# Consensus Scoring Criteria for Improving Enrichment in Virtual Screening

Jinn-Moon Yang,*[†,‡] Yen-Fu Chen,[†,‡] Tsai-Wei Shen,[†,‡] Bruce S. Kristal,[§,||] and D. Frank Hsu*[⊥,#]

Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, 30050, Taiwan, Institute of Bioinformatics, National Chiao Tung University, Hsinchu, 30050, Taiwan, Dementia Research Service, Burke Medical Research Institute, 785 Mamaroneck Ave., White Plains, New York 10605, Department of Neuroscience, Weill Medical College of Cornell University, 1300 York Ave., New York, New York 10021, Department of Computer and Information Science, Fordham University, 113 West 60th Street, LL 813, New York, New York 10023, and DIMACS Center, Rutgers University, 96 Frelinghuysen Road, Piscataway, New Jersey 08854-8018

Received January 28, 2005

**Motivation:** Virtual screening of molecular compound libraries is a potentially powerful and inexpensive method for the discovery of novel lead compounds for drug development. The major weakness of virtual screening—the inability to consistently identify true positives (leads)—is likely due to our incomplete understanding of the chemistry involved in ligand binding and the subsequently imprecise scoring algorithms. It has been demonstrated that combining multiple scoring functions (consensus scoring) improves the enrichment of true positives. Previous efforts at consensus scoring have largely focused on empirical results, but they have yet to provide a theoretical analysis that gives insight into real features of combinations and data fusion for virtual screening. **Results:** We demonstrate that combining multiple scoring functions improves the enrichment of true positives only if (a) each of the individual scoring functions has relatively high performance and (b) the individual scoring functions are distinctive. Notably, these two prediction variables are previously established criteria for the performance of data fusion approaches using either rank or score combinations. This work, thus, establishes a potential theoretical basis for the probable success of data fusion approaches to improve yields in in silico screening experiments. Furthermore, it is similarly established that the second criterion (b) can, in at least some cases, be functionally defined as the area between the rank versus score plots generated by the two (or more) algorithms. Because rank-score plots are independent of the performance of the individual scoring function, this establishes a second theoretically defined approach to determining the likely success of combining data from different predictive algorithms. This approach is, thus, useful in practical settings in the virtual screening process when the performance of at least two individual scoring functions (such as in criterion a) can be estimated as having a high likelihood of having high performance, even if no training sets are available. We provide initial validation of this theoretical approach using data from five scoring systems with two evolutionary docking algorithms on four targets, thymidine kinase, human dihydrofolate reductase, and estrogen receptors of antagonists and agonists. Our procedure is computationally efficient, able to adapt to different situations, and scalable to a large number of compounds as well as to a greater number of combinations. Results of the experiment show a fairly significant improvement (vs single algorithms) in several measures of scoring quality, specifically “goodness-of-hit” scores, false positive rates, and “enrichment”. This approach (available online at http://gemdock.life.nctu.edu.tw/dock/download.php) has practical utility for cases where the basic tools are known or believed to be generally applicable, but where specific training sets are absent.

## 1. INTRODUCTION

The average cost and time of bringing a new drug to market has been estimated to be \$802 million in year 2000 U.S. dollars and 12 years [1], respectively. Discovery of novel lead compounds through virtual screening (VS) of chemical databases against protein structures is an emerging and promising step in computer-aided drug design.[2−5] Given the structure of a target protein active site and a potential small ligand database, VS predicts the binding mode and the binding affinity for each ligand and ranks a series of candidate ligands. The VS computational method involves two basic critical elements: one or more efficient molecular docking algorithms and a means for interpreting the data derived from this algorithm, termed a scoring method. A molecular docking method for VS should be able to screen a large number of potential ligands with reasonable accuracy and speed, and scoring methods for VS should effectively discriminate between correct binding states and nonnative docked conformations during the molecular docking phase and should distinguish a small number of active compounds from hundreds of thousands of nonactive compounds during the postdocking analysis. The scoring functions that calculate

[*] Corresponding authors. E-mail: moon@cc.nctu.edu.tw (J.-M.Y). E-mail: Hsu@cis.fordham.edu (D.F.H.).

[†] Department of Biological Science and Technology, National Chiao Tung University.

[‡] Institute of Bioinformatics, National Chiao Tung University.

[§] Burke Medical Research Institute.

[||] Weill Medical College of Cornell University.

[⊥] Fordham University.

[#] Rutgers University.

10.1021/ci050034w CCC: \$30.25 © 2005 American Chemical Society
Published on Web 05/28/2005

the binding free energy mainly include knowledge-based,[6] physics-based,[7] and empirical [8] scoring functions.

In practice, the performance of a scoring function is limited by our incomplete understanding of the complex issues involved in chemical interactions. Not surprisingly, the performance of these scoring functions is, therefore, often inconsistent across different systems in a database search.[9,10] The inaccuracy of the scoring methods, that is, inadequately predicting the true binding affinity of a ligand for a receptor, is probably the major weakness for VS. The likelihood that different scoring methods might have different strengths and weaknesses raises the possibility that the simultaneous use of more than one method might increase the overall signal-noise ratio of the calculated affinity. Consistent with this concept, it has been reported that fusion among different scoring methods in VS can perform better than the average of the individual scoring functions.[11] More recently, the same phenomena has been reported in information retrieval (IR) and molecular similarity measurements.[12-17] Charifson et al.[11] presented a computational study in which they used an intersection-based consensus approach to combine scoring functions. They showed an enrichment in the ability to discriminate between active and inactive enzyme inhibitors for three different enzymes (p38 MAP kinase, inosine monophosphate dehydrogenase, and HIV protease) using two different docking methods (DOCK [18] and GAMBLER) and 13 scoring functions. Bissantz et al.[10] used three docking programs (DOCK, FlexX,[19] and GOLD [20]) in combination with seven scoring functions to assess the accuracy of VS methods against two protein targets [thymidine kinase (TK) and estrogen receptor (ER)]. Stahl and Rarey[9] presented a study of the performance of four scoring functions for library docking using the program FlexX on seven target proteins. The study in Verdonk et al.[21] addressed a number of issues on the use of VS protein-ligand docking on the basis of VS experiments against four targets (neuraminidase, ptp1b, cdk2, and ER) using the program GOLD and three scoring functions. Wang and Wang[22] presented an idealized computer experiment to explore how consensus scoring (CS) works based on the assumption that the error of a scoring function is a random number in a normal distribution. They also studied the relationship between the hit rates, the number of scoring functions, and the use of several different approaches to ranking the data (the rank-by-score, rank-by-rank, and rank-by-vote strategies) for consensus scorings.

These reported results are significant and potentially robust in that the performance results of these CS methods seem to be independent of the target receptor and the docking algorithm. The reported results seem to depend on the method of combination (by rank, by score, by intersection, by min, by max, and by voting) and the number and nature of individual scoring functions involved in the combination. Although researchers have come to realize the advantage and benefit of method combination and consensus scorings, the major issues of how and when these individual scoring functions should be combined remain a challenging problem not only for researchers but also, perhaps more importantly, for practitioners in virtual screening.

Here, we address these issues for improving the enrichment in VS using the concept of data fusion and exploring diversity on scoring characteristics between individual scoring functions. In particular, we use a function that relates the absolute score from the docking algorithm to the ranking of this score in the population of tested (potential) compounds (hereafter, the "rank/score function") as a scoring characteristic and the differences in the rank/score graph between individual scoring functions as a diversity measurement. Data fusion approaches have been proposed, developed, and implemented in information retrieval,[12,13,16,17] molecular similarity,[15] and microarray gene expression analysis,[23] where the following two general criteria have been identified for potential improvement: (a) each of the individual scoring functions has to have a relatively good performance, and (b) the scoring characteristics of each of the scoring functions have to be different. To enable us to approach CS as a problem of data fusion, we initially defined two parameters: (1) the performance ratio $P_l$ /$P_h$ ($P_l$ and $P_h$ are the low and high performance of a pairing combination, respectively), which is used as the relative performance measurement, and (2) the rank/score graph, which is used as a surrogate to mathematically describe the characteristic results of a given scoring algorithm on a given target. We will then investigate these parameters and the overall resultant quality of a VS experiment when data are combined using either rank-based or score-based consensus scoring (RCS and SCS) approaches. Our novel consensus scoring system in VS was developed and evaluated by combining five scoring functions on the four target proteins TK, human dihydrofolate reductase (DHFR), ER-antagonist receptor (ER), and ER-agonist receptor (ERA) using two docking algorithms GEMDOCK [24] and GOLD. [20]

## 2. MATERIALS AND METHODS

**2.1. Preparations of Ligand Databases and Target Proteins.** We used the ligand data set from the comparative studies of Bissantz et al.[10] to evaluate the screening accuracy of different CS on TK, DHFR, ER, and ERA. The receptors for these screens cover different receptor types and, therefore, provide a reasonable test of CS. For each target protein, the ligand database included 10 known active compounds and 990 random compounds. According to our experiments, these are some pharmaceutically relevant compounds for our test receptors in this random ligand set. In total, the database used for screening ligands against the target proteins contained 1000 molecules; that is, 990 random compounds were the same for each of these screens. For screening TK and ER, the sets of 10 known active compounds were identical to those reported earlier.[10] For screening ERA, a set of 10 known agonists was identical to that reported earlier,[25] and the 10 active compounds of DHFR were selected from the Protein Data Bank (PDB).[26]

Four complexes of the target proteins were selected for virtual screening from the PDB: TK complex (PDB code: 1kim), DHFR (PDB code: 1hfr), ER-antagonist complex (PDB code: 3ert), and ER-agonist complex (PDB code: 1gwr). These complexes were reasonable choices because their ligand-binding cavities are wide enough to accommodate a broad variety of ligands and, therefore, did not require binding site modifications. The active compound set of each target protein, target proteins, and 990 random compounds are available on the Web at http://gemdock.life.nctu.edu.tw/dock/download.php.

**2.2. Docking Methods and Scoring Functions. GEMDOCK Docking.** Our previous work[24] showed that the docking accuracy of GEMDOCK was better than that of comparative approaches, such as GOLD and FlexX, on a diverse data set of 100 protein−ligand complexes proposed by Jones et al.[20] The screening accuracies of GEMDOCK were also better than GOLD, FlexX, and DOCK on screening the ligand database from Bissantz et al. for TK[27] and ER.[28] In this study, GEMDOCK parameters in the flexible docking included the initial step sizes ($\sigma = 0.8$ and $\psi = 0.2$), family competition length ($L = 2$), population size ($N = 200$), and recombination probability ($p_c = 0.3$). For each ligand screened, GEMDOCK optimization stopped either when the convergence was below a certain threshold value or when the iterations exceeded the maximal preset value of 60. Therefore, GEMDOCK generated 800 solutions in one generation and terminated after it exhausted 48 000 solutions for each docked ligand.

GEMDOCK could use either a purely empirical (GEMDOCK−Binding) or pharmacophore-based scoring function (GEMDOCK−Pharma).[28] The empirical binding energy ($E_{\text{bind}}$) is given as

$$E_{\text{GEMDOCK-Binding}} = E_{\text{inter}} + E_{\text{intra}} \tag{1}$$

where $E_{\text{inter}}$ and $E_{\text{intra}}$ are the intermolecular and intramolecular energies, respectively.[24] The energy function, GEMDOCK−Pharma, can be dissected into the following terms:[27]

$$E_{\text{GEMDOCK-Pharma}} = E_{\text{GEMDOCK-Binding}} + E_{\text{pharma}} + E_{\text{ligpre}} \tag{2}$$

where $E_{\text{GEMDOCK-Binding}}$ is the empirical binding energy defined in eq 1, $E_{\text{pharma}}$ is the energy of binding site pharmacophores (hot spots), and $E_{\text{ligpre}}$ is a penalty value if a ligand does not satisfy the ligand preferences.[28] $E_{\text{pharma}}$ and $E_{\text{ligpre}}$ are especially useful in selecting active compounds from hundreds of thousands of nonactive compounds by excluding ligands that violate the characteristics of known active ligands, thereby improving the number of true positives. When GEMDOCK uses a pharmacophore-based scoring function, some known active ligands (more than two) are required for evolving the pharmacological consensus according to our previous results.[28]

**GOLD 2.1 Docking.** GOLD[20] is a widely used and reliable docking tool. Standard parameters of the GOLD program were used in this study. For each of the 10 genetic algorithm (GA) runs, a maximum number of 10 000 operations were performed on a population of 50 individuals. Operator weights for crossover, mutation, and migration were set to 95%, 95%, and 10%, respectively. The maximum distance between hydrogen donors and fitting points was set to 2 Å, and nonbonded van der Waals energy was cut off at 4.0 Å. To further speed up the calculation, the GA docking was stopped when the top three solutions were within 1.5 Å root mean square distance of each other. These parameters are chosen according to the standard default settings recommended by the authors for virtual screening.

GOLD offered two scoring functions that were called the GoldScore [20] and the ChemScore functions.[29] The GoldScore function was made up of three components: protein−ligand hydrogen-bond energy ($E_{\text{H\_Bond\_Energy}}$), protein−ligand van der Waals energy ($E_{\text{Complex\_Energy}}$), and ligand internal van der Waals energy and ligand torsional strain energy ($E_{\text{Internal\_Energy}}$). Here, the GoldScore function was divided into two kinds of functions (GOLD−GoldScore and GOLD−Goldinter), which were given as[20]

$$E_{\text{GOLD-GoldScore}} = -(E_{\text{H\_Bond\_Energy}} + E_{\text{Complex\_Energy}}) - E_{\text{Internal\_Energy}} \tag{3}$$

and

$$E_{\text{GOLD-Goldinter}} = -(E_{\text{H\_Bond\_Energy}} + E_{\text{Complex\_Energy}}) \tag{4}$$

The ChemScore function was derived empirically from a set of 82 protein−ligand complexes by regression against measured affinity data. The ChemScore function was defined as[29]

$$\Delta G_{\text{GOLD-ChemScore}} = \Delta G_0 + \Delta G_{\text{hbond}} + \Delta G_{\text{metal}} + \Delta G_{\text{lipo}} + \Delta G_{\text{rot}} \tag{5}$$

Each component of this equation is the product of a term dependent on the magnitude of a particular physical contribution to free energy and a scale factor determined by regression. $\Delta G_{\text{hbond}}$ was the hydrogen bond contribution, $\Delta G_{\text{metal}}$ and $\Delta G_{\text{lipo}}$ were metal−ligand and lipophilic binding contributions, respectively, and $\Delta G_{\text{rot}}$ was a term that penalized flexibility.

Here, two docking methods (GEMDOCK and GOLD) and five scoring functions (GEMDOCK−Binding, GEMDOCK−Pharma, GOLD−GoldScore, GOLD−Goldinter, and GOLD−ChemScore) were used to study the screening performance of data fusion. To analyze the performance uniformly, the fitness scores of these five scoring functions were taken as the negative of the sum of the component energy terms, so that larger fitness scores were better.

**2.3. Performance Evaluation.** It is important to have objective criteria for evaluating the overall quality (and performance) of a scoring method. Some common factors used for this purpose are false positive (FP) rate, yield (the percentage of active ligands in the hit list), enrichment, and goodness-of-hit (GH score). Suppose that $A_h$ is the number of active ligands among the $T_h$ highest-ranking compounds (i.e., the hit list), $A$ is the total number of active ligands in the database, and $T$ is the total number of compounds in the database. Then $A_h/T_h$ (%) is the hit rate and $(T_h - A_h)/(T - A)$ (%) is the FP rate, respectively. The enrichment is defined as $(A_h/T_h)/(A/T)$. The GH score is defined as[30]

$$\text{GH} = \left[\frac{A_h(3A + T_h)}{4T_hA}\right]\left(1 - \frac{T_h - A_h}{T - A}\right)$$

The GH score contains a coefficient to penalize excessive hit list size and, when evaluating hit lists, is calibrated by weighting the score with respect to the yield and coverage. The GH score ranges from 0.0 to 1.0, where 1.0 represents a perfect hit list (i.e., containing all of, and only, the active ligands). Here, we took the averages of FP rates, enrichments, and GH scores. For example, the averages of the FP rate and enrichment are defined as

$$\textstyle\sum_{i=1}^{A}(T_h^i - i)/(T - A) \tag{6}$$

and

$$[\textstyle\sum_{i=1}^{A}(i/T_h^i)/(A/T)]/A \qquad (7)$$

respectively, where $T_h^i$ is the number of compounds in a hit list containing $i$ active compounds.

**2.4. Methods of Data Fusion.** Our approach to combination methods and CS in VS is analogous to those used in IR[12,16,17] and in microarray gene expression analysis.[23] Here, we explore the fundamental question, that is, when and how two scoring functions should be combined in order to achieve a performance higher than both individual scoring functions. Because the number of compounds is in the thousands or even tens of thousands, listing all mathematically possible scoring functions would be a computationally intractable problem. Therefore, we instead chose to take a combinatorial approach to the problem that focuses on taking a group of $m$ scoring functions and evaluates the performance of all possible combinations, which are $\sum_{k=1}^{m}(C_m^k) = 2^m - 1$ (when $m$ is 5, this number is 31). In addition, when we track the performance of all combinations, we investigate specifically when and why any combinations outperform all individual scoring functions in terms of the performance and the scoring characteristics of each of the individual scoring functions.

A scoring function $S_A(x)$ of the scoring method A is a function that assigns a real number to each compound $x$ in the set of all $n$ compounds $D = \{c_1, c_2, ..., c_n\}$. Hence, the scoring function $S_A(x)$ is a function from $D \rightarrow \mathscr{R}$ (the set of real numbers). When treating $S_A(x)$ as an array of real numbers, sorting the array and assigning a rank to each of their compounds would transfer the scoring function $S_A(x)$ to a ranking function $R_A(x)$ from $D$ to $N$ where $N = \{1, 2, ..., n\}$. In the following, we elaborate on the issues of performance evaluation and methods of combination.

To fairly compare and correctly combine multiple scoring functions, one has to normalize the scores obtained by different methods. In our approach, we normalize all scoring functions $S_A(x)$: $D \rightarrow \mathscr{R}$ to the range of $x$, which is less than or equal to 1 but greater than or equal to zero; that is, $S'_A(x)$: $D \rightarrow [0, 1]$, as follows:

$$S'_A(x) = \frac{S_A(x) - S_{\min}}{S_{\max} - S_{\min}} \qquad x \in D \qquad (8)$$

where $S_{\max}$ is the maximum value and $S_{\min}$ is the minimum value of $S_A(x_j)$, respectively, where $1 \le j \le n$; $n$ is the number of compounds in the list. Here, $S_{\max}$ is the first rank and $S_{\min}$ the last rank among $n$ compounds.

**Methods of Combination.** Given a list of $m$ scoring functions, there are several different methods of combination, such as rank by voting, rank by rank, and rank by score. Rank by voting has been reported to have a poor performance.[22] In this paper, we considered two combinations using RCS and SCS. Because we distinguish the two functions [ranking function $R_A(x)$ and normalized scoring function $S'_A(x)$] for a scoring method $A$, we calculate the scoring function for RCS and SCS of the $m$ scoring methods $A_k$, where $k = 1, 2, ..., m$, as follows:

$$S_R(x) = \sum_{k=1}^{m} R_{A_k}(x)/m \qquad \text{(for RCS)} \qquad (9)$$

and

$$S_S(x) = \sum_{k=1}^{m} S_{A_k}(x)/m \qquad \text{(for SCS)} \qquad (10)$$

When we sort $S_R(x)$ and $S_S(x)$ into ascending and descending order, respectively, the ranking functions $R_R(x)$ and $R_S(x)$ can be obtained for RCS and SCS, respectively. We note that, in the two functions, we simply assign equal weight to each scoring method. Combination methods that give different weights to each individual scoring method have been reported.[31] The weighting method of scoring functions is a part of our future work.

**Rank/Score Graph.** In the process of searching for prediction variables or criteria for consensus scoring and method combination, we have defined various performance factors to evaluate scoring method A and various methods of combination. In this paper, we explore the scoring characteristics of scoring method A by calculating the rank/score function $f_A$ as follows:

$$f_A(j) = S'_A R_A^{-1}(j) = S'_A[R_A^{-1}(j)] \qquad (11)$$

where $j$ is the rank of compound $x$, which has the score $f_A(j)$; that is, $j$ is in $N = \{1, 2, 3, ..., n\}$. We note that $N$ is not the set of compounds (which is $D$) but is the set of all positive integers less than or equal to $n$. In fact, $N$ is used as the index set for the ranking function value. The rank/score function $f_A$ so defined signifies the scoring behavior of scoring method A and is independent of the compounds. The graph of the rank/score function $y = f_A(x)$ with respect to scoring method A is the rank/score graph of A. The $x$ and $y$ axes of a rank/score graph are the rank and the normalized score, respectively. The variation ($R/S_{\text{var}}$) of a rank/score graph and the relative performance measurement ($P_l$ /$P_h$) of combining two scoring functions A and B are defined as

$$R/S_{\text{var}}(f_A,f_B) = \{\sum_{j=1}^{n}[f_A(j) - f_B(j)]^2/n\}^{1/2} \qquad (12)$$

and

$$P_l/P_h = \min[P(A),P(B)]/\max[P(A),P(B)] \qquad (13)$$

where $n$ is the number of compounds in the hit list and $j$ is the rank of the compound with score $f_h(j)$, where $h$ = A or B; $P$(A) and $P$(B) are the performances (measured as GH scores and false positive rates) of methods A and B, respectively.

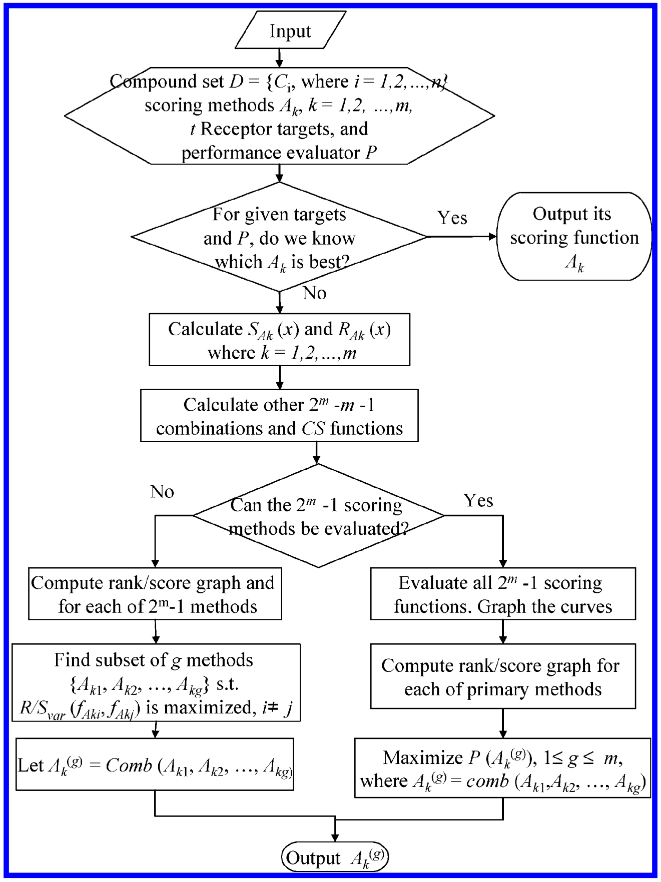
In IR, CS has been demonstrated to improve the performance when the combinations of the scoring functions involved have high performance (e.g., low FP rates or high GH scores) and their variation of the rank/score graph was large. Here, a new CS index (called CS$_{\text{index}}$), which is an indicative criterion for combining two scoring functions A and B from $m$ ($m \ge 2$) scoring methods, was developed to guide the combinations in VS and is defined as

$$\text{CS}_{\text{index}}(A,B) = g(R/S_{\text{var}}(f_A,f_B)) + g[P(A,B)]$$

where

$$g[P(A,B)] = g[P(A) + P(B) - 2P_m) + g(P_l/P_h) \qquad (14)$$

**Chart 1.** Flowchart of the RCS/SCS Algorithm


Input
Compound set $D = \{C_i$, where $i = 1,2,...,n\}$
scoring methods $A_k$, $k = 1,2, ...,m$,
$t$ Receptor targets, and
performance evaluator $P$
For given targets
and $P$, do we know
which $A_k$ is best?
Yes
Output its
scoring function
$A_k$
No
Calculate $S_{Ak}(x)$ and $R_{Ak}(x)$
where $k = 1,2,...,m$
Calculate other $2^m$ -$m$ -1
combinations and $CS$ functions
No
Can the $2^m$ -1 scoring
methods be evaluated?
Yes
Compute rank/score graph and
for each of $2^m$-1 methods
Evaluate all $2^m$ -1 scoring
functions. Graph the curves
Find subset of $g$ methods
$\{A_{k1}, A_{k2}, ..., A_{kg}\}$ s.t.
$R/S_{var}(f_{Aki}, f_{Akj})$ is maximized, $i \neq j$
Compute rank/score graph for
each of primary methods
Let $A_k^{(g)} = Comb(A_{k1}, A_{k2}, ..., A_{kg})$
Maximize $P(A_k^{(g)})$, $1 \leq g \leq m$,
where $A_k^{(g)} = comb(A_{k1}, A_{k2}, ..., A_{kg})$
Output $A_k^{(g)}$

$g(v)$ is a normalization function (i.e., $g(v)$: $v \rightarrow [0, 1]$), and $CS_{index}$ ranges between 0 and 2; $P_m$ is the mean performance of $m$ primary scoring functions [i.e., $P_m = \sum_{k=1}^{m} P(A_k)/m$].

**Algorithm.** We provided a CS procedure for both RCS and SCS to improve the screening accuracy in VS. The flowchart of the algorithm is given in Chart 1, and a more detailed description of the algorithm is included in Appendix A.

## 3. RESULTS AND DISCUSSION

Table 1 shows the overall accuracy of using two docking programs (GEMDOCK and GOLD) and five scoring functions to assess the accuracy of VS methods against four protein targets (TK, DHFR, ER, and ERA). These scoring methods, defined in eqs 1−5, were termed as GEMDOCK−Binding (Method A), GEMDOCK−Pharma (Method B), GOLD−GoldScore (Method C), GOLD−Goldinter (Method D), and GOLD−ChemScore (Method E). For each method, the first term denotes the docking tool and the second term represents the scoring function used. For example, Method A uses GEMDOCK as the docking tool and eq 1 as the scoring function; Method E uses GOLD as the docking tool and eq 5 as the scoring function. The average FP rate (eq 6) and average GH score were used to evaluate the screening accuracy. Among these five methods, the accuracy of GEMDOCK−Pharma was the best for TK and both ER receptors and GOLD−Goldinter outperformed other methods for DHFR.

Table 2 shows FP rates of GEMDOCK and four comparable approaches (Surflex,[32] DOCK,[18] FlexX,[19] and GOLD[20]) for screening ER and TK. All of these methods were tested using the same reference protein and screening database with true positive rates ranging from 80 to 100%. GEMDOCK−Pharma (GEMDOCK with pharmacological preferences) was superior to the comparative approaches and GOLD−GoldScore (GOLD using eq 3 as the scoring function) was better than FlexX and DOCK, two widely used docking tools. For example, the FP rates were 2.3% (GEMDOCK−Binding), 0.4% (GEMDOCK−Pharma), 1.6% (Surflex), 17.4% (DOCK), 70.9% (FlexX), and 8.3% (GOLD−GoldScore) when the true positive rate was 90% for ER antagonists. When known active ligands were not available, GEMDOCK could use a purely empirical scoring function (GEMDOCK−Binding, Method A), and Tables 1 and 2 show that the screening accuracy of GEMDOCK is somewhat influenced and is comparable to that of comparative methods (e.g., DOCK, FlexX, and GOLD) on these ligand data sets.

Our consensus scoring methods consist of rank combinations and score combinations on five methods, including Methods A, B, C, D, and E (Table 1). We initially used the screening of TK inhibitors to provide a perspective of the enrichment improvements that can be realized from the particular consensus approach used. Table 3 shows the ranks of 10 TK known active ligands and average accuracies using five primary methods and four pairing rank combinations to screen TK inhibitors from a data set of 1000 compounds. On the basis of the ranks of these known inhibitors and eqs 6 and 7, we can calculate the FP rates, enrichments, and GH scores of various primary methods and consensus approaches.

A summary of the results of the VS studies with various consensus methods for TK, DHFR, and ER, and ERA are summarized in Figure 1 and Tables 4 and 5. Figure 1 plots average GH scores of all 31 possible combinations including the five individual scoring functions. The $y$-axis values for each combination (including the single case) are sorted in ascending order in each group of $k$ combinations, $k = 1$, 2, 3, 4, and 5, respectively. A $k$ combination method means that it combines $k$ methods. For example, the number of 2-combination methods is 10 (i.e., $C_5^2 = 10$) in this paper. Method BD is the combination of Methods B and D, and Method CDE is the combination of Methods C, D, and E. Tables 4 (RCS) and 5 (SCS) give average FP rates and average GH scores of five kinds of $k$-combination methods for screening four targets. According to these experimental results, the behavior of RCS and SCS is similar. Therefore, we focus on the analysis of RCS in the following.

The average accuracy improved with the increase of fused methods (Figure 1, Table 4). Five individual methods for screening TK found that the best GH score and best false positive rate are 0.23 and 11.21%, respectively. When method fusions with rank combinations were carried out by combining a pair of methods one by one, the accuracies improved from 0.23 to 0.29 for the average of the overall GH score, and the average of false positive rates dropped from 11.21 to 7.77%. Fusing three and four selected methods maintained mean GH scores at 0.29 and 0.28 while decreasing the false positive rates to mean values of 5.40 and 4.04%, respectively.

The average accuracy level improves with the number of fused methods (Table 4), but strikingly, the maximum accuracy always occurs in the combination of a pair of

**Table 1.** Screening Accuracies of Five Methods for Screening TK, DHFR, ER-Antagonist Receptors, and ER-Agonist Receptors[a]

| target protein[b] | measure factor | GEMDOCK−Binding (Method A[c]) | GEMDOCK−Pharma (Method B[d]) | GOLD−GoldScore (Method C[e]) | GOLD−Goldinter (Method D[f]) | GOLD−ChemScore (Method E[g]) |
|---|---|---|---|---|---|---|
| TK | average enrichment | 12.29 | **42.27** | 10.34 | 7.09 | 1.32 |
| | average FP rate (%) | 4.11 | **0.82** | 5.04 | 7.61 | 38.48 |
| | average GH score | 0.22 | **0.45** | 0.20 | 0.17 | 0.08 |
| DHFR | average enrichment | 29.57 | 70.21 | 29.40 | **90.64** | 1.17 |
| | average FP rate (%) | 3.24 | **0.32** | 15.49 | 1.48 | 50.04 |
| | average GH score | 0.35 | 0.66 | 0.32 | **0.81** | 0.05 |
| ER-antagonist | average enrichment | 34.88 | **92.19** | 34.07 | 75.14 | 67.14 |
| | average FP rate (%) | 1.32 | **0.13** | 20.44 | 0.88 | 1.17 |
| | average GH score | 0.39 | **0.83** | 0.34 | 0.70 | 0.64 |
| ER-agonist | average enrichment | 6.94 | **45.66** | 3.50 | 15.21 | 25.09 |
| | average FP rate (%) | 7.83 | **0.75** | 21.67 | 6.40 | 5.24 |
| | average GH score | 0.17 | **0.48** | 0.12 | 0.23 | 0.31 |

[a] The bold case value is the best score. GEMDOCK−Pharma with the pharmacophore-based scoring function and GOLD with the Goldinter score are superior to the others. [b] TK: HIV-1 thymidine kinase (PDB code: 1kim); DHFR: human dihydrofolate reductase (PDB code:1hfr); ER-antagonist receptor: estrogen receptor of antagonists (PDB code: 3ert); and ER-agonist receptor: estrogen receptor of agonists (PDB code: 1gwr). [c] Method A uses GEMDOCK as the docking tool and eq 1 as the scoring function. [d] Method B uses GEMDOCK as the docking tool and eq 2 as the scoring function. [e] Method C uses GOLD as the docking tool and eq 3 as the scoring function. [f] Method D uses GOLD as the docking tool and eq 4 as the scoring function. [g] Method E uses GOLD as the docking tool and eq 5 as the scoring function.

**Table 2.** Comparison of GEMDOCK with Other Methods for Screening the ER Antagonists and TK Inhibitors by False Positive Rates (%)[a]

| target protein | true positive (%) | GEMDOCK−Binding[b] | GEMDOCK−Pharma[b] | Surflex[c] | DOCK[c] | FlexX[c] | GOLD[c] |
|---|---|---|---|---|---|---|---|
| ER-antagonists | 80 | 1.5 (15/990)[d] | **0.0 (0/990)** | 1.3 | 13.3 | 57.8 | 5.3 |
| | 90 | 2.3 (23/990) | **0.4 (4/990)** | 1.6 | 17.4 | 70.9 | 8.3 |
| | 100 | 5.2 (51/990) | **0.9 (9/990)** | 2.9 | 18.9 | *e* | 23.4 |
| thymidine kinase | 80 | 4.7 (47/990) | **0.6 (6/990)** | 0.9 | 23.4 | 8.8 | 8.3 |
| | 90 | 8.9 (88/990) | **1.3 (13/990)** | 2.8 | 25.5 | 13.3 | 9.1 |
| | 100 | 9.7 (96/990) | **2.9 (29/990)** | 3.2 | 27.0 | 19.4 | 9.3 |

[a] Using the same data set proposed by Bissantz et al.[10] The bold case value is the best score. [b] GEMDOCK uses eqs 1 and 2 as scoring functions for GEMDOCK−Binding and GEMDOCK−Pharma, respectively. [c] Directly summarized from ref 32. [d] The false positive rate from 990 random ligands. [e] FlexX could not calculate the docked solution for EST09.

**Table 3.** Ranks of 10 Known TK Inhibitors Using Five Primary Scoring Methods and Four Pairing Rank Combinations for Screening TK Inhibitors

| known ligand ID[a] | Method A[b] | Method B | Method C | Method D | Method E | Method AB[c] | Method CD | Method BC | Method BD |
|---|---|---|---|---|---|---|---|---|---|
| dt | 27 | 6 | 28 | 28 | 314 | 5 | 24 | 4 | 1 |
| idu | 30 | 10 | 15 | 53 | 362 | 7 | 28 | 2 | 7 |
| hpt | 106 | 22 | 103 | 173 | 263 | 32 | 113 | 26 | 39 |
| ahiu | 26 | 13 | 13 | 25 | 252 | 6 | 14 | 3 | 2 |
| dhbt | 40 | 14 | 94 | 78 | 325 | 12 | 70 | 20 | 11 |
| hmtt | 97 | 39 | 35 | 59 | 518 | 37 | 38 | 10 | 14 |
| mct | 23 | 9 | 70 | 52 | 188 | 4 | 51 | 12 | 6 |
| acv | 55 | 11 | 80 | 155 | 494 | 15 | 99 | 16 | 28 |
| gcv | 42 | 7 | 79 | 134 | 618 | 10 | 91 | 13 | 21 |
| pcv | 16 | 5 | 37 | 51 | 531 | 1 | 35 | 6 | 5 |
| average false positive rate (%) | 4.11 | 0.82 | 5.04 | 7.61 | 38.48 | 0.75 | 5.13 | 0.58 | 0.80 |
| average GH score | 0.22 | 0.45 | 0.20 | 0.17 | 0.08 | 0.56 | 0.20 | 0.54 | 0.58 |
| average enrichment | 12.29 | 42.27 | 10.34 | 7.09 | 1.32 | 57.48 | 9.93 | 54.56 | 59.85 |

[a] The abbreviations are as follows: dt, deoxythymidine; idu, 5-iododeoxyuridine; hpt, 6-(3-hydrody-propyl-thymine); ahiu, 5-iodouracil anhydrohexitol nucleoside; dhbt, 6-[3-hydroxy-2-(hydroxymethyl)propyl]-5-methyl-2,4(1h,3h)-pyrimidinedione [6-(dihydroxy-isobutyl)-thymine]; hmtt, 6-[6-hydroxymethy-5-methyl-2,4-dioxo-hexahydro-pyrimidin-5-yl-methyl]-5-methyl-1H-pyrimidin-2,4-dione; mct, (North)-methanocarba-thymidine; acv, aciclovir; gcv, ganciclovir; pcv, penciclovir. [b] The primary scoring methods (A, B, C, D, and E) are defined in Table 1. [c] The combining pair method is combined from five primary scoring methods (A, B, C, D, and E).

methods (Figure 1 and Table 4). Thus, the unique contribution of data fusion is most clearly observed when one individually considers the results obtained with each of the possible combinations. In all of the screening sets in this paper, the best composition consistently appeared with the combination of Methods B and D. The ER antagonists provide a clear example. For ER antagonists, the GH scores of Method A and Method C were 0.39 and 0.34 and the other three methods (Methods B, D, and E) had good GH scores with 0.83, 0.70, and 0.64, respectively (Table 1). As shown in Figure 1c, combinations with Method A or Method C may reduce the performance of an individual method. For example, Methods CD and BC performed worse than Methods B and D. One possible reason is that these less

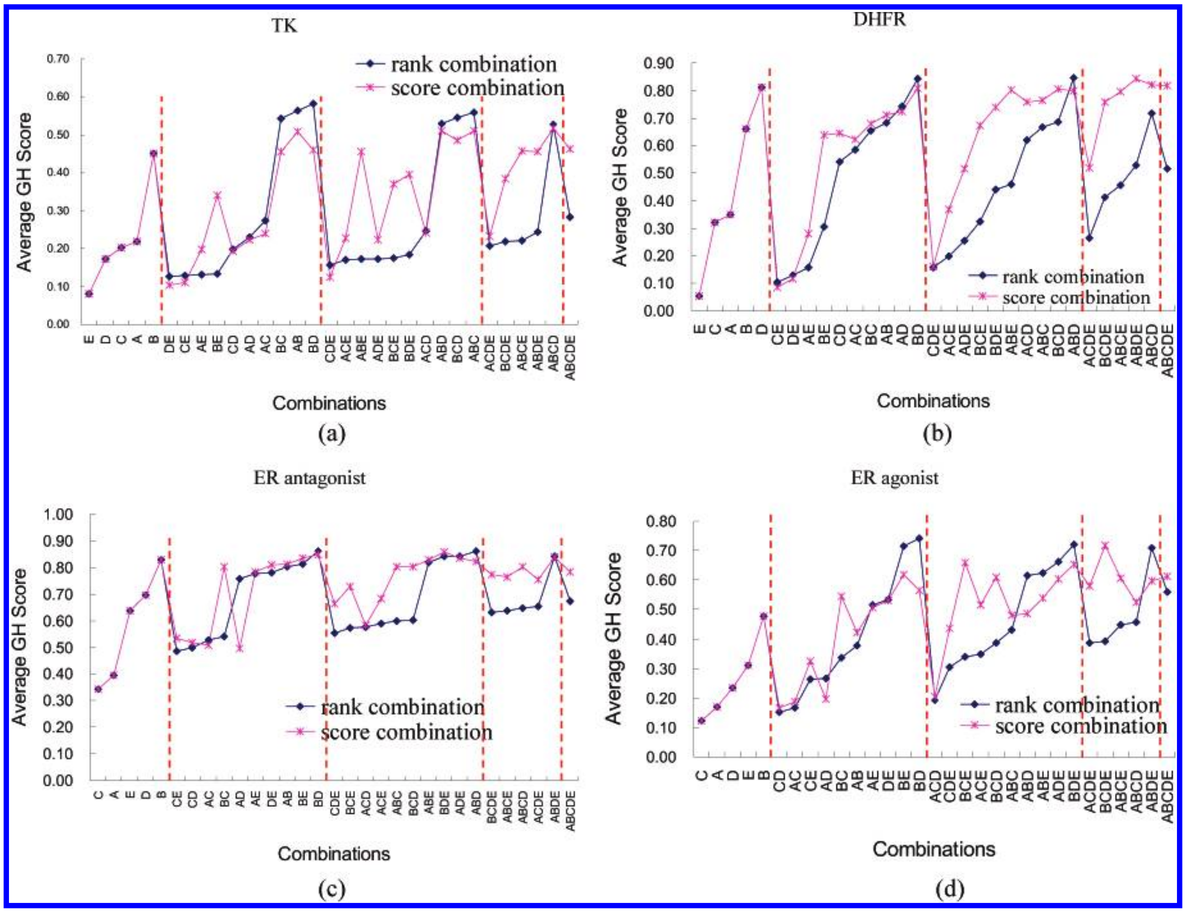


**Figure 1.** Average GH scores of 31 various rank combinations and scoring combinations of five methods for four virtual screening targets: (a) TK, (b) DHFR, (c) ER, and (d) ERA. These five methods (i.e., A, B, C, D, and E) are defined in Table 1.

accurate methods are predominantly adding noise that overwhelms the correction ability of fusion. On the other hand, combinations with Method B or Method D performed comparatively better than the other Method combinations. Method BD had the highest value (0.86) in the GH score and the lowest value in the false positive rate (0.04%). Other targets (i.e., TK, DHFR, and ERA) in Table 4 and Figure 1 show similar results. This phenomenon indicated data fusion could improve the quality of screening if each of the combination methods has relatively high performance.

Our data also indicate that the differences between methods are also important. Figure 2 shows the rank/score graphs of five individual scoring methods, and Table 6 shows the variations of rank/score graphs of 10 compositions combining two methods for four screening targets. The scoring value shown in Figure 2 was normalized through eq 8. The variation of the rank/score graph of Method AB, on average, is the smallest (i.e., the rank score graphs are the most similar) because Methods A and B used the same docking tool and similar scoring functions. Method CD has a similar phenomenon. Method B consistently outperformed Method A (Table 1), and fusion methods involving Method B are consistently better than those methods combining with Method A in four test cases (Figure 1). For DHFR, ER, and ERA, Method D is better than Method C and the fusion methods with Method D consistently outperformed the fusion methods with Method C. According to these observations, we could divide these five methods into three groups. The first group consisted of Methods A and B, the second group included Methods C and D, and the final group is GOLD−ChemScore (Method E). Notably, the greatest difference between the 10 possible pairs of rank/score graphs was that between Methods B and D (Figures 1 and 2 and Table 6), the best performing pair fusion, and Method BD also brought the best GH score for all test cases (Figure 1). In Figure 1b (DHFR), Methods B and D had the highest GH score (0.66 and 0.81, respectively) among the primary methods (Table 1), and the combination of these two methods had the best GH score (0.84) and the lowest false positive rate (0.14%) among the combinations with two methods. A similar phenomenon occurred in the ER antagonist study (Figure 1c). On the other hand, Methods A and B had the highest GH score (0.22 and 0.45) among the primary methods for TK (Figure 1a), but the best combination was Method BD among the 10 pair combinations. The critical point is that the rank/score variation between Methods B and D is larger than the rank/score variation of Methods A and B (Figure 2)—consistent with the use of different docking algorithms between B and D but not between B and A.

These experimental results using the BD model implied that the variation of the rank/score graph might be useful to improve the screening accuracy in both VS and IR. This concept is supported by observations of a similar phenomenon occurring in ER agonists (Figures 1d and 2d). Specifically, Methods B and E had the highest GH scores, but their rank/score variation is smaller than the variation of Methods B and D. The performance of Method BD was better than that of Method BE for ER agonists.

More importantly, we present evidence that, in general, a pairing combination can be expected to improve the perfor-

**Table 4.** Screening Accuracies of Different Rank Combinations of Five Methods for Screening Four Targets: TK, DHFR, ER, and ERA[a]

| measurement factors | single (5)[b] | | | | 2-com (10)[c] | | | | 3-com (10)[d] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| average false positive rate (%) | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA |
| average | 11.21 | 14.12 | 4.79 | 8.38 | 7.77 | 9.91 | 3.31 | 4.24 | 5.40 | 6.16 | 2.04 | 2.46 |
| SD | 15.44 | 20.98 | 8.76 | 7.89 | 7.34 | 9.80 | 4.09 | 3.76 | 3.63 | 4.73 | 1.72 | 2.07 |
| max value | 38.48 | 50.04 | 20.44 | 21.67 | 17.35 | 27.56 | 9.22 | 11.86 | 9.92 | 14.64 | 4.16 | 7.55 |
| min value | 0.82 | 0.32 | 0.13 | 0.75 | 0.58 | 0.14 | **0.04** | 0.99 | **0.53** | **0.07** | **0.04** | 0.72 |
| average GH score | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA |
| average | 0.23 | 0.44 | 0.58 | 0.26 | **0.29** | 0.47 | 0.68 | 0.41 | **0.29** | 0.47 | **0.69** | 0.46 |
| SD | 0.14 | 0.30 | 0.21 | 0.14 | 0.19 | 0.28 | 0.15 | 0.21 | 0.18 | 0.23 | 0.13 | 0.18 |
| max value | 0.45 | 0.81 | 0.83 | 0.48 | **0.58** | 0.84 | **0.86** | **0.74** | 0.56 | **0.85** | **0.86** | 0.72 |
| min value | 0.08 | 0.05 | 0.34 | 0.12 | 0.13 | 0.10 | 0.48 | 0.15 | 0.16 | 0.16 | 0.55 | 0.19 |

| measurement factors | 4-com (5)[e] | | | | 5-com (1)[f] | | | |
|---|---|---|---|---|---|---|---|---|
| average false positive rate (%) | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA |
| average | **4.04** | **4.11** | **1.39** | **1.52** | *g* | *g* | *g* | *g* |
| SD | 1.90 | 2.44 | 0.74 | 0.61 | *g* | *g* | *g* | *g* |
| max value | 5.88 | 6.69 | 1.86 | 1.98 | 3.10 | 3.00 | 1.04 | 1.08 |
| min value | 0.83 | 1.07 | 0.08 | 0.55 | 3.10 | 3.00 | 1.04 | 1.08 |
| average GH score | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA |
| average | 0.28 | **0.48** | 0.68 | **0.48** | *g* | *g* | *g* | *g* |
| SD | 0.14 | 0.17 | 0.09 | 0.13 | *g* | *g* | *g* | *g* |
| max value | 0.53 | 0.72 | 0.84 | 0.71 | 0.28 | 0.52 | 0.67 | 0.56 |
| min value | 0.21 | 0.27 | 0.63 | 0.39 | 0.28 | 0.52 | 0.67 | 0.56 |

[a] The methods and targets are defined in Table 1, and the bold case value is the best score. [b] Five individual methods. [c] Combination of two selected methods, 10 compositions. [d] Combination of three selected methods, 10 compositions. [e] Combination of four selected methods, five compositions. [f] Combination of five selected methods and only one composition. [g] Average and standard deviation could not be calculated when one value exists.

mance (vs its constituent members) if and only if the normalized $P_l/P_h$ and $R/S_{var}$ of a combining method both have values > 0.5. Figures 3 and 4 and Table 6 are the results of the algorithm (in Chart 1 and Appendix A) when $g = 2$, where pairing combinations were considered and $R/S_{var}(f_A, f_B)$ was used to calculate the bidiversity of methods A and B. Figure 3a shows the relationship between the GH-score improvement and the variation ($R/S_{var}$, eq 12) of the rank/score graphs of the 10 pairing combinations for each target protein. Figure 3b indicates the relationship between the GH-score improvement and the relative performance measurement ($P_l/P_h$, eq 13). These results echo those in the field of data fusion in IR, [16,17] in which studies have shown that CS improves accuracy if the multiple scoring functions involved have high performance and their rank/score variation is large. We further created an additional synthetic variable, the $CS_{index}$ (eq 14), which can be used to integrate these two criteria ($P_l$ /$P_h$ and $R/S_{var}$). Figure 5 shows the relationship between GH-score improvement and the $CS_{index}$ of the 10 pairing combinations for each target protein. A pairing combination often improves screening accuracies when its $CS_{index}$ is more than 1.5. These data suggest that one can use the $CS_{index}$ to estimate predicted benefits from fusion approaches. The overall accuracy of this approach should be readily extendable by future research, which would also be expected to increase the $CS_{index}$ itself by extending the reach to further distinct algorithms and other combination approaches beyond unweighted SCS and RCS.

**Discussion.** CS is a popular strategy for solving the scoring inaccuracy problem in VS. In this study, our CS methods addressed the use of rank combinations and score combinations on five scoring functions related to two docking algorithms. We found that some rank-based combinations outperformed (in terms of average false positive rates and average GH scores) each individual component in the fusion; that is, data fusion was beneficial. More importantly, this study of data fusion, which was based on VS results of 1000 test “compounds” and four receptor targets, suggests that a fusion method is able to improve the screening accuracy in VS only when (a) each of the individual scoring functions has a relatively good performance ($P_l$ /$P_h$ was > 0.5) and (b) the scoring characteristics of each of the scoring functions are quite different ($R/S_{var}$ was > 0.5). The observations of RCS and SCS are summarized as follows:

(a) Combining multiple scoring functions improves enrichment of true positives only if both $g(P_l/P_h) > 0.5$ and $g(R/S_{var}) > 0.5$ (Figure 4). These two prediction indicators can be combined into a single indicator, specifically $CS_{index} > 1.5$ (Figure 5). For example, in ER, the GH scores of Methods B (0.83) and D (0.70) (Table 1) are the best and Method BD (0.86) is the best among 31 combinations (Table 4). The $CS_{index}$ of Method BD is 1.68 (Table 6).

(b) The accuracy of CS was improved by increasing the scoring methods for both RCS and SCS (Tables 4 and 5), but the combination of all scoring methods did not display the best possible performance observed. For RCS, the performance of 2-combination or 3-combination methods outperformed 4-combination or 5-combination methods.

(c) The $R/S_{var}$ criterion is particularly useful and important because VS is often used to screen for compounds that

**Table 5.** Screening Accuracies of Different Score Combinations of Five Methods for Screening Four Targets: TK, DHFR, ER, and ERA[a]

| measuremen factors | single (5)[b] | | | | 2-com (10)[c] | | | | 3-com (10)[d] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| average false positive rate (%) | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA |
| average | 11.21 | 14.12 | 4.79 | 8.38 | 7.46 | 7.67 | 2.57 | 3.13 | 4.07 | 3.02 | 1.12 | 1.28 |
| SD | 15.44 | 20.98 | 8.76 | 7.89 | 9.51 | 12.25 | 4.80 | 3.26 | 5.40 | 5.47 | 2.00 | 1.57 |
| max value | 38.48 | 50.04 | 20.44 | 21.67 | 26.16 | 34.31 | 13.74 | 9.91 | 18.18 | 18.00 | 6.54 | 5.42 |
| min value | 0.82 | 0.32 | 0.13 | 0.75 | 0.73 | 0.10 | 0.07 | 0.33 | **0.64** | 0.10 | **0.07** | 0.23 |
| average GH score | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA |
| average | 0.23 | 0.44 | 0.58 | 0.26 | 0.28 | 0.53 | 0.69 | 0.41 | 0.36 | 0.64 | 0.76 | 0.52 |
| SD | 0.14 | 0.30 | 0.21 | 0.14 | 0.15 | 0.27 | 0.16 | 0.17 | 0.14 | 0.22 | 0.09 | 0.13 |
| max value | 0.45 | 0.81 | 0.83 | 0.48 | 0.51 | 0.81 | 0.85 | 0.62 | 0.51 | 0.81 | **0.86** | 0.66 |
| min value | 0.08 | 0.05 | 0.34 | 0.12 | 0.10 | 0.08 | 0.50 | 0.17 | 0.12 | 0.16 | 0.58 | 0.20 |

| measuremen factors | 4-com (5)[e] | | | | 5-com (1)[f] | | | |
|---|---|---|---|---|---|---|---|---|
| average false positive rate (%) | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA |
| average | **2.01** | **1.15** | **0.44** | **0.50** | [g] | [g] | [g] | [g] |
| SD | 1.92 | 1.45 | 0.41 | 0.27 | [g] | [g] | [g] | [g] |
| max value | 5.40 | 2.91 | 1.08 | 0.92 | 1.14 | 0.09 | 0.22 | 0.34 |
| min value | 0.69 | **0.05** | 0.10 | **0.19** | 1.14 | 0.09 | 0.22 | 0.34 |
| average GH score | TK | DHFR | ER | ERA | TK | DHFR | ER | ERA |
| average | **0.41** | **0.75** | **0.79** | **0.60** | [g] | [g] | [g] | [g] |
| SD | 0.11 | 0.13 | 0.03 | 0.07 | [g] | [g] | [g] | [g] |
| max value | **0.52** | **0.84** | 0.84 | **0.72** | 0.46 | 0.82 | 0.78 | 0.61 |
| min value | 0.23 | 0.52 | 0.75 | 0.52 | 0.46 | 0.82 | 0.78 | 0.61 |

[a] The methods and targets are defined in Table 1, and the bold case value is the best score. [b] Five individual methods. [c] Combination of two selected methods, 10 compositions. [d] Combination of three selected methods, 10 compositions. [e] Combination of four selected methods, five compositions. [f] Combination of five selected methods and only one composition. [g] Average and standard deviation could not be calculated when one value exists.

interact with receptors that have few or no known binding partners; that is, there is no adequate training set to establish an algorithm's veracity/utility. Under these common circumstances, the performance of a given individual scoring function is generally unknown and cannot be evaluated at the point it must be used. As noted above, the variation, $R/S_{\mathrm{var}}$, of a pair of rank/score graphs is a useful index to improve the screening accuracy for combining two individual methods when the individual scoring functions are quite different (or complementary, for example, normalized $R/S_{\mathrm{var}} > 0.5$). As this approach appears target-independent, our approach should be usable in different situations, whether it is running a truly blind screen, a combination screen coupling a blinded set with partial analysis and subsequent use of previous hits as a training set, or a screen with a true training set. Our approach also reveals that approaches that yield the best average GH score/FP (i.e., SCS), which are relevant for screens without training sets, are different from those approaches that optimize individual GH scores (i.e., RCS), which are applicable when a training set is available. This is reflected in the next two points.

(d) The best GH scores of RCS are consistently superior to those of SCS for these four target proteins (Figure 1). Table 4 shows the best RCS-derived individual GH scores: 0.58 (Method BD for TK), 0.85 (Method ABD for DHFR), 0.86 (Methods BD and ABD for ER), and 0.74 (Method BD for ERA). Table 5 shows the best SCS-derived individual GH scores: 0.52 (Method ABCD for TK), 0.84 (Method ABDE for DHFR), 0.86 (Methods BDE for ER), and 0.72 (Method BCDE for ERA).

(e) The best average GH scores and best average FP rates of SCS are superior to those of RCS on all target proteins (Tables 4 and 5). For example, for TK, the best average GH scores are 0.41 (SCS) and 0.29 (RCS) and the best average FP rates are 2.01% (SCS) and 4.04% (RCS). For ER antagonists, the best average GH scores are 0.79 (SCS) and 0.69 (RCS) and the best average FP rates are 0.44% (SCS) and 1.39% (RCS).

(f) For RCS methods, the moderate number of scoring functions, two or three, are the best and sufficient for the purpose of CS (Figure 1). In contrast, the number of combining methods is three or four to achieve the best performance for SCS methods. This phenomenon was also found in data fusion in IR and was consistent with the previous findings for CS.[22]

(g) When combining methods with highly differential performance, Figure 1 shows that SCS works better than RCS. For example, the combinations of BE (for TK) and ABE (for DHFR)are the best and the worst, respectively, among five primary methods. For ER and ERA, the combinations of BC (ER) and BCE (ERA) have similar results.

## 4. CONCLUSION

It has been previously shown that CS improves VS and that CS may be more robust than individual methods because each individual scoring function has strengths and weaknesses with respect to docking algorithms, receptor targets, and the database sets. Furthermore, although consensus scoring does perform better than the average performance

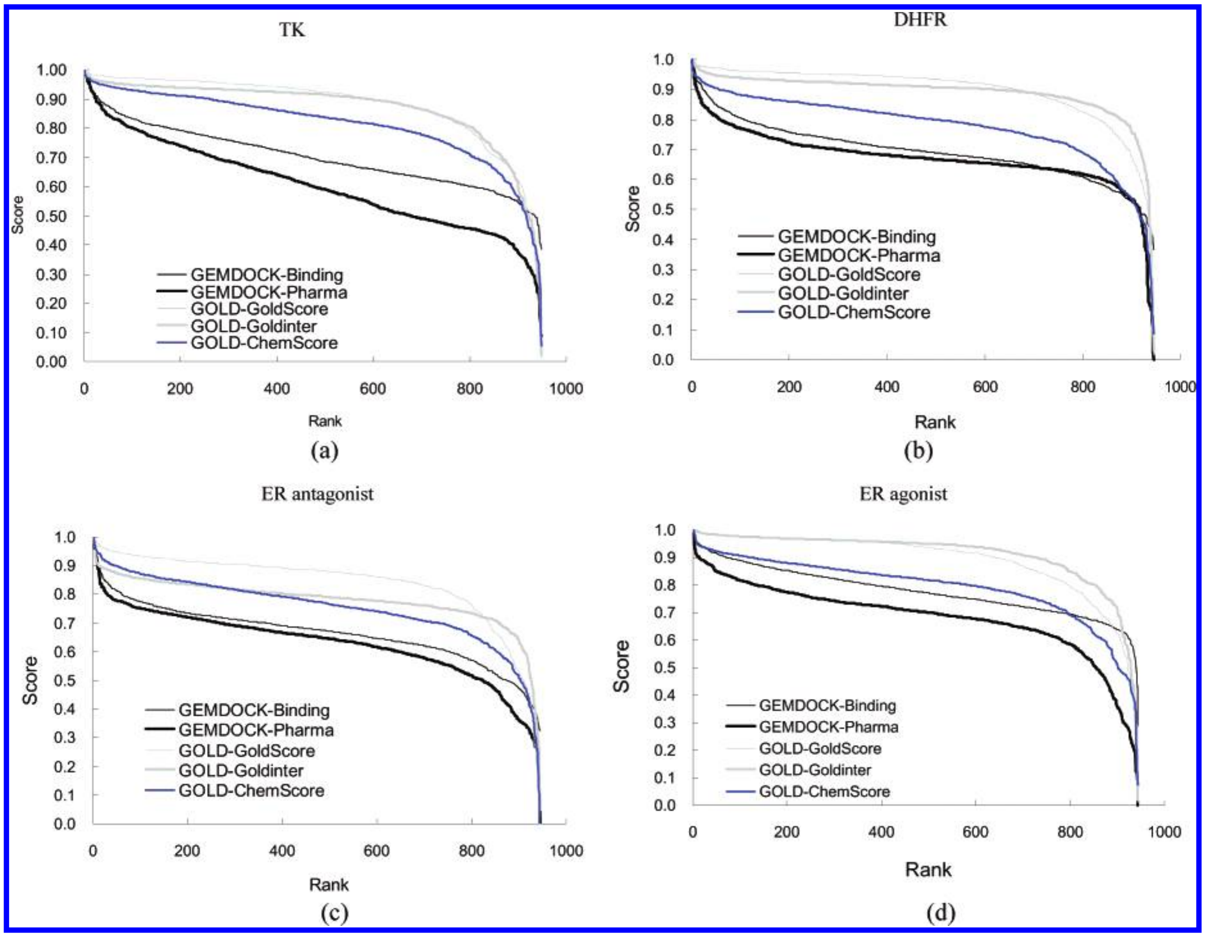


**Figure 2.** Rank/score curves of five methods (defined in Table 1) for four virtual screening targets: (a) TK, (b) DHFR, (c) ER, and (d) ERA.

**Table 6.** Relationships between the GH-Score Improvements with the Performance Ratio ($P_l/P_h$), $CS_{index}$, and the Variation ($R/S_{var}$) of Rank/Score Graph of 10 Pairing Combinations of Five Methods for Four Virtual Screening Targets

| target protein[a] | | AB[b] | AC | AD | AE | BC | BD | BE | CD | CE | DE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TK | $g(P_l/P_h)^c$ | 0.41 | 1.00 | 0.82 | 0.26 | 0.36 | 0.27 | 0.00 | 0.91 | 0.30 | 0.39 |
| | $g(R/S_{var})^d$ | 0.34 | 0.64 | 0.62 | 0.39 | 1.00 | 0.97 | 0.74 | 0.00 | 0.19 | 0.17 |
| | $CS_{index}{}^e$ | 1.34 | 1.64 | 1.37 | 0.39 | 1.92 | 1.74 | 1.03 | 0.80 | 0.19 | 0.19 |
| | RCS[f] | 0.11 | 0.06 | 0.01 | −0.09 | 0.09 | 0.13 | −0.32 | 0.00 | −0.07 | −0.05 |
| | SCS[f] | 0.06 | 0.02 | 0.00 | −0.02 | 0.01 | 0.01 | −0.11 | −0.01 | −0.09 | −0.07 |
| DHFR | $g(P_l/P_h)^c$ | 0.54 | 1.00 | 0.43 | 0.10 | 0.49 | 0.88 | 0.02 | 0.39 | 0.12 | 0.00 |
| | $g(R/S_{var})^d$ | 0.04 | 0.91 | 0.88 | 0.32 | 1.00 | 0.97 | 0.41 | 0.00 | 0.45 | 0.46 |
| | $CS_{index}$ | 0.61 | 1.56 | 1.46 | 0.32 | 1.52 | 1.97 | 0.53 | 0.54 | 0.45 | 0.65 |
| | RCS | 0.02 | 0.23 | −0.07 | −0.19 | 0.01 | 0.03 | −0.36 | −0.27 | −0.22 | −0.68 |
| | SCS | 0.05 | 0.28 | −0.09 | −0.07 | 0.02 | 0.00 | −0.02 | −0.17 | −0.23 | −0.70 |
| ER antagonists(ER) | $g(P_l/P_h)^c$ | 0.12 | 0.92 | 0.30 | 0.41 | 0.00 | 0.86 | 0.71 | 0.16 | 0.25 | 1.00 |
| | $g(R/S_{var})^d$ | 0.00 | 0.82 | 0.46 | 0.29 | 1.00 | 0.68 | 0.47 | 0.21 | 0.28 | 0.03 |
| | $CS_{index}$ | 0.15 | 1.11 | 0.62 | 0.47 | 1.01 | 1.68 | 1.30 | 0.21 | 0.30 | 0.96 |
| | RCS | −0.03 | 0.13 | 0.06 | 0.14 | −0.29 | 0.03 | −0.01 | −0.20 | −0.15 | 0.08 |
| | SCS | −0.01 | 0.11 | −0.20 | 0.15 | −0.03 | 0.02 | 0.01 | −0.18 | −0.10 | 0.11 |
| ER agonists(ERA) | $g(P_l/P_h)^c$ | 0.21 | 0.91 | 0.96 | 0.59 | 0.00 | 0.47 | 0.80 | 0.53 | 0.27 | 1.00 |
| | $g(R/S_{var})^d$ | 0.40 | 0.49 | 0.57 | 0.10 | 0.88 | 1.00 | 0.39 | 0.00 | 0.33 | 0.45 |
| | $CS_{index}$ | 0.70 | 0.77 | 1.07 | 0.43 | 0.93 | 1.61 | 1.39 | 0.08 | 0.33 | 1.22 |
| | RCS | −0.10 | 0.00 | 0.03 | 0.20 | −0.14 | 0.26 | 0.24 | −0.08 | −0.05 | 0.22 |
| | SCS | −0.06 | 0.02 | −0.04 | 0.20 | 0.07 | 0.09 | 0.14 | −0.07 | 0.01 | 0.22 |

[a] Four target proteins (TK, DHFR, ER, and ERA) are defined in Table 1. [b] There are 10 compositions of combining pair methods from five primary scoring methods (A, B, C, D, and E) defined in Table 1. [c] The normalization performance ratio (eq 13) of a pair-combination method. [d] The normalization variation (eq 12) of a rank/score graph of a pair-combination method. [e] A performance indicator (eq 14) of a pair-combination method. [f] The GH-score improvements of rank-based consensus scoring and score-based consensus scoring for RCS and SCS, respectively.

of the individual scoring methods, it does not consistently perform better than the best individual scoring function. In our experiment on the four receptors TK, DHFR, ER, and ERA, the two docking algorithms we used (GEMDOCK and GOLD) have been shown to be very good. Although performances (measured as GH score and false positive rate) of each individual scoring function do vary within each of and among the receptor targets, interesting patterns do stand

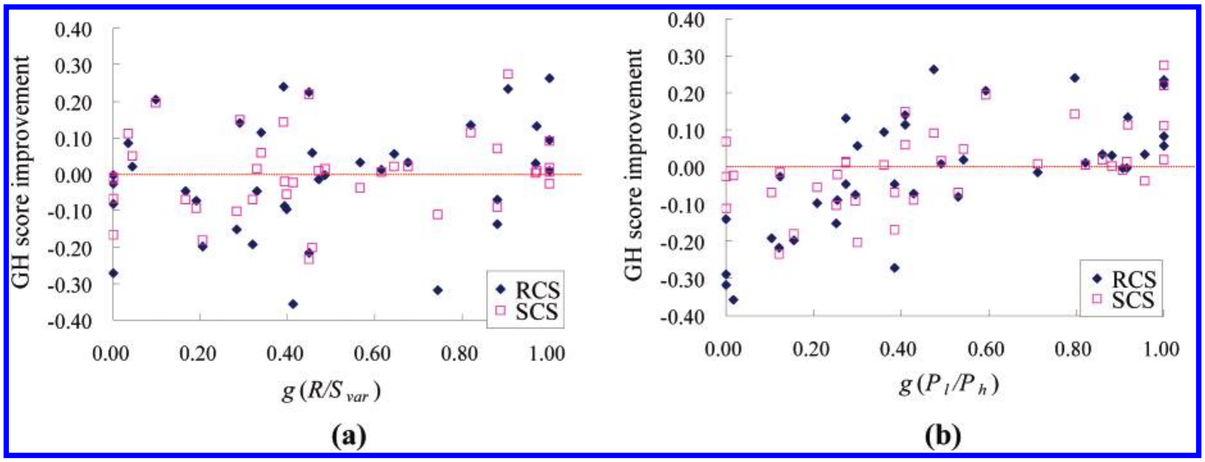


**Figure 3.** Relationships between the GH-score improvement with (a) a normalized value of variation ($R/S_{var}$) of the rank/score graph (the correlation coefficients are 0.135 for RCS and 0.131 for SCS) and (b) a normalized value of $P_l/P_h$ of 40 pairing combinations of five methods for four virtual screening targets (the correlation coefficients are 0.661 for RCS and 0.531 for SCS).

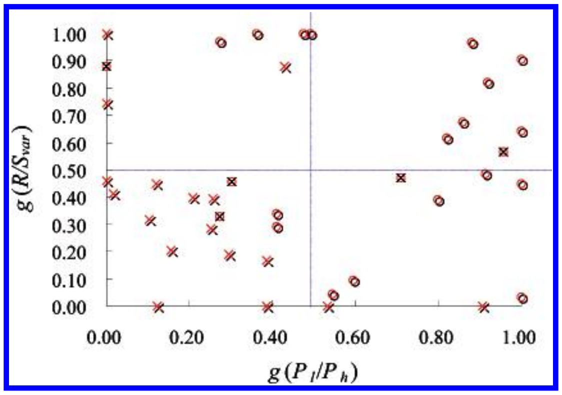


**Figure 4.** GH-score improvements with normalized variations of rank/score graphs ($R/S_{var}$) and a normalized relative performance measurement ($P_l/P_h$) of 40 RCS and SCS pairing combinations of five methods for four virtual screening targets. The positive and negative GH-score improvements are denoted with a circle and a cross, respectively. For positive cases, the mean and variance of sum of $g(R/_{var})$ and $g(P_l/P_h)$ are 1.30 and 0.399, respectively. In contrast, these two values are 0.592 and 0.271 for negative cases. The *t*-test result shows that the positive and negative cases are significantly different.

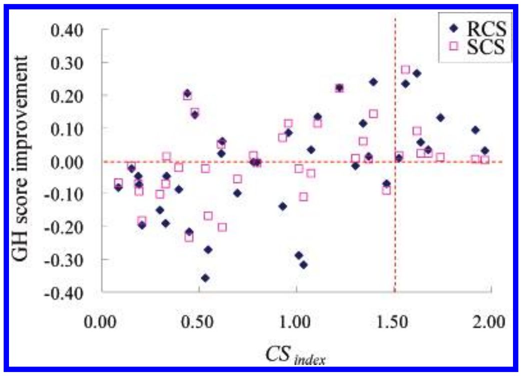


**Figure 5.** Relationships between the GH-score improvement with the $CS_{index}$ (eq 14) of 40 pairing combinations of five methods for four virtual screening targets. The correlation coefficients between the GH-score improvement and $CS_{index}$ are 0.433 (RCS) and 0.331 (SCS).

out where we showed that combinations of two scoring functions leads to significant improvement on average GH score and average FP rate.

We summarize and state the two CS criteria, which would serve as two predictive variables for improving enrichment in VS: CS that combines multiple scoring functions should only be used when (a) the scoring functions involved have high performance and (b) the scoring characteristic of each of the individual scoring functions are quite different. These two CS criteria also work for different performances between SCS and RCS. It has been reported that, on average, score combination is more effective than rank combination. However, we have demonstrated that in a majority of cases under the two CS criteria, rank combination does perform better or as good as score combination. This is analogous to the results reported in IR[12,16,17]. Our second criterion calculates the rank/score function of each scoring function and then computes the differences between the rank/score functions of the scoring methods involved. Our second criterion does not involve a performance evaluation of the combined methods. This criterion is useful because, very often, the performance of individual scoring functions is not known or cannot be evaluated. We believe that our rank-based and score-based consensus scoring (RCS and SCS) procedures and consensus criteria for improving the enrichment in VS should be useful to researchers and practitioners in VS.

Our work, thus, provides a framework to study CS criteria and a procedure (the algorithm) for both rank-based and score-based CS to improve the hit rates, FP rates, enrichment, and the GH score. The procedure is computationally efficient, able to adapt to different situations, and scalable to a large number of compounds and a greater number of combinations. Moreover, we have shown the power of two combinations (pairing combinations) and used the rank/score graph to assess the bidiversity between the two scoring methods used. Our current work represents the first of a series of investigations to explore CS criteria for improving enrichment in VS. It also engenders a whole school of issues and directions worthy of further study, which are summarized as follows:

(1) We will study the extension to three and higher number of combinations of scoring functions using the rank/score graph variation ($R/S_{var}$) as a diversity measurement for the scoring methods involved.

(2) In this paper, we used the rank/score function $f_A$ as the scoring characteristic for the scoring method A. Then, we used the variation on the rank/score function ($R/S_{var}$) to characterize the scoring diversity between two scoring methods A and B. Other parameters such as the difference between the score functions $S_A$ and $S_B$ and the difference

between the rank functions $R_A$ and $R_B$ can also be used to distinguish the scoring diversity. The rank/score graphs (Figure 2) have provided a clear visualization for characterizing the scoring diversity between individual scoring functions.

(3) In our combination (RCS and SCS) of scoring functions, we use averages to compute the scores for the rank and score combinations. Combinations using different weighting schemes can also be used. Hsu and Palumbo [31] presented work on the combination of two scoring methods using a weighted scheme with a step of $^1/_{10}$ as a proportion.

(4) In the future, we will study a more diverse set of docking tools, scoring functions (e.g., knowledge-based, physics-based, and empirical scoring functions), and receptor targets with different binding-site characteristics (e.g., hydrophobic, hydrophilic, missing loop, and highly hydrated) to systematically determine the limitations/advantages of our SCS and RCS procedures and consensus criteria for improving the enrichment in VS.

## APPENDIX A. THE RCS/SCS ALGORITHM

**Given.** A compound set $D$ with $n$ compounds in a compound database (or a hit list); $c_i \in D$; $i = 1, 2, ..., n$; $t$ receptor targets; performance evaluator $P$ (e.g., the GH score or FP rate); and $m$ scoring methods $A_k$ with scoring functions $S_{A_k}(x)$; $k = 1, 2, ..., m$.

**Output.** The best consensus scoring and combination methods for the $t$ receptor targets and the compound set $D$.

**Step 1.** If we know in advance which scoring function works better for a given target or targets, output this scoring function directly. Otherwise, execute the following steps to select the best CS.

**Step 2.** For each receptor target, calculate the scoring functions $S_{A_k}(x)$ using the $m$ scoring methods $A_k$, where $k = 1, 2, ..., m$. Obtain each ranking function $R_{A_k}(x)$ from each $S_{A_k}(x)$ by ranking the scores in $S_{A_k}(x)$ in descending order. (Note: there are the $m$ single scoring methods.)

**Step 3.** Calculate the other $2^m - m - 1$ combinations and CS using eqs 9 and 10. (Note: these are the $\binom{m}{k}$ $k$ combinations, where $k = 2, 3, ..., m$, and the scoring functions are all normalized.) If the $2^m - 1$ scoring methods can be evaluated (including both rank and score combinations), then go to Step 4. Otherwise, go to Step 5.

**Step 4.** The performance of the individual scoring function can be evaluated (i.e., the active and inactive compounds are known).

**Step 4.1.** Evaluate the performance of all of the single and combination scoring functions using evaluator $P$ (e.g., GH score or FP rate). Note that these are the ranking functions $R_{A_k}(x)$ and scoring functions $S_{A_k}(x)$, where $k = 1, 2, ..., 2^m-1$. Graph the performance curve for all of the single and combination functions using rank and score combinations. Order the performance within each of the $m$ groups with $\binom{m}{k}$ combinations, where $k = 1, 2, ..., m$.

**Step 4.2.** For each single scoring method $A$, obtain a rank/score graph using eq 11.

**Step 4.3.** Search in the space of $2^m - m - 1$ consensus scorings and find any combination method $A_k^{(g)}$ that is the combination of the $g$ single scoring methods $\{A_{k1}, A_{k2}, ..., A_{kg}\}$, where $2 \le g \le m$ and $k_j \in [1, m]$, so that (a) $P(A_{kj})$ have high performance (e.g., high GH scores or lower FP rates), (b) $f_{A_{kj}}$ and $f_{A_{ki}}$ are dissimilar and complementary for any $i$, $j$, and $i \neq j$ in $[1, g]$ [i.e., $R/S_{\mathrm{var}}(f_{A_{kj}}, f_{A_{ki}})$ is large], and (c) $P(A_k^{(g)})$ is better than or as good as $P(A_k)$, where $A_k$ are the single scoring functions and $k = 1, 2, ..., m$. The consensus scorings often improve the screening accuracy when the value $\mathrm{CS}_{\mathrm{index}}$ (eq 13) is more than 1.2.

**Step 4.4.** The combination method $A_k^{(g)}$ is the desired consensus scoring method that we seek for the receptor target and the compound set $D$. Go to Step 6.

**Step 5.** The performance of the individual scoring function is unknown (i.e., the active and inactive compounds are unknown).

**Step 5.1.** For each single scoring method $A$, obtain the rank/score graph using eq 11.

**Step 5.2.** Search in the space of the $m$ single scoring functions. Find any group of $g$ single scoring functions $A^{(g)} = \{A_{k1}, A_{k2}, ..., A_{kg}\}$, where $2 \le g \le m$ and $k_j \in [1, m]$, so that $f_{A_{kj}}$ and $f_{A_{ki}}$ are dissimilar and complementary for any $i$, $j$, and $i \neq j$ in $[1, g]$ [i.e., $R/S_{\{\mathrm{var}\}}(f_{A_{kj}}, f_{A_{ki}})$ is large].

**Step 5.3.** The combination method $\mathrm{A}_k^{(g)}$ of $g$ single scoring methods is the desired combination method for the receptor target and the compound set $D$.

**Step 6.** Output $A_k^{(g)}$, which is the desired combination method.

## ACKNOWLEDGMENT

J.-M.Y. was supported, in part, by Grant NSC-93-2113-M-009-010 from the National Science Council of Taiwan and by Grant VGHUST93-G5-05-3 from Veterans General Hospitals University System of Taiwan. B.S.K. is supported by NIH/NIA AG024232, the Hereditary Disease Foundation, and the State of New York. D.F.H. is supported, in part, by the Fordham University Faculty Fellowship, the DIMACS NSF Grant STC-1999, and the New Jersey Commission.

## REFERENCES AND NOTES

(1) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **2003**, *22*, 151−185.

(2) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Schoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213−2221.

(3) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7*, 1047−1055.

(4) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439−446.

(5) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.

(6) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(7) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.

(8) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317−324.

(9) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(10) 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*, 4759−67.

(11) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from

docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(12) Ng, K. B.; Kantor, P. B. Predicting the effectiveness of naive data fusion on the basis of system characteristics. *J. Am. Soc. Inf. Sci.* **2000**, *51*, 1177−1189.

(13) Belkin, N. J.; Kantor, P. B.; Fox, E. A.; Shaw, J. A. Combining evidence of multiple query representation for information retrieval. *Inf. Process. Manage.* **1995**, *31*, 431−448.

(14) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−85.

(15) Salim, N.; Holliday, J.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−42.

(16) Hsu, D. F.; Shapiro, J.; Taksa, I. Methods of data fusion in information retrieval: rank vs score combination. *DIMACS Technical Report* **2002**, *58*, 1−47.

(17) Hsu, D. F.; Taksa, I. Comparing rank and score combination methods for data fusion in information retrieval. *Inf. Retr.* **2005**, *8*, 449−480.

(18) 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*, 411−428.

(19) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the flexX incremental construction algorithm for protein−ligand docking. *Proteins: Struct., Funct., Bioinf.* **1999**, *37*, 228−241.

(20) 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*, 727−748.

(21) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein−ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.

(22) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−6.

(23) Chuang, H. Y.; Liu, H. F.; Chen, F. A.; Kao, C.-Y.; Hsu, D. F. In *Proceedings of the 7th International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN)* **2004**, 625−630.

(24) Yang, J.-M.; Chen, C.-C. GEMDOCK: a generic evolutionary method for molecular docking. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 288−304.

(25) van Lipzig, M. M.; ter Laak, A. M.; Jongejan, A.; Vermeulen, N. P.; Wamelink, M.; Geerke, D.; Meerman, J. H. Prediction of ligand binding affinity and orientation of xenoestrogens to the estrogen receptor by molecular dynamics simulations and the linear interaction energy method. *J. Med. Chem.* **2004**, *47*, 1018−1030.

(26) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucl. Acids Res.* **2000**, *28*, 235−242.

(27) Yang, J.-M.; Shen, T.-W.; Chen, Y.-F.; Chiu, Y.-Y. An evolutionary approach with pharmacophore-based scoring functions for virtual database screening. *Lect. Notes Comput. Sci.* **2004**, *3102*, 481−492.

(28) Yang, J.-M.; Shen, T.-W. A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 205−220.

(29) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: 1. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. -Aided Mol. Des.* **1997**, *11*, 425−445.

(30) Fisher, L. S.; Guner, O. F. Seeking novel leads through structure-based pharmacophore design. *J. Braz. Chem. Soc.* **2002**, *13*, 777−787.

(31) Hsu, D. F.; Palumbo, A. In *Proceedings of the 7th International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN)* **2004**, 557−562.

(32) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.

CI050034W