

# 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

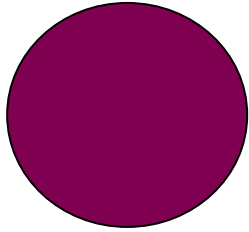
How many colors can you see in rainbow?



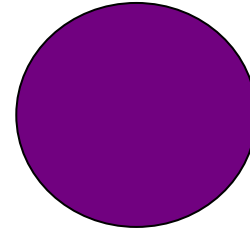
# Ontology for Rainbow Colors



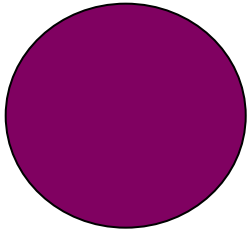
# Which are Purple?



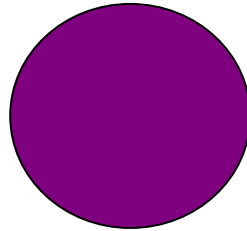
**#800050**



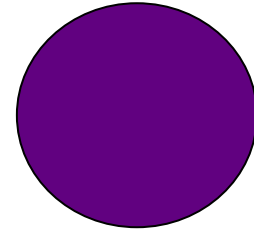
**#700080**



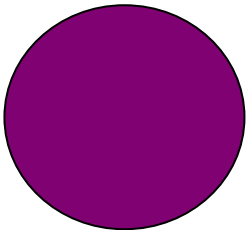
**#800060**



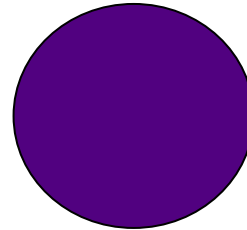
**#800080**



**#600080**

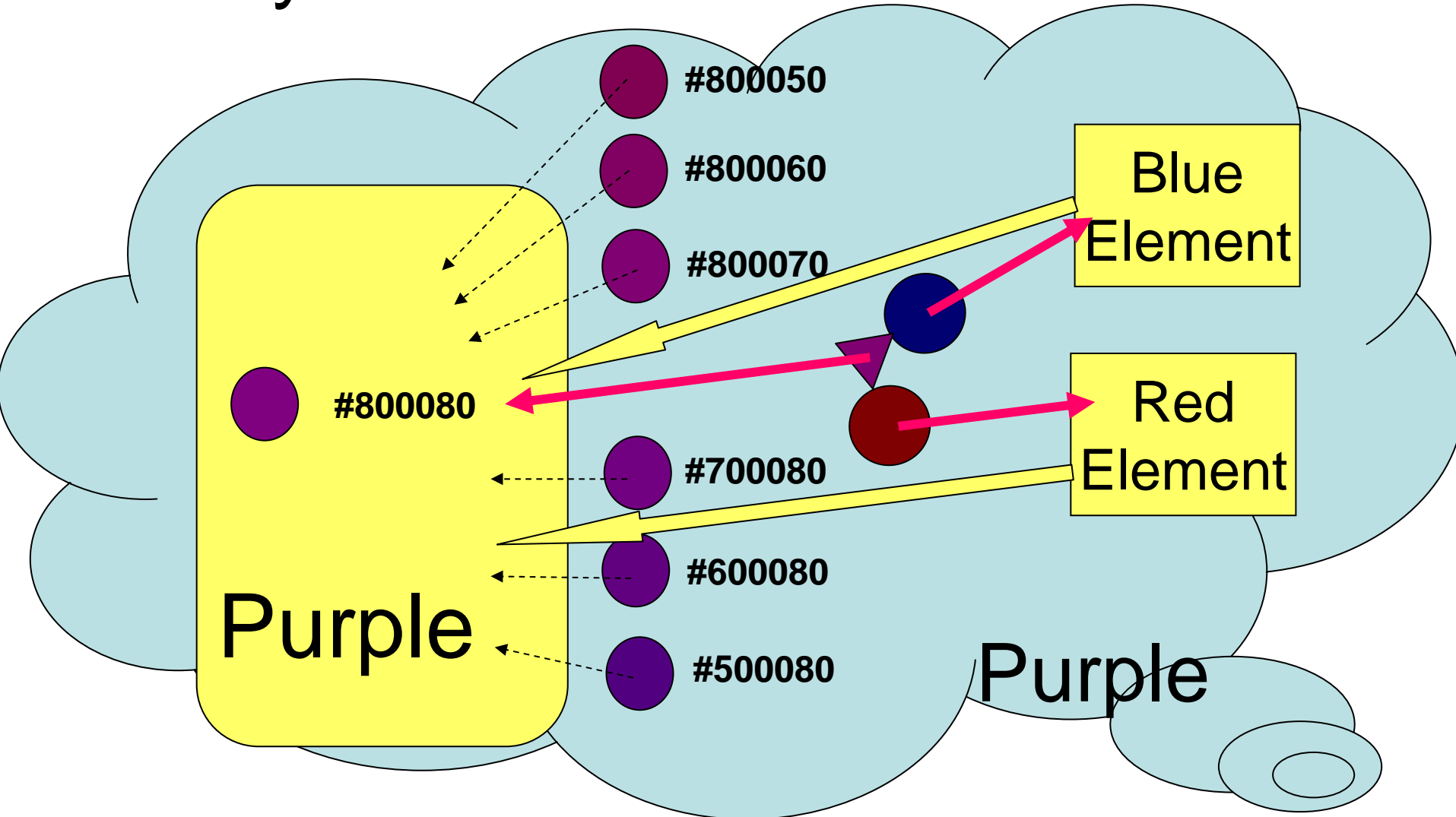


**#800070**



**#500080**

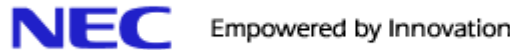
# Representation by Elements and Constructor





# Web Services for Bioinformatics

# Formulation of Community



統計数理研究所



Wakayama University



和歌山大学

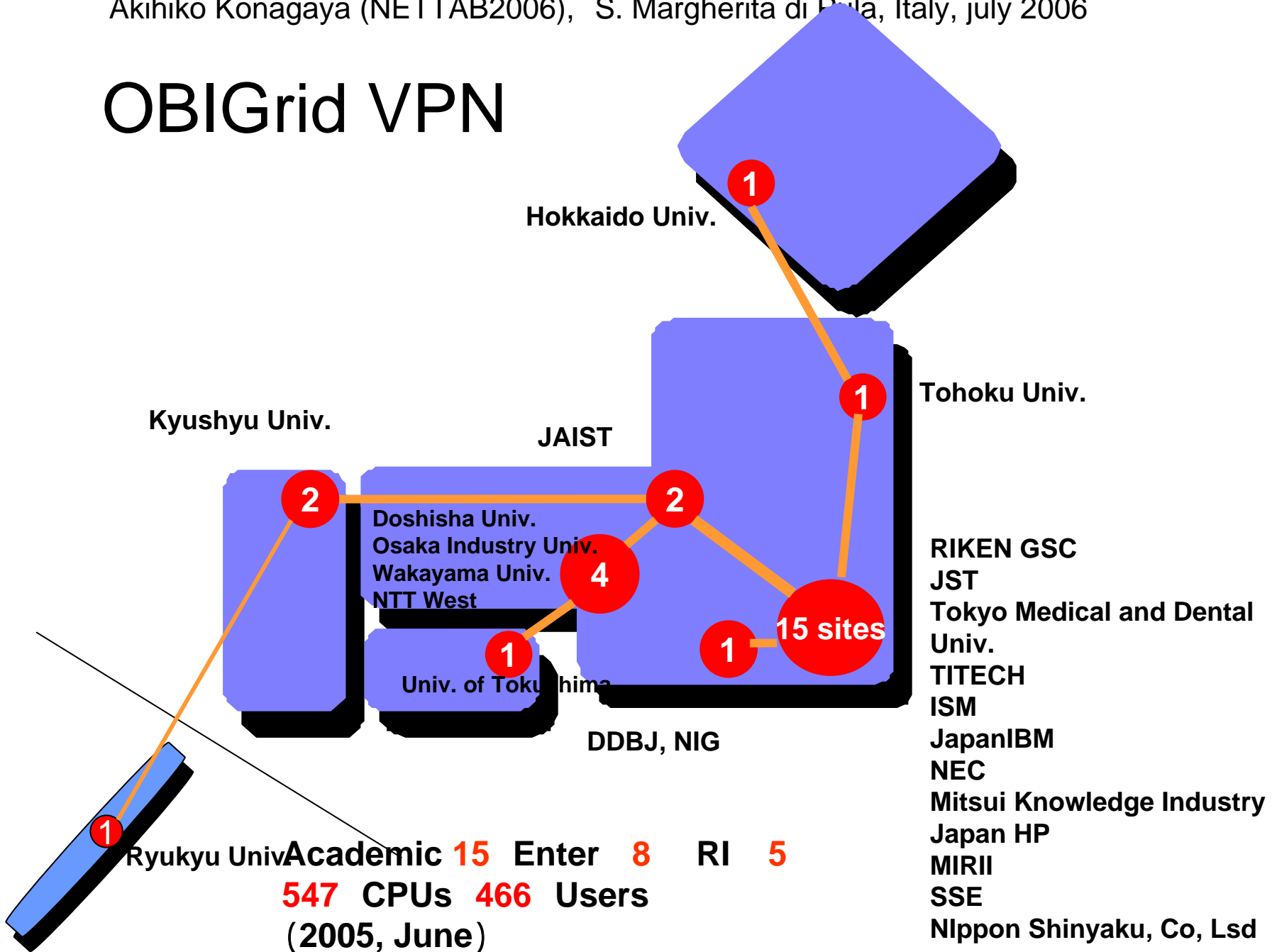


invent

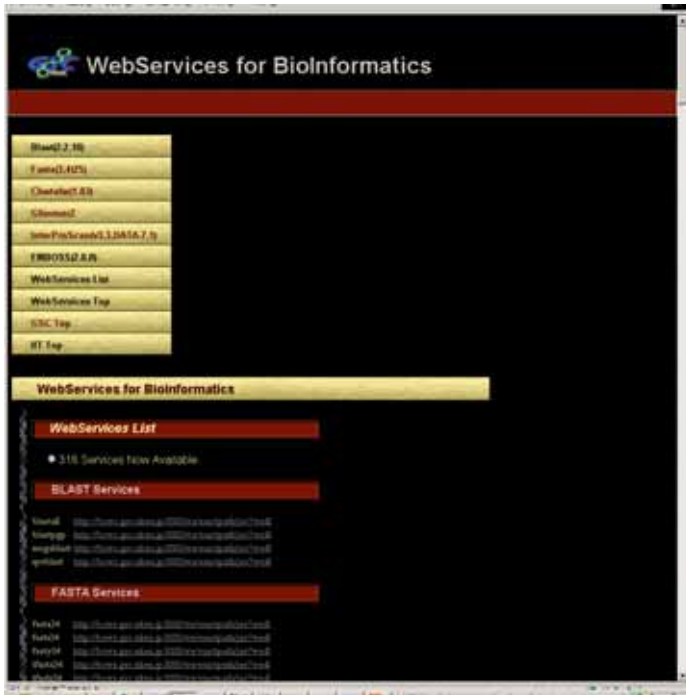
三菱総合研究所



# OBIGrid VPN



# Bioinformatics Web Services on Grid



## GRIDIFIED

BLAST, FASTA, ClustalW,  
Glimmer2, InterProScan,

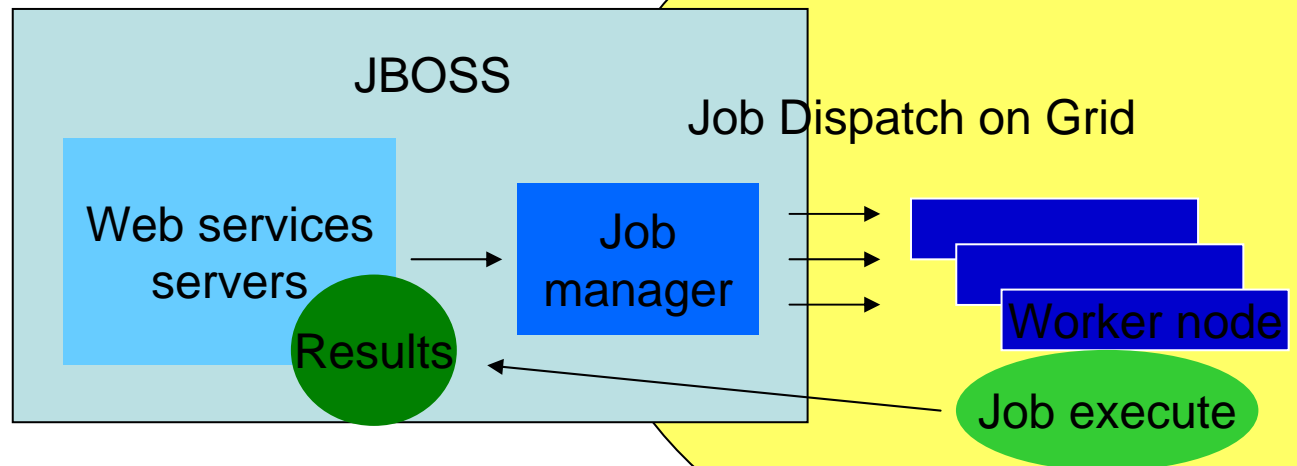
<http://jkt.gsc.riken.jp/sp/spbio/wslist.jsf>

## Client



call web services

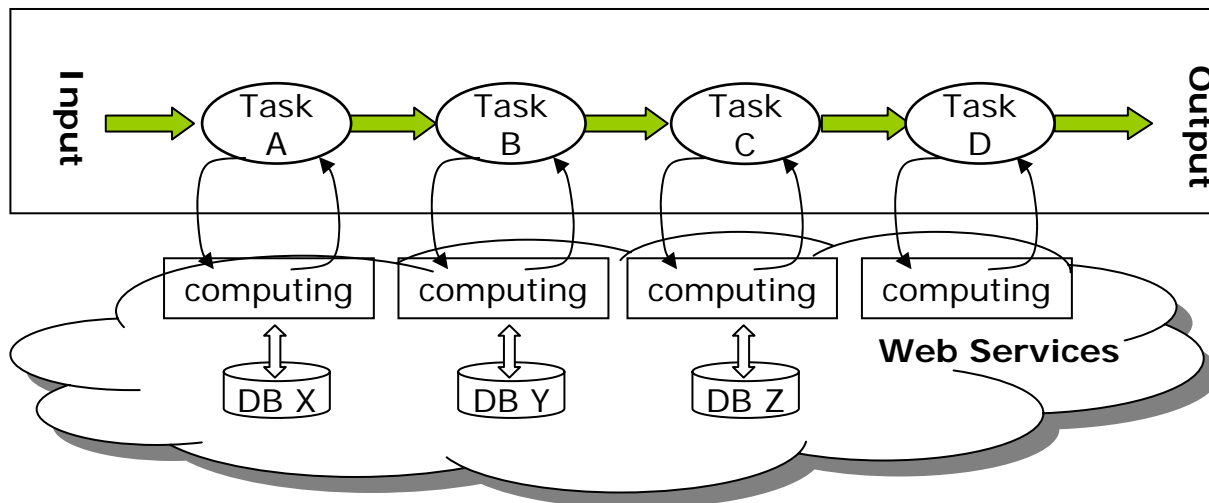
Return



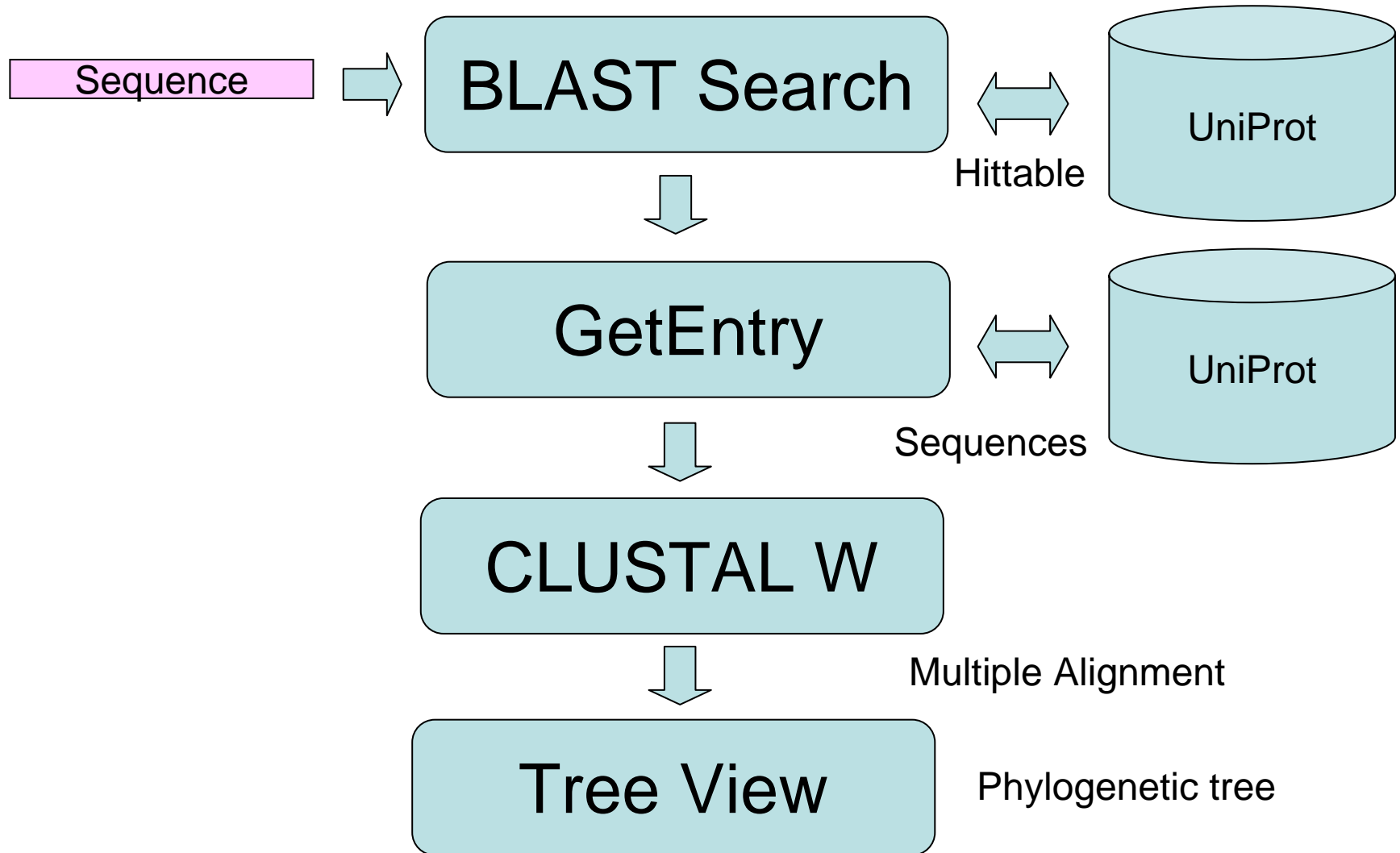
# Advantages of Web Services

- 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 illustrates a manual workflow on web applications for sequence analysis. It consists of several overlapping screenshots from a Microsoft Internet Explorer browser:

- BLAST Search Results:** A screenshot of the DDBJ BLAST search results page, showing search parameters and results.
- getentry Tool:** A screenshot of the 'getentry' tool interface, used for retrieving sequence data by accession number. It includes fields for 'ID: Accession' and 'Output Format'.
- ClustalW Alignment Tool:** A screenshot of the ClustalW alignment tool interface. It shows various options for alignment, such as 'SHOW ALIGNMENT SCORE' (checked), 'DOTSINOUTPUT' (checked), and 'ALIGN' (checked). The 'TYPE' is set to 'AUTO', 'OUTPUT' to 'clustal', and 'OUTORDER' to 'aligned'. The 'MATRIX' is set to 'blosum', and 'GAPOPEN' is set to '40'. The 'ENDGAPS' and 'NOGAPS' options are set to 'OFF'.
- PHYLIP Tree Visualization:** A screenshot of a PHYLIP tree visualization showing the relationships between several species: Rattus norvegicus, Mus musculus, Homo sapiens, Takifugu rubripes, Xenopus laevis, Gallus gallus, Anopheles gambiae, and Danio rerio. The tree shows a clear phylogenetic relationship between the species.

# 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 with several key components:

- Enactor invocation window:** Contains a 'Processor status' table with the following data:

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...	
	list	ProcessComplete	28-Jul-2004 11:39...	

- Advanced model explorer:** Shows a hierarchical view of workflow objects including 'Workflow inputs', 'Workflow outputs', and 'Processors'.
- Available services:** Lists various services such as 'Local Services', 'Soaplab', 'Biomoby', and 'WSDL'.
- Run Workflow window:** Provides options to 'Load Inputs', 'New Input', and 'New List', along with a 'Run Workflow' button.
- Workflow Graph:** A central diagram showing the flow of data between various processors and services, including 'Blast2\_program', 'comparer', 'Fasta\_to\_numbered', 'simplifier', 'ncbiblast', 'repeatmasker', 'retrieve', 'copyright', and 'list'.

# Needs Automatic Workflow Generate Tool from Very High Level Specification

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

**GetEntry** from **UniProt**

apply **CLUSTALW**

apply **TreeView**

Automatics  
Generation

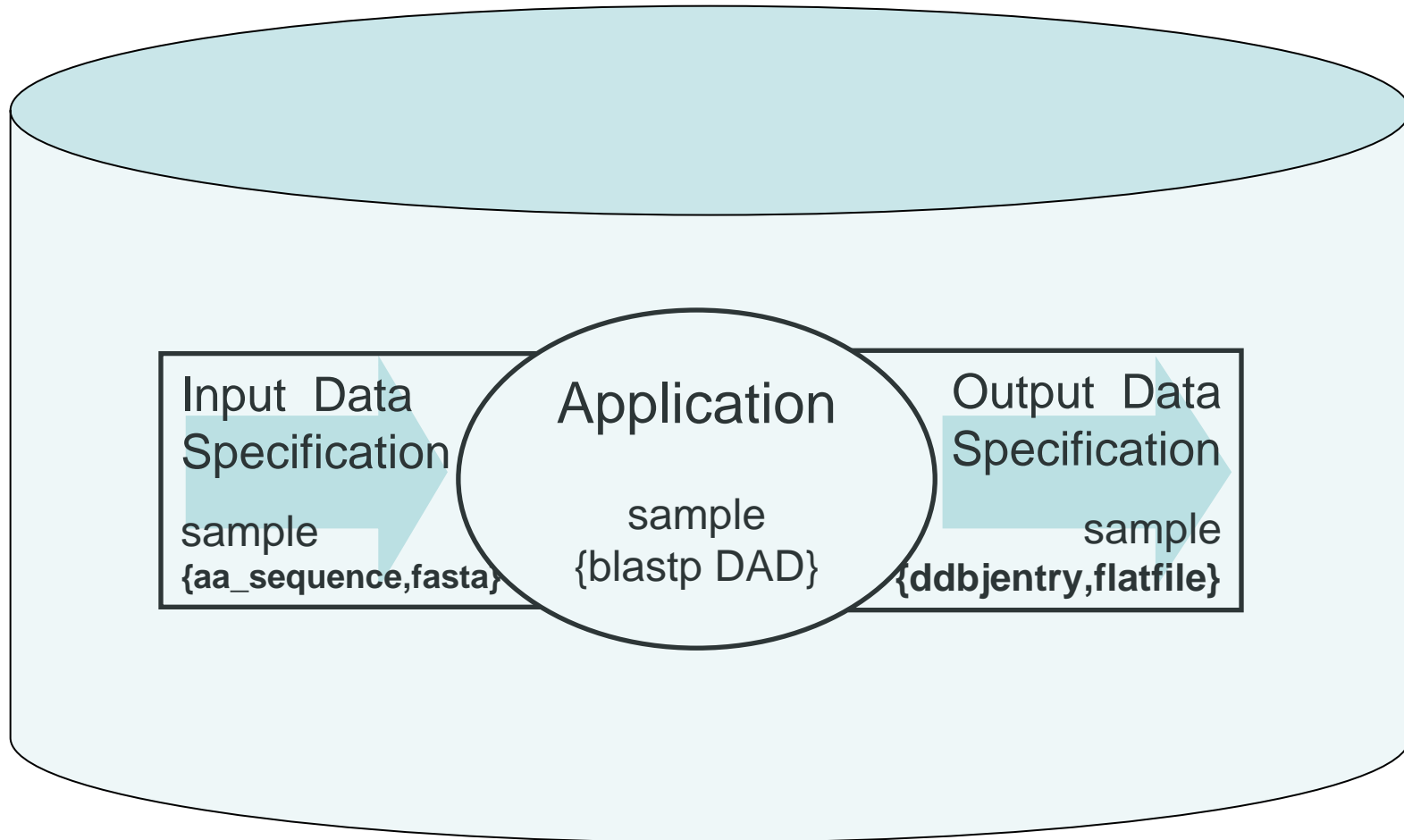


?

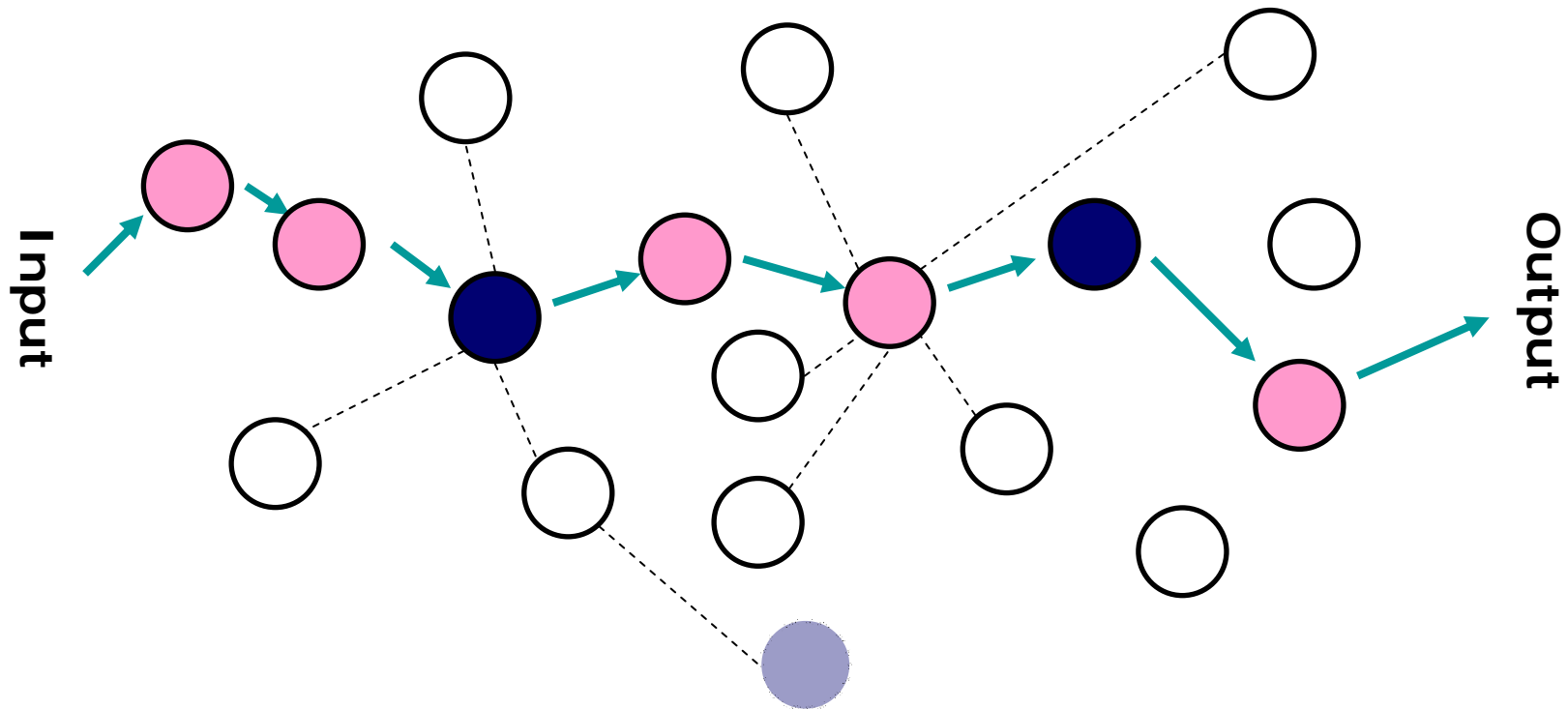
**Workflow  
for  
Bioinformatics Web Services**

# Automatic Generation of Bioinformatics Workflow

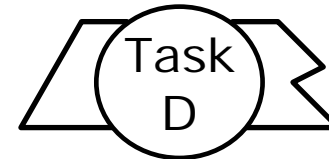
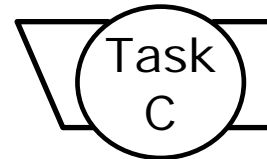
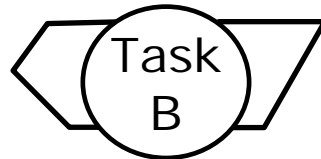
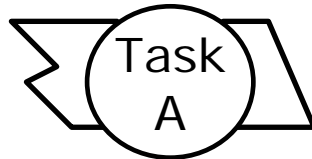
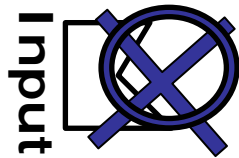
# Task as Atomic Component of Workflow



# 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}

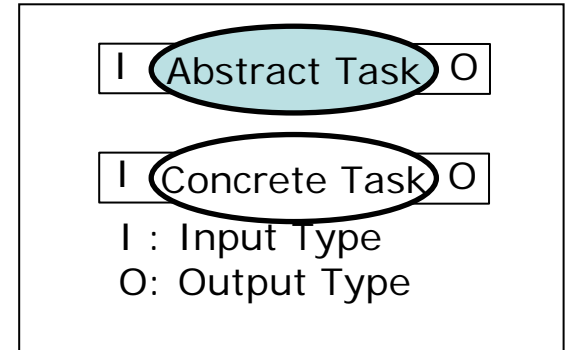
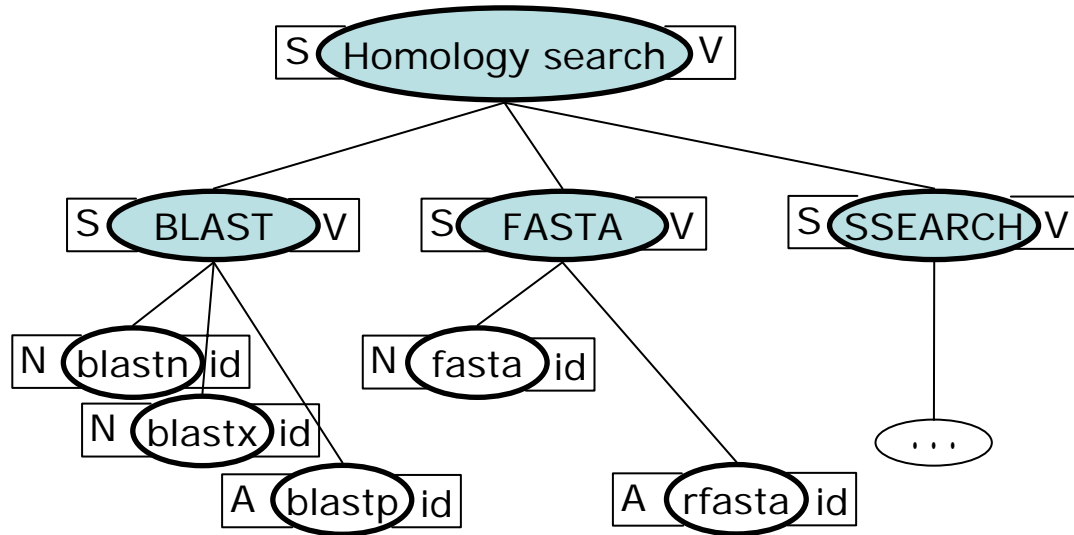
Meta Data  
for Input

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

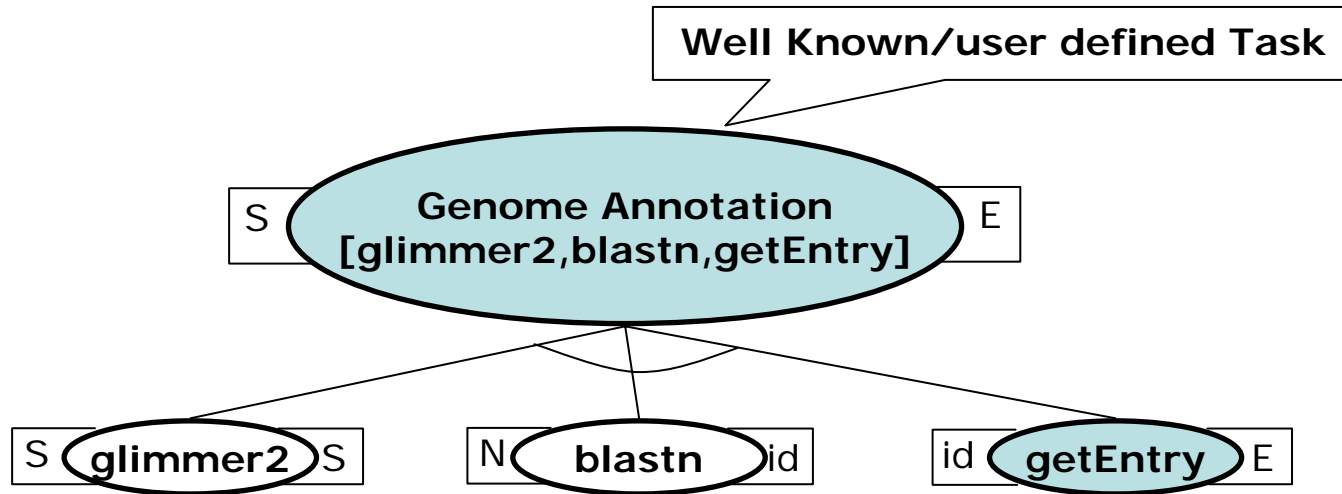
A : Amino acid  
Sequence

id : Accession ID

E : Database Entry



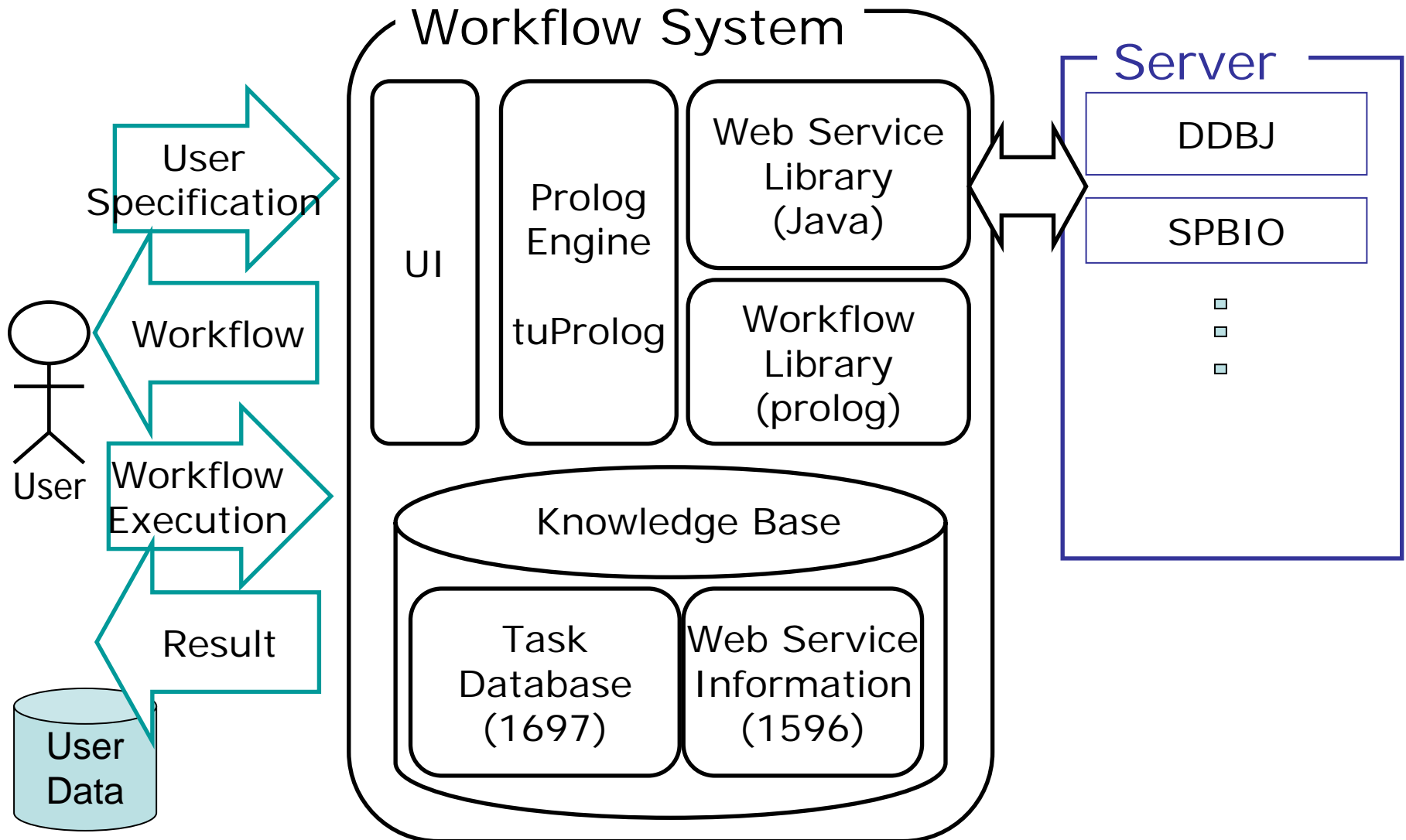
# Task Hierarchy (has\_a)



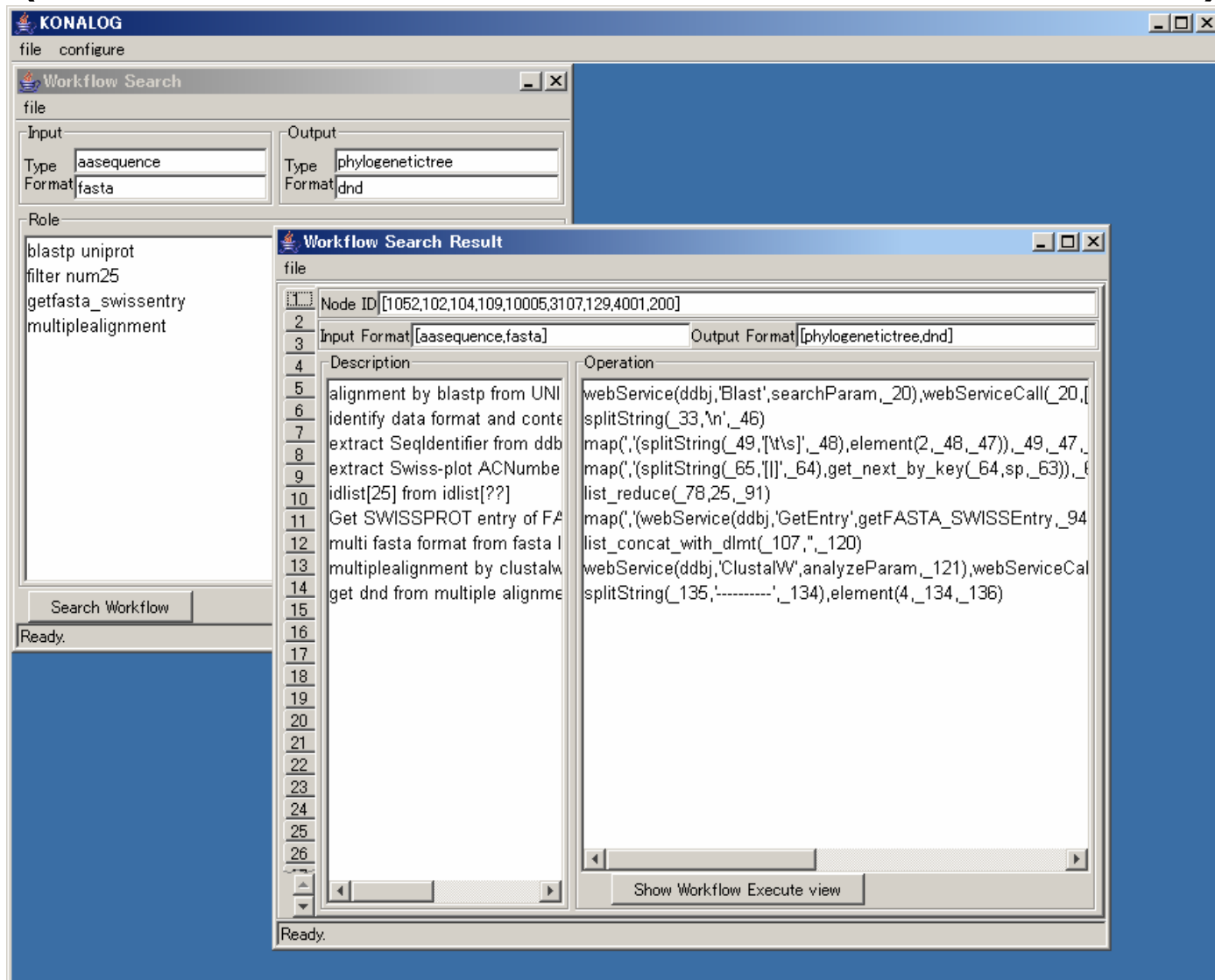
# Prototype for 'Proof of Concept'

- 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)

The screenshot displays the KONALOG Workflow Execution Console interface. The main window is titled "Workflow Execution Console" and contains several panels:

- Input Data:** A text area containing a protein sequence: `>sp|Q13541|4EBP1_HUMAN Eukaryotic translation initiation factor 4E-binding protein 1 (4E-SGGSSCSQTPSRAIPATRRVVLGDGVQLPPGDYSTTPGGTLFSTTPGGTRIYDRKFLME CRNSPVTKTPPRDLPTIPGVTPSSDEPPMEASQSHLRNSPEDKRAGGEESQFEMDI`. Below the text is an "Execute" button.
- Progress:** A list of workflow steps including "alignment by blastp from UNIPROT", "identify data format and content", "extract SeqIdentifier from ddbj BLAS...", "extract Swiss-plot ACNumber from ...", "idlist[25] from idlist[??]", "Get SWISSPROT entry of FASTA Fo...", "multi fasta format from fasta list", "multiplealignment by clustalw with bl...", and "get dnd from multiple alignment result".
- Intermediate Data:** A text area showing the number "1052" followed by a list of sequence identifiers: `'sp|Q13541|4EBP1_HUMAN sp|Q13541|4EBP1_HUM, sp|Q13541|4EBP1_HUMAN sp|Q62622|4EBP1_RAT sp|Q13541|4EBP1_HUMAN sp|Q60876|4EBP1_MOU sp|Q13541|4EBP1_HUMAN tr|Q3TII9|Q3TII9_MOUSE sp|Q13541|4EBP1_HUMAN tr|Q9BG57|Q9BG57_PIG sp|Q13541|4EBP1_HUMAN tr|Q6PFS8|Q6PFS8_BR/ sp|Q13541|4EBP1_HUMAN tr|Q3UFP6|Q3UFP6_MOI sp|Q13541|4EBP1_HUMAN sp|P70445|4EBP2_MOU: sp|Q13541|4EBP1_HUMAN tr|Q497A9|Q497A9_RAT sp|Q13541|4EBP1_HUMAN sp|Q13542|4EBP2_HUM, sp|Q13541|4EBP1_HUMAN tr|Q3UZD4|Q3UZD4_MOU sp|Q13541|4EBP1_HUMAN tr|Q6FG68|Q6FG68_HUM`.
- Result Data:** An empty text area.

The interface also includes a menu bar with "file" and "configure" options, and a status bar at the bottom.



# 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

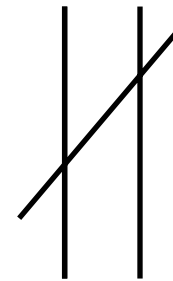
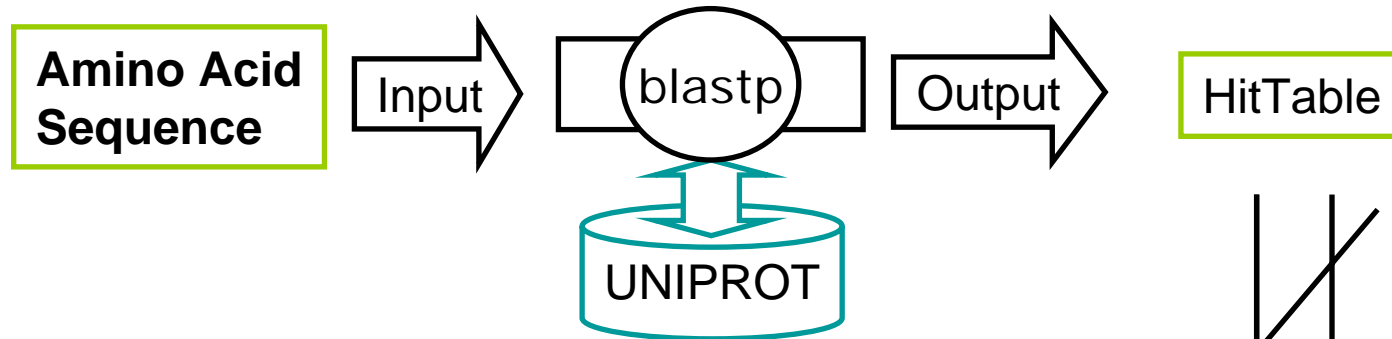
No	Workflow				Applications
	Input		Output		
	Format	Type	Format	Type	
1	fasta	aasequence	gde	aamultiplealignment	blastp uniprot
					filter num25
					getfasta_swissentry
					multiplealignment
2					blastp uniprot
					filter num25
					getfasta_swissentry
					multiplealignment
3	fasta	aasequence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

# Differences of Generated Workflow

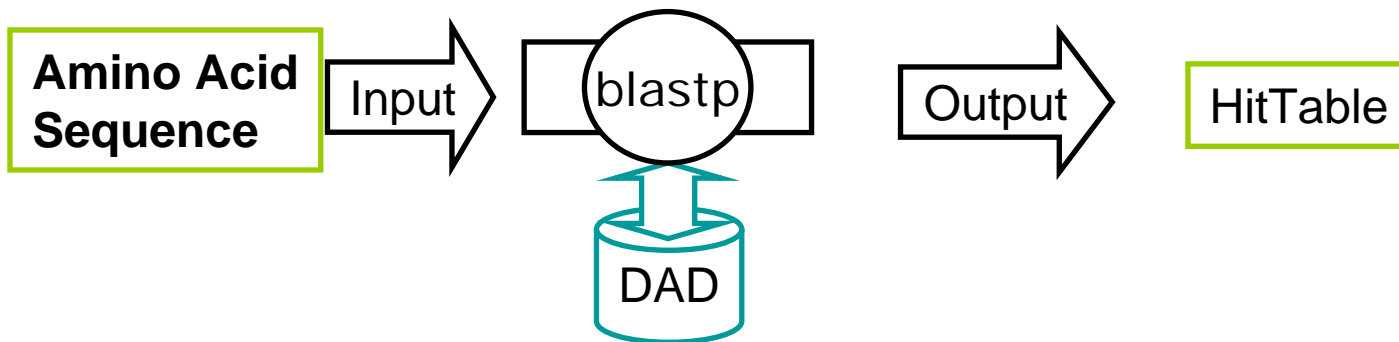
Meta Data	No.	Solution Num	time(ms)	First Match WebserviceCall	
				ID	Description
Input Database Output Full Cmds	1	8	11266	1052	alignment by blastp from UNIPROT
				102	identify data format and content.
				104	extract SeqIdentifier from ddbj BLAST result record
				109	extract Swiss-plot ACNumber 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
				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
				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?



**Lack of Interoperability Between the Web Services**



# Very Similar but not the Same Format

**Blastp for DAD**

[CLUSTALW SETUP (Graphical View<= 100 sequences) | Text View(any number of sequences)]

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= sp|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

<a href="#">L15440-1</a>	AAA59179.1	107 Homo sapiens insulin protein.	<u>177</u>	1e-43
<a href="#">BC005255-1</a>	AAH05255.1	110 Homo sapiens INS		
<a href="#">X70508-1</a>	CAA49913.1	110 Homo sapiens pre		
<a href="#">V00565-1</a>	CAA23828.1	110 Homo sapiens pre		
<a href="#">M10039-1</a>	AAA59173.1	110 Homo sapiens ins		
<a href="#">J00265-1</a>	AAA59172.1	110 Homo sapiens ins		
<a href="#">X61092-1</a>	CAA43405.1	110 Cercopithecus ae		
<a href="#">X61089-1</a>	CAA43403.1	110 Pan troglodytes		
<a href="#">J00336-1</a>	AAA36849.1	110 Macaca fascicula		
<a href="#">AY137503-1</a>	AAN06937.1	110 Pongo pygmaeus		
<a href="#">AY137500-1</a>	AAN06935.1	110 Gorilla gorilla		
<a href="#">AY137497-1</a>	AAN06933.1	110 Pan troglodyte		
<a href="#">A48810-1</a>	CAA03148.1	86 unidentified pro		
<a href="#">A11939-1</a>	CA480997.1	90 synthetic const		

**Blastp for UniProt**

[CLUSTALW SETUP (Graphical View<= 100 sequences) | 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= sp|307072|gb|AAA59179.1 (L15440) insulin [Homo sapiens] (107 letters)

Database: /db/SP/216,3

Searching.....

Sequences producing significant alignments:

	Score (bits)	E Value
sp  <a href="#">Q8HXV2</a>  INS_PONPY Insulin precursor [Contains: Insulin B chain...	<u>171</u>	4e-43
sp  <a href="#">P30410</a>  INS_PANTR Insulin precursor [Contains: Insulin B chain...	<u>171</u>	4e-43
sp  <a href="#">P30406</a>  INS_MACFA Insulin precursor [Contains: Insulin B chain...	<u>171</u>	4e-43
sp  <a href="#">P01308</a>  INS_HUMAN Insulin precursor [Contains: Insulin B chain...	<u>171</u>	4e-43
sp  <a href="#">Q6YK33</a>  INS_GORGO Insulin precursor [Contains: Insulin B chain...	<u>171</u>	4e-43
sp  <a href="#">P30407</a>  INS_CERAE Insulin precursor [Contains: Insulin B chain...	<u>171</u>	4e-43
sp  <a href="#">P01311</a>  INS_RABIT Insulin precursor [Contains: Insulin B chain...	<u>158</u>	4e-39
sp  <a href="#">Q91X13</a>  INS_SPETR Insulin precursor [Contains: Insulin B chain...	<u>156</u>	1e-38
sp  <a href="#">P01321</a>  INS_CANFA Insulin precursor [Contains: Insulin B chain...	<u>154</u>	8e-38
sp  <a href="#">P01310</a>  INS_HORSE Insulin precursor [Contains: Insulin B chain...	<u>147</u>	6e-36
sp  <a href="#">P01323</a>  INS2_RAT Insulin-2 precursor [Contains: Insulin-2 B ch...	<u>147</u>	6e-36
sp  <a href="#">P01326</a>  INS2_MOUSE Insulin-2 precursor [Contains: Insulin-2 B ...	<u>147</u>	6e-36
sp  <a href="#">P01313</a>  INS_CRIL0 Insulin precursor [Contains: Insulin B chain...	<u>147</u>	1e-35
sp  <a href="#">P01315</a>  INS_PIG Insulin precursor [Contains: Insulin B chain; ...	<u>144</u>	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**