Consensus Scoring to Improve the Predictive Power of in-silico Screening for Drug Design

Masato Okada
Faculty of Science and Technology,
Tokyo University of Science,
2641 Yamazaki, Noda,Chiba, 278-8510, Japan

Masato Tsukamoto
Faculty of Pharmaceutical Sciences,
Tokyo University of Science,
2641 Yamazaki, Noda,Chiba, 278-8510, Japan

Hayato Ohwada
Research Institute for Science and Technology,
Tokyo University of Science,
2641 Yamazaki, Noda,Chiba, 278-8510, Japan

and

Shin Aoki
Faculty of Pharmaceutical Sciences,
Tokyo University of Science,
2641 Yamazaki, Noda,Chiba, 278-8510, Japan

ABSTRACT

In this study, we implement a docking simulation system that is useful for drug development. This system uses many docking software programs. This method, called consensus scoring, improves the accuracy of docking simulation and in silico screening. In this study, we standardize two or more scores to be mutually computable. We then compute the scores. We use experimental docking simulation and compute scores with each method. This evaluation indicated that the accuracy of docking simulation can be improved.

Keywords: docking simulation, drug development, consensus scoring, in silico screening, standardize.

1. INTRODUCTION

Due to the recently improved performance of computers and software, simulation in drug discovery has become possible. In particular, the search for a compound that controls a specific protein related to sickness (target protein) has been researched. This technique is called in silico screening.

Currently, the main technique used in searching for a medicine candidate compound is high-throughput screening (HTS). In HTS, much of the compound is automatically docked with the target protein. However, when HTS docks compounds with the target protein, the cost of preparing and managing the compound is significant. To solve this problem, in silico screening is required.

In silico screening simulates the docking of the compound and the target protein conducted with HTS. The technique for simulating docking by using the structure of the protein is called in silico docking screening. Many software programs for docking simulation (e.g., FlexX [1], AutoDock [2], DOCK [3], and GOLD [4]) have been developed.

The accuracy of docking software is verified based on the results of an actual docking experiment. Necessarily, the score rises when a specific compound (known ligand) that has a binding affinity to the target protein is simulated even if which docking software is used regardless of what docking software is used.

However, it is dangerous to trust a single docking software program. This depends on compatibility among the target protein, the compound, and the software. As a result, the scores of the docking software may differ greatly.

Thus, in the present study, the problem with single software is solved by simulating two or more docking software programs, and performing consensus scoring that uses all scores. It looks for the compound that has a high score from each docking software program, as well as an already-known ligand in consensus scoring. As a result, the compound with the highest possibility of acting on the target protein can be selected.

However, the distributions of the scores from the software programs differ greatly. Thus, a process is needed for the use of these scores [5]. In the present study, we standardize each score so that it can be treated easily.

2. RELATED WORKS

Various studies have used consensus scoring. Charifson et al. [6] used two docking software programs, 13 scoring functions, and consensus among these scores. This study confirmed that consensus scoring is an effective way to obtain improved hit rates in various virtual database screening studies.

M. Okamoto [7] customized the DOCK program and used three scoring functions (DOCK4, FlexX, and PMF). Consensus scoring involved choosing the worst of the three scores. In this study, they chose 100 medicine candidate compounds and some compounds that inhibit DAP Kinase. We referred to this study for the composition of our system.

3. SOFTWARE

In this section, we describe the software used in the present study. To solve the problem of HTS, it is necessary to consider cost. For this reason, we use some software programs in screening that can be used free of charge.

3.1 AutoDock Vina

AutoDock Vina [8] (Vina) is an open source program for drug discovery. This program was based on AutoDock for precision enhancement and improved speed. For this reason, we describe both Vina and AutoDock.

AutoDock is a widely used suite of automated docking tools, and many examples of its successful application are available in the literature. Both Vina and AutoDock are free software that can be run on Linux, Mac, and Windows. AutoDockTools is the graphical front-end program. AutoDock (AutoGrid) can run on AutoDockTools, but Vina basically runs on the command line. The features of Vina are as follows.

·Specify the computational domain in cuboids.
·Support multithreading on multi-core machines.
·The file format PDBQT, created by AutoDockTools, is used.

We run Vina with various computers in our laboratory. Each computer uses Windows and or Linux. The performances of these computers are not uniform. However, the total CPU-cores exceeds 50.

In the present study, we run this software through the JAVA program for managing many files. As a result, we can complete one job after another, with maximum machine performance.

3.2 DOCK (version 6)

DOCK [3] is the oldest software for docking simulation. DOCK can be used for free for education and research purposes, but a fee is charged for commercial use. For this reason and for high accuracy, this software has been used for various studies. It can be run with Linux and Macintosh; however, it is necessary to use Cygwin in Windows.

DOCK has the following features.

·It uses a sphere on the surface of the target protein. The sphere specifies the compound position and computational domain
·It corresponds to the MPICH library.
·When docking by initialization, 20% of the compounds are not simulated correctly.
·Therefore, adjustments are necessary to decrease the error.

UCSF Chimera [9] has been prepared to create the compound file for DOCK. UCSF Chimera is a highly extensible program for interactive visualization and analysis. Chimera can add hydrogen and a force field to a compound. Afterward, it creates an MOL2 format file for DOCK.

We run this software on Linux. Therefore, we use various computers in our laboratory without Windows.

3.3 Discovery Studio (version 2.5)

Discovery Studio [10] is an application package for molecular modeling and simulation. This commercial software was developed by Accelrys, Inc. Since our university has purchased this software, we may use it free of charge for docking screening

In Discovery Studio, one docking function, LibDock, is faster than any other function. Therefore, we mainly use this function for docking screening. Although CDocker has higher accuracy, it takes ten or more times longer than LibDock. For this reason, we use CDocker only when compounds have been screened and reduced.

In the present study, we run this software in the High-Performance Computing System of the Tokyo University of Science, where we can use 64 cores or less. In addition, each CPU is a 3.0GHz Core2 generation CPU.

4. RANDOM DATA SET

To validate the accuracy, we conducted docking simulations with 1000 random compounds. These compounds were acquired from Chembl [11] at random. These files are MOL formatted. However, large compounds are excluded because they cannot become medicine.

In addition, we docked these compounds with a specific protein, X-linked Inhibitor of Apoptosis Protein (XIAP) (PDBID: 3HL5), which is related to cancer. If we can control this protein, many types of cancer can be treated. For this reason, we have been searching for a compound that can control XIAP.

5. DATA PREPROCESSING

The compound data acquired from Chembl cannot be used for docking simulation without some processing.

5.1 Three-Dimensional Transformation and Ionization

The compound data acquired from Chembl is two-dimensional. If docking simulation is conducted with this data, an error will occur, depending on the software. For this reason, the compound data must be transformed to three-dimensional data.

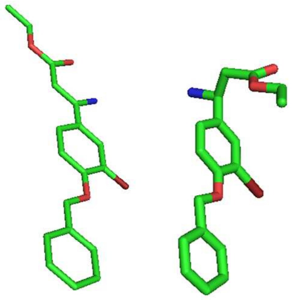

Fig. 1. 3D transformation from the MOL file

Table 1. Computing Time

Software	CPU	OS	Time(minute)
AutoDock Vina	Intel Xeon 5520 2.26GHz	Windows 7 64bit	5874
DOCK	Intel Xeon 5520 2.26GHz	Ubuntu 8.04	999
LibDock	Intel Xeon 5160 3GHz	Red Hat Enterprise Linux 6	142

Additionally, acquired data do not include charge and ion information. Docking software that cannot calculate this information cannot treat an ionic bond. Because the ionic bond is strong, this problem affects each score.

In the present study, we use Discovery Studio to transform data, ionization, and hydrogenation. The three-dimensional transformation is depicted in Fig. 1. The left figure presents the two-dimensional data acquired from Chembl. Discovery Studio transformed this data as presented in the right figure.

5.2 Format Conversion

Transformed compound data can be used for docking simulation. However, if the MOL file includes multiple compound data, an error will occur, depending on the software. For example, the software can process only the first compound but cannot identify each compound.

To solve this problem, we convert the file format to PDB and rename each compound to distinguish them. For conversion, we use the OpenBabel command line version [12].

6. DOCKING SIMULATION

After doing the processing described in section 5, we use these transformed compound for docking simulation. In the present study, we processed three kind of docking software with some computers. We used 1000 random compounds for docking simulation.

Table 1 shows the CPU using for docking simulation and computing time of each software. Because the Xeon processor is a multi-core processor, we can process the docking simulation in parallel. However, AutoDock Vina and DOCK, we used only one core and one CPU to measure the computing time. Additionally, in Xeon 5520, we cut these CPU functions, Intel Hyper-Threading Technology and Intel Turbo Boost Technology. In this experiment, we used the default status in AutoDock Vina, DOCK, and LibDock without number of CPU cores.

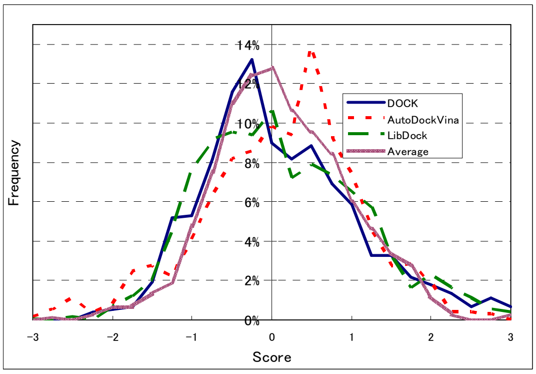

Fig. 2. Standardization

If docking simulation is processed in parallel, the computing time decrease in proportion to a parallel number. However, there is a limit at the speed improvement rate. We should make a parallel number moderate in MPICH or another multi-threading computing. Therefore, if you use a lot of compound, you should use docking software in task parallelism with many computers. These docking software, AutoDock Vina and DOCK, can be processed with a general personal computer.

7. SCORE CONVERSION

As described in section 1, the distribution of the scores from the different software programs differs greatly. For consensus scoring, these scores must be converted to be mutually computable.

7.1 Score Distribution

The range of the score obtained from each software program is as follows.

· AutoDock Vina:	-20~0
· DOCK:	-100~0
· Discovery Studio (LibDock):	0~200

The smaller score is better in Vina and DOCK, whereas the larger score is better in Discovery Studio.

7.2 Score Standardization

In the present study, we standardized each score to calculate these different distributions. For this processing, each distribution is considered to be normal. The following expressions used for standardizing the normal distribution are also used for conversion.

$$X = \frac{(x-\mu)}{\sigma}, \qquad (1)$$

Where μ is the average value of each distribution, σ is the standard deviation of each distribution, x is each score, and X is the converted score.

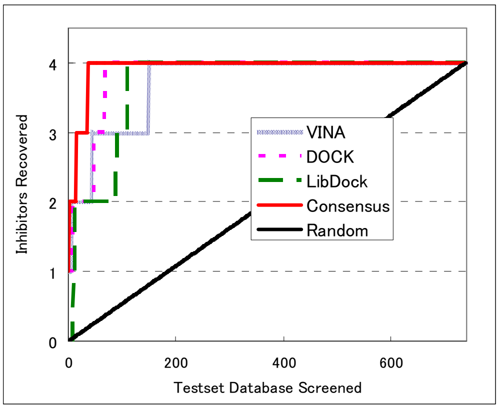

Fig. 3. Result of consensus scoring

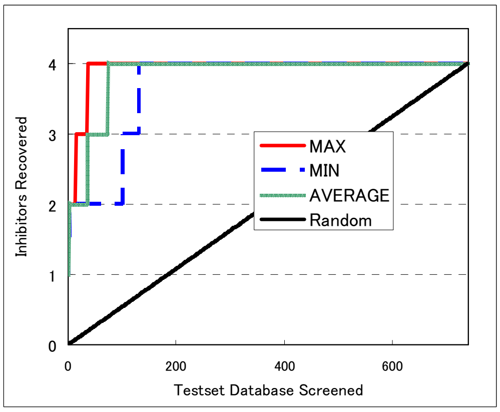

Fig. 4. Comparison of scoring methods

When each score is converted in this way, the distributions become equal. However, because the smaller score is better in Vina and DOCK, the Vina and DOCK scores must put minus be negative.

The distribution after this process is plotted in Fig. 2. The lines indicate the frequency of scores of the docking software. The average is the average score of each docking software program for each compound.

In this experiment, 1000 random compounds were docked to the target protein. As a result, 780 scores were obtained. In other words, 220 compounds were not simulated correctly. As a result of standardizing, the scores of each software program are distributed in the same range (Fig. 2). Additionally, all distributions have a similar shape.

In the present study, we use this distribution for consensus scoring.

8. METHOD OF CONSENSUS SCORING

This section describes three methods of consensus scoring. Each score is standardized, and then each compound is ranked. The important one is whether an already-known ligand is located in the high rank. It is important that a known ligand is highly ranked. In the present study, four compounds are treated as known ligands, and their activity with XIAP is already known. Additionally, information is available on the X-ray crystal structure analysis of the structure when these compounds are docked with the target protein.

In consensus scoring, the score of each software program is computed, and the compounds are ranked by consensus score. A general method for calculating the score is as follows.

The maximum (MAX) method takes the worst score of the compound. Therefore, if one software program produces a bad result, the consensus score of the compound is worse. Conversely, if the consensus score with the MAX method has a good value, the compound has a good score in all the software programs. The compound might thus affect the target protein as well as the known ligand.

The minimum (MIN) method takes the best score of the compound. Therefore, if one software program produces a good result, the consensus score of the compound is better. However, the other scores, excluding the best score, are not related to the consensus score; as a result, the consensus score rises even if there is a bad score. Because the simulation result greatly differs from the known ligand, such a compound will not affect the target protein.

The average (AVE) method takes the average score of the compound. This is an intermediate method between MAX and MIN. However, in contrast with MIN, if only one software program produces a good score, the consensus score is worse. This method moderates the weak points of MAX and MIN. For example, if a bad value is produced by chance, the consensus score becomes worse with the MAX method. With AVE, three scores are averaged, and the effect of the worst score is reduced. For this effect, AVE might use phased screening.

9. RESULTS AND DISCUSSION OF CONSENSUS SCORING

This section describes the results of consensus scoring using a random data set. To evaluate consensus scoring, we compare consensus scoring with each software program. Additionally, we compare each method of consensus scoring.

In the evaluation experiment, each compound is displayed in order of score. At this time, the rank of the known ligand is important, since accuracy is related to the rank of the known ligand. If the rank of each known ligand is lower, accuracy is better.

9.1 Comparison of Consensus Scoring with Each Software Program

Figure 3 indicates the accuracy of consensus scoring with each software program. We use the MAX method in consensus scoring. The horizontal axis represents the number of times test data is processed, and the spindle is the number of times the known ligand is detected. Each peak in the graph represents the rank of a known ligand. Therefore, if the rank of each known ligand is better and accuracy is better, the graph approaches the left. In contrast, if the graph is near the line of random nearly a random line, the software or scoring method is equally random. Using each software program with the unit, all known ligands were detected before and behind around 100th place: 69th place in DOCK, 111th place in LibDock, and 149th place in Vina. In contrast, consensus scoring with the MAX method detected all ligands in 36th place; thus, consensus scoring detected all ligands with a high rank. This data suggests that our method of consensus scoring can improve the accuracy of docking simulation.

9.2 Comparison of Methods

Figure 4 plots the accuracy of the consensus scoring methods. In this figure and in Fig. 3, one graph indicates the results of each method. Each method detected all known ligands differently: 36th place with the MAX method, 132nd place with the MIN method, and 73rd place with the AVE method. Therefore, the MAX method was the best.

10. CONCLUSION

In the present study, we focused on silico screening, which uses many docking software programs. This study demonstrated that consensus scoring improves the accuracy of docking simulation. Additionally, standardized scores can be computed easily. We conclude that consensus scoring with the MAX method is the best.

Currently, we are conducting large-scale screening. We use a commercial database that includes more than 4,000,000 compounds, and we use the target protein XIAP. Further studies are needed to improve the accuracy and speed of docking simulation. In the future, we will focus on new docking software program and machine learning.

11. ACKNOWLEDGEMENT

We thank Research Institute for Science and Technology in Tokyo University of Science for the financial support.

12. REFERENCE

[1] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. **A fast flexible docking method using an incremental construction algorithm.** J. Mol. Biol. 1996, 261, 470–489.

[2] Morris, GM.; Goodsell, DS.; Halliday, RS.; Huey, R.; Hart, WE.; et al. **Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.** Journal of Computational Chemistry 1999, 19: 1639–1662.h

[3] Ewing, T. J.; Kuntz, I. D. **Critical evaluation of search algorithms for automated molecular docking and database screening.** J. Comput. Chem. 1997, 18, 1175–1189.

[4] Jones, G.; Willett, P.; Glen, RC.; Leach, AR.; Taylor, R. **Development and Validation of a Genetic Algorithm for Flexible Docking.** Mol. Biol., 1997, 267, 727-748

[5] Oda, A.; Tsuchida, K.; Takakura, T.; Yamaotsu, N.; Hirono, S. **Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes.** J Chem Inf Model 2006, 46:380-391

[6] Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, P. **Consensus scoring: a method for obtaining improved hit ratesfrom docking databases of three-dimensional structures into proteins.** J. Med. Chem. 1999, 42, 5100–5109.

[7] Okamoto, M.; Takayama, K.; Shimizu, T.; Muroya, A.; Furuya, T. **Structure-activity relationship of novel DAPK inhibitors identified by structure-based virtual screening. Bioorg.** Med. Chem. 2010, 18, 2728–2734.

[8] Trott, O. and Olson, A. J., **AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading.** J. Comput. Chem. 31, 455–461.

[9] http://www.cgl.ucsf.edu/chimera

[10] Discovery Studio, 2.1; Accelrys Inc.: San Diego, CA 92121, U.S.A., 2008.

[11] https://www.ebi.ac.uk/chembl/

[12] http://openbabel.sourceforge.net/wiki/Main_Page