AN INTEGRATED RECEPTOR DATABASE (IRDB)

K Nakata[1*], T Takai-Igarashi[1,2], T Nakano[1] and T Kaminuma[1,3]

[1] Division of Chem-Bio Informatics, National Institute of Health Sciences, 1-18-1, Kamiyoga, Setagaya-ku, Tokyo 158-8501, Japan Email: nakata@nihs.go.jp, nakano@nihs.go.jp,
[2](Present Address) Graduate School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan
Email: takako@ims.u-tokyo.ac.jp
[3](Present Address) Bio Dynamics, Inc., 4-3-16-301, Yoga, Setagaya-ku, Tokyo 158-0097, Japan
Email: kaminuma@cbi.or.jp

ABSTRACT

Various receptor data were collected, edited and integrated into an Integrated Receptor Database (IRDB). The data stored includes structural data (amino acid sequences, their secondary-structure and three-dimensional structure), functional data, binding affinity, cell signaling data etc. The purpose of this database is to allow structural biologists, drug designers and toxicologists to analyse and elucidate receptor-ligand dockings and the resultant post-binding signal transduction pathways. IRDB is available on line (http://impact.nihs.go.jp/RDB.html)

Keywords: Receptor, Structure, Binding affinity, Cell Signaling, Drug

1 INTRODUCTION

The receptor-ligand binding triggers a series of reactions in a living system. So, detailed knowledge and data about receptors and their ligands are an important basis for understanding living systems and diseases, and for designing new drugs. Because of the advances in molecular biology, which were accelerated by the Human Genome Projects, a huge amount of DNA sequence and protein structural data have been accumulated and are available for public use. Although we had already developed a Receptor Database (Nakata, Takai & Kaminuma, 1999), we have now revised the RDB, integrating new functions and updating the data. The present version of RDB, which we call an Integrated Receptor Database (IRDB), is the updated version of our old Receptor Database (Nakata, Takai-Igarashi, Nakano & Kaminuma, 2001a).

The Internet/World Wide Web (WWW) technology had allowed us to use powerful viewers for representing retrieved data and knowledge graphically. This technology has also allowed us to link RDB dynamically to other related WWW sites. In the previous version of RDB, the goal of our system was to provide one-stop shopping on receptor data. The system uses a good viewer to represent information useful for endocrine disruptor and the drug design; information such as the structural data and binding sites, and the cell signaling pathway that is triggered by a ligand binding. IRDB includes more structural data, binding affinities, the transcription factors and regions, and single nucleotide polymorphisms (SNPs).

2 SYSTEM AND METHODS

2.1 Purpose of the database system

The purpose of RDB is to store data and knowledge on receptor proteins and properties. This data and knowledge includes protein structures and their functions, ligands and binding affinity data, cell signaling information, drug and SNPs. The database users are those who study biology and the mechanisms of disease, and those who are developing drugs based on the structure-based drug design (SBDD) approach and personalized medicine.

2.2 Hardware and software

RDB was implemented on a UNIX workstation (e.g. Silicon Graphics OCTANE). We used an object oriented database management software ACEDB (A Caenorhabditis elegans DataBase, 2001; Dunham, Durbin, Thierry-Mieg & Bentley, 1994; Stein & Thierry-Mieg, 1998) as the base system. To modify data and insert new functions, PERL and/or C programs were integrated into ACEDB. As for the method of calculating the three-dimensional structure of ligands, we used Molecular Mechanics 2 (Allinger, 1977). The three-dimensional (3D) structural image is provided by computer commands that call visualization tools, such as Chime (Martz, 2002) or RasMol (Sayle & Milner-White, 1995).

Table 1. Tree class structure

Class	Contents	(Pointer)	(Contents in pop-up window)
RG	Membrane / Nuclear Receptor		LG List
LG	Large Group List		MG List / Group List
MG	Middle Group List		Group List
Group	Sequence similarity to the main species	[MulSq]	Multiple sequence alignment
	Species	species list	Receptor information
Receptor	Protein Data (Acc. No., No. of Seq.)	[PIR ref / SP ref]	PIR[#1] entry / Swiss Prot [#2]entry
		[PIR seq / SP seq]	Transmembrane Region DNA-Binding Region Ligand-Binding Region
		[St 2D-pred]	2D Structure Prediction[#3]
	Sequence similarity to the main species	[MulSq]	Multiple sequence alignment
	3D data (Overlap region with PIR/SP)	[PDB ref]	PDB [#4]entry
		[St 3D-image]	3D image
	DNA Data (Accession No., No. of Seq.)	[GB ref]	GenBank[#5] entry
		[GB seq]	DNA Sequence
	Gene Data (Symbol, Aliases, Map pos.)	[GDB ref]	GDB[#6]entry
	SNPs information	[snp]	SNPs Collection System[#7] entry
	Drug information (Generic Name)	[drug]	Drug[#8] data
	Cell Signaling Networks information	[Signaling]	CSNDB[#9] entry
	Transcription Factor information	[Transfac]	Transfac[#10] entry
	Transcription Region information	[TRRD]	TRRD[#11] entry
	Binding Affinity information	[BindAff]	BADB[#12] entry

[#1] PIR: Protein sequence database (http://www-nbrf.georgetown.edu/pirwww/search/textpsd.html)
[#2] Swiss Prot: Protein sequence database (http://www.expasy.ch/sprot/)
[#3] 2D Struc. Prediction: Protein secondary-structure prediction, BCM Search Launcher, (http://searchlauncher.bcm.tmc.edu/seq-search/struc-predict.html)
[#4] PDB: 3D biological macromolecular structure database (http://www.rcsb.org/pdb/)
[#5] GenBank: DNA sequence database (http://www.genome.ad.jp/dbget-bin/www_bfind?genbank-today)
[#6] GDB: The Genome database (http://gdbwww.gdb.org/)
[#7] SNPs Collection System: An agent system for collecting SNPs data. (http://search.nihs.go.jp/snp/index.html)
[#8] Drug DB: Drug database (http://moldb.nihs.go.jp/moldb/)
[#9] CSNDB: Cell Signaling Networks Database (http://geo.nihs.go.jp/csndb/)
[#10] Transfac: Transcription factor database (http://transfac.gbf.de/TRANSFAC/index.html)
[#11] TRRD: Transcription Regulatory Region Database (http://wwwmgs.bionet.nsc.ru/mgs/)
[#12] BADB: Binding Affinity Database (http://moldb.nihs.go.jp/eddb/afdb/)

2.3 System configuration

In accordance with the architecture of ACEDB, all the information was stored as structured objects in tree forms. A tree can be arbitrarily extended in any direction as more information is gathered about a particular aspect of an object. Similar objects are grouped together within a "class". The class governs what can be stored in an object and how it is displayed and used. To make all the information about an object available, objects often contain labels (pointers) to other objects. They can also contain letters, numbers, objects and computer commands. The tree class structure in RDB was shown in Table 1. All information was integrated in the file for ACEDB.

The overall system configurations are shown in Figure 1. In the off-line system, we used a BLAST search (Altschul, Madden, Schaffer, Zhang, Zhang, Miller et al., 1997) and the MView program (Brown, Leroy & Sander, 1998) for studying sequence similarity. Sequences were passed to the BLAST search program and the result was modified by the MView program and stored in the RDB.

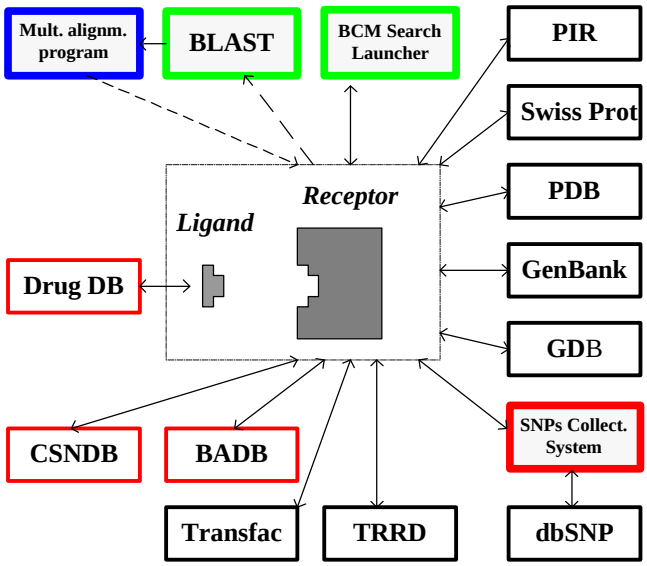

Figure 1. System configuration
↔ on-line linking; ⋯› off-line data flow,
database, in-house database, program, system, in-house system
BLAST: Sequence similarity search (http://blast.genome.ad.jp/)
BCM Search Launcher (http://searchlauncher.bcm.tmc.edu/seq-search/struc-predict.html)
dbSNP: Single Nucleotide Polymorphism database (http://www.ncbi.nlm.nih.gov/SNP/)
Mult. align. program: Multiple alignment program (off-line)

2.4 Database sources

Receptor proteins were retrieved from the Swiss Prot and PIR databases on the Internet. The sources of the RDB are classified into three categories: (1) those that are collected from various references (basic data), (2) those that are retrieved from external on-line databases by the user's request (protein structure data, SNPs collection data, etc.), (3) those that were generated by some theoretical calculations (sequence similarity data, 3D structures of the ligand, etc.).

For the protein secondary-structure prediction, the corresponding amino acid sequence data was edited in a

relevant form and passed to another analytical system, for instance the BCM search Launcher (Smith, Wiese, Wojzynski, Davison & Worley, 1996). The relevant drug data, which includes Japanese and US drugs with CAS registry numbers, chemical structures and 3D structures are compiled in a Drug Database. Japanese drugs were retrieved from JAN (Japanese Accepted Names for Pharmaceuticals, n.d), and US drugs were from USP-NF (United State Pharmacopoeia – National Formulary, n.d). 3D structures of drugs were calculated using MM2.

2.5 Database contents

The numbers of LG (Large Group) in 'Membrane and Nuclear Receptor' are 36 and 5, respectively. At present, the total number of receptor proteins in the RDB is 1780. Each receptor protein has labels for the PIR / Swiss Prot entry, functional region and the secondary-structure prediction. The numbers of the DNA binding sites and ligand binding sites are 250 and 170, respectively. An aligned sequences chart for the different species was stored for the main receptors. There are 410 entries for 3D structure data in the RDB.

DNA sequences, which are translated into the receptor proteins, are available for each receptor protein. Gene data and SNPs information are included for most human receptor proteins. Data about drugs that bind to the receptors, is included as an example .

Cell signaling information is available only for human receptors (Takai-Igarashi, Nadaoka & Kaminuma, 1998; Takai-Igarashi & Kaminuma, 1999). Transcription information (Wingender, Chen, Fricke, Geffers, Hehl, Liebich et al., 2001; Kolchanov, Ignatieva, Ananko, Podkolodnaya, Stepanenko, Merkulova et al., 2002) covered all species. The binding affinity data was included only for endocrine disruptor related receptors (Kaminuma, Takai-Igarashi, Nakano & Nakata, 2000).

2.6 Automatic genetic variation data collection

An agent system of collecting Single Nucleotide Polymorphisms (SNPs) data on the Internet, was developed to search for and retrieve SNPs data related to those genes and proteins pre-registered in the system (Nakata, Takai-Igarashi, Nakano & Kaminuma, 2001b). The related gene names were previously input into the agent system and linked to IRDB. The position of any allelic frame-shift in the DNA sequence, the corresponding amino acid offset, and the converted amino acids are represented in the SNPs information.

3 DISCUSSION

IRDB was designed to be one part of the pharmaco-informatics infrastructure for genome-based personalized medicine (Kaminuma, Nakata, Nakano & Takai-Igarashi, 2001). A drug or its metabolite binding to target biomolecules, such as membrane receptors, cytoplasm enzymes, and nuclear receptors, triggers a series of reactions. Although these target molecules are not yet fully identified, it was estimated that nearly half of them are receptors (Drews, 1998).

For structure-based drug design, exact 3D structures of receptors and ligands are essential. Although only the 3D structures of a few receptors have been identified, theoretically predicted secondary-structures and aligned sequence- charts for different species are available in RDB. Although only endocrine disruptor related data is now included in BADB, much more experimental binding-affinity data are still required and theoretically calculated binding-affinity values could be included in the database in the future.

The signal pathways, which are the post-binding effects of the receptor and ligand, can be retrieved via CSNDB. The signal tranduction and transcription information may help in understanding the effects of various chemicals, such as drugs or environmental chemicals, on the living system via gene expression. SNPs data for receptors is essential for personalized medicine, in areas such as drug responses and common disease predisposition. We intend to include a link to OMIM in the near future, relating the receptor and ligand docking and the signal pathway flow. We expect this information to be useful in the basic research of drug design and for understanding living systems.

Our Receptor Database is open to the public. Access is not restricted by any firewall. By installing a free

visualization tools, such as Chime (Martz, 2002), the user can look at a three-dimensional image of the protein. No other tools are needed to look at the information in IRDB. The waiting time may be long for some sites, and sometimes there may be bad connections to the Analytical sites. To improve this, we intend to have the Analytical system on our site. Because of the huge number of amino acid sequences (in PIR and Swiss Prot), DNA sequences (in GenBank) and protein structural data (in PDB), we did not store them on our computer disk, but provided links to the original Web sites.

The whole IRDB system is constructed of many in-house sub-systems and independent systems. We do not intend to provide the software itself, because maintaining whole system would be very complicated and difficult.

4 ACKNOWLEDGMENTS

We appreciate for useful discussion with Dr. E. Wingender (GBF, Braunschweig), Dr. N. A. Kolchanov (ICG, Novosibirsk) and Dr. H. Toh (BERI, Osaka) on this subject. We are also grateful to those who provided sites on the Internet, which are relevant to our study. We also thank Mr. M. Hayakawa for PERL programming and Ms. S. Hasegawa for painstaking data input. This work was partly supported by Science Research Promotion Fund from Science and Technology Agency from 1996 to 1998.

5 REFERENCES

ACEDB (n.d) Homepage of A Caenorhabditis elegans DataBase. Available from: http://www.acedb.org/

Allinger, N. L. (1977) Conformational Analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc. 99,* 8127-8134.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped
BLAST & PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25*, 3389-3402.

BLAST (n.d) Available from *NCBI, NIH* Web Site: http://www.ncbi.nlm.nih.gov/BLAST/

BCM Search Launcher (n.d) Available from *Baylor College of Medicine HGSC* Web site: http://searchlauncher.bcm.tmc.edu/

Brown, N. P., Leroy, C. & Sander, C. (1998) MView: a web-compatible database search or multiple alignment viewer.
Bioinformatics 14, 380-381

Drews, J. (1998) Technological Advances and a Paradigm Shift in Drug Research, In Drews, J. *In Quest of Tomorrow's Medicines* (pp 72-82), New York: Springer-Verlag

Dunham, I., Durbin, R., Thierry-Mieg, J. & Bentley, D. R. (1994) Physical mapping projects and ACEDB. In Bishop,
M. J. (Ed.), *Guide to Human Genome Computing* (pp.111-158), London: Academic Press

JAN (Japanese Accepted Names for Pharmaceuticals) Homepage of
Available from DCBI, NIHS (Japan) Website: http://moldb.nihs.go.jp/jan/index.html

Kaminuma, T., Takai-Igarashi, T., Nakano, T. & Nakata, K. (2000) Modeling of Signaling Pathways for Endocrine
Disruptors, *BioSystems. 55,* 23-31.

Kaminuma, T., Nakata, K., Nakano, T. & Takai-Igarashi, T. (2001) Pharmacoinformatics Infrastructure for

Genome-
based Personalized Medicine. *Chem-Bio Informatics Journal, 1,* 1-17.

Kolchanov, N. A., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Stepanenko, I. L., Merkulova, T. I., Pozdnyakov, M. A., Podkolodny, N. L., Naumochkin, A. N. & Romashchenko, A. G.. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002, *Nucleic Acids Res, 30,* 312-317.

Martz, E. (2002) Protein Explorer. Easy yet powerful macromolecular visualization. *Trends Biochem Sci. 27,* 107-109.

Martz (n.d) *Chime Resources*. Retrieved from the University of Massachussets Website: http://www.umass.edu/microbio/chime/

Nakata, K., Takai, T. & Kaminuma, T. (1999) Development of The Receptor Database (RDB): Application to The
Endocrine Disruptor Problem. *Bioinformatics. 15,* 544-552.

Nakata, K., Takai-Igarashi, T., Nakano, T. & Kaminuma, T. (2001a) Extension of The Receptor Database (RDB).
Chem-Bio Informatics Journal, 1, 115-119.

Nakata, K., Tokunaga, M., Toda,K., Takai-Igarashi, T. & Kaminuma, T. (2001b) An Agent System for Collecting SNPs
Data on the Internet. *Chem-Bio Informatics Journal, 1,* 120-123.

Smith, R. F., Wiese, B. A., Wojzynski, M. K., Davison, D. B. & Worley, K. C. (1996) BCM Search Launcher—an
integrated interface to molecular biology data base search and analysis services available on the World Wide Web.
Genome Res. 6, 454-462.

Sayle, R. A. & Milner-White, E. J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem Sci, 20,* 374-376.
http://www.umass.edu/microbio/rasmol/index2.htm

Stein L. D. & Thierry-Mieg, J. (1998) Scriptable access to the caenorhabditis elegans genome sequence and other
ACEDB databases. *Genome Res. 12,* 1308-1315.

Takai-Igarashi, T., Nadaoka, Y. & Kaminuma, T. (1998) A database for cell signaling networks. *J. Comp. Biol. 5,*
747-754

Takai-Igarashi, T. & Kaminuma, T. (1999) A pathway finding system for the cell signaling networks database. *In Silico*
Biology, 1, 129-146

USP-NF (United State Pharmacopoeia – National Formulary) Homepage of U.S. Pharmacopeia.
Available from: http://www.usp.org/

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Drull, M., Matys, V., Michael, H., Ohnhauser,
R., Prus, M., Schachere, F., Thiele, S. & Urbach, S. (2001) The TRANSFAC system on gene expression regulation,
Nucleic Acids Res, 29, 281-283.