Pharmacoinformatics Infrastructure for Genome-based Personalized Medicine

Tsuguchika Kaminuma[1], Kotoko Nakata[1], Tatsuya Nakano[1],
Takako Takai-Igarashi[2]

[1]Division of Chem-Bio Informatics, National Institute of Health Sciences, 18-1, Kamiyoga 1-chome,
Setagaya-ku, Tokyo 158-8501 Japan
[2]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai,
Minato-ku, Tokyo 108-8639 Japan

kaminuma@nihs.go.jp

(Received January 9, 2001; Accepted January 20, 2001)

Synopsis

One of the most important areas for the application of genomic science and technology is said to be personalized medicine (which is often called tailor-made medicine in Japan). Personalized medicine is an idealized medical practice aiming to give the right drugs to the right patients at the right times. It has been widely admitted that many projects for finding Single nucleotide polymorphisms (SNPs) are the basis for such practice. However, such knowledge alone is by no means sufficient, for good practice must be supported by well trained medical professionals who can easily access a wide range of relevant data and knowledge at their clinical sites. New informatics are also needed in order to utilize these data and knowledge effectively. Such an informational environment, i.e., data and knowledge bases and computational tools, would be called the infrastructure for personalized medicine. The infrastructure would also be useful in pharmaceutical research for finding leads and analyzing detailed mechanisms of drug actions. As the only national institution for pharmaceutical research in Japan, we have started to implement some components of the infrastructure and put their prototypes on the Web. Further research initiatives are being discussed in the Chem-Bio Informatics Society as components of its Grand Challenge Projects.

Key Words: pharmacoinformatics, personalized medicine, SNPs, ADME

1. Introduction

Application of genomic science have succeeded in unveiling the complete genomes of many micro organisms, model organisms, and human [1-1], [1-2], [1-3]. Genome science and its related technologies are revolutionizing many fields of medical as well as pharmaceutical sciences and practices. One of the most important targets for application of genomic science and technology is said to be personalized medicine (which is often called tailor-made medicine in Japan).

Personalized medicine is the idealized medical practice aiming to give the right drugs to the right patients in the right amounts at the right times. Good practices must be supported by sound relevant data and information in addition to the technical skills of medical professionals including physicians, nurses, pharmacists, and other paramedical staff. Drugs must be chosen based on a fine differential diagnosis, and the dosage amounts should be determined based on the pharmacokinetic and pharmacodynamic data taking polymorphisms of each patient's drug metabolic enzymes and transporters into consideration. Computer-based pharmacokinetic simulation will become a routine means for adjusting drug dosage amounts by monitoring various clinical markers. Therefore, finding Single nucleotide polymorphisms (SNPs) for drug target biomolecules and drug metabolism enzymes is considered as a premise for personalized medicine; however, it is by no means the sufficient condition.

In short, personalized medicine requires more extensive and sophisticated data and knowledge on drugs and their effects on patients. What kinds of databases, knowledge bases, and simulation tools are needed for such practice, how to develop these systems, and how to utilize these resources effectively for clinical practice are the subjects of this paper.

The starting point of our consideration is the information on registered drugs in advanced countries: Europe, USA, and Japan (for example, British Approved Names (BAN), United States Adopted Names (USAN), Japanese Accepted Names (JAN)). These countries have signed a mutual agreement, ICH (International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use), on import and export of their drugs. The WHO has organized a working committee to harmonize the names of these drugs, which is called the International Nonproprietary Names (INN). Each country provides a drug name list with associated identification information such as CAS (Chemical Abstract Service) Numbers, molecular formulas, and chemical structures. Unfortunately, not all of this fundamental information on drugs is yet fully computerized either in Japan or in other countries. However, discussion is ongoing to computerize the lists and to add more relevant information such as drug target and ADME (Absorption, Distribution, Metabolism, and Excretion) data.

Thus it was considered that the making of a complete list of the names of all the drugs that have ever been registered in Japan, with their associated ADME data and their target molecule information is most important to provide a fundamental information source for good practice of drug regimen.

These data and knowledge are also considered as the platform for pharmocokinetics and pharmacodynamics studies. With these data and knowledge a pharmaceutical researcher can examine in silico how drugs are metabolized, bind to their main target molecules, and stimulate signals that affect either non-genetic bioreactions or gene expressions. Adverse effects and drug-drug interaction can be studied at the molecular level, and prediction and prevention of these effects may become more accurate. Moreover these data and knowledge are also useful in the search for new lead compounds, for the existing drugs are good paradigms for drug-like compounds. Thus structure databases for accepted drugs are good source for drugablity analysis.

We can thus conclude that if we make complete databases for existing drugs, link these databases to drug target molecule databases and ADME databases, and further link them to bioreaction pathway databases, the resultant linked systems may form the fundamental platform or the infrastructure for pharmaceutical research and personalized medicine. Of course to utilize these data and knowledge effectively, we should provide various computational tools some of which already exist and some of which should be newly developed.

Taking this conclusion as the working hypothesis we started to implement some of these systems at the National Institute of Health Sciences (NIHS), which is the only national institution for pharmaceutical research in Japan. Our efforts started from providing a complete catalog of drugs that had been given accepted (which is also called approved or adopted in other countries) names in Japan (JAN).

This paper introduces the concept and birds eye view of the pharmacoinformatics infrastructure, its prototype component systems that we are developing, and examples of relevant systems and tools developed by other research groups.

2. Overview of the pharmacoinformatics infrastructure

Figure 1 shows the skeletal components comprising the pharmainformatics infrastructure and their relations [2]. Each component corresponds to a database, knowledge base, or computational tool. The central databases are the complete catalog databases of drugs that are accepted by the governments of advanced countries including Japan.

The ADME Knowledge Base is the database for absorption, distribution, metabolism, and excretion of drugs. Cytochrome P-450 (CYP) is one of the most well known drug metabolism isozymes that have polymorphisms. Another well known enzyme is N-acetyltransferase, and nearly a hundred enzymes have been identified as drug metabolism enzymes. Transporters such as ABC transporters are responsible for transporting (pumping in and out) the drugs in the body. These data and knowledge are important to know how drugs are distributed to various tissues and organs and how they are chemically modified. Thus they are essential to estimate the bioavailability of the drugs, to calculate reasonable amounts for administration, consider adverse effects, and predict drug-drug interactions.

The Drug Target Database includes the information on the primary targets and modes of action of the drugs. Binding to the targets are the most important characteristics of the drugs that lead to their expected actions. If the targets are membrane receptors, signals caused by drug binding may be transmitted through the membrane and then passed to the transcriptional machinery by the head on molecular signal transductions [3]. These transcriptions cause synthesis of proteins that take some further actions. If the drug targets are metabolic enzymes, drugs binding to these enzymes cause some perturbation in metabolic reactions and may control the reaction chains to more desirable directions.

Drug target information was not well provided in former days. So there still are many even well known drugs, such as aspirin, whose targets have not yet been so clearly identified. Metabolic pathways were well studied even before the development of sequence technology or recombinat DNA technology [4]. CYP data have also been accumulated for more than two decades. Signal pathway data have been accumulated for only a decade or two. In any case, the success of many genome sequence and analysis projects and their successor projects such as determination of SNPs, detection of gene expression by microarrays (transcriptomics), proteomics, structural genomics, protein-protein interaction analysis, and pathway finding are now vastly enhancing those accumulated data and knowledge. In short, it is a good time to start to catalogue these data and knowledge systematically and edit them into the forms of data or knowledge bases.

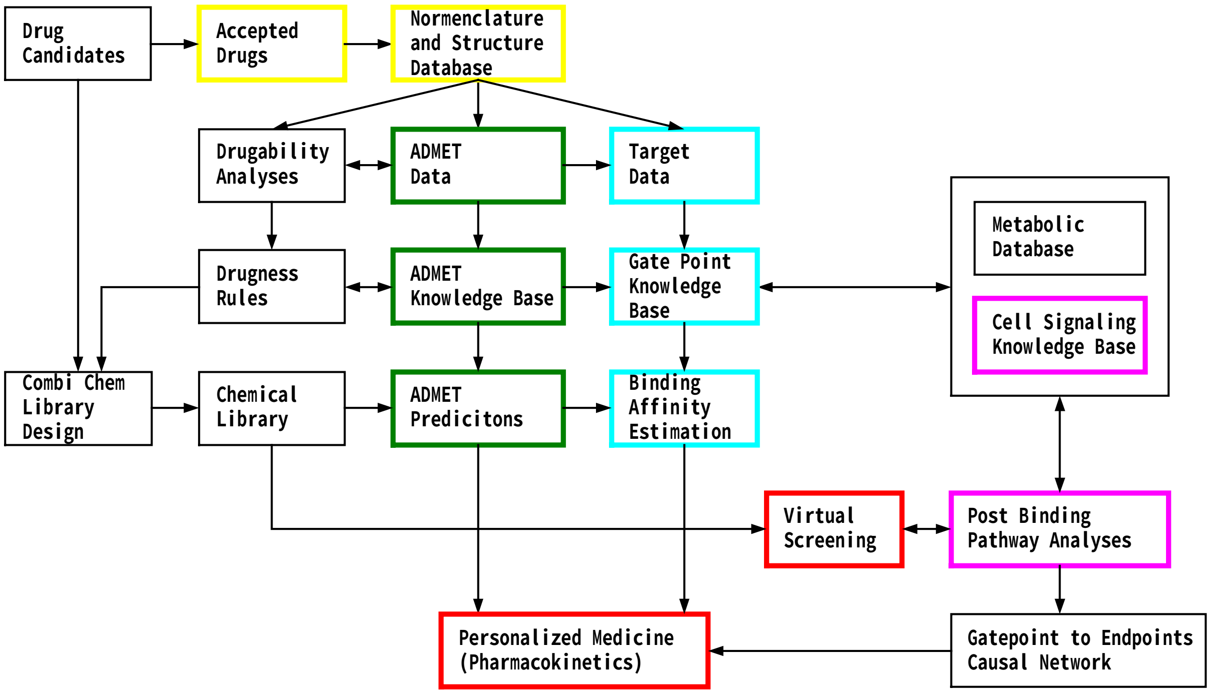

Drug Candidates
Accepted Drugs
Normenclature and Structure Database
Drugability Analyses
ADMET Data
Target Data
Metabolic Database
Drugness Rules
ADMET Knowledge Base
Gate Point Knowledge Base
Cell Signaling Knowledge Base
Combi Chem Library Design
Chemical Library
ADMET Predicitons
Binding Affinity Estimation
Virtual Screening
Post Binding Pathway Analyses
Personalized Medicine (Pharmacokinetics)
Gatepoint to Endpoints Causal Network

Figure 1. Overview of Pharmacoinformatics Infrastructure. The figure shows the components for the infrastructure. The colored boxes (except red ones) corresponds to the components for which the authors are implementing some prototype systems. The infrastructure will be relevant for designing more efficient virtual screening systems and a better personalized drug regimen (red boxes).

Figure 1 shows that the eventual goal of developing these data and knowledge bases and their associated computational tools is to enable virtual screening and personalized medicine. Virtual screening is a computational system to discover drug candidate compounds (leads) from a large random collection of chemicals. For this purpose, one must first make a model of the fixed receptor or the target protein molecule for a (potential) drug and then calculate binding affinity between the target molecule and the screening compound by a docking model. Such screening based on a docking model is often called in silico assay. The problem for docking studies is to identify the model receptor (the target) and select good candidate compounds. Pharmaceutical companies have their own libraries of chemicals for screening, and there are many commercial vendors who supply these compounds, as well. Combinatorial chemistry (Combi Chem) is the use of a process to prepare large numbers of structurally diverse sets of organic compounds by combining sets of chemical building blocks (called monomers) often in every possible combination. How to design such a library is a challenging theme for drug designers and computational chemists. Several methods have been proposed to utilize the existing drugs as good paradigms for elucidating characteristic features of new drugs. This means that the databases of the existing drugs are not only relevant for rational drug regimen but also for developing compound libraries for efficient drug screening.

Until recently, the docking study was the goal of the rational drug designers. But things are changing very drastically, for vast amounts of new knowledge on signal pathways and transcription mechanisms and their variations due to genetic polymorphisms are allowing researchers to examine in more detail what is going on after the drug molecules bind to their primary targets. It may not be exceptional that a drug has more than a single type of target, and there may be cross-talks between signals triggered from the multi-target-hits. It becomes very realistic to start thinking about tracing the causal network of drug actions from the multi-target bindings to the endpoints by combining the data and knowledge on metabolic and signal transduction pathways, transcription factors, and cell-cell communications. Undoubtedly, diagnosis by knowing the genotypes of the proteins involved in this causal network is useful for personalized medicine.

There already exist information and computing resources that correspond to some of these infrastructure components. Yet some essential components such as JAN or the ADME database had not been developed in Japan, and the integration of the component systems for drug research or personalized medicine has not been carried out even in Europe or the United States. In the next section, we will review the existing relevant systems and the systems that we have developed.

3. Component Systems

3.1 Drug candidate database

There already exists a number of chemical substance databases for drug research. Pharmaceutical companies keep huge collections of compounds for screening in their in-house compound libraries. They also have the ability to screen several hundred thousands to several millions of chemicals within days or weeks. There are many commercial out-sourcing vendors who supply large volumes of compound libraries. A well known commercial database system that also contains compound libraries is ISIS, which contains a public testing drug database from the National Cancer Institute (NCI).

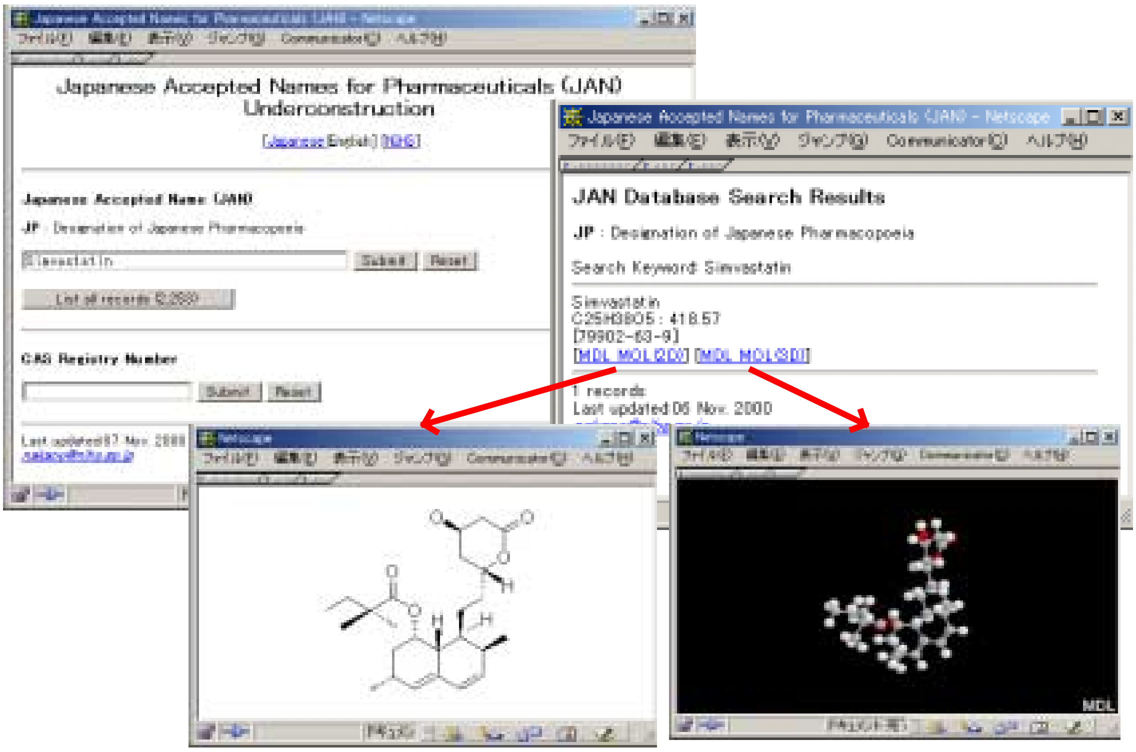
A proposal was made to the Chem-Bio Informatics (CBI) Society from a researcher from one of its member companies to make an information exchange database for chemicals synthesized and maintained by academic researchers. The idea, which is to let university and government institute researchers register the interesting chemicals they synthesized on the CBI WWW, is under consideration. Similar ideas may apply for natural chemicals and medicinal plants. For example, Satake of NIHS is now developing a medicinal plant resource database on the Web (http://www.nihs.go.jp/dpp/dppdb/).

3.2 The Digital JAN

Advanced countries provide official documents that specify the drugs that have been formally accepted and may have the possibility of being used. These documents are called phamacopeias. Examples are the Japanese Pharmacopoeia, US Pharmacopoeia, British Pharmacopoeia, and European Pharmacopoeia. In addition, each government provides more comprehensive, yet simple, name list documents for registered drugs. These are JAN, USAN, BAN, and WHO/INN. These documents are also supplied on media such as CD-ROMs and/or on Web sites, but not yet by databases.

Collaborating with N. Miyata of NIHS we are developing a nomenclature and structure database for the drugs in JAN [5]. This database which is called the Digital JAN, contains the following data items: INN, Japanese Accepted Names, CAS No, chemical (IUPAC) names, chemical structures, three dimensional coordinates, and administrative information such as registered data. The three dimensional coordinates are either obtained from CSD (the Cambridge Structural Database for X-ray crystallography) or by model calculation based on the ab initio MO method (Gaussian 98).

Data contents for the Digital JAN Database were first edited on PC-based ACCESS and have already been put on our Web server for test usage (See Figure 2). It is planted to include all JAN information in this database by the end of March 2001.



Japanese Accepted Names for Pharmaceuticals (JAN)
Underconstruction

[Japanese English] [NIHS]

Japanese Accepted Name (JAN)

JP : Designation of Japanese Pharmacopeia

Simvastatin

CAS Registry Number

Last updated 07 Nov. 2000

JAN Database Search Results

JP : Designation of Japanese Pharmacopoeia

Search Keyword: Simvastatin

Simvastatin
C25H38O5 : 418.57
[79902-63-9]
[MDL MOL(2D)] [MDL MOL(3D)]

1 records
Last updated 06 Nov. 2000

Japanese Accepted Names for Pharmaceuticals (JAN) - Netscape
ファイル(F) 編集(E) 表示(V) ジャンプ(G) Communicator(C) ヘルプ(H)
Submit
Reset
List all records
Netscape
H
O
MDL

Figure 2. The Digital JAN. A nomenclature and structure database has been developed based on the Japanese Accepted Names for Pharmaceuticals documents. The database has links to the ADME Knowledge Base and The Drug Target Database.

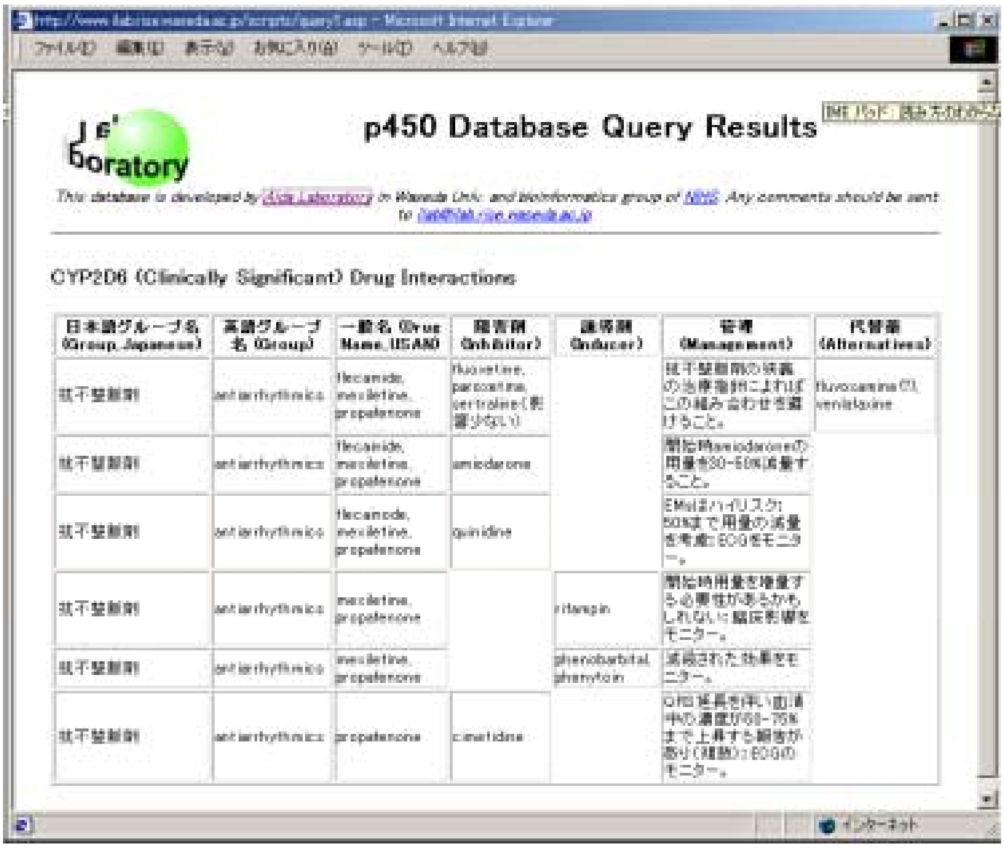
3.3 The ADME Knowledge Base

The ADME data are important for predicting the so-called poor metabolizers and excess metabolizers. They are also important for predicting adverse effects due to drug (or food and drug) interactions so that they may be prevented. Although there are many enzymes and transporters to cover, we are limiting our prototype database, the ADME Knowledge Base, to contain cytochrome P-450 (CYP) data.

By surveying literature [6] and Web sites, we have edited the collected CYP data for each drug and isozyme of CYPs into tables. Basic data are the relationships between drugs and CYP isozyme members.　For each CYP isozyme we have identified drugs that induce this isozyme, for each drug we have also identified CYP isozymes that metabolize the drug and the drugs that inhibit this enzyme action.　These relationships are easily represented by tables, which are formed into a database. This prototype system has been implemented by our collaborators and put on the Web (http://www.ilab.rise.waseda.ac.jp/).

It is straightforward to build a simple consultation system that predicts or warns of drug adverse effects and drug interactions based on this database.　Such a database/consultation system is already on the Internet, but it is limited to only US drugs.

Some efforts are being carried out to collect more experimental data for CYPs and drug-drug interactions, by the HAB consortium and the Drug Interaction Database Research Group in which researchers from pharmaceutical companies play important roles, for example.　We have contacted with these groups and planned to collaborate to provide their data on the open databases on the Internet.


http://www.ilab.rise.waseda.ac.jp/scripts/query1.asp - Microsoft Internet Explorer
ファイル(F)　編集(E)　表示(V)　お気に入り(A)　ツール(T)　ヘルプ(H)
Laboratory
インターネット

p450 Database Query Results

This database is developed by Aida Laboratory in Waseda Univ. and bioinformatics group of NIHS. Any comments should be sent to [illegible]

CYP2D6 (Clinically Significant) Drug Interactions

日本語グループ名 (Group_Japanese)	英語グループ名 (Group)	一般名 (Drug Name_USAN)	阻害剤 (Inhibitor)	誘導剤 (Inducer)	管理 (Management)	代替薬 (Alternatives)
抗不整脈剤	antiarrhythmics	flecanide, mexiletine, propafenone	fluoxetine, paroxetine, sertraline(影響少ない)		抗不整脈剤の併用の治療指針によればこの組み合わせを避けること。	fluvoxamine(?), venlafaxine
抗不整脈剤	antiarrhythmics	flecanide, mexiletine, propafenone	amiodarone		開始時amiodaroneの用量を30-50%減量すること。	
抗不整脈剤	antiarrhythmics	flecanide, mexiletine, propafenone	quinidine		EMsはハイリスク: 50%まで用量の減量を考慮; ECGをモニター。	
抗不整脈剤	antiarrhythmics	mexiletine, propafenone		rifampin	開始時用量を増量する必要性があるかもしれない; 臨床影響をモニター。	
抗不整脈剤	antiarrhythmics	mexiletine, propafenone		phenobarbital, phenytoin	減弱された効果をモニター。	
抗不整脈剤	antiarrhythmics	propafenone	cimetidine		QRS延長を伴い血清中の濃度が50-75%まで上昇する報告があり(稀); ECGのモニター。	

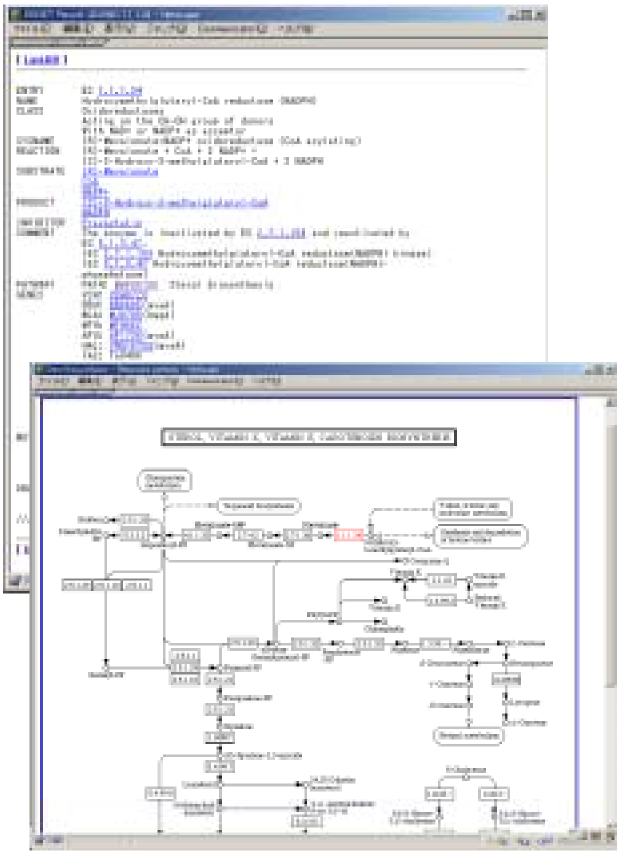
Figure 3. The ADME Knowledge Base.　A simple model database was implemented for CYP data by H. Ochiai at Waseda University.

3.4 The Drug Target Database

A drug or its metabolite binds to some of the biomolecules that trigger a series of reactions and end up with the effects aimed at by the prescriber. Although these target molecules are not yet identified for all the drugs, it was estimated that nearly half of them are receptors, one fourth of them are enzymes, and a few of them are transcriptional factors [7]. Nevertheless, target molecules are not yet fully identified nor described in the explanation documents for all drugs.

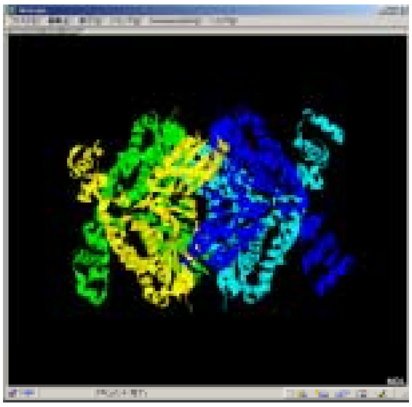
We are planning to survey target molecules for all drugs in the basic drug database (the Digital JAN). Since it may take time we decided to work first on some paradigmatic drug receptors and the drugs that bind them. These drugs and targets include the serotonin receptor and serotonin-related drugs, opioid receptors, ion channel receptors, some metabolic enzymes like HMG CoA reductase, CYP aromatases, and protease inhibitors, some cell signal transduction related protein kinase inhibitors, some nuclear receptors including estrogen receptors, transcription factors, DNA abducts and DNA intercalaters, and antibiotics.

The Drug Target Knowledge Base has been implemented mostly for receptors. Enzyme targets are still under development, however, detailed knowledge was accumulated for an important nuclear receptor, i.e., estrogen receptors. Target molecules are linked both to the Receptor Database (RDB; http://impact.nihs.go.jp/RDB.html) [8] and Cell Signaling Network Database (CSNDB; http://geo.nihs.go.jp/csndb) [9]. A SNPs data collector was implemented on the CYP data and is being tested for receptor proteins in the Receptor Database.

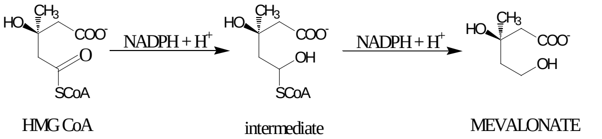


[LinkDB]

```
ENTRY       EC 1.1.1.34
NAME        Hydroxymethylglutaryl-CoA reductase (NADPH)
CLASS       Oxidoreductases
            Acting on the CH-OH group of donors
            With NAD+ or NADP+ as acceptor
SYSNAME     (R)-Mevalonate:NADP+ oxidoreductase (CoA acylating)
REACTION    (R)-Mevalonate + CoA + 2 NADP+ =
            (S)-3-Hydroxy-3-methylglutaryl-CoA + 2 NADPH
SUBSTRATE   (R)-Mevalonate
PRODUCT     (S)-3-Hydroxy-3-methylglutaryl-CoA
COMMENT     The enzyme is inactivated by EC 2.7.1.109 and reactivated by
PATHWAY     Sterol biosynthesis
```

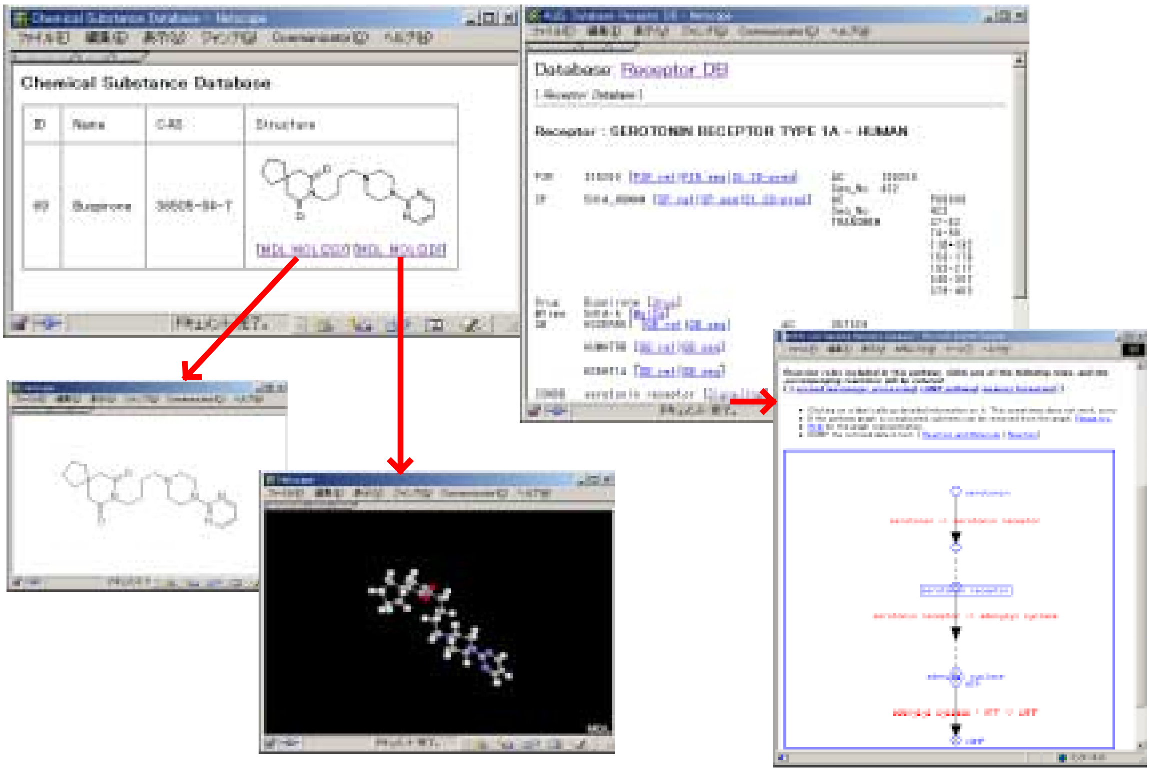


Human HMG-CoA reductase (1DQA.pdb)



[Reaction catalyzed HMG CoA reductase]

Figure 4. HMG CoA reductase information in the Target Molecule Database. There are links to the metabolic pathway database KEGG and PDB.



Chemical Substance Database

ID	Name	CAS	Structure
49	Buspirone	36505-84-7	[MDL MOLCODE] [MDL MOLCODE]

Database: Receptor DB

[Receptor Database]

Receptor : SEROTONIN RECEPTOR TYPE 1A - HUMAN

Figure 5. The Cell Signaling Networks Database (CSNDB) linked to the Recepter Database. These two databases are experimentally linked together so that one can trace signal flows from the substance.

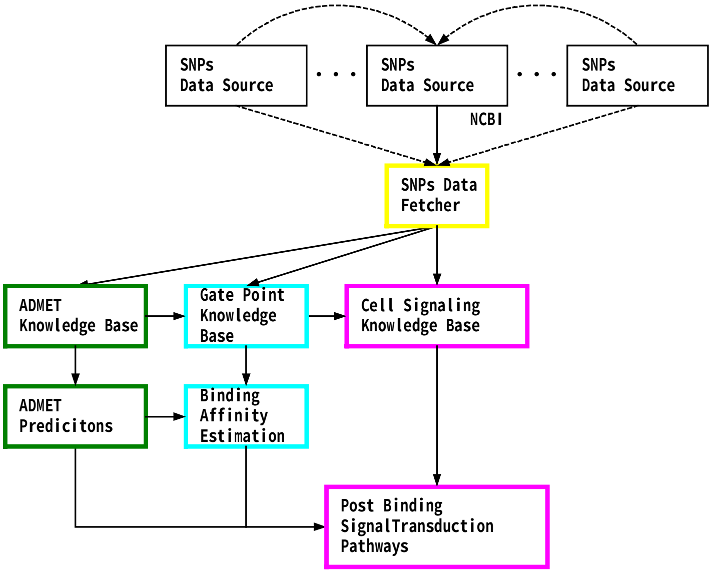

SNPs Data Source
SNPs Data Source
SNPs Data Source
NCBI
SNPs Data Fetcher
ADMET Knowledge Base
Gate Point Knowledge Base
Cell Signaling Knowledge Base
ADMET Predicitons
Binding Affinity Estimation
Post Binding SignalTransduction Pathways

Figure 6. SNPs Data Collection System. The figure shows the concept of the interface system to collect SNPs data for registered proteins (CYPs and drug target molecules) from open data sources of SNPs. Since NCBI collects these data from distributed sources, the current system takes data only from the NCBI.

3.5 The Binding Affinity Database

The affinity of a drug to the target molecule is the basis for elucidating its principal actions. Affinity can be measured by in vitro binding experiment of ligands to the receptors. Ideally these data should be linked to every target molecules in the Target Molecule Database. However, it is laborious work to survey experimental data for binding affinities, adjust their values so that they are comparable with each other, and put them into a database. We have so far developed such a database only for the estrogen receptors. Here affinity values were adjusted to relative values taking the natural estrogen value as the reference.

3.6 Gate Point Knowledge Base and Analysis

One of the eventual goals of theoretical drug designers is to make a good model for drug-target interactions, and to analyze the interaction by calculations based on some molecular computational method. For that they should have good experimental background data on target protein molecular structure, binding modes and affinities with various ligands, and powerful computers to perform the calculation.

The Target Molecule Database and the Binding Affinity Database provide relevant information for the docking studies, but they are not enough for modelers. The most crucial data are fine structures of the ligand-receptor complex obtained either by X-ray crystallography or by NMR analysis. Although there exist many proteins which are considered to be potential drug targets, not many of their structures have been determined. Thus theorists, dream of elucidating the structure solely by computation; but current computational power is too weak to accomplish this goal [10]. In the meantime, the functional genomics project will determine protein structures at a faster pace [11].

Though we have not yet tried, we think it better to develop a database that supplies various background data and knowledge for docking studies taking the data in the Target Molecule Database and the Binding Affinity Database into consideration. This database should contain such information on the target molecules as (1) the sequences, the consensus sequences, the active sites, three dimensional structure data of the molecular family members, (2) chemical structure and the three dimensional coordinate data of the ligands, and (3) the structures of ligand-target complexes with binding affinities, if any.

3.7 Pathway Databases

The modified or unmodified drug molecules are transported to the target organs and tissues by the blood stream and are perceived by their cells. No matter where the molecules are captured, they may affect either biosynsthesis or gene expression of the cells.

If we can provide the complete intra cellular metabolic pathway database and signal pathway database of these cells, one can trace the effects of the drug on the cells. There already exist a number of such databases. Many of them are now commercialized, but there some reliable public databases still remain. We chose Metabolic Pathways in KEGG (Kyoto Encyclopedia of Genes and Genomes) as the reference database [12] and CSNDB as the signal pathway database reference to which the Target Molecule Database is linked. Figure 5 shows an example of such links.

These two databases are well designed and maintained as public databases, but they are still very incomplete from our viewpoint. First, both databases compile data and knowledge of a fictitious cell, a cell that does not have differential character. In reality each cell in an organism has its own character and thus pathways are different if the cell types are different. Cells are also different according to their positions in the cell lineages. Knowledge was accumulated for a typical case, but not for specific cases.

Second, the present metabolic pathway databases were developed in order to computerize standard reaction pathways which are described in the textbooks [4]. However, to trace the effects of drugs on biosynthetic reaction pathways, we need more specific and detailed reaction data. Some efforts are needed to cover relevant metabolic pathways for studying drug actions.

Third, gene expression is controlled by complex transcription machinery, whose mechanism and component molecules, the transcription factors, are not yet fully unveiled. There already exist a few transcription factor databases, but they are still in the developmental phase. CSNDB was experimentally linked to one of these databases, TRANSFAC [13], but both contents must be greatly enriched in the future.

There is a hope, however, that the newly rising technology, the microarray (DNA chips) or more probably the yeast two-hybrid, may contribute to identifying genes that are responsible for the effects of drugs. This is one of the hot spots of current research. Chip technology is being used to clarify the concept of "gene network. " Interaction between genes are not only controlled by their upstream and downstream relations, but they are indirectly controlled by their products (proteins) via metabolic and signal pathways, their complicated cross talks, and the feed back loops.

This fact suggests that sooner or later the current databases or knowledge databases of pathways and gene networks should be linked to each other or should be integrated into a more inter-related complex system.

3.8 The Pharmacokinetic Database

The gist of personalized medicine is to identify genotypes of drug metabolic enzymes and transporters and utilize those data for drug regimen. In real clinical practice one needs more quantitative data for estimating drug concentrations, enzyme activities, and their interactions in various compartments. Additional parameters such as absorption and excretion (clearance) of these chemicals into and from each compartment are needed for pharmacokinetic simulation calculations. Y. Sugiyama and others, including pharmaceutical industry researchers, are collecting these data and trying to put them into a database [14].

4. Integration of Component Systems

4.1 Linkage between Component Systems

We have tried to link some component systems by the WWW hyperlink technique. Currently we are planning to link the Digital JAN, the Target Molecule Database, the ADMET Knowledge Base, the metabolic pathway database (KEGG), and CSNDB, which was already linked to TRANSFAC. The Binding Affinity Database remains independent. Except for the metabolic pathway database all systems mentioned in this paper are implemented on our unix server machines, which are connected to the internal network.

In the future, we expect that development and maintenance of the component systems will be carried out by research groups at multiple sites on the Internet. Such an extension of the project is straightforward from hardware and software view points.

4.2 Embedding SNPs Data

Although SNPs information is presently being accumulated at a rapid pace, it is difficult for individual researchers to keep collecting the most updated information. We have tried to develop an agent system for collecting existing SNPs data on the Internet.

The agent system is an interface program that fetches SNPs data for those biomolecules that are preassigned in the system from public databases such as NCBI or Swiss-Prot. We have tried to test this system for collecting SNPs data for receptors in the Receptor Database and for CYP isozymes in the ADME Knowledge Base.

The system successfully brings in domain-specific and regularly updated SNPs information, and we find it will be generally useful to those doing genome-related research.

Unfortunately, the presently publicly available SNPs are lacking in detailed information on the clinical conditions of the patients or the volunteers who supplied the samples. Moreover, terminologies for describing the clinical conditions should be controlled by some standardized terminology (thesaurus) which is under the control of some committee. No such system yet exist in Japan.

4.3 Public Open Usage vs Private Closed Usage

We believe that data or knowledge contained in most of the component systems of the infrastructure should be open to the public and available at almost no charge; for these data and knowledge are vitally important for citizens, consumers, and medical service clients. However, the computer systems on which these knowledge and databases are operated may be commercialized and charged, otherwise it is very difficult to develop user friendly systems. Moreover some commercial companies may put some of these open data and knowledge into their in-house systems for closed usage by the pharmaceutical industry.

One practical approach to cope with this discrepancy is to separate the contents such as data and knowledge from systems that operate the data and knowledge. The raw data and knowledge should be opened to the public, while the system would be kept closed.

5. Applications

5.1 Drugability Analysis and Combinatorial Chemistry Designing

In recent years, the sources of drug leads and screening methods in the pharmaceutical industry have changed dramatically. Combinatorial chemistry allowed efficient automatic synthesis of massive numbers of compounds for screening, and high throughput screening (HTS) systems have enabled the screening of several hundreds thousands of compound in few days or a week. How to design the screening (Combinatorial) library became an important research topic to which a theoretical approach seems promising.

N. Hirayama pointed out that the drugs on the market are good paradigms, when we think about new leads and showed the usefulness of analyzing the JAN to extract relevant features as drugs [15]. C. A. Lipinski studied same topic and formulated the "rule of 5", whereby poor absorption or permeation is more likely when there are more than 5 H-bond donors, 10H-bond acceptors, the molecular weight is less than 500, and the calculated Log P is greater than 5 [16]. Such knowledge is now referenced when one designs a candidate compound library.

These works proved that the database of accepted drugs has usefulness for finding characteristic features of even new drugs. Since our Digital JAN will be updated regularly, it will be a good source for computational studies for drugability analysis.

5.2 Binding Analysis by Computation

The docking study examines the binding forces and reactive sites of ligands, and makes predictions on the binding structure and their energy. Although it is difficult to model the complex structure of the ligands and target molecules (proteins) de novo, we can elucidate the structure by computation if we have experimental data for the structures of some ligand-target complexes.

Recently, we have tried to elucidate the docking of estrogen alpha and beta receptors with estrogen and estrogen-like ligands, which are suspected of being the so-called endocrine disruptors. With a new computational tool called the Fragment Molecular Orbital Method; we get good correlation between calculated binding energies and experimental binding affinities. FMO method, which was developed by K. Kitaura and T. Nakano, is an highly exact approximation method for the ab initio MO Method [17-1], [17-2], [17-3]. These calculations were carried out by running several hours on a supercomputer. Similar calculations would be possible for many combinations of drugs and their target molecules.

5.3 Virtual Screening

Currently, what people call virtual screening is in fact mere docking simulation by computer. In real applications, the number of chemicals subjected for screening is on the order of several hundred thousand or more. Thus computation must be fast enough to cover this number of chemicals. There are a number of such docking simulation programs: Dock, ADM, and DOT.

Although fast, these simulations are dependent on empirical parameters that vary with the expertise of the researcher. On the contrary, docking studies based on FMO method do not use any empirical parameters. When the computers become fast enough to carry out simulation calculations based on FMO method, the screening power will be vastly improved.

However, there still remains the problem to obtain reasonably fine structures of the target molecules. For that we should have a tool to model the three dimensional structure of proteins from their sequences. IBM is developing peta flops machines for protein folding problem under the code name of Blue Gene [18]. Blue Gene machines will be available within five years. From that time de novo modeling of three dimensional protein structures will become a realistic research subject.

5.4 Gate Point to Endpoint Causal Analysis

Binding to its receptor (target molecules) is the starting point for a drug’s action. No matter where cells perceive the drug signal, they are transduced to perturb some reaction or gene expression, the effects which may cascade, cross-talk with each other, and feedback to trigger bigger effects in the cell and on the other cells by cell-cell or tissue-tissue communication. This is a very complicated causal network. We have tried to challenge this theme by using the endocrine disruptors problem as a model [19].

One of the famous riddles of endocrine disruptors is the effects of the phytoestrogens, genistein or daidzein, for example. Such chemicals bind to estrogen receptors as do other environmental pollutants, yet epidemiologists found that eating soy beans contribute to the prevention of breast cancer. To explain this controversy, we assume that phytoestrogens have multi-gates, that is they may stimulate cells in multiple ways, many of which are not yet unveiled. Such analysis can only be carried out by combination of yeast two-hybrid, microarray and proteomics technologies.

5.5 Personalized Medicine

We have already introduced some of the component systems necessary to realize personalized medicine. One important premise for personalized medicine is that it is possible to know patient genetic profiles, which will be more easily obtained by the new DNA diagnosis chips or SNPs profiling in the future. By referring to these diagnostic data with the data and knowledge bases in the infrastructure, physicians may be aided in making good decisions to choose the right drugs for right patients.

The required new data and knowledge sources for good practice are the drug target, the binding affinity, ADMET(ADME plus Toxicity), gene networks, signal pathways, and protein-protein interactions. These resources give the clinical experts good insight, not only for the drug target and how to bind to it, but also for the post-binding signal transductions, transcription factors, and endpoints. Unfortunately, these data or knowledge bases have not yet been directly linked to the real drugs that are used in today's clinical settings in Japan. The infrastructure we are proposing will overcome this problem.

The problem still remains on how to give the right amounts of drugs at the right times; for that, more pharmacokinetic data are needed. The Pharmacokinetic Database will supply some relevant information. However, the states of the patients change from time to time, and clinical monitoring data are needed to adjust the amounts of drugs or even change the drugs. Such judgments require expertise and skills.

In any case, in order to realize effective personalized medicine, data and knowledge in the pharmacoinformatics infrastructure should easily be accessible from any clinical site. In addition, physicians and other clinical staff must be trained on how to access these relevant information bases and how to understand and make use of them.

It is also desirable to have good communications between the physicians who are often using certain drugs and the clinical pharmacists who are interested in studying that type of drug. If such communications become routine, drugs can be well traced from their birth to after the market. In this Internet age, it is quite possible to build the communicational infrastructure connecting those who provide the resource with those who have the knowledge and skill and those who perform services.

6. Futurescope

Full scale implementation of all the components shown in Figure 1 is likely to be beyond the scope of any single research group. In fact, these components are of heterogynous character. Many of them are precompetitive in their nature, but some of them are very competitive. It may be that the government or a US-FDA like government institution should provide the accepted names databases.

Pharmaceutical companies are required to attach relevant information to the packaging of each drug on the market, but such information is not so easy to use nor sufficient for scientific reasoning. It is very time-consuming, painstaking, and costly work to add new data and information and to edit those sources into a more clear-cut scientific knowledge form. Thus development of the ADME and drug target molecule database would be the task of some consortium comprising of government, academic, and industry researchers sponsored jointly by government, NGOs, and pharmaceutical industries.

There already exist many companies that are intending to sell pathway databases. However, it is more desirable if such data and knowledge bases are developed by some nonprofit consortium, for such data or knowledge which are relevant for understanding how drugs affect our bodies and cure our diseases must be open to the public particularly to medical clients. Furthermore, since these data or knowledge are indispensable for clinical practice, if they are ever charged, they should not be so expensive that consumers or clients cannot have access to them.

On the other hand, drug candidate chemical database and combinatorial chemistry library are the sources of competitive power in the pharmaceutical industries, and so are the virtual screening systems. Computational tools for estimating binding energies between ligands and receptors (or drugs and target molecules) are the by-product of general molecular computing tools. These software should be delivered in such a price range that any academic researcher, including graduate students, can easily use them. Pharmacokinetic simulation software must be of same character. Here academic researchers in universities or governmental institutions would be the main contributors of development.

Despite this heterogeneity, all of the information and computational resources are inter-related as was shown by the lines in the figure. This is one reason why we have called them the infrastructure. We have started to develop prototypes of some of the component systems. Except for the Digital JAN, our systems are still very limited in scope and much remains for further development. There also remains the problem of maintenance or the constant addition of data and knowledge which we expect to come bursting streams as the genome science advances.

We thus propose that some research association nonprofit in nature should take the initiatives for the full scale development, maintenance, and user education of such infrastructure in the near future. We consider that the CBI Society is the ideal association to take these initiatives, for CBI is made up of researchers from industries, academics, and the government and it has long history of information exchange and collaboration in computational chemistry and bioinformatics aiming at drug design [20]. We hope that the CBI Society will be able to take immediate action for at least initiating the full-scale implementation of non-competitive components of the infrastructure under collaboration with other societies like the Japan Pharmaceutical Society in the near future. We are grateful if this paper is useful for CBI to take such an initiative.

Acknowledgements

The future goals in this vision paper came into the authors minds through discussions with many researchers including Prof. N. Hirayama of Tokai University, Dr. N. Miyata of the National Institute of Health Sciences, Prof. Y. Sugiyama of Tokyo University, and Dr. T. Okabe of Banyu Pharmaceutical Co. The authors also thank to those who collaborated for implementing some of the data and knowledge bases. These includes Mr. T. Kawaide and Ms. S. Suzuki of Data Index Inc, Dr. Y. Iida of Waseda University (Pharmacia Upjohn), and our colleagues, Dr. M. Yamamoto, Ms. S. Hasegawa, Ms. H. Taki, Mr. K. Komine, and Mr. K. Koyano. The authors also thank to Ms. J. S. Miwa for wording.

References

[1] Some reports on landmark in genome sequencing are:
[1-1] *C. elegans* in *Science*, **282**, 11 December (1998);
[1-2] *Drosophila* in *Science*, **287**, 24 March (2000);
[1-3] *Arabidopsis* in *Nature*, **408**, 14 December (2000).

[2] Web Information for Figure 1 is (http://www.nihs.go.jp/DCBI/dcbi-www/Pharmainfra.htm), which is a clickable map, and each box leads to web sites related to that components.

[3] T. Hunter, *Signaling: 2000 and Beyond*, *Cell*, **100**, 113-127; January 7, (2000).

[4] A. L. Leihninger, *Biochemistry* (2nd Ed.), Worth, NY (1976).

[5] Japanese Accepted Names for Pharmaceuticals (JAN) (http://moldb.nihs.go.jp/jan/index.html).

[6] Michalets EL., *Pharmacotherapy* **18**(1):84-112 (1998); Cytochrome P450 Drug Interaction Table, The Department of Pharmacology Georgetown University Medical Center, (http://www.dml.georgetown.edu/depts/pharmacology/davetab.html).

[7] Drew, J.: *Inquest of Tommorrow's Medicine*, Springer, (1998).

[8] Nakata, K., Takai-Igarashi, T. and Kaminuma, T.: *Development of A Receptor Database Bioinformatics*, **15**, 544-552 (1999).

[9] Takai-Igarashi, T. and Kaminuma, T.: A Pathway Finding System for the Cell Signaling Networks Database, In *Silico Biology*, **1**, 129-146 (1999).

[10] Baker, D.: A Surprising Simplicity to Protein Folding, *Nature*, **405**, 39-42 (2000).

[11] Sanchez, R., Pieper, U., et al.: Structure Modeling for Structural Genomics, *Nature Structural Biology*, Structural Genomics Supplement, 986-990 (2000).

[12] KEGG (http://bioinformatics.weizmann.ac.il:3456/kegg/kegg2.html).

[13] Heinemeyer, T. et al., Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanism, *Nucleic Acicds Res*., **27**, 318-322 (1999). (http://www.cbi.pku.edu.cn/TRANSFAC/doc/toc.html)

[14] Y. Sugiyama private communication.

[15] Fujii I., et al., *CBI Journal*, **1**, 18-22(2001).

[16] Lipinski, C. A., et al: Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings, *Advanced Drug Delivery Reviews*, **23**, 3-25 (1997).

[17] [17-1]Kitaura, K., Sawai, T., Asada, T., Nakano, T. and Uebayasi, M.: *Chem. Phys. Letters*, **312**,319-324(1999).
[17-2]Kitaura, K., Ikeo, E., Asada, T., Nakano, T. and Uebayasi, M.: *Chem. Phys. Letters,* **313**, 701-706(1999).
[17-3] Nakano, T., Kaminuma, T., Sato, T., Akiyama, Y., Uebayasi, M. and Kitaura, K. ; *Chem. Phys. Letters*, **318**, 614-618(2000).

[18] IBM Blue Gene (http://www.research.ibm.com/bluegene/index.html).

[19] Kaminuma, T., Takai-Igarashi, T., Nakano, T., and Nakata, K.: Modeling of Signaling Pathways for Endocrine Disruptors, *BioSystems*, 5521-31 (2000).

[20] Kaminuma, T: CBI Grand Challenge, *Proc. CBI Millennium Symposium*, July 2000, Tokyo.

ゲノムを基礎としたパ-—ソナライズド・メディシンのための
医薬情報学的な基盤

神沼二眞[1]、中田琴子[1]、中野達也[1]、高井貴子[2]

[1]国立医薬品食品衛生研究所　化学物質情報部
〒158-8501 東京都世田谷区上用賀1-18-1
[2]東京大学 医科学研究所 ヒトゲノム解析センター　ゲノムデータベース分野
〒108-8639 東京都港区白金台4-6-1

kaminuma@nihs.go.jp

要旨

急展の著しいゲノム科学とゲノム技術の次なる最重要課題のひとつは、（日本ではテーラーメード医療とも言われる）パーソナライズド・メディシンである。パーソナライズド・メディシンとは、正しく選択された薬を正しく選択された患者に適切な量を投与する実践行為である。SNPsを見つけるプロジェクトは多数あり、それらがパーソナライズド・メディシンの基礎であるように言われている。しかし、 SNPsデータだけでは十分ではない。パーソナライズド・メディシンを実現するためには、よくトレーニングされたメディカル・スタッフの存在が前提であり、彼らが必要とする知識やデータに臨床の現場から容易にアクセスできる環境が整備されていなければならない。われわれはこのような知識基盤が如何なるものであり、それを活用するためにどのような計算ツールが必要かを考察した。それらは医薬品の一般名、構造式、3次元構造、薬物代謝酵素、標的分子、標的結合後の分子信号の流れなどに関するデータベースや知識ベース、ファーマコキネティクスのシミュレータ、分子計算パッケージなどである。われわれはその一部を開発しているが、全体の開発はCBI学会を母体としたコンソシアムなどでチャレンジすべき課題である。

キーワード: ファルマコインフォマティクス、パーソナライズド・メディシン、テーラード・メディシン、SNPs, ADME