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

Masato Kitajima[1,2] 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.*
E-mail: matsuda@ist.osaka-u.ac.jp

Abstract

Now that the human genome project has been completed and the human genes have been identified, genome-based drug discovery is starting to play a major role. It involves the process of target and lead identifications; identifying proteins (or drug targets) causing the disease, and identifying the lead compounds that would counteract them. We have developed a system for accelerating drug discovery, especially focusing on the lead identification. The system is based on Globus Toolkit3/OGSA-DAI and is accessible through the OGSA Grid Data Service. In this system, we have designed meta-databases for integrating information on disease, proteins (drug targets) and ligands (lead compounds). Using this system, one could easily screen a large library of compounds for ligands of a given protein, just by specifying the protein sequence. The effectiveness of our system is demonstrated by measuring the performance of the lead identification of several target proteins, such as nuclear receptors.

1. Introduction

Since the completion of the human genome sequence in 2001, genome-based drug discovery has started to play an important role. Genome-based drug discovery is a long process involving a series of stages: target identification/validation, lead identification/optimization and preclinical/clinical trials. These different stages require different databases (see Fig. 1). Thus it is needed to integrate all these databases into one big database for analyzing the entire drug discovery process, but the cost of it would be too expensive and thus it is impractical to do so. To solve this problem, we have devised a method for integrating their databases by connecting them all together using meta-databases (see Fig. 2). This connected network of databases makes use of the grid technology for delivering high performance searches of the databases.

Open Grid Services Architecture (OGSA) [1] enhances web service technology with advanced functions such as state management. The utilization of these functions standardized by Global Grid Forum (GGF) [2] makes it possible to realize how efficient our approach is for integrating heterogeneous databases that are necessary for lead identification. The effectiveness of the system is demonstrated by applying the method to identifying lead compounds of the glucocorticoid receptor protein.

Lead identification is a very crucial stage of drug discovery since it decides the fate of selected compounds in the latter phases of drug discovery. It is estimated that only 1 in 5,000 compounds investigated in preclinical discovery stages becomes a clinical lead, and about 1 in 10 drug candidates ever reaches the market [3]. Careless selection of the compounds in this stage could lead to detrimental results that would cause a very great loss; both in time and money. Therefore environments and workflows like the ones discussed in this paper are needed for assisting in the drug discovery process and expediting a part of the chain.

2. Approach for heterogeneous database federation

2.1. Metadata-based database federation

Genome-based drug discovery requires comprehensive knowledge and information in different disciplines such as

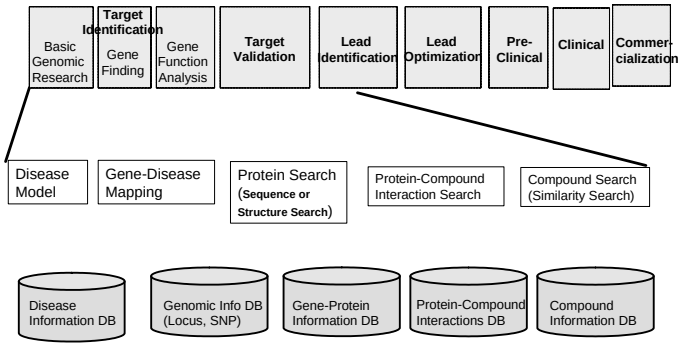

Figure 1. Databases needed in genome-based drug discovery.

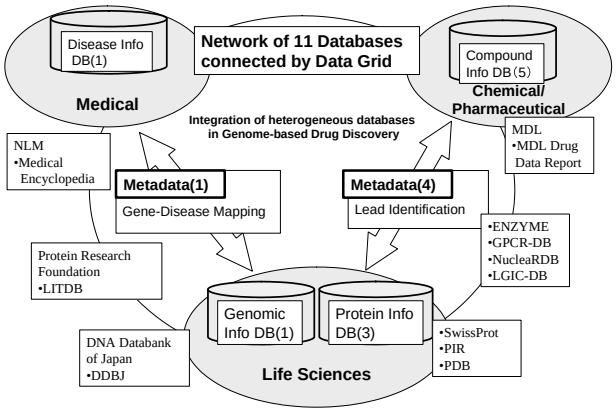

Figure 2. A network of databases for information integration in drug discovery.

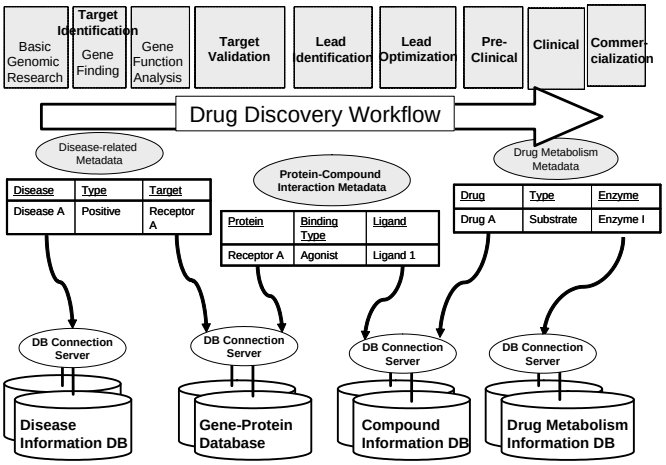

Figure 3. A workflow for lead identification with metadata-based database federation.

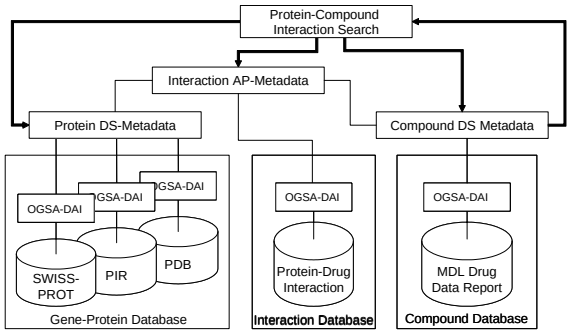

Figure 4. Two levels of database integration using AP (application) and DS (data-service) Metadatas.

genomics, medical science, molecular biology and pharmaceutics. Integration of information from these separate disciplines is essential in every stage for accelerating the drug discovery process. Several databases specific to each discipline are either publicly or commercially available. These databases were focused separately on disease-related, proteins-related or compounds-related databases. Success in the integration of these databases is very useful in the field of lead identification.

In this paper we introduce an integrated heterogeneous database system based on the drug discovery process workflow which is suitable for use in lead identification (see Fig. 3). The databases were grouped according to the domain of their contents, *i.e.*, disease information, gene-protein information, compound information and drug metabolism information. Two types of metadata were designed to bridge information among the databases, *i.e.*, application metadata (AP-Metadata) and data service metadata (DS-Metadata). AP-Metadata links information between database groups, while DS-Metadata links information within the same group. Figure 4 shows the relationship between AP-Metadata and DS-Metadata for an example of protein-compound interaction search services.

AP-Metadata provides a link between information found in databases of different domains, for example, a link between a "protein" entry in NucleaRDB[4] and a "ligand" entry in MDDR[8]. In contrast, DS-Metadata provides a link between information found in databases of the same domain, for example a link between the "protein" entries in SWISS-PROT[9], PIR[10] and PDB[11]. DS-Metadata provides a unified format for all the databases in the same group. Each entry in the databases is assigned a unique identifier (id) according to the unified format. AP-Metadata

and DS-Metadata keep reference pointers to the original databases. The reference pointers include their database names and their database IDs for database entries.

2.2. Open Grid Services Architecture

We use the grid technology as one of the most promising technologies that enable us to efficiently integrate heterogeneous resources for lead identification. Recently GGF has proposed OGSA [1]. OGSA has prescribed uniform grid service interfaces as an extension of traditional web services with new functions such as state management and life cycle management. Every service is expected to have an interface that is described with XML, exchange messages in an XML format via Internet-based protocol (*e.g.,* SOAP).

As a tool which is constructed by OGSA, OGSA-DAI (Open Grid Service Architecture Data Access and Integration)[12] has been developed in the e-Science project[13]. OGSA-DAI is a set of grid services that enables us to make various data resources accessible as grid services. It will support DB2, Oracle, MySQL and Xindice. By using OGSA-DAI, the database will be integrated virtually using web mechanisms such as SOAP to enable database services to operate within the XML scheme. The architecture was proposed in detail in our previous papers[14, 15]. In our system, the grid services are integrated by using Globus Toolkit 3 with OGSA-DAI (see Fig. 4).

3. Application in lead identification

Our prototype system has been built giving special focus on its actual use in genome-based drug discovery. In particular, we see its use in the process of lead identification; *i.e.,* in selection of the most suitable chemical compounds for further development into drugs that are both safe and effective. Lead identification is important since it decides the fate of compounds to be refined and tested in later phases of drug discovery.

Our approach to lead identification could briefly be described in four steps; they are as follows:

1. Searching for disease-related target proteins (Protein Search).
2. Searching for homologous proteins of the target (Homologous Search).
3. Searching for compounds interacting with the selected proteins (Protein-Compound Interaction Search).
4. Searching for compounds that are similar in structure as the above, from a large set (Similarity Search).

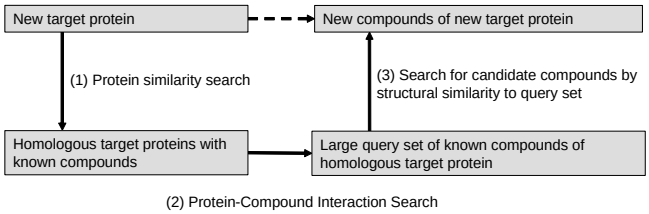

Figure 5. Lead identification steps using various database searches.

The first step involves the selection of target proteins related to the disease that is in question. In our prototype system, the user selects a disease and the system will automatically retrieve all target proteins it could find in the database related to the specified disease.

Next, the user may select all or some of these target proteins to further find more proteins that may be structurally or functionally similar to them. Homology search is done using specialized tools such as the NCBI-BLAST [16]. Only homologous proteins that passed a certain threshold value for the BLAST E-value are selected.

After finding and selecting the homologous proteins, the user may search for compounds that can alter the action of any of these proteins. This is done by searching from a protein-compound interactions database. Schuffenhauer and coworkers introduced a compound-protein interaction database[17]. The database covers compounds with hierarchical levels of protein classification based on pharmacological activity. The classification used for a protein-compound interaction search is as shown in Fig. 5. An advantage of the method is that it is able to predict the lead compounds for a specific target protein that has unknown biological function.

The last step involves searching for lead compounds from a very large set (candidate set) that are similar in structure with the compounds found above (reference set). Structural similarity between two compounds is evaluated using the Tanimoto coefficient [19].

In our prototype system, both reference and candidate sets are included in the same database. Currently the reference set is selected manually based on user interest. The candidate set is obtained by removing the reference set from the database. All compounds in the candidate set were ranked by their similarity.

4. Implementation

The following is the configuration of the system used for lead identification:

- OS: Redhat Linux 9

- CPU: Pentium4 2.4 GHz
- Memory Size: 4GB
- Java: Java SDK 1.4.1 03 b02
- Globus Toolkit: Globus Toolkit 3.0.2
- OGSA-DAI: Release 2.5
- Container: Jakarta Tomcat 4.1.24
- DBMS: MySQL 3.23.54

11 bio-related databases are aggregated and categorized based on their types for efficiency. The categories are disease, genome, protein, compound and interaction as shown in Fig. 2. Each database category is available for query through our grid service. The data and program for protein-compound interaction searching are installed in the same machine.

4.1. Databases

We use the SWISS-PROT Release 39.17 of 27-Apr-2001 for the proteins database and MDL Drug Data Report (MDDR) Release 2003.2 [8] for the compounds database. SWISS-PROT contains 137,885 protein entries while MDDR includes 142,553 compound entries. Proteins are denoted by their SWISS-PROT accession numbers (*e.g.*, P014050) and compounds are denoted by their MDDR registry id (*e.g.*, 209035).

We bridge the protein entries to the compound entries using the NucleaRDB relational database for a protein-compound interaction database, which is annotated by protein classification [18]. All these databases are stored in MySQL. Protein similarity searching is also available using NCBI-BLAST Version 2.2.6.

We removed redundant compounds existing in the MDDR database by selecting only compounds that have the same value for both the 'MDDR registry id' and 'PREF.NUMBER' fields. The 'PREF.NUMBER' field contains the Prous Entry Number of a compound, which indicates if it has the greatest biological activity or is the representative compound in the series of derivative compounds.

4.2. Search Results

As an example, the human glucocorticoid receptor (GR) was used as a target protein to identify lead compounds in the database. First, the homologous proteins of the GR protein in the SWISS-PROT database were searched by using the NCBI-BLAST program.

Search results are displayed in order of decreasing similarity of the homologous proteins as measured by their BLAST E-values. Progesterone receptor (PR), androgen receptor (AR) and estrogen receptor (ER) from human were selected from the results of the homology search.

The protein-compound interaction search were then applied to the selected proteins, and gave 347 compounds that are known to be active to these proteins from the interaction database.

Out of the 347, only 5 compounds were selected (115029, 170262, 315962, 322129 and 329279) whose interaction types are agonist, and their similar structures were searched against the MDDR database by setting the Tanimoto coefficient threshold to 0.9, yielding a final set of 26 compounds. The processing time of the query is about 7 seconds.

To evaluate processing times by using OGSA-DAI in the prototype system, we selected estrogen-like receptors for our target. First, we searched for the number of diseases related to estrogen-like proteins. From the database, we found 36 diseases that are related to estrogen-like receptors and a total of 50 proteins related to these retrieved diseases. Calculating the average, we found that 1.39 target proteins are related per disease relating to estrogen-like receptors. Next, we searched for all compounds that have interactions with the above target proteins and found that, on average, there are 86.7 compounds related per target protein. The average searching time for the four target proteins is 1,847 msecs, while the average searching time for 347 compounds is 21.3 msecs.

5. Conclusions and future works

We have introduced a system for identifying lead compounds of a disease-related target protein based on a heterogeneous database federation with Globus Toolkit 3 and OGSA-DAI. Specific example of the application of the system was demonstrated using the human glucocorticoid receptor as the target protein.

Current implementation of the system makes use of a relational database for storing bio-related data. An XML native database may be introduced in order to express the hierarchical structure of these data. Further improvement of the system is needed to solve security issues in practical drug discovery environments.

Enhancements must also be done to automate the setting of the number of results to be displayed to the user. In addition, design of new compound descriptors is needed in order to recognize more characteristic substructure patterns contributing to the increase in accuracy of substructure searches.

Acknowledgments

This study was performed through IT-program of Min-

istry of Education, Culture, Sports, Science and Technology. The authors thank the Biogrid project members.

References

[1] I. Foster, C. Kesselman, J.M. Nick, and S. Tuecke, "The Physiology of the Grid. An Open Grid Services Architecture for Distributed Systems Integration," Open Grid Service Infrastructure WG, Global Grid Forum, 2002, http://www.ggf.org/

[2] Global Grid Forum, http://www.gridforum.org/

[3] G. Schneider and S.-S. So, "Modeling Structure-Activity Relationships," G. Schneider and S.-S. So (ed.), *Adaptive Systems in Drug Design*, Landes Bioscience, Georgetown TX, 2002.

[4] NucleaRDB: An Information System for Nuclear Receptors, http://receptors.ucsf.edu/NR/

[5] Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzyme-Catalysed Reactions (NC-IUBMB), http://www.chem.qmw.ac.uk/iubmb/enzyme/.

[6] GPCRDB: Information system for G protein-coupled receptors (GPCRs), http://www.gpcr.org/

[7] LGICdb: The Ligand Gated Ion Channel Database, http://www.pasteur.fr/recherche/banques/LGIC/LGIC.html

[8] MDL Drug Data Report Version 2003.2, MDL ISIS/HOST software, MDL Information Systems, Inc. San Leandro, CA, http://www.mdl.com.

[9] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The SWISS-PROT Protein Knowledgebase and its Supplement TrEMBL in 2003," *Nucleic Acids Research*, vol.31, no.1, 2003, pp.365–370.

[10] C.H. Wu, L.-S.L. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R.S. Ledley, B.E. Suzek, C.R. Vinayaka, J. Zhang, and W.C. Barker, "The Protein Information Resource," *Nucleic Acids Research*, vol.31, no.1, 2003, pp.345–347.

[11] P.E. Bourne, K.J. Addess, W.F. Bluhm, L. Chen, N. Deshpande, Z. Feng, W. Fleri, R. Green, J.C. Merino-Ott, W. Townsend-Merino, H. Weissig, J. Westbrook, and H.M. Berman, "The Distribution and Query Systems of the RCSB Protein Data Bank," *Nucleic Acids Research*, vol.32, no.1 (Database issue), 2004, pp.D223–D225.

[12] OGSA-DAI Project, http://www.ogsadai.org

[13] UK e-Science (GRID) core programme, http://www.escience-grid.org.uk/

[14] H. Nakamura, S. Date, H. Matsuda, and S. Shimojo, "A Challenge towards Next-Generation Research Infrastructure for Advanced Life Science," *New Generation Computing*, vol.22, no.2, 2004, pp.157–166.

[15] T. Kosaka, Y. Tohsato, S. Date, H. Hatsuda, and S. Shimojo, "An OGSA-Based Integration of Life Scientific Resources toward Drug Discovery," Proc. of HealthGRID 2004, Clermont-Ferrand, 2004.

[16] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs," *Nucleic Acids Research*, vol.25, no.17, 1997, pp.3389–3402.

[17] A. Schuffenhauer, J. Zimmermann, R. Stoop, J.-J. van der Vyver, S. Lecchini, and E. Jacoby, "An Ontology for Pharmaceutical Ligands and its Application for in silico Screening and Library Design," *Journal of Chemical Information and Computer Sciences*, vol.42, no.4, 2002, pp.947–955.

[18] A. Schuffenhauer, P. Floersheim, P. Acklin, and E. Jacoby, "Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins," *Journal of Chemical Information and Computer Sciences*, vol.43, no.2, 2003, pp.391–405.

[19] P. Willett and V.A. Winterman, "Comparison of Some Measures for the Determination of Intermolecular Structural Similarity," *Quant. Struct. Act. Relat.* vol.5, 1986, pp.18–25.