Development of a Database System for Drug Discovery by Employing Grid Technology

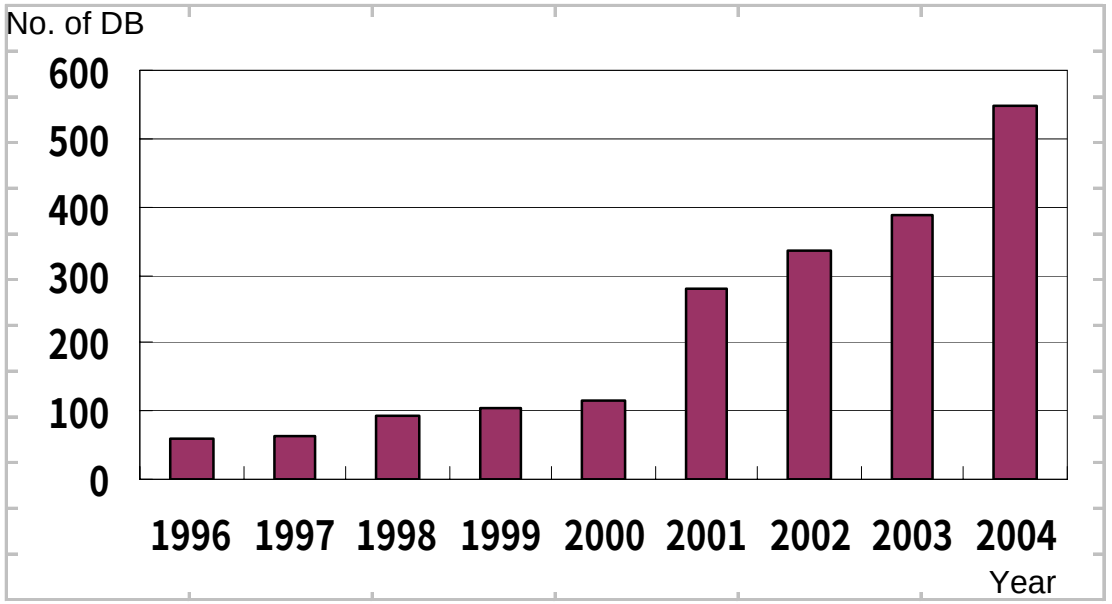
July 21,2004

Masato Kitajima1,2 Yukako Tohsato 1, Takahiro Kosaka 1,
Kazuto Yamazaki 3,Reiji Teramoto 3, Susumu Date 1, Shinji Shimojo 4, Hideo Matsuda 1

1 Graduate School of Information Science and Technology, Osaka University.
2 Fujitsu Kyushu System Engineering Limited.
3 Research Division, Sumitomo Pharmaceuticals Co., Ltd.
4 Cybermedia Center, Osaka University.

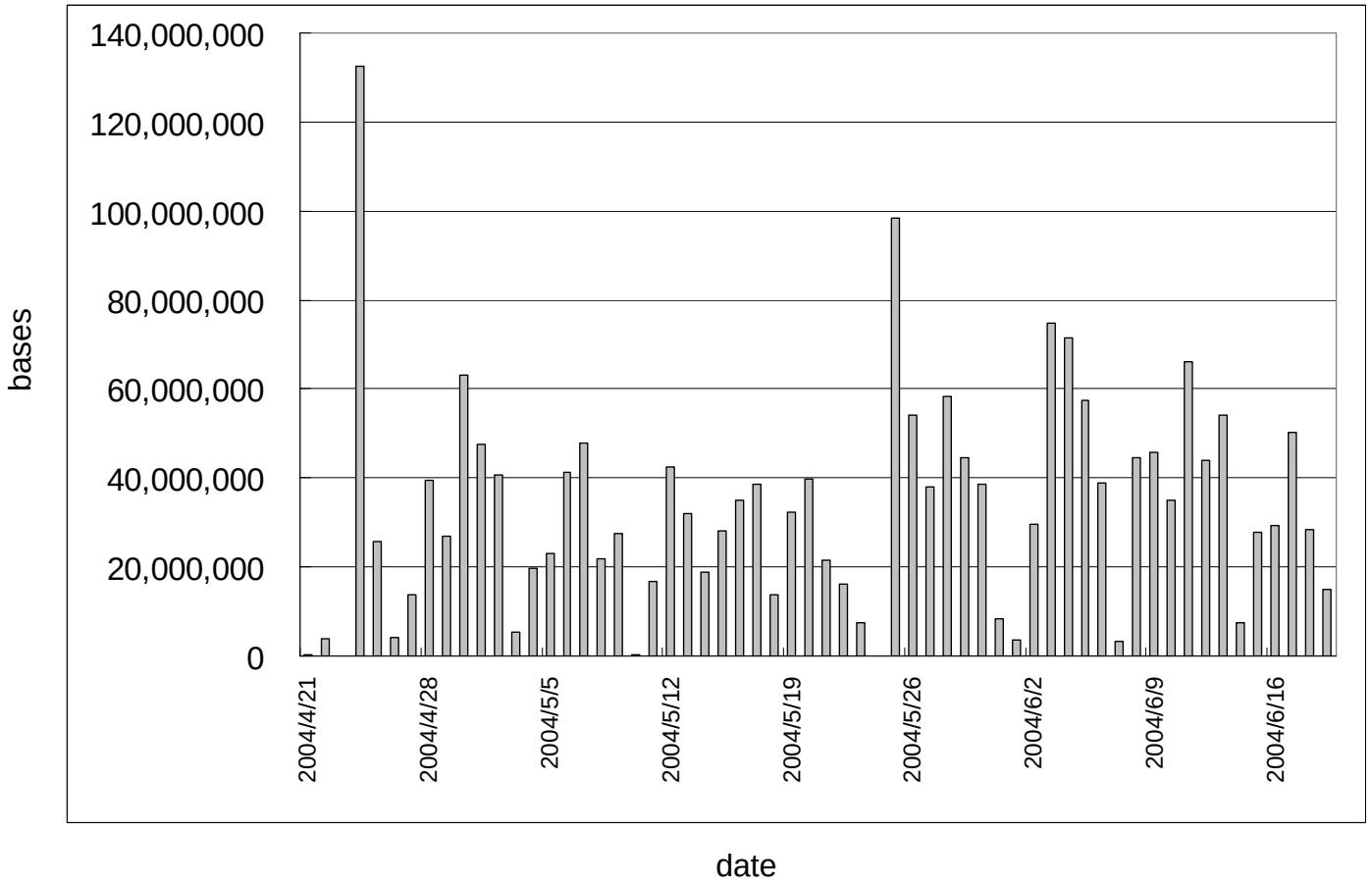
Databases in the Life Sciences

The amount of data and the number of databases in life science have dramatically increased in just a few years



Nucleic Acids Research DB Issue

Amount of updates in two months of a DNA database



Common Database Problems in the Life Sciences

- Increase in the amount of data puts a great load to the administrator who updates the database
- A slight change in the schema of one of the databases requires a complete rebuild of the whole system
- A considerable amount of time and resources wasted in just updating the database

Different Ways Of Integrating Distributed Databases

- **Hyperlinked Database**
  - **Most commonly used for linking databases**
  - **Hyperlinks cannot carry special meanings**

- **Integrated Database (ex. NCBI's Entrez)**
  - **User only needs to access a single database**
  - **Changes in the schema of one database will prompt the rebuilding of the whole database system**

- **Heterogeneous Database (ex. Stanford Univ.'s TSIMMIS)**
  - **Builds a "wrapper" on each of the databases to be accessed by a mediator (Changes in the schema of one database, only requires a change in the wrapper for that database)**
  - **Databases that use authentications and functionalities specific to life sciences(like homology searching and similarity searching) pose a problem in integration**

Common Problems in Linking the Databases

- **Unorganized structure of information**

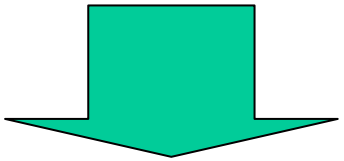
- **Data in unformatted text**

- **Inconsistent use of terms on different databases**

- **Building of relationships between the databases could only be done manually**

Proposal of a New Database System

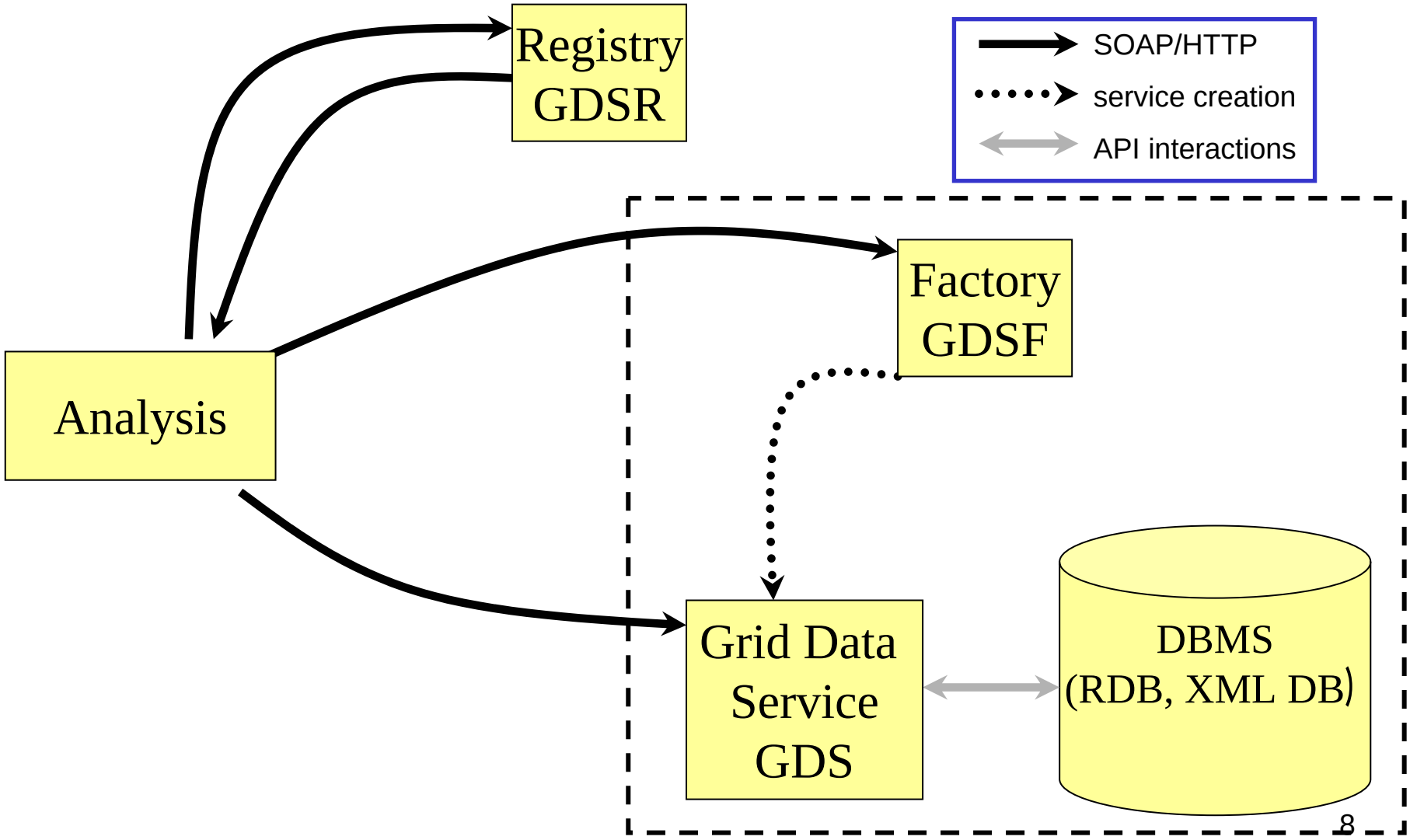
Use of grid technology
and
Introduction of the concept of metadata



Greatly helped in building mutual data relationships between databases in a distributed system

Overview of OGSA-DAI

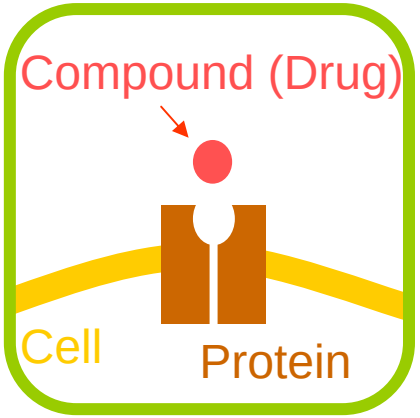
OGSA (Open Grid Service Architecture Data Access and Integration)



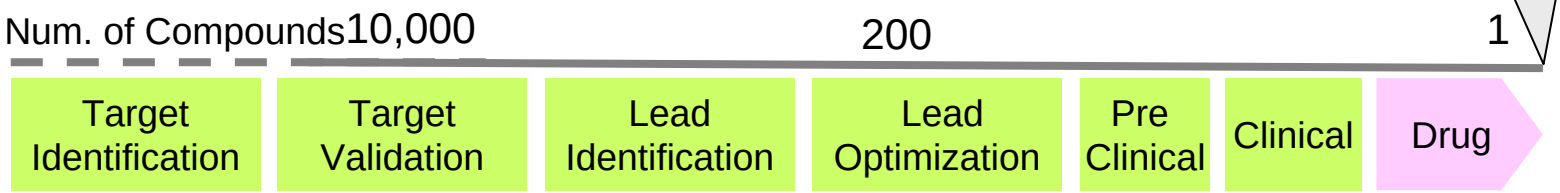
Genome-based Drug Discovery Process

Application to the **drug discovery process**

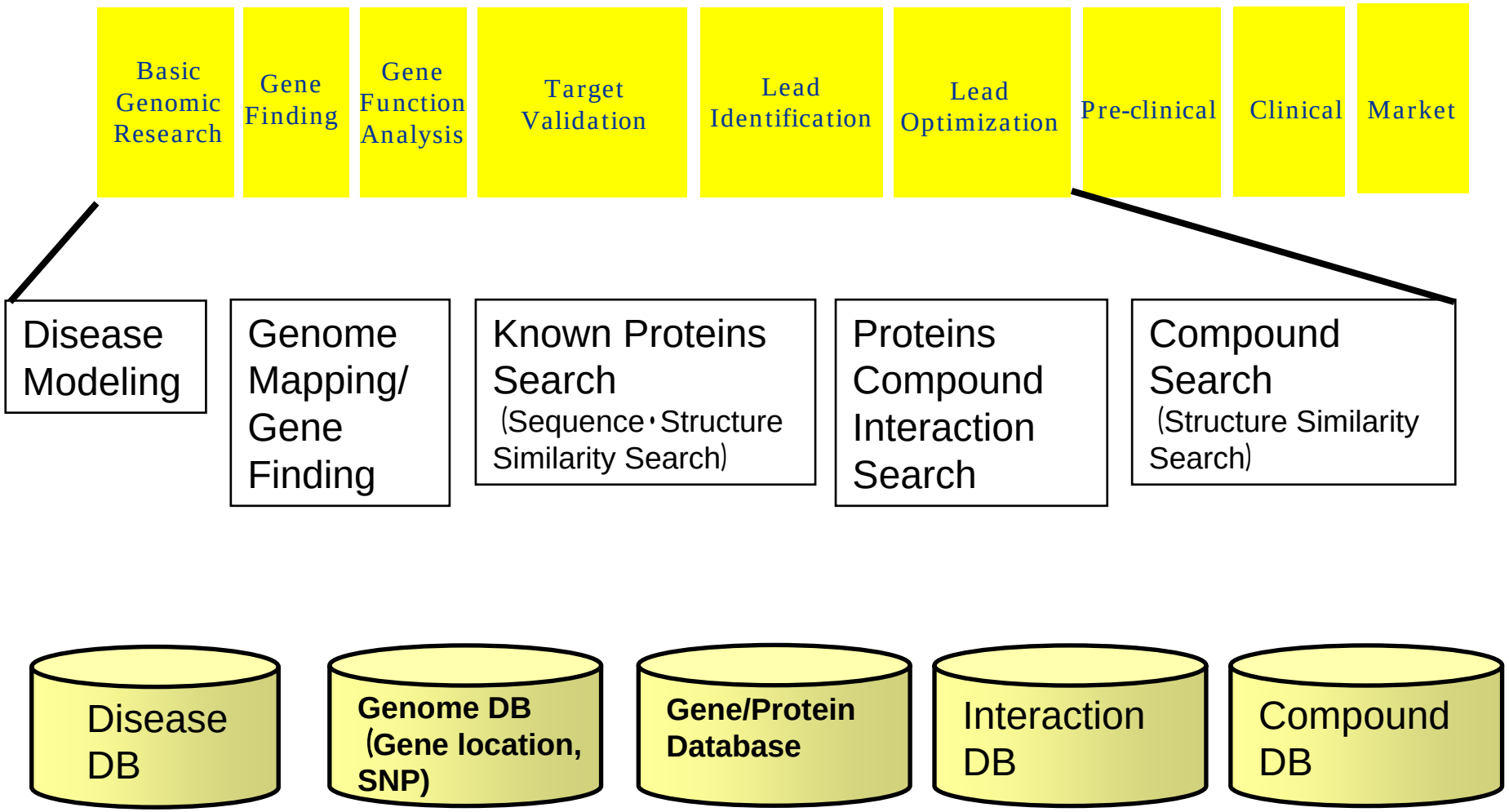
- Compounds (drugs) are activated by binding to proteins in a cell.
- **Drug Discovery Process** is to find chemical compounds that have good effects on their target proteins.
- The process is time-consuming and expensive.



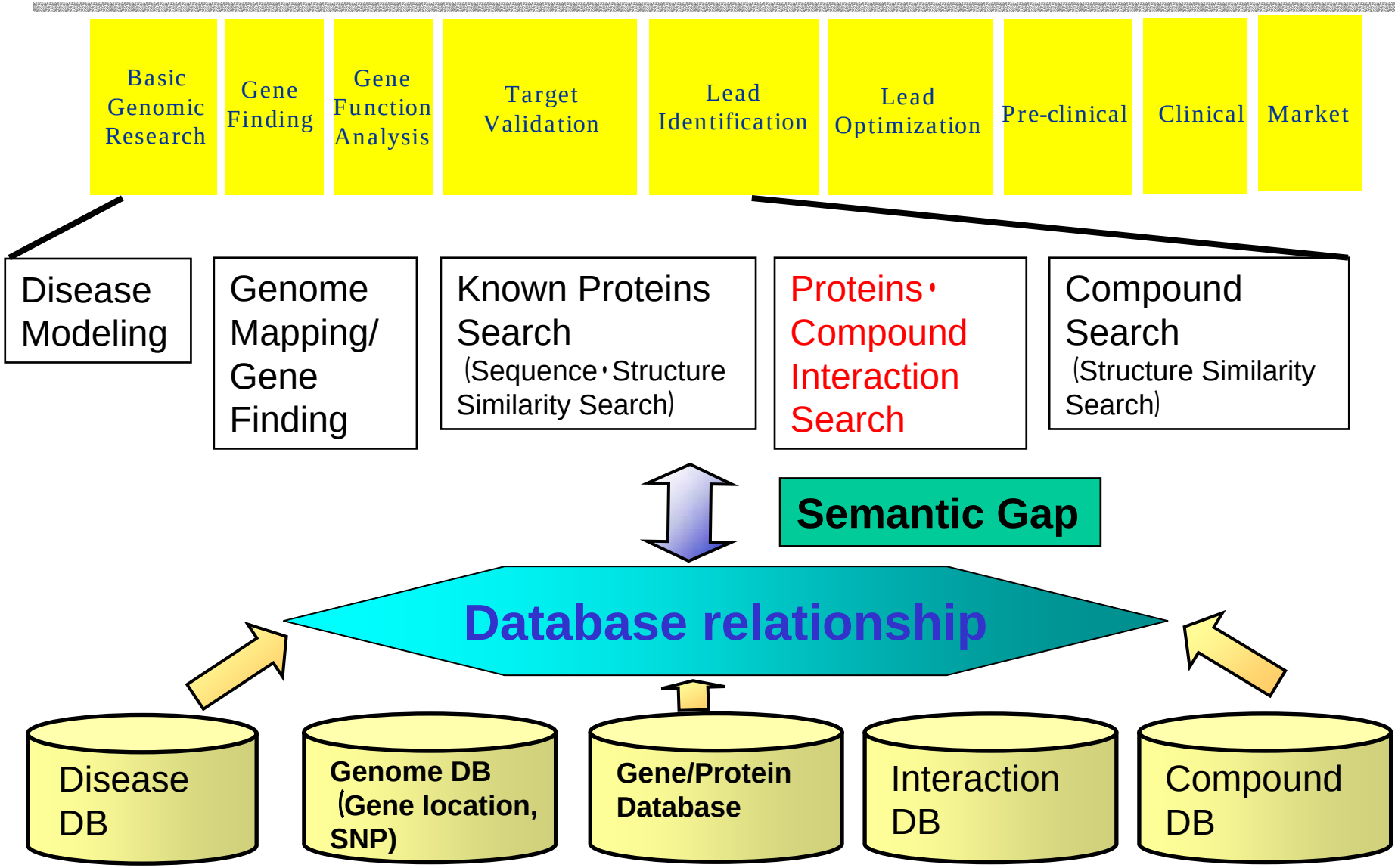
5~10 million $
(10~15 years)



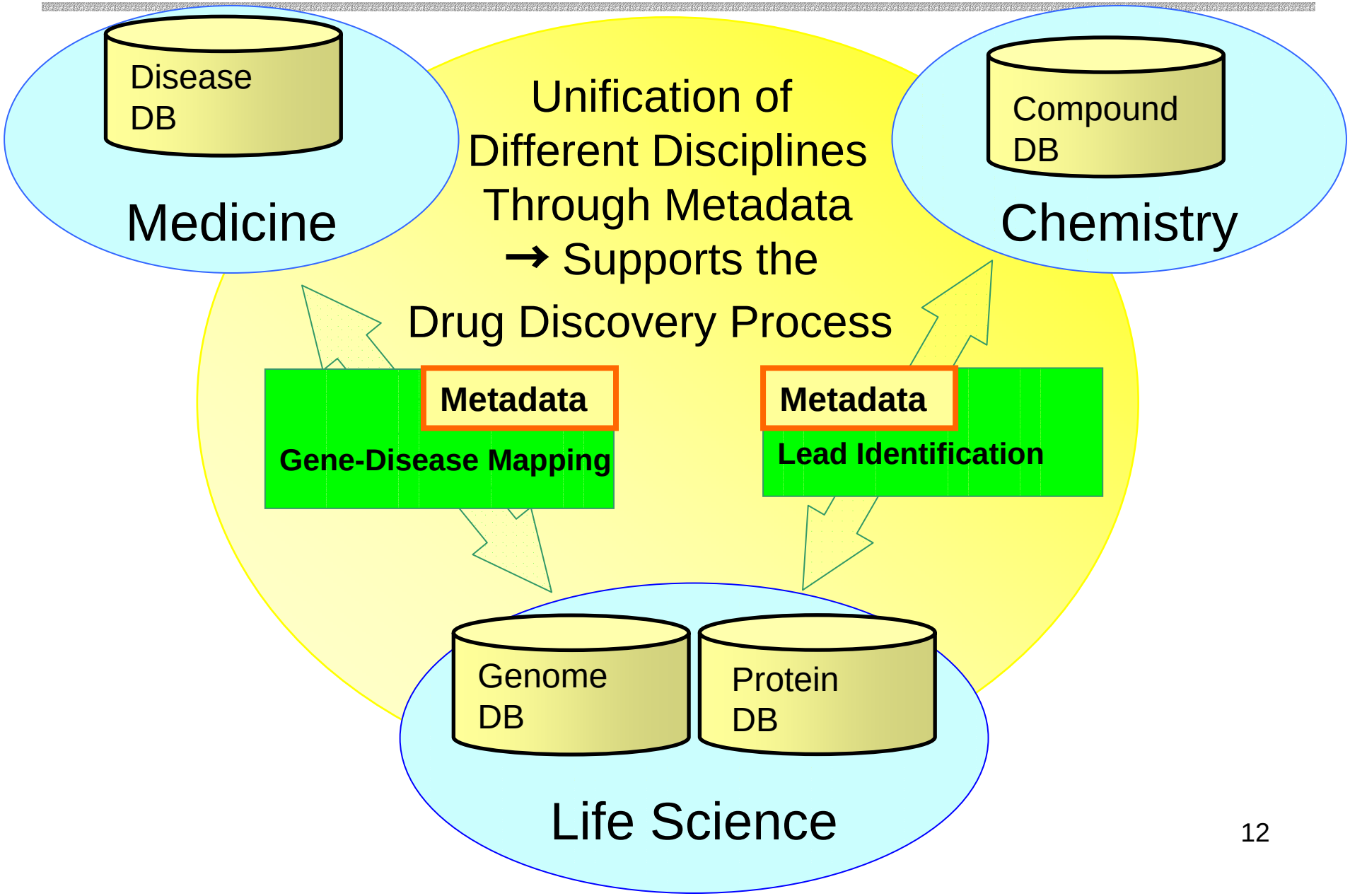
Databases Needed in Genome-based Drug Discovery



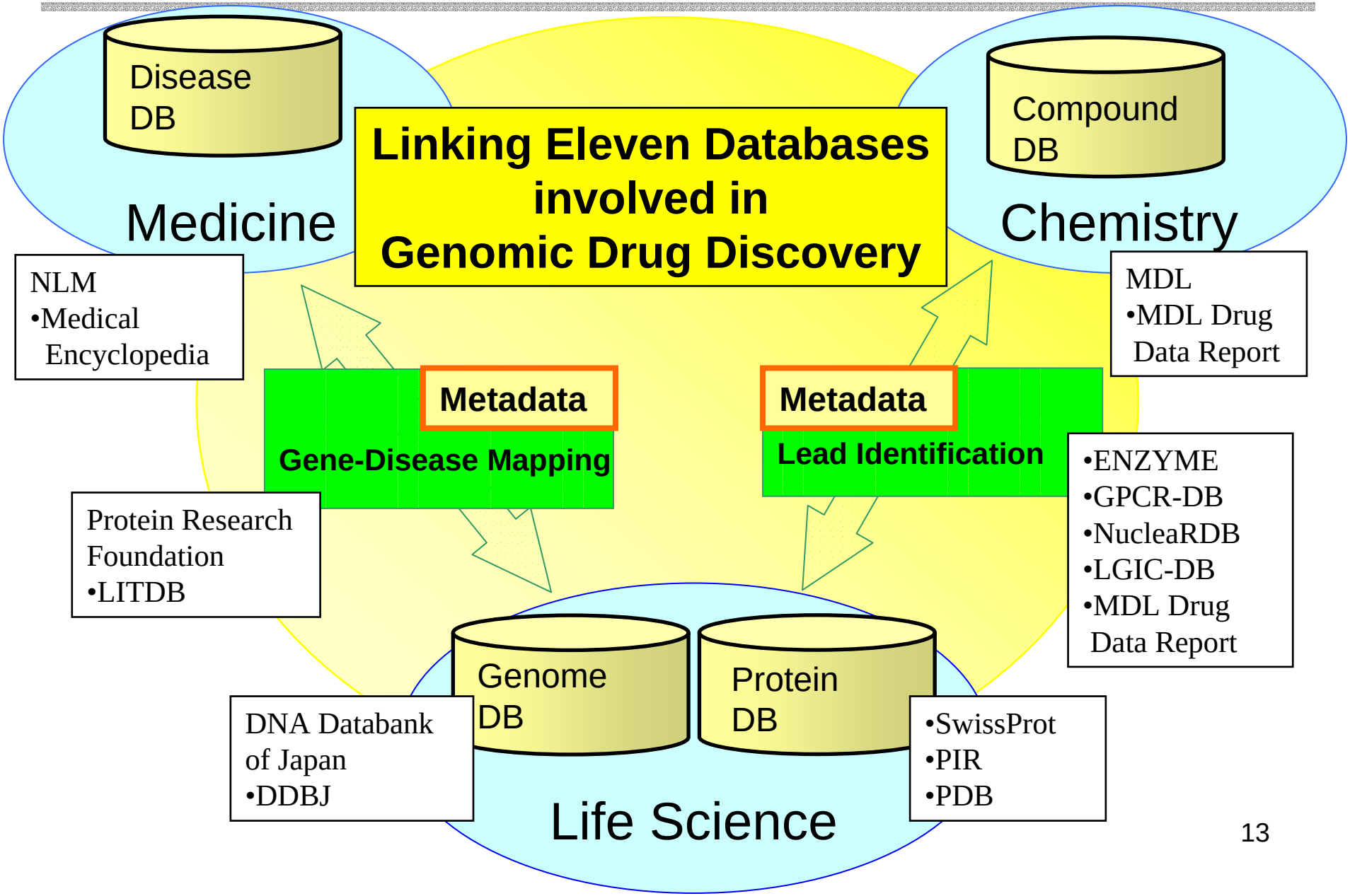
Semantic Gap Exists Between Databases and Their Corresponding Disciplines



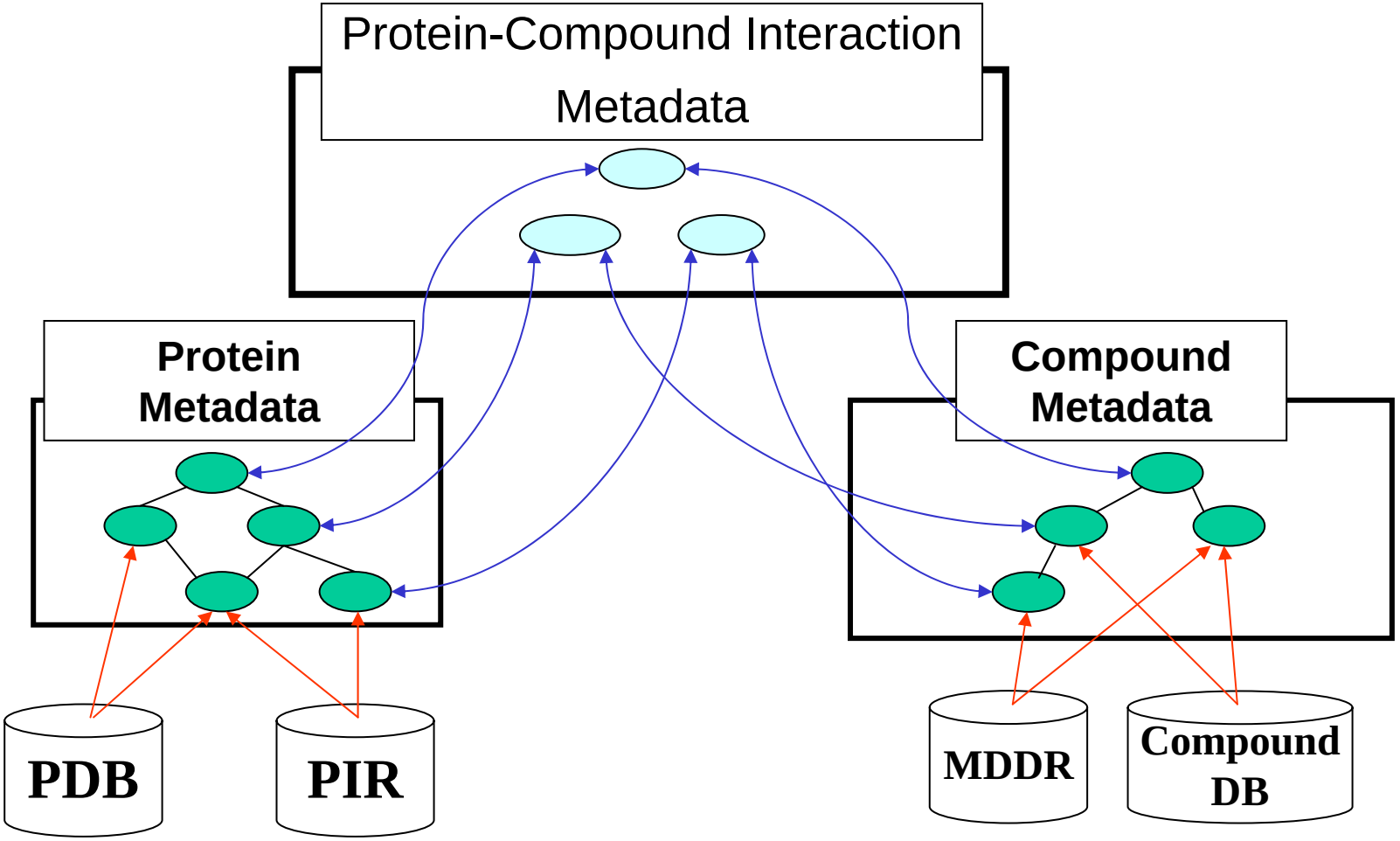
Linking Databases in Different Disciplines



Linking Databases in Different Disciplines

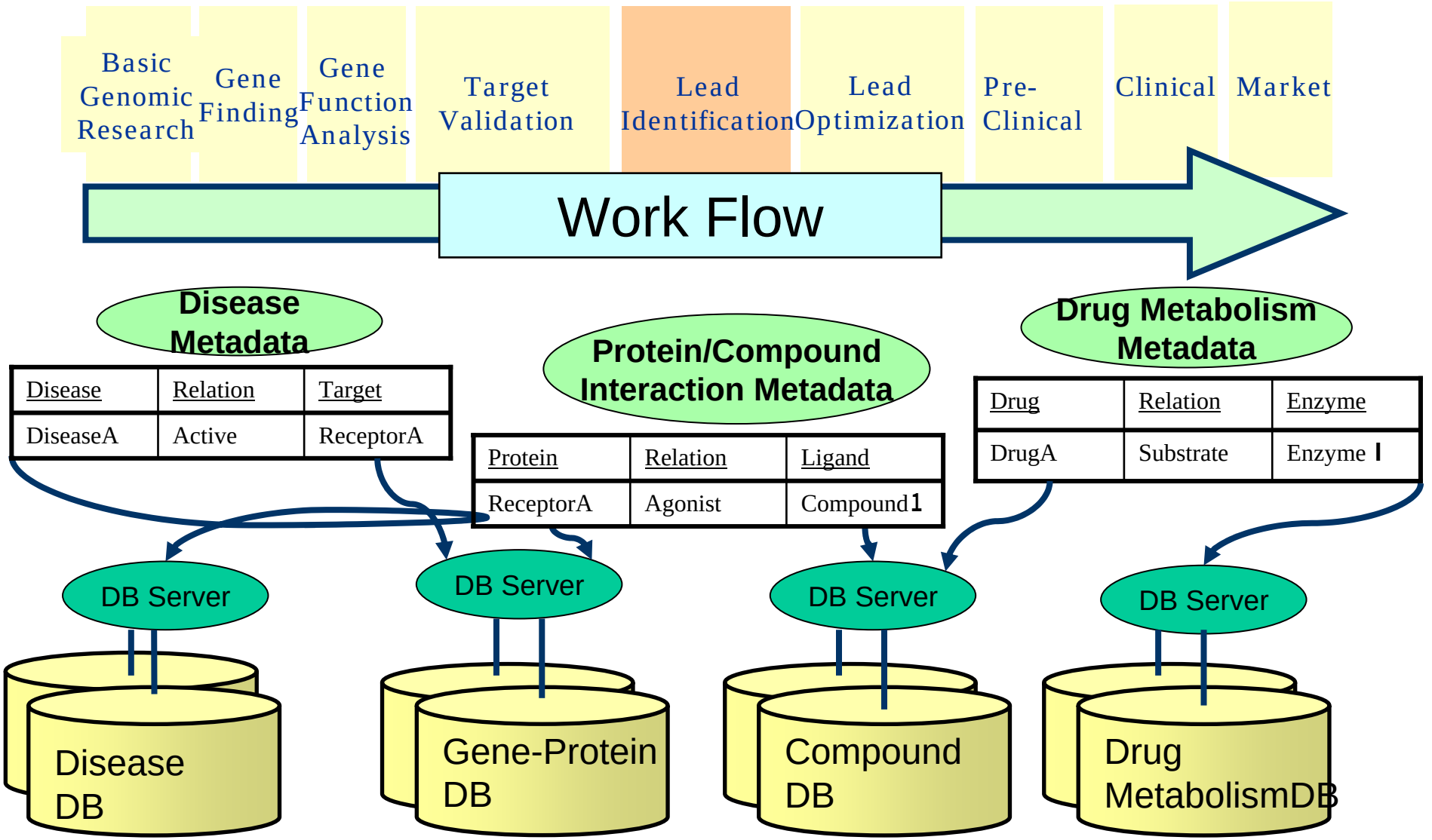


Two-Level Implementation of the Metadata

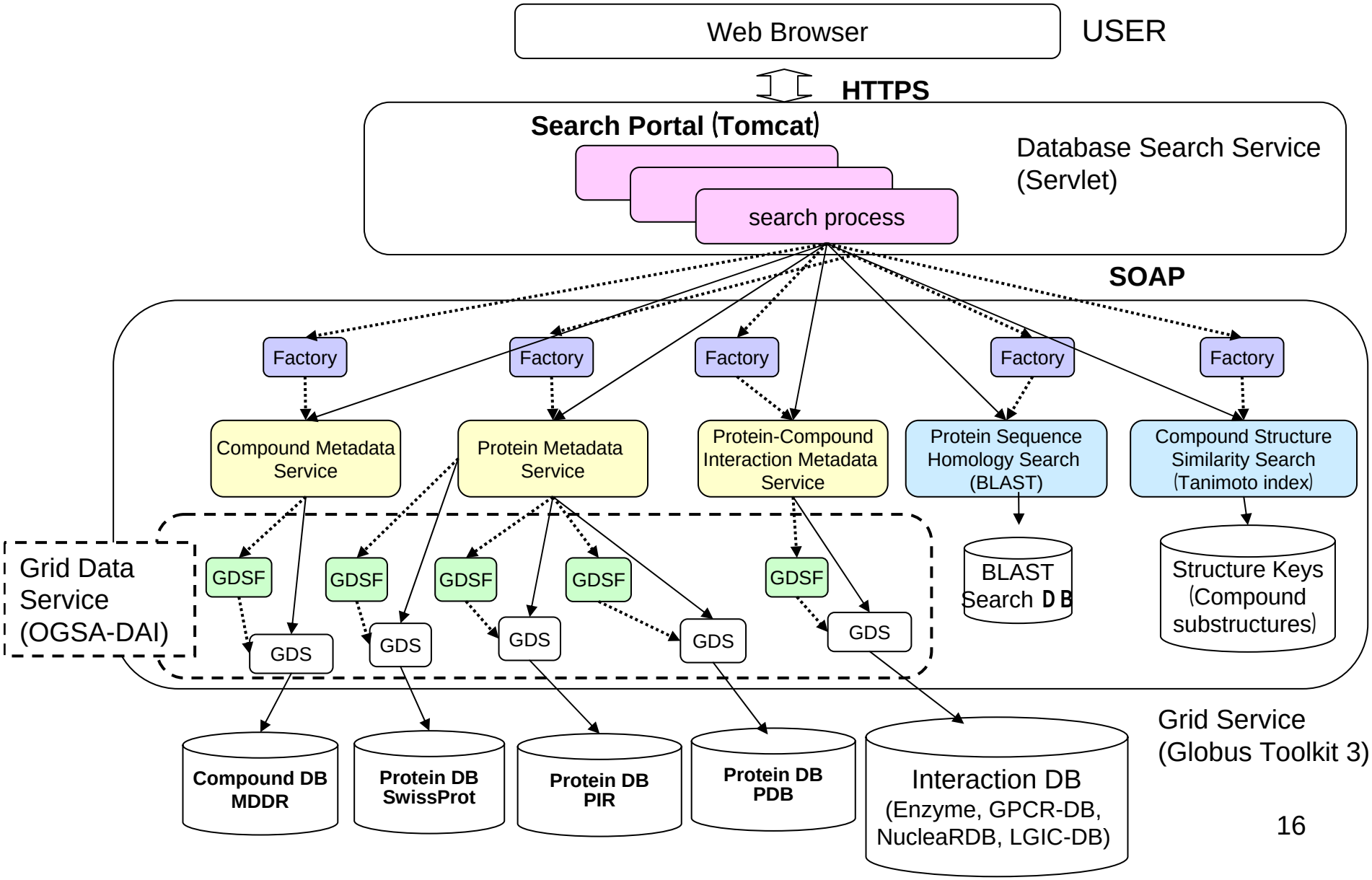


The relationship between groups in each category level of Protein Metadata and Compound Metadata

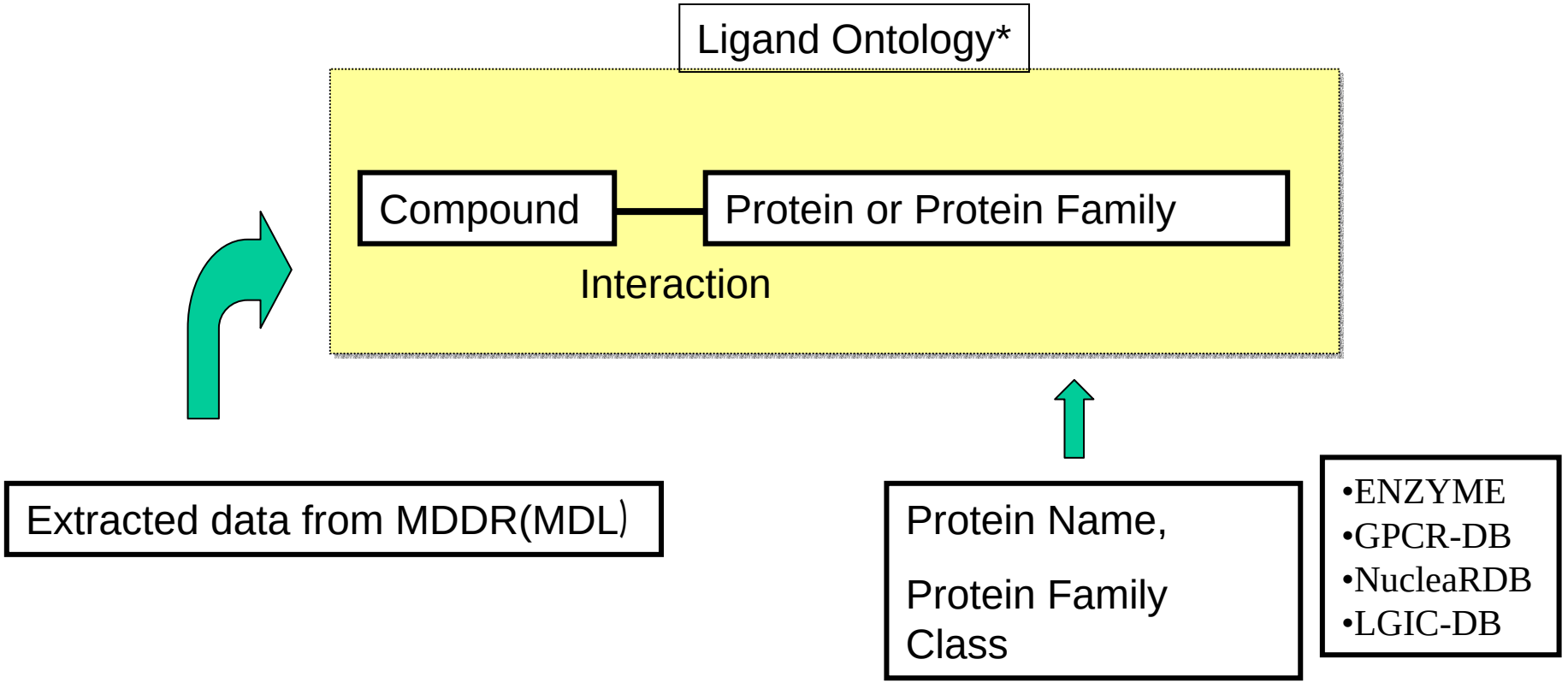
Metadata as Implemented on the Drug Discovery Workflow



Database System for Protein-Compound Interaction Search

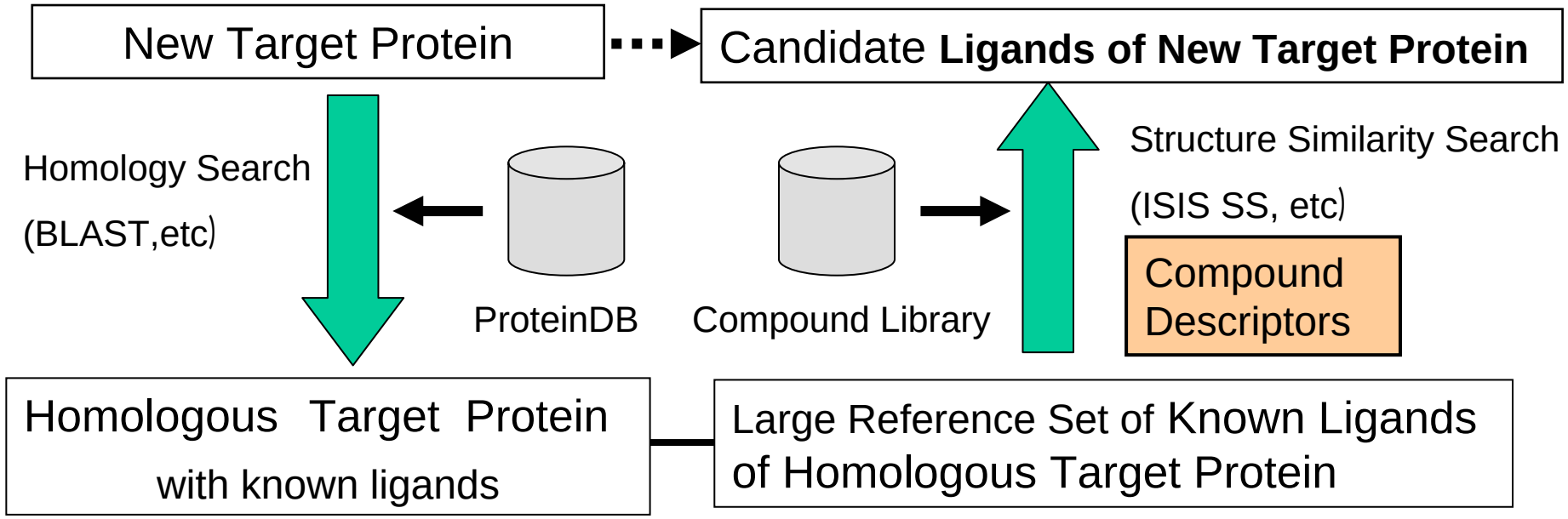


Strategy Used in Protein-Compound Interaction Search



* Schuffenhauer A, Zimmermann J, Stoop R, van der Vyver JJ, Lecchini S, Jacoby E. "An ontology for pharmaceutical ligands and its application for in silico screening and library design," *J Chem Inf Comput Sci.* 2002 Jul-Aug;42(4):947-55.

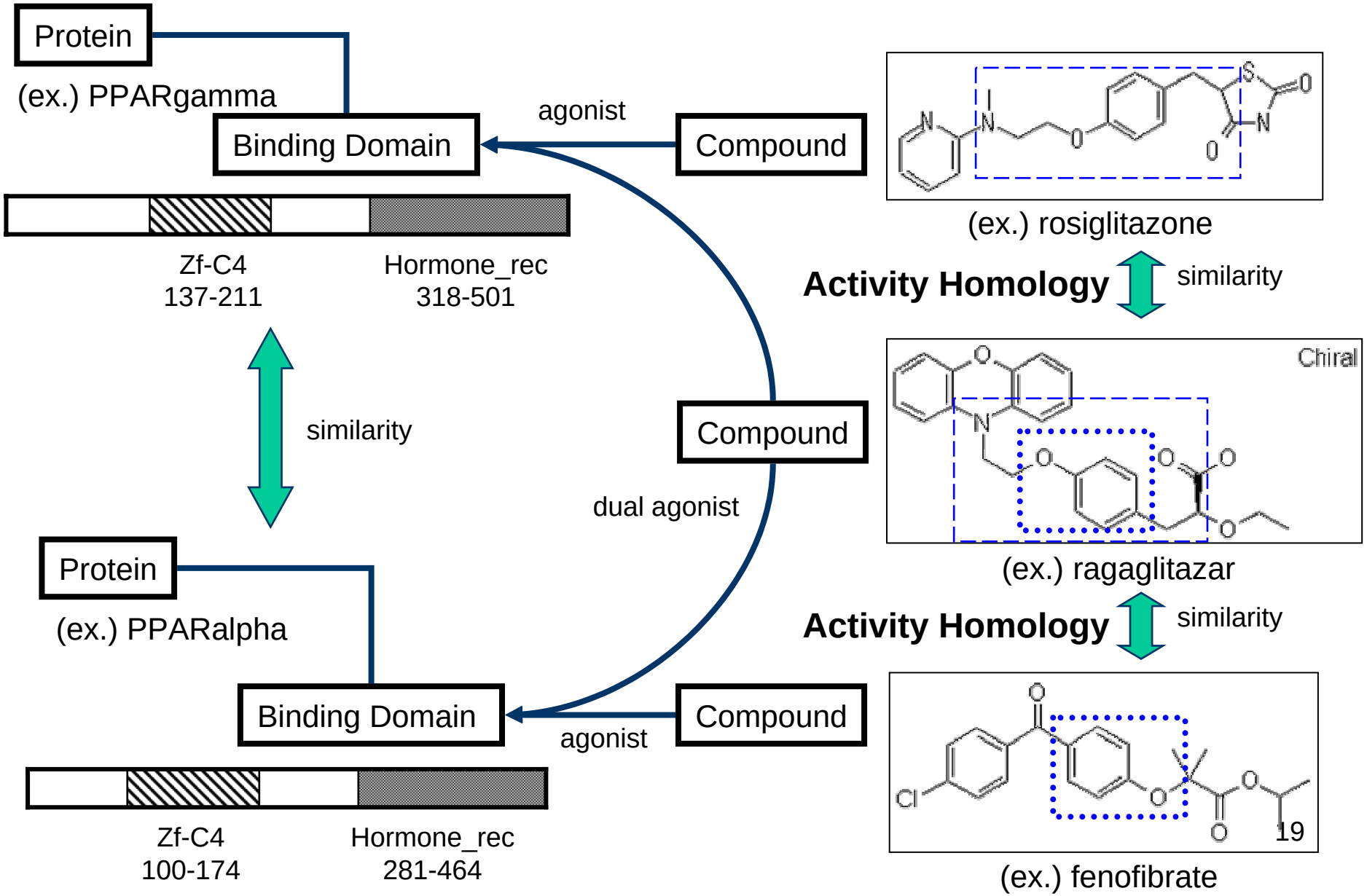
Process Flow in Protein-Compound Interaction Search



Interactions Search

Schuffenhauer A, Floersheim P, Acklin P, Jacoby E.,
"Similarity metrics for ligands reflecting the similarity of the target proteins",
J Chem Inf Comput Sci. 2003 Mar-Apr;43(2):391-405.

Example of Protein-Compound Interaction Search



Protein-Compound Interaction Search System Website

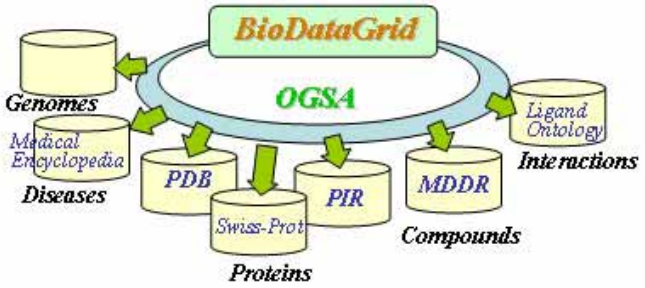
BioDataGrid

Protein-Compound Interaction Search

Overview

The BioDataGrid provides a cooperative search among molecular biology databases. This system is compliant to the OGSA which is a standard architecture of grid technologies. You need not to being aware of location or heterogeneity of databases.

Protein-Compound Interaction Search is an application on the BioDataGrid, to find interactions between proteins and comounds from protein view, disease view, or comound view.



Available Databases

Category	Database	Amount
Disease	Medical Encyclopedia	3079 entries
Genome	DDBJ	Human 7037852 entries, 10176023644 bases Mouse 5063486 entries, 6071844270 bases
Protein	Swiss-Prot	137885 entries, 50735179 amino acids
	PIR	283227 entries, 96134583 amino acids
	PDB	23073 entries
Compound	MDL Drug Data Report (MDDR-3D)	142553 entries

Applications Available to the User

- **Protein Sequence Search :**

  Retrieve the target protein's sequence by specifying its Protein ID.

- **Homology Search** :

  Search for proteins homologous to the target in the Protein DB.

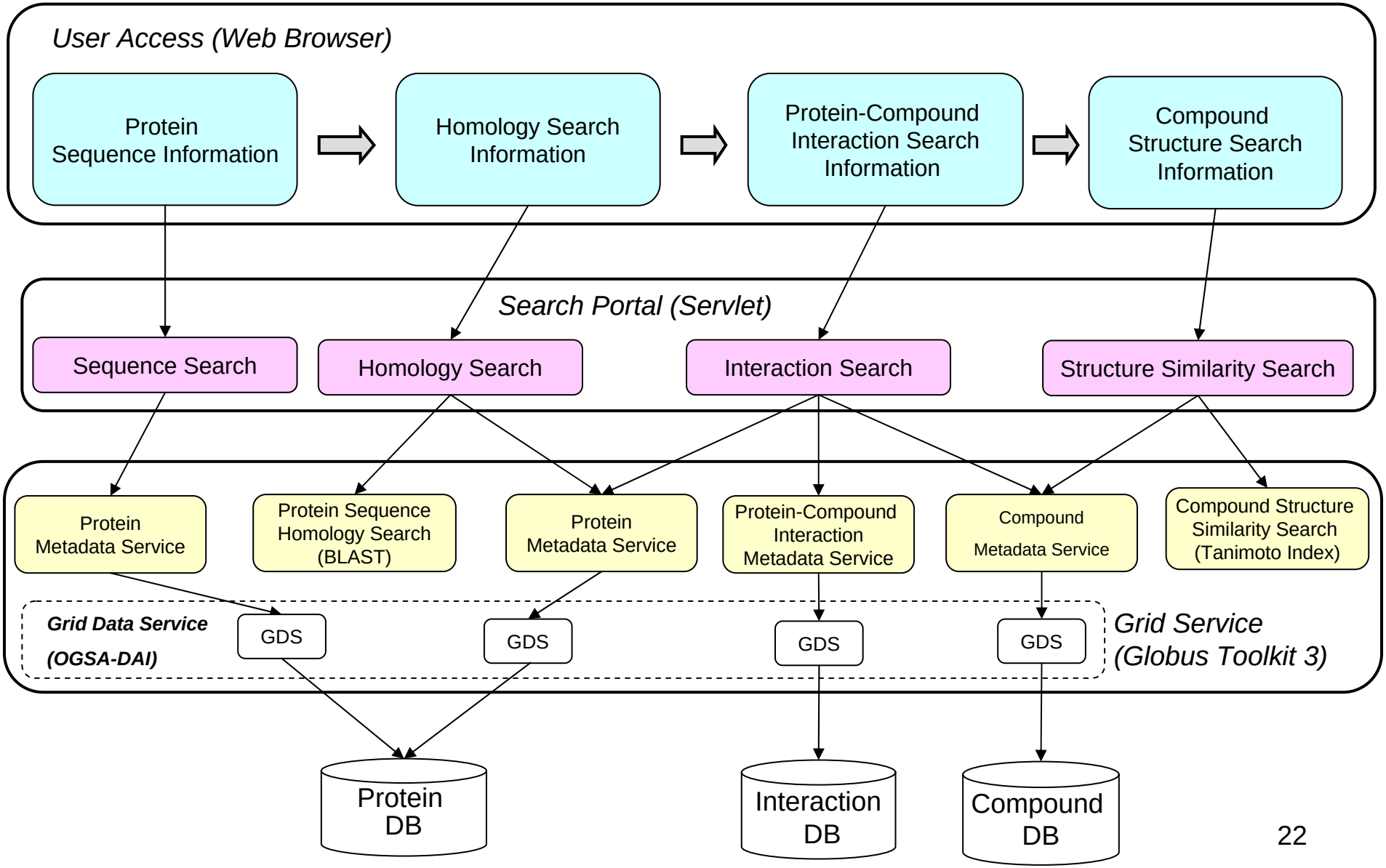
- **Protein-Compound Interaction Search** :

  Extract ligands that bind to the homologous proteins.

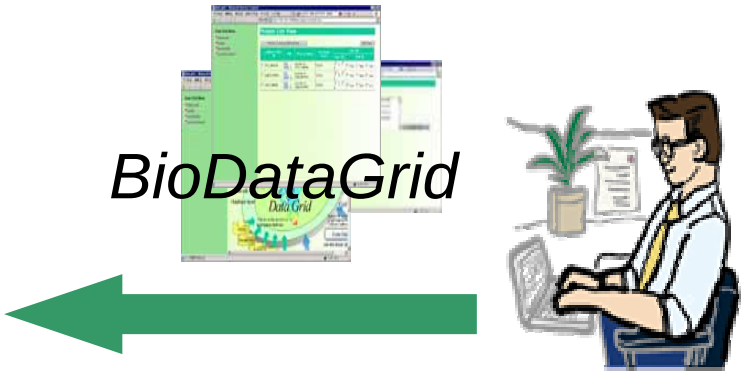
- **Compound Search** :

  Search for new compounds that may possibly interact with the target protein, by structural similarity to the extracted ligands.

Flow of User Access and Grid Service Execution

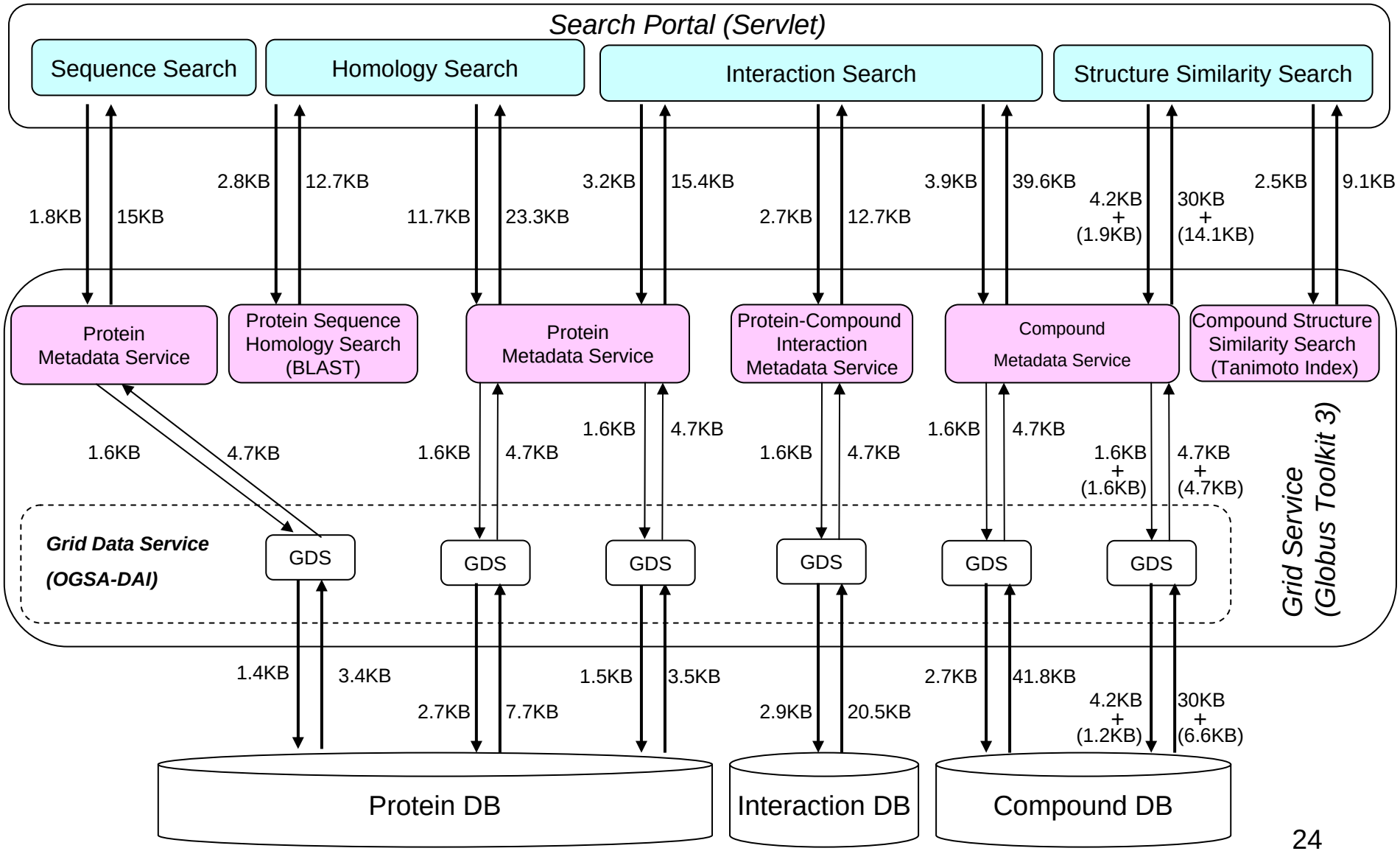


Actual Implementation of the Search System



- OS:Red Hat Linux 9
- CPU：Pentium4 (2.4GHz)
- Memory Size:4GB
- Java SDK 1.4.1
- Globus Toolkit 3 beta
- OGSA-DAI :Release 2.5
- Jakarta Tomcat 4.1.24
- MySQL 3.23.54

Average Amount of Data Flow for Each Grid Services



*values in parenthesis show average for displaying a single compound

Conclusion

- A data grid system that links together online databases was proposed
- Actual linking of 11 databases in the Life Sciences was explained
- An integrated heterogeneous database system based on the workflow of the genome-based drug discovery process was discussed
- Use of the latest grid technology like Globus Toolkit 3/OGSA-DAI in linking distributed databases was successfully proven

Future Works

- Implementation of security technologies
- Implementation of XML DBMS technologies
- Improvement of the search program

Acknowledgment

This study was conducted under the IT-Program of the Japanese Ministry of Education, Culture, Sports, Science and Technology.

The authors thank the Biogrid Project members for all their support.