

## Universal Optimizations of Scoring Functions for Virtual Screening

Kenji Onodera\*, Shunsuke Kamijo

*Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro, Tokyo 153-8505, Japan.*

*\*E-mail: onodera.kenji@gmail.com*

(Received April 19, 2010; accepted June 23, 2010; published online July 10, 2010)

### Abstract

Structure-based virtual screening is gaining popularity in drug discovery. A number of molecular docking programs and scoring functions have been developed in the community, but they had not fulfilled the demands for the improved accuracy, yet. In order to improve the accuracy, the consensus scoring method has been developed. It combines docking scores from various scoring functions without considering characteristics of the docking scores. In this study, we adopted the concepts of the consensus scoring, and improved the docking score from each docking programs, DOCK, FRED or GOLD, for virtual screening. Instead using simple sum of score components in those docking scores, weight factors of the score components were introduced and adjusted for better predictions of active ligands during virtual screening. Several optimization processes were tested to find the best optimization methods of the docking scores using a wide variety of 113 target proteins with over 2000 diverse decoys. Finally, the optimizations improved the chance to discover the active ligands by up to 52.4% (e.g. from 36.8% to 56.1% using GOLD) for the test set. Additionally, the combination of the optimized scores using GOLD and FRED improved success rate in the test set by 77.2%, and approximately 70% of ligands for target proteins were predictable in the test set with 20 times enrichment.

**Key Words:** Docking, virtual screening, scoring function, optimization, consensus scoring

**Area of Interest:** Information and Computing Infrastructure for Drug Design and Toxicology

## 1. Introduction

Virtual screening by molecular docking is an essential process for drug discovery [1]. More new leads are found based on three-dimensional structures of the target proteins [2]. However, current virtual screening methods are not satisfactory, and there are persistent demands for improvement in the accuracy of virtual screening [2][3][4]. In structure-based virtual screening, the three dimensional structure of a target protein is pre-required, and the screening is carried out in three steps; docking, scoring, and ranking of ligands for the target protein. Thus, both accurate docking prediction and scoring of ligand-protein complex are keys for better screening.

For the docking step, molecular docking programs are used to predict the structures of ligands and target protein complexes. By the year 2007, over 60 molecular docking programs [4] have been disclosed including DOCK [5], GOLD [6], FRED [7], AutoDock [8], FlexX [9], Glide [10], and ICM [11]. New developments for molecular docking are published every year. Since a number of molecular docking programs are available, and it is difficult to select the one suitable for the projects. Thus, their evaluation is helpful and many evaluation tests have been reported [4][12]. Those evaluation studies show that many of the molecular docking programs perform well. For example, a molecular docking program called GOLD could predict 70% to 80% of ligand-protein complexes successfully confirmed either by visual inspection, or within 2 Å in root-mean-square distance (RMSD) of ligand atom positions between the crystal and the predicted positions [6][13].

For accurate scoring, scoring functions have also been developed vigorously. As of 2007, more than 30 scoring functions have been disclosed [4]. Even though the complex structure is predicted and scored, such scoring can be useless for virtual screening without differentiating true actives from non-binders. There are no correct complex structures for non-binders, since they do not bind to the targets. Thus, it is impossible to predict the correct positions for the non-binders by the molecular docking programs. However, these programs generate complex structures and docking scores for the non-binders that are thus irrelevant. When the scoring function is optimized to calculate the binding affinities of true actives, docking scores for non-binders cannot be calculated correctly from the complex structures of the non-binders and their target proteins.

The aims also differ between the molecular docking programs and virtual screening. Molecular docking programs are aimed at selecting one of the best conformations from a number of generated conformations of 'a ligand'. On the other hand, virtual screening requires selecting the best ligands among 'a number of ligands'. There is an essential difference in the goals for the molecular docking programs and virtual screening, and thus, optimization is required for the latter.

The improvement for virtual screening performance can be achieved by two approaches, consensus scoring method [14][15] and optimization with decoys [16]. The consensus scoring method combines ranks or scores from two or more docking scores, and it improves performances of virtual screening by compensating for the deficiencies of each scoring function. The consensus score can be improved even more if each individual docking score performs better, and thus, the optimization of individual docking scores can be one of the fundamental solutions for the improvement of virtual screening. In this study, we selected three molecular docking programs: one popular program (FRED) free for academic users and two programs (GOLD and DOCK) marked good screening performances in our previous evaluation study [12]. Then, we examined the optimization of the docking scores from the molecular docking programs using known active ligands and decoys for virtual screening. The optimization and the evaluation were performed by 113 diverse target proteins to achieve general improvement of virtual screenings. Although optimization for the specific targets leads to better performances, we believe that the universal

optimizations for the general targets are useful especially for the orphan receptors or targets with a few known active ligands.

Evaluations of docking programs are often done using several active ligands for one or a few targets. In this study, we target on the general improvement of virtual screening method that not restricting ourselves to specific target protein. For the general improvement and its evaluation, a number of ligand-target protein complexes are required to avoid bias toward specific target proteins. Moreover, a diversity of known active ligands is also an important factor for the optimization, because the optimization for specific active ligands also leads bias toward specific ligand types. The diversity of ligands in test set also affects evaluation results. Thus, optimization and evaluation processes were performed using a single active ligand per target protein to secure the diversity of both active ligands and target proteins in the data set.

## 2. Materials and Methods

In this study, co-crystallized ligands from 113 complexes were used as active ligands in the screenings. Binding sites for the docking calculations were defined by surrounding areas of the co-crystallized ligands as described later.

### 2.1 Training and Test Sets

A total of 113 ligand-protein complexes were obtained from the Protein Data Bank (PDB) as training and test sets. The 113 complexes were sorted alphabetically by PDBID, and simply divided into two parts: first 56 complexes for the training and next 57 complexes for the test sets (Table 1). The training set was used only for finding the best scoring parameters for virtual screening. The test set was not used for training, but only for evaluations of the obtained scoring parameters. Otherwise, both the training and the testing sets were treated equally from ligand and target file preparations to the docking calculations.

**Table 1.** List of PDBIDs for training and test sets

Training Set (56 complexes)						
1aaq	1abe	1acj	1ack	1acm	1aco	1aec
1aha	1apt	1ase	1atl	1azm	1baf	1blh
1bma	1byb	1cbs	1cbx	1cil	1com	1coy
1cps	1dbb	1dbj	1did	1die	1dwd	1eap
1epb	1eta	1etr	1fen	1fkg	1fki	1frp
1ghb	1glp	1glq	1hdc	1hdy	1hef	1hfc
1hri	1hsl	1icn	1ida	1igj	1imb	1ive
1lah	1lcp	1lic	1lmo	1lna	1lpm	1lst
Test Set (57 complexes)						
1mcr	1mdr	1mmq	1mrg	1mrk	1mup	1nco
1nis	1pbd	1pha	1phd	1rds	1rne	1rob
1slt	1srj	1stp	1tdb	1tka	1tng	1tni
1tnl	1tpp	1tyl	1ulb	1wap	1xie	2ada
2ak3	2cgr	2cht	2cmd	2ctc	2dbl	2gbp
2lgs	2mcp	2mth	2phh	2pk4	2plv	2r07
2sim	2yhx	3aah	3cla	3cpa	3gch	3hvt
3ptb	3tpi	4dfr	4phv	5p2p	6abp	6rnt
8gch						

The sets of ligand-target protein complexes were selected for the evaluation of GOLD by their developers [6]. They contain a wide variety of proteins including oxidoreductases, transferases, hydrolases, lyases, isomerases, ligases, virus coat proteins, and antibodies. According to the authors of the GOLD evaluation, the complexes were selected on the basis of drug-likeness of ligands and pharmacological interests. Those complexes were popular and used in many evaluations of molecular docking programs [17]. We believe that the set of the complexes is also suitable for our purpose, because it is beneficial to optimize virtual screening system for the pharmacological usage.

## 2.2 Preparation of Protein and Ligand Structure Files

The PDB files of ligand-target protein complexes were simply separated into ligand and target protein files. Both ligands and protein files were processed to add hydrogen atoms by SYBYL [18]. The ligands were optimized to have energy minimized structures by SYBYL. Thus, coordinates and conformations of input ligands differ from those in the co-crystallized ligands.

## 2.3 Screening Library

The number of compound structures in the screening library is 2103, which consists of 113 ligands from the ligand-target protein complexes from PDB described above, and 1990 compounds from NCI diversity set [19]. The NCI diversity set was prepared by the National Cancer Institute (NCI) with acceptable conformations for the compounds. All compounds in the NCI diversity set were simply treated as decoys in this study. Only one out of 2103 compounds in the screening library was considered as an active ligand (referred as a correct ligand) for each target protein, and all other compounds were treated as decoys.

Unexpected binders may exist in the screening library. However, such ligands should not be frequent in the screening library. The correct ligands were counted as success if it ranked within 100<sup>th</sup> among the compounds in the screening library in this study. Thus, the unexpected binders are ignorable as long as they occupy only a small part of the screening library. Properties of the compounds in the screening library were similar between the ligands from the complexes and the decoys from the NCI diversity set. The molecular weight ranged from 90 to 1297 (Avg. 312) in the NCI diversity set and from 114 to 776 (Avg. 306) in the ligands from the complexes, and most fulfilled Lipinski's Rule of Five (95.9% in the NCI diversity set and 91.4% in the ligands from the complexes).

## 2.4 Binding Site Definition

Binding sites were defined as areas to cover all heavy atoms of the ligands of the complexes with 5 Å cushion. Required shape of binding site differs in each molecular docking program. Binding site definitions of DOCK and FRED can be rectangular while GOLD can not. Thus, sizes of binding site can be slightly differed.

For DOCK and FRED, spheres were generated using SPHGEN [5], and all spheres within 5 Å from all heavy atoms of a ligand were selected. The binding site was defined as the smallest rectangle which covers all the selected spheres. For GOLD, a minimum-size sphere that covers all the heavy atoms with 5 Å cushion was defined as a binding site.

## 2.5 Molecular Docking Programs and Processes

All 113 target proteins were docked and scored with 2103 compounds in the screening library. Bissantz, et al. [20] reported the default settings performed generally well for DOCK, FlexX, and

GOLD. Thus, the parameters used for this study were basically their default settings as described in the next paragraph. Molecular docking programs return several solutions for ligand-target protein complexes in the calculations, but only the best scored solution was selected as a docking result of each ligand-target protein complex in each molecular docking program. In this study, the following molecular docking programs and their settings were used for docking calculations.

DOCK 4.0.1 is based on an incremental construction and random search algorithm. Both input ligand and protein structures were passed to DOCK as mol2 files generated by SYBYL with partial charge (Gasteiger and Marsili). Since DOCK does not possess standardized default settings, the parameter settings in a demo, supplied with the DOCK suite, were used for the docking calculations. First, GRID [21], which is also supplied with the DOCK suite, was applied to compute interaction energies of the binding sites of the target proteins. Next, docking calculations were processed using DOCK. Modifications made to the demo settings were to set 'rank\_ligands' as NO, to reduce 'anchor\_size' from 10 to 5, 'heavy\_atoms\_minimum' from 5 to 0, 'energy\_cutoff\_distance' from 10 to 5 and 'maximum\_iterations' from 100 to 2, and increase 'maximum\_orientations' from 10 to 1000, 'ligands\_maximum' from 100 to 880000, and 'heavy\_atoms\_maximum' from 35 to infinity. Finally, as results of docking calculations, Energy score (a force field based score) and Contact score (an empirical score) were obtained.

GOLD 4.0.1 is based on a genetic algorithm. GOLD uses mol2 file format for both ligands and proteins, and the mol2 files generated at the previous preparation step with SYBYL were applied for the docking calculations. Calculations were processed using their default settings from a preset of '7-8 times speed-up.' Only minor modifications were made to reduce output file sizes of GOLD; 'clean\_up\_option save\_top\_n\_solutions' was 1, and 'n\_top\_solutions' was 3. Three scoring functions, GoldScore (a force field based score), ChemScore (an empirical score), and ASP (Astex Statistical Potential; a knowledge based score), were used for dockings separately.

FRED 2.2.4 is based on a shape-based docking method. FRED requires a set of input conformers for each ligand in their in-house format. Thus, the screening library was processed with OMEGA [22] to generate a single binary file for all input ligands. FRED was simply run in command line with their default settings. After shape fittings, ligands were optimized by one of three scoring functions; Chemgauss3 (a knowledge based score), ChemScore (an empirical score), or PLP (Piecewise Linear Potential; an empirical score).

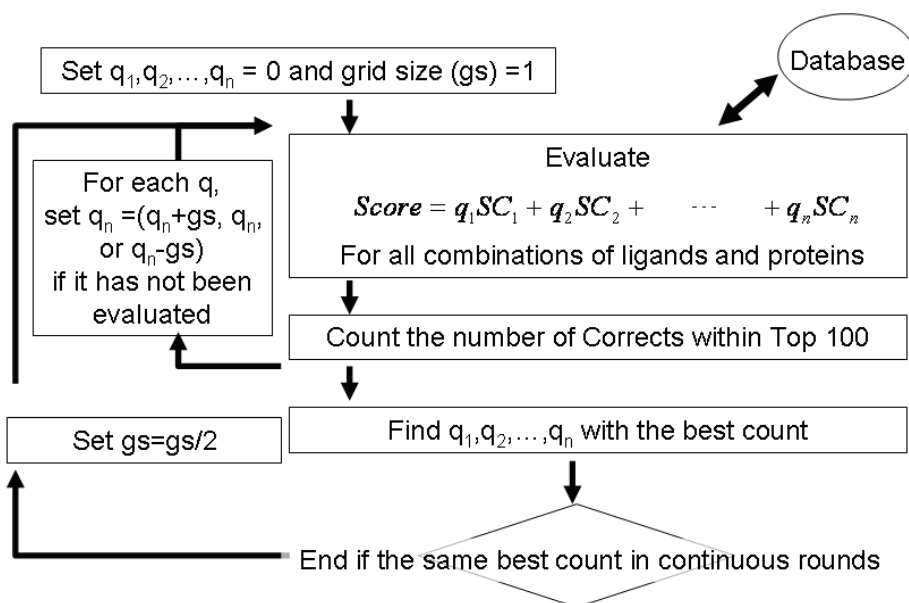
All docking calculations were performed using computers running Ubuntu 4.2.4 (Linux) with Intel® Core™2 CPU 6700 at 2.66GHz and Red Hat 4.0.2 with Intel® Xeon™ 5110 at 1.60GHz.

## 2.6 Searches for the Best Scoring Parameters

In this study, we defined the scoring with weight factors in the following equation (Eq. 1).

$$Score = q_1 SC_1 + q_2 SC_2 + \cdots + q_n SC_n \quad (1)$$

In the equation,  $q_1, q_2, \dots$ , and  $q_n$  stand for weight factors, and  $SC_1, SC_2, \dots$ , and  $SC_n$  stand for selected score components. A flow chart for searching the best scoring parameters is shown in Fig. 1. First, scores were calculated for all combinations of the weight factors,  $q_1, q_2, \dots$ , and  $q_n$ , for each docking result when the weight factors  $q$  were -1, 0 or 1, and the grid size was 1. The obtained scores were converted to ranks of docked ligands. The total numbers of correct ligands within top 100 compounds (Top 100 count) was counted, and the best combination of the weight factors with the largest Top 100 count was selected for the round. The combination of the weight factors becomes a starting point for the following round. Also, from the following rounds, the grid size was set to half of the previous round. The computations of the best combinations were repeated until the local minimum was identified by obtaining the same Top 100 count in two consecutive rounds.



**Figure 1.** Flow chart for the multivariate grid search

## 2.7 Results Database

Docking scores and their score component values were obtained from the molecular docking programs. All of those scores and values were collected and stored in a database using MySQL version 5.0.51a running under Ubuntu 4.2.4 with Intel® Core™2 CPU 6700 at 2.66GHz. Optimization of docking scores and calculations of new docking scores and ranks were performed under MySQL database with Perl v5.8.8.

## 3. Results and Discussions

### 3.1 Docking Calculations and Run Times

A total of 113 target proteins were processed for the docking (Table 1). Three docking programs with eight scoring functions were processed for all combinations of 113 target proteins and 2103 compounds in the screening library. As a result, approximately 1.9 million docking calculations were performed, and all docking results were stored in a MySQL database for analysis.

Total processing times for the docking calculations were 32.6, 16.0, and 6.06 days for GOLD with GoldScore, DOCK with Energy score, and FRED with Chemgauss3 for all combinations of ligands and target proteins, respectively (Table 2).

**Table 2.** Processing times for docking and optimizations

		DOCK	FRED	GOLD
Docking Calculation (days)		16.0	6.06	32.6
Optimization	3 grid points (sec)	551	1154	1220
	9 grid points (sec)	32606	284361	193786
3 grids/9 grids		59.2	158.8	246.4

### 3.2 Ranking Performances of the Original Scoring Functions

Docking scores from various molecular docking programs cannot be compared directly, since base units and scales of the docking scores differ among scoring functions [12]. Even with the same scoring function, score range for one target protein differs from those from other target proteins. Direct comparison of docking scores is difficult. Thus, docking scores from predicted ligand-protein complexes were converted to ranks in the screening library by sorting the docking scores. For evaluations in virtual screening, top 5% is a popular cutoff for measurements of successes in virtual screening [23]. Thus, performances of scoring were determined by the total numbers of correct ligands within top 100 (Top 4.8%) out of 2103 compounds (referred as Top 100 count) in this study.

The Top 100 counts by the eight scoring functions of the three molecular docking programs are shown in Table 3. Three scoring functions were tested for GOLD, the best was GoldScore. Two scoring functions were tested for DOCK, and the best was Energy Score. Three scoring functions were tested for FRED, and the best was Chemgauss3. Top 100 counts for the other scoring functions were much less than the best function for each molecular docking program. It seems that those other functions have difficulties to find the correct ligands in virtual screening, and probably we cannot expect better optimization results from those.

**Table 3.** The number of ligands correctly predicted by the original docking scores

		Training Set	Test Set	Total
DOCK	Energy Score	19	20	39
	Contact Score	13	11	24
FRED	Chemgauss3	18	22	40
	ChemScore	10	10	20
	PLP	10	16	26
GOLD	GoldScore	18	21	39
	ChemScore	11	10	21
	ASP	11	18	29

There were only small differences in Top 100 counts in between the training and the test sets for most scoring functions, and the major differences were observed only in GOLD with ASP and FRED with PLP. This probably indicates that distributions of target proteins in the training and test sets were not similar. However, they cannot be the same as long as the limited numbers of target proteins were used, and the differences were small for most of the scoring functions. Thus, we

decided to use the training and the test sets for optimization, and simply selected three pairs of the molecular docking programs and the scoring functions; GOLD with GoldScore, DOCK with Energy Score, and FRED with Chemgauss3 for the further optimization processes.

### 3.3 Searches for the Best Scoring Parameters

Optimizations of the scoring parameters were examined using a method based on the multivariate grid search method, a classical method used in finding minimum potential energy for a molecular surface [24]. Docking scores for binding affinity estimations usually consists of several score component values, and a simple sum of score component values is the docking scores. Instead of simple summation for the score components, we introduced weight factors for the score components to improve performance of docking scores (Eq. 1).

During optimization of docking scores, several combinations of weight factors often showed the same numbers of Top 100 counts. Thus, average ranks were also considered besides the Top 100 counts for differentiations. Optimizations were tested in two ways (Table 4). One was to have more Top 100 count and less average rank of correct ligands. Another was to have less average ranks without considering the Top 100 count to avoid over-fitting to the training set.

**Table 4.** The number of ligands correctly predicted by the optimized scores

The percentages in the table indicate that success rates out of all 57 correct ligands in the test set. All but '9 Grid Points' were optimized using three grid points. In 'All for Training', all 113 complexes were used for training, and Top 100 counts were listed here as a reference.

		Training Set	Test Set	Total
DOCK	Unoptimized	19	20 (35.1%)	39
	Average only	16	25 (43.9%)	41
	Top 100 count	23	25 (43.9%)	48
	9 Grid points	23	26 (45.6%)	49
	(All for Training)	(51)	—	(51)
FRED	Unoptimized	18	22 (38.6%)	40
	Average only	27	29 (50.9%)	57
	Top 100 count	28	32 (56.1%)	61
	9 Grid points	29	31 (54.4%)	61
	(All for Training)	(62)	—	(62)
GOLD	Unoptimized	18	21 (36.8%)	39
	Average only	27	22 (38.6%)	50
	Top 100 count	31	32 (56.1%)	64
	9 Grid points	31	30 (52.6%)	62
	(All for Training)	(63)	—	(63)

Top 100 counts for the original docking scores (without optimization) were almost the same among the molecular docking programs, but Top 100 counts after optimization varied. For GOLD, optimization improved performances by 52.4% (from 36.8% in the original score to 56.1% in the optimized score). FRED was also improved by 45.5%. The weakest improvement was DOCK, but it still improved for 25.0%. This is a reasonable result, since previous evaluation study showed



performance of DOCK was good for virtual screening although precisions of docked complexes were not very satisfying [12]. Thus, it probably indicates that Energy Score for DOCK was somewhat close to the optimum without optimization.

The optimization with Top 100 count improved better than the optimization solely using average ranks. It is quite reasonable, because the average improved more for the optimization from rank 2000<sup>th</sup> to 1000<sup>th</sup> than from rank 200<sup>th</sup> to 100<sup>th</sup>, and thus, the optimization with Top 100 count performed better.

The optimized parameters of DOCK/Energy Score, FRED/Chemgauss3 and GOLD/GoldScore are listed in Eq. 2 to Eq. 4.

$$\begin{aligned} \text{Optimized DOCK} = & -1.125 \times [\text{Internal Electro}] - 0.5 \times [\text{Internal vdw}] \\ & - 0.375 \times [\text{Intra Electro}] - 0.5 \times [\text{Intra vdw}] \quad (2) \end{aligned}$$

$$\begin{aligned} \text{Optimized FRED} = & 0.25 \times [\text{Desolvation}] - [\text{Steric}] \\ & - 1.75 \times [\text{Metal}] - 1.25 \times [\text{Donor}] - [\text{Acc}] \quad (3) \end{aligned}$$

$$\begin{aligned} \text{Optimized GOLD} = & [\text{External Hbond}] + 0.5 \times [\text{External vdw}] - 1.5 \times [\text{Internal Hbond}] \\ & + [\text{Internal Torsion}] + 0.5 \times [\text{Internal vdw}] \quad (4) \end{aligned}$$

In the optimized scores, polar atom-oriented score components were heavy weighted. ‘Electrostatic’ in DOCK is calculated by coulombic electrostatic energy. ‘Hbond’ in GOLD stands for hydrogen bond energy. ‘Acc’ and ‘Donor’ in FRED are interaction energies of Acceptor and Donor atoms in ligands, and they are polar atom-oriented score components. Such polar atom-oriented score components were weighted up to twice more than van der Waals related components (‘vwd’ in GOLD and DOCK, and ‘Steric’ in FRED). Probably, van der Waals energy was essential to decide ligand positions to avoid overlapping or too close contacts between a ligand and a target protein, because van der Waals depends exponentially on the distance. On the other hand, it was less important for scoring after docked probably due to weak attractions by van der Waals at long range. Thus, optimizations without van der Waals components were also examined. We found that success rates lowered by 9.4% to 21.3% without van der Waals components (data not shown). This indicates that importance of van der Waals is less for scoring, but it is not negligible. Van der Waals energy defines attractive and dispersive interactions, and polar atom-oriented energy accounts for interaction of polar atoms only. Importance of van der Waals energy in the optimized scores probably indicates importance of non-polar interactions during the scoring. Therefore, van der Waals energy is still important for the optimized score.

Hydrogen bond or polar atom-oriented energies ‘Within’ and ‘Between’ molecules were also weighted differently in the optimized scores. Energy within a ligand or a protein is indicated by ‘Internal Electro’ in DOCK and ‘External Hbond’ in GOLD, whereas energy between a ligand and a protein is indicated by ‘Intra Electro’ in DOCK and ‘Internal Hbond’ in GOLD. In the optimized DOCK, energy between a ligand and a protein was weighted three times more than that within a ligand or a protein (Eq. 2). In the optimized GOLD, larger and positive values in scores indicate better scores, and the weight factor for energy within a ligand or a protein was a negative value, whereas all the others were positive (Eq. 4). Thus, hydrogen bonds ‘within’ a ligand or a protein were less important for virtual screening. Probably, such hydrogen bonds interfere with complex structure more than binding affinity, and thus, it was less important in the optimized scores.

During optimization of the weight factors, three grid points were used for each parameter in

each round of the search. Since using more grid points increase the parameters to be tested, it may increase the chances for finding better local minimum than using only three grid points. Thus, the weight factor  $q$  in Eq. 1 were optimized with nine grid points of -4, -3, -2, -1, 0, +1, +2, +3, and +4 in each round. Although coverage of the weight factors was increased with the nine grid points, Top 100 counts in the training set did not improved (Table 4). At the same time, processing times were dramatically increased from 20 minutes with three grid points to two days with the nine grid points in GOLD (Table 2). Thus, the three grid points are sufficient for optimization and computationally inexpensive.

We also tested an optimization using all 113 target proteins (Table 4), because more target proteins in the training set may lead to the better optimization results. We note that it cannot be used as an actual optimization result when all 113 target proteins are used for training, and it is for reference purpose only. As a result, the total numbers of Top 100 counts were actually similar between 56 and 113 target proteins used for the training. It supports that our dividing ratio of the training to the test sets for the 113 target proteins was satisfactory in this study.

### 3.4 Combinations of the Optimized Scores

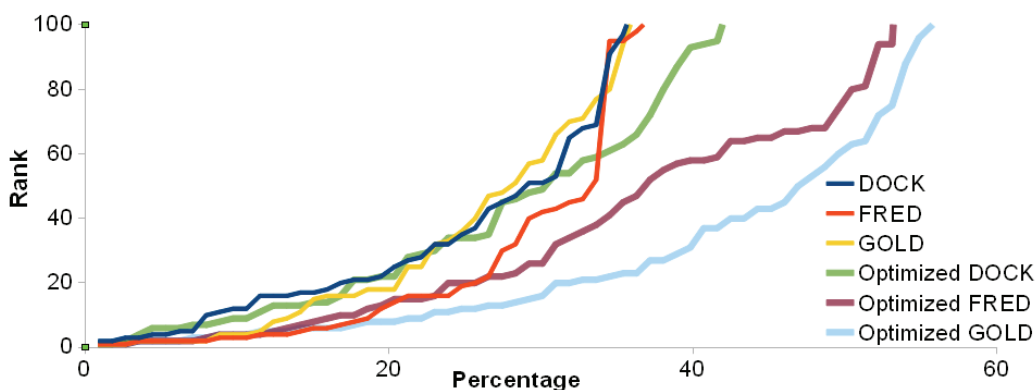
By optimization, the docking score improved performance of virtual screening by up to 52.4%. Performance of virtual screening should improve even more when several docking scores are combined, since combinations can cover deficiencies of each molecular docking program. Consensus score is a popular method to improve performance of virtual screening. However, Xing, et al. reported decrease performance of the consensus scoring, when single docking engine was used for dockings and rescored by several scoring functions [25]. Errors in predictions of complex structures will lead to wrong scoring results, and worsen consensus score results. Thus, in this study, each docking scores were calculated by each docking program individually. Then, consensus scores were calculated by a method called 'score by score'. In 'score by score', average docking scores from several docking programs are used as consensus scores. Docking scores were not simply added or averaged in this study. Instead, weighted averages were applied to the combinations of three original and three optimized scores. Then, the consensus scores were optimized to improve Top 100 count.

Although the original scores were optimized by 'score by score', it improved only 6.9% from the unoptimized scores by FRED with Chemgauss3 in the test set. This percentage was far smaller than those for the single optimized scores. On the other hand, consensus scores for the optimized docking scores did not improve Top 100 count at all. Actually the best combination of the optimized scores was GOLD alone. It means the combination in 'score by score' should not be used for the optimized scores in this study.

Since 'score by score' consensus scores failed to improve performance of virtual screening, rankings were used for combinations of the optimized scores instead. According to docking enrichment plots (Fig. 2), correct ligands were not distributed equally over rank 1<sup>st</sup> to 100<sup>th</sup>, but they were concentrated more in higher ranks. Thus, certain numbers of top ranked compounds from each optimized score were combined to make a list of total 100 compounds, and examined for how many correct ligands existed in the list. Because FRED and GOLD showed good optimization results, a combination of those two should show good optimization result. When more scoring functions are added to combinations, it may cause over-fitting. Since DOCK show weak improvement, including DOCK for a combination is probably not necessary.

As expected, the combination of the optimized scores of FRED and GOLD showed smaller counts in the training set and larger counts in the test set than that of all three optimized scores probably due to over-fitting (Table 5).

When top 39 and top 61 ligands were obtained from the optimized GOLD and FRED, respectively, the combination showed an improvement by 77.2% from the original scores, and approximately 70% of target proteins in the test set were performed successfully. Both GOLD and DOCK are good at small binding sites while FRED is good at large binding sites [13], and a pair of GOLD and FRED was the best combination to cover deficiencies of each molecular docking program.



**Figure 2.** Docking enrichment plots for the original scores and the optimized scores

**Table 5.** The number of ligands correctly predicted by the combinations of docking scores  
The percentages in the table indicate that success rates out of all 57 correct ligands in the test set.

Opt. GOLD	Opt. FRED	Opt. DOCK	Training Set	Test Set	Total
Top 60	Top 5	Top 35	37	34 (60.0%)	71
Top 39	Top 61	—	35	39 (68.4%)	74
Top 83	—	Top 17	33	33 (57.8%)	66
—	Top 39	Top 61	31	33 (57.8%)	64

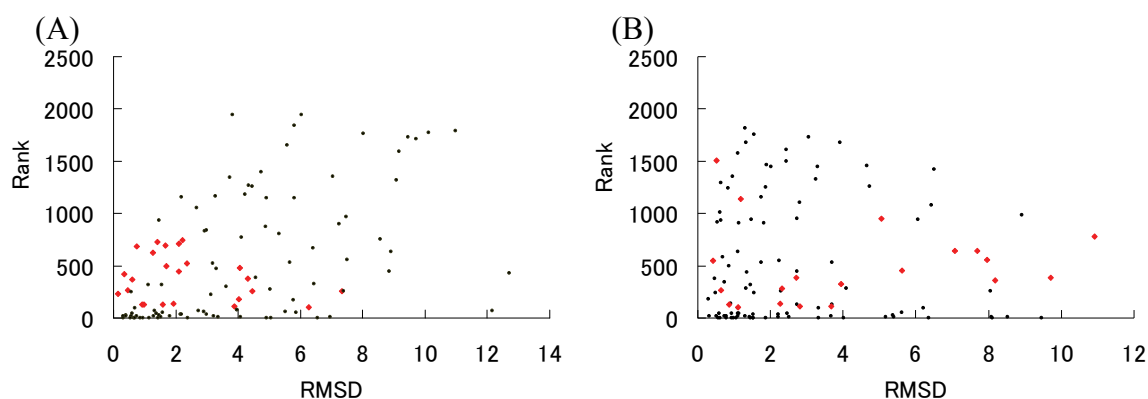
### 3.5 RMSD and the Optimizations

At least one or more complexes were ranked over 1500<sup>th</sup> or worse both before and after the optimization in each molecular docking program. These were mostly due to failures in docking predictions. We listed ten ligand-protein complexes in the training set, which were not ranked within top 100 by any of docking scores even after the optimizations (Table 6). RMSDs were widely distributed for such complexes, and some RMSDs were under 2 Å, which is usually counted as success cases in dockings. At first glance, it seems that there were no strong correlation between docking accuracies and predictions of the correct ligands as Bissantz, et al. [20] described. However, there are no RMSD under 1 Å for such complexes. Thus, screening probably works well if dockings performed as perfect as under 1 Å in RMSD.

**Table 6.** The complexes in the training set failed by all docking programs and docking accuracies (Å) in RMSD

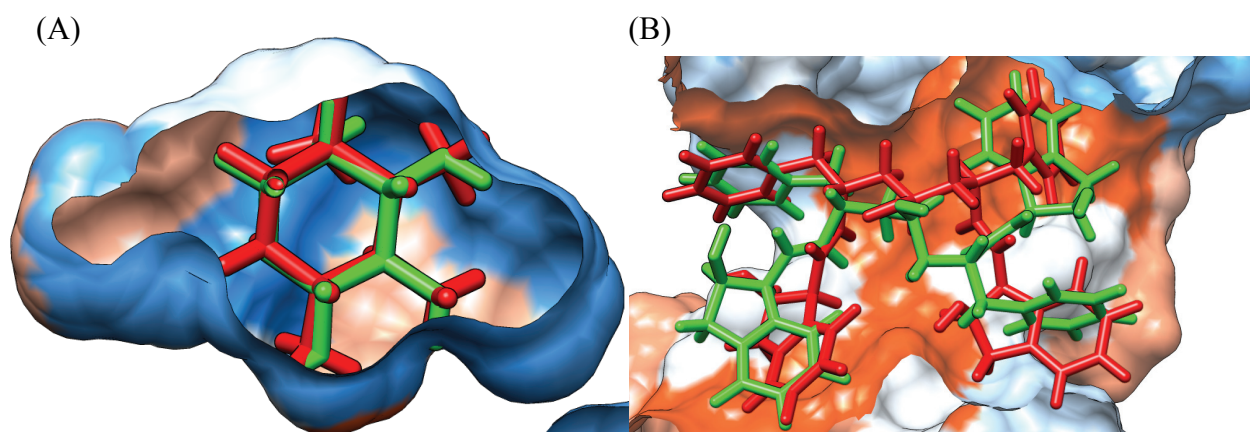
PDBID	DOCK	FRED	GOLD
1acj	1.25	0.62	3.18
1bma	12.25	4.74	5.57
1cbs	6.79	1.55	9.09
1die	3.19	3.92	4.23
1epb	12.45	2.81	1.56
1fen	1.89	1.75	1.13
1fkg	2.36	2.02	5.80
1fki	5.64	2.51	7.40
1hdc	9.47	1.53	9.16
1lpm	7.08	2.21	7.49

The relations between RMSDs and the ranks were analyzed more (Fig. 3). The most of success cases after optimization were complexes with around 2 Å or less in RMSD for GOLD with GoldScore (Fig. 3A). Higher RMSDs have less success cases after optimization. For example, PDBID: 1abe was docked well and 0.15 Å in RMSD, but not ranked well until optimization (Fig. 4A). Binding area for 1abe was small and surrounded by the target protein. In such cases, the docked positions can be predicted easier without accurate docking scorings.

**Figure 3.** The relations between RMSDs and Ranks in GOLD with GOLDScore (A) and in FRED with ChemGauss3 (B)

The red points indicate that the complexes ranked within top 100 after the optimizations.

For FRED, many complexes with low RMSDs also became within top 100 after optimization (Fig. 3B). However, more success cases after optimization in FRED has higher RMSD than those in GOLD. After visual inspections of such complexes, we found that many predicted ligand positions were similar to co-crystallized positions, even though RMSD values were high. For example, PDBID: 4phv had 8.92 Å in RMSD, but overall placements of predicted and co-crystallized ligand position were somewhat similar (Fig. 4B). The orientation differed by rotation of the right half of Fig. 4B by 180 degree, and permutation between the indole and the phenyl rings in the right increased the RMSD value. In such cases, RMSD values underestimate docking accuracies sometimes [26][27], but we can say that there are some correlations between docking and ranking accuracies in our results. The better docking will generate the better ranking in the optimized scores.



**Figure 4.** The ligands position for predicted (Red) and co-crystallized (Green) for PDBID: 1abe by GOLD in the right (A) and 4phv by FRED in the left (B)

#### 4. Conclusion

Optimization of docking scores could improve performance of virtual screening. In this paper, we proposed optimization methods of the docking scores for the universal improvement of virtual screening. Optimization used co-crystallized ligands to reach high varieties of protein targets. Screenings of such ligands may be easier than other ligands, but performances of the original docking scores for co-crystallized ligands were currently unacceptably low. Thus, we believe that the improvements for universal targets even using co-crystallized ligands are the first priority to refurbish low screening efficiencies of molecular docking programs. The performances were improved by up to 52.4% for single docking scores alone, and 77.2% in the combination. Now, 56.1% and 68.4% of ligand screenings were performed successfully by the single docking scores alone and the combination, respectively.

Efficiencies of DOCK were similar to the other two molecular docking programs before the optimization. However, optimization of docking score in DOCK showed the weakest improvement, while the molecular docking programs with better complex predictions, GOLD and FRED, were improved dramatically. DOCK showed the lowest docking accuracies based on RMSD. Molecular docking programs with better docking accuracies could have better virtual screening performance after the optimization. Thus, for further improvement of virtual screening, ordinary but necessary suggestion is improvement of docking accuracy.

Approximately 70% target proteins were correctly enriched over 20 times (Top 100 out of 2103 compounds) in the combination of FRED and GOLD. The success rate for the combination was at least 10% more than that for the single docking scores. Of course, it is burden and time consuming that using two molecular docking programs for each virtual screening. However, it is less costly than using more chemicals and testing in the downstream of drug developments. FRED was more than five times faster than GOLD. Thus, additional processing time in the combination of FRED and GOLD is just 19% more than when GOLD is used alone.

This is a result of the universal optimizations of docking scores that not restricting ourselves to specific target protein. Thus, we need to note that it is not maximum screening performances of the molecular docking programs for virtual screening. Usually, several active ligands and some other information about its binding site are known for a target. For instance, physical properties of

molecules, such as size, may differentiate possible active ligands from inactive compounds. Knowledge of pharmacophores of target proteins also helps inactive compounds to be screened out in a large screening library. With such information, adjustable parameters in molecular docking programs can be modified for better dockings. Visual inspections of compounds and complex structures are probably unavoidable tasks in virtual screening, and they also help for further enrichments of hits in hit lists. Universal optimization of docking scores certainly reduces chances to enrich wrong compounds in the hit list, and then, it increases chances to find novel compounds for drug discovery. Our methods for optimization can enhance performance of virtual screening using molecular docking programs currently available in our community.

We thank Dr. D. Fourmy for critical comments on the manuscript. This work is supported by grant-in-aid for Scientific Research (19710173). Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081) [28].

## References

- [1] Jorgensen, W. L. The many roles of computation in drug discovery. *Science*. **2004**, *303* (5665), 1813-1818.
- [2] Bailey, D.; Brown, D. High-throughput chemistry and structure-based design: survival of the smartest. *Drug Discov Today*. **2001**, *6* (2), 57-59.
- [3] Waszkowycz, B. Towards improving compound selection in structure-based virtual screening. *Drug Discov Today*. **2008**, *13* (5-6), 219-226.
- [4] Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br J Pharmacol*. **2008**, *153 Suppl 1*, S7-26.
- [5] Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J Mol Biol*. **1982**, *161* (2), 269-288.
- [6] Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*. **1997**, *267* (3), 727-748.
- [7] McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian Docking Functions. *Biopolymers*. **2003**, *68*, 76-90.
- [8] Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. **2009**.
- [9] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*. **1996**, *261* (3), 470-489.
- [10] Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*. **2004**, *47* (7), 1739-1749.
- [11] Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*. **1994**, *15* (5), 488-506.
- [12] Onodera, K.; Satou, K.; Hirota, H. Evaluations of molecular docking programs for virtual screening. *J Chem Inf Model*. **2007**, *47* (4), 1609-1618.
- [13] Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins*. **2004**, *57* (2), 225-242.

- [14] Teramoto, R.; Fukunishi, H. Consensus scoring with feature selection for structure-based virtual screening. *J Chem Inf Model.* **2008**, *48* (2), 288-295.
- [15] Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *J Chem Inf Model.* **2005**, *45* (4), 1134-1146.
- [16] Reid, D.; Sadjad, B. S.; Zsoldos, Z.; Simon, A. LASSO-ligand activity by surface similarity order: a new tool for ligand based virtual screening. *J Comput Aided Mol Des.* **2008**, *22* (6-7), 479-487.
- [17] Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins.* **2002**, *49* (4), 457-471.
- [18] SYBYL, Version 6.92; Tripos, Inc.: St. Louis, MO, 2004.
- [19] DTP - Diversity Set Information. [http://dtp.nci.nih.gov/branches/dscb/diversity\\_explanation.html](http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html) (Accessed on Aug. 19, 2009).
- [20] Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem.* **2000**, *43* (25), 4759-4767.
- [21] Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des.* **2001**, *15* (5), 411-428.
- [22] OMEGA, Version 2.3.1; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2008.
- [23] Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins.* **2005**, *60* (3), 325-332.
- [24] Hinchliffe, A. *Molecular Modelling for Beginners* Wiley: Hoboken, NJ, 2003.
- [25] Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and application of multiple scoring functions for a virtual screening experiment. *J Comput Aided Mol Des.* **2004**, *18* (5), 333-344.
- [26] Baber, J. C.; Thompson, D. C.; Cross, J. B.; Humblet, C. GARD: a Generally Applicable Replacement for RMSD. *J Chem Inf Model.* **2009**, *49* (8), 1889-1900.
- [27] Yusuf, D.; Davis, A. M.; Kleywegt, G. J.; Schmitt, S. An alternative method for the evaluation of docking performance: RSR vs RMSD. *J Chem Inf Model.* **2008**, *48* (7), 1411-1422.
- [28] Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem.* **2004**, *25* (13), 1605-1612.