

Bioinformatics Data Integration

Hideo Matsuda^{1,2},

Susumu Date¹, Shinji Shimojo^{1,2}

(1 Osaka University, 2 NAREGI)

Kentaro Wakatsuki³, Takehiro Furudate³

Gen Kawamura^{1,4}, Yoshiyuki Kido^{1,5}

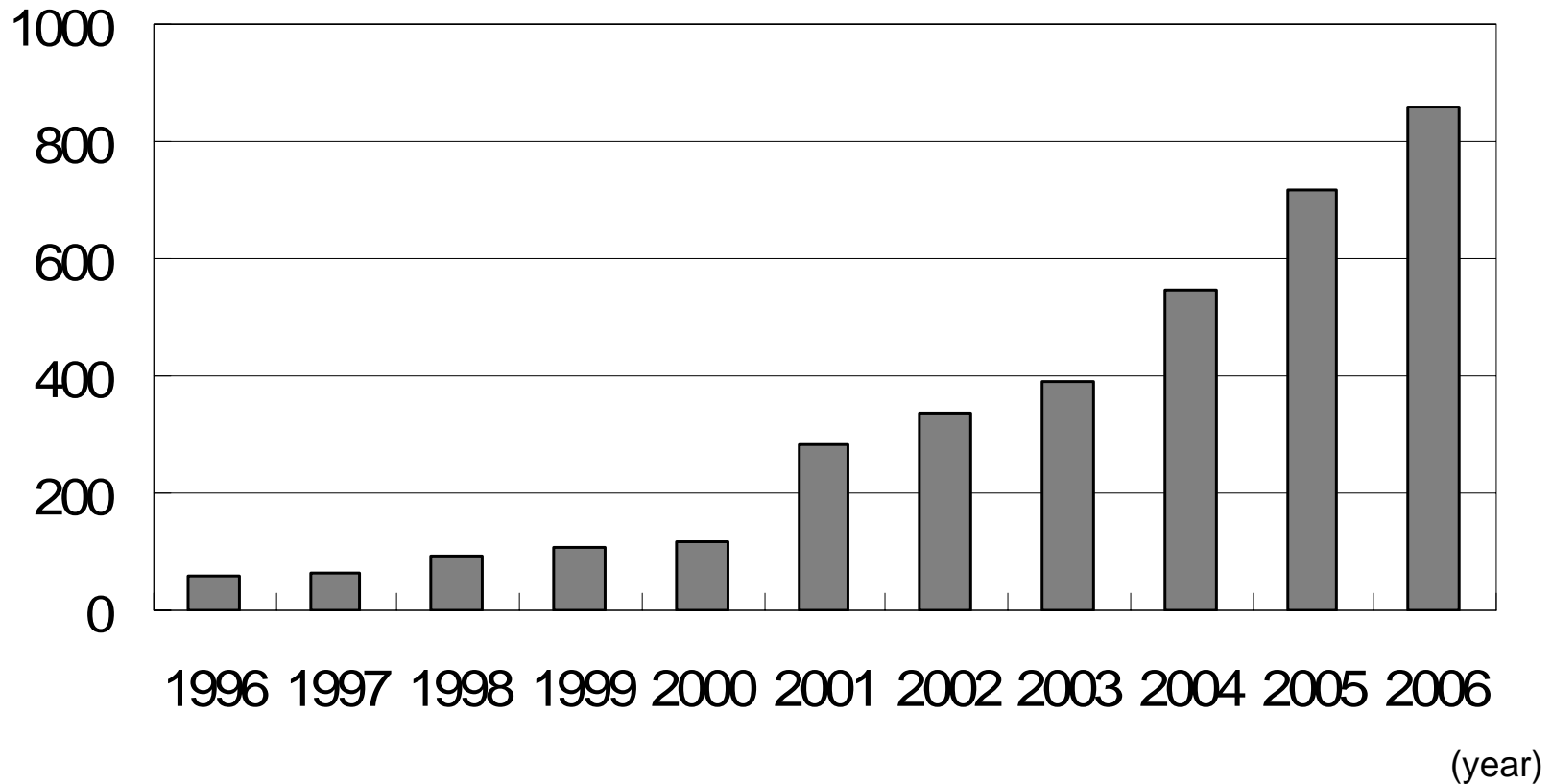
(3 Hitachi Software, 4 Aztec System, 5 Mitsui Knowledge Industry)

Biological DBs

- Not so huge amount of data compared to high-energy physics, astronomy, etc.
- Large number of DBs (> 800 DBs)
- Highly heterogeneous and complex in data description

Number of Biological DBs

- The number of biological DBs is very rapidly increasing.



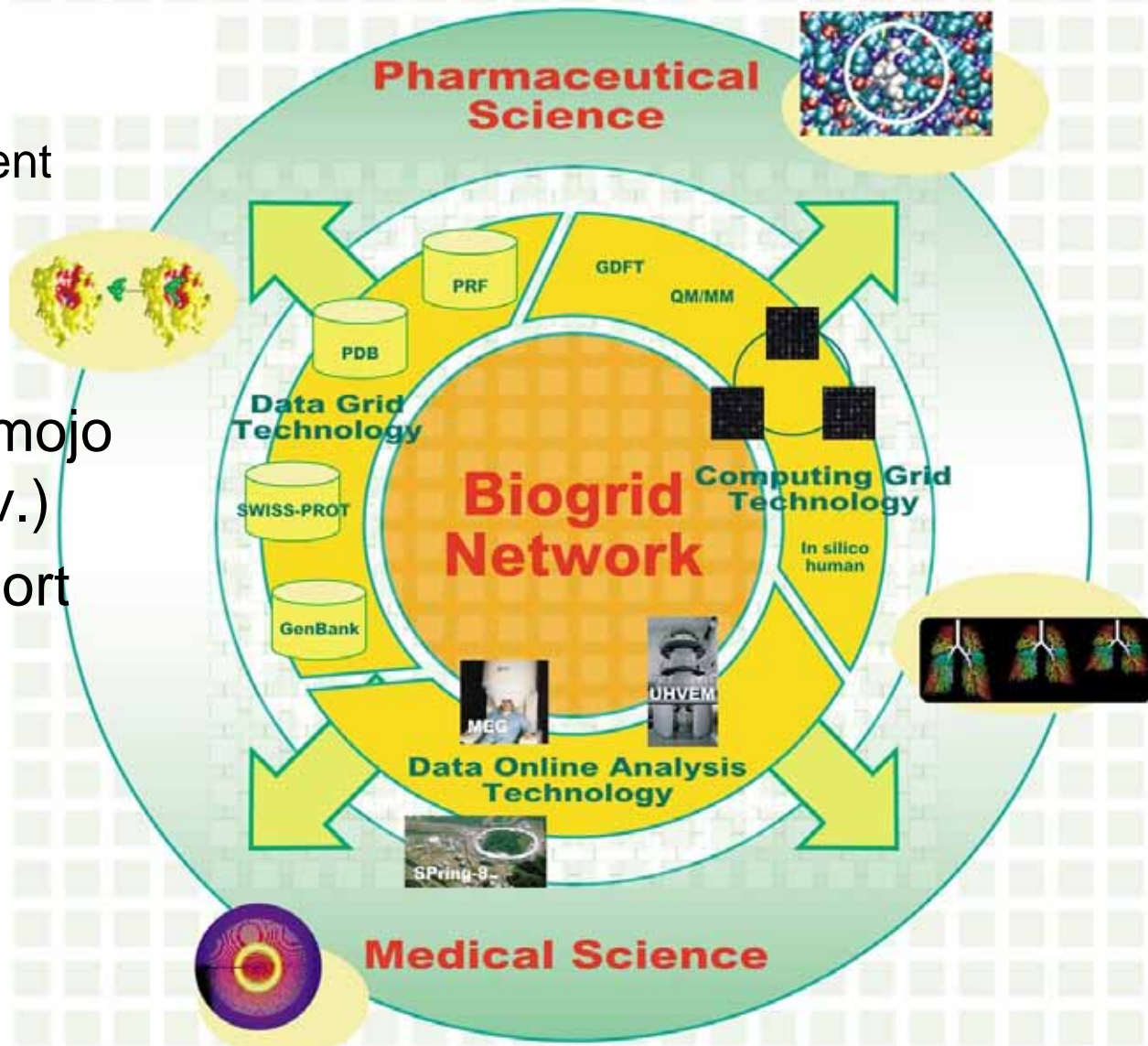
of Biological DBs in Nucleic Acids Research DB issue.

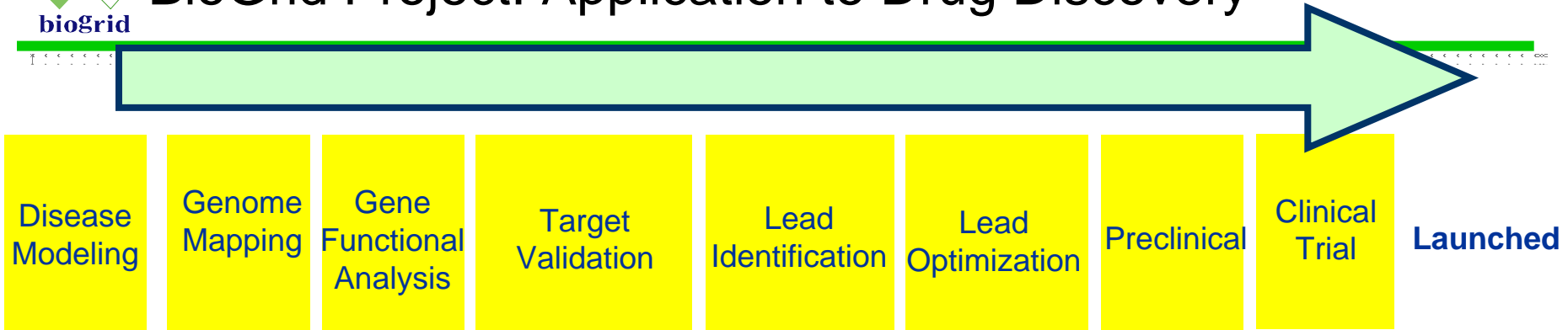
- Project: 2002 - 2006
- Goals

Technology Development
for: Pharmaceutical
(Drug Discovery) and
Medical Sciences.

- Leader: Shinji Shimojo
(CMC, Osaka Univ.)
- Government Support
(MEXT): 5years
(1~4M\$/year)
- Web site:

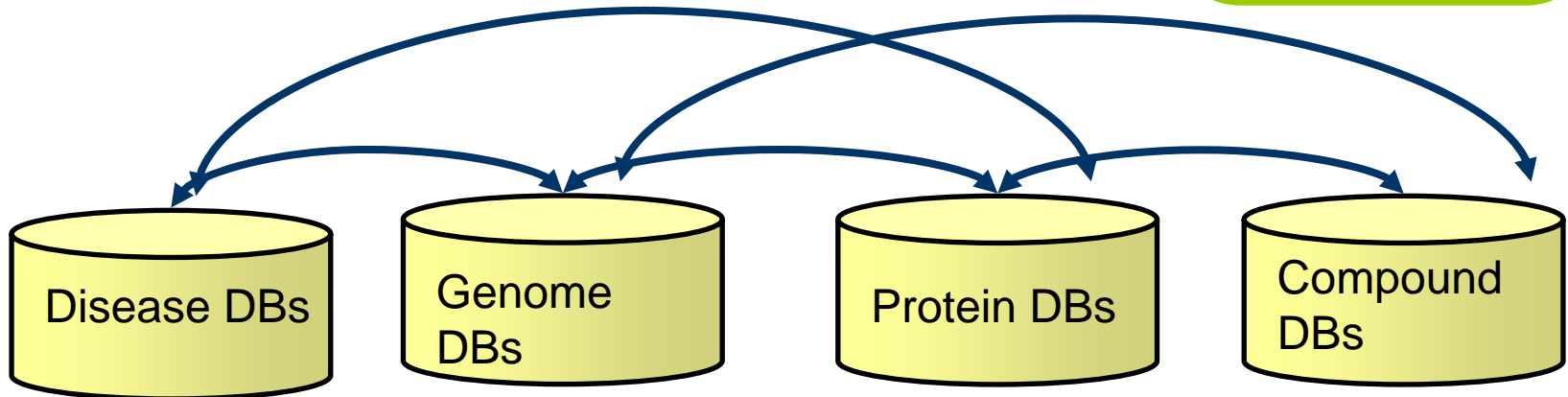
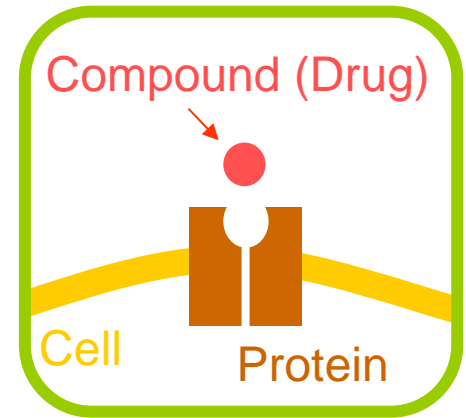
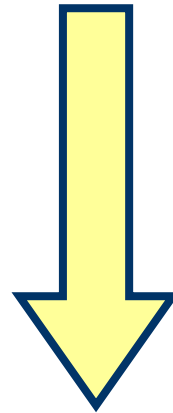
<http://www.biogrid.jp>

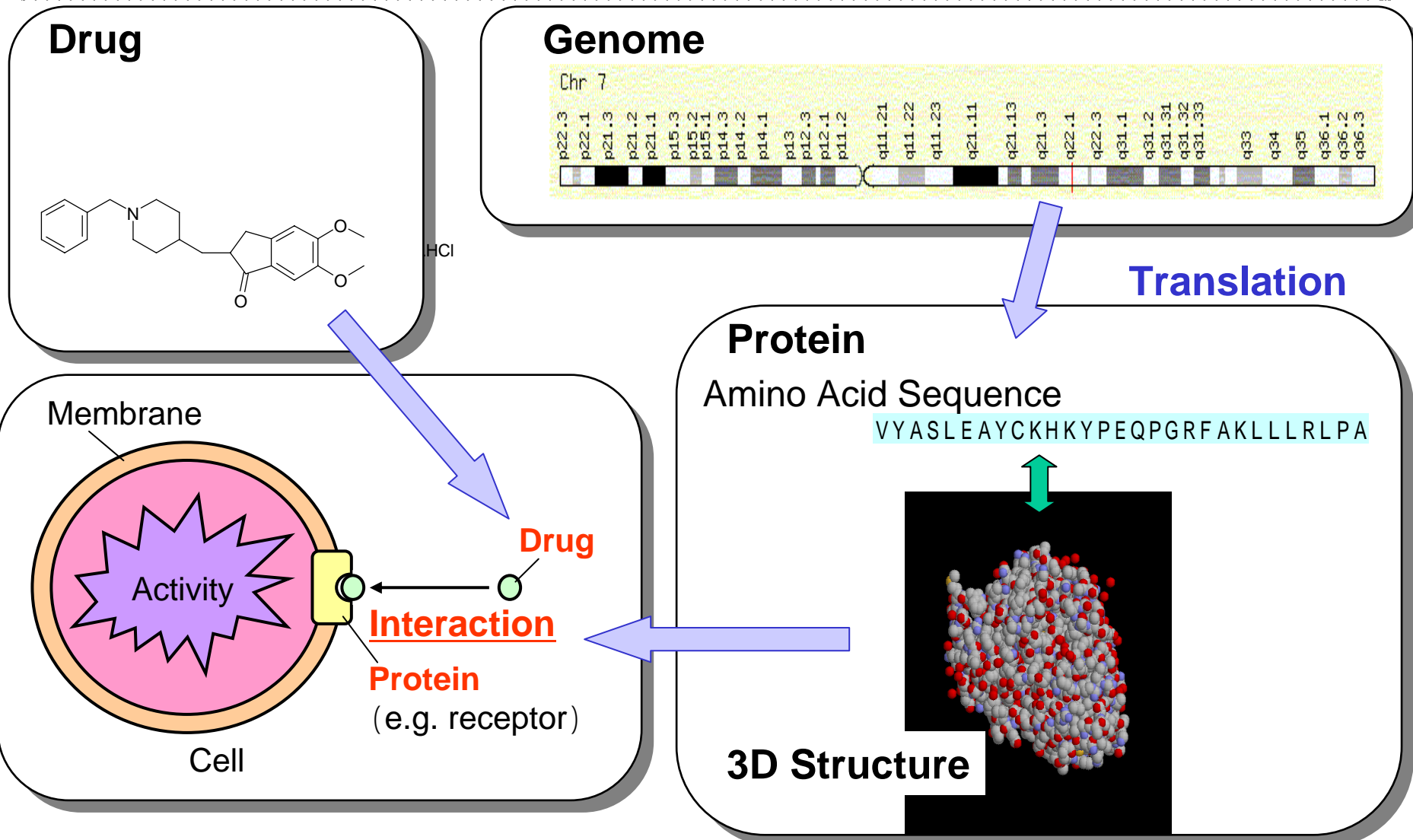




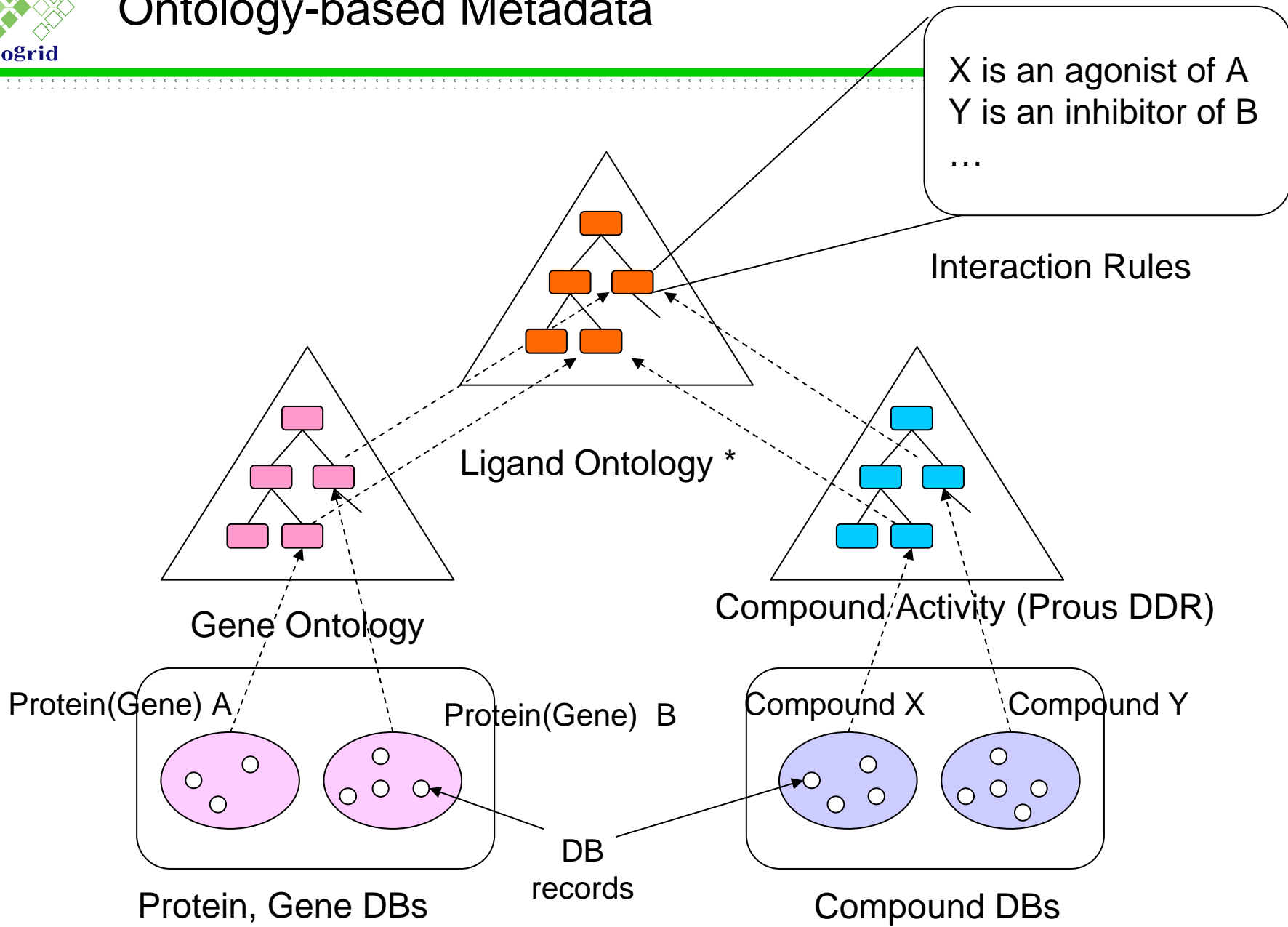
Data Grid Issues:

1. Heterogeneity of data descriptions.
2. Large number of relationships



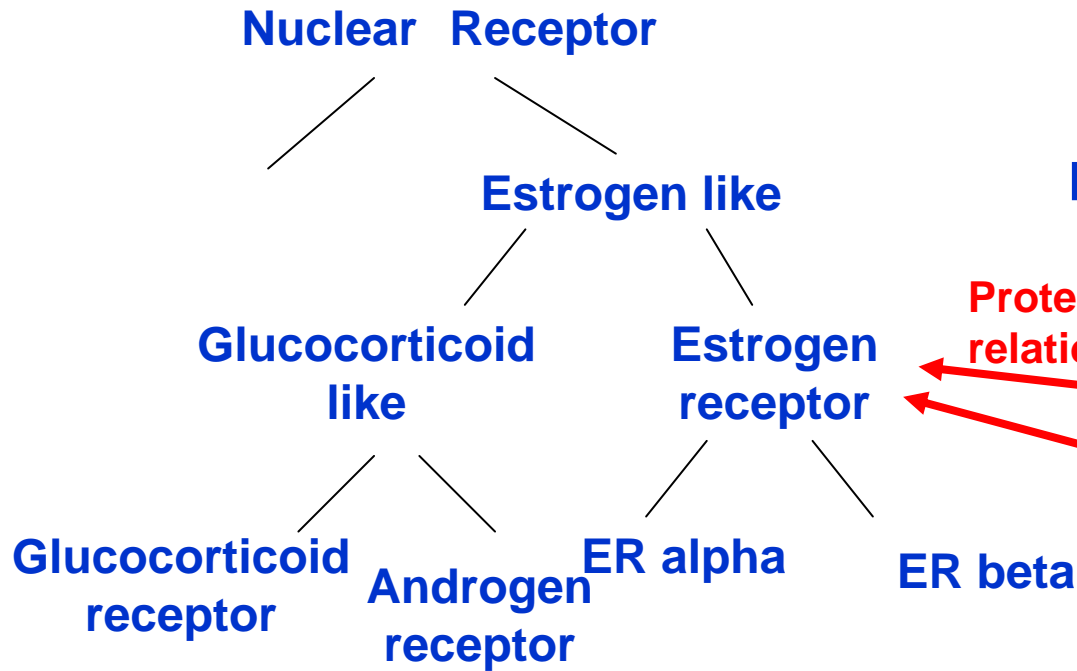


Protein-Compound Interaction Search is one of the most important technologies in drug discovery.

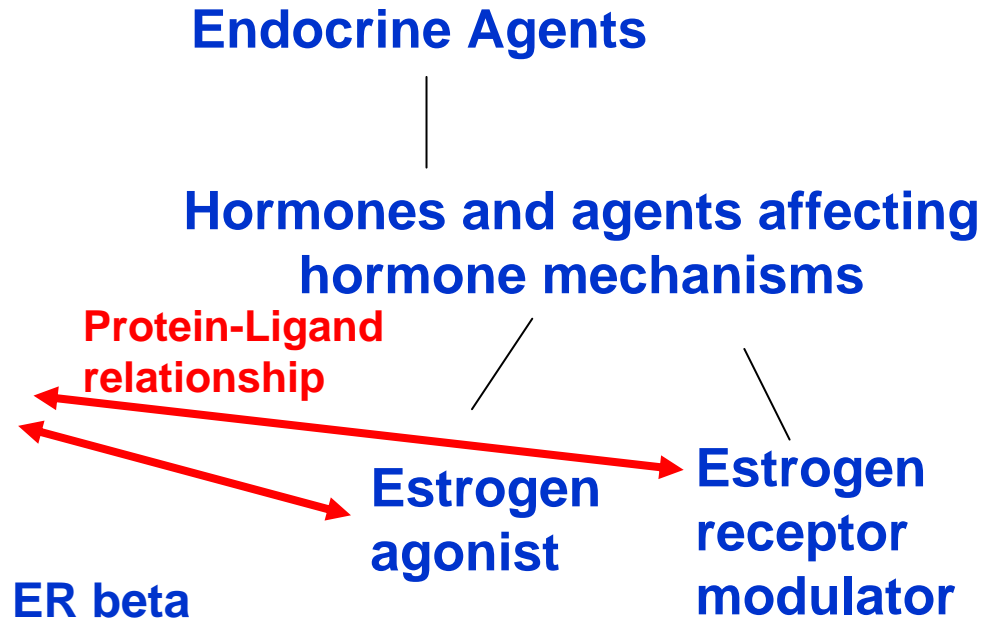


* Schuffenhauer, A., et al., J. Chem. Inf. Comp. Sci., 42, 947-955 (2002).

Protein Functional Classification (e.g., Gene Ontology)



Compound Activity Classification (e.g., Activity Class (DDR))



- Extract relationships between protein functions and compound activities by traversing hierarchical classifications (or *ontologies*)

Protein Compound Interaction View

Save Compound

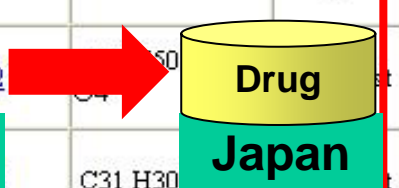
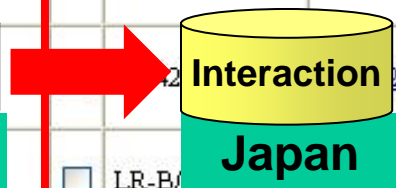
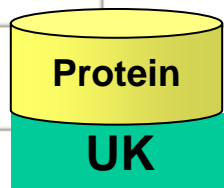
Compound Similarity

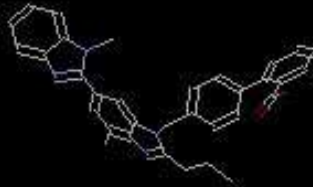
Top Compounds

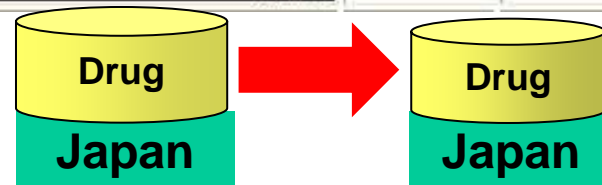
E-value

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

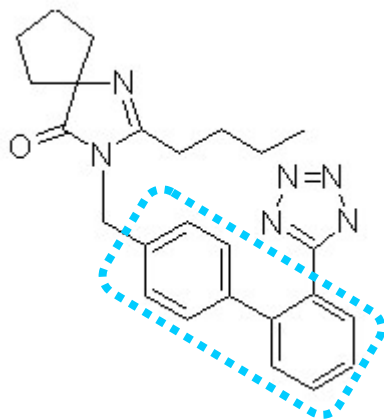
SWISS-PROT ID	SWISS-PROT Acc.	PIR	Protein Name	Homology Score	Homology Evalue	Compound	MDDR EXTREG	Molformula	Type
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> 212740	212740	C26 H25 N7 O	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> 260825	260825	C28 H30 N8 O5	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> 317215	317215	C27 H24 F3 N5 O3 S	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> FK-739	193909	C24 H22 N7 Na	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> XR-510	211727	C39 H47 F N5 O6 S K	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> UR-7198	225409	C27 H29 N3 O2	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> 212740	212740	C26 H25 N7 O	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> 260825	260825	C28 H30 N8 O5	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> 317215	317215	C27 H24 F3 N5 O3 S	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> FK-739	193909	C24 H22 N7 Na	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> XR-510	211727	C39 H47 F N5 O6 S K	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> UR-7198	225409	C27 H29 N3 O2	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	<input type="checkbox"/> LR-B...	...	C31 H30	antagonist



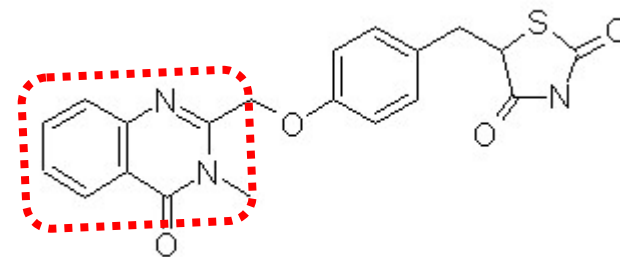
Similar Compound				Score	Inquiry Compound		
	Compound Name	Compound ID	Molformula			Compound Name	Compound ID
	Ondansetron hydrochloride	30006044	C18H24ClN3O3	0.686		Telmisartan	30005957
	ro 32-0432	10000554	C28H28N4O2	0.676		Telmisartan	30005957



Compounds possibly-interacted to the target protein

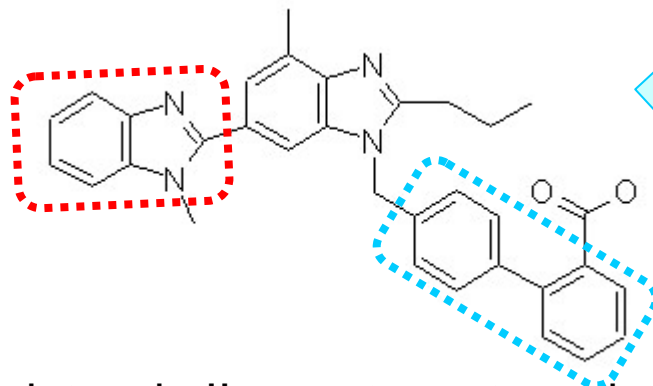
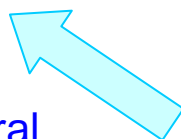


Irbesartan (angiotensin II receptor antagonists (hypertension))

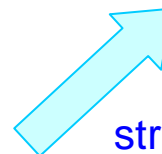


Balagritazone
(PPAR_γ agonist (diabetes))

structural
similarity



structural
similarity



Telmisartan (originally, angiotensin II receptor antagonists (hypertension); but recently peroxisome proliferator-activated receptor (PPAR)_γ partial agonist (diabetes))

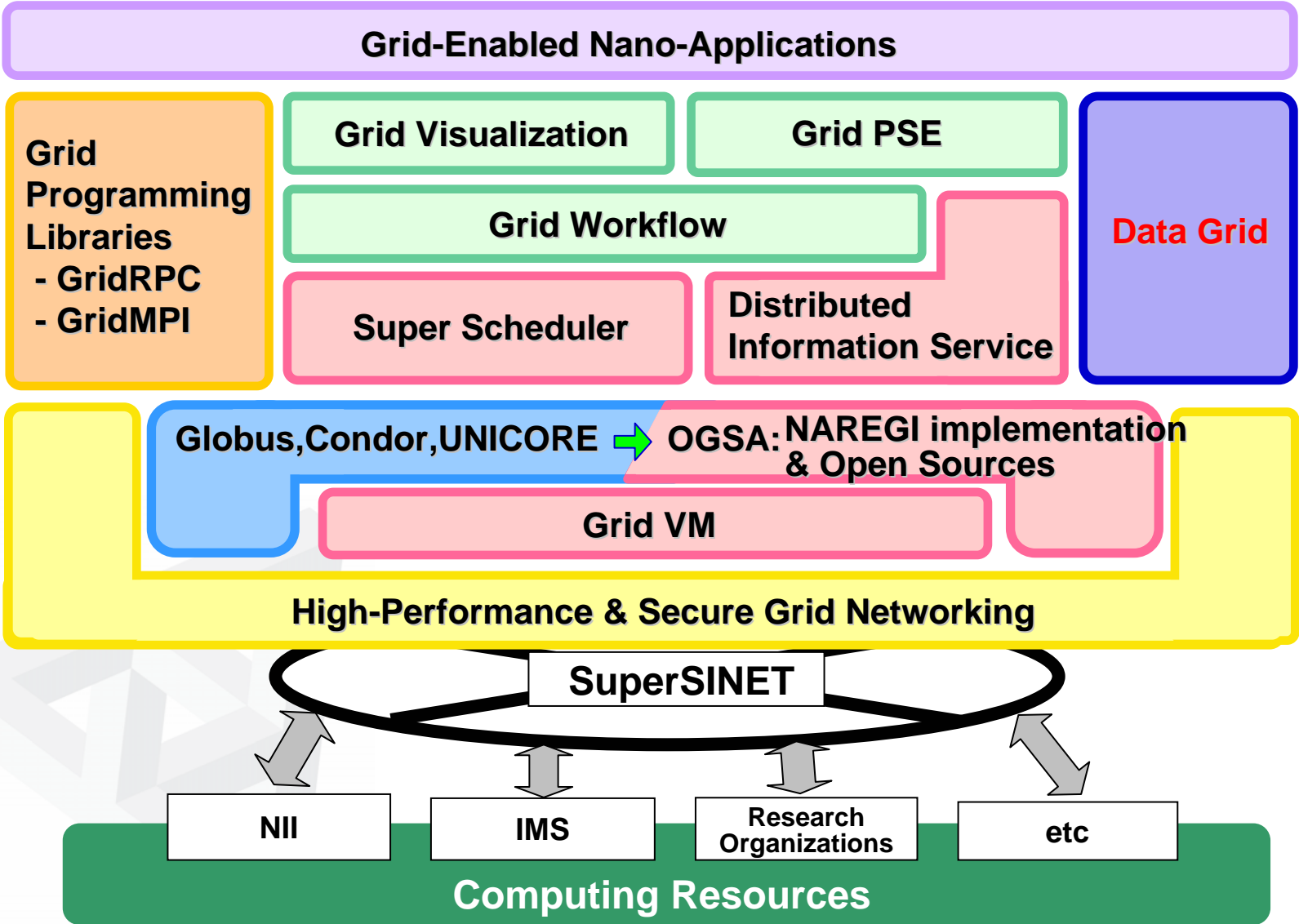
Farley, S., *Nature Reviews Drug Discovery*. 3(6), 475, 2004.

Outline of the NAREGI Project

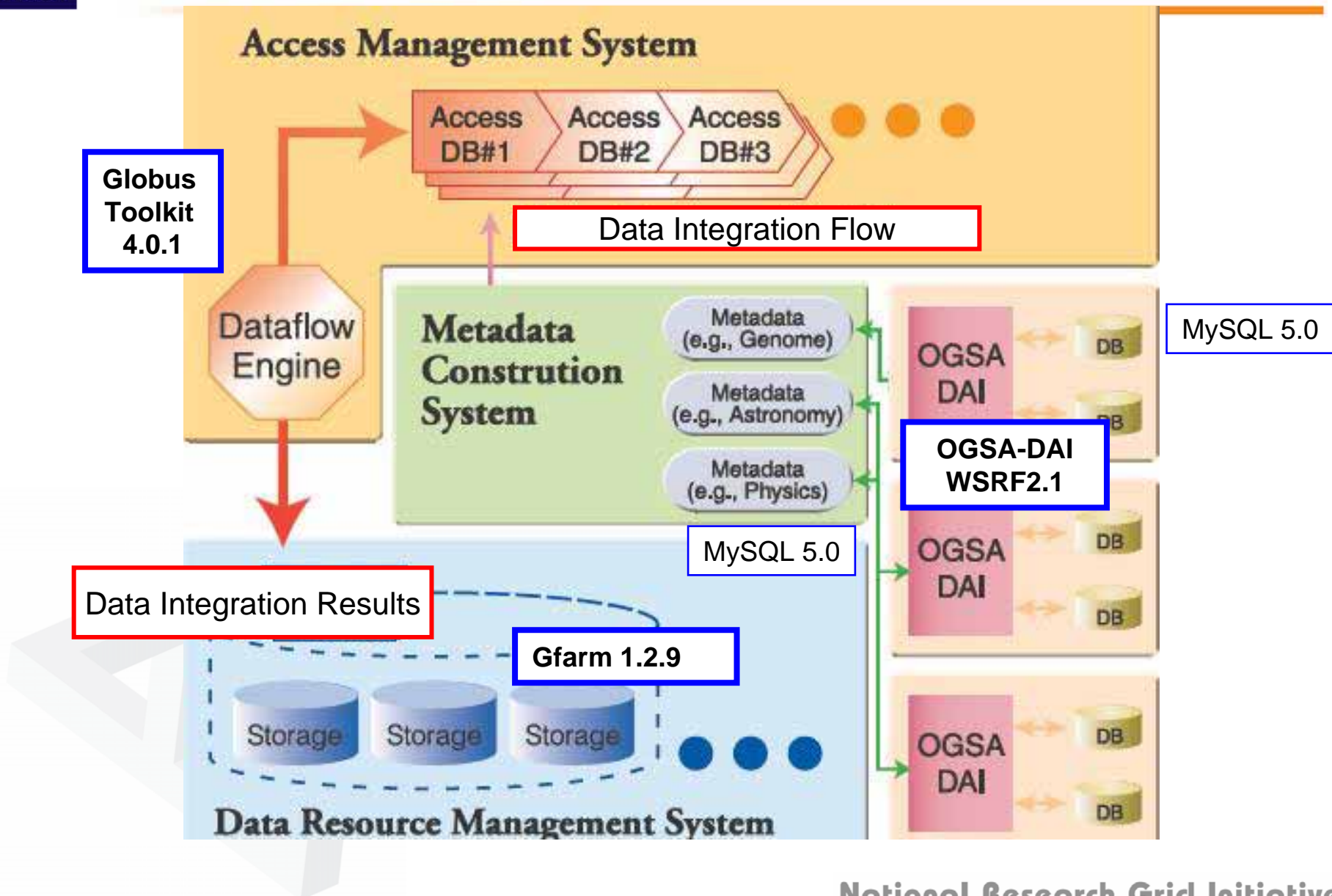
- NAREGI: National Research Grid Initiative in Japan
- Funded by Japanese Government (MEXT)
- Started from April 2003
- Two main sites:
 - R&D: NII (National Institute for Informatics)
 - Nano Science Application: IMS (Institute for Molecular Science).



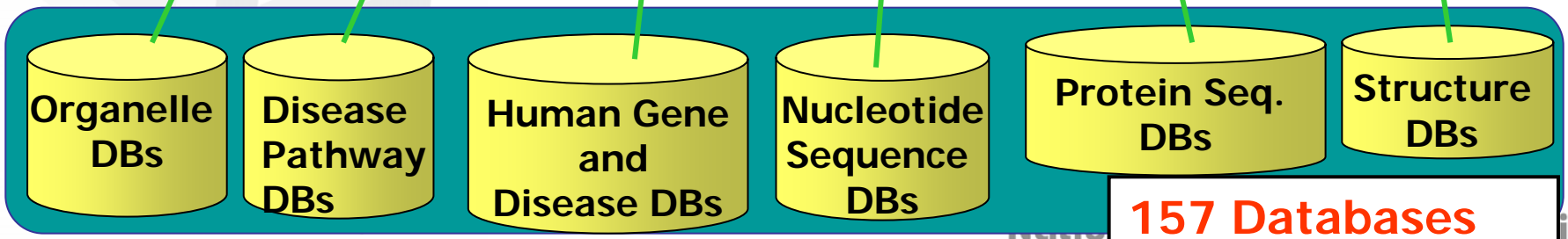
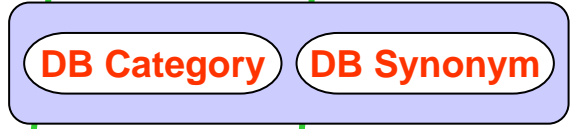
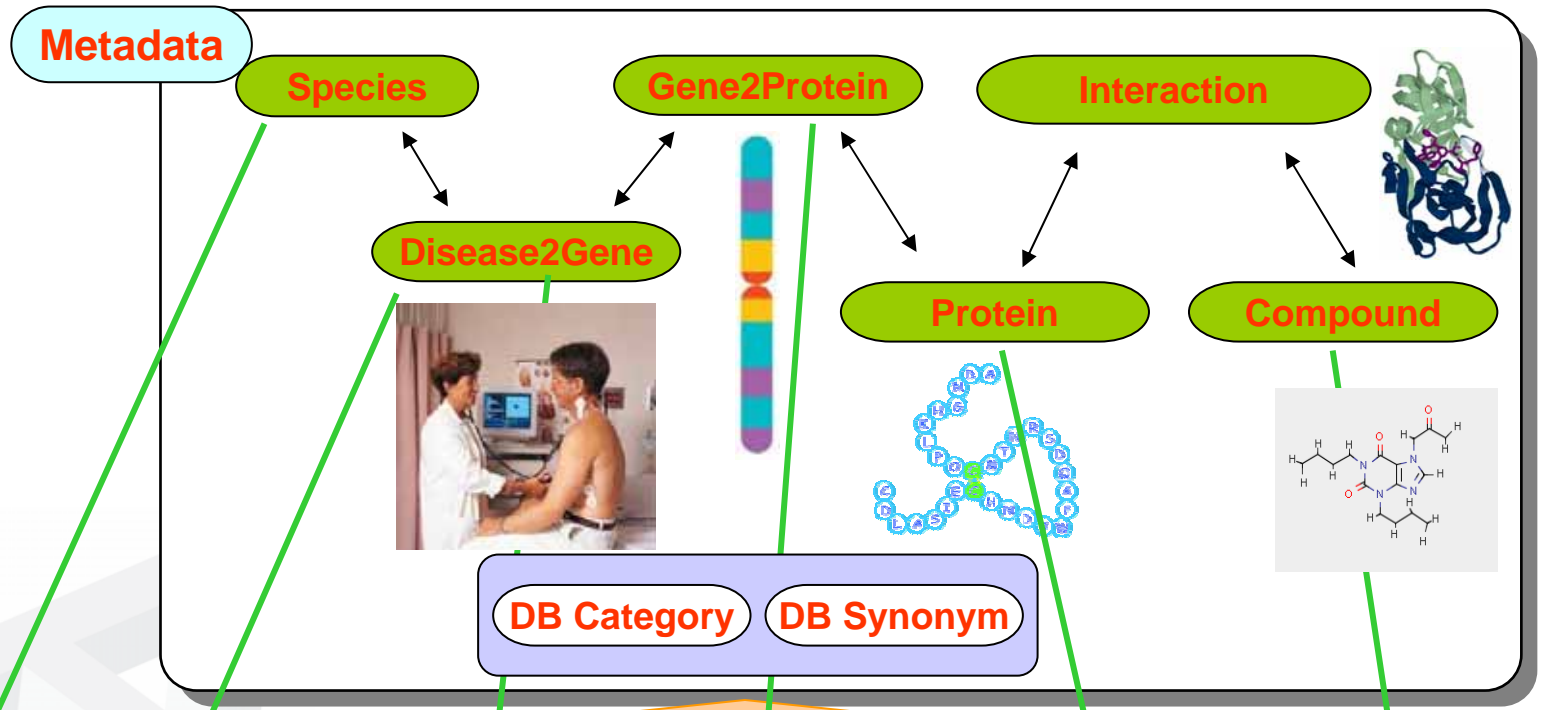
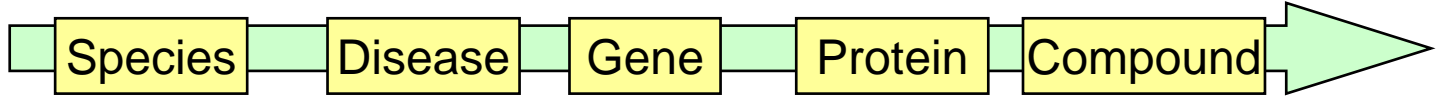
NAREGI Software Stack



NAREGI Data Integration w/ Workflow



Data Integration Workflow using Metadata

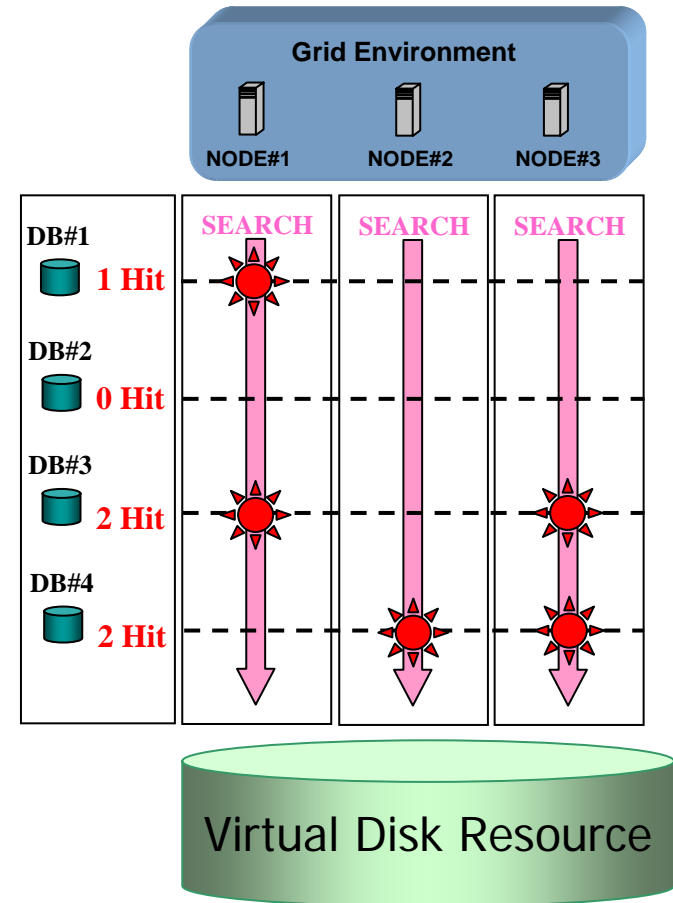


157 Databases

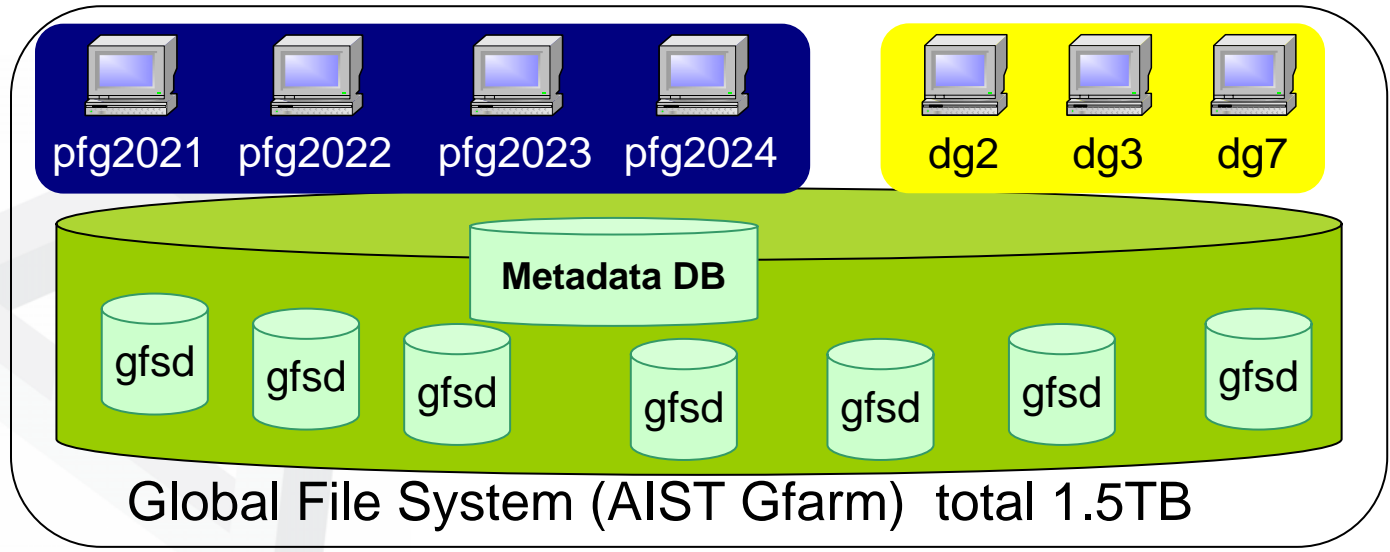
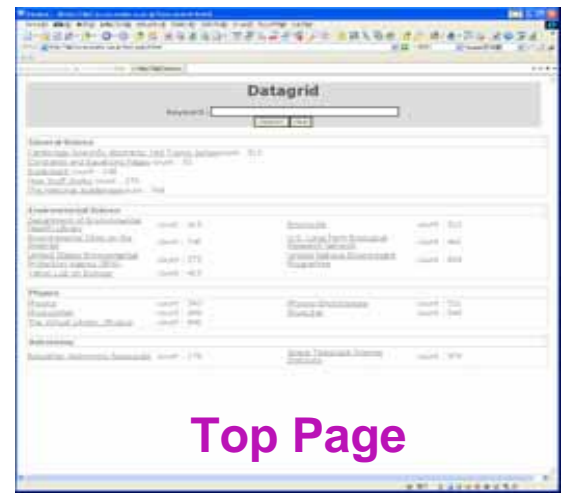
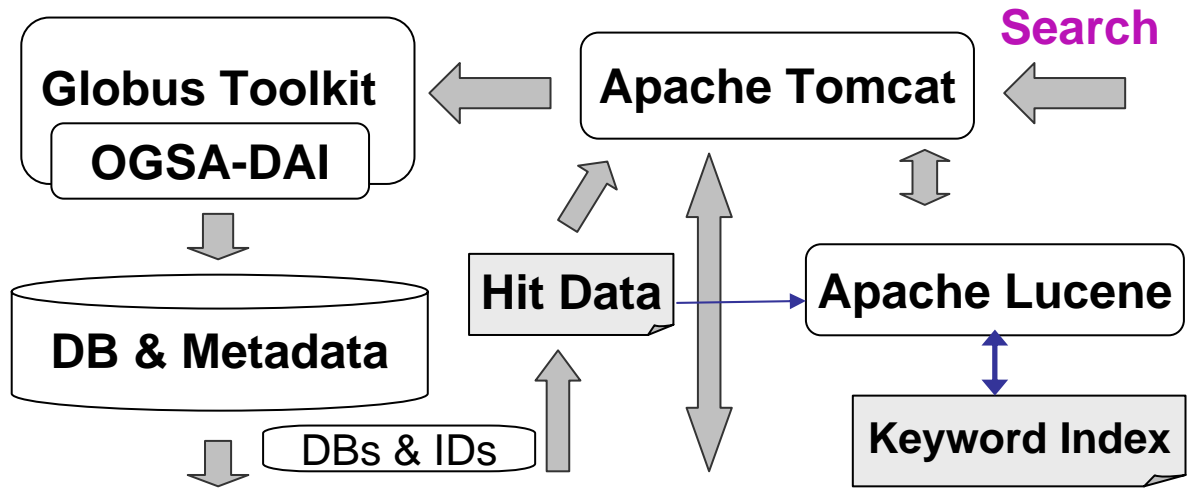
Parallel Search across DBs

- Retrieve in parallel against lifescience DBs by using Globus Toolkit, OGSA-DAI, and Gfarm.
- We extract the information of related records among different lifescience DBs and integrate their results.

Search across Databases



Data Integration System Overview



Keyword Search 1

アドレス http://dg7.ics.es.osaka-u.ac.jp:8080/dg/Top

Data Integration Flow

keyword :

Protein Sequence Databases

- [Blocks](#)
- [EMBLCONEXP](#)
- [InterPro](#)
- [PFAMA](#)
- [PFAMHMMFS](#)
- [PFAMSEED](#)
- [ProDom](#)
- [REMTREMBL](#)
- [Swiss-Prot](#)
- [UniProt](#)
- [UNIREF50](#)

- [EMBLCON](#)
- [ENSEMBLCPG](#)
- [IPRMATCHES](#) [ENSEMBL](#)
- [PFAMB](#)
- [PFAMHMMLS](#)
- [PIR-PSD](#)
- [PROSITEDOC](#)
- [SPTREMBL](#)
- [TrEMBL](#)
- [UNIREF100](#)
- [UNIREF90](#)

Human Genes And Diseases

- [GENOMEREVIEWS](#)
- [HGBASE_HAPLOTYPE](#)
- [HUMAN_MITBASE](#)
- [MUTRES](#)
- [OMIMOFFSET](#)
- [PRF](#)
- [TAXONOMY](#)

- [HGBASE](#)
- [HGBASE_SUBMITTER](#)
- [HUMUT](#)
- [OMIM](#)
- [P53LINK](#)
- [SWISSCHANGE](#)

Input Keyword

Name of Database in Our System

Protein Sequence Databases

Human Genes And Diseases

EMBLCON
ENSEMBLCPG
IPRMATCHES ENSEMBL
PFAMB
PFAMHMMLS
PIR-PSD
PROSITEDOC
SPTREMBL
TrEMBL
UNIREF100
UNIREF90

HGBASE
HGBASE SUBMITTER
HUMUT
OMIM
P53LINK
SWISSCHANGE

Keyword Search 2

アドレス http://dg7.ics.es.osaka-u.ac.jp:8080/dg/Top 移動 リンク Norton AntiVirus

Data Integration Flow

Click Category Anchor

Count of Retrieval Results

Keyword: submit clear

Protein Sequence Databases

Blocks	count : 0
EMBLCONEXP	count : 0
InterPro	count : 0
PFAMA	count : 0
PFAMHMMFS	count : 0
PFAMSEED	count : 0
ProDom	count : 0
REMTREMBL	count : 0
Swiss-Prot	count : 43
UniProt	count : 0
UNIREF50	count : 0

EMBLCON	count : 0	
ENSEMBLCPG	count : 0	
IPRMATCHES	ENSEMBL	count : 0
PFAMB	count : 0	
PFAMHMMLS	count : 0	
PIR-PSD	count : 0	
PROSITEDOC	count : 0	
SPTREMBL	count : 0	
TrEMBL	count : 0	
UNIREF100	count : 0	
UNIREF90	count : 0	

Human Genes And Diseases

GENOMEREVIEWS	count : 0
HGBASE_HAPLOTYPE	count : 0
HUMAN_MITBASE	count : 0
MUTRES	count : 0
OMIMOFFSET	count : 0
PRF	count : 1
TAXONOMY	count : 0

HGBASE	count : 0
HGBASE_SUBMITTER	count : 0
HUMUT	count : 0
OMIM	count : 392
P53LINK	count : 0
SWISSCHANGE	count : 0

Data Integration Flow 1

Select dataflow template

The screenshot shows a web browser window with the title "Data Integration Flow". Below the title are "Fire" and "Back" buttons. The main content area displays three dataflow templates, each consisting of a sequence of database names connected by double arrows (>>>). The first template is highlighted with a pink rounded rectangle.

Template	Database 1	Database 2	Database 3	Database 4
1 (Highlighted)	Protein Sequence Databases	Human Genes And Diseases	Structure Databases	
2	Protein Sequence Databases	RNA Sequence Databases	Microarray Data And Other Gene Expression Database	
3	Protein Sequence Databases	Immunological Databases	Human Genes And Diseases	Proteo Resou

Data Integration Flow 2

アドレス http://dg7.ics.es.osaka-u.ac.jp:8080/dg/Dataflow

移動 リンク Norton AntiVirus

template

Protein Sequence Databases

Human Genes And Diseases

Structure Databases

1

*	18 databases/ 78 entries					
	1	1	1	1	1	1
	1	2	2	2	2	2
	2	2	2	2	2	2
	2	2	3	4	4	4
	4	4	4	4	4	5
	6	7	7	8	8	8
	8	8	8	8	8	8
	8	8	8	8	8	8
	8	8	9	9	10	11
	11	12	13	14	15	16
	16	16	16	16	16	16
	16	16	16	16	16	16
	16	16	16	16	17	18
*	17 databases/ 66 entries					
	*	*	*	*	*	*

Icon color shows status of process

- :running
- :finished
- :failed

Summary and Future Works

- We have developed a system for data integration of >40 lifescience DBs.
- Globus Toolkit and OGSA-DAI integrates distributed DBs and metadata.
- DB keywords and indices are stored in a grid filesystem (Gfarm).

Future Works

- Most of data are currently downloaded. Need to develop web-service (WSDL/SOAP) interfaces.