Development and Applications of Virtual Screening System —Xsi ver. 1.0—

Sumitomo Pharmaceuticals Co., Ltd.
Research Division
Chemistry Research Laboratories
Kazuto YAMAZAKI
Genomic Science Laboratories
Masaharu KANAOKA

Virtual screening, a computational method to identify bioactive compounds among a vast number of chemicals, is in use with a great expectation of saving time and cost necessary for actual screening of bioactive compounds. Virtual screening techniques can be categorized in two, namely Ligand-based Drug Design (LBDD) and Structure-based Drug Design (SBDD). Conceptually LBDD and SBDD can be complementary with each other, but have been developed and applied independently. In order to improve the current methods, we developed 'Multiple Docking' method which utilized LBDD and SBDD complementarily. In this review we describe this new method, along with a computer software system 'Xsi' developed to realize the method.

This paper is translated from R&D Report, "SUMITOMO KAGAKU", vol. 2005-I.

Introduction

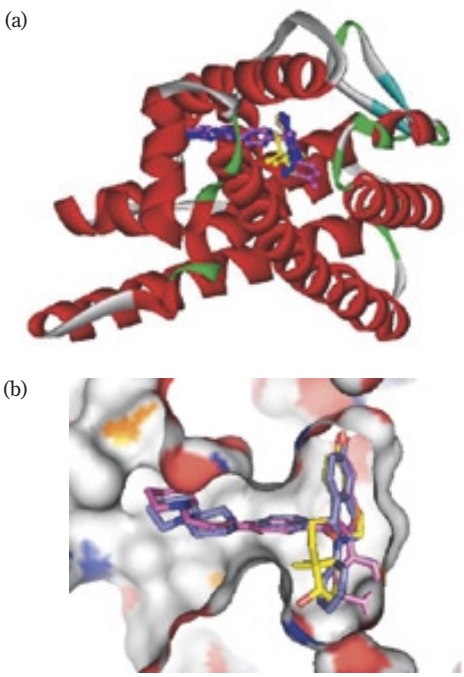
Screening for lead identification in drug discovery must be carried out on a large number of compounds to find the biologically active chemical compounds. Through the progress in combinatorial synthesis technology and high-throughput screening technology, there has been an improvement in the supplying of large number of compounds and the efficiency of pharmacological evaluation testing, but even so the discovery of drug molecules requires a great deal of time and expense. Therefore, there are great expectations for virtual screening that carries out the screening on computers.

When virtual screening was first used, the main current was procedures based on structure-activity relationship data obtained experimentally. This method was based on empirical rules to the effect that compounds with the same biological activities have common structures and physiochemical properties. QSAR analysis and superposition analysis have been known for a long time, and in recent years, they have been developed as chemoinformatics. Since these are methods based on pharmacologically active compounds, that is, ligand information, this is called Ligand-based Drug Design (LBDD). There has been rapid improvement in the techniques in LBDD in the last decade, and it has become an indispensable tool for drug discovery research. Since these procedures are based on structure-activity relationship information, they are used in close connection with pharmacological activity evaluation testing, and the predictive abilities are improved with the accumulation of experimental data. The use of LBDD reduces the time and effort needed for discovery of drug molecules, and it effectively contributes to the improvement of the probability of success. However, there is an fundamental problem in that predictions cannot be made when prior experimental information is insufficient, that is, in areas where the structure-activity relationship is unknown.

On the other hand, with the developments in post-genome research, a large number of tertiary structures of proteins, which are the targets of drug compounds, have been elucidated. As a result, virtual screening based on the tertiary structures of proteins is becoming more widely used. These techniques are based on the knowledge of complementary relationships ("key and keyhole" relationships) between interacting proteins and chemical compounds. Specifically, docking studies, molecular simulations and the like are called Structure-based Drug Design (SBDD). Due to intensive research over the past few years, SBDD has finally reached a

practical level. As opposed to LBDD, this technique does not require prior experimental information, so that there are great expectations that, in principle, it can predict unknown parts. However, even though it seemingly can replace experiments, it cannot be said to have actually reached that level. In addition, the fact that it is not predicated on prior experimental information means that, conversely, even if experimental data, such as structure-activity relationships and X-ray crystal structures, are accumulated, the predictive performance of SBDD cannot be improved. Therefore, in actual drug discovery research LBDD and SBDD have to be used along with pharmacological experiments in such a way optimal for the particular research program.

Fig. 1 shows a superimposition of three ligand structures bound to an estrogen receptor obtained from X-ray structural analyses.[1), 2)] It illustrates that each of the ligands is complementary to the receptor, as well as that the ligands themselves are similar to each other. In other words, LBDD and SBDD have each made for independent improvements in technique, but essentially, they do no more than analyze the same phenomena from different points of view. In addition, the characteristics and problems of the two methods as described above are complementary. Therefore, virtual screening that integrates the two techniques should be coming to practical usage. In recent years, there have been a number of reports on docking studies carried out on multiple compounds with specific proteins followed by three-dimensional QSAR analysis based on the models obtained.[3) – 6)] This procedure is certainly a combination of the two techniques, but it is no more than simply using them sequentially. The integrated analytical technique that is really required should address the problems of both, but in actual fact neither the prediction of unknown parts by structure-activity relationship, which is the problem with LBDD, nor the improvement of docking precision, which is the problem with SBDD, has been improved.



(a) X-ray structure of the ligand binding domain of estrogen receptor α with Tetrahydroisoquinoline (1XQC, blue), Raloxifene (1ERR, pink) and 17β-Estradiol (1ERE, yellow)
(b) Ligand binding site of estrogen receptor α with the three ligands

Fig. 1 Structures of three ligands bound to estrogen receptor α

The authors have developed the Xsi ver. 1.0 integrated virtual screening system that can carry out virtual screening by seamlessly linking the LBDD and SBDD techniques jointly with Mizuho Information & Research Institute, Inc. (formerly Fuji Research Institute Corporation).[7), 8)] In addition, the multiple docking analysis method was invented as a technique that solves the problems with LBDD and SBDD.[9), 10)] The utility of this was verified using a specific example that simulated actual drug discovery research.

Development of the Integrated Virtual Screening System

Typical integrated software presently on the market includes SYBYL® (Tripos, Inc.)[11)] and MOE™ (Chemical Computing Group Inc.).[12)] SYBYL® is an integrated molecular design software with a long history to which new modules can be added as necessary. It started as a system for low molecular weight organic molecules, but it is now multifunctional integrated software that includes calculation functions for proteins. MOE™ has been developed with the goal of providing a suitable programming environment for molecular calculations. Various calculation functions are provided as add-on programs working under that environment. Both are equipped with a Graphical User Interface (GUI), and the executing of calculations and performing of analyses is done intuitively. In addition, a dedicated programming language is provided, and complex, continuous calculation processing is possible.

However, the convenience and extensibility of the

two types of software have their merits and demerits. When functions are grouped to a certain extent and provided as modules as with SYBYL®, there is a high level of convenience for the user. In other words, the users do not need to develop the modules themselves, and effort can be focused on carrying out the appropriate analysis by combining existing functions. On the other hand, environments where one can develop the modules themselves as with MOE™ have superior extensibility for changing specifications and adding functions. To realize the multiple docking analysis that will be discussed later, it is necessary not only to supplement the functions that are insufficient in existing modules but also to make complex combinations of multiple modules. Therefore, since we wanted integrated software with both convenience and extensibility, we decided to carry out our own software development.

The integrated software was developed jointly with Mizuho Information & Research Institute, Inc., a domestic software development company. Our intention was that it would not only be used in house, but also made as a commercial product.

Collaboration with a software company was beneficial, for we could utilize their program assets as well as the high-level IT technology of experts. In addition, it is easier to consign the software maintenance envisioned after operations begin as well as the development for changes in specifications and extensions of functions. With a domestic company, we could have smooth communications from the beginning and a quicker response than with overseas companies that tend to give the large pharmaceutical companies priority. It is also important that software reliability be assured for its being sold as a product. This integrated software was released at the end of January 2004 as Xsi ver. 1.0.[7), 8)]

1. Overview of Xsi

Like SYBYL®, Xsi is composed of multiple modules. Each module consists of one class for storing data and multiple functions that specify the data operations and calculation processing for that class. Xsi execution is carried out based on a Character User Interface (CUI) through dedicated scripts. The task of writing a dedicated script is very simple, basically it is sufficient to list a class and function pair on each line. Conditional expressions, repeat expressions and a variety of operators can be used as needed. Every class can be handled as a multidimensional array. **Fig. 2** shows an example of a script for outputting the results of the continuous optimization of structures using molecular mechanical calculations on multiple molecules read from an SD file. Excluding the comments added to make the script easy to read, it can be written with a script of fewer than 20 lines.

```
MoleculeSetFileSD sdfile;                    } Declaration of class and array
Array<Molecule> am;
MolecularMechanics mm;
Integer i;
//
sdfile.setFileName ("test.sdf");             } Import molecules from an SD file
am = sdfile.getMolecule();
//
i = 0;                                       } Energy minimization by molecular mechanics
while (i < am.size()) {¥
mm.setMolecule(am[ i ]);¥
mm.minimize();¥
mm.clear();¥
i = i + 1;¥
};
//
sdfile.clear();                              } Output molecules to an SD file
sdfile.setFileName("test_mm.sdf");
sdfile.setMolecule(am);
sdfile.output();
quit;
```

Fig. 2 Xsi script for energy minimization by molecular mechanics

Execution through CUI using scripts is convenient for both computational chemists working on procedures by trial and error and researchers that are not specialists in computational chemistry who just execute routine calculations. The former can achieve high productivity, finding optimal conditions or procedures by making repeated modifications to parts of scripts they have created or transferring the parts to other applications. On the other hand, since implemented on a CUI basis, Web services can be provided readily, which is more desirable than a dedicated GUI that the latter is not used to.

This program is written in C++ and runs on Linux. By using this general-purpose technology, it is easily adapted to a new computer environment where there are continuous, rapid improvements. Since some of the classes are adapted to parallel computations, it runs on PC clusters made up of multiple CPUs and can perform large-scale calculations. Recently, compatibility with grid calculation environment also has been completed. At Sumitomo Pharmaceuticals, it is used on a dispersed memory PC cluster made up of one master node and

four slave nodes. Each node is equipped with an Intel Corp. Xeon 2.4 GHz CPU and 2 Gbytes of RAM, and it is run on SuSE Linux 7.3.

2. Module Configuration

The computational chemistry techniques used in drug discovery research are extremely diverse, and there is a lot of freeware and commercial software. Most of it is targeted at a specific type of analysis, and for convenience is equipped with necessary related functions. As a result, a duplication of functions arises among the pieces of software for different analytical targets. For example, with software for QSAR analysis based on three-dimensional structural descriptors and superposition analysis for multiple compounds, both require conformational analysis function. In integrated software targeted at general-purpose analysis, not only does this duplication of functions increase the trouble in development and maintenance of the software, but also the difference in specifications affects the results of analysis. With Xsi, objects are created such that the necessary functions are included, but without duplication, and classes and functions are constructed to be able to handle a variety of analytical goals by changing the combination of these.

On the other hand, in terms of individual functions, there are a considerable number of freeware and inexpensive commercial software solutions that are quite superior. These pieces of software are typically widely disseminated, and many researchers are familiar with them. The authors decided to actively use this existing software.

Xsi is made up of the 27 classes and 374 functions given in **Table 1**. Four classes have been prepared for file input functions. MOL and SDF formats are used for low molecular weight organic molecules, PDB format for proteins and CSV format for text data. Each of the formats is the standard file format for that application. Molecule and Universe are set up as the classes that store molecular information. The former stores a single molecule, and the latter stores amino acid sequence of proteins and other related data. In addition, classes for storing the constraint information for various molecular calculations are prepared for both. Ten types of

Table 1 Xsi function modules

Category	Xsi_Class	Xsi_Function	Description
File I/O	MoleculeFileMol	8	I/O for MOL format file (for single small molecule)
	MoleculeSetFileSD	9	I/O for SDF format file (for multiple small molecules)
	MoleculeSetFilePDB	11	I/O for PDB format file (for protein)
	CSVFile	22	I/O for CSV format file (for text data)
Molecular Information	Molecule	31	Information of small molecule
	Universe	9	Information of protein or multiple molecules
	Constraints	12	Constraints information for small molecule
	UniverseConstraints	20	Constraints information for protein or multiple molecules
Molecular Calculation	MolecularMechanics	19	Molecular Mechanics calculation
	MolecularOrbitalCNDO	2	Molecular Orbital calculation
	MonteCarlo	48	Monte Carlo simulation
	Docking	38	Docking simulation
	LigandAlignment	53	Ligand alignment
	RotationAngleReservoir	2	Storage of dihedral angles
	Descriptor	5	Strctural descriptor calculation
	Field	5	Projection of molecular properties onto grid cross sections
	PotentialField	8	Projection of potential energy onto grid cross sections
	RMSMinimizer	7	RMS calculation
Miscellaneous	Integer	6	Integer variable
	Real	5	Real number variable
	String	6	Character string variable
	RandomNumberGenerator	5	Gneration of randum numbers
	ClusterAnalyzer	10	Cluster analysis
	Similarity	5	Similarity search
	Clique	6	Clique search
	Statistics	10	Regression and discriminant analyses, etc.
	Utility	12	Operation of array etc.

classes are set up for molecular calculations. Among these, molecular mechanics, Monte Carlo calculations and docking analysis are provided for energy calculations. In these calculations, MMFF94s force-field parameters[13] and, for solvent effects, distance dependent dielectric constants and GB/SA models[14] are used. Docking analysis can carry out high-speed calculations using potential fields in addition to methods based on normal energy calculations. To compare molecules, coordinate system dependent characteristic values, such as pharmacophore and shape reflected on grid cross sections are used, as well as coordinate system independent ones (WHIM descriptors[15]) derived from the main component analysis on above. The characteristic values for molecules that Xsi can calculate are mainly based on three-dimensional structures. External programs should be used for structural descriptors computed from 2D structures such as LogP and the topological index.

Classes for numerical values and character strings other than these have been prepared in Xsi. These classes can be the subjects of four-rule operations using scripts. In addition, multidimensional arrays can be handled in a flexible manner, and Utility class has been prepared for conveniently handling these array operations. In Xsi, the results of molecular calculations and the characteristic values for molecules are output as CSV files, and it is assumed that statistical analysis and informatics analysis will be performed by external programs. For many statisticians, there is a great merit in being able to use the statistical analysis software they are used to using or the latest informatics techniques. Most of these external programs use CSV format, and it is easy to have the results of analysis reflected in the original molecular information taken up by Xsi. Some statistics and infromatics techniques frequently used in connection with molecular calculations are implemented in Xsi. Linear regression analysis, linear discrimination analysis, similarity calculations, cluster analysis and clique searches are among them.

3. General-Purpose Molecular Calculation Functions

The authors created scripts to carry out the calculations used frequently in drug discovery research (**Table 2**).

Generating a tertiary structure of a molecule and analyzing the possible conformations are typically the first processes in molecular calculations. Continuous, high-speed processing for these is needed because many molecules must be processed in a short period of time. Therefore, all of the scripts are written so as to be able to process multiple molecules continuously. In addition, those scripts covers wide variety of functions from just generating tertiary structures having standard bond lengths and bond angles to detailed conformation analysis to find minimum energy structures.

Table 2 General purpose functions for drug design programmed by Xsi scripts

LBDD/SBDD	Tertiary structure generator
	Residual conformation optimizer
	Conformation search for diverse set
	Confromation search for multiple minimum set
	Combinatorial conformation generator
	Substructure conformation search
LBDD	Structural descriptors calculation
	Molecular surface calculation (polar/non-polar/total)
	Similarity or k-nearest neighbor search
	Similarity matrix calculation
	Cluster analysis and diversity extraction
	Pharmacophore mapping with output of MOL format file
	Multiple ligand alignment based on pharmacophore similarity
	Regression and descriminant analysis with selection of explanatory variable
	Grid-projected descriptors calculation
	3d-QSAR (CoMFA/CoMSIA)
	Ligand alignment with conformation adjustment
SBDD	Random structure generator in active site of protein
	Water molecule mapping in active site of protein
	Potential field generator with output of PDB format file
	Grid-based Docking
	Cartesian-based Docking
	Substructure conformation search in active site of protein
	Binding energy calculation
	Molecular surface calculation for ligand/protein complex

Furthermore, functions for conformational analysis limited to specific substituents and generation of combinatorial libraries have been implemented.

In terms of scripts related to LBDD, variety of functions, such as calculation of the characteristic values of molecules, chemoinformatics, pharmacophore analysis, QSAR analysis and ligand alignment, have been implemented. It is possible to make output for pharmacophore in MOL format, which can be used to make displays corresponding to the molecular structure using a molecular structure browser such as PyMol.[16] In addition, advanced analysis is possible by giving the characteristic values for the molecules to various informatics analysis programs via CSV format.

It is as well with SBDD scripts, and many functions are covered. For docking analysis, we have implement-

ed both high-speed analysis using potential fields and detailed analysis based on accurate energy calculations that include protein motion. Docking analysis with modification only at a specific substituent is possible. Furthermore, based on the binding models obtained from docking analysis, calculations of binding energy coupled with the free energy of hydration and calculations of various surface areas that are indicated as correlating with binding energy can be carried out.

4. Complementary Use of External Software

The external software that is used to complement Xsi is listed in **Table 3**. These pieces of software carry out data exchange using the general-purpose file formats of SDF, PDB and CSV. For other file formats, the use of BABEL and other conversion programs is envisioned.

Table 3 External softwares used with Xsi

Category	Function	Software
Molecular Viewer		PyMol
Molecular Editor	2D	ISIS Draw, Chem Draw
	3D	DS ViewerPro
	R-substitution	Accord for Excel
	Database	ISIS Base, ChemFinder
	Protein	PDB Viewer
File Format Converter		BABEL
Analysis	Statistics	SPSS, S-PLUS
	Infomatics	RandomForest, LibSVM, NEUROSIM, BayesiaLab
Molecular Calculation	Molecular Properties	Dragon, WSKOW
	Molecular Dynamics	TINKER
	Molecular Orbital	MOPAC

In terms of advanced utilization of external software in connection with Xsi, there is a case of carrying out docking studies continuously on derivatives where transformations with only specific substituents are carried out based on compounds with known or predicted binding modes for the target protein. In this instance, there is a need to prepare the common substructure and a set of substituents for the transformation. It is necessary to modify the molecular structure while maintaining the original binding mode, using DS Viewer Pro.[17] It is convenient to use the "R Group Table" function of the Accord for Excel[17] for the preparation of the substituents set. Normally, molecular fragments introduced at specific sites are selected from reagents having suitable functional groups in the synthesis scheme. Therefore, if multiple compounds extracted from a reagent database are taken up in the Accord for Excel and applicable functional groups specified in the "R Group Table" function, the fragments to be introduce and the binding sites are all automatically set for the molecule. If the conformation data for the common core substructure above is also given to Xsi, not only can suitable molecules be constructed, but also conformational analysis and docking while changing only the introduced fragment parts can be carried out automatically and continuously.

Multiple Docking Analysis

To address problems with LBDD and SBDD simultaneously, namely prediction of unknown parts in structure-activity relationships and improvement of docking precision, the authors devised the multiple docking analysis method.[9), 10)] The characteristic of this method is obtaining binding models for multiple ligands with known biological activity simultaneously. Various evaluation functions are calculated for the following docking studies and biological activity predictions based on the binding models obtained for the multiple molecules. Virtual screening can be carried out using these evaluation functions for compounds with unknown biological activity.

1. Docking Studies for Multiple Molecules

From many examples, it is known that derivatives with common core substructures exhibit similar binding modes for the same target protein (Fig. 1). Using this knowledge, we start with simultaneously obtaining binding models for multiple known ligands having common core substructures in the multiple docking analysis.

Multiple molecule docking studies carry out stepwise searches of binding models for common core substructures and peripheral substituents (**Fig. 3**). The calculation procedures at each stage are shown in **Fig. 4**, but both stages are based on the same calculation procedures. In other words, binding model candidates for each molecule are listed according to the target protein complementarity, and the optimal binding model combination for the entire molecules is found based on the similarity of their pharmacophore, shape and the like. A graph algorithm technique called "clique search" is used for the latter, and combinations of binding model candidates where the multiple molecules are more similar to each other are searched for. To make

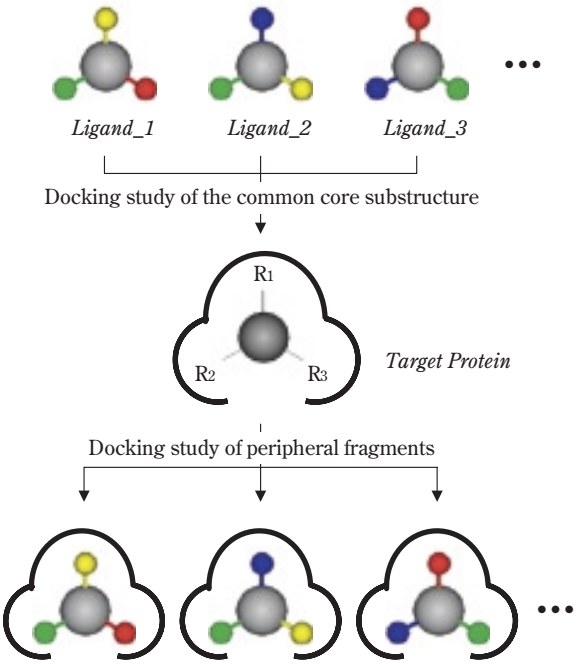


Fig. 3 Scheme for the stepwise docking study

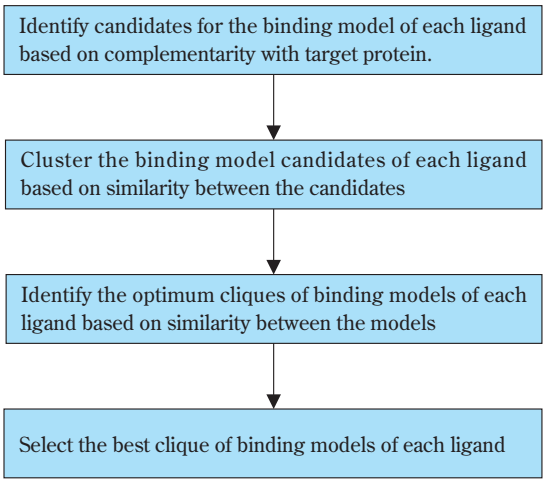
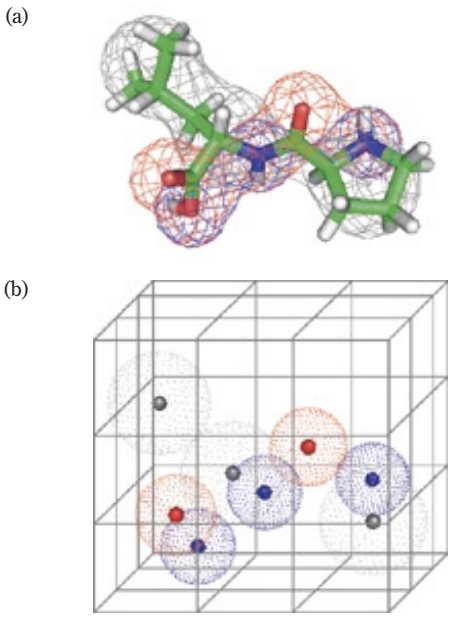


Fig. 4 Procedure of docking study for both common core substructure and peripheral fragments



Hydrogen bond acceptor (red), Hydrogen bond donor (blue) and Hydrophobic center (gray).
(a) Molecular properties mapped on a model compound.
(b) Molecular properties projected onto grid cross sections

Fig. 5 Examples of molecular properties

this search easier, there is a need to extract the diverse binding model candidates for each molecule first. Similarities in pharmacophore, shape and the like are calculated based on characteristic values such as the hydrogen binding ability, the charge and the hydrophobic properties for each binding model projected onto common grid cross sections (**Fig. 5**).

The binding model searches on common core substructure and peripheral substituent are conducted using the same calculation procedures in this manner, but the methods for each procedure are somewhat different for the two. First of all, in the procedure for finding binding model candidates from how they are complementary with the target protein, the former conducts a rough search over a wide range in the neighborhood of the active site, while the latter searches a limited area in detail. Therefore, in the former, high-speed but lower precision calculations carry out searches based on the potential field,[18] and in the latter highly precise but slower calculations carry out searches based on the Cartesian coordinate systems. The latter also take into account changes to the protein structure as necessary. In the procedure for extracting a variety of binding model candidates for each molecule, the former uses the center of each cluster obtained from a cluster analysis, while the latter uses the minimum energy binding model candidates. In the method for determining the optimal binding model combination, similarity in common core substructure is used as the index for the former, and the latter is based on indices such as correlation of the binding energy with the biological activity that will be discussed later.

Each of the binding models of the multiple molecules obtained from the analysis above is complementary to the target protein, and at the same time there are good correspondences between the models. In other words, it is a method that supplements the insufficiency of the evaluation functions and optimal solu-

tion searches in general docking methods through similarity among binding models. In addition, it is possible to efficiently carry out wide-ranging and detailed searches through stepwise searching for common core substructures and peripheral substituents and the selection of calculation methods according to the goals of each procedure.

2. Creation of Evaluation Functions and Virtual Screening

Evaluation functions for carrying out virtual screening of compounds having unknown biological activity are created based on the binding models obtained from the multiple molecule docking study. Virtual screening is carried out in two steps, docking for the target protein and prediction of the biological activity based on the binding models obtained (**Fig. 6**). Therefore, evaluation functions are created for both steps.

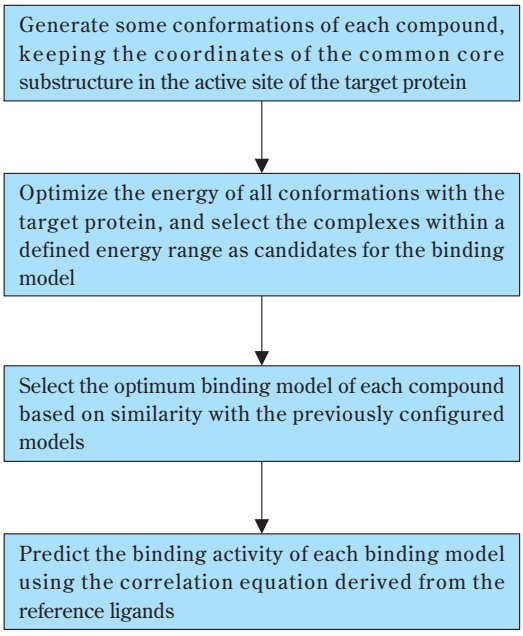


Fig. 6 Procedure of virtual screeing based on the multiple docking model

In the docking studies, the complementarity with the target protein and the similarity with existing binding models are evaluated stepwise or simultaneously. The purpose of the former is to list the binding model candidates, so a general-purpose evaluation function is used as is. By the latter it is determined if the molecule in question can bind to the target protein, and, if it can, the optimal binding model for the molecule is selected. Therefore, the docking evaluation function is to evaluate the similarity with existing binding models. It is possible to add a complementarity score for the target protein into this evaluation function.

Molecular characteristics projected onto common grid cross sections are used in evaluating the similarity with existing binding models. Comparatively simple methods can be used, such as the K-Nearest Neighbor method[19] and methods that evaluate the similarity to the averaged characteristic values for multiple molecules projected onto the grid cross sections. More advanced evaluation methods, like such informatics analytical methods as the One-Class SVM method[20] and SOM method[21], are also applicable. In any event, existing binding models are used as learning data, and methods with fewer false negatives are used. Adding binding model candidates that differ from the true binding state and binding model candidates for known inactive molecules to the learning data might be effective to reduce false positives. These methods lead to an accumulation of structure-activity relationship information and perfect the learning model, and in the end they can be expected to improve the predictive precision. If the structure of the complex for the molecule and the target protein is obtained experimentally, it is possible to correct the binding models for other molecules by substituting the experimental data for the binding model of that molecule.

Quantitative prediction of biological activity based on binding models is a challenging problem. The binding activity of ligands for the target protein is determined by the difference in the free energy between the states of binding and not binding in a reversible equilibrium state (**Fig. 7**). However, it requires a large amount of calculations to accurately estimate the entropy contribution and electrostatic interactions for the complex

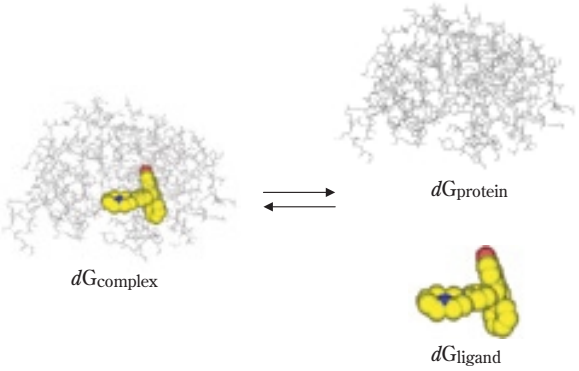


Fig. 7 Free energy of binding between ligand and target protein

system that includes the solvent molecules. In recent years, calculations[22] of free energy through generalized-ensemble molecular dynamics simulations, and accurate estimates[23] of interaction energy using molecular orbital calculations have been attempted, but their throughput is far low from that required for virtual screening. For realistic resolution, the authors calculate approximate values for binding energy from the potential energy using force field functions and hydration free energy based on continuous system model. In this calculation, a term corresponding to the interaction energy between the target protein and the ligand and a term corresponding to the hydration energy difference are separated, and a regression formula is configured with known biological activity as the index. There are often cases where the force field functions cannot make an accurate estimation of electrostatic interactions. To circumvent this problem, a regression formula using the surface area exposed to the solvent with and without polarity can sometimes make a good prediction of the biological activity.[24] After testing these two methods , one that made more accurate predictions of the existing structure-activity relationships should be used.

3. Usefulness of Multiple Docking Analysis

The virtual screening technique described above performs analysis by suitably combining the complementarity between the target protein and the compound with similarity between multiple compounds. Only an integrated software like Xsi, which can seamlessly joins the functions of SBDD and LBDD, has made it possible for the first time.

This technique differs from typical docking studies, as some structure-activity relationship information on the derivative compounds is necessary as prior experimental data. But cases where such information is not available are rare, because usually, based on the structures of HTS hit compounds or other biologically active compounds, derivatives are synthesized and tested for bioactivity. Naturally, the tertiary structure of the target protein is necessary for the multiple docking analysis. When coordinate data that has been obtained experimentally cannot be used, it can be constructed using the homology modeling. In typical docking studies, modeled protein structure is used without questioning the accuracy of the structure. In addition, only insufficient consideration is given to changes in the protein structure that depend on the binding ligand. But with the multiple docking analysis, because an evaluation function is constructed using known structure-activity relationship information, it indirectly verifies the suitability of the target protein structure for carrying out virtual screening. In addition, unlike general docking studies, accumulation of structure-activity relationship information or structure determination of protein-ligand complexes using X-ray crystallography can improve the predictive precision.

Application in Phosphodiesterase-4 Inhibitors

To verify the usefulness of the multiple docking analysis, we carried out a model experiment simulating an actual drug discovery. The object selected for the model experiment was Phosphodiesterase-4 (PDE-4) inhibitor for which research and development is progressing globally with expectations as a drug for treating asthma and other diseases. Rolipram has been known for a long time (**Fig. 8**) as a compound that inhibits PDE-4, and the results of docking studies using commercial SBDD software (Dock, FlexX and AutoDock) from multiple groups were reported in 2002.[25, 26] The following year, the structure of the complex from X-ray structural analysis was reported (PDB, 10YN), revealing the binding models were completely different from the true binding mode.[27] In other words,

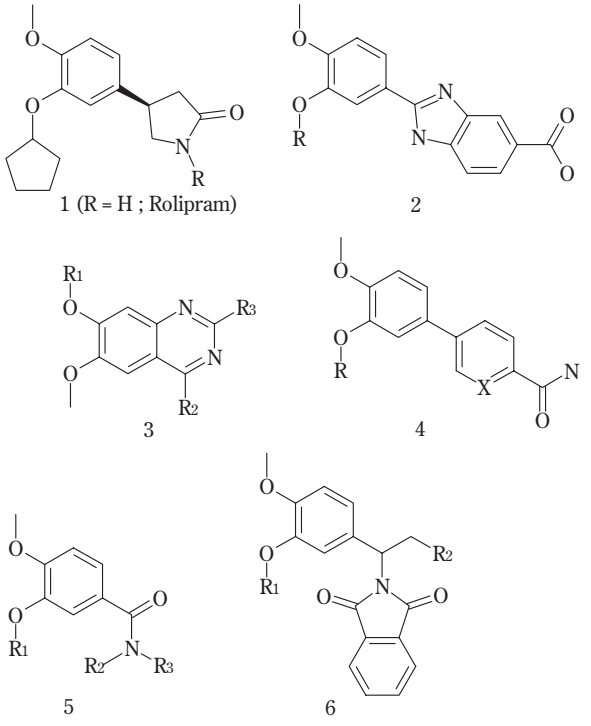


Fig. 8 Analogs of PDE-4 inhibitors

PDE-4 is a difficult target for predicting the binding mode of inhibitors computationally.

In the model experiment, 78 compounds, from six types of catechol derivatives exhibiting PDE-4 inhibitory activity were selected as the objects of the analysis (Fig. 8).[28)–33)] The multiple compound docking study was carried out for a total of 12 compounds out of these, two randomly chosen compounds from each derivative type. As in the referenced reports, ligand-free state obtained from X-ray structural analysis was used for the PDE-4 coordinate data (PDB; 1FOJ).[34)] When the analysis was carried out following the previously described procedure (Fig. 4), good binding models were obtained, except for one compound (**Fig. 9**). Rolipram was included in these 11 compounds, so a comparison was made with the structure of the complex obtained from X-ray structural analysis. As a result, the binding model obtained from the calculation was good reproduction of the true binding state, and the RMS value for both hetero atoms was 1.09 Å (**Fig. 10**).

Based on the binding models for the 11 compounds, virtual screening was carried out on the remaining 66 compounds (including 8 inactive compounds). The evaluations of similarity with the binding models for the 11 compounds in the docking study were carried out based on averaged information for the characteristic values for multiple molecules projected onto grid cross sections. The validity of the binding models was determined based on energy for the complex with PDE-4 and the interaction energy and hydration energy that consist the complex formation. In other words, compounds with either of the energy values very large were judged to be inactive. As a result, binding models were obtained for 48 compounds (83%) out of the 58 active compounds.

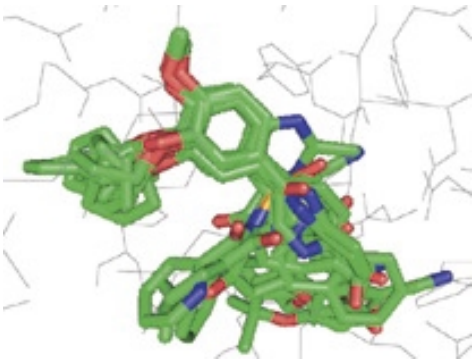


Fig. 9 Binding models of 11 kwown PDE-4 inhibitors

Fig. 10 The X-ray structure of Rolipram bound to PDE-4 and the binding model by multiple docking analysis

We investigated predicting the biological activity values for these 48 binding models. First of all, we considered the binding energy calculated from the interaction energy and the free energy for hydration, but determined that it was not suitable because its correlation with the biological activity was a negative correlation. Then we focused only on the interaction energy, but it was still insufficient for predicting the biological activity even though a positive correlation was observed. Finally, when we tried the method based on the polar and non-polar surface area exposed to the solvent, we were able to obtain comparatively good results. In more detail, we obtained the regression formula in Equation (1) as the result of a linear regression analysis carried out with the difference between polar and non-polar surface area exposed to the solvent used as a descriptive variable and the biological activity value as the index. The correlation coefficient (R) for this regression formula was 0.733, and five compounds were eliminated from the analysis as outliers. **Fig. 11** shows the relationship between the predicted values for biological activity and the values obtained experimentally. While a little variation can be seen, we can see that good predictions were obtained for the pIC_{50} values across a little under a 3rd order range. The lower limit of this range is the level of biological activity intensity expected in hit compounds in HTS or other bioassays, and the upper limit is a level that can make for candidates of development. In other words, the results is sufficient enough to be used in a drug discovery research program. On the other hand, of the eight inactive compounds included for the virtual screening, biological activity values were calculated from Equation (1) for the six compounds for which binding models were obtained. As a result, two compounds gave pIC_{50} val-

ues around 6.5, which indicated comparatively strong activity, but the other four compounds were predicted with weak values of 5.5 or less. The pIC_{50} values from experiments for these compounds were 5 or less, and they might have, if any, weak biological activity. In any event, since this virtual screening is used for the purpose of activity improvement, it is sufficient for discriminating inactive compounds or ones with weak activity.

$$pIC50 = 1.303 + 0.00441 * Polar\_ASA + 0.00494 * Non\ Polar\_ASA,\ n = 43,\ R = 0.733 \quad (1)$$

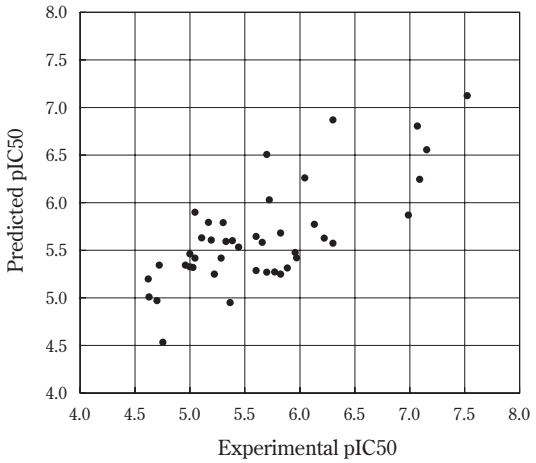


Fig. 11 Correlation between experimental and predicted pIC50's of PDE-4 inhibitors

Conclusion

We have described an overview of the development of Xsi, which is integrated software for the various LBDD and SBDD functions and an example of application. The example shown here is no more than a model experiment, but Xsi has already been used in various drug discovery research programs, and the record shows its validity. In addition, though we did not touch upon it in this paper, the applications of Xsi are not limited to virtual screening, and it is being used for supporting ADME and toxicity predictions and reverse proteomics research. The functions of Xsi can be extended as needed, and development of the next version and planning for the version after that are in progress. In the future, along with improving the reliability of the program by increasing the users, we expect the appearance of even more advanced virtual screening methods through the use of this software.

Finally, we would like to express our deep gratitude to those involved with the Biotechnology Laboratory at our joint development partner for Xsi, Mizuho Information & Research Institute, Inc.

Reference

1) A. M. Brzozowski, A. C. W. Pike, Z. Dauter, R. E. Hubbard, T. Bonn, O. Engstrom, L. Ohman, G. L. Greene, J. A. Gustafsson, and M. Carlquist, *Nature*, **389**, 753-758 (1997)
2) J. Renaud, S. F. Bischoff, T. Buhl, P. Floersheim, B. Fournier, M. Geiser, C. Halleux, J. Kallen, H. Keller, and P. Ramage, *J. Med. Chem.*, **48**, 364-379 (2005)
3) C. L. Kuo, H. Assefa, S. Kamath, Z. Brzozowski, J. Slawinski, F. Saczewski, J. K. Buolamwini, and N. Neamati, *J. Med. Chem.*, **47**, 385-399 (2004)
4) K. Jozwiak, S. Ravichandran, J. R. Collins, and I. W. Wainer, *J. Med. Chem.*, **47**, 4008-4021 (2004)
5) A. A. Soderholm, P. T. Lehtovuori, and T. H. Nyronen, *J. Med. Chem.*, **48**, 917-925 (2005)
6) R. Ragno, M. Artico, G. D. Martino, G. L. Regina, A. Coluccia, A. D. Pasquali, and R. Silvestri, *J. Med. Chem.*, **48**, 213-223 (2005)
7) http://www.mizuho-ir.co.jp/science/xsi/index.html
8) Y. Inagaki, M. Hamada, K. Yamazaki, M. Kanaoka, and H. Chuman, Chem-Bio Informatics Society 2004 Proceedings (2004)
9) K. Yamazaki, and M. Kanaoka, PCT/JP02/11401 (Sumitomo Pharmaceuticals)
10) K. Yamazaki, M. Kanaoka, and Y. Inagaki, Chem-Bio Informatics Society 2004 Proceedings (2004)
11) http://www.tripos.com/
12) http://www.chemcomp.com/
13) T. A. Halgren, *J. Comp. Chem.*, **17**, 490-641 (1996)
14) W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, *J. Am. Chem. Soc.*, **112**, 6127-6129 (1990)
15) R. Todeschini, M. Lasagni, and E. Marengo, *J. Chemometrics*, **8**, 263-272 (1994)
16) http://pymol.sourceforge.net/
17) http://www.accelrys.com/
18) I. D. Kunts, *Science*, **257**, 1078-1082 (1992)
19) D. Chema, and A. Goldblum, *J. Chem. Inf. Comput. Sci*, **43**, 208-217 (2003)
20) E. Byvatov, and G. Schneider, *J. Chem. Inf. Comput. Sci.*, **44**, 993-999 (2004)
21) Z. R. Yang, and K. C. Chou, *J. Chem. Inf. Comput. Sci.*, **43**, 1748-1753 (2003)
22) A. Mitsutake, Y. Sugita, and Y. Okamoto, *Biopoly-*

mers, **60**, 96-123 (2001)
23) K. Kitaura, E. Ikeo, T. Asada, T. Nakano, and M. Uebayasi, *Chem. Phys. Lett.*, **313**, 701-706 (1999)
24) J. S. Bardi, I. Luque, and E. Freire, *Biochemistry*, **36**, 6588-6596 (1997)
25) O. Dym, I. Xenarios, H. Ke, and J. Colicelli, *Mol. Pharmacol.*, **61**, 20-25 (2002)
26) P. Pospisil, T. Kuoni, L. Scapozza, and G. Folkers, *J. Recept. Sig. Trans.*, **22**, 141-154 (2002)
27) Q. Huai, H. Wang, Y. Sun, H. Y. Kim, Y. Liu, and H. Ke, *Structure*, **11**, 865-867 (2003)
28) I. L. Pinto, D. R. Buckle, S. A. Readshaw, and D. G. Smith, *Bioorg. Med. Chem. Lett.*, **3**, 1743-1746 (1993)
29) J. B. Cheng, K. Cooper, A. J. Duplantier, J. F. Eggler, K. G. Kraus, S. C. Marshall, A. Marfat, H. Masamune, J. T. Shirley, J. E. Tickner, and J. P. Umland, *Bioorg. Med. Chem. Lett.*, **5**, 1969-1972 (1995)
30) R. J. Chambers, A. Marfat, J. B. Cheng, V. L. Cohan, D. B. Damon, A. J. Duplantier, T. A. Hibbs, T. H. Jenkinson, K. L. Johnson, K. G. Kraus, E. R. Pettipher, E. D. Salter, J. T. Shirley, and J. P. Umland, *Bioorg. Med. Chem. Lett.*, **7**, 739-744 (1997)
31) J. G. Montana, G. M. Buckley, N. Cooper, H. J. Dyke, L. Gowers, J. P. Gregory, P. G. Hellewell, H. J. Kendall, C. Lowe, R. Maxey, J. Miotla, R. J. Naylor, K. A. Runcie, B. Tuladhar, and J. B. H. Warneck, *Bioorg. Med. Chem. Lett.*, **8**, 2635-2640 (1998)
32) G. W. Muller, M. G. Shire, L. M. Wong, L. G. Corral, R. T. Patterson, Y. Chen, and D. I. Stirling, *Bioorg. Med. Chem. Lett.*, **8**, 2669-2674 (1998)
33) B. Charpiot, J. Brun, I. Donze, R. Naef, M. Stefani, and T. Mueller, *Bioorg. Med. Chem. Lett.*, **8**, 2891-2896 (1998)
34) R. X. Xu, A. M. Hassell, D. Vanderwall, M. H. Lambert, W. D. Holmes, M. A. Luther, W. J. Rocque, M. V. Milburn, Y. Zhao, H. Ke, R. T. Nolte, *Science*, **288**, 1822-1825 (2000)

PROFILE

Kazuto YAMAZAKI

Sumitomo Pharmaceuticals Co., Ltd.
Research Division
Chemistry Research Laboratories
Molecular Design Research Group
Research Scientist

Masaharu KANAOKA

Sumitomo Pharmaceuticals Co., Ltd.
Research Division
Genomic Science Laboratories
(formerly, Molecular Design Research Group)
Research Director Ph. D