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'.

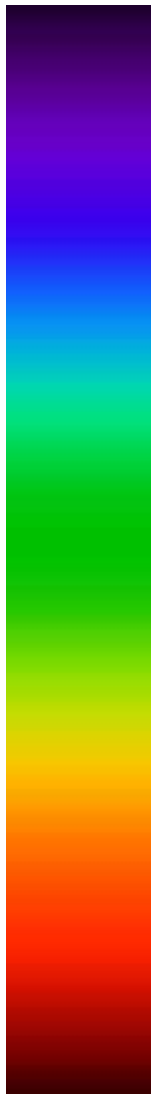
Michael Polanyi, "The Tacit Dimension" 1967

Michael Polanyi (1891-1976)

Rainbow Color

How many colors can you see in rainbow?

Ontology for Rainbow Colors



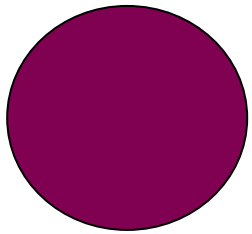
From 360 nm〜400 nm

All the colors you can see with your own eyes!

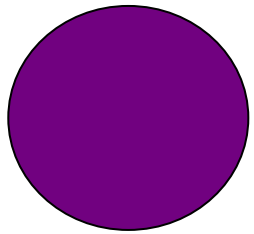
to 760 nm〜830 nm

	RGB Value
Purple	**#800080**
Indigo	**#000080**
Blue	**#0000FF**
Green	**#008000**
Yellow	**#FFFF00**
Orange	**#FF8000**
Red	**#FF0000**

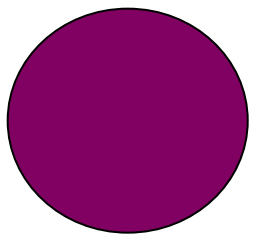
Which are Purple?



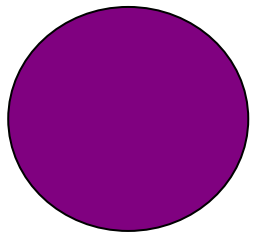
#800050



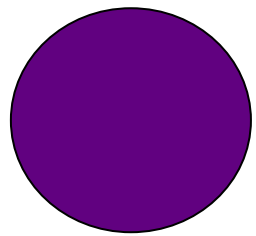
#700080



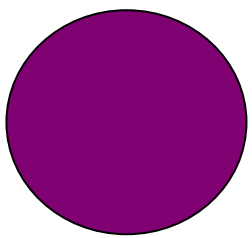
#800060



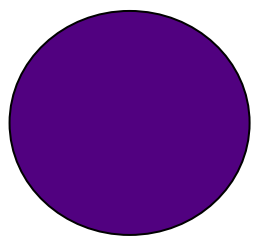
#800080



#600080

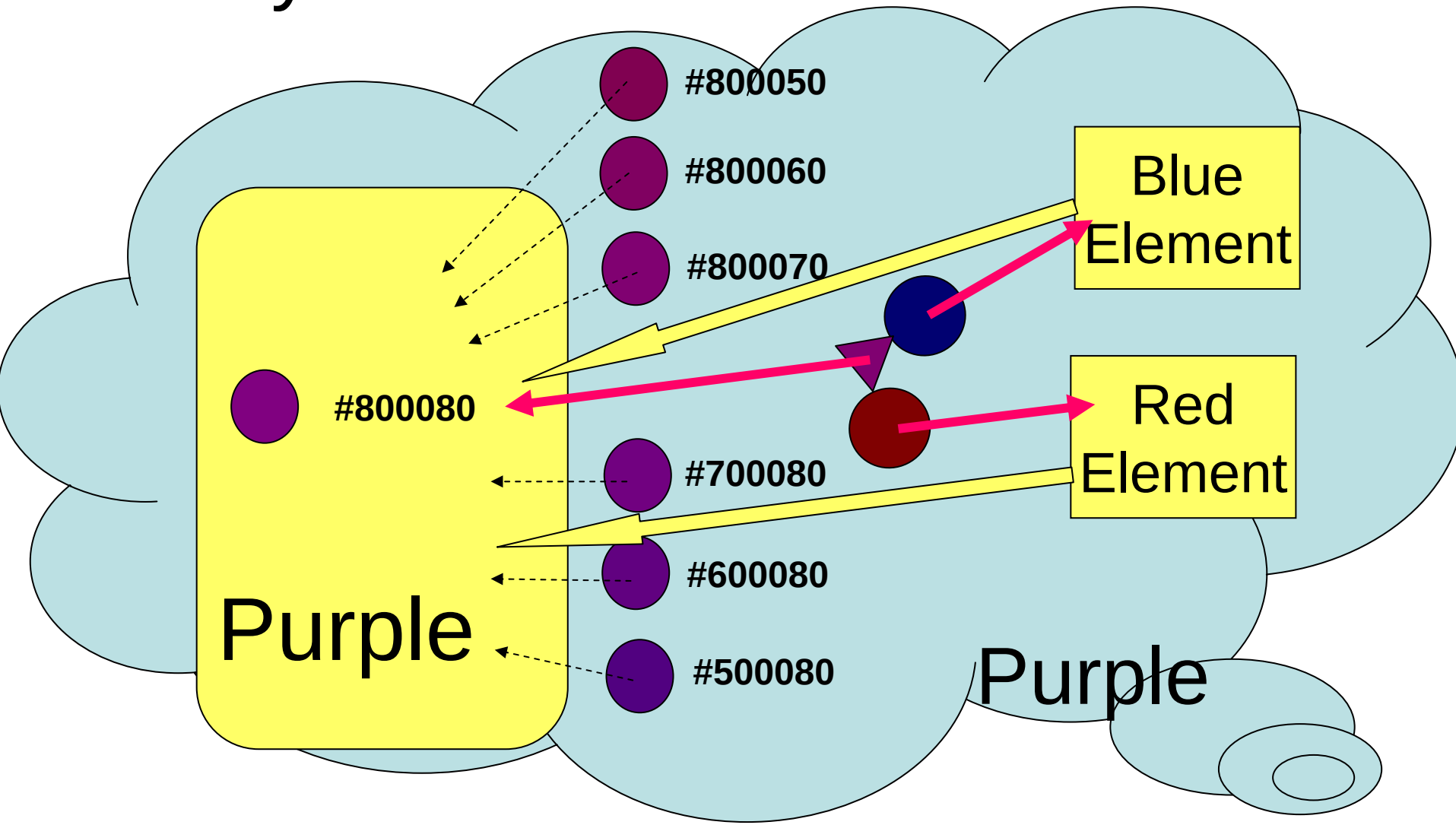


#800070



#500080

Representation by Elements and Constructor



Web Services for Bioinformatics

Formulation of Community

東京医科歯科大学
Tokyo Medical and Dental University

IBM

科学技術振興事業団
Japan Science and Technology Corporation

NEC
Empowered by Innovation

国立遺伝学研究所
NATIONAL INSTITUTE OF GENETICS

JAIST

GSC RIKEN ゲノム科学総合研究センター

統計数理研究所

大阪産業大学
Osaka Sangyo University

徳島大学

MKI 三井情報開発株式会社
Mitsui Knowledge Industry

SSE

MIZUHO

DOSHISHA UNIVERSITY 同志社大学

Wakayama University
和歌山大学

NTT西日本
"光"、ひろがる。ひびきあう。

東京工業大学
Tokyo Institute of Technology

hp
invent

三菱総合研究所

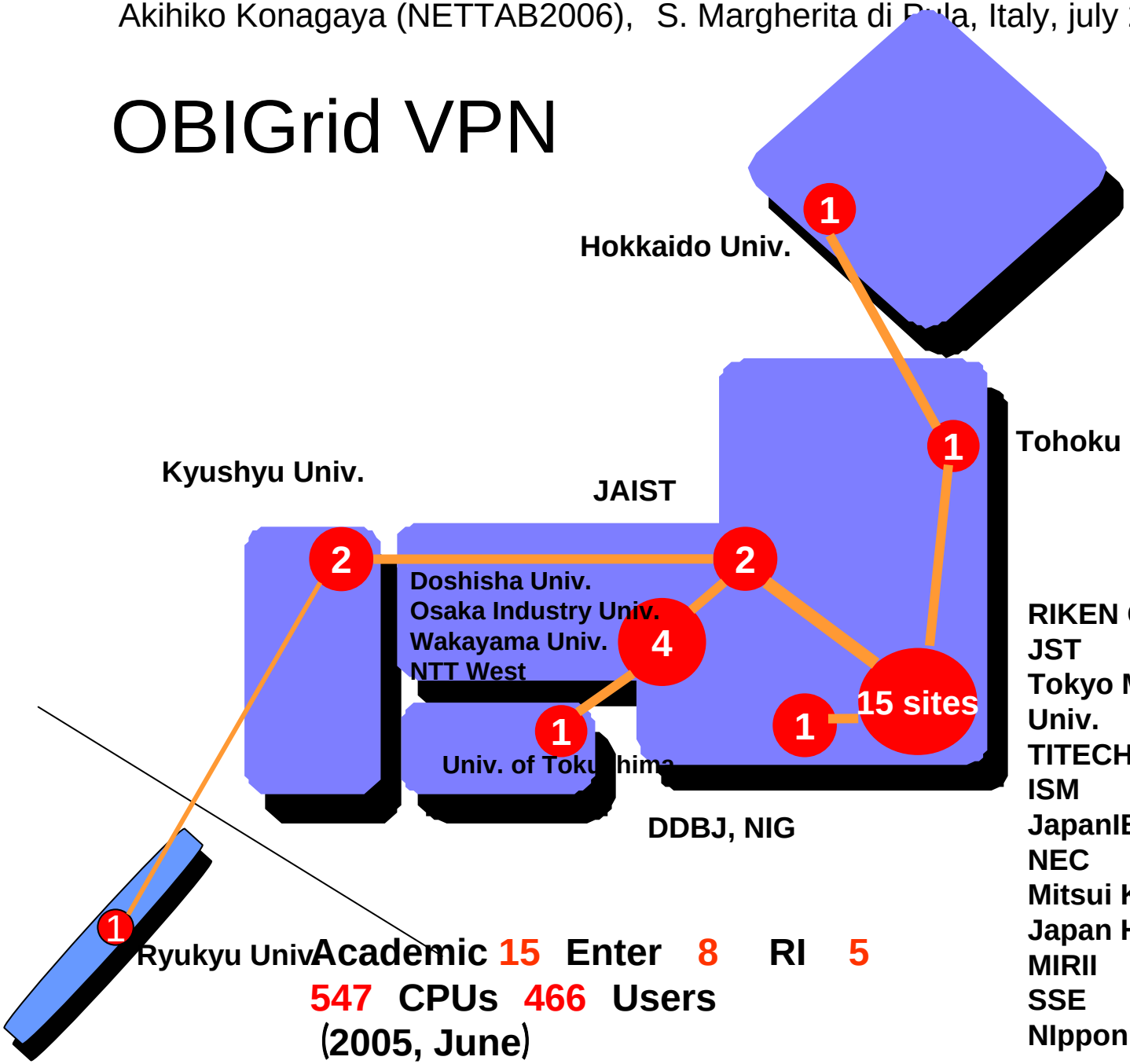
富士総合研究所

九州大学
Kyushu University

日本新薬

琉球大学 University of the Ryukyus

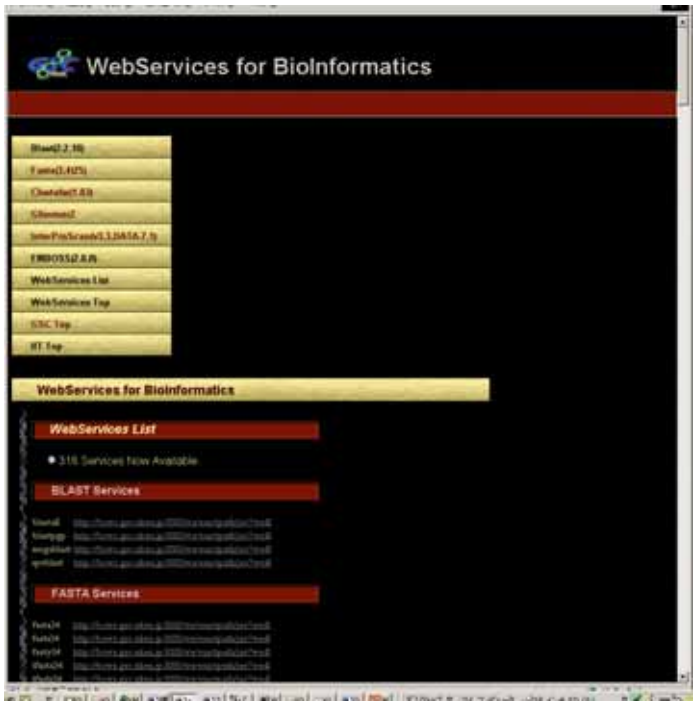
OBIGrid VPN



RIKEN GSC
JST
Tokyo Medical and Dental Univ.
TITECH
ISM
JapanIBM
NEC
Mitsui Knowledge Industry
Japan HP
MIRII
SSE
NIppon Shinyaku, Co, Lsd

Academic 15 Enter 8 RI 5
547 CPUs 466 Users
(2005, June)

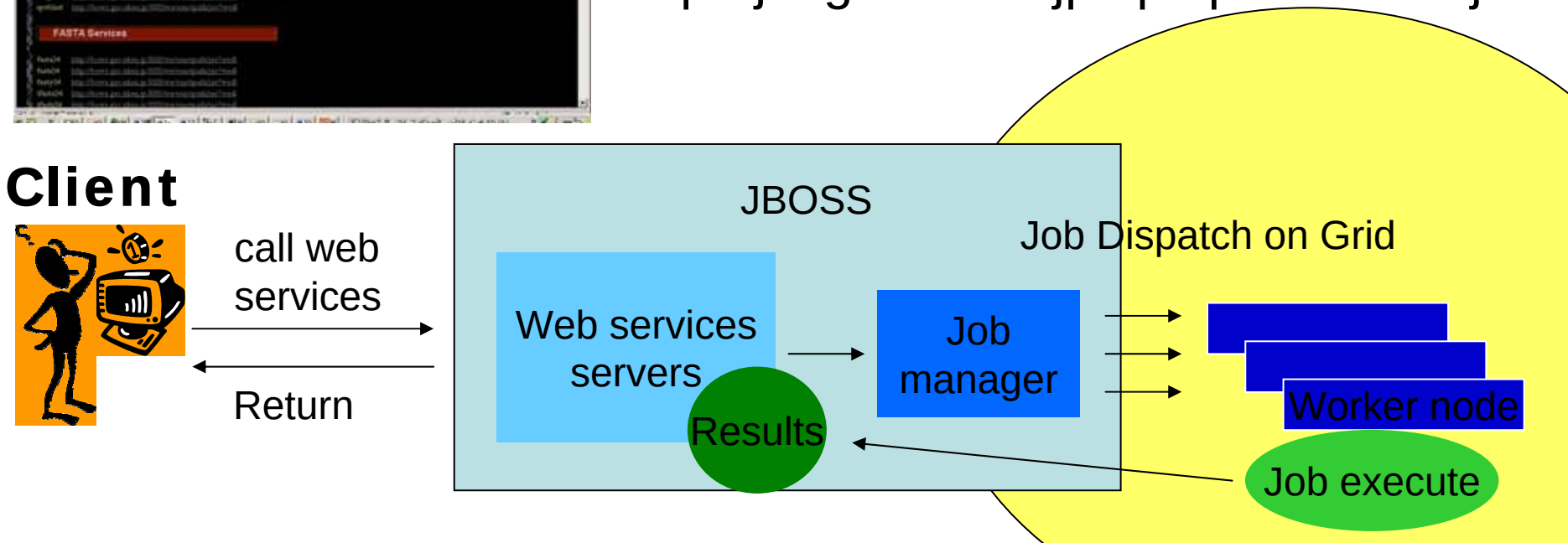
Bioinformatics Web Services on Grid



GRIDIFIED

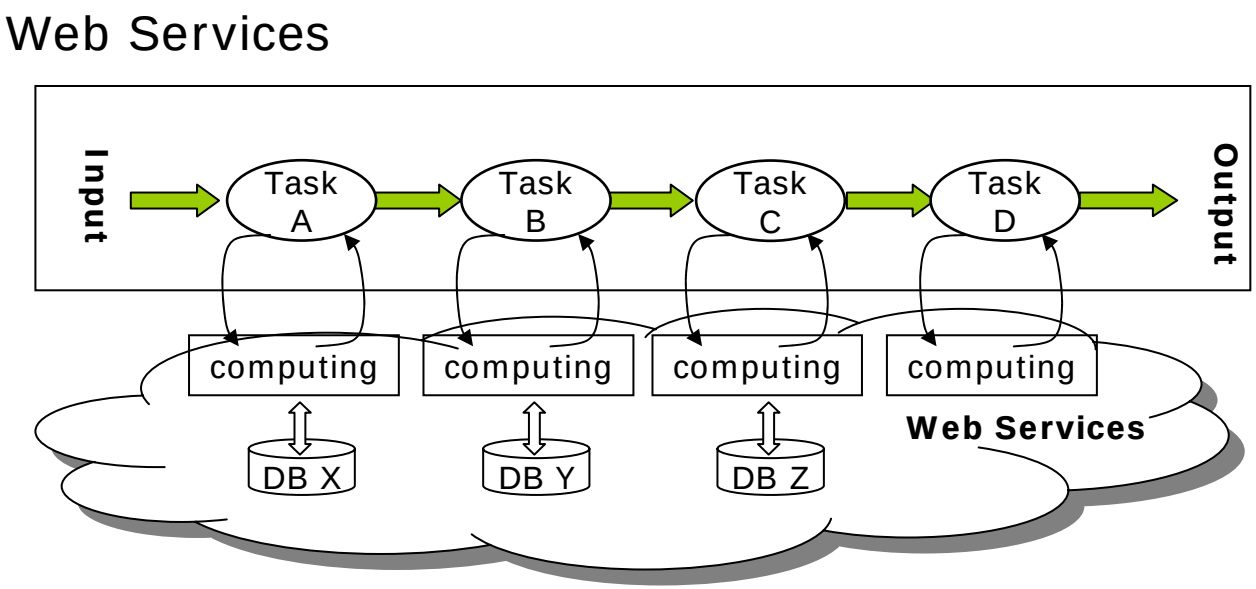
BLAST,FASTA, ClustalW, Glimmer2, InterProScan,

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

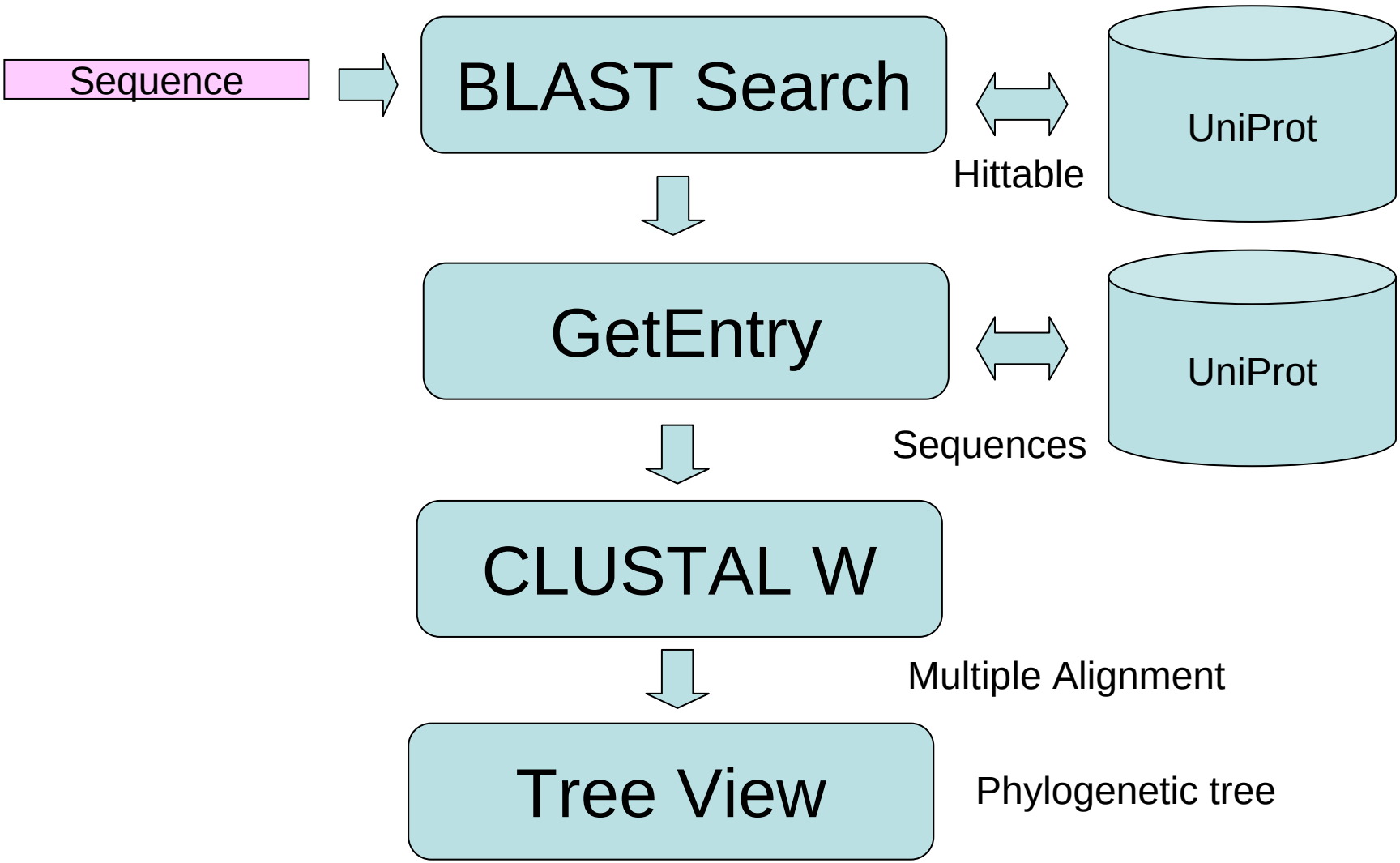


Advantages of Web Services

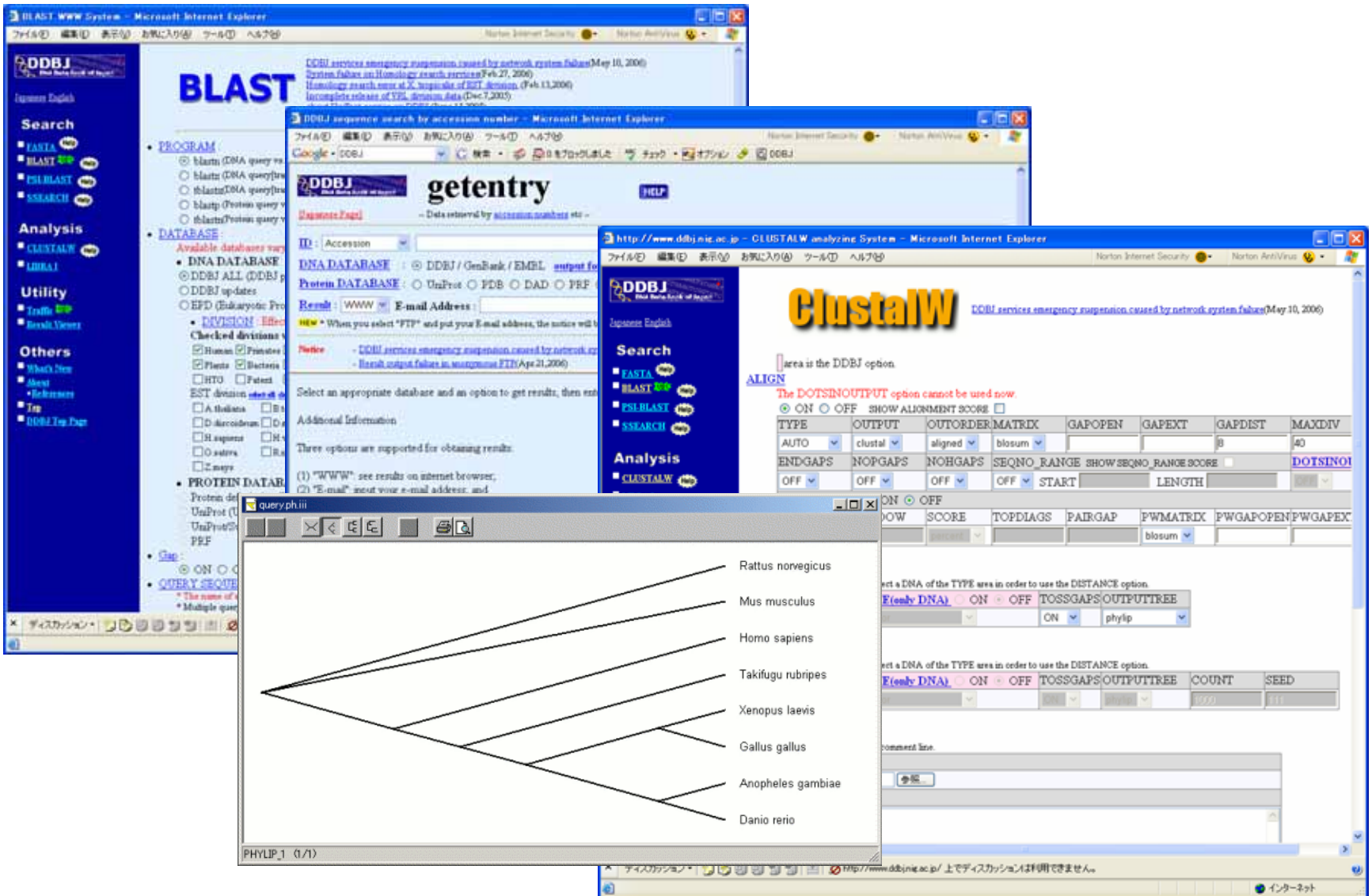
- Liberating from the maintenance of biological databases and tools
- Scalability of computational resources
- High-level application programming interface



Very Simple Work Flow



Manual Workflow on Web Apps

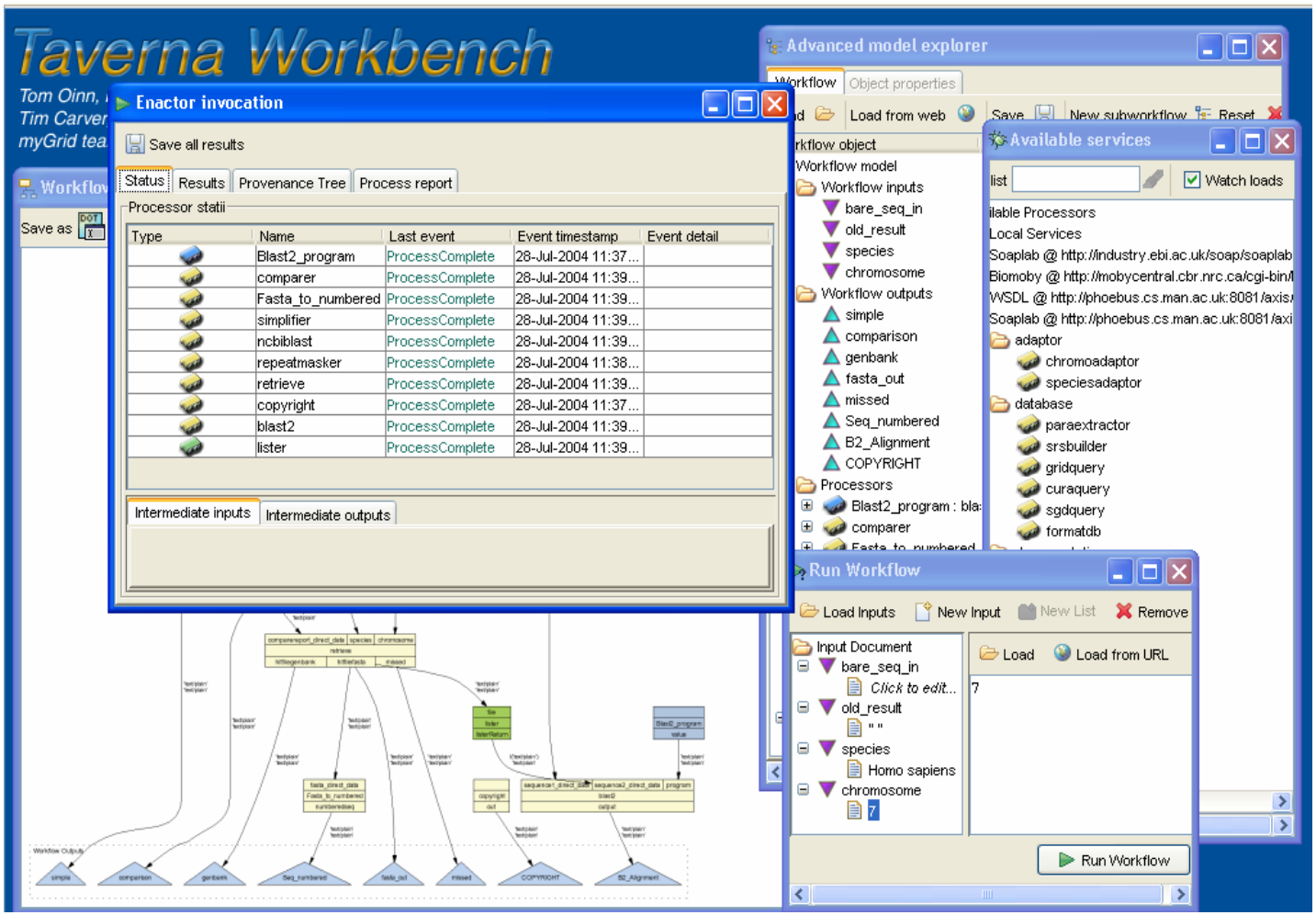


Web Service Programming

```
#!/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?



http://www.cyclonic.org/Taverna_and_myGrid.ppt

Needs Automatic Workflow Generate Tool from Very High Level Specification

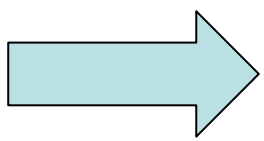
apply **Blastp** to **UniProt**

GetEntry from **UniProt**

apply **CLUSTALW**

apply **TreeView**

Automatics Generation

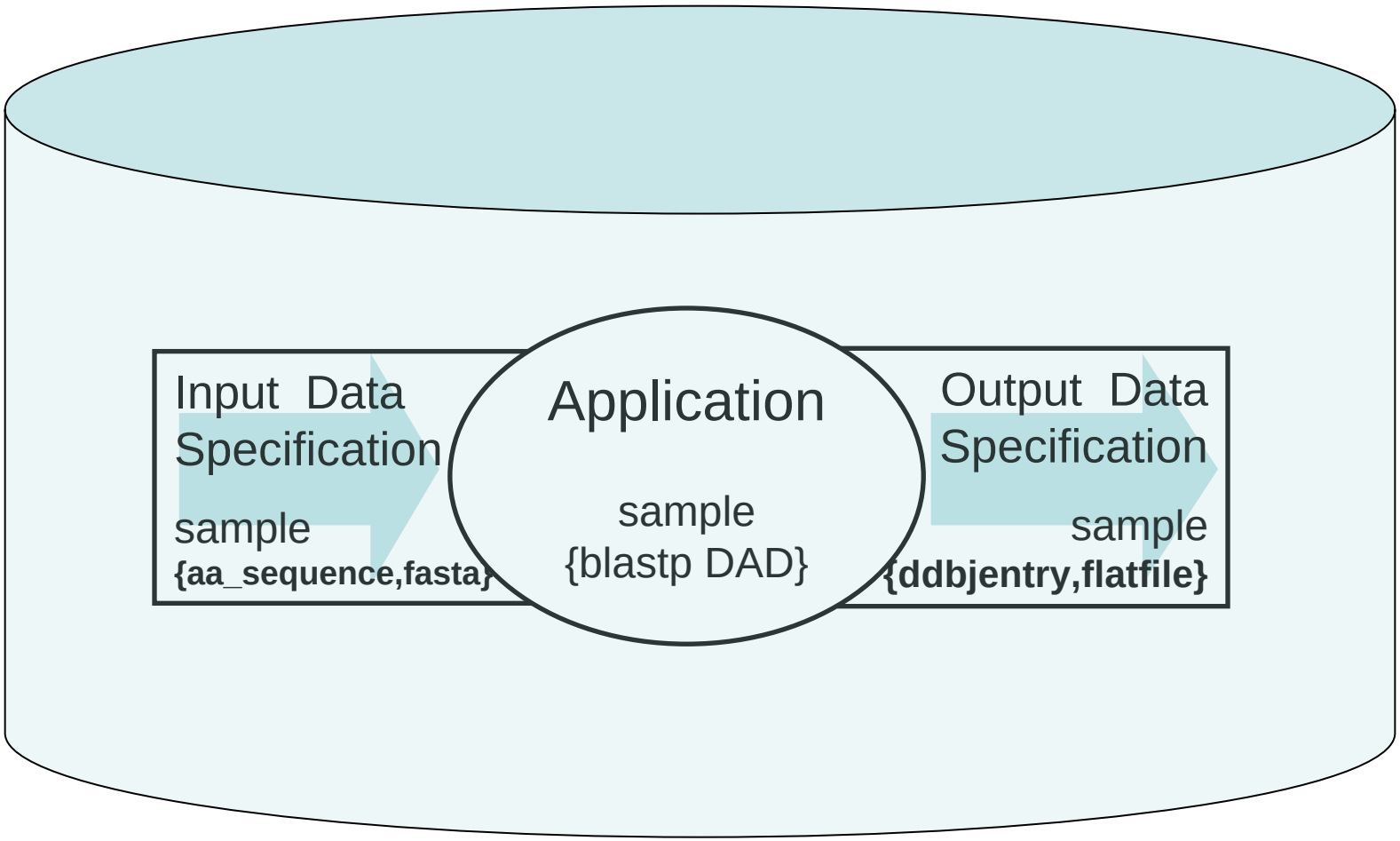


?

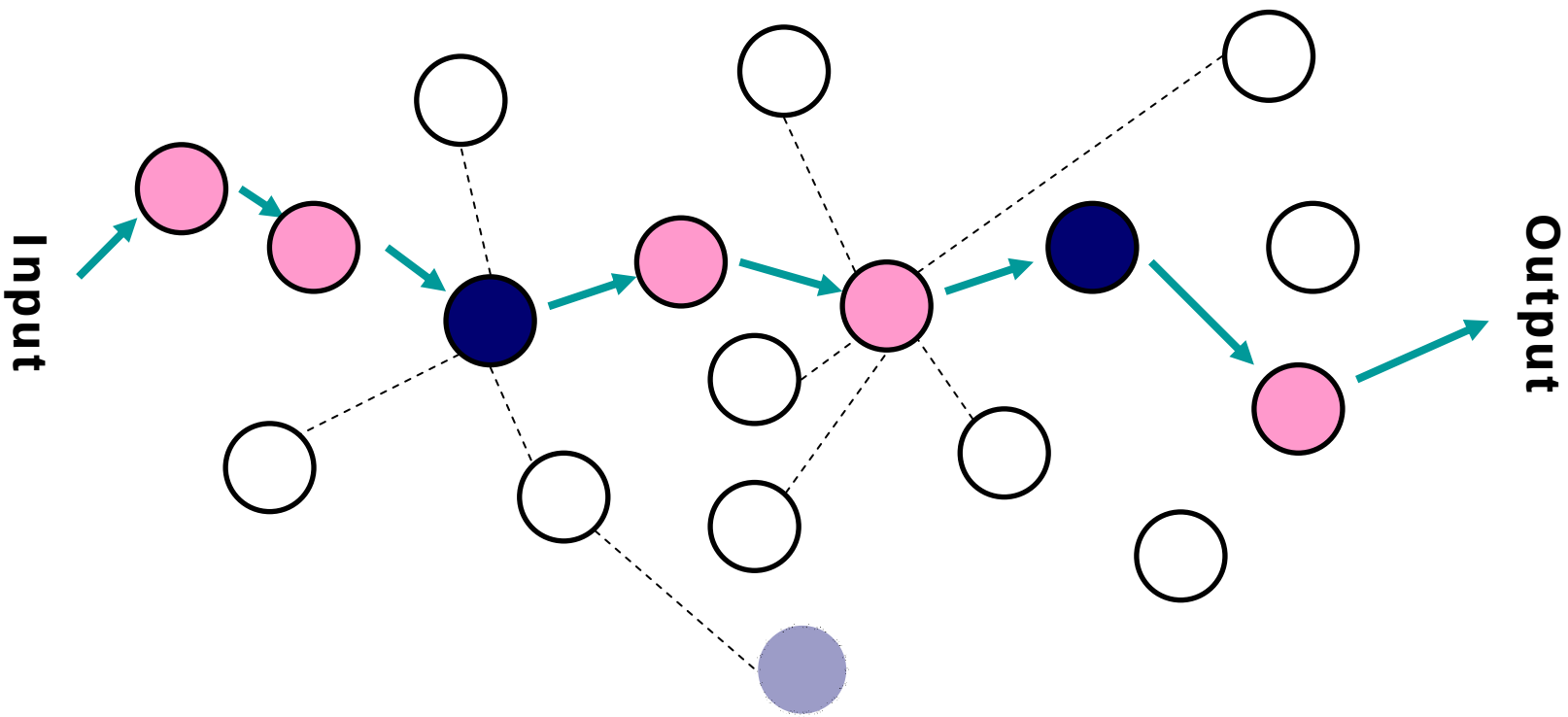
Workflow for Bioinformatics Web Services

Automatic Generation of Bioinformatics Workflow

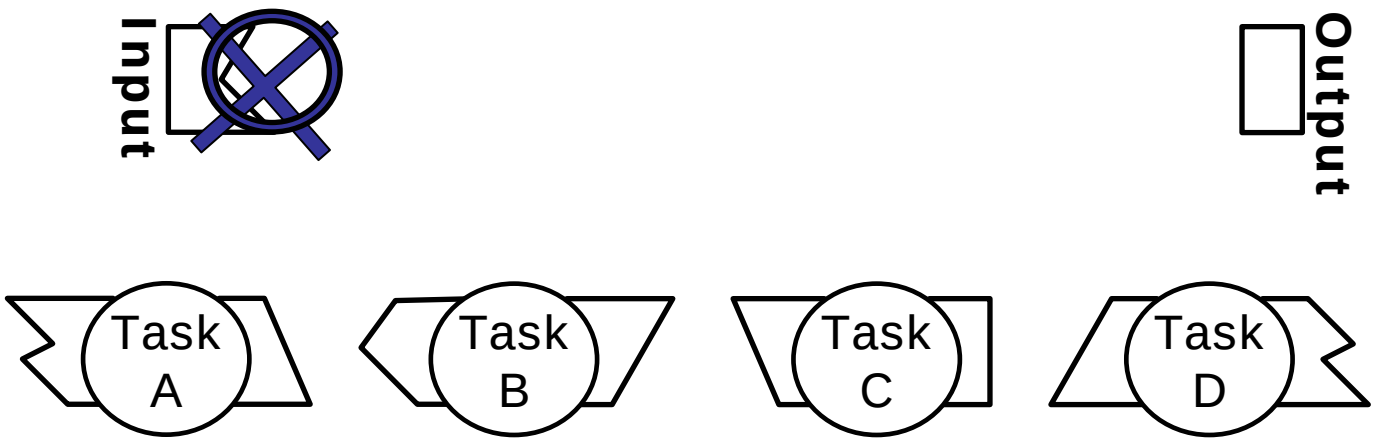
Task as Atomic Component of Workflow



Workflow as a Sequence of Tasks

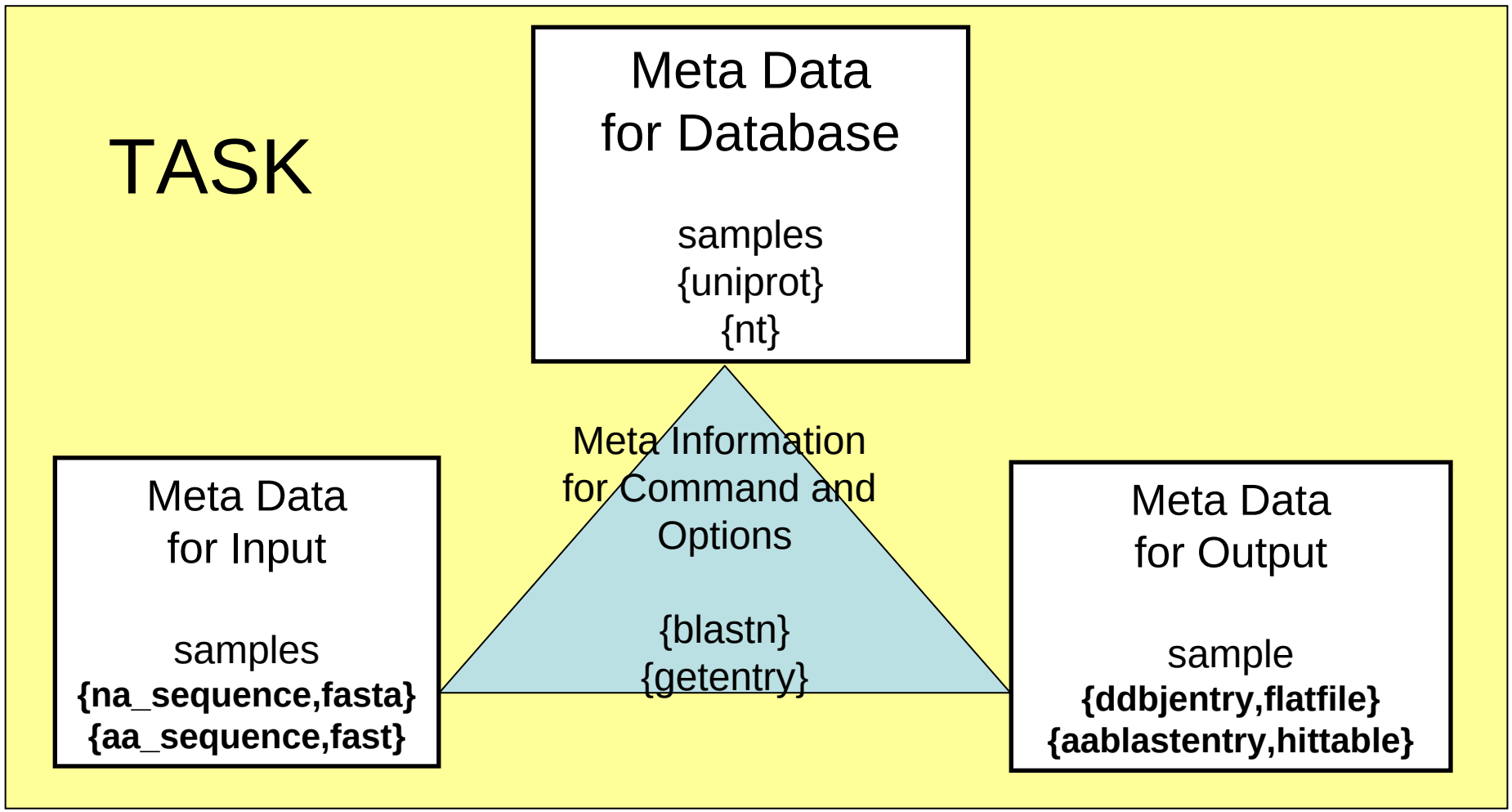


Automatic Generation of Workflow from Given Input and Output Data Specification and Tasks

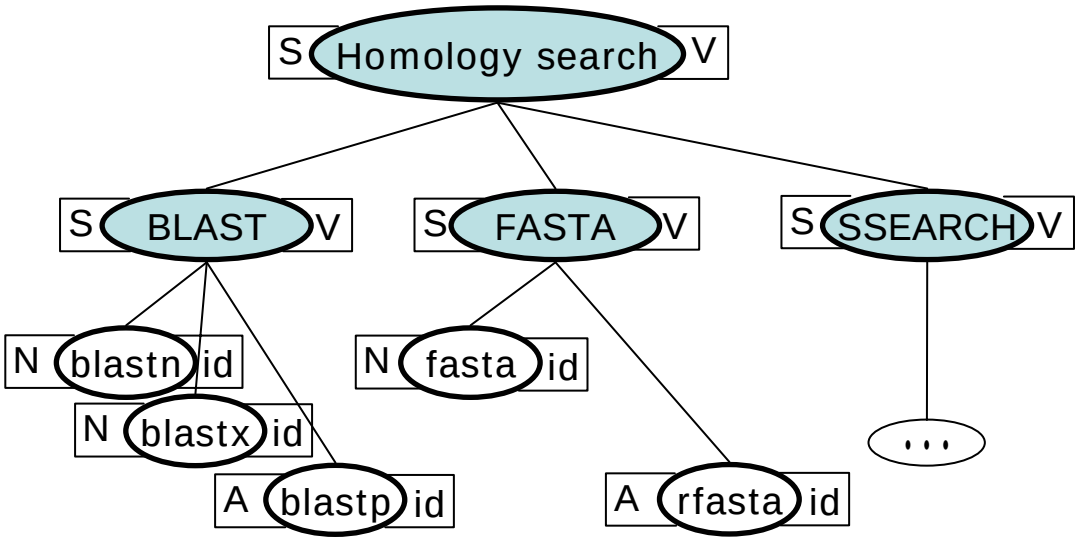


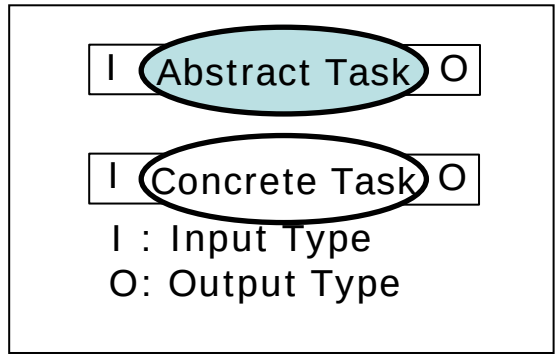
- Path Finding using Meta Information

Meta Information to Specify the Functionality of Task



Task Hierarchy (is_a)





S : Sequence or Sequence Name

V : Various Type

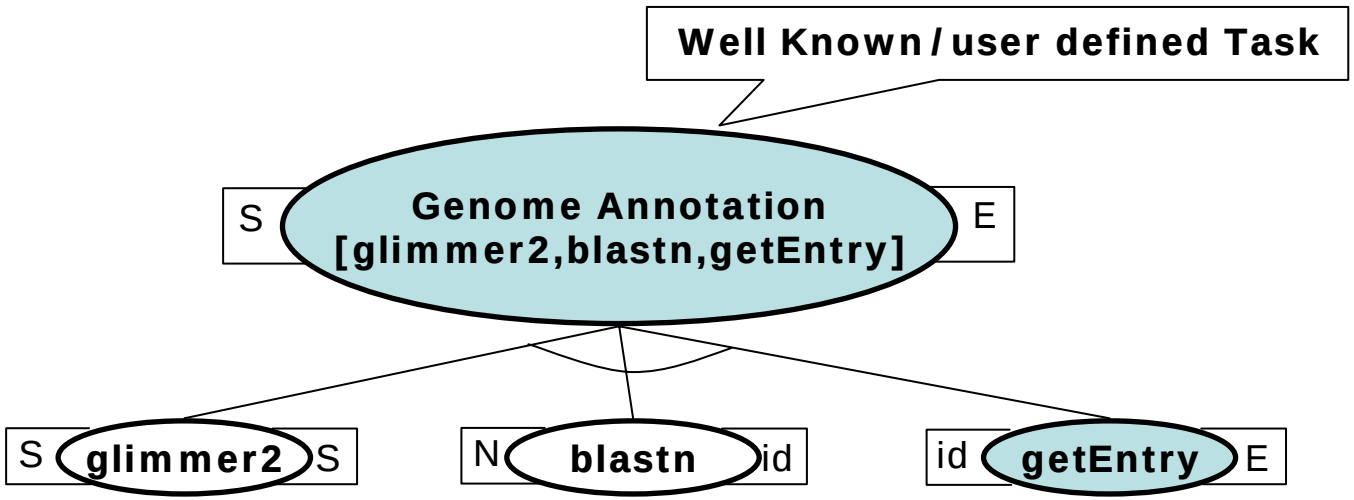
N : Nucleoside Sequence

A : Amino acid Sequence

id : Accession ID

E : Database Entry

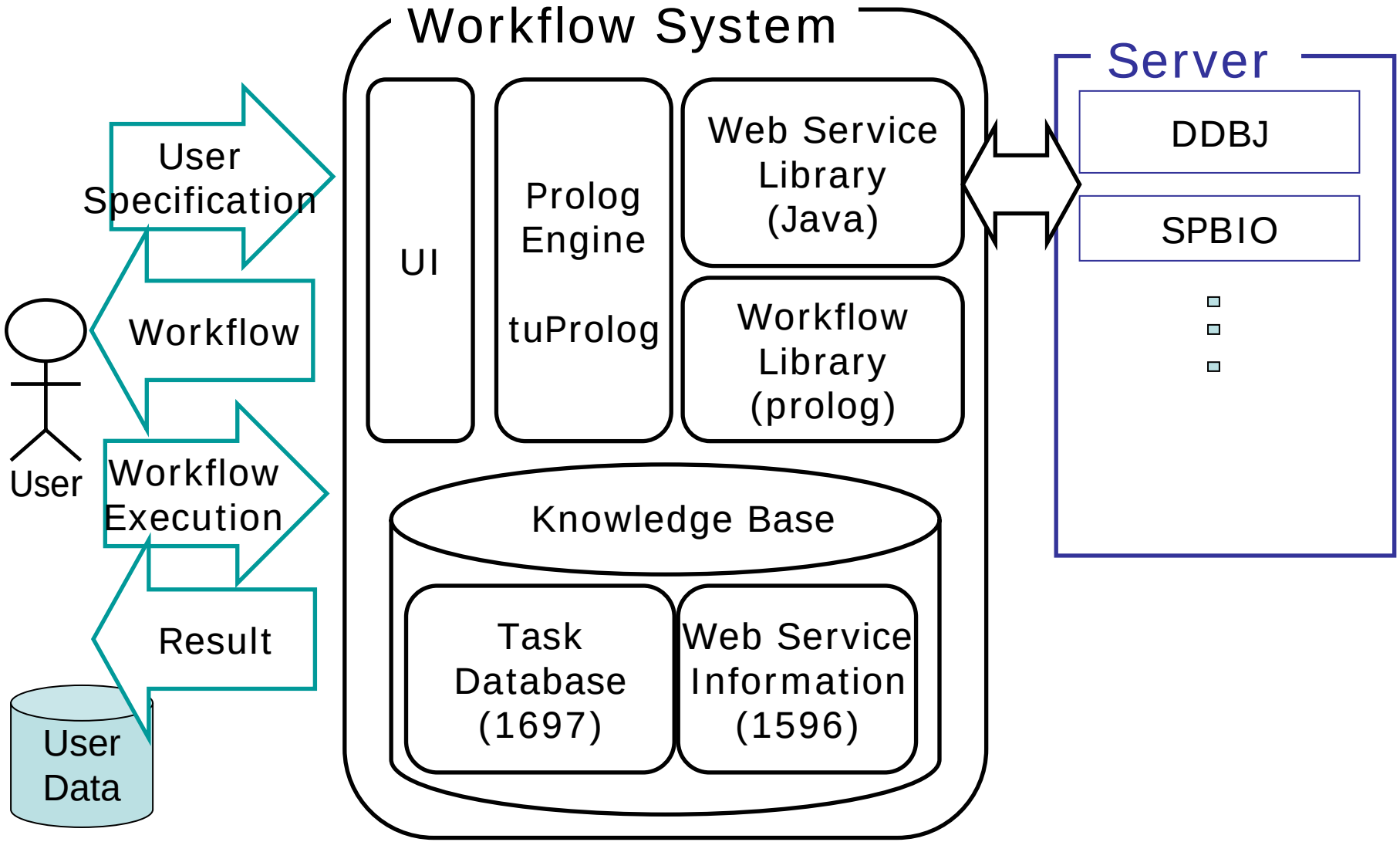
Task Hierarchy (has_a)



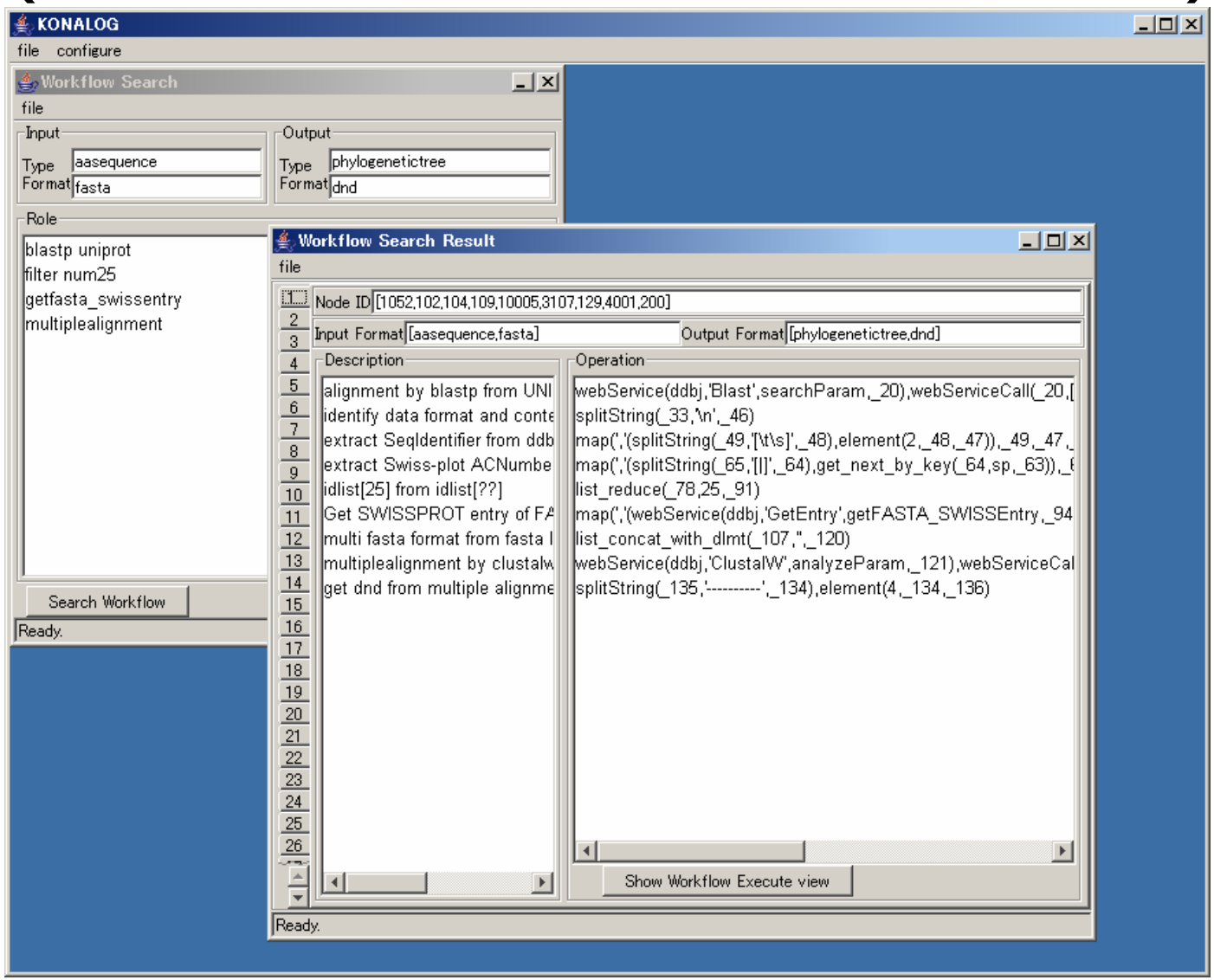
Prototype for ‘Proof of Concept’

- 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

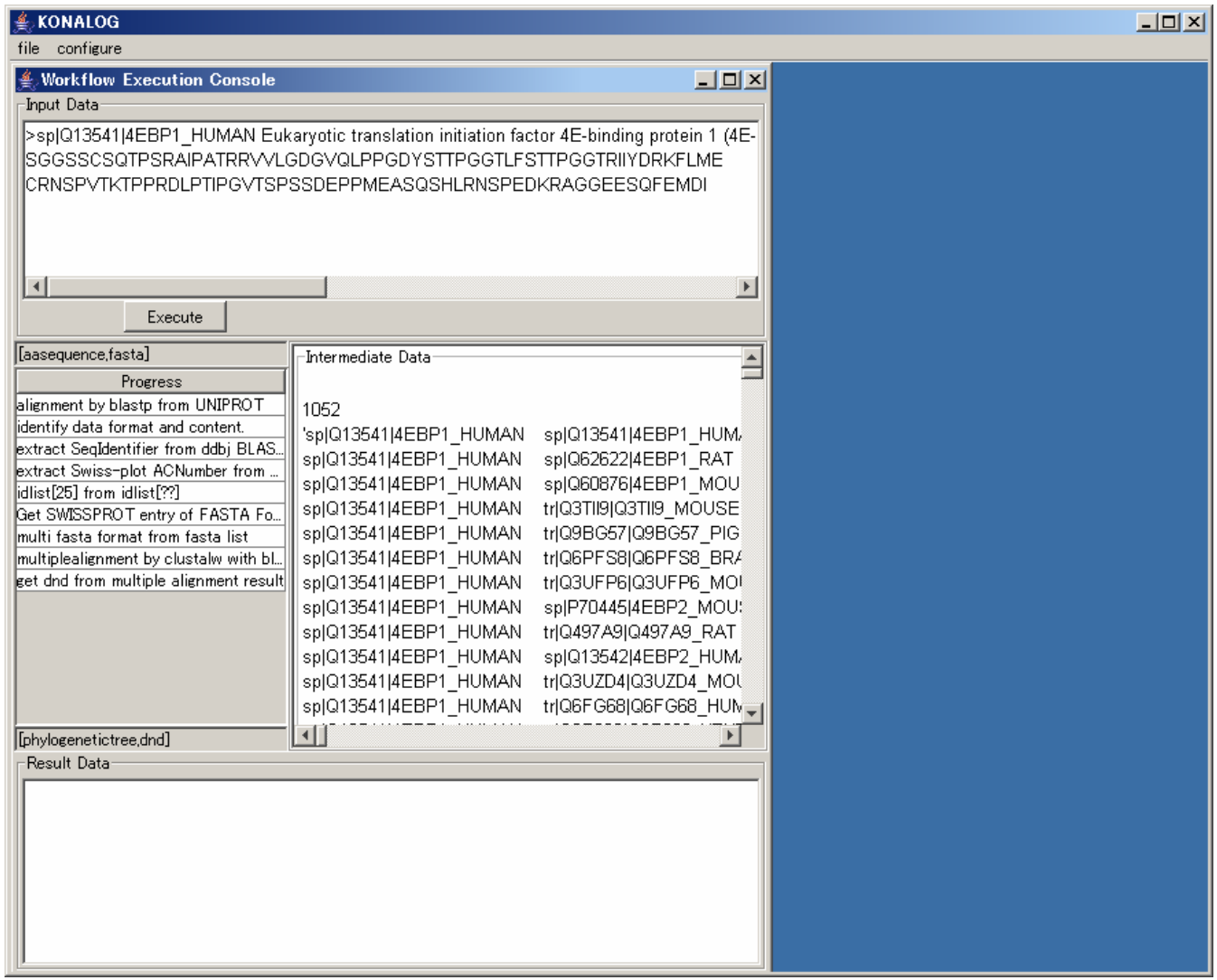
System Overview



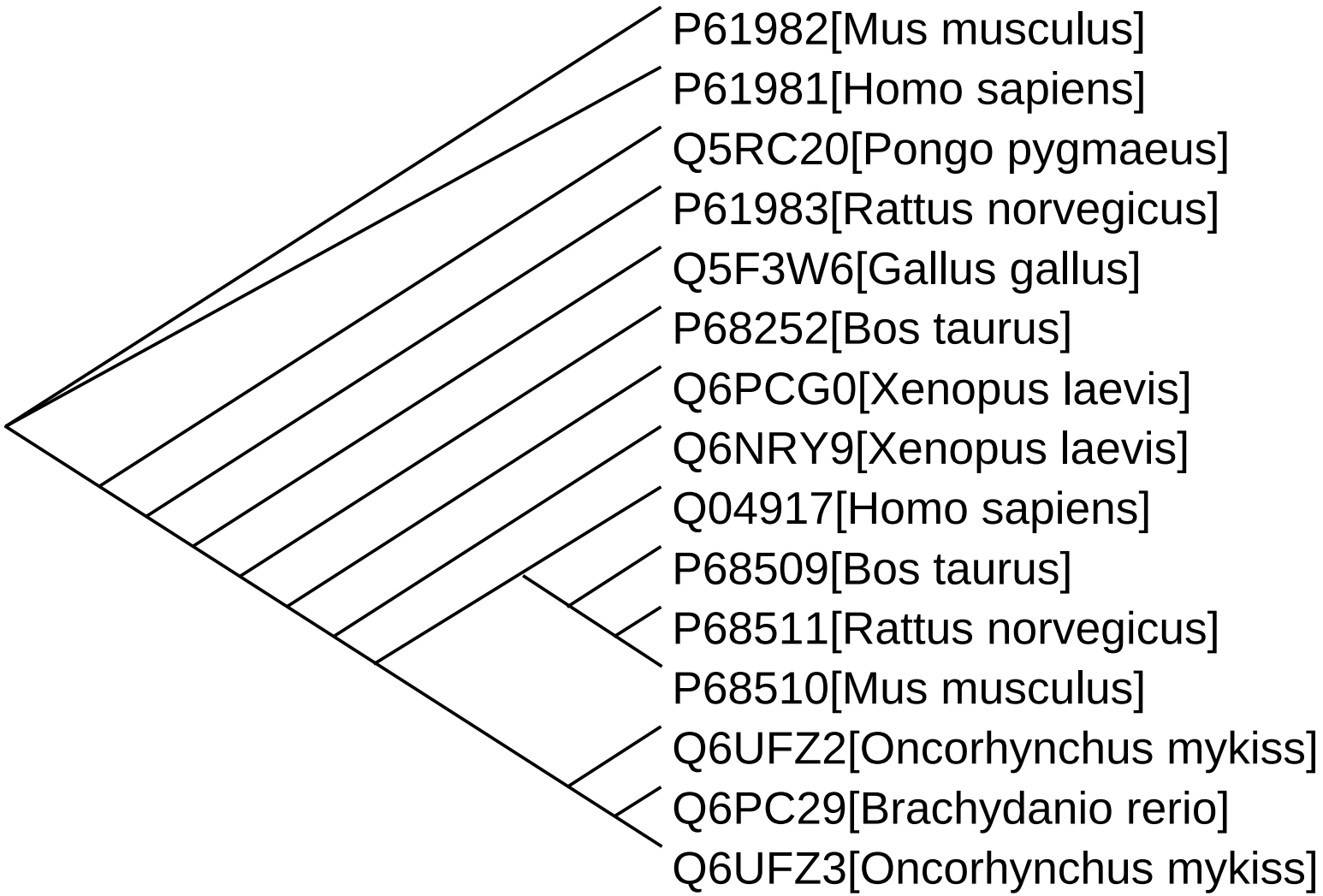
Screen Snapshot (Workflow Generation Phase)



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

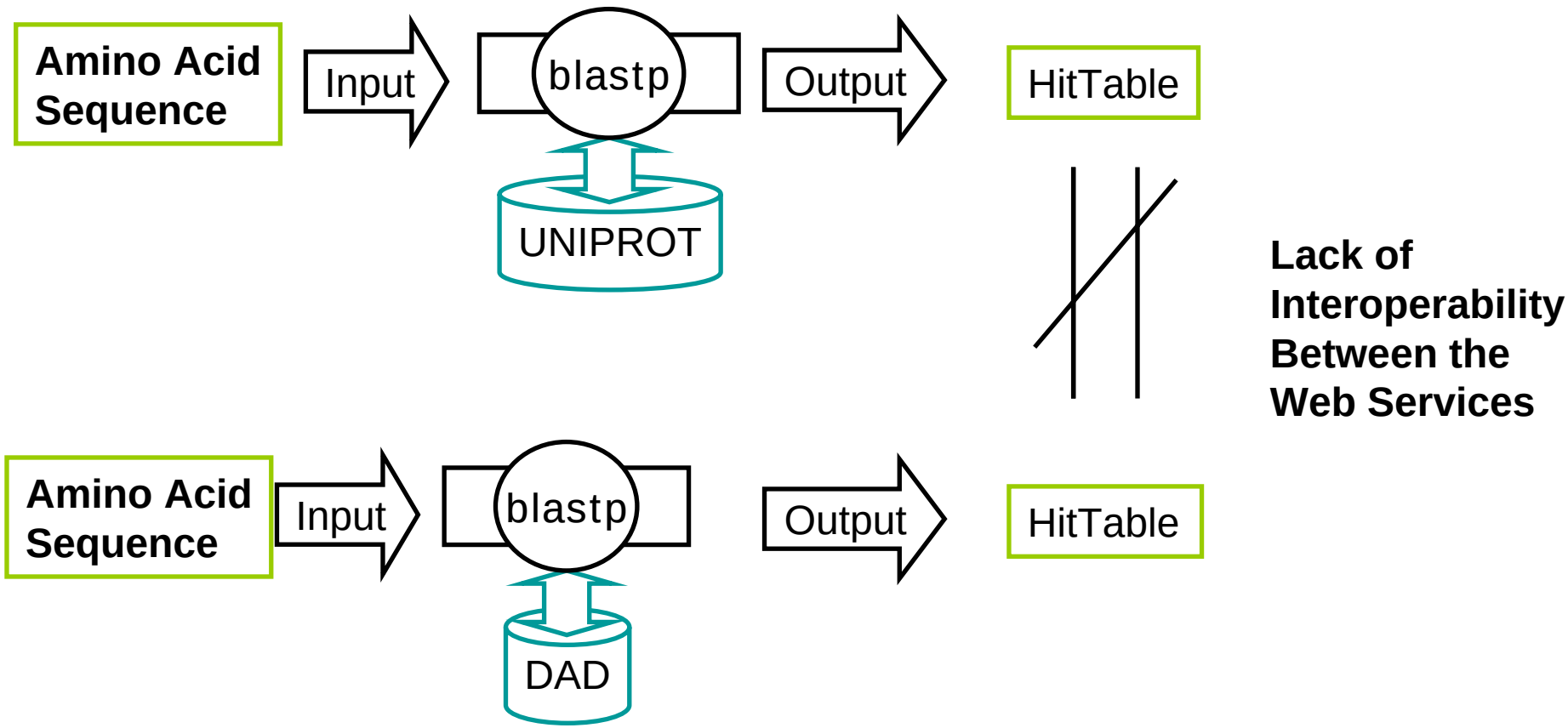
Test Set of Specification

No	Workflow				
	Input		Output		Aplications
	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

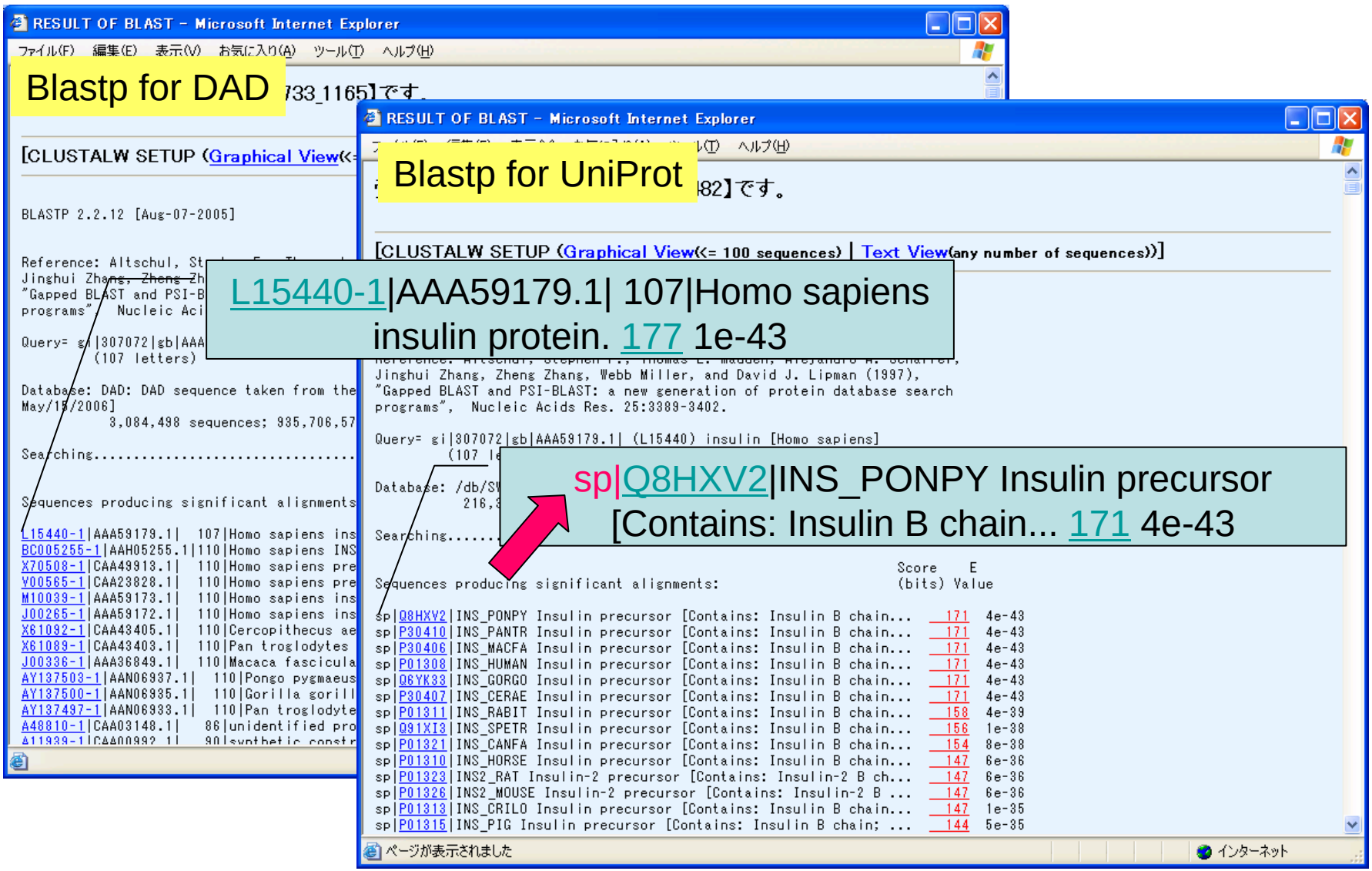
Differences of Generated Workflow

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?


Amino Acid Sequence
Input
blastp
Output
HitTable
UNIPROT
Lack of Interoperability Between the Web Services
Amino Acid Sequence
Input
blastp
Output
HitTable
DAD

Very Similar but not the Same Format



Conclusion

- Web Services have great potential to share Bioinformatics Data ant 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