Better than Random? The Chemotype Enrichment Problem

Mark D. Mackey* and James L. Melville

Cresset BioMolecular Discovery Ltd., BioPark Hertfordshire, Broadwater Road, Welwyn Garden City, Hertfordshire AL7 3AX, United Kingdom

Received October 27, 2008

Chemotype enrichment is increasingly recognized as an important measure of virtual screening performance. However, little attention has been paid to producing metrics which can quantify chemotype retrieval. Here, we examine two different protocols for analyzing chemotype retrieval: "cluster averaging", where the contribution of each active to the scoring metric is proportional to the number of other actives with the same chemotype, and "first found", where only the first active for a given chemotype contributes to the score. We demonstrate that this latter analysis, common in the qualitative analysis used in the current literature, has important drawbacks when combined with quantitative metrics.

INTRODUCTION

Virtual Screening (VS) is a valuable tool for drug discovery.[1] Given the rapid proliferation of different approaches to VS,[2,3] validation of these techniques is becoming increasingly important.[4−7] One of the largest scale evaluations to date was carried out by researchers at GlaxoSmithKline,[8] who evaluated ten docking programs against eight proteins, and required the input of experts both for the targets and for the docking programs used. Given the effort and manpower required to carry out such evaluations it is therefore vital that as much information as possible can be extracted from the exercise. The difficulties of doing VS reliably, especially in a docking context, have recently been summarized.[9] These concerns have motivated a great deal of research into how best to evaluate VS experiments.[10−16]

Regardless of whether structure-based techniques such as docking or ligand-based techniques such as similarity searching are employed for VS, two universal considerations are the metrics used to measure success, and the composition of the data set. While structure-based methods such as docking provide an estimate of binding affinity, these are widely acknowledged to be, at best, only weakly correlated with experimental data.[8,17] Therefore, the more important goal is to aim to separate a database into actives and inactives ("decoys"), normally by returning a ranked list of molecules ordered by decreasing likelihood of activity.

It is now recognized that the choice of decoys must be carried out carefully, to avoid trivial separation of actives and decoys due to large differences in physicochemical properties (e.g., molecular weight or charge).[18] An additional consideration is the desirability of "scaffold hopping", i.e. the retrieval of actives which are diverse with respect to their basic molecular frameworks. If only structurally very similar molecules are required, then well-established 2D methods will suffice, and the extra complexity of more advanced virtual screening is wasted effort. Therefore, increasing attention has been paid to the structural diversity of the actives, and additional measures of success have focused on the number of different chemotypes retrieved by a method.[8]

The best way to quantify the success of retrospective virtual screening evaluations is yet another area of disagreement. In most real-world scenarios, only a small fraction of the best scoring molecules from a ranked database would be experimentally screened, so an obvious measure of success is the number of actives found in some small fraction of the ranked database, e.g. the top 1% of the database. Normalizing this value with respect to the number of actives that would be retrieved by random selection yields the enrichment factor (EF),[19] which has been one of the most popular measures of success since its introduction. However, the EF is sensitive to the cutoff value, especially when this results in a small number of molecules being used in the calculation. Alternatively, the use of the Area Under the Receiver Operating Characteristic curve (AUC) metric has been advocated.[20] This has the advantage of taking the entire database into account and has well-known statistical properties. A disadvantage of the AUC is that no emphasis is placed on retrieval in the top part of the database. Various attempts have been made to merge the good qualities of the EF and AUC. The Robust Initial Enhancement (RIE) considers the entire data set but applies a decreasing exponential weight, effectively smoothing out the effect of the cutoff.[21] A further modification, the Boltzmann Enhanced Discrimination of ROC (BEDROC), applies a range scaling of the RIE to produce a metric which varies between zero and one.[22]

It is usually desired that new lead molecules are structurally dissimilar from known actives, both for patent reasons and to ensure that the new leads have different DMPK (Drug Metabolism and Pharmacokinetics) properties to the extant ones. For this reason, virtual screens are usually only seen as successful if they locate previously unknown chemotypes. Measurement of chemotype retrieval would thus seem to be important if the degree of success of VS methods is to be ascertained. Despite the proliferation of metrics for assessing VS performance, little attention has been paid to the use of these metrics when moving from a molecule-centric to a chemotype-centric view of retrieval.

* Corresponding author phone +44 1707 356120; e-mail: mark@cresset-bmd.com.

10.1021/ci8003978 CCC: $40.75
Published on Web 04/27/2009

By far the most common approach to chemotype analysis is to count the number of chemotypes "found" at a certain fraction of the screen. A chemotype is denoted as "found" so long as at least one of its members has been "found". Commonly, a graphical analysis is presented, showing the number of chemotypes found at different percentages of the database screened.[23−25] Similarly, Warren and co-workers presented a table showing the percentage of the database required to find at least one member of each chemotype.[8]

This "First Found" (FF) method has the advantage that it corresponds closely to our intuitive understanding of how VS is used in practice: simple 2D similarity searching should uncover the close structural analogues of any active molecule in a given chemotype; therefore, only one member of each chemotype needs to be present in the relevant portion of the ranked database.

Alternative chemotype analysis methods have been proposed by Clark and Webster-Clark in the context of calculating ROC curve areas.[26] They propose setting the contribution of each active to the final score to be inversely proportional to the number of other members of its cluster. Each cluster thus has the same overall weighting in the final ROC area. This method turns out to be mathematically equivalent to performing multiple analyses, each time picking just one active from each cluster, and computing the final metric as the average over all possible selections. We will call this method the "Cluster Average" method (CA).

Clark and Webster-Clark also propose a hybrid of the CA and FF methods which they call the "Harmonic Average" method (HA).[26] The first active found from each cluster is assigned a weight of one, the second a weight of one-half, the third a weight of one-third, and so forth. The rationale comes from information theory: the first active found contains all the information known to date about that chemotype and so should have the same weight as a singleton. The second active found now contains only half of the known information about that chemotype, so it is assigned a weight of 0.5 and so forth. This method is related to the FF method in that the ranking of the highest-ranked active in a cluster is the most important. However, it is not as sensitive to the exact composition of the data set as the FF method, as the positions of all of the actives are used to determine the final score.

In this paper we will present formulas for applying the FF and CA methods to a variety of widely used metrics. Based on these, we argue that the widely used FF method has serious flaws (as does the HA method) which are not shared by the CA method.

METHODS

There is no widely accepted formal definition of a chemotype: division of sets of molecules into chemotypes has in the past been accomplished both by automated methods[25,27] and by manual inspection.[8,23] For this study we will simply assume that actives can be partitioned into disjoint clusters by some method, where each cluster represents a separate chemotype.

Many different enrichment metrics have been used in the literature: a good analysis of the commonly used ones is given by Truchon et al.[22] We will focus here on the EF, ROC, RIE, and BEDROC metrics, and present formulas for calculating these metrics with both cluster average (CA) and first found (FF) chemotype analysis. We will also briefly discuss the harmonic average (HA) method. These cluster analysis methods can all be expressed as weighting schemes, as shown in Table 1.

Table 1. Weighting Schemes That Have Been Used for Clustered VS Data

name	weight assigned to each active
Unclustered	1
Cluster Average (CA)[26]	1/(cluster size)
First Found (FF)[23]	1 for the first active found in each cluster, 0 for the others
Harmonic Average (HA)[26]	1 for the first active found in each cluster, 1/2 for the second, 1/3 for the third...

In the following discussion, N is the total number of compounds, n_a is the number of actives, and n_d is the number of decoys (so $n_d + n_a = N$). When discussing unclustered data x_i is the relative rank of the i^{th} active, and f_i is the fraction of decoys ranked better than active i. For clustered data, the actives are clustered into m clusters of size $c_1 \ldots c_m$, x_{jk} refers to the relative rank of the k^{th} active from the j^{th} cluster, and f_{jk} refers to the fraction of decoys ranked higher than the k^{th} active from the j^{th} cluster.

Enrichment Factors. The enrichment factor (EF) has been the most widely used metric for VS analysis, despite its shortcomings.[28] The EF at a fraction χ of the data set is calculated in the nonclustered case as the number of actives found divided by the expected number of actives from a random ranking

$$EF(\chi) = \frac{nfound}{nrandom} = \frac{\sum_{i=1}^{n_a} \delta(x_i)}{\left\langle \sum_{i=1}^{n_a} \delta(x_i^R) \right\rangle} = \frac{\sum_{i=1}^{n_a} \delta(x_i)}{n_a \chi} \qquad \delta(x_i) = \begin{cases} 1, x_i \le \chi \\ 0, x_i > \chi \end{cases} \tag{1}$$

where $x^R{}_i$ indicates the rank of the ith active if the actives were randomly assigned ranks with a uniform distribution. The angle brackets indicate the expectation value (i.e., the mean over many such randomized distributions). This formula can be rearranged to give an alternative interpretation

$$EF(\chi) = \frac{\sum_{i=1}^{n_a} \delta(x_i)}{n_a \chi} = \frac{1}{n_a} \sum_{i=1}^{n_a} \frac{\delta(x_i)}{\chi} \tag{2}$$

i.e. we calculate a "per active" enrichment factor, and the average of these is the enrichment factor for the data set. This makes generalizing to the clustered case straightforward: the enrichment factor for the data set is the weighted average enrichment factor for each cluster.

For methods with uniform intracluster weights (CA), the clustered modification of the EF formula is simple. The enrichment for each cluster is the number of cluster members found divided by the expected number of cluster members found by random selection. This is averaged over all clusters to give the final metric

$$EF_{CA}(\chi) = \frac{1}{m}\sum_{j=1}^{m}\frac{\sum_{k=1}^{c_j}\delta(x_{jk})}{c_j\chi} \tag{3}$$

The CA enrichment factor has similar properties to the unclustered enrichment factor metric: values range from ~0 to $1/\chi$, a random distribution of actives yields an EF of 1, and the calculated value can be interpreted as "how much better are we doing than random selection?" The essential difference is the contribution of each active to the final score: in CA this is weighted inversely to the size of the cluster to which it belongs: a singleton is given 16 times more weight than any individual active from a cluster of 16.

Calculating EFs which make use of the FF weighting is slightly more complicated. Again, we take the overall EF to be the arithmetic mean of the EFs from each cluster. To obtain the EF for a cluster, we note that the probability that at least one active from a cluster with c members is selected in a random fraction χ of the data set is the complement of the probability that no actives are found

$$p(\chi) = 1 - (1 - \chi)^c \tag{4}$$

Taking the definition of enrichment factor as the number of actives found divided by the expected number of actives found by random selection, we obtain

$$EF_{FF}(\chi) = \frac{1}{m}\sum_{j=1}^{m}\frac{\delta(\min x_j)}{1 - (1 - \chi)^{c_j}} \tag{5}$$

where "min x_j" is the minimum relative rank for cluster j, i.e. the relative rank of the first active found from that cluster. In other words, the enrichment factor for an individual cluster is 0 if no members occur above the cutoff (χ), and the inverse of the probability that we would have found that cluster by random selection otherwise.

It has been pointed out that the enrichment factor metric is flawed due to its reliance on extensive quantities (the number of actives and decoys) and that a better measure can be obtained by considering the number of actives found at a particular fraction of the decoys screened, rather than a particular fraction of the whole data set.[16,28] This "ROC enrichment" metric can be analyzed in exactly the same way as the traditional EF metric. In eqs 1−5, we simply use f_i (the fraction of decoys scoring better than active i) rather than x_i and replace χ with χ_d, defined as the fraction of the decoys screened. The conclusions drawn later in this paper about EF_{FF} vs EF_{CA} are unchanged if we use ROC enrichment rather than EF.

Area Under Curve for Receiver Operator Characteristic (AUC). The AUC is the area under the ROC curve. It can be interpreted as the probability that a randomly chosen active is ranked higher than a randomly chosen decoy.[29] Given that the probability of a specific active being ranked higher than a randomly chosen decoy is simply the relative rank of that active, the AUC can therefore be calculated as the average relative rank of the actives

$$AUC = \frac{1}{n_a}\sum_{i=1}^{n_a}(1 - f_i) \tag{6}$$

Clark and Webster-Clark extend this to the clustered case with uniform intracluster weights (e.g., the CA method)[26]

$$AUC_{CA} = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{c_j}\sum_{k=1}^{c_j}(1 - f_{jk}) \tag{7}$$

They also provided a formula for AUC_{HA} (8) which would suggest an analogous formula for AUC_{FF} by simply replacing the weighting scheme with a weight of 1 for the best-ranked active and 0 for the others

$$AUC_{HA} = \frac{\sum_{j=1}^{m}\sum_{k=1}^{c_j}\frac{(1 - f_{jk})}{k}}{\sum_{j=1}^{m}\sum_{k=1}^{c_j}\frac{1}{k}} \tag{8}$$

$$AUC_{FF'} = \frac{1}{m}\sum_{j=1}^{m}(1 - \min f_j) \tag{9}$$

Here, we use the symbol FF′ to differentiate this version of AUC_{FF} from a superior formulation which we shall derive in the following discussion. The problem with both of these formulas is that they make no correction for the fact that larger clusters are more likely to have members found early by random chance. The value obtained for a random distribution thus depends on the size distribution of the clusters. This makes these metrics impossible to interpret in the general case: it cannot even be said whether a particular metric value is better than or worse than random without detailed knowledge of the distribution of cluster sizes in the experiment.

Equations 8 and 9 might possibly be modified by applying a correction factor to scale the expected value of the AUC for a random distribution of actives back to 0.5 (the expected value of the unclustered AUC for a random distribution). However, any such linear correction would affect the range. The problems of transferability and interpretability would remain.

A better "first found" formulation of the ROC AUC would use its interpretation as the probability that a randomly chosen active is better ranked than a randomly chosen inactive: for a single active i this value is simply $(1-f_i)$, and the ROC AUC is just the arithmetic mean over all actives. The analogous measure for a cluster of c compounds would be the probability that that cluster is better ranked than a set of c randomly chosen inactives. If we define the overall AUC_{FF} metric to be the arithmetic mean of the cluster values, then we get

$$AUC_{FF} = \frac{1}{m}\sum_{j=1}^{m}(1 - \min f_j)^{c_j} \tag{10}$$

This measure, unlike that from 9, is corrected for the cluster sizes and hence preserves the expected AUC range of [0,1] and the expected value for a random data set of 0.5. In what follows, when referring to AUC_{FF}, we mean this corrected version of the AUC, not eq 9. A similar improvement in the formulation of the HA metric would measure the probability that the weighted sum of scores for a cluster is greater than the weighted sum of scores for a random selection of inactives. The calculations for this are very complex, however.

Robust Initial Enrichment and BEDROC. The RIE metric[21] can be formulated in a way analogous to the EF

$$RIE(\alpha) = \frac{\sum_{i=1}^{n_a} e^{-\alpha x_i}}{\left\langle \sum_{i=1}^{n_a} e^{-\alpha x_i^R} \right\rangle} \tag{11}$$

Both metrics can be viewed as the sum of a "score" for each active, divided by the expected sum of scores for a random distribution. For EF(χ) the score is 1 if the active is above the cutoff χ and 0 otherwise, while for RIE the score is an exponential function of the relative rank, with the α parameter controlling how much the score is weighted toward early retrieval. In the same way as EF, the RIE for the data set can be expressed as the average of an RIE value for each active, calculated as the score for that active divided by the expected score from a uniform distribution

$$RIE(\alpha) = \frac{1}{n_a} \sum_{i=1}^{n_a} \frac{e^{-\alpha x_i}}{\langle e^{-\alpha x^R} \rangle} \tag{12}$$

where the expected score from a uniform distribution is[22]

$$\langle e^{-\alpha x^R} \rangle = \frac{1}{N} \left(\frac{1 - e^{-\alpha}}{1 - e^{-\alpha/N}} \right) \tag{13}$$

Adjusting this for the cluster-average (CA) method is simple: the score for a cluster is the average score of all of the members of the cluster. The expected value for this average from a uniform distribution is the same as the expected value from a single active, giving

$$RIE_{CA}(\alpha) = \frac{1}{m} \sum_{j=1}^{m} \frac{\frac{1}{c_j} \sum_{k=1}^{c_j} e^{-\alpha x_{jk}}}{\langle e^{-\alpha x^R} \rangle} \tag{14}$$

For the first found (FF) method, the appropriate measure for each cluster is the score of the first active found, divided by the expected score of the first active found

$$RIE_{FF}(\alpha) = \frac{1}{m} \sum_{j=1}^{m} \frac{e^{-\alpha \min(x_j)}}{\langle e^{-\alpha \min(x_j^R)} \rangle_{c_j}} \tag{15}$$

where $\min(x_j)$ is the minimum relative rank for the jth cluster. The denominator represents the expected score for the lowest of a set of c_j ranks selected at random from a uniform distribution. This expected score can be calculated by multiplying the probability that the first cluster member found is at position k by the score for position k. The probability that the first cluster member found is at position k is given by 16, as one member is at position k and there are $c_j - 1$ other cluster members in $N - k$ positions. This leads to the formula for the expected score given in 17

$$p = \frac{\text{no. ways of arranging } c_j - 1 \text{ actives in } N - k \text{ positions}}{\text{no. ways of arranging } c_j \text{ actives in } N \text{ positions}} = \frac{\binom{N-k}{c_j - 1}}{\binom{N}{c_j}} \tag{16}$$

$$\langle e^{-\alpha \min(x_j^R)} \rangle_{c_j} = \frac{\sum_{k=1}^{N} \binom{N-k}{c_j - 1} e^{-\alpha k/N}}{\binom{N}{c_j}} \tag{17}$$

We were unable to determine an analytical formula for this sum, so it needs to be numerically calculated in order to calculate RIE_{FF} values.

The BEDROC metric is obtained by linearly scaling the RIE metric to [0,1]. The $BEDROC_{CA}$ and $BEDROC_{FF}$ metrics can be defined analogously

$$BEDROC_{FF} = \frac{RIE_{FF} - \min RIE_{FF}}{\max RIE_{FF} - \min RIE_{FF}} \tag{18}$$

$$BEDROC_{CA} = \frac{RIE_{CA} - \min RIE_{CA}}{\max RIE_{CA} - \min RIE_{CA}} \tag{19}$$

For the FF analysis method, the maximum RIE_{FF} value occurs when the first m ranks are occupied by a single member of each of the m clusters: this can be calculated according to the formulas published by Truchon and Bayly,[22] using the number of clusters rather than the number of overall actives. The minimum RIE_{FF} value is more complicated. It occurs when all clusters are found as late as possible, with all members of larger clusters being found before all members of smaller clusters, and the exact value depends on the number of members of each cluster. However, provided that either α is large (>10) or the proportion of actives in the data set R_a is small, the minimum RIE is well approximated by 20

$$\min RIE_{FF} \approx \frac{\alpha}{e^{\alpha} - 1} \tag{20}$$

In the cluster-average case, the maximum RIE_{CA} value is difficult to calculate for the same reason as the minimum RIE_{FF}: it depends on the order in which the clustered actives are found. The absolute maximum RIE_{CA} value occurs when all singletons are found first, followed by all clusters of two actives, and so forth. In the case where $\alpha R_a \ll 1$, the maximum RIE_{CA} value will approximate the maximum RIE value where the number of actives is set equal to the number of clusters. An upper bound on the error of this approximation can be obtained by considering the difference in RIE score between a singleton and a cluster with n_a members all assigned the score of the n^{th} member

$$\text{relative error in max } RIE_{CA} < 1 - e^{-\alpha R_a} \tag{21}$$

If this error is large, then it is simplest just to construct the "perfect" data set and calculate its RIE_{CA} value. The minimum RIE_{CA} value can be approximated using eq 20 where α is large or R_a is small: the error on this value is bounded by

$$\text{relative error in min } RIE_{CA} < e^{-\alpha(1-R_a)} \tag{22}$$

It should be noted that since the maximum RIE_{CA} and RIE_{FF} values are dependent on the order in which the clustered actives are found, in the majority of cases a "perfect" result in which all actives are found before all decoys will get a $BEDROC_{CA}$ or $BEDROC_{FF}$ value less than one, and often significantly less than one unless the ratio of decoys to actives is very high.

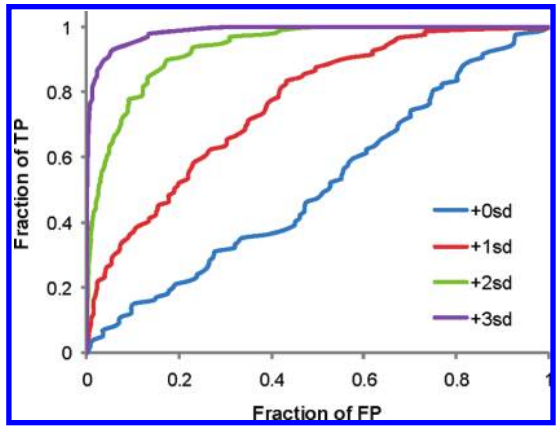

Figure 1. ROC curves for synthetic data, 100 actives, 10000 decoys, active and decoy scores from normal distribution with the mean of the active distribution shifted from the mean of the inactive distribution by the specified amount.

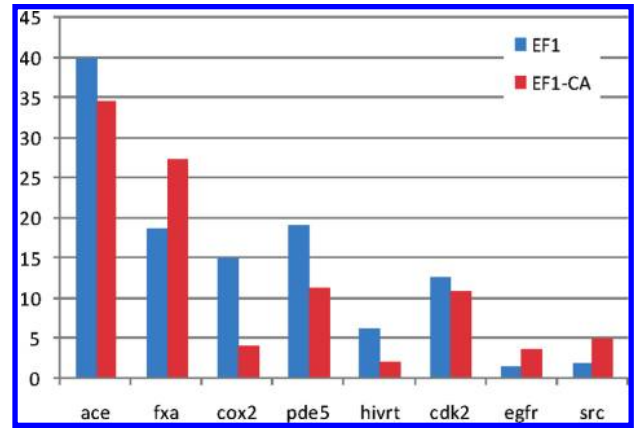

Figure 2. Enrichment factors at 1% for the DOCK results on the DUD data set with and without cluster-average chemotype correction.

Simulated Data Sets. In order to evaluate fully the effect of data set size, clustering, score distribution, ratio of actives to decoys, etc. on each metric, it is useful to generate simulated data sets where these parameters can be explicitly varied. The use of exponential cumulative distribution functions (CDFs) has been suggested for this purpose.[22] However, in our hands, these had problems with saturation when using small data set sizes: the suggested protocol uses the CDF to generate a rank for an active repeatedly until an unused rank is generated. This introduces biases when the number of available ranks is not much larger than the number of actives.

Instead, we generate simulated data sets by assuming that the scores from the actives and the decoys are both normally distributed. The active and decoy list are then merged and sorted based on these scores and then assigned ranks based on their position in the sorted list. The degree of discrimination demonstrated by the VS method is then controlled by changing the mean of the active distribution relative to the decoy distribution. For example, a weak enrichment of actives can be simulated by assigning to the decoys scores randomly generated from a normal distribution with mean 0 and standard deviation 1, while assigning to the actives scores randomly generated from a normal distribution with mean 1 and standard deviation 1 (the +1sd line in Figure 1). This approach removes the saturation problem and allows the simulation of highly enriched data sets with small numbers of decoys. Example ROC curves generated with this algorithm are shown in Figure 1.

RESULTS

Application to the DUD Data Set. Before analyzing in detail the metrics that have been mentioned, the question arises as to whether chemotype corrections actually have a significant effect in practice. To demonstrate that they do, we have reanalyzed the results provided by Huang et al. for the performance of DOCK against the DUD data set,[18] downloaded from the DUD Web site, http://dud.docking.org/r2/energies (accessed August 25, 2007). We have also made use of the data as processed by Good[27] to produce a set of more druglike actives, partitioned into chemotypes by means of reduced graphs.[30] Additionally, where a molecule was present in more than one tautomer, we used only the best scoring structure. We have focused on 8 data sets where there were 17 or more active clusters and where DOCK showed reasonable retrieval performance: ace, cdk2, cox2, egfr, fxa, hivrt, pde5, and src. These data sets show a range of clustering behavior with the ratio of the number of actives to the number of clusters ranging across nearly an order of magnitude, 1.18 (for pde5) to 9.13 (for egfr).

Figure 2 shows the results of the analysis, using the "all" decoy set,[18] with both the unclustered and the CA metrics. For some data sets this difference between the two methods is large. Most notably, the cox2 data set has an enrichment factor of 15, significantly better than random. However, correctly accounting for the chemotype bias yields a substantially smaller enrichment value of 4.1. This large reduction can be traced to the observation that although there are 212 actives in the cox2 data set, there are only 44 clusters, and around half of the actives are found in one cluster. The apparently excellent early retrieval for this target is comprised mainly of retrieval of actives from this one cluster, which are all very structurally similar to the ligand crystallized in the protein structure that was used. As a result, the adjusted cox2 results show that DOCK is significantly less useful for scaffold-hopping purposes on the cox2 data set than the original EF values suggest.

Conversely, applying the CA correction indicates that DOCK performs significantly better at retrieving different chemotypes for fxa than was apparent from the plain EF results. 19% of the actives were retrieved in the top 1% of the results, but these represented a wide variety of chemotypes such that 27% of the chemotype diversity was found. So although the uncorrected EF metric suggested that the performance of DOCK for cox2 and fxa was roughly comparable, the corrected metric shows that markedly better performance was obtained on the fxa data set.

If instead of enrichment factors we use a whole-data-set metric such as BEDROC, we see a similarly marked change on applying chemotype correction (Figure 3). It can be seen that, based on the BEDROC values, the results for the cox2 and pde5 targets are clearly superior to those obtained for hivrt, cdk2, egfr, and src. However, taking a chemotype enrichment view and using the $BEDROC_{CA}$ values, the performances for cox2 and pde5 are substantially reduced, and it is more difficult to discern a difference in performance in these six targets.

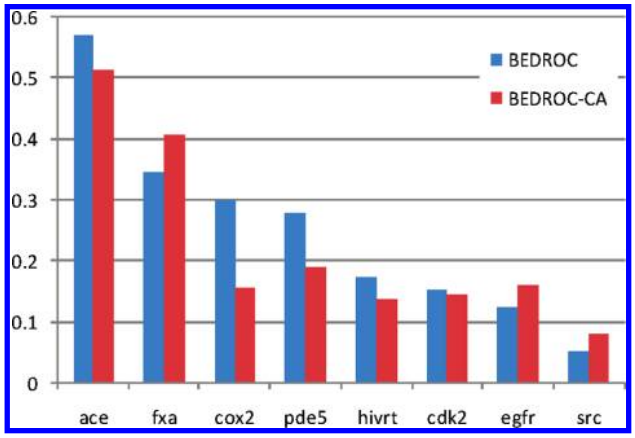

Figure 3. BEDROC ($\alpha = 20$) values for the DOCK results on the DUD data set with and without cluster-average chemotype correction.

This demonstrates that clustering of actives can have a measurable and substantial effect on the conclusions drawn about the success of VS in a scaffold hopping context, both in absolute terms for a particular target, and in evaluating success between targets. We stress that these effects have been observed with DOCK, a structure-based VS tool. This underlines that chemotype bias is an issue that must be addressed for structure-based as well as ligand-based methods, contrary to claims made in the literature.[31]

Application to Simulated Data. To further demonstrate the effect of the cluster corrections, we generated synthetic data sets for both clustered and unclustered data and calculated the various metrics. 1000 simulations were performed, each with 100 actives in a total of 10000 compounds, with corrections based on the CA and FF method. The scores for the decoys are generated with mean 0 and standard deviation 1. The scores for the actives are generated with standard deviation 1 and differing mean values reflecting differing levels of discrimination between actives and decoys. The results are shown in Table 2, along with standard metric values ignoring any clustering. The cluster sizes in data sets F, G, and H are those in the cox2 DUD data set as clustered by Good.[27] This data set provides a real-life example of highly clustered data: it should be noted that other data sets in DUD are even more highly clustered.

The importance of correcting for clustering can be seen from data set G in Table 2: here one cluster representing ~40% of the actives is enriched, while the other 43 clusters of actives are not discriminated from the decoys at all. This kind of distribution is not unknown for 2D searches on data sets where a proportion of the actives are structurally related to the search query. The raw EF metric gives an enrichment of 15.6 in this case, which significantly overestimates the actual ability to distinguish actives from inactives. Clustering the data set and using the cluster-corrected metrics gives much lower values, correctly indicating that the ordered data set has only been slightly enriched in chemotypes.

The dependence of the FF metrics on cluster size can be seen in data sets C, D, and E. Although the scores of all of the actives were generated independently of the clustering, the EF_{FF}, AUC_{FF}, and $BEDROC_{FF}$ values obtained depend strongly on the cluster size. This is a highly undesirable property, as it means that the interpretation of a metric value

Table 2. Enrichment Factors at 1% and ROC AUC Values and BEDROC Values for Synthetic Data Sets[c]

data set	mean active score	clusters	EF(0.01), no clustering	$EF_{CA}(0.01)$	$EF_{FF}(0.01)$	AUC, no clustering	AUC_{CA}	AUC_{FF}	BEDROC, no clustering	$BEDROC_{CA}$	$BEDROC_{FF}$
A	1	100 singletons	8.8 ± 2.7	8.8 ± 2.7	8.8 ± 2.7	0.76 ± 0.02	0.76 ± 0.02	0.76 ± 0.02	0.24 ± 0.03	0.24 ± 0.03	0.24 ± 0.03
B	1	20 clusters of 5	8.7 ± 2.7	8.7 ± 2.7	7.4 ± 2.2	0.76 ± 0.03	0.76 ± 0.03	0.86 ± 0.04	0.24 ± 0.03	0.24 ± 0.03	0.64 ± 0.07
C	2	100 singletons	32.0 ± 4.2	32.0 ± 4.2	32.0 ± 4.2	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01	0.58 ± 0.04	0.58 ± 0.04	0.58 ± 0.04
D	2	20 clusters of 5	31.9 ± 4.1	31.9 ± 4.1	17.5 ± 1.5	0.92 ± 0.01	0.92 ± 0.01	0.98 ± 0.01	0.58 ± 0.04	0.58 ± 0.04	0.93 ± 0.03
E	2	10 clusters of 10	31.9 ± 4.3	31.9 ± 4.3	10.2 ± 0.5	0.92 ± 0.01	0.92 ± 0.01	0.99 ± 0.01	0.58 ± 0.04	0.58 ± 0.04	0.98 ± 0.01
F	2	cox2[a]	25.0 ± 1.9	24.9 ± 4.8	21.8 ± 4.8	0.92 ± 0.01	0.92 ± 0.02	0.94 ± 0.02	0.61 ± 0.02	0.51 ± 0.04	0.54 ± 0.05
G	2,0[b]	cox2[a]	15.5 ± 1.7	1.2 ± 0.9	0.6 ± 0.9	0.71 ± 0.02	0.51 ± 0.04	0.51 ± 0.04	0.35 ± 0.02	0.06 ± 0.02	0.06 ± 0.02
H	0	cox2[a]	1.0 ± 0.6	1.0 ± 1.1	1.0 ± 1.1	0.50 ± 0.02	0.50 ± 0.04	0.50 ± 0.04	0.06 ± 0.01	0.05 ± 0.02	0.07 ± 0.03

[a] Cluster sizes from analysis of the cox2 DUD data set: one cluster of 125 compounds, one each of 16, 13, 12, 7, 6, and 5, three of 4, four of 3, five of 2, and 24 singletons. [b] A mean of 2 was used for the cluster with 125 members. A mean of 0 (same as the decoys) was used for the other compounds. [c] Values are given as mean ± standard deviation on 1000 repetitions.

depends on the details of how it is clustered, so that comparison of metric values between data sets becomes difficult.

DISCUSSION

The retrospective evaluation of virtual screening methods almost invariably relies on collecting a set of known actives for a target of interest. We are then trying to gain an understanding of the performance of VS methods in terms of enrichment of the actives with respect to a set of decoys. Recent papers have focused attention on the decoys: much care is needed to ensure that they match the actives in physicochemical properties to avoid trivial distinction between actives and inactives.

One serious flaw in the majority of these retrospective studies is the provenance of the actives. Typically they are either gleaned from the literature or from internal data from a pharmaceutical drug discovery program. In both cases, due to the way that medicinal chemistry works, the actives tend to be clustered into series with only minor structural differences between members of each series. The actives thus do not represent independent samples throughout "activity space". Our examination of chemotype retrieval on the DUD data set has shown that the effect of this in the interpretation of retrospective VS experiments can be large. This is true even for structure-based methods, contrary to received wisdom.[31]

A number of authors have analyzed VS experiments in the context of chemotypes. However, the majority of these studies have used (explicitly or implicitly) the "first found" method. Our analysis has shown that this method introduces a number of severe problems. The origin of these is rooted in the fact that that the FF metrics only consider the best-ranked compounds in each cluster. As the cluster sizes get larger, the probability that a member of that cluster will occur very early in the ranking just by random chance gets large. For example, with a cluster size of 100 molecules, the probability of picking at least one in a random selection of 1% of the data set is 0.634. As a result, the maximum enrichment factor at 1% obtainable from finding this cluster is only 1.58. The presence of this cluster in the data set imparts very little information about the effectiveness of a VS method. As cluster sizes increase, it therefore becomes more and more difficult to distinguish real enrichment from random chance.

On top of this, there is a second effect at play. VS metrics are attempts to quantify how well a VS method discriminates actives from inactives. The probability that a randomly chosen active scores better than a randomly chosen inactive is related to the difference in the mean of the two populations. If we take multiple samples from each and compare only the best-ranked sample, then this introduces a bias making the difference seem larger than it actually is. For example, if the scores for the inactives have mean 0, standard deviation 1, and the scores for the actives have mean 1, standard deviation 1, then the probability that a randomly chosen active will score better than a randomly chosen inactive is 0.76. The probability that the best of 10 samples from the actives has a better score than the best of 10 samples from the inactives is 0.89. With 100 samples, the probability increases to 0.95.

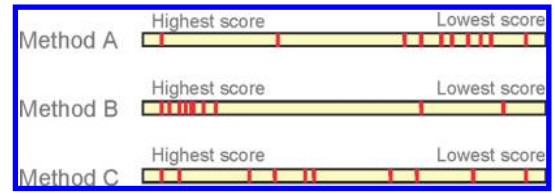

Figure 4. Example rankings for a cluster of 10 molecules.

The effect of the FF method is thus to magnify any discrimination between actives and inactives. As cluster sizes increase it is very difficult to distinguish between moderate and strong enrichment and hence to derive meaningful comparisons between VS methods.

In addition to these mathematical difficulties, there are also philosophical arguments against the FF method. The logical pitfalls associated with focusing attention on the extrema of a population rather than examining the whole distribution have been pointed out in fields as diverse as ecology, palaeobiology, and sporting statistics: the properties of the extrema (the size of the largest animals, the best baseball batting average in a given year) are generally not good guides to the properties of the whole population.[32] By analogy, focusing attention on only the highest rank attained for each chemotype will not provide a good guide to the overall performance of the VS method.

The FF method has been used in the past largely because it is seen as intuitive. However, it does not necessarily accord with our intuitition on closer inspection. Given the distribution of ranks for a cluster of 10 molecules shown in Figure 4, the obvious conclusion that method B is giving a moderate enrichment, method C is more-or-less random, and method A is worse than random. However, all three methods would be determined to be equally successful using the FF analysis.

The FF method also behaves in a counterintuitive manner as we increase the size of the data set. As more examples of each cluster are introduced, the chance that the cluster will have a good ranking just by random chance increases. As a result it becomes harder both to distinguish whether a VS method is performing better than random, and whether one VS method is performing appreciably better than another. Thus, the more information that is added to the experimental setup the less information is gained from the experiment.

A final philosophical argument against the FF method becomes apparent when we express it as a weighting scheme. The weights are assigned *post hoc*: it is only **after** the experiment that a weight is determined for each active. The weight for an active is not independent of its rank: only if no other actives in the same cluster have a higher rank do we include it in the analysis. The final metric value is thus not a function of a set of independent measurements but a set of highly correlated ones. Effectively, we are repeating an experiment multiple times, choosing the best result, and then trying to correct for the statistical biases introduced. This is not generally considered good experimental design, and it makes analytical error analysis extremely difficult.

It has been suggested that a useful metric should ideally possess the following properties:[28] (i) independence to extensive variables, (ii) robustness, (iii) straightforward assessment of error bounds, (iv) no free parameters, and (v) easily understood and interpretable. When considering clustered data, some additional desirable criteria can be

suggested: (vi) invariance on uncorrelated clustering and (vii) consistency of interpretation.

Criterion (vi) says that if the clustering method is completely independent of the VS method (i.e., their covariance is zero), then the metric value should not change on clustering. This makes intuitive sense: if the distribution of scores within each cluster is the same as the overall distribution of scores among the complete set of actives, then the clustering is irrelevant to the analysis, and the metric value should not change. In particular, the score for a "random" data set should not change upon clustering, and the score for a data set should not (on average) change on clustering if cluster membership is assigned randomly.

Criterion (vii) simply says that the interpretation of the metric should not change on clustering. If a metric value of 0.3 is "poor" and 0.8 is "good", then these scores should be "poor" and "good" using the clustered metric. If a metric value of 2.0 can be interpreted as "twice as good as random", then a clustered metric value of 2.0 should mean the same thing.

It becomes obvious from analysis of both the corrected and uncorrected metrics presented in this paper that the FF analysis method violates (ii), (iii) and (vi) and partially violates (vii). The FF analysis is highly sensitive to the rank of some compounds in the data set but not others, it is virtually impossible to assess analytical error bounds, and the values obtained depend strongly on the average cluster size. Comparison between data sets thus becomes extremely difficult. Note that these issues also plague the HA method as it also assigns *post hoc* weightings to cluster members.

The FF method has been the most widely used in the literature, with the explanation that it corresponds best to real-life situations: we perform a virtual screen, find some hits, and then do a 2D search around each of the hits to find additional examples within that chemotype. However, in a retrospective virtual screening experiment, the aim is not simply to ascertain how well a VS method performs on the particular data set but to draw conclusions about its general performance. The FF method fails at this, despite its intuitive appeal.

In contrast, the CA metrics suffer from neither the mathematical nor the philosophical problems of the FF metrics. The CA analysis method is robust, uses all of the data from the actives, and still satisfies the goal of chemotype analysis in that it removes chemotype bias (data set G, Table 2). Usefully, chemotype bias is only removed to the extent that the VS ranking method being tested actually correlates with the clustering method. If the scoring and the clustering are completely independent, then the CA metrics give the same value as the unclustered metric (Table 2, data sets A through F).

The CA method is relatively simple to apply to most of the commonly used enrichment metrics. The exception is the BEDROC metric. Most of the complexity comes from the fact that BEDROC is formulated in terms of ranks, with the concomitant restriction that no two compounds may have the same rank. This means that even if all actives are found before all inactives, different metric values will be obtained depending on the order that the actives were found. This effect is small if the ratio of actives to inactives is very small, and given the known susceptibility of BEDROC to saturation effects the BEDROC metric should only be used under these circumstances in any case.[16,22]

CONCLUSIONS

To evaluate a VS method for use in scaffold hopping, the effects of chemotype bias in data sets cannot be ignored. Several methods have been proposed in the literature for dealing with this. The "first found" method, while intuitively appealing, cannot sensibly be used in a quantitative manner. Either the metric range or the metric's value for random data become dependent on how many and how large the clusters are, or moderate enrichments become indistinguishable from strong enrichments. In contrast, the "cluster average" method is a safe, intuitive alternative, which makes use of all of the active data.

Analysis of the DOCK DUD results demonstrates that these considerations have practical significance for both ligand-based and structure-based virtual screening. The relative performance of DOCK on various targets differs greatly depending on whether we assess raw retrieval of actives or retrieval of chemotypes. As the latter is more important in VS experiments, we recommend that VS evaluations explicitly measure chemotype retrieval using the CA method.

Finally, we strongly suggest that authors of new VS methods provide the ranked output of their validation experiments, along with the structures of all of the actives and the cluster membership information (if any), to enable the calculation of enrichment metrics and facilitate comparisons between studies. To encourage this effort, we have made software freely available to calculate the metrics discussed in this paper, both with and without chemotype corrections.

ACKNOWLEDGMENT

We thank Dr. Andy Vinter and Dr. Timothy Cheeseright of Cresset BMD for helpful discussions and for proof-reading this manuscript.

Supporting Information Available: Perl code for calculating all of the metrics mentioned in this paper and program for generating and analyzing the synthetic data sets from Figure 1 and Table 2. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) *Virtual Screening in Drug Discovery*; Alvarez, J., Shoichet, B., Eds.; CRC Press: Boca Raton, 2005.
(2) Willett, P. Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.
(3) Coupez, B.; Lewis, R. A. Docking and scoring - Theoretically easy, practically impossible. *Curr. Med. Chem.* **2006**, *13*, 2995–3003.
(4) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 235–249.
(5) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of Automated Docking Programs as Virtual Screening Tools. *J. Med. Chem.* **2005**, *48*, 962–976.
(6) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model.* **2006**, *46*, 401–415.
(7) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74–82.

(8) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(9) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580–594.

(10) Pan, Y. P.; Huang, N.; Cho, S.; MacKerell, A. D. Consideration of Molecular Weight during Compound Selection in Virtual Target-Based Database Screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267–272.

(11) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; 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.

(12) Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369–1375.

(13) Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 325–332.

(14) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179–190.

(15) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.

(16) Jain, A. N. Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201–212.

(17) Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energies for Protein-Ligand Binding. *J. Med. Chem.* **2006**, *49*, 5880–5884.

(18) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(19) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118–127.

(20) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.

(21) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395–1406.

(22) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the Early Recognition Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(23) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 529–536.

(24) Pérez-Nueno, V. I.; Ritchie, D. W.; Rabal, O.; Pascual, R.; Borrell, J. I.; Teixidó, J. Comparison of Ligand-Based and Receptor-Based Virtual Screening of HIV Entry Inhibitors for the CXCR4 and CCR5 Receptors Using 3D Ligand Shape Matching and Ligand-Receptor Docking. *J. Chem. Inf. Model.* **2008**, *48*, 509–533.

(25) Moffat, K.; Gillet, V. J.; Whittle, M.; Bravi, G.; Leach, A. R. A Comparison of Field-Based Similarity Searching Methods: CatShape, FBSS, and ROCS. *J. Chem. Inf. Model.* **2008**, *48*, 719–729.

(26) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141–146.

(27) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.

(28) Nicholls, A. What do we know and when do we know it. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.

(29) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.

(30) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(31) Irwin, J. J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193–199.

(32) *Life's Grandeur: The Spread of Excellence from Plato to Darwin*; Gould, S. J. Vintage Press: Boca Raton, 1997.

CI8003978