JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

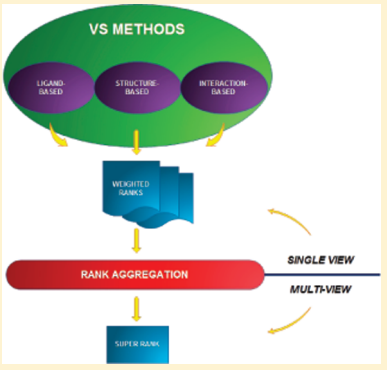
Virtual Drug Screen Schema Based on Multiview Similarity Integration and Ranking Aggregation

Hong Kang, Zhen Sheng, Ruixin Zhu, Qi Huang, Qi Liu,* and Zhiwei Cao*

School of Life Sciences and Technology, Tongji University, 200092, China

Supporting Information

ABSTRACT: The current drug virtual screen (VS) methods mainly include two categories. i.e., ligand/target structure-based virtual screen and that, utilizing protein−ligand interaction fingerprint information based on the large number of complex structures. Since the former one focuses on the one-side information while the later one focuses on the whole complex structure, they are thus complementary and can be boosted by each other. However, a common problem faced here is how to present a comprehensive understanding and evaluation of the various virtual screen results derived from various VS methods. Furthermore, there is still an urgent need for developing an efficient approach to fully integrate various VS methods from a comprehensive multiview perspective. In this study, our virtual screen schema based on multiview similarity integration and ranking aggregation was tested comprehensively with statistical evaluations, providing several novel and useful clues on how to perform drug VS from multiple heterogeneous data sources. (1) 18 complex structures of HIV-1 protease with ligands from the PDB were curated as a test data set and the VS was performed with five different drug representations. Ritonavir (1HXW) was selected as the query in VS and the weighted ranks of the query results were aggregated from multiple views through four similarity integration approaches. (2) Further, one of the ranking aggregation methods was used to integrate the similarity ranks calculated by gene ontology (GO) fingerprint and structural fingerprint on the data set from connectivity map, and two typical HDAC and HSP90 inhibitors were chosen as the queries. The results show that rank aggregation can enhance the result of similarity searching in VS when two or more descriptions are involved and provide a more reasonable similarity rank result. Our study shows that integrated VS based on multiple data fusion can achieve a remarkable better performance compared to that from individual ones and, thus, serves as a promising way for efficient drug screening, taking advantages of the rapidly accumulated molecule representations and heterogeneous data in the pharmacological area.



INTRODUCTION

It is well-known that virtual screening (VS) has played an important role in drug discovery and medicinal chemistry research in recent years.[1] "Automatically evaluating very large libraries of compounds using computer programs", which was defined by Walters et al. in the nineteen century,[2] has indicated the superiority and advantage of VS in filtering high throughput screening (HTS) data. On the one side, VS approaches are generally either ligand-based[3−7] with the concept of the similarity property principle[8] that states similar molecules may have similar biologic activities or target structure-based[3,5,9−12] which evaluates energetic and geometric criteria based on the knowledge of 3D target structure.[10−12] Among them, structure-based methods are often considered as the first choice when a target protein structure at high atomic resolution is available. Meanwhile, several other studies have pointed to the fact that ligand-based methods may offer a strong alternative to structure-based ones.[13] So far, more and more efforts have been made to develop various VS methods that fully integrate the ligand- and structure-based methods together, taking the advantages of their complementarities in nature to further boost the VS effect.[14−21] On the other side, a few recent works have also explored interaction-based VS methods based on available complex structures and interaction information[22] accumulated in the Protein Data Bank (PDB).[23]

With the increasing emerging VS methods, this gives rise to a new important issue to be addressed, i.e., how to integrate different VS results derived from different VS methods, since different results are obtained from different perspectives and it is hard to decide which one is better. This is particularly useful for HTS data analysis, since a comprehensive data integration model to aggregate all the VS results that are generated by multiple views, like the ligand-based, target structure-based, and interaction-based approaches and others, is expected to enhance single VS substantially. It has actually become a ubiquitous problem in most research areas: how to evaluate the same objects based on different evaluation systems.

Naturally as implicated by its name, multiview learning combines models from different aspects of one identical entity to obtain an overall and comprehensive representation. The concept of integration of different information sources has been

Received: October 9, 2011
Published: February 14, 2012

dx.doi.org/10.1021/ci200481c | J. Chem. Inf. Model. 2012, 52, 834−843

developed for years in the field of information retrieval.[24−26] Multiview clustering algorithms, as an unsupervised-learning method, can be divided into two categories in general:[27] (1) fusion of similarity data by deriving a convex combination of similarities from different views to minimize a given penalty error[28,29] and (2) fusion of clustering decision derived from each view separately.[30] Rank aggregation can be categorized to the second one, and it does not need complicated parameter selection, which makes it more efficient, repeatable, and applicable for high throughput data than those belonging to the previous category such as matrix fusion. More specifically, compared with the matrix fusion method which carries out similarity searching after aggregating individual matrixes from different views, rank aggregation is straightforward and easier to handle by simple integration of each single evaluation rather than complex matrix fusion. It should be noted that rank aggregation and data integration methods have already been applied in a few fields in bioinformatics and chemoinformatics. Several pioneer works have been carried, such as integrating the individual gene ranking list from various microarray studies[31,32] and evaluating the fusion clustering for miRNA targets predictions.[33,34] Recently, with the inspiring of iterative theory from rank fusion, Lange et al. provided an efficient way to combine similarity data originating from multiple sources for object grouping.[28] Such a method along with its idea may be extended to the area of pharmacology to help us to integrate the weighted ranks resulting from multiview similarity-based queries for drug VS.

In this study, four kinds of ranking aggregation (RA) methods were adopted for drug VS and were tested comprehensively with statistical evaluations. They are derived from two popular computational frameworks as well as two kinds of similarity measurements for data samples, i.e., the cross-entropy-based and genetic-algorithm-based computational frameworks, together with the data similarity represented in Spearman and Kendall's tau distance between ordered lists,[35] in total 4 combinations, denoted as cross-entropy Monte Carlo algorithm with Spearman's distance (CES), genetic algorithm with Spearman's distance (GAS), cross-entropy Monte Carlo algorithm with Kendall's tau distance (CEK), and genetic algorithm with Kendall's tau distance (GAK). Our experiment contains two parts: (1) Since more and more studies have been focused on HIV and its related data sets are widely accumulated, a commonly used HIV data set was used in our study,[36] which includes 18 crystal structures of HIV-1 protease with ligands from PDB. RA was tested based on different compound representation for drug query. (2) Furthermore, a novo type of descriptor for drugs: gene ontology (GO) fingerprint[37] was used in our study, which reduces the high dimensions and noises in microarray data. And such a descriptor was applied for drug description in a biological activity view. Then, an RA method was used to integrate the similarity ranks calculated by GO fingerprint and structural fingerprint on the data set from Connectivity Map.[38] Two typical inhibitors involved in Lamb's work were used as the queries to demonstrate our multiview representation ability in drug screening.

DATA AND METHODS

HIV-1 Protease Data Set. Several strict rules were set to screen the HIV-1 proteases from PDB (Protein Data Bank), which include crystal structure (1) with resolution <2.5 Å, (2) which is mutation-free, (3) with a clear K_i value, and (4) with representative K_i (to reduce the fluctuating errors resulting from different experimental platforms). With Molecular Operating Environment (MOE),[39] a diverse subset method based on MACCS keys and K_i values was used for the original data set with 229 complexes (Dec 27, 2011, at 4:00 p.m. PST). Eighteen representative samples remained as a golden testing data set for the first experiment. These data were also applied in our former study.[36] The data set contains the following protease complexes: 1AAQ,[40] 1AJV,[41] 1GNO,[42] 1HBV,[43] 1HIH,[44] 1HOS,[45] 1HPV,[46] 1HSG,[47] 1HVH,[48] 1HWR,[49] 1HXW,[50] 1ODY,[51] 1SBG,[52] 1XL2,[53] 1XL5,[53] 2AQU,[54] 2BPY,[55] and 7UPJ.[56] Among the 18 complexes, 4 ligands are known to be the marketed drugs against HIV. They are Ritonavir (1HXW), Indinavir (1HSG), Atazanavir (2AQU), and Amprenavir (1HPV). Ritonavir was used as a query in our VS study due to its good curative effect and the lower K_i value.

Connectivity Map Data Set. Lamb et al. had created the first installment of a reference collection of gene-expression profiles from cultured human cells simulated with bioactive small molecules, together with the pattern-matching algorithm to mine these data.[38] To date, this "Connectivity Map" contains approximately 7100 expression profiles representing 1309 compounds. In this study, gene ontology (GO) fingerprint,[37] a convenient representation to indicate the activity of compound, was generated to reduce the high dimensions and noises in these microarray data by means of the transformation from gene chip data to bits arrays through enrichment analysis of gene ontology. Such a descriptor was used for drug description in a biological activity view. The main computational procedures were summarized as follows:

(1) Select the functional modules of gene ontology.
(2) Calculate the expression differentiation among functional modules by Kolmogorov−Smirnov score.[57]
(3) Significance test functional modules.
(4) Generate GO fingerprint.
(5) Delete redundancy bits and rectification.

Multiview Representations for VS. Our multiview representations for VS are listed in the following:

Fingerprint. A ligand-based fingerprint was chosen in this drug representation, which is widely used in the high-speed structural screening as previously defined by Daylight.[58] Different from traditional structural keys in which each bit represents the presence (TRUE) or absence (FALSE) of a specific structural feature (pattern), the key of this method is splitting a molecule into small patterns which were served as a seed to a pseudorandom number generator. The output of each pattern is a set of bits (typically 4 or 5 bits per pattern) and the set of bits is combined together (with a logical OR) as the final fingerprint. This fingerprint, which is denoted as hashed fingerprint (h-FP) in our study, is a Boolean array of 1024 bits. The Tanimoto coefficient was used to measure the similarity for a given query to all the left samples. h-FP has several advantages over structural keys: (1). Since fingerprints have no predefined set of patterns, one fingerprinting system can easily serve all databases and all types of queries. (2) A fingerprint can save more memory than a structural key while keeping the comparable discriminating power. (3) The more complex a molecule is, the more accurate its fingerprint holds in its characterization of the molecule.

Energy-Based Scoring. The purpose of docking is to search favorable binding configurations between small to medium-sized ligands and a nonflexible macromolecular protein target. For each ligand, a number of configurations called poses are

generated and scored in an effort to determine favorable binding modes. Optionally, poses can be constrained to fit a pharmacophore query. The top scoring poses are written out to a database for further analysis. Usually, the most feasible poses are determined by considering the scores and binding modes simultaneously. In our study, the London dG scoring function integrated by MOE was used, which serves as a more robust energy-based scoring function for docking between small molecules and proteins. Such a score is also adopted to measure the similarity between a query and other samples by ranking the absolute value of their difference.[59] It should be noted that when the protein−ligand complex formed at a special conformation, the docking energy is unique in theory.

PLIF. PLIF is a residue-based interaction fingerprint calculated by MOE, distributed by Chemical Computing Group, Inc. This representation defines six types of intermolecular interactions: hydrogen bond with side chain donor, hydrogen bond with side chain acceptor, hydrogen bond with backbone donor, hydrogen bond with backbone acceptor, ionic attraction, and surface contact. For each residue, a weak interaction and a strong interaction of the six types are encoded by 12 bits to represent the presence or absence of the six types. When a query is chosen, a Tanimoto coefficient is also applied to measure the similarity between PLIFs.

RANK AGGREGATION

Two Philosophies and Two Distance Functions. There are two radically different philosophies on rank aggregation due to distinct definition of distance to measure the similarity between ordered lists. The first one is based on equalitarianism. It approaches to seek the consensus among individual rank ordered lists and is usually based on the form of rank averaging. Following this philosophical paradigm, the Spearman's footrule distance[60] between two ordered lists can be defined as

$$S(L_i, L_j) = \sum_{t \in L_i \cup L_j} |r^{L_i}(t) - r^{L_j}(t)| \tag{1}$$

Furthermore, the weighted Spearman's footrule distance between L_i and L_j is given by the following weighted sum representation

$$\mathrm{WS}(L_i, L_j) = \sum_{t \in L_i \cup L_j} |M(r^{L_i}(t)) - M(r^{L_j}(t))| \times |r^{L_i}(t) - r^{L_j}(t)| \tag{2}$$

An alternative to ranking aggregation is similar to voting, which is based upon majoritarian principles and tends to attach importance to the majority of individual preferences rather than relatively seldom ones. The Kendall's tau distance[61] is applied here for distance measuring between two ranked lists through majoritarian way. The Kendall's tau distance is defined as the summation of all the partial distances of every two elements of the list

$$K(L_i, L_j) = \sum_{t,u \in L_i \cup L_j} K_{tu}^{p} \tag{3}$$

where

$$K_{tu}^{p} = \begin{cases} 0 \quad \text{if} \quad r^{L_i}(t) < r^{L_i}(u),\ r^{L_j}(t) < r^{L_j}(u) | r^{L_i}(t) > r^{L_i}(u),\ r^{L_j}(t) > r^{L_j}(u) \\ 1 \quad \text{if} \quad r^{L_i}(t) > r^{L_i}(u),\ r^{L_j}(t) < r^{L_j}(u) | r^{L_i}(t) < r^{L_i}(u),\ r^{L_j}(t) > r^{L_j}(u) \\ p \quad \text{if} \quad r^{L_i}(t) = r^{L_i}(u) = k+1 | r^{L_j}(t) = r^{L_j}(u) = k+1 \end{cases} \tag{4}$$

Similar to that of Spearman distance, the weighted Kendall's tau distance is defined as

$$\mathrm{WK}(L_i, L_j) = \sum_{t,u \in L_i \cup L_j} (|M(r^{L_i}(t)) - M(r^{L_i}(u))| + |M(r^{L_j}(t)) - M(r^{L_j}(u))|) \times K_{tu}^{p} \tag{5}$$

It should be noted that calculations of the weighted Spearman or Kendall's tau distance will cause disproportionate contribution involved by disunity of scores' dimensions. To eliminate the deviation, a simple normalization which spread the scores of different ordered lists evenly between 0 and 1 is used. The new score after normalization is defined as

$$M_i^{*} = \frac{M_i - \min(M_i)}{\max(M_i) - \min(M_i)}, \qquad i = 1, \ldots, n \tag{6}$$

Optimization for Ranking Aggregation. To discover a superlist that would be simultaneously as close as possible to all the given ordered lists, an optimization function is defined as follows:

$$\delta^{*} = \arg\min \Phi(\delta) \tag{7}$$

$$\text{where} \qquad \Phi(\delta) = \sum_{i=1}^{m} w_i d(\delta, L_i) \tag{8}$$

w_i is the importance weight of ordered list L_i. Parameter d, which is defined by either Spearman or Kendall's tau distances, is the distance between "super list" δ^{*} and L_i. The goal of the ranking aggregation is to find δ^{*} which would minimize the total distance between the super list and every ordered list. In our study, two methods were adopted for combining the ordered lists: the cross-entropy method (CE)[33,62] and genetic algorithm (GA).[63,64]

Discounted Cumulative Gain. In our study, discounted cumulative gain (DCG)[65,66] is used to evaluate the ranking performance compared to the control, which is initially a measure of effectiveness of a web search engine algorithm. The premise of DCG is that highly relevant items appearing lower in the search result should be penalized as the graded relevance value is reduced logarithmically proportional to the position of

dx.doi.org/10.1021/ci200481c | J. Chem. Inf. Model. 2012, 52, 834−843

Table 1. Results for VS Using 1HXW as the Query[a]

rank	pK_i	h-FP	c-PLIF	rd-PLIF	c-ES	rd-ES
1	1HXW	1HXW(1)	1HXW(1)	1HXW(1)	1HXW(1)	1HXW(1)
2	1HSG	1SBG(12)	1HIH(10)	2AQU(3)	1HSG(2)	1AJV(13)
3	2AQU	2BPY(15)	7UPJ(6)	1ODY(9)	1AAQ(7)	1HOS(5)
4	1HPV	1ODY(9)	1HPV(4)	1HIH(10)	2BPY(15)	1ODY(9)
5	1HOS	1AAQ(7)	2AQU(3)	1HWR(8)	2AQU(3)	1AAQ(7)
6	7UPJ	2AQU(3)	1GNO(14)	1HPV(4)	1ODY(9)	1XL2(16)
7	1AAQ	1HWR(8)	1XL5(17)	1AJV(13)	1AJV(13)	1HVH(11)
8	1HWR	1HSG(2)	1SBG(12)	1HSG(2)	1HVH(11)	1GNO(14)
9	1ODY	1HIH(10)	1ODY(9)	1HBV(18)	1HBV(18)	2AQU(3)
10	1HIH	1HVH(11)	1AAQ(7)	1AAQ(7)	1XL5(17)	1XL5(17)
11	1HVH	1HOS(5)	2BPY(15)	1GNO(14)	1GNO(14)	1HIH(10)
12	1SBG	1XL5(17)	1HSG(2)	2BPY(15)	1HOS(5)	2BPY(15)
13	1AJV	1HBV(18)	1HWR(8)	1XL2(16)	1HIH(10)	1HPV(4)
14	1GNO	1HPV(4)	1XL2(16)	1HVH(11)	1XL2(16)	1HBV(18)
15	2BPY	1GNO(14)	1AJV(13)	7UPJ(6)	1SBG(12)	1SBG(12)
16	1XL2	7UPJ(6)	1HOS(5)	1XL5(17)	7UPJ(6)	7UPJ(6)
17	1XL5	1XL2(16)	1HVH(11)	1SBG(12)	1HPV(4)	1HSG(2)
18	1HBV	1AJV(13)	1HBV(18)	1HOS(5)	1HWR(8)	1HWR(8)
nDCG scores	**1.00**	**.8299**	**.8972**	**.9158**	**.9068**	**.8450**

[a]Using 1HXW as a query, 18 HIV-1 proteases with ligand were ranked by 5 VS methods and the rank order with pK_i was taken as the control, and the rank values in control were shown in the brackets. Also the nDCGs were calculated.

the result. The discounted CG accumulated at a particular rank position p is defined as

$$\mathrm{DCG}_p = \mathrm{rel}_1 + \sum_{i=2}^{p} \frac{\mathrm{rel}_i}{\log_2 i} \tag{9}$$

where rel_i is the graded relevance of the result at position i.

Comparing the similarity to the ideal list cannot be consistently achieved using DCG alone, due to the fact that different lists vary in length depending on the query. Therefore, the cumulative gain of each list is normalized by producing an ideal DCG. For a given query, the normalized discounted cumulative gain, or nDCG, is computed as

$$\mathrm{nDCG}_p = \frac{\mathrm{DCG}_p}{\mathrm{IDCG}_p} \tag{10}$$

The nDCG values for all lists can be averaged to obtain a measure of the average performance of one of the ranking algorithms.

RESULTS AND DISCUSSION

Experiment 1: Evaluation for Ranking Aggregation of VS. *Individual Ranking List.* With Ritonavir (1HXW) as the query, the energy-based scores and PLIF for both crystal structures and redocked structures by separating the ligands and proteins were calculated respectively for the 18 complex structures of HIV-1 protease. Therefore, in total there are five weighted ordered lists to be aggregated for the drug VS: ligand-based hashed fingerprints (h-FP), crystal structures' energy-based scores (c-ES), crystal structure-based PLIFs (c-PLIF), redocked structures' energy-based scores (rd-ES), and redocked structure-based PLIFs (rd-PLIF). The same parameters were used on all complexes for every method. It should be noted that the redocked structure-based methodology was only used for the cases when the crystal structures of the complex were not available.

Table 1 presents the ranking result for five different VS methods (h-FP, c-PLIF, rd-PLIF, c-ES, and rd-ES) with 1HXW as a query. The pK_i value calculated from K_i is used as a control to represent the real ranking result. For each measurement, 18 HIV-1 proteases with ligand can be ranked based on the scores which are sorted either in ascending or descending order, depending on whether a larger or smaller score corresponds to a better performance or not. In our study, h-FP and PLIF present higher scores with better performances while a smaller score is desired for ES to indicate better performance. The scores of all measurements and the control value are listed in the Supporting Information (Table S1). To evaluate the ranking result, for each of them the nDCG value was calculated compared to the control pK_i value. The nDCG value of control ranking is 1. For other ranking lists, the larger value indicates a better performance.

Table 1 gives us several insights on the performance of individual VS method. It indicates that (1) h-FP has obtained the worst score with the nDCG value of 0.8299. This is not surprising, since the control pK_i is the measurement of binding affinity of the inhibitor, while h-FP is a ligand-based fingerprint which is only generated by the chemical construction of the drugs. Thus compared with the control pK_i, h-FP may be not efficiently representative. (2) The performance of PLIF is relative higher, regardless of crystal structures-based (c-PLIF) or redocked ones (rd-PLIF). Such results suggest that the PLIF may serve as a well-defined stable VS method. (3) A large margin accuracy gap is shown between c-ES and rd-ES. The reason is speculated that, in molecular docking, there's no perfect solution to validate the conformation sampling and scoring for drugs, which induces the instability of sorting the docking result directly.

Aggregated Rank Lists. It should be noted that it is not trivial to select the general best similarity searching method for VS here. However, rank aggregation is extremely helpful in reconciling the onesidedness from individual methods and generating a well-defined super list to rank all the elements based on their performance as determined by multiview measurements

dx.doi.org/10.1021/ci200481c | J. Chem. Inf. Model. 2012, 52, 834−843

simultaneously (Figure 1). In our study, four ranking aggregation methods (CES, GAS, CEK, and GAK) were evaluated to investigate their integration abilities for VS[35]

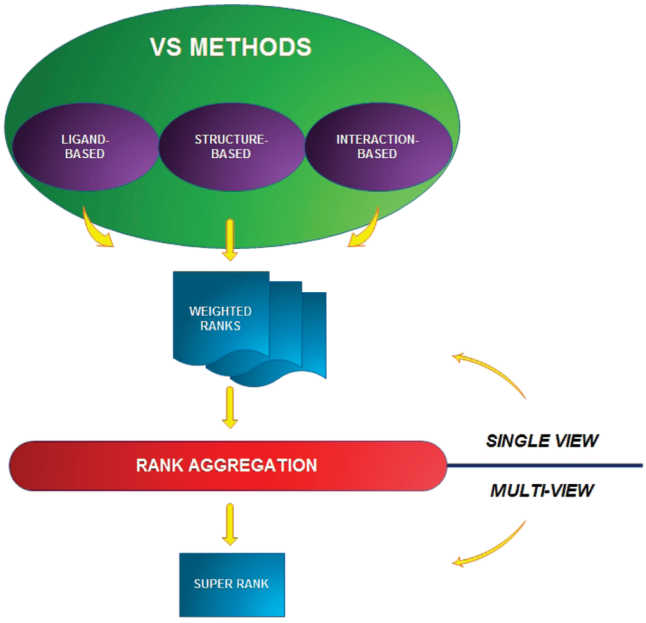


Figure 1. Concept of multiview learning in VS. Rank aggregation was used to shift the perspective from single view to multiview.

As an example, the five rankings were aggregated by CES. In this case, the final rank list was ordered as 1HXW 1ODY 1AAQ 1AJV 2AQU 1HIH 2BPY 1HVH 1GNO 1HOS 1XL2 1XL5 1HSG 1HPV 1SBG 1HBV 7UPJ 1HWR with the minimum total distance of 11.399. It is interesting to see that although 1ODY does not obtain the best place in the individual ranking lists, it finally achieves the best performance after ranking aggregation, which indicates that the aggregation can boost the VS result which may have failed by individual searching. A visual representation of aggregation results was given in the Supporting Information (Figure S1).

Similar, the genetic algorithm (GA) can also be used with either of the two distances. Figure S2 provides the visual representation of rank aggregation using the GA algorithm with the weighted Spearman distance. The top five of final rank is 1ODY 1AJV 1AAQ 2AQU 1HIH, which is slightly different from that of CES in the order of 1AJV and 1AAQ. In addition, it can be seen that the GA algorithm takes a significantly larger amount of cycles to converge but less time to operate. Detailed comparisons of different aggregation methods will be illustrated later.

To assess the effect of the aggregations, all reasonable combinations of the noted VS methods from three different categories were integrated by CES, GAS, CEK, and GAK. As the five methods belong to three different categories, namely ligand-based (h-FP), energy-based (ES), or interaction-based (PLIF), only that from a different class could form a meaningful combination to be aggregated. Different combinations and corresponding nDCG scores of each final ranking list were given in Figure 2a. Figure 2a is highlighted in the following color scheme: scores with red suggest that they are higher than each member of the combination. Scores with blue suggest that they are higher than one or some of the members' scores but lower than that of the others. The black ones are the worst cases in that the aggregation score is lower than that of all the individuals.

As shown in Figure 2, several useful clues may be identified for VS integration: (1) combinations with rd-ES obtained worse performance through all the methods, which further proves the instability and low-quality of rd-ES in drug VS. (2) If the nDCG scores of individual ranks are close, the super list after aggregating tends to be better. (3) It is apparent that the more members that participate in the aggregation, the better the effect of fusion obtained (the combination with rd-ES must not be counted). This may indicate that the aggregated VS result tends to be better when more VS methods are involved. (4) Noted that in most cases of VS research, crystal structures are rarely provided. Consequently, c-PLIF and c-ES are hard to obtain in most of the similarity searching scenarios. However, our study has indicated that the rd-PLIF is served as an equivalent substitution to c-PLIF with generally similar VS results, thus it is recommended to directly aggregate the ranks of h-FP and rd-PLIF to get a satisfied super list when there's no crystal structures data available for VS.

Methodology Comparison. In order to compare the four ranking aggregation methods (CES, GAS, CEK, and GAK) comprehensively, several important features in the aggregation like nDCG scores, minimum values, and consuming time were evaluated statistically. Table 2 provides a general view of the evaluation results.

In our study, the paired-samples T test was used to evaluate the CE and GA methods via comparing their minimum values of the optimal objectives. In total, the four ranking aggregations were compared in two pairs: CES−GAS and CEK−GAK. The minimum distance values of the four methods in all combinations and the analyzing result of paired-samples T test is present in Supporting Information Tables S2 and S3, in which it is indicated that the minimum values are significantly different for CES−GAS ($t = -7.084$, sig < 0.01), while they are less significant for CEK−GAK ($t = -2.432$, 0.01 < sig < 0.05). Generally with the same distance, the less the minimum value is, the better the aggregation. Therefore, we conclude that the cross-entropy method performs better than genetic algorithm in VS when using minimum distance values as an evaluation criterion.

A cross classification analysis of variance (ANOVA) is applied to compare the four methods (CES, CEK, GAS, and GAK) by using "combinations" and "methods" as the fixed factors, with "nDCG values" as the dependent variable. In this univariate situation, we are only concerned with the main effects to reduce the random error. The significance level is set to 0.05. For the "corrected model", the value of d*f* and *sum of squares* are the sum of those from fixed factors with $F = 23.402$ and $P < 0.01$. This indicates that the analytical model has statistical significance. For the fixed factor method, $F = 0.844$ and P (sig) = 0.478 > 0.05, which demonstrates that there is no significant difference among different methods by comparing the nDCG scores. Meanwhile, the P value of factor combination is extremely smaller than 0.01, which means that the difference among different combinations is highly significant (Supporting Information Table S4). In summary, the following can be concluded: different combinations influence the effect of aggregation much more than the that of the method itself. Furthermore, to analyze the difference of each method, the LSD multiple comparisons were adopted (Supporting Information Table S5) and no significant difference was obtained.

(a)

		CES	GAS	CEK	GAK
Combinations	h-FP c-PLIF	.8968	.9009	.9002	.9045
	h-FP rd-PLIF	.9184	.9191	.9283	.9178
	h-FP c-ES	.9026	.9249	.8826	.9076
	h-FP rd-ES	.8424	.8555	.8656	.8584
	c-PLIF c-ES	.9488	.9455	.9101	.9201
	rd-PLIF c-ES	.9399	.9371	.9451	.9192
	c-PLIF rd-ES	.8473	.8463	.8574	.8494
	rd-PLIF rd-ES	.8519	.8530	.8424	.8414
	h-FP c-PLIF c-ES	.9180	.9303	.9113	.9186
	h-FP rd-PLIF c-ES	.9459	.9405	.9155	.9181
	h-FP c-PLIF rd-ES	.8491	.8740	.8696	.8745
	h-FP rd-PLIF rd-ES	.8539	.8527	.8954	.8753
	No rd-ES	.9217	.9229	.9224	.9261
	All	.8645	.8954	.8937	.8901

(b)

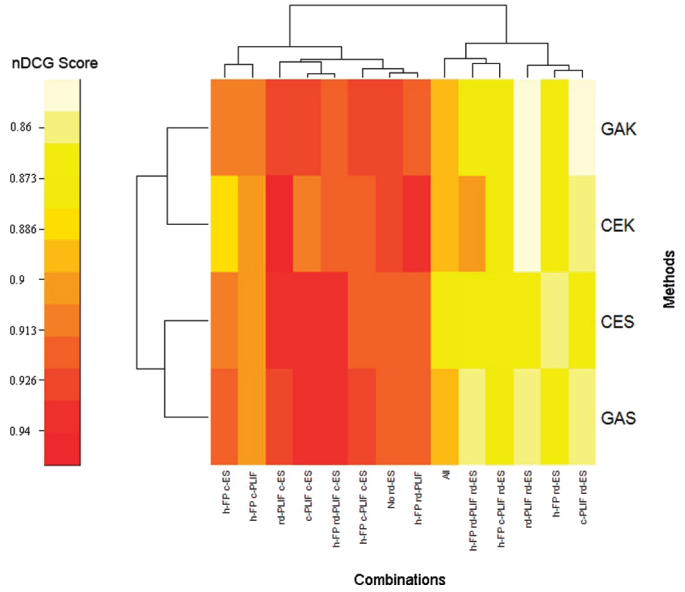


Figure 2. Fusion results for VS using rank aggregation. (a) Fusion evaluation by nDCG. nDCG scores were calculated for every super list fused by different combinations and rank aggregation methods. If the score is higher than each member of the combination, it is highlighted by a red color. Scores with blue suggested that they are higher than one or some of the members' scores but lower than that of the others. The black ones are the worst cases that the aggregation score is lower than that of all the individuals. (b) Heat map of the performance of rank aggregation measured by nDCG scores.

Table 2. General Result for Comparison of the Four Ranking Aggregation Methods[a]

evaluated feature	statistical method	which one is better?	sig (*P*-value)
minimum distance	paired *T* test	CES and GAS	0.000 and 0.030
nDCG score	ANOVA	no difference	0.478
consuming time	ANOVA	CES	0.000

[a]The four rank aggregation methods were compared by evaluating three important features statistically.

The same statistic approach was used to compare the four integration methods by investigating their consuming time. The analysis indicated that there is extreme significance of method ($P < 0.01$) and common significance of combination ($0.01 < P < 0.05$). Obviously, the statistical differences among these methods are significant in terms of time efficiency. With the average consuming time of 23.62 s, CES seems to be the most preferable method among others (Supporting Information Table S6 and S7).

In summary, it is statistically proven that: (1) CE performs better than GA in the way of optimization; (2) Spearman distance is more time efficient than that of Kendall. On the basis of such conclusions, the cross-entropy algorithm with Spearman distance (CES) is the most preferable choice of rank aggregation for drug VS.

Since the first experiment was designed to comprehensive evaluate the applicability of rank aggregation, the data set was expected to be available for all the given virtual screening methods as we mentioned in the method part (fingerprint, energy-based scoring and PLIF). Therefore, we set several strict rules to screen the collected data, and 18 representative samples remained. Indeed, more testing data and independent data sets would be helpful to test the robustness of the mixture model, but this is difficult to find. We had tried to search other families

dx.doi.org/10.1021/ci200481c | J. Chem. Inf. Model. 2012, 52, 834–843

(e.g., CDK), but only a few proteases with both crystal structural and K_i value were available. However, fewer but curated samples could help us to evaluate and compare the parameters in rank aggregation more efficiently, especially when the tests consumed a large amount of time. In order to test the robustness of rank aggregation, the high throughput C-map data was used to make a further evaluation in the second experiment.

Experiment 2: Further Evaluation on HTS Data using CES. To further evaluate the performance of RA, we tried to integrate gene expression data and structural descriptors of compounds in virtual drug screen. In this study, approximately 7100 expression profiles representing 1309 compounds from the Connectivity Map[38] were compiled as a data set, and both GO fingerprint[37] and structural fingerprint were calculated for every single compound. Two typical HDAC and HSP90 inhibitors which involved in Lamb's work were chosen as the queries to test our experiment. For each query, the ranks of similarity searching derived by the two kinds of fingerprints were aggregated by CES method, which has been demonstrated previously as the most preferable choice. Details follow here.

First, trichostatin A (TSA), a typical HDAC inhibitor,[67] was used as the query of similarity searching based on GO fingerprint and structural fingerprint respectively. By comparing the ranks before and after the aggregation, vorinostat and scriptaid, two strong HDAC inhibitors[68,69] were successfully retrieved in the top 3% (rank 17 and 31 of the 1309 compounds) after RA, which achieved a relatively good result compared to that using a structure descriptor only, as shown in Table 3. Meanwhile, cobalt chloride, which is apparently not the HDAC inhibitor,[70] and another valproic acid, which is an HDAC inhibitor with poor specificity, were successfully pushed down in the list (rank 939 and 425) compared to that using GO fingerprint. Similarly, another HSP90 inhibitor geldanamycin[71] was also used as the query only. As a result, alvespimycin and tanespimycin, both the derivants of geldanamycin, which are also typical HSP90 inhibitors,[72,73] were ranked in first and third place after RA. In addition, monorden which is another common HSP90 inhibitor[74] was successfully ranked in the top 3% (rank 22). On the other side, sirolimus, accepted not as the HSP90 inhibitor,[75,76] was ranked 367th, a relatively lower rank compared to that found using GO fingerprint only.

Table 3. Performance of RA in HDAC and HSP90 Inhibitors' Issue[a]

compound	rank (GO)	rank (structural)	rank (RA)	inhibitors or not	IC_{50} (μM)
HDAC inhibitor					
trichostatin A	query	query	query	yes	0.1−0.3[77,78]
vorinostat	1	82	17	yes	<0.086[79]
scriptaid	2	293	31	yes	0.6−1[78]
valproic acid	53	1170	425	no	N/A
cobalt chloride	35	1266	939	no	N/A
HSP90 inhibitor					
geldanamycin	query	query	query	yes	4.2[80]
tanespimycin	1	1	3	yes	3.5[80]
alvespimycin	3	2	1	yes	0.024[81]
monorden	8	123	22	yes	0.02−0.2[81−84]
sirolimus	452	4	367	no	N/A

[a]Approximately 7100 expression profiles representing 1309 compounds from Connectivity Map were involved as a data set. Several compounds reported as inhibitors or not were chosen to show and evaluate the fusion effect. In addition, the IC_{50} of every inhibitor was listed for reference.

Conclusions from Table 3 include (1) RA made the final ranks more reasonable when there was a divergence between the two evaluation systems (vorinostat, scriptaid, valproic acid, and cobalt chloride in HDAC's case; monorden and sirolimus in HSP90's case). (2) RA can preserve or even enhance the ranks of compounds which were identified in agreement by both of the two systems and unify the result when there is a disagreement.

The purpose of the second experiment was to reduce the unilateral prejudice by combining the structural and gene expression information using rank aggregation. Our result indicated that both the expected inhibitors and noninhibitors were ranked in much more reasonable positions after fusion than those from individual views. It should be noted that in the common virtual screening methods, only the top 10% candidates in the ranking list are focused, and taken as potential targets with more likelihood. Within this threshold, the exact ranking position for a specific sample becomes meaningless since all the samples within this threshold can be viewed in a relatively equal significance for virtual screening. From this point of view, our ranking result is convincing and helpful, since it can be seen in Table 3 that all the known inhibitors were successfully ranked into the top 3% in the aggregated ranking list, and at the same time, all the known noninhibitors were pushed out lower than the top 30%. Such a result indicated that the list produced by rank aggregation had a great classification ability to distinguish inhibitors between noninhibitors, while this successful discrimination could not be achieved by using a single view (structural fingerprint or GO fingerprint) individually. In addition, the C-Map data has its great practicability to become an objective virtual screening benchmark data for its advantages as free downloading and regular updating. The fast and standard high throughput technologies would provide increasing assistance in drug screening fields.

Noted that when the results derived from different methods were aggregated based on a multiview concept, a better result would be produced if there are complementary between or among the methods. For example, the structure and activity information of a given compound is complementary. On the basis of this idea, we designed the second experiment and obtained better results with rank aggregation. However, although more members participating in the aggregation creates a better effect of the fusion obtained in theory, if we simply aggregate the results derived from different methods that belong to the same category, the result might not be advanced or even be worse since the aggregated information may be redundant.

In summary, all of the experiments indicated the reliability and applicability of rank aggregation on the small samples and high throughput screening data. Rank aggregation was recommended in virtual screening, since it can aggregate the advantages of different VS methods and produce a comprehensive result which is more close to the ground truth.

CONCLUSIONS

It is proven that rank aggregation obtains remarkably better performance than the individual ones in VS. Furthermore, by comparing of h-FP and rd-ES, it is indicated that the more reliable and complementary the individual rank, the higher quality the fused rank obtained. In order to measure the similarity in VS comprehensively, multiview rank aggregation based on more VS methods is essential. In summary, rank aggregation is an effective

and convenient method to integrate various measurements for drug VS.

Moreover, a comparison of different integration methods is performed by concerning several key features in aggregation. There is no evident difference in the quality of aggregation between CE and GA if their individual parameters (in particular the sample size N for CE and the crossover and mutation probabilities for GA) are set appropriately. However, CE presents higher stability than GA, with Spearman distance running more efficiently than Kendall's tau distance. Our study finally indicates that the cross-entropy Monte Carlo algorithm with Spearman distance (CES) will be applied preferably if there are no special requirements for large-scale drug VS integration. Further, CES was used in HDAC and HSP90 inhibitors screening and showed a remarkablly robust ranking performance.

In summary, our study indicates that integrated VS based on multiple data fusion can achieve a much better performance compared to that from individual ones and, thus, serves as a promising ranking way for efficient drug screening, taking advantage of the rapidly accumulated molecule representations and heterogeneous data in the pharmacological area.

■ ASSOCIATED CONTENT

Supporting Information

Figures S1 and S2 and Tables S1–S7. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: qiliu@tongji.edu.cn. Tel.: +8602165980296. Fax: +8602165980296 (Q.L.). E-mail: zwcao@tongji.edu.cn. Tel.: +8602165980296. Fax: +860216598029 (Z.C.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by grants from Ministry of Science and Technology China (2010CB833601), National Natural Science Foundation of China (31171272, 31100956, and 61173117), Research Fund for the Doctoral Program of Higher Education of China (20100072110008, 20110072120048), and Shanghai Pujiang talent funding (11PJ1407400).

■ REFERENCES

(1) Stahl, M.; Guba, W.; Kansy, M. Integrating molecular design resources within modern drug discovery research: the Roche experience. *Drug Discoveryery Today* **2006**, *11* (7–8), 326–33.

(2) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening--an overview. *Drug Discoveryery Today* **1998**, *3* (4), 160–178.

(3) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discoveryery* **2002**, *1* (11), 882–94.

(4) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discoveryery Today* **2007**, *12* (5–6), 225–33.

(5) Jain, A. N. Virtual screening in lead discovery and optimization. *Curr. Opin. Drug Discovery Dev.* **2004**, *7* (4), 396–403.

(6) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Design* **2005**, *11* (9), 1189–202.

(7) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11* (23–24), 1046–53.

(8) Varnek, A.; Tropsha, A. *Chemoinformatics Approaches To Virtual Screening*; Royal Society Chemistry: London, 2008; p 125.

(9) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47* (4), 409–43.

(10) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3* (11), 935–49.

(11) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432* (7019), 862–5.

(12) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11* (13–14), 580–94.

(13) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903–911.

(14) Bissantz, C.; Schalon, C.; Guba, W.; Stahl, M. Focused library design in GPCR projects on the example of 5-HT(2c) agonists: comparison of structure-based virtual screening with ligand-based search methods. *Proteins* **2005**, *61* (4), 938–52.

(15) Wei, D. Q.; Zhang, R.; Du, Q. S.; Gao, W. N.; Li, Y.; Gao, H.; Wang, S. Q.; Zhang, X.; Li, A. X.; Sirois, S.; Chou, K. C. Anti-SARS drug screening by molecular docking. *Amino Acids* **2006**, *31* (1), 73–80.

(16) Lin, T. W.; Melgar, M. M.; Kurth, D.; Swamidass, S. J.; Purdon, J.; Tseng, T.; Gago, G.; Baldi, P.; Gramajo, H.; Tsai, S. C. Structure-based inhibitor design of AccD5, an essential acyl-CoA carboxylase carboxyltransferase domain of Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. USA* **2006**, *103* (9), 3072–7.

(17) Vidal, D.; Thormann, M.; Pons, M. A novel search engine for virtual screening of very large databases. *J. Chem. Inf. Model.* **2006**, *46* (2), 836–43.

(18) Barreiro, G.; Guimaraes, C. R.; Tubert-Brohman, I.; Lyons, T. M.; Tirado-Rives, J.; Jorgensen, W. L. Search for non-nucleoside inhibitors of HIV-1 reverse transcriptase using chemical similarity, molecular docking, and MM-GB/SA scoring. *J. Chem. Inf. Model.* **2007**, *47* (6), 2416–28.

(19) Tikhonova, I. G.; Sum, C. S.; Neumann, S.; Engel, S.; Raaka, B. M.; Costanzi, S.; Gershengorn, M. C. Discoveryery of novel agonists and antagonists of the free fatty acid receptor 1 (FFAR1) using virtual screening. *J. Med. Chem.* **2008**, *51* (3), 625–33.

(20) Tan, L.; Geppert, H.; Sisay, M. T.; Gutschow, M.; Bajorath, J. Integrating structure- and ligand-based virtual screening: comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets. *ChemMedChem* **2008**, *3* (10), 1566–71.

(21) Svensson, F.; Karlen, A.; Skold, C. Virtual Screening Data Fusion Using Both Structure- and Ligand-Based Methods. *J. Chem. Inf. Model.* **2011**, DOI: 10.1021/ci2004835.

(22) Tan, L.; Batista, J.; Bajorath, J. Computational methodologies for compound database searching that utilize experimental protein-ligand interaction information. *Chem. Biol. Drug Design* **2010**, *76* (3), 191–200.

(23) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. http://www.rcsb.org/pdb/home/home.do (accessed Nov 28th, 2011).

(24) Ghani, R. Combining labeled and unlabeled data for multiclass text categorization. *Proceedings of the nineteenth international conference on machine learning*, The University of New South Wales, Sydney, Australia, July 8–12, 2002, Morgan Kaufmann Publishers Inc.: 2002; pp 187–194.

(25) Brefeld, U.; Scheffer, T. Co-EM support vector learning. *Proceedings of the twenty-first international conference on machine learning*; Banff, Alberta, Canada, July 4–8, 2004 ACM: Banff, Alberta, Canada, 2004; p 16.

(26) Zhou, Z.; Li, M. Semi-supervised regression with co-training. *Proceedings of the International Joint Conference on artificial intelligence*, Edinburgh, Scotland, UK, July 30–August 5, 2005.

(27) Bruno, E.; Marchand-Maillet, S. Multiview clustering: A late fusion approach using latent models. *Proceedings 32nd annual international ACM SIGIR conference on research and development in*

information retrieval; Boston, MA, USA, July 19−23, 2009; pp 736−737 and 870.
(28) Lange, T.; Buhmann, J. M. Fusion of Similarity Data in Clustering. *Adv. Neural Inf. Process. Syst.* **2006**, *18*, 723−730.
(29) Long, B.; Yu, P. S.; Zhang, Z. M. A general model for multiple view unsupervised learning. *Proceedings of the 8th SIAM international conference on data mining*, Atlanta, Georgia, USA, April 24−26, 2008; pp 822−833.
(30) Bickel, S.; Scheffer, T. Multi-view clustering. *Proceeding of IEEE data mining conference*, Brighton, UK, November 1−4, 2004; pp 19−26.
(31) DeConde, R. P.; Hawley, S.; Falcon, S.; Clegg, N.; Knudsen, B.; Etzioni, R. Combining results of microarray experiments: a rank aggregation approach. *Stat. appl. Genetics Molec. Biol.* **2006**, *5*, No. article 15.
(32) Pihur, V.; Datta, S.; Datta, S. Finding common genes in multiple cancer types through meta-analysis of microarray experiments: a rank aggregation approach. *Genomics* **2008**, *92* (6), 400−3.
(33) Lin, S.; Ding, J.; Zhou, J. Rank Aggregation of putative microRNA targets with Cross-Entropy Monte Carlo Methods. *Preprint, Presented at the IBC 2006 conference*, Montreal, Quebec, Canada, July 16−21, 2006.
(34) Lin, S.; Ding, J. Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA Studies. *Biometrics* **2009**, *65* (1), 9−18.
(35) Pihur, V.; Datta, S.; Datta, S. RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics* **2009**, *10*, 62.
(36) Huang, Q.; Kang, H.; Zhang, D.; Sheng, Z.; Liu, Q.; Zhu, R.; Cao, Z. Comparison of Ligand-, Target Structure-, and Protein-Ligand Interaction Fingerprint-based Virtual Screening Methods. *Acta Chim. Sinica* **2011**, *69* (5), 515−522.
(37) Sheng, Z.; Huang, Q.; Kang, H.; Liu, Q.; Cao, Z.; Zhu, R. A New Fingerprint of Chemical Compounds and Its Application to Drugs Virtual Screening. *Acta Chim. Sinica* **2011**, *69* (16), 1845−1850.
(38) Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J. P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **2006**, *313* (5795), 1929−35.
(39) C.C.G., Inc. *Molecular Operation Environment*, 2008.10; Montreal, Quebec, Canada, 2008.
(40) Dreyer, G. B.; Lambert, D. M.; Meek, T. D.; Carr, T. J.; Tomaszek, T. A. Jr.; Fernandez, A. V.; Bartus, H.; Cacciavillani, E.; Hassell, A. M.; Minnich, M.; et al. Hydroxyethylene isostere inhibitors of human immunodeficiency virus-1 protease: structure-activity analysis using enzyme kinetics, X-ray crystallography, and infected T-cell assays. *Biochemistry* **1992**, *31* (29), 6646−59.
(41) Backbro, K.; Lowgren, S.; Osterlund, K.; Atepo, J.; Unge, T.; Hulten, J.; Bonham, N. M.; Schaal, W.; Karlen, A.; Hallberg, A. Unexpected binding mode of a cyclic sulfamide HIV-1 protease inhibitor. *J. Med. Chem.* **1997**, *40* (6), 898−902.
(42) Hong, L.; Treharne, A.; Hartsuck, J. A.; Foundling, S.; Tang, J. Crystal structures of complexes of a peptidic inhibitor with wild-type and two mutant HIV-1 proteases. *Biochemistry* **1996**, *35* (33), 10627−33.
(43) Hoog, S. S.; Zhao, B.; Winborne, E.; Fisher, S.; Green, D. W.; DesJarlais, R. L.; Newlander, K. A.; Callahan, J. F.; Moore, M. L.; Huffman, W. F.; et al. A check on rational drug design: crystal structure of a complex of human immunodeficiency virus type 1 protease with a novel gamma-turn mimetic inhibitor. *J. Med. Chem.* **1995**, *38* (17), 3246−52.
(44) Priestle, J. P.; Fassler, A.; Rosel, J.; Tintelnot-Blomley, M.; Strop, P.; Grutter, M. G. Comparative analysis of the X-ray structures of HIV-1 and HIV-2 proteases in complex with CGP 53820, a novel pseudosymmetric inhibitor. *Structure* **1995**, *3* (4), 381−9.
(45) Abdel-Meguid, S. S.; Zhao, B.; Murthy, K. H.; Winborne, E.; Choi, J. K.; DesJarlais, R. L.; Minnich, M. D.; Culp, J. S.; Debouck, C.; Tomaszek, T. A. Jr.; et al. Inhibition of human immunodeficiency virus-1 protease by a C2-symmetric phosphinate. Synthesis and crystallographic analysis. *Biochemistry* **1993**, *32* (31), 7972−80.
(46) Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rao, B. G.; Tung, R. D.; Navia, M. A. Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *J. Am. Chem. Soc.* **1995**, *117* (3), 1181−1182.
(47) Chen, Z.; Li, Y.; Chen, E.; Hall, D. L.; Darke, P. L.; Culberson, C.; Shafer, J. A.; Kuo, L. C. Crystal structure at 1.9-A resolution of human immunodeficiency virus (HIV) II protease complexed with L-735,524, an orally bioavailable inhibitor of the HIV proteases. *J. Biol. Chem.* **1994**, *269* (42), 26344−26348.
(48) Jadhav, P. K.; Woerner, F. J.; Lam, P. Y.; Hodge, C. N.; Eyermann, C. J.; Man, H. W.; Daneker, W. F.; Bacheler, L. T.; Rayner, M. M.; Meek, J. L.; Erickson-Viitanen, S.; Jackson, D. A.; Calabrese, J. C.; Schadt, M.; Chang, C. H. Nonpeptide cyclic cyanoguanidines as HIV-1 protease inhibitors: synthesis, structure-activity relationships, and X-ray crystal structure studies. *J. Med. Chem.* **1998**, *41* (9), 1446−55.
(49) Ala, P. J.; DeLoskey, R. J.; Huston, E. E.; Jadhav, P. K.; Lam, P. Y.; Eyermann, C. J.; Hodge, C. N.; Schadt, M. C.; Lewandowski, F. A.; Weber, P. C.; McCabe, D. D.; Duke, J. L.; Chang, C. H. Molecular recognition of cyclic urea HIV-1 protease inhibitors. *J. Biol. Chem.* **1998**, *273* (20), 12325−31.
(50) Kempf, D. J.; Marsh, K. C.; Denissen, J. F.; McDonald, E.; Vasavanonda, S.; Flentge, C. A.; Green, B. E.; Fino, L.; Park, C. H.; Kong, X. P.; et al. ABT-538 is a potent inhibitor of human immunodeficiency virus protease and has high oral bioavailability in humans. *Proc. Natl. Acad. Sci. USA* **1995**, *92* (7), 2484−8.
(51) Kervinen, J.; Lubkowski, J.; Zdanov, A.; Bhatt, D.; Dunn, B. M.; Hui, K. Y.; Powell, D. J.; Kay, J.; Wlodawer, A.; Gustchina, A. Toward a universal inhibitor of retroviral proteases: comparative analysis of the interactions of LP-130 complexed with proteases from HIV-1, FIV, and EIAV. *Protein Sci.* **1998**, *7* (11), 2314−23.
(52) Abdel-Meguid, S. S.; Metcalf, B. W.; Carr, T. J.; Demarsh, P.; DesJarlais, R. L.; Fisher, S.; Green, D. W.; Ivanoff, L.; Lambert, D. M.; Murthy, K. H.; et al. An orally bioavailable HIV-1 protease inhibitor containing an imidazole-derived peptide bond replacement: crystallographic and pharmacokinetic analysis. *Biochemistry* **1994**, *33* (39), 11671−7.
(53) Specker, E.; Bottcher, J.; Lilie, H.; Heine, A.; Schoop, A.; Muller, G.; Griebenow, N.; Klebe, G. An old target revisited: two new privileged skeletons and an unexpected binding mode for HIV-protease inhibitors. *Angew. Chem., Int. Ed.* **2005**, *44* (20), 3140−4.
(54) Clemente, J. C.; Coman, R. M.; Thiaville, M. M.; Janka, L. K.; Jeung, J. A.; Nukoolkarn, S.; Govindasamy, L.; Agbandje-McKenna, M.; McKenna, R.; Leelamanit, W.; Goodenow, M. M.; Dunn, B. M. Analysis of HIV-1 CRF_01 A/E protease inhibitor resistance: structural determinants for maintaining sensitivity and developing resistance to atazanavir. *Biochemistry* **2006**, *45* (17), 5468−77.
(55) Munshi, S.; Chen, Z.; Li, Y.; Olsen, D. B.; Fraley, M. E.; Hungate, R. W.; Kuo, L. C. Rapid X-ray diffraction analysis of HIV-1 protease-inhibitor complexes: inhibitor exchange in single crystals of the bound enzyme. *Acta Cryst.* **1998**, *54* (Pt 5), 1053−60.
(56) Skulnick, H. I.; Johnson, P. D.; Aristoff, P. A.; Morris, J. K.; Lovasz, K. D.; Howe, W. J.; Watenpaugh, K. D.; Janakiraman, M. N.; Anderson, D. J.; Reischer, R. J.; Schwartz, T. M.; Banitt, L. S.; Tomich, P. K.; Lynn, J. C.; Horng, M. M.; Chong, K. T.; Hinshaw, R. R.; Dolak, L. A.; Seest, E. P.; Schwende, F. J.; Rush, B. D.; Howard, G. M.; Toth, L. N.; Wilkinson, K. R.; Romines, K. R.; et al. Structure-based design of nonpeptidic HIV protease inhibitors: the sulfonamide-substituted cyclooctylpyramones. *J. Med. Chem.* **1997**, *40* (7), 1149−64.
(57) Jain, R.; Wagner, M. Kolmogorov-Smirnov scores and intrinsic mass tolerances for peptide mass fingerprinting. *J. Proteome Res.* **2010**, *9* (2), 737−42.
(58) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 493−500.
(59) Zhu, R.; Hu, L.; Li, H.; Su, J.; Cao, Z.; Zhang, W. Novel natural inhibitors of CYP1A2 identified by in silico and in vitro screening. *Int. J. Mol. Sci.* **2011**, *12* (5), 3250−62.

(60) Fagin, R.; Kumar, R.; Sivakumar, D. Comparing Top k Lists. *SIAM J. on Discrete Math.* **2003**, *17* (1), 134−160.
(61) Paluszewski, M.; Karplus, K. Model quality assessment using distance constraints from alignments. *Proteins* **2009**, *75* (3), 540−9.
(62) Pihur, V.; Datta, S.; Datta, S. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics* **2007**, *23* (13), 1607−15.
(63) Davidor, Y.; Schwefel, H.-P.; Männer, R.; Eiben, A.; Raué, P.; Ruttkay, Z., Genetic algorithms with multi-parent recombination. In *Parallel Problem Solving from Nature—PPSN III*, Springer: Berlin/Heidelberg, 1994; Vol. *866*, pp 78−87.
(64) Capcarrere, M. S.; Freitas, A. A.; Bentley, P. J.; Johnson, C. G.; Timmis, J.; Ting, C.-K., On the Mean Convergence Time of Multi-parent Genetic Algorithms Without Selection. In *Advances in Artificial Life*, Springer: Berlin/Heidelberg, 2005; Vol. *3630*, pp 403−412.
(65) Jarvelin, K.; Kekalainen, J., Cumulated gain-based evaluation of IR techniques. *TOIS* **2002**, *20* (4).
(66) Liu, Q.; Huang, J.; Liu, H.; Wan, P.; Ye, X.; Xu, Y. Analyses of domains and domain fusions in human proto-oncogenes. *BMC Bioinf.* **2009**, *10*, 88.
(67) Vanhaecke, T.; Papeleu, P.; Elaut, G.; Rogiers, V. Trichostatin A-like hydroxamate histone deacetylase inhibitors as therapeutic agents: toxicological point of view. *Curr. Med. Chem.* **2004**, *11* (12), 1629−43.
(68) Ree, A. H.; Dueland, S.; Folkvord, S.; Hole, K. H.; Seierstad, T.; Johansen, M.; Abrahamsen, T. W.; Flatmark, K. Vorinostat, a histone deacetylase inhibitor, combined with pelvic palliative radiotherapy for gastrointestinal carcinoma: the Pelvic Radiation and Vorinostat (PRAVO) phase 1 study. *Lancet Oncol.* **2010**, *11* (5), 459−64.
(69) Keen, J. C.; Yan, L.; Mack, K. M.; Pettit, C.; Smith, D.; Sharma, D.; Davidson, N. E. A novel histone deacetylase inhibitor, scriptaid, enhances expression of functional estrogen receptor alpha (ER) in ER negative human breast cancer cells in combination with 5-aza 2′-deoxycytidine. *Breast Cancer Res. Treat.* **2003**, *81* (3), 177−86.
(70) Goda, N.; Dozier, S. J.; Johnson, R. S. HIF-1 in cell cycle regulation, apoptosis, and tumor progression. *Antioxid. Redox Signal* **2003**, *5* (4), 467−73.
(71) Li, Y. H.; Tao, P. Z.; Liu, Y. Z.; Jiang, J. D. Geldanamycin, a ligand of heat shock protein 90, inhibits the replication of herpes simplex virus type 1 in vitro. *Antimicrob. Agents Chemother.* **2004**, *48* (3), 867−72.
(72) Yao, J. Q.; Liu, Q. H.; Chen, X.; Yang, Q.; Xu, Z. Y.; Hu, F.; Wang, L.; Li, J. M. Hsp90 inhibitor 17-allylamino-17-demethoxy-geldanamycin inhibits the proliferation of ARPE-19 cells. *J. Biomed. Sci.* **2010**, *17*, 30.
(73) Ramanathan, R. K.; Egorin, M. J.; Erlichman, C.; Remick, S. C.; Ramalingam, S. S.; Naret, C.; Holleran, J. L.; TenEyck, C. J.; Ivy, S. P.; Belani, C. P. Phase I pharmacokinetic and pharmacodynamic study of 17-dimethylaminoethylamino-17-demethoxygeldanamycin, an inhibitor of heat-shock protein 90, in patients with advanced solid tumors. *J. Clin. Oncol.* **2010**, *28* (9), 1520−6.
(74) Wang, S.; Xu, Y.; Maine, E. A.; Wijeratne, E. M.; Espinosa-Artiles, P.; Gunatilaka, A. A.; Molnar, I. Functional characterization of the biosynthesis of radicicol, an Hsp90 inhibitor resorcylic acid lactone from Chaetomium chiversii. *Chem. Biol.* **2008**, *15* (12), 1328−38.
(75) Kino, T.; Hatanaka, H.; Hashimoto, M.; Nishiyama, M.; Goto, T.; Okuhara, M.; Kohsaka, M.; Aoki, H.; Imanaka, H. FK-506, a novel immunosuppressant isolated from a Streptomyces. I. Fermentation, isolation, and physico-chemical and biological characteristics. *J. Antibiot. (Tokyo)* **1987**, *40* (9), 1249−55.
(76) Sehgal, S. N.; Baker, H.; Vezina, C. Rapamycin (AY-22,989), a new antifungal antibiotic. II. Fermentation, isolation and characterization. *J. Antibiot. (Tokyo)* **1975**, *28* (10), 727−32.
(77) Desai, D.; Salli, U.; Vrana, K. E.; Amin, S. SelSA, selenium analogs of SAHA as potent histone deacetylase inhibitors. *Bioorg. Med. Chem. Lett.* **2010**, *20* (6), 2044−7.
(78) Hu, E.; Dul, E.; Sung, C. M.; Chen, Z.; Kirkpatrick, R.; Zhang, G. F.; Johanson, K.; Liu, R.; Lago, A.; Hofmann, G.; Macarron, R.; de los Frailes, M.; Perez, P.; Krawiec, J.; Winkler, J.; Jaye, M. Identification of novel isoform-selective inhibitors within class I histone deacetylases. *J. Pharmacol. Exp. Ther.* **2003**, *307* (2), 720−8.
(79) Wishart, D. DrugBank. http://www.drugbank.ca/drugs/DB02546 (accessed Dec 14th, 2011).
(80) Cheung, K. M.; Matthews, T. P.; James, K.; Rowlands, M. G.; Boxall, K. J.; Sharp, S. Y.; Maloney, A.; Roe, S. M.; Prodromou, C.; Pearl, L. H.; Aherne, G. W.; McDonald, E.; Workman, P. The identification, synthesis, protein crystal structure and in vitro biochemical evaluation of a new 3,4-diarylpyrazole class of Hsp90 inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *15* (14), 3338−43.
(81) Taldone, T.; Sun, W.; Chiosis, G. Discoveryery and development of heat shock protein 90 inhibitors. *Bioorg. Med. Chem.* **2009**, *17* (6), 2225−35.
(82) Proisy, N.; Sharp, S. Y.; Boxall, K.; Connelly, S.; Roe, S. M.; Prodromou, C.; Slawin, A. M.; Pearl, L. H.; Workman, P.; Moody, C. J. Inhibition of Hsp90 with synthetic macrolactones: synthesis and structural and biological evaluation of ring and conformational analogs of radicicol. *Chem. Biol.* **2006**, *13* (11), 1203−15.
(83) Barril, X.; Brough, P.; Drysdale, M.; Hubbard, R. E.; Massey, A.; Surgenor, A.; Wright, L. Structure-based discovery of a new class of Hsp90 inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *15* (23), 5187−91.
(84) Dymock, B. W.; Barril, X.; Brough, P. A.; Cansfield, J. E.; Massey, A.; McDonald, E.; Hubbard, R. E.; Surgenor, A.; Roughley, S. D.; Webb, P.; Workman, P.; Wright, L.; Drysdale, M. J. Novel, potent small-molecule inhibitors of the molecular chaperone Hsp90 discovered through structure-based design. *J. Med. Chem.* **2005**, *48* (13), 4212−5.