Advanced in-silico drug screening to achieve high hit ratio

— Development of 3D-compound database —

**Yoshifumi Fukunishi[1] *, Yuusuke Sugihara[2], Yoshiaki Mikami[3],
Kohta Sakai[2], Hiroshi Kusudo[3] and Haruki Nakamura[4]**

[Translation from *Synthesiology*, Vol.2, No.1, p.60-68 (2009)]

Every year, several million compounds for drug screening are released by numerous vendors around the world. The information provided on these compounds is in the form of their two-dimensional (2D) structures. We have developed a software system to generate a database of three-dimensional (3D) structures of these compounds and have distributed this database. We have also developed a database of protein-compound docking scores of 180 proteins with respect to these millions of compounds. These databases make it possible to identify new active compounds for many drug targets.

Keywords **:** compound database, myPresto, virtual screening, *in-silico* drug screening, compound library

1 Introduction

One of the primary objectives in the post-genome era is the innovation of drug discovery. However, compared with the dramatic advancements in genetic analysis technology, drug discovery processes have been experiencing continuous difficulties and the expected results have not been achieved. In this situation, computational drug screening (*in-silico* or virtual screening (VS)) is considered to be one of the strategies for streamlining the drug discovery process. VS is used to computationally select seed molecules from existing molecules for pharmaceutical applications. Thus, VS requires a computationally accessible database of chemical compounds with 3D molecular structures (hereafter referred to as “compound DB”). Although overseas compound DB products are available, we developed an in-house DB due to the issues of price, quality, and management of results. We developed a compound DB by eliminating duplicated data using a chemical informatics approach, constructing 3D compound structures by a molecular force-field method, and computing atomic charges by quantum chemical calculations. These methods are described in chapter 4. In addition, we also developed a novel DB that predicts the binding energy of large numbers of predefined proteins and compounds. Using these DBs, it is now possible to predict active compounds with respect to target proteins for drug discovery with a high probability of success.

2 Objective

Our objective was to develop a compound DB usable for VS based on several million compounds marketed every year and to make it immediately available to the pharmaceutical industry. Several dozen major reagent vendors in the world distribute electronic files of reagent catalogs listing their 2D molecular structures, but 3D stereoscopic molecular structures rather than 2D chemical structures are necessary for VS. Therefore, the purpose of our study was to construct 3D molecular structures from the 2D molecular structures of millions of compounds in these catalogs and distribute them by compiling a database.

3 Benefits

Irrespective of the method used, the development of pharmaceutical compounds begins with a search for candidate compounds that could bind to the target proteins from compound databases. Computational searching in compound databases is a necessity in the modern drug discovery process. However, there are several issues in this process:

(1) Although compound DBs for VS have been developed and marketed for pharmaceutical companies by overseas software developers since the 1980s, license fees are expensive, with a license costing 4–6 million yen per year[1].

(2) Software products for the development of compound DBs are also marketed by overseas software developers[2]. Having used costly software to develop compound DBs for VS, we experienced several issues in terms of their quality such as frequent representations of incorrect 3D molecular structures, incorrect appositions of hydrogen atoms, and the generation

1. Biomedicinal Information Research Center, AIST Aomi 2-41-6, Koto-ku 135-0064, Japan * E-mail : y-fukunishi@aist.go.jp, 2. FUJITSU KYUSHU SYSTEM ENGINEERING LIMITED Life Science System Dept. PLM Solution Div. Nakase 1-9-3, Mihama-ku, Chiba 261-8588, Japan, 3. Hitachi East Japan Solutions, Ltd. Ekimaehonchou 12-1, Kawasaki-ku, Kawasaki 210-0007, Japan, 4. Institute for Protein Research, Osaka University Yamadaoka 3-2, Suita 565-0871, Japan

Received original manuscript October 28,2008, Revisions received December 9,2008, Accepted December 9,2008

of structures with a low probability of existence.

(3) The distribution of data generated using commercially available software for developing compound DBs is prohibited due to the licensing policy.

As described later in this paper, our objective was to develop and distribute novel DBs for VS using protein-compound affinity matrices based on the compound DBs. However, this cannot be achieved using commercially available software. The development of in-house software that generates compound data and compound DBs will resolve these issues. Upon distribution, the use of VS can be encouraged for users who find it difficult to afford costly licenses such as small- and medium-sized enterprises and academic researchers, and novel and advanced VS methods can be widely disseminated even to large enterprises. Economic and technological benefits will be obtained as a result.

4 Processes

4.1 Overall perspective

The overall software development process consisted of approximately 10 steps, as follows (Fig. 1). First, we eliminated duplicated compounds listed in 2D SD files provided by reagent vendors (for example, methanol is sold by any vendor). Since hydrogen atoms (protons) are normally omitted in 2D structures, protons were added. Parameters such as distances and bond angles between atoms were assigned to all of the atoms. The 3D structural coordinates, as well as enantiomers if they existed, were reproduced from the 2D coordinates based on this information. Atomic charges were then evaluated by quantum chemical calculations so that equivalent atoms exhibited equivalent charges. The generated 3D data were compiled into a relational database. We developed our software, avoiding violations of the patents on a number of commercially available software products for each step. Each development step is explained in detail below.

4.2 Handling of massive data sets

It is difficult to handle massive data sets. If millions of items of compound information are stored in a single file, the file size will exceed the limit that a computer can handle, whereas if one item of compound information is stored in each file, the millions of files produced cannot be contained in a single folder due to the constraints imposed by the computer system. Thus, the information on a single compound was stored in a file and the data for approximately 10,000 compounds were contained in each of several hundred folders prepared in order to handle millions of items of compound information using a hierarchical structure. The developed compound DB could be stored as a single relational database in a system with a 64-bit architecture.

4.3 Exclusion of duplicated compounds: Determination of compound identity

It is necessary to determine whether or not two molecules are identical. Since the identification of 4 million compounds requires the square of 4 million comparisons, we developed a high-speed discrimination method that consists of several steps as described later. We prioritized speed over accuracy by sacrificing a certain degree of discrimination accuracy. Since a few percent of commercially available compounds are different from the actual structures due to incorrect structure identifications and insufficient quality control, excessive pursuit of mathematical strictness would be meaningless.

4.3.1 Determination of chemical compositions based on pseudo-molecular mass weight

The chemical composition is a description of the type and number of atoms contained in a molecule; in the case of methanol (CH_3-OH), for example, it will be $C_1O_1H_4$. Comparison of chemical compositions is a quick method of discriminating compounds. No further discrimination is necessary if the chemical compositions of two molecules differ from each other. However, the character-string comparison of chemical compositions takes too much time. We therefore evaluated the molecular mass weight using the atomic mass weight with three places after the decimal point for each atom, and obtained a six-digit number for each molecule. This realized an accurate comparison in practical terms of chemical compositions by a single computation of molecular mass weight without comparing their character strings.

4.3.2 Identification of molecular topology based on graph invariants

The structural formulas of two compounds may differ even when their chemical compositions match. Although molecules can be graphically compared by superimposing their graphs, the graphical superposition of molecules is a nondeterministic polynomial time (NP)-complete problem, in which the computation time cannot be described as a polynomial of the number of atoms. In general, high-speed algorithms exist for problems with polynomial computation times; however, no effective algorithm exists in the case of

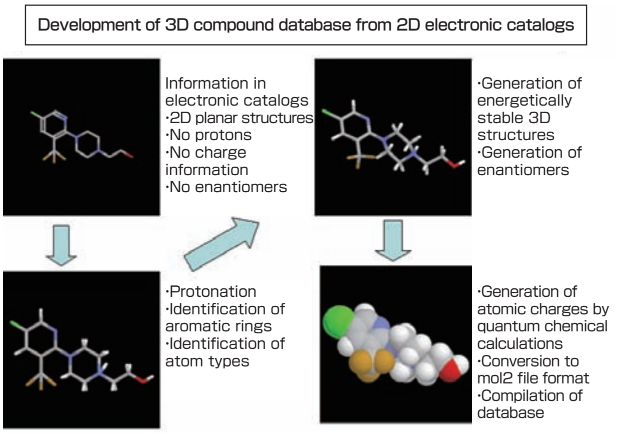

Development of 3D compound database from 2D electronic catalogs
Information in electronic catalogs
·2D planar structures
·No protons
·No charge information
·No enantiomers
·Generation of energetically stable 3D structures
·Generation of enantiomers
·Protonation
·Identification of aromatic rings
·Identification of atom types
·Generation of atomic charges by quantum chemical calculations
·Conversion to mol2 file format
·Compilation of database

Fig. 1 Process of development of 3D compound structures.

NP-complete problems, which thus require a prolonged period of computation time[3]. Hence, we developed a method that compares the topology of molecules using a molecular edge matrix, M, where M (i, j) = 1 if atoms i and j bind to each other and M (i, j) = 0 if not (Fig. 2).

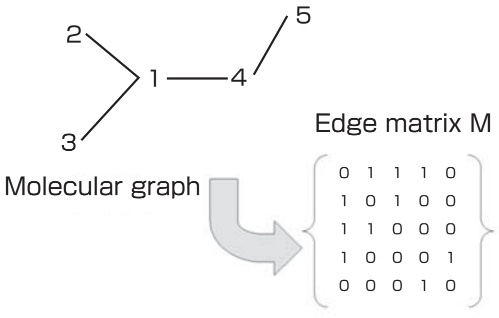
In molecular graphs, the sequence of atomic numbers is meaningless, and graph invariants should thus be evaluated. Here, since the edge matrix is a Hermitian matrix, eigenvalues will be the graph invariants. Although the Hosoya index is known as a method to evaluate graph invariants, it is computationally cumbersome[4]. Eigenvalue evaluation is a practical approach because its computation time is N^3 for the number of atoms, N. Protons were eliminated to reduce the matrix dimensions by half and the atomic number of each atom was substituted in the diagonal terms in order to reflect the type of atom.

4.3.3 Identification of geometric isomers

Although the method described in Section 4.3.2 makes it possible to identify the graphical topology of molecules with reasonable accuracy, it is incapable of discriminating geometric isomers such as cis and trans isomers. Thus, we developed graph invariants that can discriminate geometric isomers. First, for atoms i and j bound by a double bond, each graph fragment on the edge of four bonds is sequentially numbered from the maximum eigenvalue of a partial graph matrix as 1, 2 and 1', 2' (Fig. 3). Geometric isomers can thus be identified from the eigenvalues of the whole graph matrix by assigning −2 for the i–j component if vectors 1→2 and 1'→ 2' are parallel and +2 if they are anti-parallel.

4.4 Protonation

The number of absent protons in atoms such as C, N, O, and S in 2D structures was predicted from the bond order, and plausible coordinates of these protons were evaluated from the positional relationship with adjacent atoms and appended to the molecules. Although software that appends protons such as babel[5] and openbabel[6] is available, such software is not necessarily accurate. We investigated the protonated states of various functional groups and devised an algorithm so that it reproduces a molecule with a dominant ion forms under a vacuum and in water (near pH 7.0). Since accurate prediction of ionic configuration for a whole molecule is difficult, representative ionic configurations were applied for each functional group. Moreover, since the 2D chemical structures are simply diagrams, the actual atomic distance may be 1 Å or 10 Å. The average distance of the chemical bonds was therefore scaled to 1.5 Å.



$$\begin{Bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{Bmatrix}$$

Fig. 2 Construction of a binding matrix, M, from a molecular graph.

4.5 Addition of force field parameters

The generation of 3D molecular structures from their 2D counterparts was conducted using a molecular force field. Our compound DB applied a general amber force field (GAFF)[7] to generate 3D structures. Since the parameters of a GAFF are not available for most molecules, molecular structures cannot be determined. We therefore obtained accurate molecular structures by optimization calculations based on ab-initio calculations of quantum mechanics using CSD[8], a crystal structure database, and manually constructed the structures of 660 molecules. We then developed an algorithm that assigns atom types and force field parameters to all of the atoms if the parameters are absent, thereby making it possible to handle more than 99.9 % of molecules. Moreover, in addition to the consolidation of force field parameters, we developed tplgeneL, a software that assigns force field parameters to general compounds. This software is also capable of assigning parameters to the transition states of chemical reactions, which is useful for enzyme research.

4.6 Generation of 3D structures

Once force field parameters have been provided for the molecules, the 3D molecular structures can be generated. We applied cosgene[9], a software that we had previously developed for simulating molecular dynamics, to generate 3D structures by energy optimizations. 3D molecular structures cannot be generated unless a random displacement is applied on the initial coordinates, because a force in the Z-axis direction will not be generated in a 2D structure containing only X and Y coordinates. The structural adequacy of the generated 3D molecular structures (such as atomic distances and binding angles) was assessed by software, and if a distorted structure was generated, the initial coordinates were

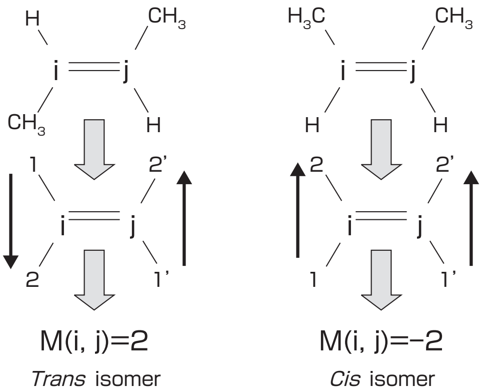


Fig. 3 Identification of geometric isomers. The thin black arrows indicate the sequence of eigenvalues assigned on the partial graphs.

reconstructed and the 3D structure was regenerated.

4.7 Identification of enantiomers and generation of isomers

If each of four chemical bonds of an atom such as carbon binds to different molecular fragments, the central atom will be the chiral center. Hence, discrimination of the four molecular fragments bound to the central atom will be necessary. In the method that we have developed, the bonds to the central atom are cut off and the molecular fragments are identified through comparison using a method similar to the algorithm described in section 3.2. Although the process will be slightly more complex if the central atom is part of a ring structure, a similar discrimination method is used. If only a single chiral center exists in a molecule, the enantiomers can be generated by converting the coordinate of each atom, (X, Y, Z), to (X, Y, −Z). If two or more chiral centers exist in a molecule, the bonds need to be reconnected; we used confgeneC, a newly developed software, for this purpose.

4.8 Computation of atomic charges by quantum chemical calculations

In quantum chemical calculations, electron spins and charges of molecules are necessary in addition to the molecular structures. Molecules used for drug development should not be radicals and also rarely exhibit magnetic properties; hence, the molecules were assumed to be closed-shell with zero spin. We developed an algorithm that automatically computes the molecular charge that stabilizes the system from the information on chemical bonds. The charge of the whole molecule is considered to be the sum of the formal charge of each atom. For example, the formal charge of a carbon atom is considered to be zero if the sum of the chemical bonds is four and +1 if it is three; the charge of nitrogen is +1 if the sum of the chemical bonds is four and zero if it is three; and the charge of oxygen is zero if the sum of the chemical bonds is two and −1 if it is one. The molecular charges were then evaluated by summing the formal charges obtained in this manner from the whole molecule.

There are several methods of computing atomic charges. The Gasteiger method[10] assigns electron affinity to each atom and evaluates the equilibrium electron distribution where atoms pull their electrons with each other based on the organic electron theory. A rough estimation requires less than a second for most molecules. In semi-empirical quantum chemical calculations, the AM1 and PM3 models (recently, PM7) of MOPAC[11] are well-known. The PM3 model is an excellent method in which an effective Hamiltonian is derived by fitting parameters so that the heat of formation can be represented; however, structures commonly observed in pharmaceuticals such as an amide bond cannot be accurately computed. The AM1 model also evaluates an effective Hamiltonian by fitting parameters; although it is inaccurate in predicting heat of formation, most structures such as an amide bond can be calculated correctly. However, this method occasionally fails to accurately predict the atomic charges of ring structures that contain nitrogen atoms. If the molecular structure is defined, the computation time will normally be several to several dozen seconds and is approximately proportional to N^3 for the size of an atom, N. The computational accuracy of the charge is very high. In ab-initio calculations of quantum chemistry, wave functions and partial atomic charges are generally computed by the RHF/6-31G* and restrained electrostatic potential (RESP) methods[12], respectively. Although this approach evaluates charges extremely accurately, the computation time will normally be several to several dozen minutes if the molecular structure is defined and is proportional to N^4 for the size of an atom, N.

Atomic charges will be meaningless unless protein-compound bindings are accurately computed. Thus, docking calculations of 132 protein-compound complexes were performed by sievgene[13], our protein-compound docking simulation software*[term1]. As a result, accurate structures were obtained with a probability of 56 % by RHF/6-31G* (with an accuracy of 2 Å), with 2–3 % lower probability by MOPAC AM1, and with about 5 % lower probability by the Gasteiger method. A small-scale drug screening experiment using approximately 10,000 compounds was also performed targeting several proteins such as cyclooxygenase-2 (COX-2) and thermolysin. It was found that the hit rate was higher if the molecular charges were more accurate, and the hit rate was only a few percent lower even when the Gasteiger method was applied.

Since the atomic charges of several million molecules will be calculated, the computation time should be prioritized. In addition, it was found to be unnecessary to use a method that is as accurate as RHF/6-31G*. Hence, we decided to employ the MOPAC AM1 method for the computation of charges because the compound DB will be the overall foundation. Although MOPAC generally requires a MOPAC-specific input format, we modified it so that we can input and output the mol2 file format, which is a standard format to represent compounds in the field of drug discovery. For this purpose, we are distributing a patch file to modify MOPAC free of charge.

4.9 Determination of equivalent atoms

The charge of three protons in a methyl group should be configured to be chemically equivalent. The determination of atomic equivalency is necessary for computing atomic charges.

The equivalency of arbitrary atoms i and j is considered to be as follows: If i = j, the atoms are obviously equivalent. However, if this is not the case, and if atoms i and j do not directly bind to each other, all of the atoms binding to atom i should be equivalent to those binding to atom j, whereas if atoms i and j do bind, all of the other binding atoms should be equivalent to each other.

The method of discriminating equivalent atoms is as follows: Arbitrary atoms i and j are selected for marking as "already checked." If the atoms are equivalent, they are marked as "already checked." If i = j, the atoms are equivalent and, in this case, no further validation is necessary. If atoms i and j are bound by a single bond, both i and j bind to the "already checked" atoms, and their atomic symbols are the same, then i and j are considered to be equivalent.

All of atoms m_i and m_j binding to i and j, respectively, are temporarily marked as "already checked" and their equivalency is tested by the abovementioned procedure. Subsequently, if m_i and m_j are not equivalent, the "already checked" mark is removed. However, if all of m_i and m_j are determined to be equivalent, atoms i and j are considered to be equivalent.

When equivalency is tested from atoms i and j, the atoms that are to be tested are indicated as gray-filled circles and the atoms that will finally be tested are indicated as black-filled circles in Fig. 4. The equivalency test is necessary up to the point where the routes from i and j meet each other (black-filled circles), and the whole graph is not necessarily tested.

4.10 Compilation of database and downloading of files

The compound DB is structured as a relational database. The schema includes the information on compound mol2 files (atomic names, 3D coordinates, atomic charges, chemical bond orders, etc.) in addition to the molecular weight, HOMO/LUMO energy in the MOPAC AM1 model, and solvation free energy per molecule and per atom calculated by the GBSA model. The solvation free energy per atom is useful for identifying the location of a compound in the chemical space of compounds (compound space), and is thus used as a parameter that indicates its diversity (the degree of diversity in the collected compounds) in a DB[14]. Compound information can be downloaded in the form of mol2 files from the compound DB.

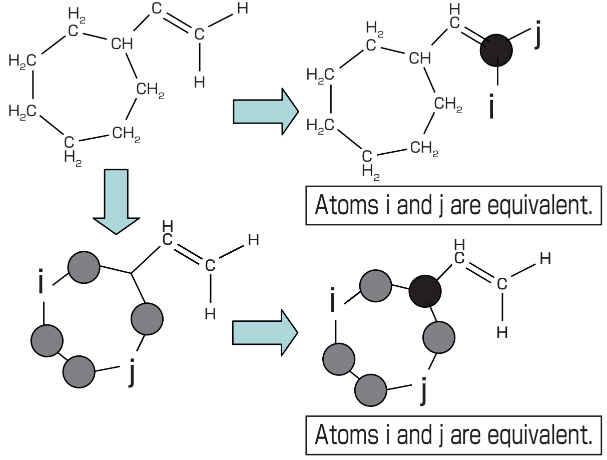


Fig. 4 Determination of equivalent atoms.
"●" indicates "already checked" atoms.

4.11 Computation of protein-compound affinity matrix

We selected a large number of proteins other than the target proteins, performed combinatorial docking calculations against the compound library in the compound DB, constructed a protein-compound affinity matrix, and compiled it into a database. This is the basic DB for the drug screening methods we developed, the multiple target screening (MTS) method[15] and docking score index (DSI) method[14], which will be described later, and is a crucial resource for our VS (Fig. 5).

If the compounds that bind to the target proteins are selected in the order of the higher docking scores*term2 (scores) calculated by general VS, the hit rate is low. A compound that exhibits a high score against a target protein occasionally also exhibits high scores against other proteins, which indicates that the associativity of the compound with respect to the target protein is not specific. In contrast, only one compound is focused on in the MTS method; the proteins that bind to the compound are searched for and the compounds that bind to the target proteins with the highest score are selected as candidate hit compounds.

The accuracy of the score can also be improved by using the protein-compound affinity matrix. The free energies of binding for a particular compound to bind to analogous proteins are considered to be close in value. Errors in the score can be reduced by averaging the weighted scores depending on the similarity of the proteins; the details are reported elsewhere[16]. In particular, the scores were corrected by the following equation:

$$s^{new^i}{}_a = \frac{\sum_b s^i_b R^b_a}{\sum_b R^b_a} \qquad (1)$$

where $s^{new\ i}{}_a$, s_b^i, and R_a^b are the newly defined score between protein a and compound i, the score between protein b and compound i, and the correlation coefficient of protein a and protein b, respectively.

In addition, if known active compounds exist in the compound list, the scores can also be corrected so that the known active compounds will be preferentially predicted. As shown in the following equation, the corrected scores are described as a linear combination of the scores and the coefficient M_a^b was evaluated so that the database enrichment*term3 was maximized:

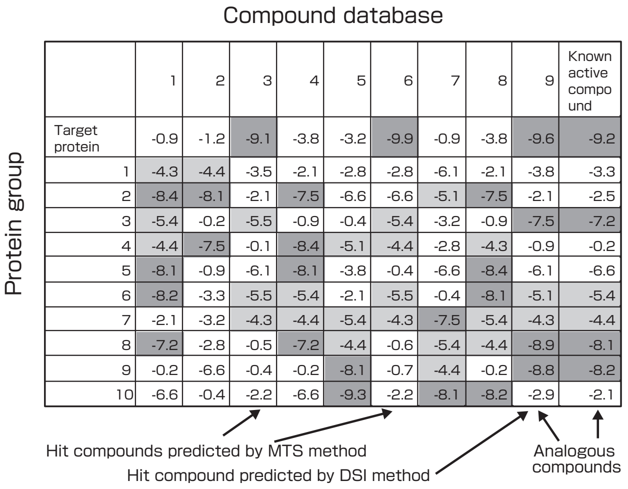
$$s^{new^i}{}_a = \sum_b s^i_b M^b_a \qquad (2)$$

As a result of applying the MTS method to 12 target proteins including COX-2 and HIV-1 protease and selecting the top 1 % of compounds predicted from the compound library, the discovery rate was improved approximately 40-fold compared with random screening[16].

The DSI method searches for compounds analogous to known active compounds using the protein-compound interaction matrix. Even different compounds that bind to the same protein are considered analogous (Fig. 5). The DSI method does not require the 3D structure of target proteins and can thus be applied to target proteins with unknown 3D structures such as G protein-coupled receptors (GPCRs). In addition, similarly to the MTS method, the DSI method can be combined with methods that correct scores to maximize the discovery rate of known compounds. As a result of applying the DSI method to a total of 14 target proteins including the proteins mentioned above and GPCRs and selecting the top 1 % of compounds predicted from the compound library, the discovery rate was improved approximately 70-fold on average compared with random screening[17].

5 Degree of achievement

We have currently achieved more than 90 % of the initial objective. Our first compound DB was released in 2004 and immediately used for compound screening against TNF-α converting enzyme. The MTS and DSI methods were applied using a protein-compound affinity matrix containing 182 proteins and 1 million compounds. Among 900 compounds subsequently purchased, 35 were found to be active compounds. The discovery rate was approximately 500-fold higher than the previously conducted screenings in which seven active compounds were obtained by randomly screening 100,000 compounds. In addition, no active compound was found after purchasing 700 compounds following screening by Glide, a commercially available software; hence, the discovery rate was dramatically improved by our methods. Since then, the compound DB has been annually renewed and the 2007 version is the latest. We have conducted direct screenings with respect to 10 target proteins over a period of six years and obtained active compounds with a probability range of a few to 20 %. This rate is several hundred to one thousand times higher than that achieved by random screening. Moreover, every year the compound DB and the protein-compound affinity matrix have been distributed to 10 to 20 institutions, primarily pharmaceutical companies, in Japan and overseas. The software and the compound DB have been partially released as myPresto[18] and LigandBox[19], respectively.



Compound database

	1	2	3	4	5	6	7	8	9	Known active compo und
Target protein	-0.9	-1.2	-9.1	-3.8	-3.2	-9.9	-0.9	-3.8	-9.6	-9.2
1	-4.3	-4.4	-3.5	-2.1	-2.8	-2.8	-6.1	-2.1	-3.8	-3.3
2	-8.4	-8.1	-2.1	-7.5	-6.6	-6.6	-5.1	-7.5	-2.1	-2.5
3	-5.4	-0.2	-5.5	-0.9	-0.4	-5.4	-3.2	-0.9	-7.5	-7.2
4	-4.4	-7.5	-0.1	-8.4	-5.1	-4.4	-2.8	-4.3	-0.9	-0.2
5	-8.1	-0.9	-6.1	-8.1	-3.8	-0.4	-6.6	-8.4	-6.1	-6.6
6	-8.2	-3.3	-5.5	-5.4	-2.1	-5.5	-0.4	-8.1	-5.1	-5.4
7	-2.1	-3.2	-4.3	-4.4	-5.4	-4.3	-7.5	-5.4	-4.3	-4.4
8	-7.2	-2.8	-0.5	-7.2	-4.4	-0.6	-5.4	-4.4	-8.9	-8.1
9	-0.2	-6.6	-0.4	-0.2	-8.1	-0.7	-4.4	-0.2	-8.8	-8.2
10	-6.6	-0.4	-2.2	-6.6	-9.3	-2.2	-8.1	-8.2	-2.9	-2.1

Protein group
Hit compounds predicted by MTS method
Hit compound predicted by DSI method
Analogous compounds

Fig. 5 Diagram of MTS and DSI methods.
The numbers in the table indicate the scores. Higher scores are indicated by a deeper color.

6 Future work

Firstly, our compound DB is not suited to screening of inhibitors of metalloproteinases containing metals such as zinc. The ion form of the molecules exhibits a predominant configuration under water; however, it will be different when the molecule binds to metals. For example, while a thiol (-SH) is normally configured as -SH under water, it is deprotonated and becomes -S$^-$ in the case of coordination with a metal. Changes in the ion form of molecules due to coordination with metals are observed in various functional groups. We found that the discovery rate strongly depends on the ion form of compounds through the VS of metalloproteinases. Accordingly, we plan to develop a compound DB for metalloproteinases.

Secondly, our compound DB does not include inorganic compounds. Inorganic compounds such as metal complexes are considered to be unsuitable for drugs and are generally excluded from the compound DB. However, zinc complex was recently discovered to be an active compound with respect to insulin receptor protein, for which no active compound has previously been known except peptides, and this has attracted attention to inorganic compounds as novel therapeutic agents. The development of a DB for inorganic compounds is therefore necessary in order to examine the possible applications of inorganic compounds.

Thirdly, distribution of our compound DB has depended solely on word-of-mouth publicity and it has not gained recognition by means of journal articles or websites. This is because our compound DB depends on catalog data provided by commercial firms. Catalog distribution is restricted to the marketing of reagents and advertisements of reagent vendors should be posted. For example, the free downloading of ZINC[20] was realized by posting advertisements of reagent vendors on university websites as a result of direct negotiations with reagent vendors. However, the advertising of private companies is prohibited at the National Institute of Advanced Industrial Science and Technology (AIST), and free downloading therefore cannot be realized. We are consequently distributing our compound DB on the assumption that AIST has compiled a database from catalogs that the users have independently obtained. It is also possible for incorporated associations, our collaborators who support our research, to distribute our package, but not to negotiate with reagent vendors. Collaboration with private corporations is being promoted through the encouragement of industry-government-academia coordination, however, and this issue is therefore also considered to be a future task.

Acknowledgement

This study was supported by the New Energy and Industrial Technology Development Organization (NEDO) and the Ministry of Economy, Trade and Industry (METI) of Japan.

Terminology

Term 1. Protein-compound docking software: Software that computationally predicts the most feasible and energetically stable structure of protein-compound complexes by allocating a compound adjacent to the surface of a 3D protein structure. The docking simulation takes several seconds to a minute in drug screening. Typical software includes DOCK, AutoDock, and myPresto.

Term 2. Docking scores: Values that represent the strength of a protein-compound interaction estimated by docking software, and generally correspond to the free energy of binding.

Term 3. Enrichment: The ratio of the number of correct hit compounds to the number of candidate compounds predicted by computations in drug screening. In general, one out of 10,000 compounds hit in a random screening; thus, if one out of 100 compounds predicted by computational analysis was found to be a hit compound, the enrichment with respect to the random experiment would be 100-fold.

References

[1] http://www.mdl.com/jp/products/experiment/cims/index.jsp

[2] http://www.molecular-networks.com/software/corina/index.html

[3] M. Hattori, Y. Okuno, S. Goto and M. Kanahisa: Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.,* 125 (39), 11853-65 (2003).

[4] J. Gasteiger and T. Engel: *Chemoinformatics: A textbook*. WILEY-VCH: Weinheim. (2003).

[5] http://www.lmcp.jussieu.fr/sincris-top/logiciel/prg-babel.html

[6] http://openbabel.org/wiki/Main_Page

[7] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case: Development and testing of a general amber force field. *J. Compt. Chem.,* 25 (9), 1157-1174 (2004).

[8] http://www.ccdc.cam.ac.uk/products/csd/

[9] Y. Fukunishi, Y. Mikami and H. Nakamura: The filling potential method: A method for estimating the free energy surface for protein-ligand docking. *J. Phys. Chem. B.* 107 (47), 13201-13210 (2003).

[10] J. Gasteiger and M. Marsili: A new model for calculating atomic charges in molecules. *Tetrahedron Lett*., 3181-3184 (1978).

[11] http://openmopac.net/index.html

[12] J. Wang, P. Cieplak and P.A. Kollman: How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem*., 21 (12), 1049-1074 (2000).

[13] Y. Fukunishi, Y. Mikami and H. Nakamura: Similarities among receptor pockets and among compounds: Analysis and application to *in silico* ligand screening. *J. Mol. Graph. and Model*., 24 (1), 34-45 (2005).

[14] Y. Fukunishi, Y. Mikami, K. Takedomi, M. Yamanouchi, H. Shima and H. Nakamura: Classification of chemical compounds by protein-compound docking for use in designing a focused library. *J. Med. Chem*., 49 (2), 523-533 (2006).

[15] Y. Fukunishi, Y. Mikami, S. Kubota and H. Nakamura: Multiple target screening method for robust and accurate in silico ligand screening. *J. Mol. Graph. and Model*. 25 (1), 61-70 (2005).

[16] Y. Fukunishi, S. Kubota and H. Nakamura: Noise reduction method for molecular interaction energy: application to in silico drug screening and in silico target protein screening. *J. Chem. Info. Mod*., 46 (5), 2071-2084 (2006).

[17] Y. Fukunishi, S. Hojo and H. Nakamura: An efficient in silico screening method based on the protein-compound affinity matrix and its application to the design of a focused library for cytochrome P450 (CYP) ligands. *J. Chem. Info. Mod.,* 46 (6), 2610-2622 (2006).

[18] http://presto.protein.osaka-u.ac.jp/myPresto4/index_e.html

[19] http://presto.protein.osaka-u.ac.jp/LigandBox/web_search.cgi

[20] J. J. Irwin and B. K. Shoichet: ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.,* 45 (1), 177-82 (2005).

Authors

Yoshifumi Fukunishi

Dr. Fukunishi received his Ph.D. from the Graduate School of Engineering, Kyoto University, in 1994 and served as an adjunct researcher at the National Institute for Advanced Interdisciplinary Research, Ministry of International Trade and Industry (MITI). He then worked as an HFSP fellow, a post-doctoral researcher at Rutgers University, a JST post-doctoral fellow at RIKEN, and at Hitachi, Ltd., and has held the position of senior research scientist at the Biomedicinal Information Research Center (BIRC), AIST, since 2000. He specializes in computational chemistry and was in charge of developing prototype models, devising algorithms, and designing the overall research in this study

Yuusuke Sugihara

Mr. Sugihara received his M.S. from the Macromolecular Chemistry Course, Department of Chemistry, Graduate School of Science, Hiroshima University, in 1996, and joined Arakawa Chemical Industries, Ltd. in the same year. After leaving that company in 2000, he joined Fujitsu Kyushu System Engineering, Limited in 2001. He was primarily in charge of developing 3D structures from cataloged compounds in this study.

Yoshiaki Mikami

Mr. Mikami joined Hitachi East Japan Solutions, Ltd. in 1987 and is currently engaged in system development in such areas as virtual screenings as well as in consultation business. He is a member of the Information Processing Society of Japan and was primarily in charge of developing protein-compound interaction matrices in this study.

Kohta Sakai
Mr. Sakai received his M.S. from the Macromolecular Chemistry Course, Department of Chemistry, Graduate School of Science, Kyushu University, in 1989, and joined Fujitsu Kyushu System Engineering, Limited in the same year. He was primarily in charge of developing 3D structures from the cataloged compounds in this study.

Hiroshi Kusudo
Mr. Kusudo joined Hitachi East Japan Solutions, Ltd. in 2002 and is currently engaged in the development of parallel computation systems and research support business. He is a member of the Information Processing Society of Japan and was primarily in charge of developing protein-compound interaction matrices in this study.

Haruki Nakamura
Dr. Nakamura received his Ph.D. from the Graduate School of Science, the University of Tokyo, in 1980. He was an assistant professor at the Faculty of Engineering, the University of Tokyo, and served at the Protein Engineering Research Institute and the Biomolecular Engineering Research Institute. He has held the position of professor at the Institute for Protein Research, Osaka University, since 1999. He specializes in biophysics and was primarily in charge of collecting public data sets and supervising the overall research in this study.

Discussion with Reviewers

1 Significance of developing the compound database

Question and comment (Akira Ono)

The authors stated a clear research aim and comprehensibly portrayed a scenario to select elemental technologies as shown in Fig. 1, which were then integrated into a practically operational database. This is a typical *Type 2 Basic Research* and is also an excellent example of *Product Realization Research*. It is expected that the database developed in this study will be highly valued through its applications to drug screening for target proteins.

Answer (Yoshifumi Fukunishi)

A compound database is a repository of knowledge regarding the “synthetic easiness” of compounds synthesized in the past. We expect that the database will not only be directly used for drug discovery but also function as a basis to understand easily synthesizable compounds and even previously inconceivable compounds, which may lead to establish a new research field.

2 Design of unknown active compounds

Question and comment (Akira Ono)

The compound database is intended for efficiently examining the strength of chemical bonds between a particular target protein and numerous predefined compounds, thereby dramatically improved a hit ratio of drugs. Now, is it possible for the users to predict unknown compounds that could bind with a particular target protein even more strongly by using the database?

Answer (Yoshifumi Fukunishi)

There is a possibility that the users can predict previously unknown active compounds that are not included in the compound database. There was a trend that active compounds were classified into several clusters according to their chemical properties as a result of screening compounds that potentially bind to a particular target protein. Therefore, it is considered feasible to synthetically design an unknown active compound that exhibits the properties of a particular active compound group.

3 Possible improvement of the database

Question and comment (Akira Ono)

As mentioned in section 4.3, it is important to recognize that excessive pursuit of mathematical strictness is meaningless in developing a compound database. In this regard, is there still a room for optimizing the database according to needs by reassessing its development process?

Answer (Yoshifumi Fukunishi)

Redesigning of the database according to needs, *i.e.*, ad-hoc database, is considered possible. For example, we adopted a dominant configuration of compounds in water (regarding the protonation state of a carboxylic acid, for example, $-COO^-$ was adopted instead of -COOH) for the current version of the database; however, the molecular configuration could be changed when the compounds bind with proteins. Recently, it is often discussed that a docking simulation of a compound targeting a highly-charged protein pocket is extremely difficult. A dominant ion form of carboxylic acid could occasionally be -COOH inside the negatively-charged protein pocket. Hence, it will be important to develop a target-oriented compound database in future.

4 Comparison with existing overseas databases

Question and comment (Akira Ono)

It is mentioned in section 4.11 that when the current database was employed, the enrichment factor was improved approximately 40- or 70-fold compared with random screening. In contrast, how superior is the developed database over the precedent overseas compound databases in terms of enrichment or enrichment factors?

Answer (Yoshifumi Fukunishi)

In general, unsuccessful predictions of computationally screened compounds are not published as research articles; thus, it is difficult to compare databases in detail. In our case, the hit ratio was 3–30 % when computationally predicted 100–300 compounds were purchased. So far, only one out of five cases showed 0 % hit ratio. Hit ratios reported on other literatures are mostly 10 % at best and 50 % of targets show 0 % hit ratio. Therefore, the developed database combined with our prediction method is considered more effective than the existing overseas databases.

5 Consideration of tautomers and ion forms

Question and comment (Takatsugu Hirokawa)

While commercially-produced overseas databases and software are acquiring the major share in the field, it is noteworthy that such a high-quality compound database and an unprecedentedly unique protein-compound affinity matrix were released by Japanese researchers. Various issues involved in digitally processing compound data were fully addressed in each development process discussed in the article, and this ensures that the database can be used reliably by researchers. This work should also be highly acclaimed as *Product Realization Research*.

Regarding the protonation of compounds, how are the tautomers and ion forms considered (such as whether pH7.0 is assumed) besides the statement, “We investigated the protonatation states of various functional groups...under a vacuum and in water (near pH7.0)” of section 4.4.

Answer (Yoshifumi Fukunishi)

Protonation status is based on the assumption of pH7.0. However, since it is difficult to predict accurate pKa, a dominant configuration of each functional group contained in a molecule at pH7.0, rather than the pKa of the whole molecule, was adopted.

We applied the same scheme on tautomers. Accordingly, although the compound configurations are still not chemically strict, the ion forms of our compound database are more reliable than typical open software such as babel and openbabel, which frequently generate high-energy tautomers.

6 Prediction of non-selective compounds

Question and comment (Takatsugu Hirokawa)

As application examples of protein-compound affinity matrix discussed in section 4.11, non-selective or low-selective compounds (or highly selective compounds), which nonselectively bind to many target proteins, may possibly be predicted by using the database. If so, this may realize highly unique annotations such as *in-silico* frequent hitters or *in-silico* chemical alerts for target selectivity. Please discuss about the possibility of predicting non-selective compounds based on the database, although this might have been conducted already.

Answer (Yoshifumi Fukunishi)

It is an insightful question. Frequent hitters account for several tens of percent in VS and thus increase the cost and are the bottleneck of screening processes. Approximately 20 % of the predicted compounds are frequent hitters in our screenings. Recently, we are collecting several dozens of compounds that are considered the frequent hitters from literatures (*J. Med. Chem.* 2003, vol. 46, page 4477-4486, *J. Med. Chem.* 2002, vol45, page 137-142) and developing their 3D structures. Once the data have been prepared and included in the computation of protein-compound affinity matrix, we may be able to find a property that contributes to a high score against any target protein. However, there is a report that most frequent hitters form micelle colloids when observed under an electron microscope, and thus, the adsorption of the micelle to proteins could be a cause of the “frequent hit” (*J. Med. Chem.* 2002, vol. 45, page 1712-1722). If the frequent hitters exhibit non-selectivity against target proteins as unimolecular compounds, it will be possible to distinguish frequent hitters by docking simulations. However, if the micelle formation is the cause of the “frequent hit,” docking simulations, which assume infinite dilution condition, will not be able to discriminate frequent hitters. Thus far, as a result of applying a solubility prediction based on molecular descriptors to analyze the aqueous solubility of frequent hitters, it was found that highly hydrophobic molecules tended to be frequent hitters and thus could be distinguished from drug molecules that are not frequent hitters (P1-06 at the Chem-Bio Informatics Society (CBI) Annual Meeting 2008 International Symposium). If the water solubility of a compound determines its likeliness to be a frequent hitter, the micelle formation should be the primary cause of “frequent hit.” Nevertheless, there remains a possibility that docking simulations could be more effective to predict frequent hitters. Although it will take time, we will continue to characterize frequent hitters.

Although we analyzed the side effects caused by the non-selectivity of compounds by MTS and DSI methods, thus far, no clear association between the side effects and non-selectivity of compounds was found. COX2, a typical target of nonsteroid anti-inflammatory drug (NSAID), and COX1, functioning as a gastric mucosal protector, are the enzymes with 60 % homology in their amino acid sequences; thus, it was an issue of concern that NSAID could cause gastric ulcers by inhibiting not only COX2 but also COX1. Recently, COX2 selective NSAIDs (e.g. coxibs) have been developed. Even though we then examined whether selective and non-selective NSAIDs can be distinguished by using protein-compound affinity matrix, it was not successful. In fact, the COX2 selectivity of coxibs is relatively low; at a concentration of 80 % inhibition of COX2, selective and non-selective NSAIDs inhibited approximately 20 % and 80 %, respectively, of COX1 activity (*Proc. Natl. Acad. Sci.* USA. 1999, vol. 96, page 7563-7568). Hence, the drug selectivity, in this case, is not a black-and-white property but a matter of degree. We consider it possible to distinguish highly selective and non-selective compounds by using protein-compound affinity matrix and currently prepare the structures of approximately 1500 proteins for docking simulations. Although an actual analysis cannot be performed due to the limitation of computational capacity, it will be possible anytime soon.