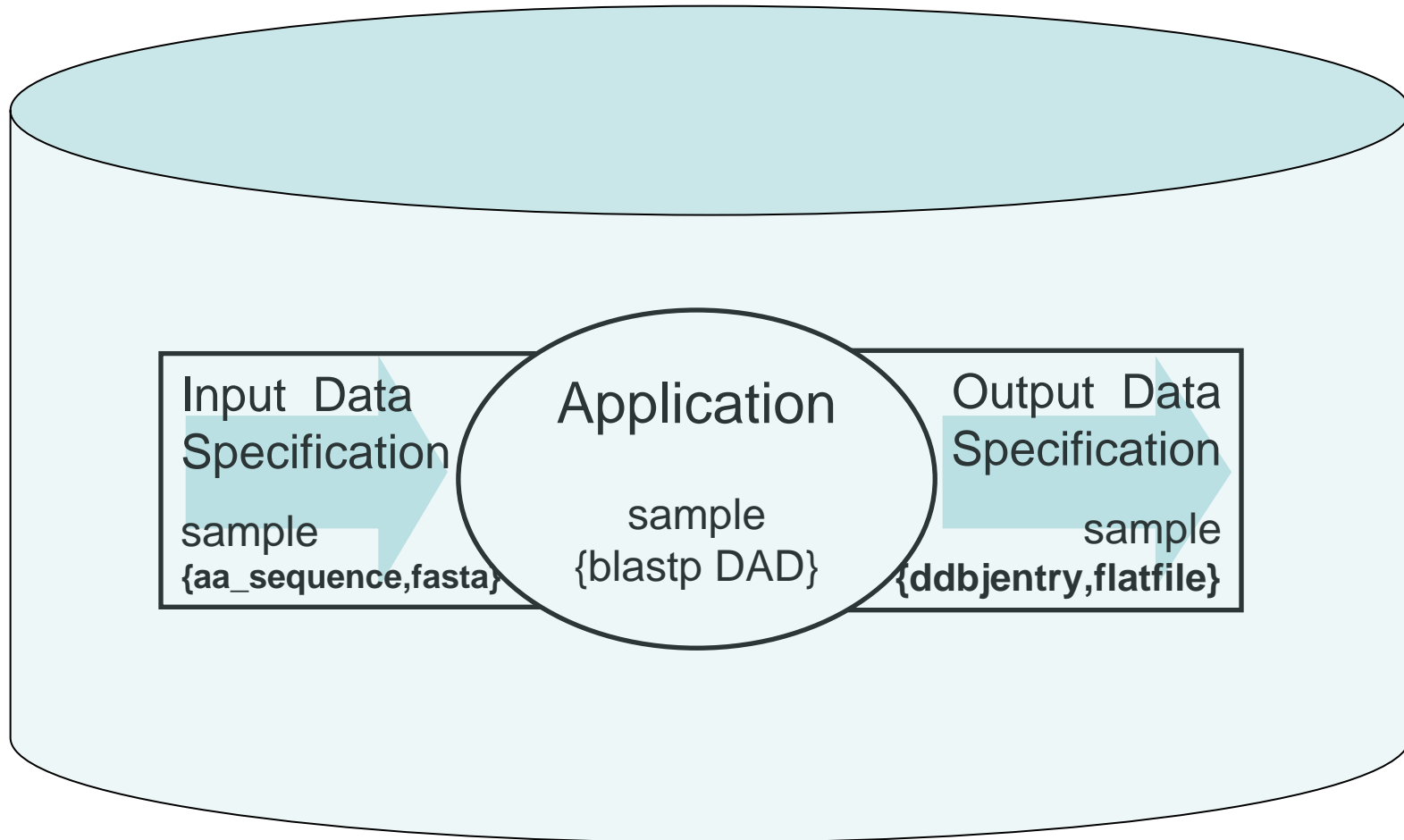


?

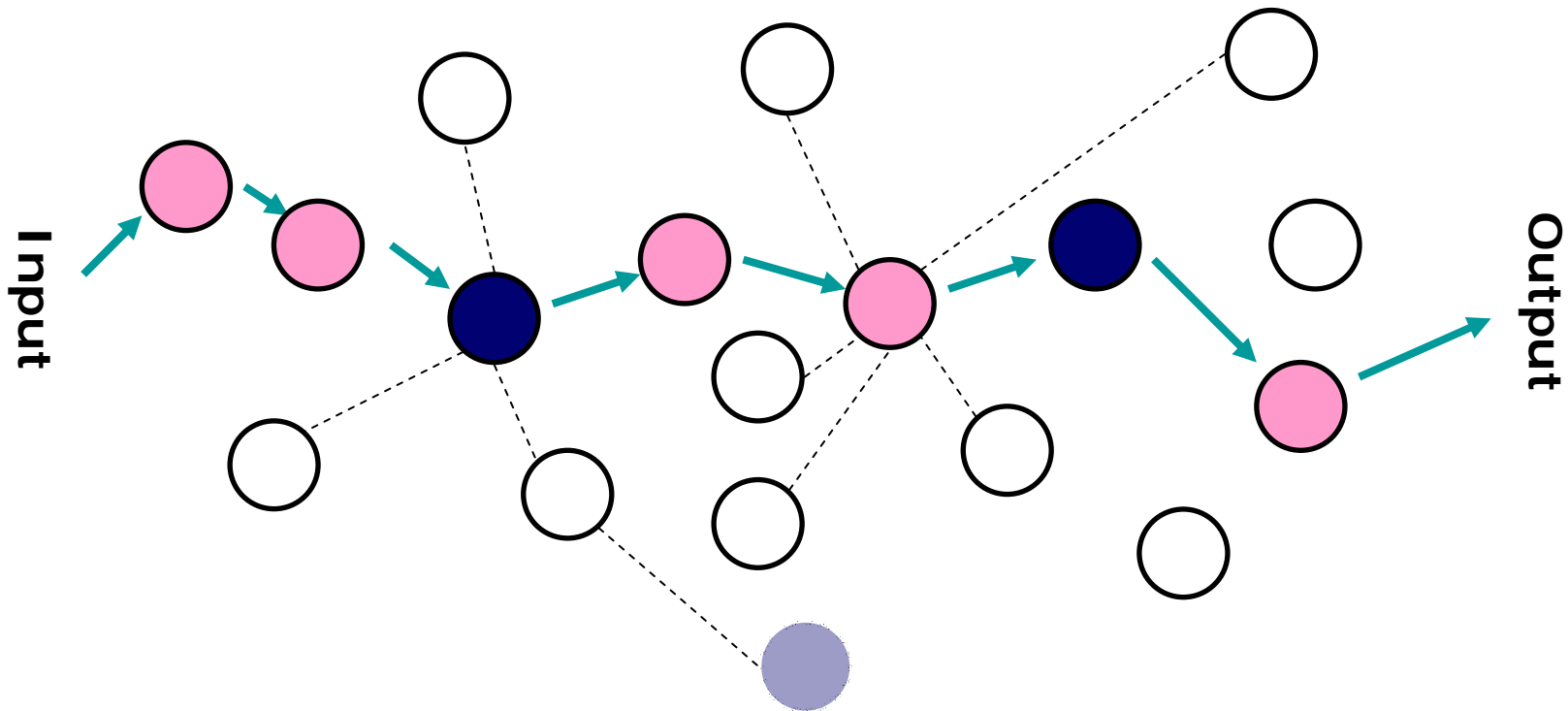
**Workflow
for
Bioinformatics Web Services**

Automatic Generation of Bioinformatics Workflow

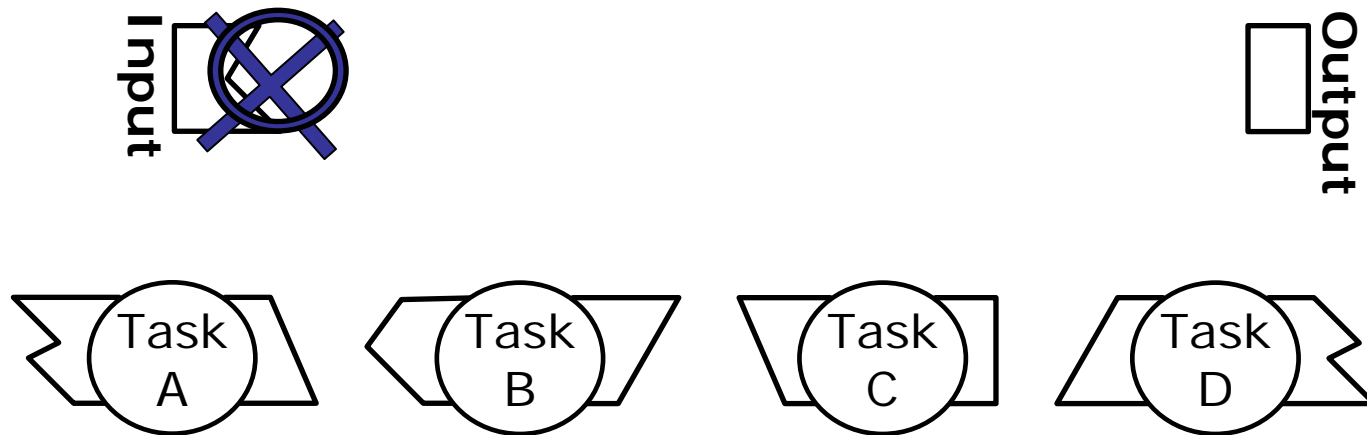
Task as Atomic Component of Workflow



Workflow as a Sequence of Tasks



Automatic Generation of Workflow from Given Input and Output Data Specification and Tasks



- Path Finding using Meta Information

Meta Information to Specify the Functionality of Task

TASK

Meta Data
for Database

samples
{uniprot}
{nt}

Meta Information
for Command and
Options

{blastn}
{getentry}

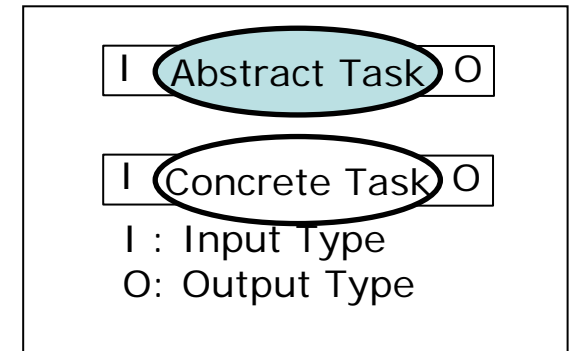
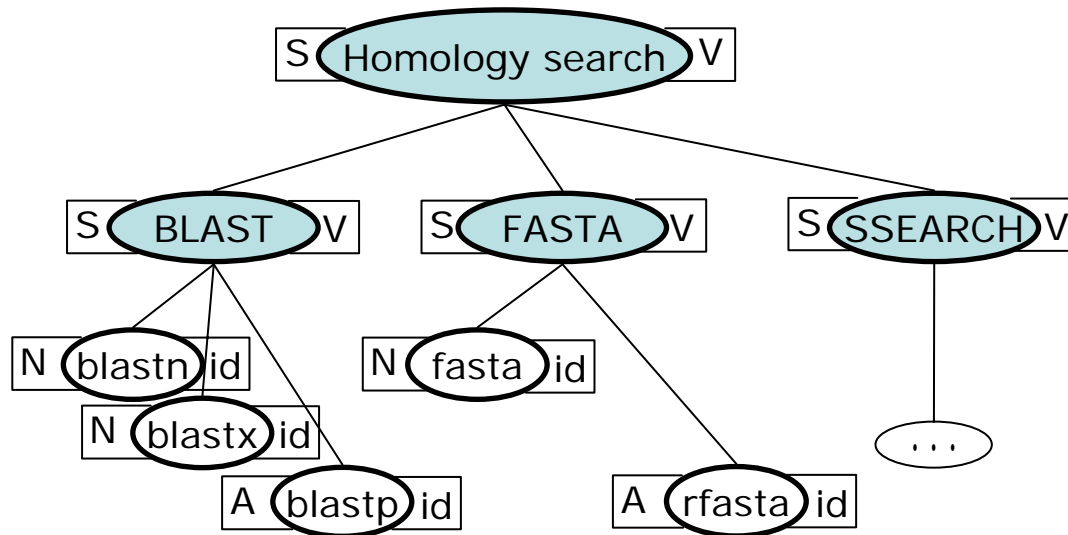
Meta Data
for Input

samples
{na_sequence,fasta}
{aa_sequence,fast}

Meta Data
for Output

sample
{ddbjentry,flatfile}
{aablantentry,hittable}

Task Hierarchy (is_a)



S : Sequence or
Sequence Name

V : Various Type

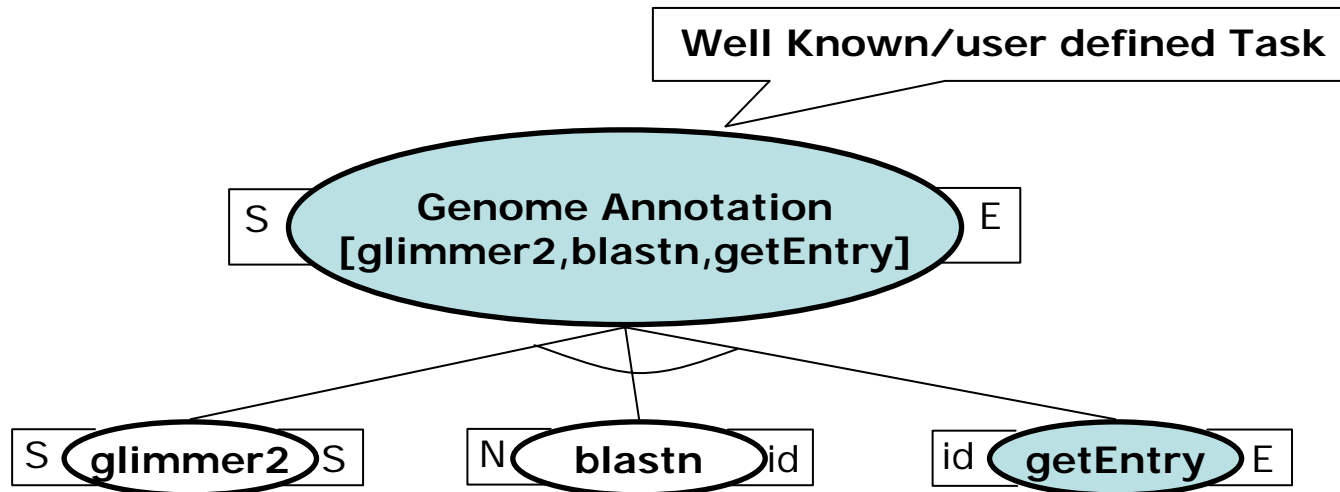
N : Nucleoside
Sequence

A : Amino acid
Sequence

id : Accession ID

E : Database Entry

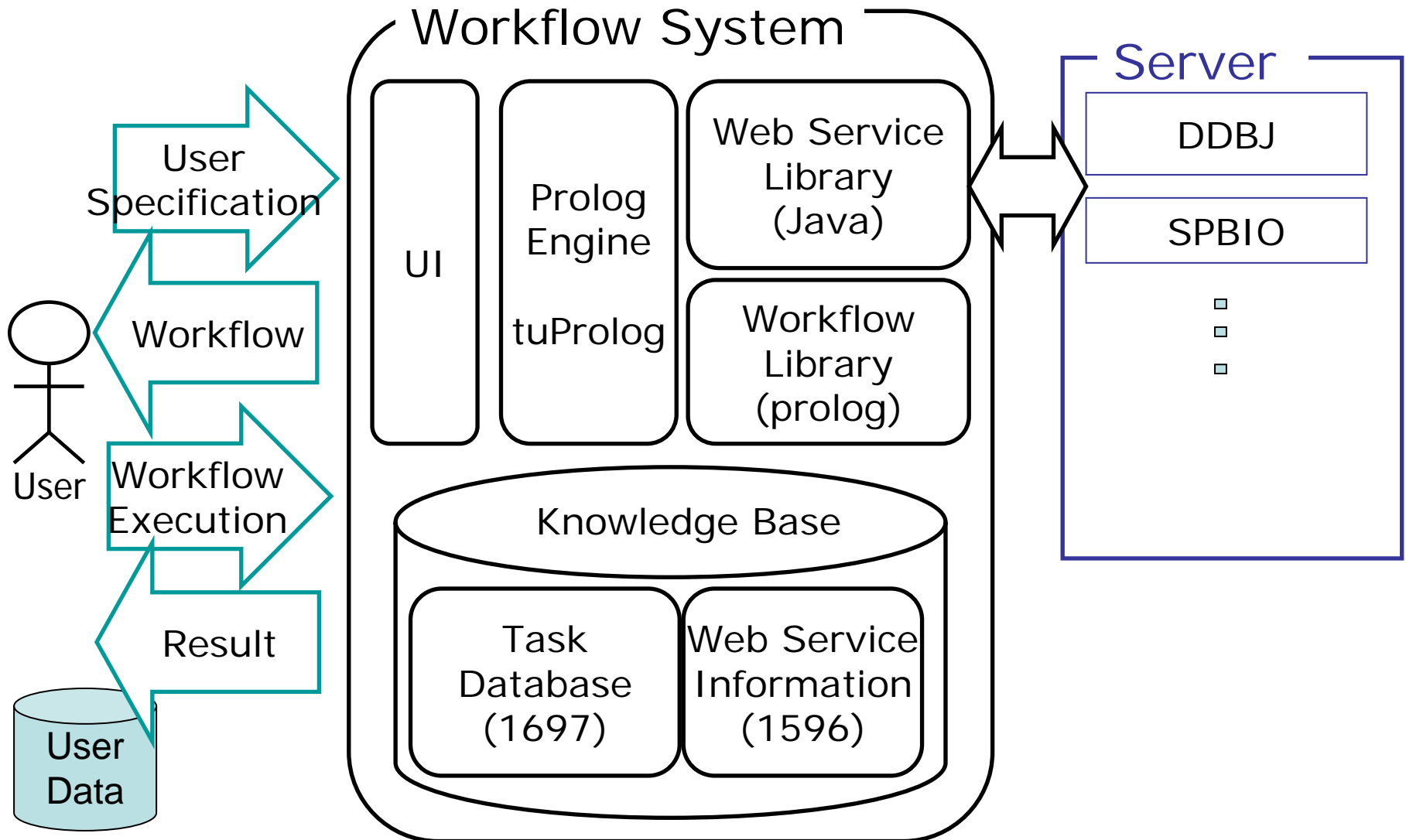
Task Hierarchy (has_a)



Prototype for 'Proof of Concept'

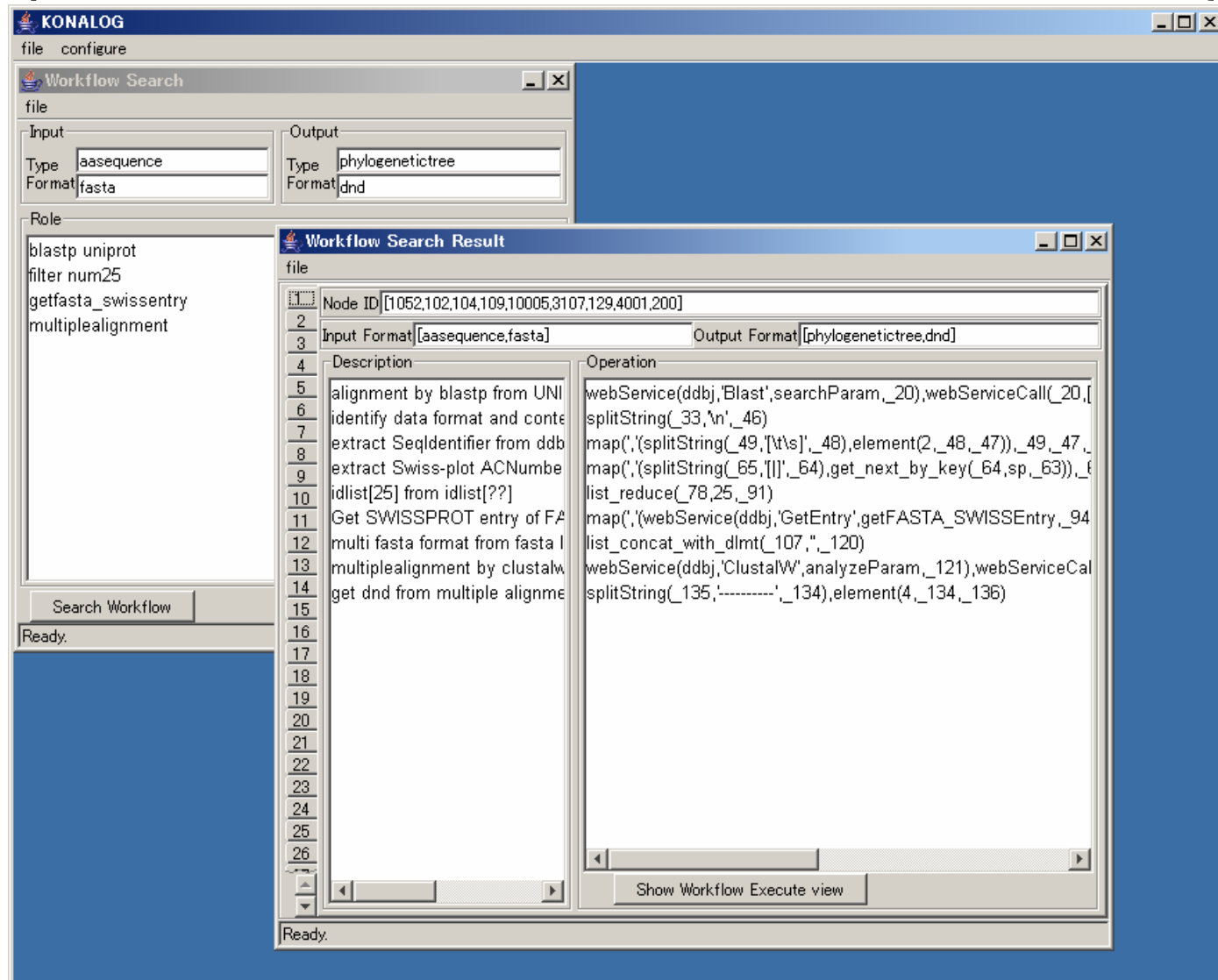
- Language tuProlog
 - Java to Prolog
 - Prolog to Java
 - Web Service Interface through JAVA API
- Task Database
 - Prolog Clause Database
- Optimal Path Finding
 - Bidirectional Breadth First Search Algorithm

System Overview

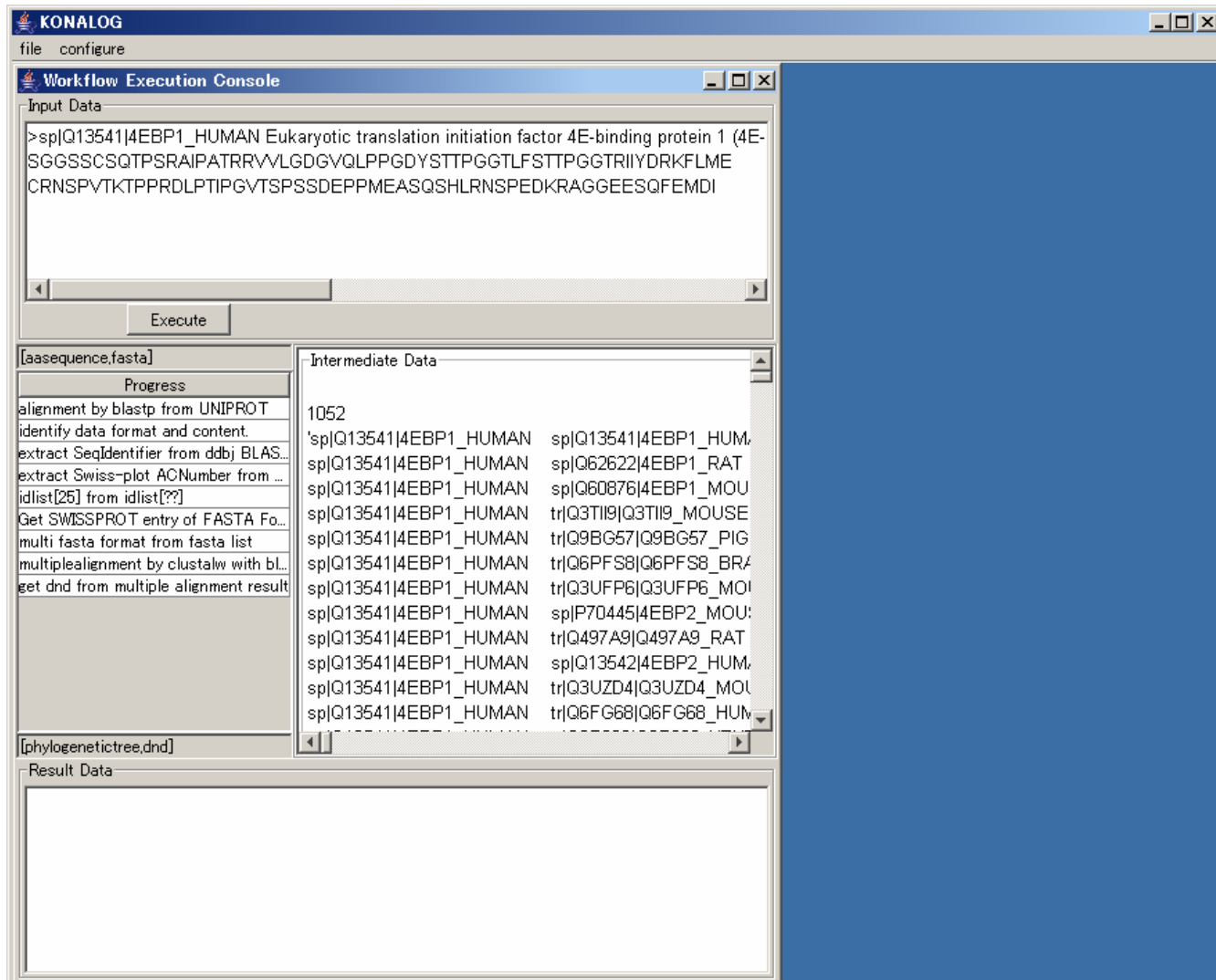


Screen Snapshot

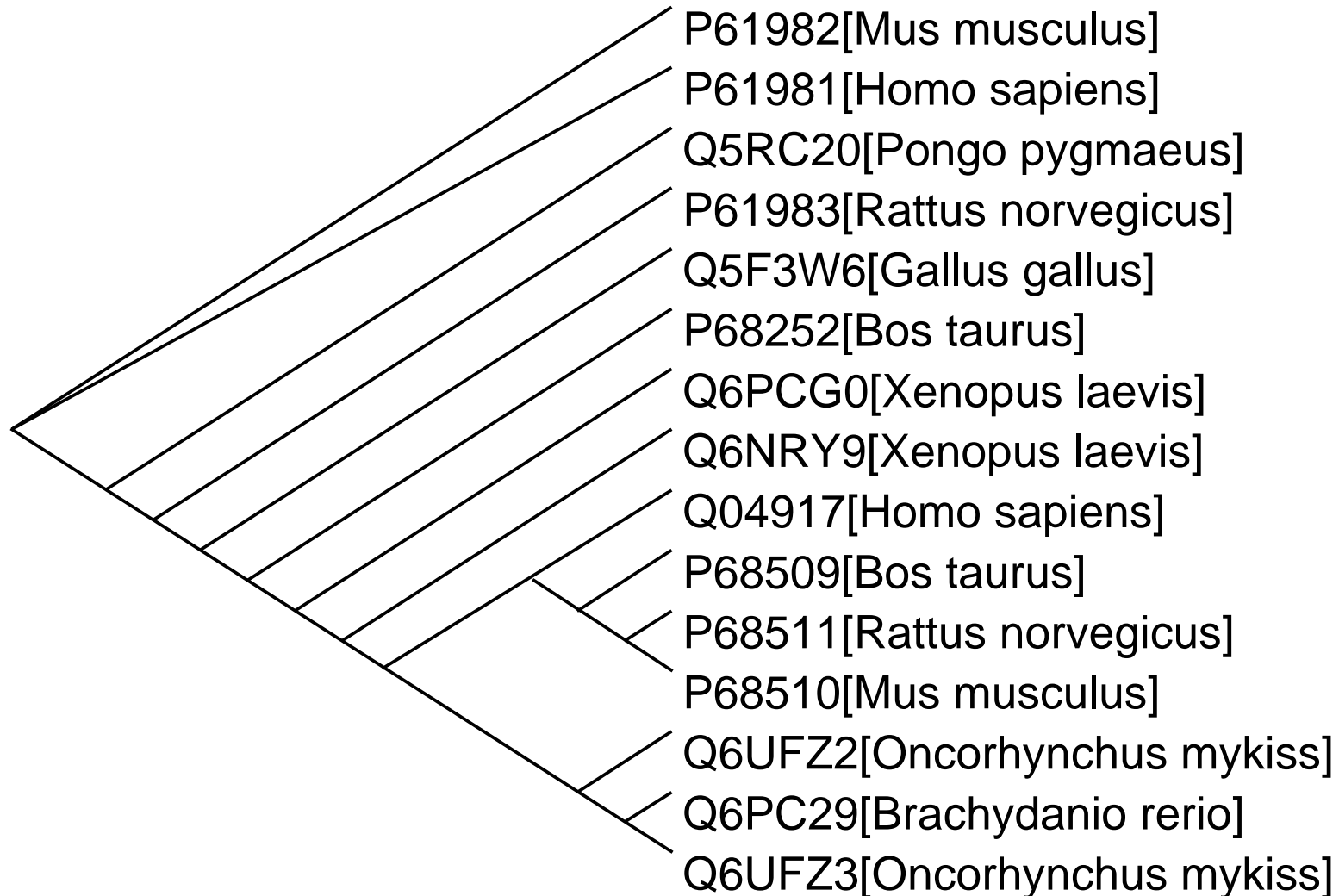
(Workflow Generation Phase)



Screen Snapshot (Workflow Execution Phase)



Obtained Phylogenetic Tree by a generated workflow when applying to a Human Insulin Sequence



Lessons from our First Experience

Task Database (prototype)

Web Service Call

DDBJ Blast	453
DDBJ SRS	638
DDBJ GetEntry	38
DDBJ ClustalW	62
SPBIO Blast	405

Format Transformation	56
-----------------------	----

Data Selection	45
----------------	----

In Total	1697
----------	------

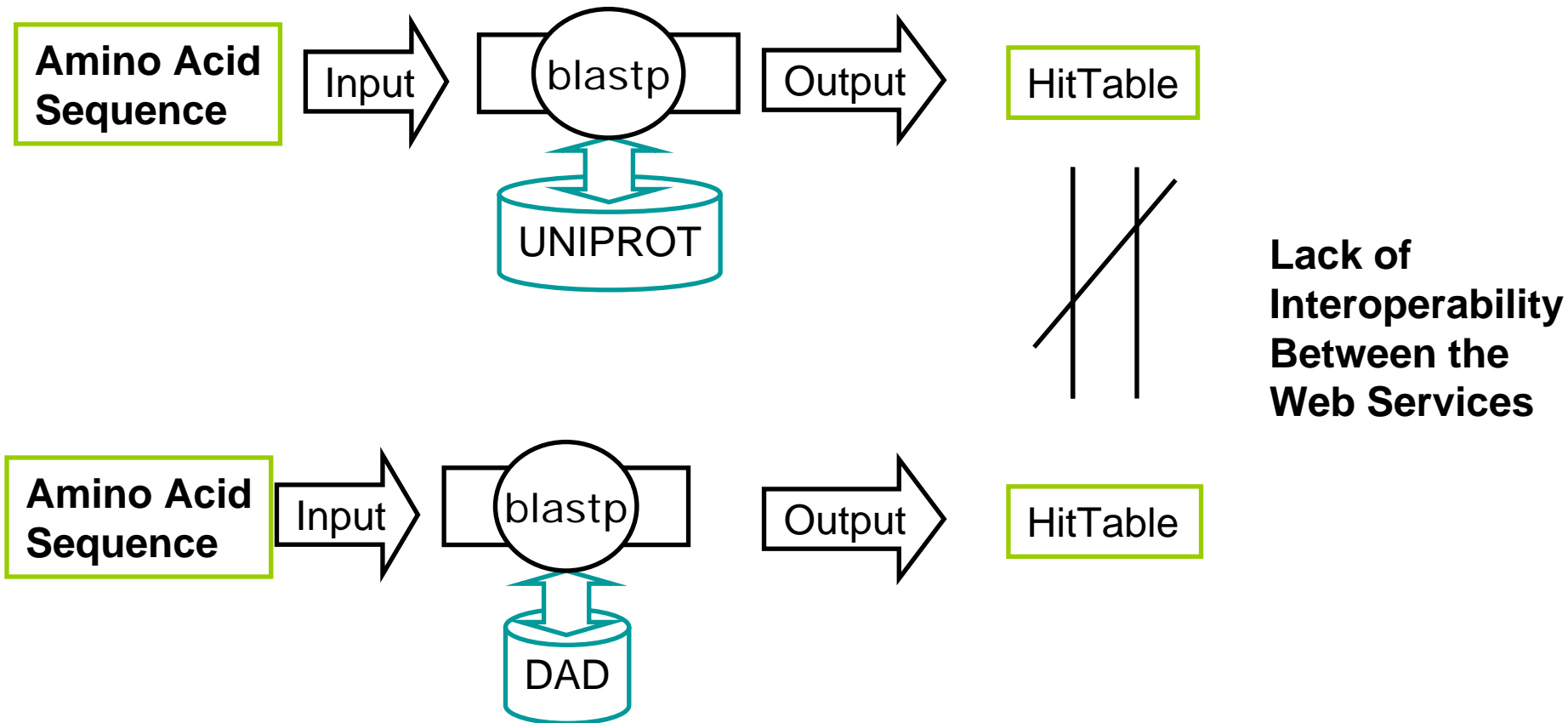
Test Set of Specification

No	Workflow				
	Input		Output		Applications
	Format	Type	Format	Type	
1	fasta	aasequence	gde	aamultiplealignment	blastp uniprot
					filter num25
					getfasta_swissentry
					multiplealignment
2					blastp uniprot
					filter num25
					getfasta_swissentry
					multiplealignment
3	fasta	aasequence	gde	aamultiplealignment	alignmentsearch
					filter
					getentry
					multiplealignment
4	fasta	aasequence			filter
					getentry
					multiplealignment

Differences of Generated Workflow

Meta Data	No.	Solution Num	time(ms)	First Match WebServiceCall		
				ID	Description	
Input Database Output Full Cmds	1	8	11266	1052	alignment by blastp from UNIPROT	○
				102	identify data format and content.	
				104	extract SeqIdentifier from ddbj BLAST result record	
				109	extract Swiss-plot ACNumber from SequenceIdentifier	
				10005	idlist[25] from idlist[??]	
				3107	Get SWISSPROT entry of FASTA Format by Accession Number.	
				129	multi fasta format from fasta list	
				4018	multiplealignment by clustalw with blosum	
No input Database No output Full Cmds	2	41	46704	1052	alignment by blastp from UNIPROT	○
				102	identify data format and content.	
				104	extract SeqIdentifier from ddbj BLAST result record	
				109	extract Swiss-plot ACNumber from SequenceIdentifier	
				1005	idlist[25] from idlist[??]	
				3107	Get SWISSPROT entry of FASTA Format by Accession Number.	
				129	multi fasta format from fasta list	
				4001	multiplealignment by clustalw with blosum	
Input No DB Output Partial Cmds	3	100 over	249906	1043	alignment by blastp from DAD	X?
				102	identify data format and content.	
				104	extract SeqIdentifier from ddbj BLAST result record	
				109	extract Swiss-plot ACNumber from SequenceIdentifier	
				10001	idlist[25] from idlist[??]	
				3107	Get SWISSPROT entry of FASTA Format by Accession Number.	
				129	multi fasta format from fasta list	
				4018	multiplealignment by clustalw with blosum	
input No DB No output Partial Cmds	4	100 over	25297	1043	alignment by blastp from DAD	X?
				102	identify data format and content.	
				104	extract SeqIdentifier from ddbj BLAST result record	
				109	extract Swiss-plot ACNumber from SequenceIdentifier	
				10001	idlist[5] from idlist[??]	
				3107	Get SWISSPROT entry of FASTA Format by Accession Number.	
				129	multi fasta format from fasta list	
				4001	multiplealignment by clustalw with blosum	

Why Failed?



Very Similar but not the Same Format

RESULT OF BLAST - Microsoft Internet Explorer

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

Blastp for DAD 733_1165]です。

[CLUSTALW SETUP ([Graphical View](#))]

BLASTP 2.2.12 [Aug-07-2005]

Reference: Altschul, Stephen F., et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= gi|307072|gb|AAA59179.1| (107 letters)

Database: DAD: DAD sequence taken from the May/15/2006

3,084,498 sequences; 935,706,57

Searching.....

Sequences producing significant alignments

Accession	Species	Score	E-value
L15440-1	AAA59179.1	107	Homo sapiens insulin protein. 177 1e-43
BC005255-1	AAH05255.1	110	Homo sapiens INS
X70508-1	CAA49913.1	110	Homo sapiens pre
V00565-1	CAA23828.1	110	Homo sapiens pre
M10039-1	AAA59173.1	110	Homo sapiens ins
J00265-1	AAA59172.1	110	Homo sapiens ins
X61092-1	CAA43405.1	110	Cercopithecus ae
X61089-1	CAA43403.1	110	Pan troglodytes
J00336-1	AAA36849.1	110	Macaca fascicula
AY137503-1	AAH06937.1	110	Pongo pygmaeus
AY137500-1	AAH06935.1	110	Gorilla gorilla
AY137497-1	AAH06933.1	110	Pan troglodyte
A48810-1	CAA03148.1	86	unidentified pro
A11939-1	CAA00992.1	80	synthetic constr

RESULT OF BLAST - Microsoft Internet Explorer

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

Blastp for UniProt 82]です。

[CLUSTALW SETUP ([Graphical View](#))] | [Text View](#) (any number of sequences)]

Reference: Altschul, Stephen F., et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= gi|307072|gb|AAA59179.1| (L15440) insulin [Homo sapiens] (107 letters)

Database: /db/SP/216,3

Searching.....

Sequences producing significant alignments:

Accession	Species	Score	E-value
sp Q8HXV2 INS_PONPY	INS_PONPY Insulin precursor	171	4e-43
sp P30410 INS_PANTR	INS_PANTR Insulin precursor	171	4e-43
sp P30406 INS_MACFA	INS_MACFA Insulin precursor	171	4e-43
sp P01308 INS_HUMAN	INS_HUMAN Insulin precursor	171	4e-43
sp Q6YK33 INS_GORGO	INS_GORGO Insulin precursor	171	4e-43
sp P01321 INS_CANFA	INS_CANFA Insulin precursor	171	4e-43
sp P01311 INS_RABIT	INS_RABIT Insulin precursor	158	4e-39
sp Q91X13 INS_SPETR	INS_SPETR Insulin precursor	156	1e-38
sp P01321 INS_CANFA	INS_CANFA Insulin precursor	154	8e-38
sp P01310 INS_HORSE	INS_HORSE Insulin precursor	147	6e-36
sp P01323 INS2_RAT	INS2_RAT Insulin-2 precursor	147	6e-36
sp P01326 INS2_MOUSE	INS2_MOUSE Insulin-2 precursor	147	6e-36
sp P01313 INS_CRILQ	INS_CRILQ Insulin precursor	147	1e-35
sp P01315 INS_PIG	INS_PIG Insulin precursor	144	5e-35

sp|[Q8HXV2](#)|INS_PONPY Insulin precursor [Contains: Insulin B chain... 171 4e-43]

Conclusion

- **Web Services** have great potential to share Bioinformatics Data and Tools in all over the world
- Needs **Automatic Workflow Generation Tools** to make full use of Web Services
- **Bioinformatics Ontology** is a key to establish Interoperability among Bioinformatics Web Services

Acknowledgement

- Daisuke Shinbara Tokyo Institute of Technology (Hitachi, Ltd.)
- Sumi Yoshikawa RIKEN GSC, TITECH

References

Akihiko Konagaya: “Bioinformatics Ontology: Towards the Automatics Generation of Bioinformatics Workflow for Web Services,” in Proc. of Distributed Applications, Web Services, Tools and GRID Infrastructures for Bioinformatics (NETTAB2006), S. Margherita di Pula, Italy (<http://www.nettab.org/2006/>), pp.75-82 (2006)

Akihiko Konagaya: “OBIGrid: Towards the 'Ba' for Sharing Resources, Services and Knowledge for Bioinformatics”, in Proc. of Fourth International Workshop on Biomedical Computations on the Grid (BioGrid), Singapore ([CCGRID 2006](#)), 37 (2006)

Thank You for Listening