Assessing the Discriminatory Power of Scoring Functions for Virtual Screening

Markus H. J. Seifert*

4SC AG, Am Klopferspitz 19a, D-82152 Martinsried, Germany

Received January 24, 2006

The efficiency of scoring functions for hit identification is usually quantified in terms of enrichment factors and enrichment curves. Close inspection of simulated and real score distributions from virtual screening, however, suggests that 'analysis of variance' (ANOVA) is a more reliable method for assessing their performance. Using ANOVA to quantify the discriminatory power of scoring functions with respect to ligands, decoys, and a reproducible reference database has the potential to facilitate the advancement of scoring functions significantly.

INTRODUCTION

Virtual screening is an established method to identify hit molecules that serve as starting points for medicinal chemistry. A large number of studies demonstrate that virtual screening of molecular databases is able to enrich hit molecules in ranking lists of relatively small size, although an obvious target dependence exists.[1,2] Numerous studies compare the performance of scoring functions and docking programs used for virtual screening on various target proteins.[3−9] To obtain a high throughput for virtual screening (vHTS) simple, mostly empirical scoring functions are used to estimate the binding affinity of the screened molecules.[10] Interestingly, in most cases no significant correlation of experimental binding affinities and in silico scores is found for the virtual screening hit molecules, giving rise to the question why virtual screening obviously works pretty well even though the scoring functions apparently fail to predict affinity.

From a theoretical point of view, correlation is a measure suitable for two continuous scales, so-called ratio scales. The results of virtual screening are usually verified experimentally. However, in most cases only the affinity (IC50, K_i) of true hit molecules is determined but not of weak-binding or nonbinding false-positives (so-called decoys[11]). Most of the biochemical assays do not even permit the measurement of weak binding constants, with the exception of biophysical methods such as e.g. NMR, isothermal calorimetry, and surface plasmon resonance. Since no defined affinities exist for the decoy molecules, the available affinity data are basically known on an ordinal scale, i.e., one knows about the assignment of molecules to two groups: ligands and decoys. It is therefore reasonable to evaluate methods for the analysis of virtual screening that are able to handle an ordinal scale of affinity, although one piece of data is disregarded in this type of analysis: the actual affinity data of the ligands. A technique that is able to work on ordinal data is analysis of variance (ANOVA). ANOVA is well established in social science, psychology, and clinical sciences for discriminating, e.g., different treatments of patients.[12] Additionally, ANOVA is simple to implement e.g. in spreadsheets and requires no time-consuming computations even for large data sets.

Here we develop a simple theoretical model that allows for the exploration of the relationship between enrichment curves, the underlying score distributions, and the discriminatory power of scoring functions. Krumrine et al.[13] and Klon et al.[14] had a look at score distributions before. But here the distributions are examined more in detail. Warren et al. state that neither enrichment nor lead identification alone are sufficient for optimizing algorithm performance during development.[1] Correlation analysis is hampered by the narrow range of affinity data usually at hand. Cole et al.[15] suggested using an analysis of variance (ANOVA) for establishing a statistically significant variation when analyzing various protein−ligand docking programs. Here we would like to extend this idea and propose ANOVA as a general methodology to measure the performance of virtual screening. First, the η^2 and F coefficients from ANOVA will be introduced in order to measure the discriminatory power and the signal-to-noise ratio of scoring algorithms for ligands and decoys. Second, it will be demonstrated by in silico experiments that η^2 is less artifact-prone and therefore more suitable for algorithm development compared to enrichment plots and enrichment factors, which in turn are more diagnostic for the practical performance of virtual screening. Third, a practical example will be given by analyzing the results of virtual screening for inhibitors of human dihydroorotate dehydrogenase (DHODH), which is an important target for treating rheumatoid arthritis. Next, an explanation is given for the common observation that in silico scores and experimentally determined affinities do not correlate significantly. Finally, some factors are described that are important in obtaining reliable results, and a procedure for selecting a reference database is suggested.

EXPERIMENTAL SECTION

Simulation of Enrichment Diagrams. The distribution of scores (s) for ligands and decoys are modeled as Gaussian distributions with means $\bar{S}_L$ and $\bar{S}_D$ and standard deviations of σ_L and σ_D. The Gaussian distributions G(s) are calculated using octave[16] and binned into 400 bins b_i over a range of

* Corresponding author phone: +49-89-700763-0; fax: +49-89-700763-29; e-mail: markus.seifert@4sc.com.

10.1021/ci060027n CCC: $33.50
Published on Web 04/26/2006

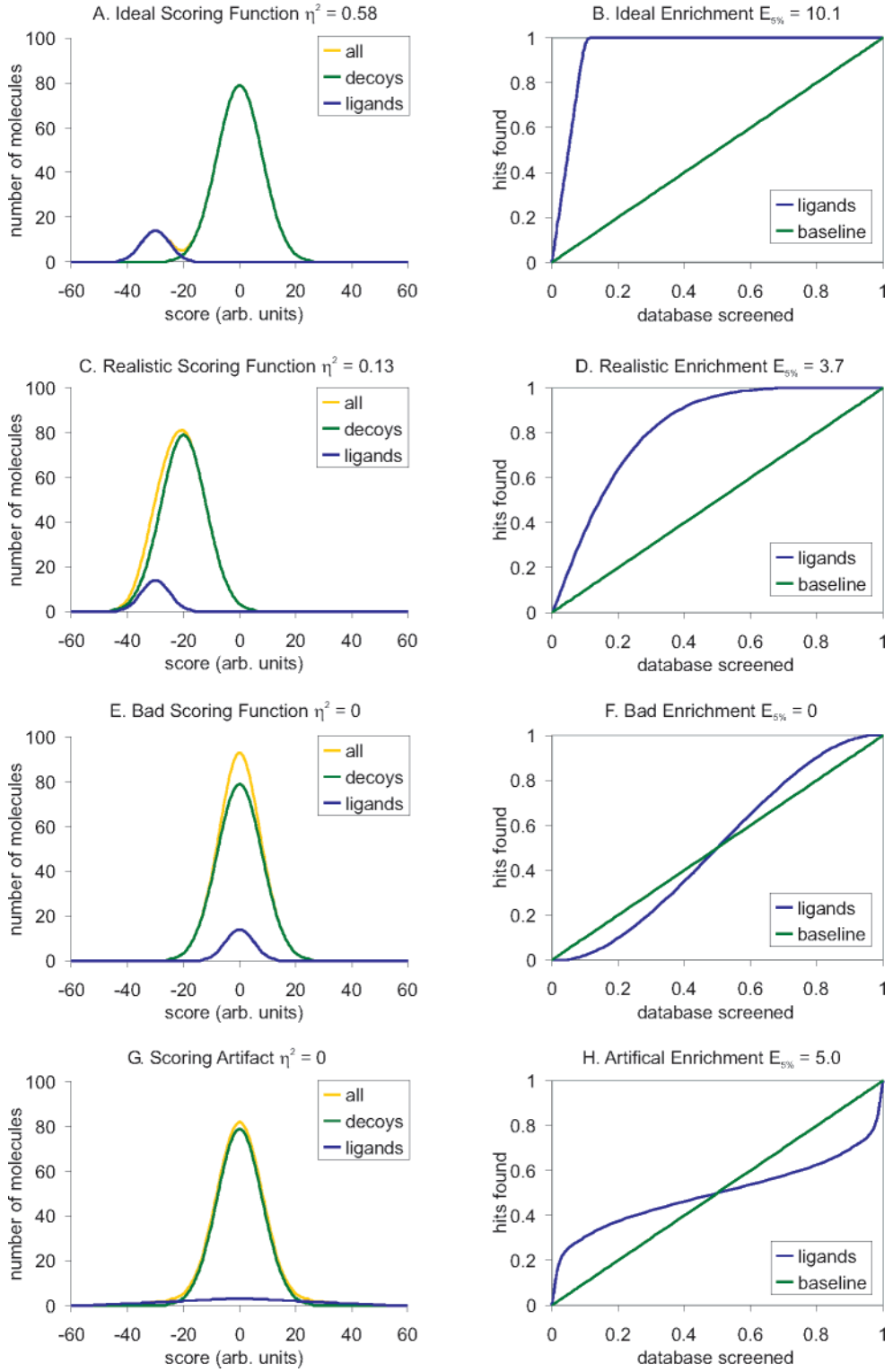

Figure 1. Relationship between score distribution functions (SDF, or 'score spectra') and enrichment curves. Lower scores indicate higher affinity. Mean value and standard deviation of the ligand and decoy distributions directly determine the enrichment diagram and enrichment factor. An example of ideal (A, B), realistic (C, D), bad (E, F), and artificially good (G, H) scoring and enrichment is given. The baseline in B, D, F, and H indicates the enrichment by pure chance.

scores from −80 up to +80, without loss of generality. Integrating the score distribution function of ligand and decoy molecules and rescaling the axes in order to plot the percentage of databases already screened (x) vs the percentage of ligands already found (y) leads to the well-known enrichment curves $E(x)$. The enrichment factor EF is given by $E(x)/x$. Some fixed parameters—without loss of generality—are used for this model: the percentage of the screened database used for calculating the enrichment factor (5%), the total number of molecules in the database (4400), and the ratio of ligands to decoys in the database (10%). The distributions and enrichments shown in Figure 1 have been computed with the parameters given in Table 1.

Analysis of Variance. Analysis of variance (ANOVA) is applied to determine the proportion of the variability within the observed or simulated scores (s) that is due to the scoring

Table 1. Mean Score Values and Standard Deviations Used for Generating the Gaussians Shown in Figure 1

	ideal	realistic	bad	artifact
$\bar{S}_L$	−30	−30	0	0
$\bar{S}_D$	0	−20	0	0
σ_L	5	5	5	25
σ_D	8	8	8	8

method itself as opposed to variability caused by random errors. The null hypothesis is that there is no difference in scores between ligands and decoys. The rejection of the null hypothesis indicates that the scoring method is able to discriminate between ligands and decoys on a specific significance level. The independent variable for this analysis is the assignment to 2 groups, either ligands or decoys ($p = 2$, $i = 1 \ldots p$). N and n_i denote the total number of molecules and the number of molecules per group, respectively. The analysis proceeds as follows:[12]

1. Compute

$$S_1 = \frac{1}{N}\left(\sum_{i=1}^{p}\sum_{m=1}^{n_i} s_{mi}\right)^2,\ S_2 = \sum_{i=1}^{p}\sum_{m=1}^{n_i} {s_{mi}}^2,\ \text{and}\ S_3 = \sum_{i=1}^{p}\frac{1}{n_i}\left(\sum_{m=1}^{n_i} s_{mi}\right)^2$$

2. Compute the total sum of the squares $S_{tot} = S_2 - S_1$, which measures the overall variability, $S_{treat} = S_3 - S_1$, which measures the variability due to the scoring method, and $S_{error} = S_2 - S_3$, which is the variability due to random errors.

3. The degrees of freedom associated with these sums of squares are $df_{tot} = N - 1$, $df_{treat} = p - 1$, and $df_{error} = N - p$, respectively.

4. The total variance, treatment variance, and error variance are $\sigma^2_{tot} = S_{tot}/df_{tot}$, $\sigma^2_{treat} = S_{treat}/df_{treat}$, and $\sigma^2_{error} = S_{error}/df_{error}$, respectively.

5. Measures for the discrimination of the two groups by a scoring method are given by the coefficients $\eta^2 = S_{treat}/S_{tot}$ and $\omega^2 = S_{treat} - (p-1)\cdot\sigma^2_{error}/S_{tot} + \sigma^2_{error}$. η^2 is the fraction of variance within the actual data set that is explained by the scoring function. ω^2 is an estimate for the explained variance within the population (i.e. the complete database) from which the actual data are derived.

6. The strength of the effect is measured by $\epsilon = \sqrt{\eta^2/1-\eta^2}$. ϵ is related to the difference of the mean values divided by the variance within the data. A weak, medium, and strong effect is characterized by $\epsilon \approx 0.1, \eta^2 \approx 1\%$; $\epsilon \approx 0.25, \eta^2 \approx 6\%$; and $\epsilon \approx 0.4, \eta^2 \approx 14\%$, respectively.

7. The significance has to be verified by $F = \sigma^2_{treat}/\sigma^2_{error}$ in comparison to the F distribution. Due to this definition, F is a measure for the signal-to-noise ratio of the virtual screening.

Analysis of Score Distributions of DHODH Inhibitors. Virtual screening, biochemical testing, and medicinal chemistry of DHODH inhibitors have been described in detail elsewhere.[17−20] Two hundred forty-four of the active compounds (ligands) with IC50 values ranging from below 1 nM up to ~1 mM and 240 less active or inactive compounds (experimentally verified decoys) have been selected for ANOVA analysis. This series of molecules includes cyclic aliphatic dicarboxylic acids, in which one carboxylic group is amide bound with an aromatic biphenyl aniline. Two independent random subsets (subset 1 and 2, $N = 5000$) were selected from an in-house database of 5 million commercially available compounds (molecular weight MW: mean 444 D, rmsd 105 D; topological polar surface area TPSA:[21] mean 84 Å^2, rmsd 31 Å^2; number of H-bond acceptors ACC: mean 4, rmsd 1.8; number of H-bond donors DON: mean 1.4, rmsd 1.0). These random selections act as reference databases and are used to determine the enrichment curves. Further independent random subsets were extracted by narrowing the range of descriptors (MW 250-500 D, TPSA 21-133, ACC 1-5, and DON 1-4) and additionally limiting the compounds to aromatic compounds (subset 3, $N = 2000$) and acidic compounds with at least one aromatic moiety (subset 4, $N = 500$). These descriptor ranges reflect the properties present in the series of active compounds. The random subsets were seeded with the known ligands and decoys in order to create small databases for the analysis of score distributions. Two hundred forty-three ligands, 224 decoys, and 4166, 4133, 1960, and 460 molecules of subsets 1−4, respectively, were docked successfully into the active site of DHODH (PDB entry 1D3G) using ProPose[22,23] with application of Böhm's scoring function.[24,25] In general, if docking into the active site was not successful—i.e., no base fragment placement or incremental construction is possible inside the active site—no score is given by ProPose, and the molecule was disregarded in the analysis. The ligands, decoys, and subset 1 were aligned onto brequinar—the ligand cocrystallized in PDB 1D3G—using the ProPose alignment functionality.[23] The target dependence was investigated by docking subset 1 into fatty acid binding protein (FABP, PDB entry 1HMT), cyclooxygenase-2 (COX2, PDB entry 1CX2), herpes simplex virus thymidine kinase (TK, PDB entry 1KIM), protein kinase A (PKA, PDB entry 1STC), and thrombin (THR, PDB entry 1OYT) using identical docking parameters. The dependence of the score distributions on molecular weight were investigated by docking 7 independent subsets of size 5000 into DHODH: subset 50 (MW 50−100 D), subset 100 (MW 100−200 D), subset 200 (MW 200−300 D), subset 300 (MW 300−400 D), subset 400 (MW 400−500 D), subset 500 (MW 500−600 D), and subset 600 (MW 600−700 D).

RESULTS AND DISCUSSION

Analysis of Variance. Figure 1 shows the simulated effect of variations in the score distribution functions (SDF, or 'score spectra') for ligands and decoys on the resulting enrichment diagram and the enrichment factor. An almost ideal scoring function is expected to generate two, clearly distinct distributions of scores for active and inactive molecules (Figure 1A), which is reflected by the ANOVA coefficient $\eta^2 = 58\%$. η^2 is a direct measure of the variance within the data set that is due to the discrimination of ligands from decoys—which is caused by our scoring function—in relation to the total variance. $\eta^2 = 1$ corresponds to two well-separated, infinitely narrow peaks for ligands and decoys, where all of the variance is explained by the scoring function. In contrast, $\eta^2 = 0$ corresponds to two peaks with identical mean value, where the variance is only due to database composition. Therefore the 'explained variance' η^2 is a measure for the 'discriminatory power' of the scoring function, both of which will be used as synonyms in the

following paragraphs.[26] Integration of the score distributions directly results in the well-known enrichment diagrams.[27] The enrichment diagram here shows a nearly perfect enrichment with an enrichment factor of $E_{5\%} = 10.1$ (Figure 1B). A more realistic case is shown in Figure 1C,D: the distributions of scores overlap, but still a significant proportion of the variance (discriminatory power $\eta^2 = 13\%$) is explained by scoring and the enrichment is acceptable $E_{5\%} = 3.7$. Many examples of this curve shape are depicted in refs 1 and 14. Figure 1E,F depicts the effect of an incorrect scoring function: there is no discrimination between the mean score of the ligands and of the decoys, resulting in an enrichment of $E_{5\%} = 0$. This case is observed—for example—by Warren et al. (ref 1, Figure 8, Fred, ScreenScore)[1] and Klon et al. (ref 14, Figure 7C, Dock/Bayes Model).[14] Figure 1G,H shows a potential artifact in the interpretation of the enrichment factor: here, the scoring function clearly is not able to discriminate between ligands and decoys. But due to the much larger variance of ligand scores, an artificial enrichment is produced ($E_{5\%} = 5.0$). A somewhat similar curve shape is shown by Warren et al. (ref 1, Figure 6B, LigFit, Dreiding, DScore).[1] This effect may be tolerable for practical applications; however, for developing scoring functions and algorithms this would point to the completely wrong direction. Notably, the discriminatory power η^2 is able to reveal the truth, that this scoring function is defective ($\eta^2 = 0$). Therefore it may be advisable not only to look for good enrichment factors but also additionally to check the results by an analysis of variance. Warren et al. found that—for some target/docking program combinations—good enrichments are obtained, while the cognate compounds were poorly docked.[1] One possible explanation for this troubling result is that poor docking might have a strong impact on the variance of the scores, thereby generating an artificial enrichment.

For reasons of clearness, the simulations shown in Figure 1 assume that all active molecules are predicted nearly equally well, resulting in a single peak of the distribution function. More complicated situations may occur if scoring functions discriminate between, e.g. different ligand classes. Such effects can be incorporated easily into the simulations by adding multiple peaks for the active molecules. The simulations shown in Figure 1 additionally assume that the rank ordering of the molecules is predicted correctly by the scoring function, i.e., ligands exhibit—on average—lower scores than decoys. If the opposite is true, i.e., the scoring function assigns higher scores to ligands compared to decoys, the scoring function would effectively deplete the ligands within the ranking list. However, η^2 still would be larger than zero in this case. A large η^2 is therefore a sign of a good scoring function only if the rank ordering is correct on average. This has to be checked before interpreting different η^2 values. A practical workaround for this case is to report $\mathrm{sign}(\bar{S}_D - \bar{S}_L)\cdot\eta^2$, giving positive values for good scoring functions and negative values for bad scoring functions. The assumption of a Gaussian distribution here does not limit the applicability of the above considerations: the same will be true for any more or less symmetric, two-tailed distribution. However, the Gaussian distribution is not far from reality, as will be shown in the following paragraph. ANOVA itself does not depend on the specific distribution function: the dissection of the total square sum S_{tot} into S_{treat} and S_{error} is not tied to any preconditions.[12] The test for significance using the F-distribution is linked to some preconditions: the components contributing to the noise have to be normally distributed, independent and similar in variance. However, when large numbers of ligands and decoys ($N > 100$) are used, the effect of disobeying some preconditions is limited and ANOVA is regarded as a robust method.[12] Only for small numbers of molecules per group ($n_i < 10$) and very different numbers of group members special precautions have to be taken, for example by using nonparametric methods such as the Kruskal-Wallis test.[28]

Analysis of DHODH Virtual Screening. The analysis of virtual screening data for DHODH revealed that the correlation of the scores of the 243 ligands—which were docked successfully by ProPose[22,23] with application of Böhm's scoring function[24,25]—with their experimentally determined IC50 values is weak (correlation coefficient $R = 0.2-0.3$ for different subsets of the data; see Figure 2). Due to the lack of unambiguous correlation, we have to ask whether the scoring function is able to predict affinity at all. Compounds with good affinities, however, were found using virtual screening. This apparent discrepancy has to be clarified. ANOVA was used to analyze the score distributions resulting from docking the molecules into the active site of DHODH. The score distributions of the random subsets 1 and 2 are very similar to a Gaussian distribution (Figure 3A). Therefore our assumption in the theoretical part was correct. To check for artifacts due to molecules completely unacceptable for DHODH inhibition, more independent random subsets (subset 3 and 4) were selected which comprise molecules similar to known DHODH inhibitors with respect to size and polarity.[30] Interestingly, the variance and the mean value of docking scores does not change significantly for molecules which are more similar to real inhibitors, as measured by molecular descriptors (Figure 3A). This result suggests that descriptor distances are not directly related to docking results, at least for simple descriptors derived from a molecular 2D structure.

Figure 3B shows the score distribution of ligands and decoys (i.e. experimentally verified false positives) in comparison to independent random subsets 1 and 2 from our in-house database. The ligand scores are more asymmetric and have a significantly lower mean value with a long tail in the direction of higher scores. The decoy scores have a slightly higher mean value and a more symmetric distribution compared to the ligands. These observations were quantified using ANOVA (Table 2). ANOVA reveals that even the relatively simple, empirical scoring function of Böhm is able to discriminate significantly not only between random molecules and real ligands (discriminatory power $\eta^2 = 9.3\%$, signal-to-noise ratio $F = 452$) but also additionally between ligands and experimentally verified decoys (discriminatory power $\eta^2 = 15.8\%$, signal-to-noise ratio $F = 87.4$). Despite the strong overlap of the score distributions of ligands, decoys, and random molecules the discriminatory power translates into an acceptable enrichment of ligands ($E_{5\%} = 5.7$, Figure 3C). Please note that the enrichment diagram obscures the fine structure of the score distributions, making it less informative than the corresponding score spectra. This result clearly shows that a significant correlation between the experimental binding affinities of the active molecules and their in silico scores is not necessary for obtaining useful

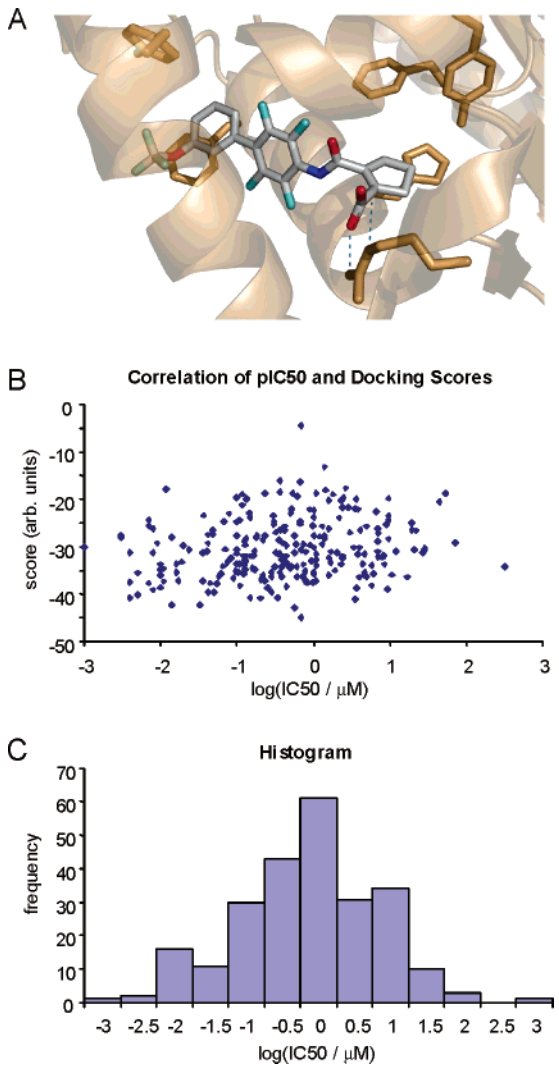

Figure 2. Example of the ProPose binding mode of a cyclopentene inhibitor (in element colors) in the active site of human DHODH (A). The carboxylic group is bound by a salt bridge to Arg-136, and the biphenyl moiety reaches into a hydrophobic pocket. This figure was created by PyMol.[29] A nonsignificant correlation is found for ProPose docking scores and logarithmic IC50 values for DHODH ligands (B). The correlation coefficient $R = 0.20$ is not significant for the set of active molecules ($N = 243$, see Experimental Section). Some subsets of these data, which were measured under the same assay conditions, give rise to correlations in the range of $R = 0.2-0.3$. The histogram of log(IC50) values shows a symmetric distribution with a mean activity of ca. 1 μM (C).

results from virtual screening. Due to the small range of scores covered by the active molecules—only 50% in comparison with the overall distribution—and the 'noisy' scoring functions it is rather unlikely to find significant correlations. For example, Ferrara et al. found no correlation for several scoring functions applied to several targets, except for serine and metalloproteases.[6] However, this is more likely due to a strong correlation of experimental binding affinities with molecular weight which is also present in most scoring functions.

Dependence on Target. The target dependence of the score distribution is illustrated in Figure 3D. Docking into fatty acid binding protein (FABP, 4575 successfully docked molecules) provides a negative control for ANOVA: the active site of FABP binds negatively charged, hydrophobic ligands, similar to DHODH. But there is only a very weak discrimination between DHODH ligands and random molecules ($\eta^2 = 2\%$, $F = 103$) and between DHODH ligands and DHODH decoys ($\eta^2 = 4\%$, $F = 16.8$), as expected (Figure 3E). The enrichment curve (Figure 3F) indicates a similar result. To investigate the variation of the score distribution with the target, subset 1 was docked into thymidine kinase (TK, 1450 successfully docked molecules), cyclooxygenase-2 (Cox2, 3797 successfully docked molecules), protein kinase A (PKA, 4835 successfully docked molecules), and thrombin (THR, 4820 successfully docked molecules). Surprisingly, the mean value and the variance of the scores are roughly constant for DHODH, FABP, and Cox2, despite large differences in the active site of the target proteins. Larger deviations are observed for TK, PKA, and thrombin. But the score distributions are still of Gaussian shape. Interestingly, Godden et al. showed that the score distributions of DOCK are positively skewed if the binding site was small.[31] They found a positive skew for PKA, whereas the ProPose/Böhm scores do not exhibit such a skew for PKA. The active sites for ProPose docking, however, were designed such that ligand binding outside the active site was minimized, which was probably the origin of the long tail of bad scores in the DOCK score distribution for small active sites. The TK binding site is particularly small, leading to a large number of failed dockings, a more rugged score distribution, and more positive scores compared to the other targets. The PKA and thrombin scores, in contrast, are shifted toward the more negative direction. In summary, the deviations with respect to Gaussian shape, mean value, and variance are rather limited for Böhm's scoring function, if only successful dockings are considered. However, even larger deviations would be no limitation to ANOVA itself, since the test for significance can be performed by the Kruskal-Wallis test as well.

Figure 4 shows another interesting case study: using ligand−ligand alignment instead of protein−ligand docking for virtual screening results in a strong discrimination between two series of molecules. Molecules containing the biphenyl motif are aligned well onto brequinar, another DHODH inhibitor.[32] Molecules with a benzyloxy-phenyl moiety, however, receive an inferior score since they obviously do not fit onto brequinar. Similar results may occur even for protein−ligand docking when the scoring function favors distinct features within the molecules. The two peaks corresponding to these series of molecules are well separated, allowing for the application of ANOVA for each single peak separately. The biphenyl compounds achieve an excellent discriminatory power of $\eta^2 = 28\%$ ($F = 2019$), whereas the benzyloxy-phenyl compounds strongly overlap with the 'noise' molecules ($\eta^2 = 1\%$, $F = 46$). This pronounced discrimination of chemical substructures leads, on one hand, to a strong enrichment of such compounds (see Figure 4B) but, on the other hand, reduces the applicability of pure ligand−ligand alignment for scaffold hopping.[23]

Dependence on Group Assignment. ANOVA relies on the assignment of molecules to distinct groups. This assignment is ambiguous in some cases: for example, molecules with affinities in the range of 100 μM may be regarded as ligands or decoys, depending on the point of view. To evaluate the sensitivity of the discriminatory power η^2 to potential misclassifications of molecules, the assignment of random subsets of DHODH ligands and decoys was inter-

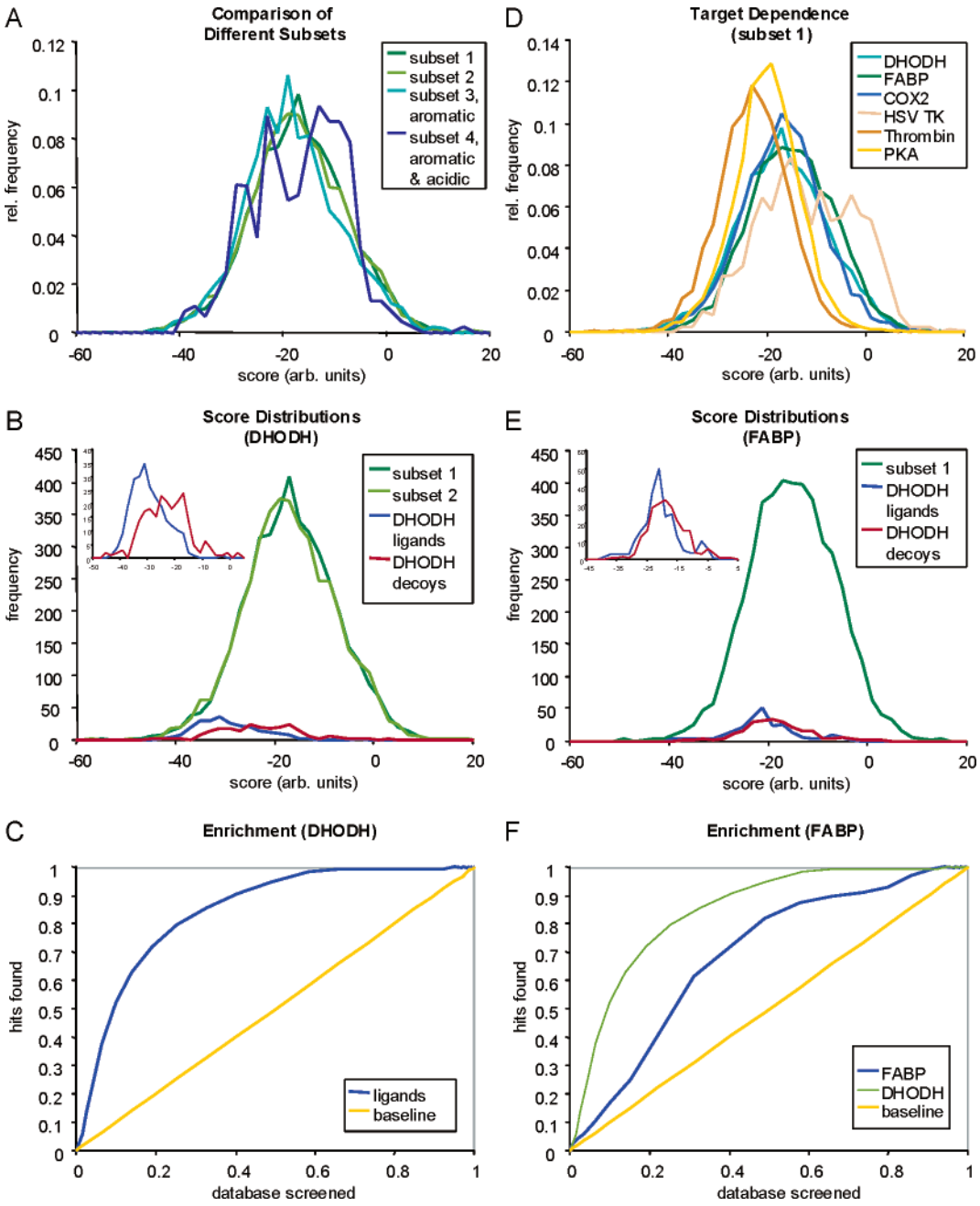

Figure 3. Different subsets of a large molecular database exhibit a similar variance of docking scores (A). Subsets 1 and 2 ($N = 4166$ and 4133, respectively) are randomly selected from a database of commercially available organic compounds. Subset 3 ($N = 1960$) is limited to aromatic compounds similar to known ligands with respect to MW, TPSA, and donor/acceptor moieties. Subset 4 ($N = 460$) only contains acidic molecules with at least one aromatic moiety. The score spectra (B) of DHODH ligands, experimentally verified DHODH decoys, and 2 sets of random decoy molecules are compared to the enrichment (C) obtained by docking into DHODH and—as negative control—FABP (D,E,F). The random decoy molecules (subset 1) were additionally docked into PKA, thrombin, HSV TK, and Cox2 (D).

changed. This evaluation showed that even if 10% of the molecules are assigned wrongly, the results are still highly significant. The discriminatory power η^2 for ligands and random molecules (and for ligands and decoys) drops from 9.3% to 8.7% (15.8% to 11.7%) on average, but the results remain highly significant with $F > 410$ (45) in any case. The sensitivity toward misclassifications obviously depends on the sample size, i.e., the number of molecules. Considering the typical number of vHTS molecules being evaluated in vitro—which is on the order of 10^2—the analysis of variance represents a rather robust method for measuring the discriminatory power of scoring functions.

Selection of a Reproducible Reference Database. To measure the discriminatory power η^2 in a reliable manner, the variance of scores within the reference database consisting of random molecules has to be calibrated: for example, score distributions derived from additive scoring functions quite generally seem to exhibit a trivial dependence on molecular weight MW (see Figure 5A and ref 30, Figure 3A showing a plot of ChemScore vs heavy atom count HAC, which is highly correlated to MW). The score distributions (Figure 5A) and the ANOVA parameters (Figure 5B), however, converge when a high upper limit of molecular weight is applied. This observation can be explained easily

Table 2. ANOVA Results for DHODH Score Distributions[a]

	subset 1 vs ligands	subset 1 vs decoys	ligands vs decoys
S_{tot}	389548	367251	29841
S_{treat}	36248	8172	4720
S_{error}	353300	359078	25120
df_{tot}	4408	4389	466
df_{treat}	1	1	1
df_{error}	4407	4388	465
σ^2_{tot}	88.4	83.7	64.0
σ^2_{treat}	36248.2	8172.7	4720.7
σ^2_{error}	80.2	81.8	54.0
η^2	0.093	0.022	0.158
ω^2	0.093	0.022	0.156
ϵ	0.320	0.151	0.433
F	452[b]	99.9[b]	87.4[b]

[a] Total sum of squares S_{tot}, treatment sum of squares S_{treat}, sum of squares of the errors S_{error}, total degrees of freedom df_{tot}, degrees of freedom of treatment df_{treat}, degrees of freedom of the errors df_{error}, associated variances σ^2_{tot}, σ^2_{treat}, and σ^2_{error}, the variance explained by the scoring function with respect to the sample data (η^2) and the population (ω^2), the strength of the effect ϵ, and the F-value for testing the significance are shown. [b] Indicates an error probability well below 0.1%.

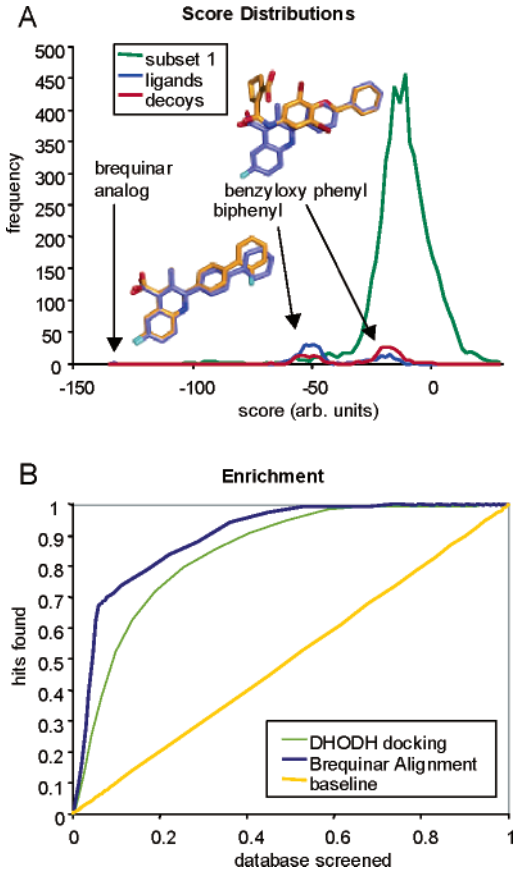

Figure 4. Alignment of subset 1, DHODH ligands, and decoys onto the native ligand of PDB structure 1D3G, brequinar, shows a strong discrimination between two different series of molecules: biphenyl and benzyloxy phenyl compounds (A). This discrimination can be anticipated by the corresponding bend of the enrichment curve (B). The score spectrum, however, is much more indicative.

by examination of the factors entering Böhm's scoring function: Böhm's scoring function is a weighted sum of the number of salt bridges, hydrogen bonds, aromatic and hydrophobic interactions, minus an entropy penalty depending on the number of rotatable bonds. All these factors correlate with the ligand's molecular weight (number of hydrogen bond acceptors $R^2 = 0.24$, number of aromatic

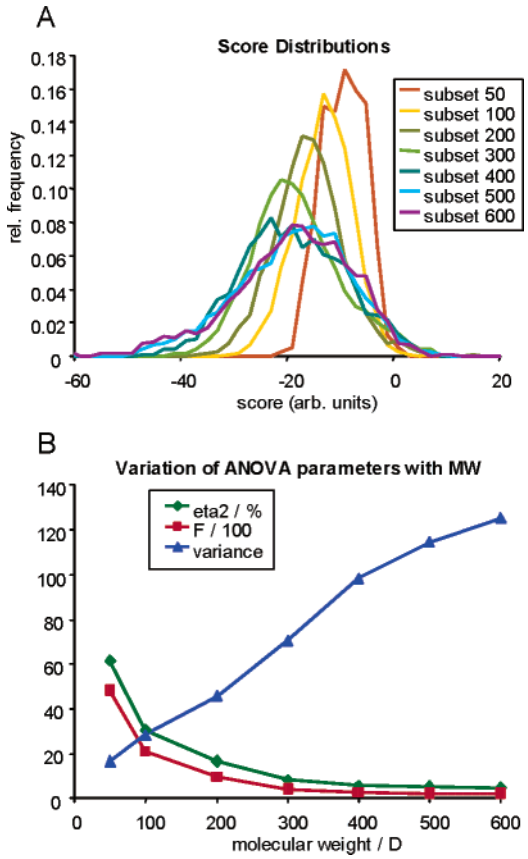

Figure 5. The range of molecular weights in the reference database strongly influences the ANOVA parameters. Sets containing only high molecular weight molecules exhibit a larger variance and lower mean value of scores in comparison to sets of low molecular weight compounds (A). A large variance within the reference database leads to a decrease in η^2 and F, which converge at high molecular weights (B).

atoms $R^2 = 0.33$, number of hydrophobic atoms $R^2 = 0.70$, number of rotatable bonds $R^2 = 0.42$, etc.), which leads—on average—to an increase of the score for heavier molecules. But the active site of a protein provides only a specific number of partners for the interactions. Once the average molecular weight of the ligands is large enough—i.e., the molecules contain a lot of diverse functional groups—to potentially saturate most of the receptor interactions no further decrease of the score is expected. This provides a plausible explanation for the convergence at high molecular weights: a sufficiently large reference database provides enough molecules with features to saturate the active sites of different targets. Since all empirical scoring functions—including potentials of mean force—are weighted sums of some properties, it is expected that MW-dependence is a relatively conserved feature of scoring functions. Since ANOVA parameters as well as enrichment curves crucially depend on the score distribution of the reference database, special attention has to be paid to the molecular weight distribution within the reference database.

Obviously, the publication of a set of, e.g. 5000, molecular structures would provide a standardized reference set for future investigations. It seems, however, that is not necessary. The databases of commercial suppliers—and others—seem to be rather consistent with respect to the distribution of molecular descriptors (see Figure 6). This consistency translates into comparable score distributions (Figure 7). The

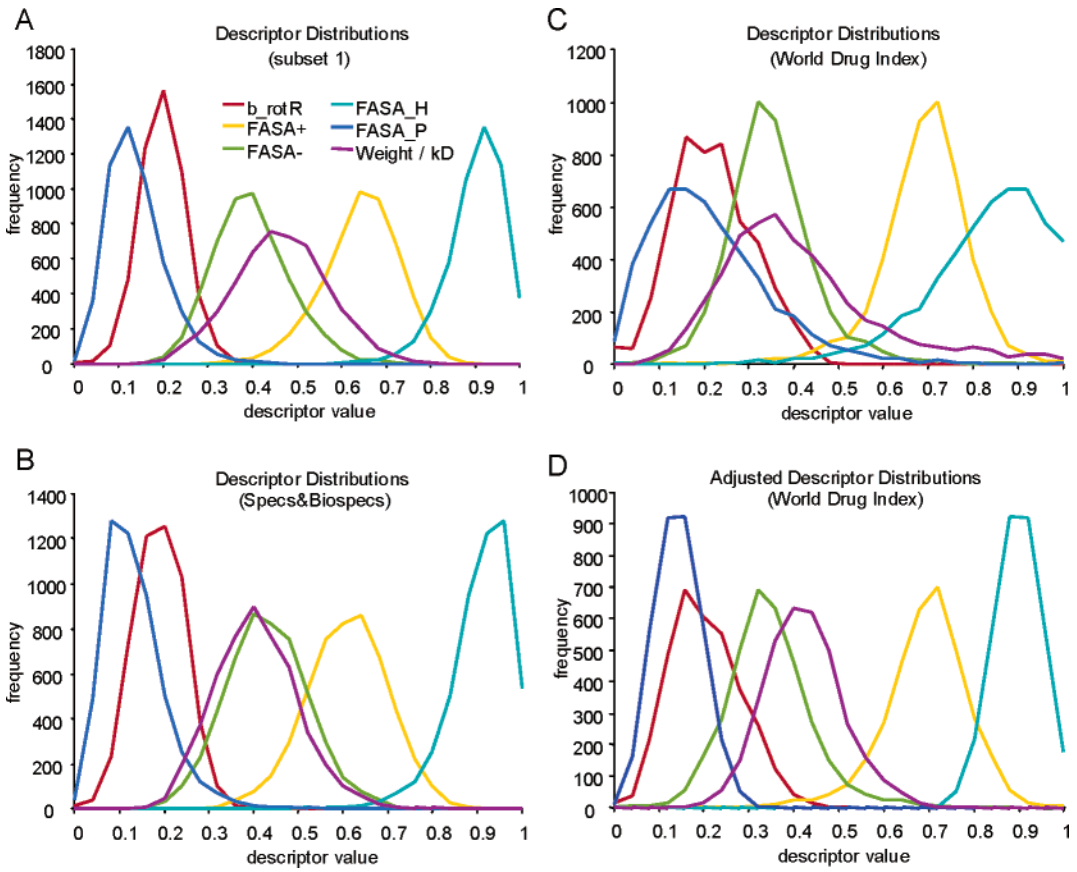

Figure 6. Example for the differences in descriptor distributions from different databases (subset 1 in comparison to 5000 molecules randomly selected from the Specs&Biospecs catalog and the World Drug Index). By adjusting the selection criterion with a Gaussian probability function the descriptor distribution can be modified to be similar to the reference subset 1. This is demonstrated using the WDI, where only MW and FASA_P were forced to be similar to subset 1.

discriminatory power η^2 of our scoring function with respect to DHODH ligands and subset 1, Specs&Biospecs, and the World Drug Index (WDI) now is 9.3% ($F = 452$), 7.7% ($F = 383$), and 15.4% ($F = 662$), respectively. The choice of descriptors for characterizing a database here is based on the fact that most molecular descriptors are more or less correlated to molecular weight (MW). For example, out of 191 2D and 3Di descriptors computed by MOE (Molecular Operating Environment, Chemical Computing Group, Montreal, Canada, http://www.chemcomp.com/) only very few do not correlate with molecular weight ($R < 0.1$): ratio of rotatable (single) bonds to total number of bonds (b_1rotR and b_rotR), density, fractional accessible surface areas (FASA ±/H/P), and the binned PEOE_VSA and Q_VSA descriptors. This leads to the selection of the following descriptors for reference selection: MW, b_rotR, FASA+, and FASA_P. This set of descriptors describes the most important molecular features: size, conformational flexibility, and polarity. FASA+/− and FASA_H/_P sum up to one; therefore, one descriptor of both pairs is redundant. MW, b_rotR, FASA+, and FASA_P are not cross-correlated ($R^2 < 0.07$).

There may be databases with significantly different descriptor distributions. By modifying the selection criterion, however, selections from such databases can be created that exhibit the desired descriptor distributions: usually random numbers R_i are used for selecting molecules M_i from a large molecular database. If R_i is smaller than a constant C—which reflects the percentage of molecules which should be extracted—molecule M_i is selected. This selection criterion is now modified by a factor derived from a Gaussian probability distribution with the desired mean μ and standard deviation σ of the target descriptor distribution. If $R_i < C \cdot \exp(-(x - \mu)^2/2\sigma^2)$ then molecule M_i is extracted. For example, adjusting the MW and FASA_P distribution of the selection from the WDI to 444 ± 104 and 0.118 ± 0.064, respectively, results in descriptor (see Figure 6D) and score distributions (see Figure 7A) more similar to subset 1, with the exception of a small region around scores −40 to −35. Accordingly, the ANOVA parameter η^2 stays rather constant ($\eta^2 = 16.5\%$), whereas F now is closer to the values obtained for subset 1 ($F = 523$). The adjustment of the score distribution is therefore not ideal, which is however an artifact due to the much smaller size of the WDI (ca. 58k) compared to, e.g., the Specs&Biospecs catalog (ca. 200k). Once a common standard of reference descriptor means and standard deviations is agreed on, everybody is able to generate comparable reference data sets, at least up to an accuracy of some percent. This would increase the comparability of results when evaluating different docking programs and scoring functions. A useful set of descriptor means and standard deviations is given in the following: MW 400 ± 100; b_rotR 0.2 ± 0.05; FASA+ 0.6 ± 0.1; and FASA_P 0.12 ± 0.06.

SUMMARY

In summary, it has been shown that an ambiguous correlation between experimental affinities and in silico

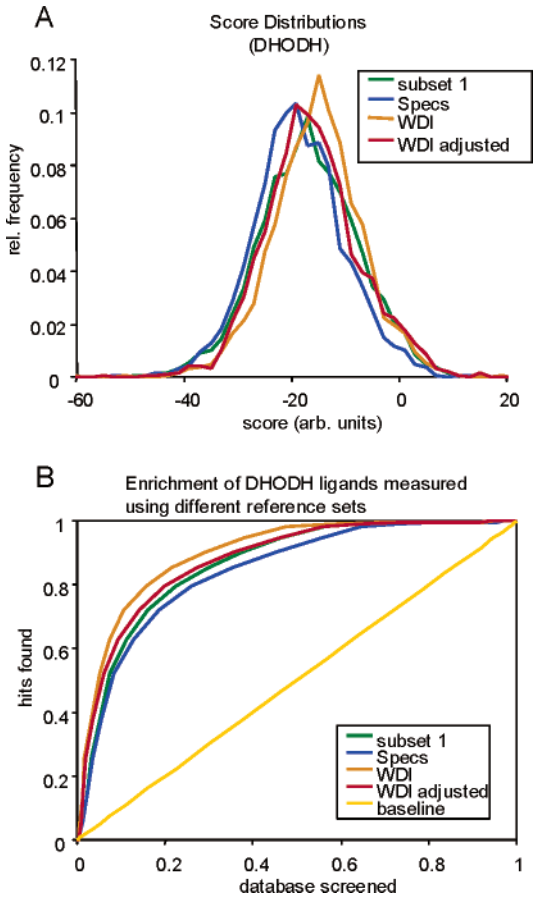

Figure 7. The DHODH score distributions resulting from the reference databases characterized in Figure 6. The score distributions are surprisingly similar, despite the differences in the descriptor distributions. After adjustment of the selection from the WDI, only minor differences remain compared to subset 1. This is reflected in similar enrichment curves. Please note that the curves of the adjusted WDI reference and of subset 1 are almost identical, with the exception of a small region around scores −40 to −35.

scores is not necessarily a sign of bad scoring function for virtual screening. Such ambiguous correlations are not only caused by imperfect scoring functions but also additionally are an artifact due to a small within-group variance, i.e., the molecules with known activities cover only a small range of score values compared to the range of possible scores. By taking into account the information about decoys—without having to worry about their 'affinities'—ANOVA of the score distribution functions (SDFs) is a much more promising method for evaluating scoring functions compared to simple correlation analysis or enrichment curves: (i) In contrast to enrichments factors and enrichment curves, ANOVA not only is able to differentiate the effect of scoring from the "noise" due to database composition quantitatively but also additionally is able to test this effect for significance. (ii) The discriminatory power η^2 is independent of any arbitrary cutoff, in contrast to enrichment factors. (iii) It detects artificial enrichments due to defective scoring functions, (iv) is easy to implement and compute, (v) can be extended easily to a larger number of groups, e.g. high-affinity ligands, medium-affinity ligands, and decoys, and (vi) is rather robust with respect to the misclassification of molecules, especially when larger data sets are available. Enrichment factors are still valuable for measuring the efficiency of practical applications of virtual screening. But for developing better scoring functions and for avoiding artificially good results, the analysis of variance of the SDFs seems to provide more reliable parameters which may then be used as objective functions for optimization.[33] The application of ANOVA may influence the weighting of parameters which are regarded to be important for successful virtual screening: for example shifting the focus from binding mode reproduction to sound statistical approaches may pave the way to identify the parameters which really separate the ligands from the decoys. It is too early to draw a definitive conclusion, but offering some more freedom of thought to scoring function designers may be beneficial.

Based on the presented results, a generic procedure to measure scoring function performance for small molecule virtual screening is suggested:

- Randomly select molecules (upper limit of MW ∼600 D or higher) from a large database of commercially available organic compounds (containing $>10^5$ unique compounds) and compile them in a reference database of reasonable size (>5000).
- Adjust the distribution of MW, b_rotR, FASA+, and FASA_P within the reference database, if necessary.
- Randomly select the desired number of reference compounds (e.g. 5000).
- Dock known ligands, known decoys, and the reference molecules into the target protein of interest.
- Apply ANOVA to quantify the discriminatory power η^2 and the signal-to-noise ratio F of the applied scoring function with respect to the scores of ligands, decoys, and reference molecules.
- A useful scoring function for a specific target will discriminate significantly between ligands and decoys and between ligands and random reference molecules.

The current arguments in favor of ANOVA (and nonparametric versions thereof) are largely based on results derived with Böhm's scoring function and several data available from literature. The general methodology, however, is applicable without any change to arbitrary scoring functions. But future investigations obviously have to examine the differences of other scoring functions with respect to their ANOVA parameters. All considerations presented above apply to the virtual screening in the hit identification phase. There is no doubt that more accurate scoring methods have to be developed for lead optimization, where a within-group correlation is an absolute must.

In conclusion, classical statistics provide a well-established machinery of methods to analyze the results of virtual screening. ANOVA in particular has the potential to increase the comparability of results reported in the literature and to facilitate the advancement of scoring functions.

Abbreviations. ANOVA, analysis of variance; Cox2, cyclooxygenase-2; DHODH, dihydroorotate dehydrogenase; FABP, fatty acid binding protein; FASA, fractional accessible surface area; HSV, herpes simplex virus; MW, molecular weight; PKA, protein kinase A; SDF, score distribution function; THR, thrombin; TK, thymidine kinase; TPSA, topological polar surface area; vHTS, virtual high-throughput screening.

ACKNOWLEDGMENT

I would like to thank 4SC's DHODH team for pushing forward this project; Daniel Vitt and Bernd Kramer for their

unwavering support; and Thomas Herz, Kristina Wolf, Matthias Busemann, and Jürgen Kraus for stimulating discussions. Special thanks go to Stefan Duschek for bringing η^2 to my attention.

REFERENCES AND NOTES

(1) 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.* **2005**, ASAP.

(2) 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*, 935−949.

(3) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56* (2), 235−249.

(4) Kellenberger, E.; Rodrigo, J.; Müller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57* (2), 225−242.

(5) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47* (3), 558−565.

(6) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks III, C. L. Assessing Scoring Functions for Protein−Ligand Interactions. *J. Med. Chem.* **2004**, *47* (12), 3032−3047.

(7) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model. (Online)* **2003**, *9* (1), 47−57.

(8) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44* (7), 1035−1042.

(9) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43* (25), 4759−4767.

(10) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed. Engl.* **2002**, *41* (15), 2644−2676.

(11) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for Docking. *J. Med. Chem.* **2005**, *48* (11), 3714−3728.

(12) Bortz, J. *Statistik*, 6th ed.; Springer Medizin Verlag: Heidelberg, 2005; pp 247−288.

(13) Krumrine, J. R.; Maynard, A. T.; Lerman, C. L. Statistical Tools for Virtual Screening. *J. Med. Chem.* **2005**, *48* (23), 7477−7481.

(14) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results. *J. Med. Chem.* **2004**, *47* (11), 2743−2749.

(15) Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Comparing protein−ligand docking programs is difficult. *Proteins* **2005**, *60* (3), 325−332.

(16) Eaton, J. W.; et al. Octave. A high-level language, primarily intended for numerical computations. http://www.octave.org.

(17) Baumgartner, R.; Walloschek, M.; Kralik, M.; Gotschlich, A.; Tasler, S.; Mies, J.; Leban, J. Dual binding mode of a novel series of DHODH inhibitors. *J. Med. Chem.* **2006**, *49* (4), 1239−1247.

(18) Leban, J.; Kralik, M.; Mies, J.; Baumgartner, R.; Gassen, M.; Tasler, S. Biphenyl-4-ylcarbamoyl thiophene carboxylic acids as potent DHODH inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16* (2), 267−370.

(19) Leban, J.; Kralik, M.; Mies, J.; Gassen, M.; Tentschert, K.; Baumgartner, R. SAR, species specificity, and cellular activity of cyclopentene dicarboxylic acid amides as DHODH inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *15* (21), 4854−4857.

(20) Leban, J.; Saeb, W.; Garcia, G.; Baumgartner, R.; Kramer, B. Discovery of a novel series of DHODH inhibitors by a docking procedure and QSAR refinement. *Bioorg. Med. Chem. Lett.* **2004**, *14* (1), 55−58.

(21) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(22) Seifert, M. H.; Schmitt, F.; Herz, T.; Kramer, B. ProPose: a docking engine based on a fully configurable protein-ligand interaction model. *J. Mol. Model.* **2004**, *10*, 342−357.

(23) Seifert, M. H. ProPose: steered virtual screening by simultaneous protein−ligand docking and ligand-ligand alignment. *J. Chem. Inf. Model.* **2005**, *45* (2), 449−460.

(24) Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput.-Aided. Mol. Des.* **1994**, *8* (3), 243−256.

(25) Böhm, H. J. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided. Mol. Des.* **1998**, *12* (4), 309−323.

(26) An alternative, however artifact-prone measure of the performance of a scoring function is the so-called 'Gini' coefficient, which is proportional to the area between the diagonal and the enrichment curve (Wilkie, A. D. Measures for comparing scoring systems. In *Credit Scoring and Credit Control*; Thomas, L. C., Crook, J. N., Edelman, D. B., Eds.; Clarendon Press: Oxford, 1992; pp 123−138).

(27) Enrichment curves are also known as 'Lorenz' curves in the area of credit scoring, as receiver-operating curves (Hand, D. J.; Henley, W.E. Statistical classification methods in consumer credit scoring: a review. *J. Royal Stat. Soc., Series A* **1997**, *160*, 523−54), and as performance curves (Gourieroux, C.; Jasiak, J. *Financial Econometrics: Problems, Models, and Methods*; Princeton University Press: Princeton, NJ, 2001, Chapter 4).

(28) Kruskal, W. H.; Wallis, A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47* (260), 583−621.

(29) DeLano, W. L. *The PyMol Molecular Graphics System 2000;* Open Source Version, DeLano Scientific: San Carlos, CA.

(30) 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* (3), 793−806.

(31) Godden, J. W.; Stahura, F. L.; Bajorath, J. Statistical analysis of computational docking of large compound databases to distinct protein binding sites. *J. Comput. Chem.* **1999**, *20* (15), 1634−1643.

(32) Liu, S.; Neidhardt, E. A.; Grossman, T. H.; Ocain, T.; Clardy, J. Structures of human dihydroorotate dehydrogenase in complex with antiproliferative agents. *Structure* **2000**, *8* (1), 25−33.

(33) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein−ligand interactions using negative training data. *J. Med. Chem.* **2006**, ASAP.

CI060027N