J Comput Aided Mol Des (2008) 22:257–265
DOI 10.1007/s10822-008-9168-9

Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results

Robert P. Sheridan · Georgia B. McGaughey · Wendy D. Cornell

Received: 30 October 2007 / Accepted: 28 December 2007 / Published online: 14 February 2008

Abstract As an extension to a previous published study (McGaughey et al., J Chem Inf Model 47:1504–1519, 2007) comparing 2D and 3D similarity methods to docking, we apply a subset of those virtual screening methods (TOPOSIM, SQW, ROCS-color, and Glide) to a set of protein/ligand pairs where the protein is the target for docking and the cocrystallized ligand is the target for the similarity methods. Each protein is represented by a maximum of five crystal structures. We search a diverse subset of the MDDR as well as a diverse small subset of the MCIDB, Merck's proprietary database. It is seen that the relative effectiveness of virtual screening methods, as measured by the enrichment factor, is highly dependent on the particular crystal structure or ligand, and on the database being searched. 2D similarity methods appear very good for the MDDR, but poor for the MCIDB. However, ROCS-color (a 3D similarity method) does well for both databases.

Keywords 2D similarity · 3D similarity · Docking · BEDROC · ROC · Glide · ROCS · SQ · SQW · TOPOSIM

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9168-9) contains supplementary material, which is available to authorized users.

R. P. Sheridan (✉) · W. D. Cornell
Molecular Systems Department, Merck Research Laboratories, RY50SW-100, Rahway, NJ 07065, USA
e-mail: sheridan@merck.com

G. B. McGaughey
Molecular Systems Department, Merck Research Laboratories, WP53F-301, West Point, PA 19486, USA

Introduction

In an earlier publication, McGaughey et al. [1], we benchmarked three commonly used approaches for virtual screening: 2D similarity (represented by TOPOSIM, and Daylight), 3D similarity (represented by SQW, SQW-shape, ROCS, and ROCS-color), and docking (represented by FLOG, Glide, and FRED). We applied these methods to a set of 11 targets, represented by a protein (for the docking methods) or a cocrystallized ligand (for the similarity methods). Our general conclusion was that, although there is not one method that is better for all targets, the similarity methods seemed to be generally superior to docking methods in selecting actives from diverse databases. This is interesting because the similarity methods, especially the 2D ones, require very little information to construct a query, are computationally very inexpensive, and have the fewest adjustable parameters. The diversity of the actives selected was about the same for docking and 3D similarity methods, and the diversity for the 2D similarity methods was less, but still respectable.

During the review of that manuscript an anonymous reviewer noted that the coupling between a particular protein structure and a particular ligand is arbitrary, so it may not be meaningful to compare docking results on a protein from a crystal structure to the similarity results on the ligand from the same crystal structure. Instead one should look at a variety of ligands for the same "activity." For instance, for the activity "HIV protease inhibitor" there are dozens of crystal structures to choose from as a docking target, and each ligand therein could be considered a potential target for similarity searches. This paper presents the investigation into that suggestion. It is an obvious extension to do docking using more than one protein structure, and we look at that as well.

Methods

Selection of target protein/ligand pairs

In McGaughey et al. we examined a set of 11 target protein/ligand pairs. As part of the follow up, we looked for PDB datasets where there were other versions of exactly the same protein as in McGaughey et al. (including species and subtype) that had drug-like ligands. We chose up to 5 alternative versions of that protein, including the original one from McGaughey et al., such that the ligands were very different (<0.6 similarity using the AP descriptor and the Dice index of similarity [2]). Note that the choice was made before any virtual screening was done so as not to bias the choice toward better targets. We did not pay particular attention to resolution of the crystal structures during selection, but as will be shown, resolution does not seem to systematically affect enrichment. In some cases we could not find as many as 5 examples and settled for as few as 3 (as for TS). DHFR (dihydrofolate reductase from L. casei) and CA_I (human carbonic anyhydrase I) had fewer than 3 examples and were eliminated. The final PDB datasets are listed in Table 1. The dataset used in McGaughey et al. for each target class is always listed first. The structures of the ligands are in Fig. 1.

Databases and actives

A retrospective evaluation such as ours requires a set of targets and databases to search. The databases, in turn, consist of a carefully chosen set of actives and a large number of decoys (presumed inactives). We use the same databases as in McGaughey et al.: First, a diverse (no two compounds with similarity >0.7 by AP Dice) subset of the MDDR [3] (a licensable database compiled from patent literature). This database contained ~25,000 compounds. Second, a diverse subset of the MCIDB (Merck's corporate database) consisted of ~9,800 compounds. The idea of diverse databases is that, since close analogs have been eliminated, the selection of actives can be considered a form of "lead-hopping." The list of actives in McGaughey et al. for the MDDR were derived for the most part from "activity keywords" for the MDDR. For the MCDIB actives are molecules where the $IC_{50} < 1$ μM for the primary in vitro assay. Here we use the same list of actives as in McGaughey et al., except that NEURAMINIDASE and PTP1B were removed as MDDR target classes; their actives did not come from activity keywords and the number of actives was very small. There are no known TS or NEURAMINIDASE inhibitors in the MCIDB, so those searches were not done. The numbers of actives we use here are shown in Table 2.

Treatment of the protein and ligand portions

For all targets, the ligand was separated from the protein, and all waters were eliminated. The entire protein was kept and there was no attempt to carve out an active site. Alternative side-chain positions were deleted. The bond orders of the ligands were corrected by hand. For the 3D similarity methods, the target was the cocrystallized conformation of the ligand.

Virtual screening (VS) methods

We decided that it was not necessary to look at all the methods in McGaughey et al., but just to use "examplar methods" of each type that we thought at the time to be the "best": TOPOSIM-AP for 2D similarity, ROCS-color (version 2.2) [4, 5] for 3D similarity and Glide [6, 7] for docking. In all cases we used the default parameters for each particular method and no target-specific constraints, our philosophy being that this is the only fair way of comparing the different classes of VS methods. Where it was necessary to use multiple conformers for database compounds (SQW and ROCS-color), these were generated using the conformation-generation method normally used for the VS method. We refer the reader to McGaughey et al. for details. We note the following changes and additions: TOPOSIM-AP, which we referred to as simply "TOPOSIM" in McGaughey et al., uses the Carhart atom pair descriptor [2] and Dice similarity index. Subsequent studies in our laboratory on the McGaughey et al. target set showed that the descriptor combination TTDT seems especially good at lead-hopping. This combination of the topological torsion TT [8] and the "binding point torsion" DT is described in Kearsley et al. [9] (it was called "ttbt" in that reference). Since it appeared that we may have used a less than optimal topological method in McGaughey et al., we applied TOPOSIM-TTDT on the expanded target set. Similarly, we wanted to compare our in-house 3D similarity method SQW (a variant on SQ [10]) on the expanded target set because ROCS-color appeared to have a slight edge over SQW on the target set in McGaughey et al. for MDDR and a large lead over SQW for MCIDB. In McGaughey et al. we used Glide version 3.0. The Glide version we had at the time of the follow-up study is 4.0. In the new version we found that the "box size" needed to be at least 10 Å around the native ligand in order for Glide to return scores for almost all the database entries.

Enrichment metrics

How "good" a VS method is usually depends on a series of retrospective screening experiments. We have a target t, a

Table 1 EF at 1%

Target class	PDB	Resolution	MDDR TOPOSIM-AP	TOPOSIM-TTDT	SQW	ROCS-color	Glide	MCIDB TOPOSIM-AP	TOPOSIM-TTDT	SQW	ROCS-color	Glide
CDK2	1aq1	2.0	22.2	20.8	10.4	18.2	10.4	1.9	0.0	0.0	1.9	3.9
CDK2	2c5x	2.9	5.2	7.8	0.0	7.8	9.1	1.9	4.8	7.7	14.5	4.8
CDK2	1g5s	2.6	1.3	1.3	1.3	3.9	7.8	12.6	1.9	1.9	7.7	3.9
CDK2	1e9h	2.5	6.5	6.5	2.6	1.3	1.3	2.9	2.9	1.0	5.8	4.8
CDK2	2c5t	2.1	5.2	7.8	1.3	9.1	5.2	21.3	18.4	1.9	18.4	4.8
COX2	1cx2	3.0	21.1	19.1	12.5	25.4	9.4	17.1	4.0	5.0	28.2	12.1
COX2	3pgh	2.8	7.0	2.0	2.3	3.5	0.8	13.1	10.1	6.0	13.1	7.0
COX2	4cox	2.9	7.0	9.0	2.0	6.6	7.0	7.0	7.3	10.1	12.1	7.0
COX2	1pxx	2.9	6.6	5.5	2.0	6.2	5.5	5.0	4.0	1.0	8.1	6.0
ER	3ert	1.9	14.9	20.3	27.1	21.7	14.9	10.4	11.7	12.1	17.3	6.5
ER	1l2i	2.0	12.2	28.5	28.5	28.5	13.6	23.8	17.7	17.7	21.6	21.6
ER	1sj0	1.9	10.8	8.1	9.5	16.3	13.6	14.7	18.2	8.2	13.8	13.0
ER	1gwq	2.5	10.8	13.6	20.3	25.8	9.5	13.0	15.1	18.6	13.4	12.5
ER	2b1z	1.8	12.2	36.6	27.1	23.0	13.6	22.5	13.0	12.5	16.4	20.3
HIV_pr	1hsh	1.9	31.7	24.3	5.2	12.5	13.3	22.3	26.3	10.7	18.7	7.1
HIV_pr	1kzk	1.1	24.3	22.1	14.8	15.5	4.4	6.2	1.3	7.6	10.2	3.1
HIV_pr	1ohr	2.1	28.0	21.4	10.3	20.7	11.1	20.1	16.5	9.4	24.1	8.9
HIV_pr	2bqv	2.1	26.6	31.0	21.4	17.0	13.3	12.0	21.8	16.5	17.4	3.6
HIV_pr	1ajv	2.0	18.4	12.5	5.9	19.9	7.4	4.5	2.7	1.8	6.7	4.0
HIV_rt	1ep4	2.5	2.7	3.4	7.4	2.0	2.7	2.3	0.5	1.4	1.8	2.3
HIV_rt	2be2	2.4	6.1	3.4	11.4	7.4	6.7	4.2	1.8	3.7	2.8	3.7
HIV_rt	1jkh	2.5	16.2	14.8	20.2	23.6	11.4	18.5	20.8	15.2	25.9	10.2
HIV_rt	1hpz	3.0	3.4	3.4	6.7	6.7	8.1	3.2	1.4	4.6	5.1	9.7
HIV_rt	1eet	2.7	4.7	7.4	12.8	8.8	13.5	2.3	0.9	3.7	9.2	12.9
PTP1B	1c87	2.1	–	–	–	–	–	0.0	1.0	0.0	0.0	12.7
PTP1B	1xbo	2.5	–	–	–	–	–	0.0	2.0	3.9	3.9	12.7
PTP1B	2cmc	2.2	–	–	–	–	–	8.8	33.2	29.3	45.0	66.5
PTP1B	1q6s	2.2	–	–	–	–	–	24.4	39.1	24.4	25.4	73.3
PTP1B	2cng	1.9	–	–	–	–	–	4.9	1.0	0.0	4.9	20.5
THROMBIN	1dwc	3.0	28.6	32.6	27.6	21.1	12.0	8.9	12.0	5.7	5.1	6.7
THROMBIN	1ta2	2.3	8.5	21.1	11.5	27.6	14.0	6.9	13.0	12.0	18.2	10.9
THROMBIN	1ypl	1.9	29.6	28.1	37.1	45.1	14.0	9.5	15.4	11.3	18.4	10.7
THROMBIN	1riw	2.0	13.5	17.1	33.6	35.6	16.1	2.8	4.9	9.1	17.2	7.7
THROMBIN	1bmm	2.6	31.6	22.6	32.1	31.6	15.0	7.3	7.3	10.5	13.2	10.1
TS	2bbq	2.3	22.7	64.7	58.3	6.5	29.1	–	–	–	–	–
TS	1axw	1.7	48.5	55.0	25.9	48.5	25.9	–	–	–	–	–
TS	1syn	2.0	38.8	48.5	25.9	22.7	22.7	–	–	–	–	–
meanEF			16.5	19.4	16.1	17.8	11.3	9.9	10.4	8.4	13.7	12.5
meanRank			3.0	2.6	3.2	2.5	3.7	3.1	3.4	3.7	2.0	2.9

method m, and a database d consisting of known actives and known (or presumed) inactives. One scores each entry in the database against the target, and sorts the database entries in order of the VS score (ascending or descending depending on whether high scores are more or less likely to be associated with activity). Sometimes a method cannot produce a score for a given compound; in McGaughey et al. these were given arbitrarily poor scores and ended up toward the end of the sorted list. One then "tests" the database entries in that order and notes the total number of actives found as a function of the number of database entries tested. If the method is useful, the front of the list is enriched in actives relative to a list where the actives are randomly scattered. There are many metrics to measure the

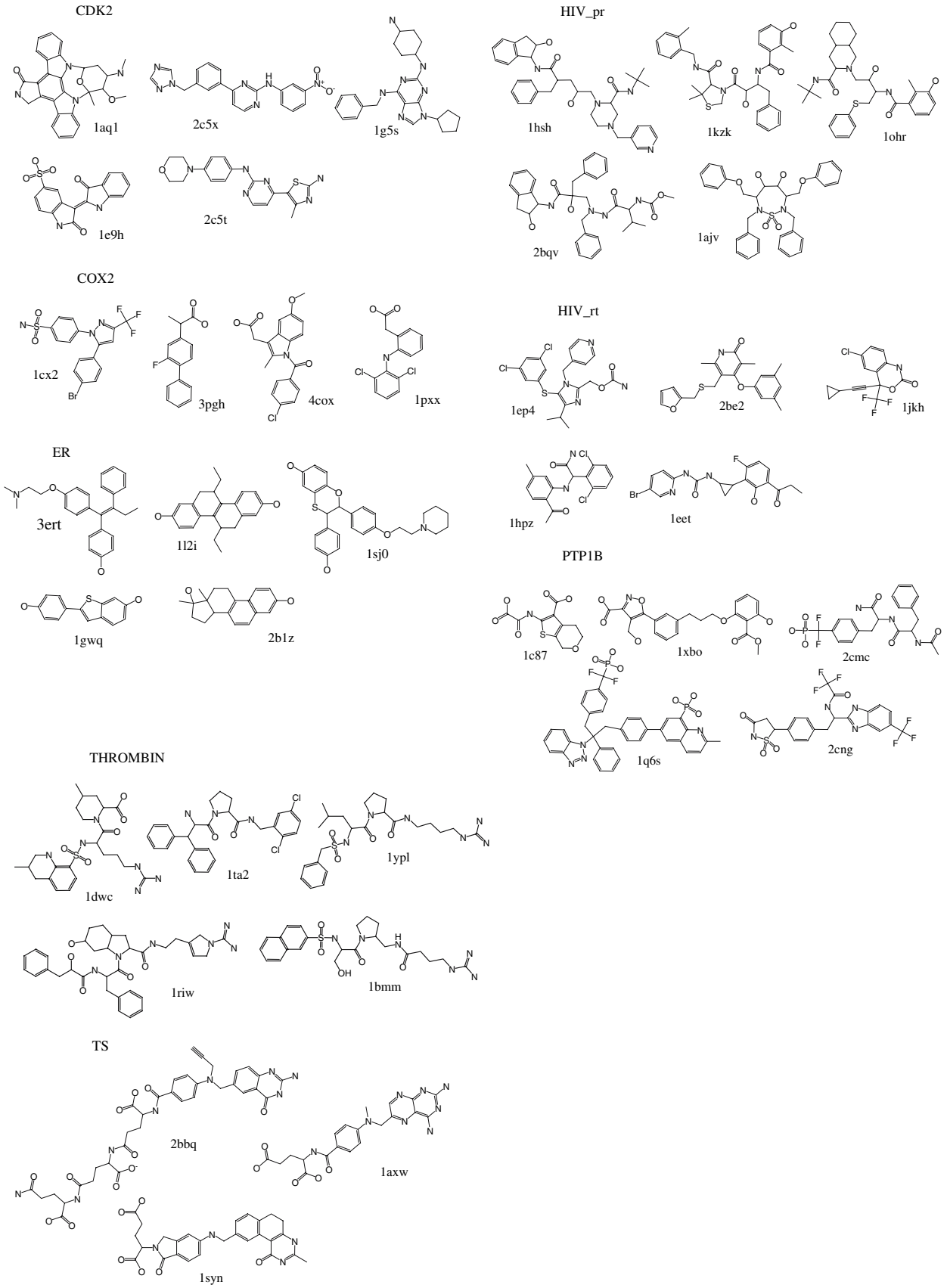

Fig. 1 Structures of the ligands used in this study. The PDB dataset from which they were extracted is shown

Table 2 The number of actives

Target class	Name	MDDR activities	Actives in MDDR set	Actives in MCIDB set
CDK2	Human protein kinase CDK2	"Protein Kinase C Inhibitor"[a]	77	104
COX2	Human cyclooxygenase 2	"Cyclooxygenase Inhibitor" "Cyclooxygenase-2 Inhibitor"	257	100
ER	Human estrogen receptor (alpha)	"Antiestrogen" "Estrogens"	74	233
HIV-pr	HIV-1 protease	"HIV-1 Pr Inhibitor"	136	226
HIV-rt	HIV reverse transcriptase	"Reverse Transcriptase Inhibitor"	149	218
PTP1B	Human protein tyrosine phosphatase 1B	–	–	103
THROMBIN	Human thrombin	"Thrombin Inhibitors"	200	510
TS	*E. coli* thymidylate synthase	"Thymidylate Synthetase Inhibitor"	31	–

[a] There are no known "CDK2" inhibitors in the MDDR. The closest related kinase is "Protein Kinase C" and that key word was used to search for actives

enrichment [11–15]. In McGaughey et al. we used enrichment factor (EF) at 1% of the database as the metric, that is, how many times as many actives are found after 1% of the database is tested than would be expected if the actives were randomly distributed in the database. We made this choice because the EF is the most popular metric in the literature, and is easy to understand. While EF at 10% of the database is often used in the literature, we felt that 1% was more realistic, because in practice only a small fraction of most corporated database is tested based on VS. Truchon and Bayly [15] formulated a new metric called BEDROC, which they claim has the well-behaved statistical properties of area under the ROC curve [13] (ROCAUC), but keeps the same property of "early recognition" one sees in EF, i.e. only the very front of the sorted list is heavily weighted. Truchon and Bayly recommend using BEDROC with $\alpha = 20$. On the other hand, several speakers at the "Evaluation of Computational Methods: Insights, Philosophies and Recommendations" session of the Fall 2007 ACS meeting (COMP Division) recommended ROCAUC as a metric. For ROCAUC, the entire sorted list is equally weighted. Here we will stay with EF at 1% to be consistent with McGaughey et al., of which this work is an extension. A detailed comparison of EF, BEDROC, and ROCAUC is beyond the scope of this paper. However, we provide ROCAUC and BEDROC values in Supporting Material should others want to make such a comparison. Also we will comment on the differences they make to the final conclusions.

Possible global goodness metrics

In most virtual screening studies, the enrichment varies from target to target, and there is no one method that is best for all targets. It is necessary to generate a number that summarizes the goodness of a method considering all the targets, i.e. the "global goodness". The simplest, the one we used in McGaughey et al., and the one used by some others in the literature, is what we will call here "meanEF(m)", that is the average EF for method m over the targets. The larger the meanEF(m), presumably the better the method. We later realized that meanEF(m) is problematical in that it is dominated by the targets where there is most variation between methods, and those targets are usually the ones with the highest EFs. The same applies to medianEF(m). A new paper from our laboratory [16] discusses alternative global goodness metrics that do not have this issue. One we will use here is meanRank. For every target, one ranks the methods. Say, for target CDK2/1aq1, TOPOSIM-AP has the highest EF, so it would be rank 1. ROCS-color has the next highest EF for CDK2/1aq1, so it would be rank 2, etc. The metric meanRank(m) is the average rank over all targets, the lower the better. A method that was always the best would have a meanRank(m) of 1, a method that was always the worst would have a meanRank(m) = M, where M is the number of methods. In practice, VS methods are hardly ever the best or worst on all targets, so meanRank(m) is likely to be closer to the middle of the range than near 1 or M. meanRank has the advantage as a global goodness metric that each target is self-scaled so that all targets are weighted approximately equally and the disadvantage that the numerical value depends on the set of methods being compared.

Database and target set perturbation

Whatever the global goodness metric, it will depend on the exact composition of the database and the selection of targets. One way of measuring the robustness of the conclusions of a study is to perturb the database and set of

targets to see how much the global goodness changes [16]. One way of perturbing the database is to randomly delete 20% of the molecules as suggested by Truchon and Bayly [15], recalculate the individual EFs, and then recalculate the global goodness metrics. Also one can perturb the target set by omitting a randomly selected subset of targets, say 10–20%, and recalculating the global goodness metrics. If one does this for a number of perturbation trials, say 10–20, one can generate two types of "error bars" for the global goodness. The idea is that if the error bars significantly overlap between two methods, the difference between them is very sensitive to the exact composition of the database and/or the exact set of targets and is less likely to be "real."

Results

Accumulation curves and metrics

Table 1 shows the (unperturbed) EFs at 1% for all targets and methods. We provide ROCAUC and BEDROC values in Supporting Material. Again, a detailed comparison of enrichment metrics is beyond the scope of this paper, but it is important to emphasize that for many target/method/database combinations EF does not necessarily track with ROCAUC. There are some cases where having a high EF, but a low ROCAUC is due to an artifact. For some 3D methods a score cannot be returned for a large fraction of the actives in a target class, and the actives end up at the end of the list, making the ROCAUC very low. However, when we confine ourselves to cases where >99% of the actives are scored, we still see real differences between the metrics. These are explainable by the non-ideal shape of some accumulation curves. These occur for a number of different methods and targets. Some examples are shown in Fig. 2.

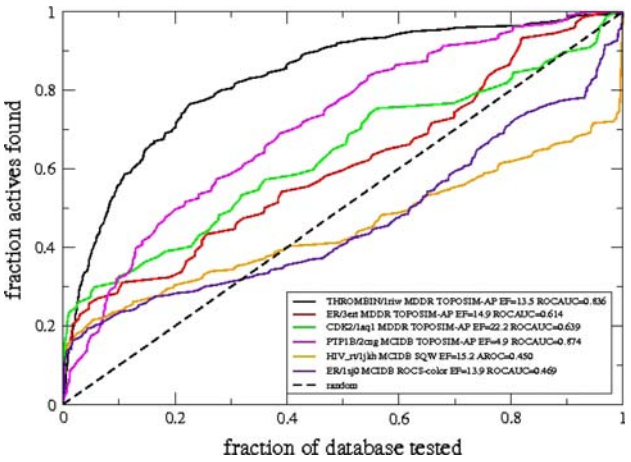

Fig. 2 Selected accumulation curves demonstrating the variability in shape that gives rise to disparity of EF and area under the ROC curve (ROCAUC). At this scale, the ROC curves are indistinguishable from the accumulation curves

At the scale of the plot, accumulation curves and ROC curves are indistinguishable and the observations apply equally well. The black line (THROMBIN/1dwc TOPOSIM-AP) is a more or less ideal example of a hyperbolic accumulation curve having a high EF and a high ROCAUC. Surprisingly common is the real "bimodal" situation shown by the red, green, orange, and blue lines, where there are a significant number of actives at the very front of the list, hence a high EF, but after the front of the list, the actives do not accumulate faster than random, so that the ROCAUC is near the "random" value 0.50. In the case of CDK2/1aq1 TOPOSIM-AP, for instance, the ligand target is staurosporine. Some of the actives (protein kinase inhibitors) in the MDDR resemble staurosporine topologically, and are at the front of the list, but most of the actives do not resemble staurosporine more than the inactives do. Less common is the situation shown by the magenta line (PTP1B/2cng TOPOSIM-AP), where distribution of actives is far from random in a favorable direction, hence a high ROCAUC, but there are very few actives at the beginning of the list, hence a low EF. EF and BEDROC are generally more correlated to each other than either is to ROCAUC, not surprising since EF and BEDROC are "early recognition" methods, with the qualification that BEDROC with $\alpha = 20$ effectively weights more of the list than the first 1%. The fact that very many real accumulation curves are non-ideal supports the idea that EF/BEDROC and ROCAUC are not interchangeable measures. Truchon and Bayly argue that only early recognition metrics are relevant for VS, where one can experimentally screen only a small fraction of a database, and that ROCAUC, while perhaps more statistically rigorous than EF, is a less relevant metric for VS. We would agree.

The variation of EFs over protein/ligands

We note the following observations based on the (unperturbed) EFs in Table 1:

1. In a given target class (example CDK2), EF may vary widely between PDB datasets. One can find examples in 2D and 3D similarity as well as for docking. That is, enrichments strongly depend on exactly which protein structure and/or ligand is used.
2. In some cases, the choice of ligand we made in McGaughey et al. proved to have the best topological similarity within that target class (e.g. MDDR TOPOSIM-AP for CDK2/1aq1), sometimes it was the worst (e.g. MDDR TOPOSIM-AP for TS/2bbq).
3. Within a target class there is generally no significant correlation between the EFs of the similarity methods and the EFs for Glide, indicating that the choice of protein and the choice of ligand are effectively "uncoupled," as suggested by the reviewer. Figure 3

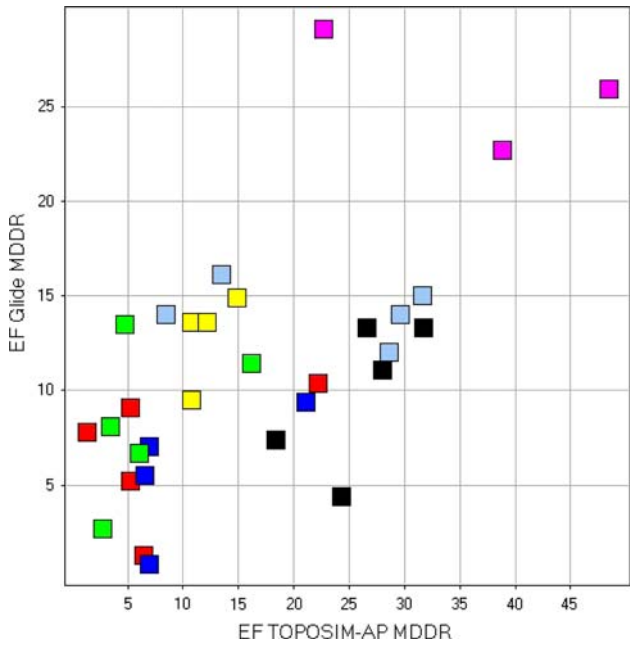

Fig. 3 The relationship between EF 1% for a docking method (Glide) and the EF for topological similarity (TOPOSIM-AP) for the MDDR database. Target classes are represented by colors: red = CDK2, light blue = COX2, yellow = ER, black = HIV_pr, green = HIV_rt, dark blue = THROMBIN, magenta = TS

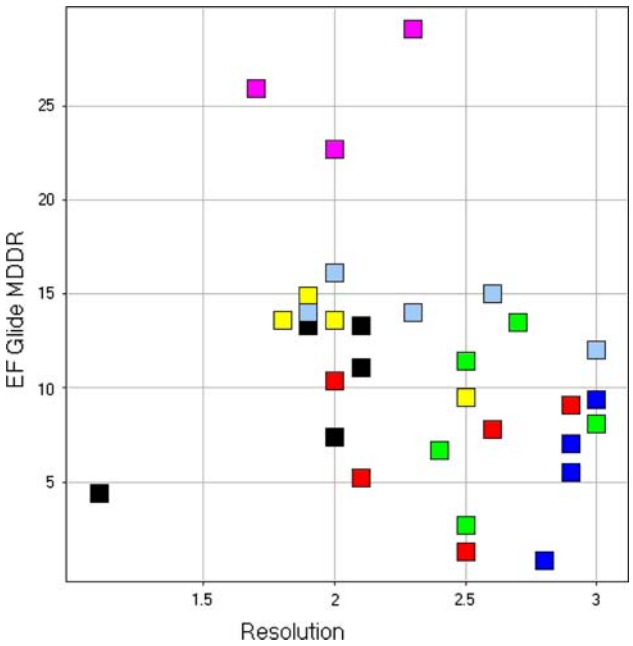

Fig. 4 The relationship between EF 1% for a docking method (Glide) and the resolution of the crystal structure. The MDDR database is shown. Target classes are represented by colors: red = CDK2, light blue = COX2, yellow = ER, black = HIV_pr, green = HIV_rt, dark blue = THROMBIN, magenta = TS

shows the example of Glide versus TOPOSIM-AP for the MDDR database. The other similarity methods give similar graphs for MDDR, and the situation is the same for the MCIDB.

4. Within a target class, there is generally no correlation between the EFs for Glide and the resolution of the crystal structure. Figure 4 shows the situation for the MDDR database. It is the same for the MCIDB. No correlation is seen with other measures of the quality of the crystal structures like R-Value and R-Free (not shown here, but provided in Supporting Material).
5. Within each target class and any given method, generally speaking there is no consistent correlation between the EFs for MDDR versus the same for MCIDB. Figure 5 shows the situation with TOPOSIM-AP, but the other methods give similar graphs.

The global goodness of methods

For this exercise, we generated 20 perturbed sets for each type of perturbation: in one we excluded 20% of the database and in the other we excluded 20% of the targets without regard to which target class they were from. We treated the MDDR and MCIDB separately. Figure 6 shows the global goodness metrics with "error bars" over the 20 trials, red for the database perturbation and black for the target set perturbation. meanEF for MDDR is basically consistent with

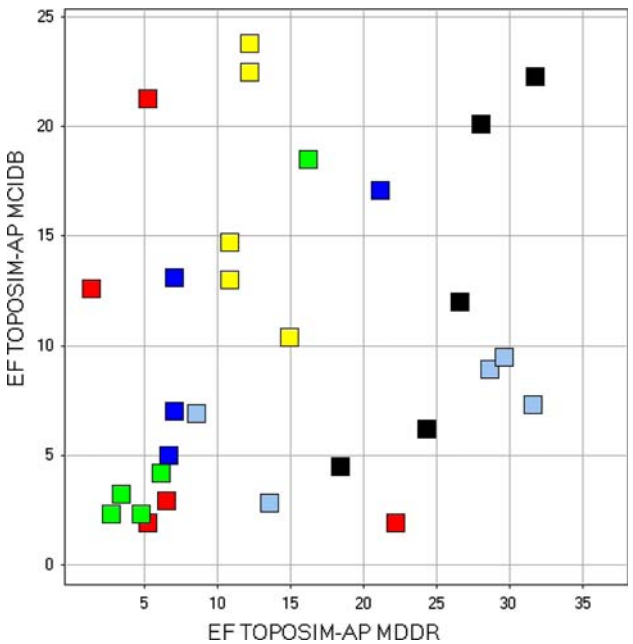

Fig. 5 The relationship between EF 1% for MCIDB versus MDDR for a topological similarity method (TOPOSIM-AP). Target classes are represented by colors: red = CDK2, light blue = COX2, yellow = ER, black = HIV_pr, green = HIV_rt, dark blue = THROMBIN, magenta = TS

the conclusions from McGaughey et al. TOPOSIM-AP, SQW and ROCS-color are about equivalent, as seen in their partial overlap of error-bars, and all are better than Glide.

The error bars on Glide for MCIDB are so large because in the MCIDB, PTP1B/1xbo, PTP1B/2cmc and PTP1B/1q6s have especially large EFs, such that the meanEF is perturbed when they are omitted. TOPOSIM-TTDT appears to be somewhat better than TOPOSIM-AP only for MDDR.

If we compare the MCIDB set, Glide has roughly the same meanEF as for MDDR. However, all the similarity methods have been lowered, with ROCS-color lowered the least. Why should there be that strong a difference between databases? We noted some specific cases in McGaughey et al. For example, the biggest difference between databases is for CDK2/1aq1, where the MDDR TOPOSIM-AP EF = 20.8 and the MCIDB TOPOSIM-AP EF = 1.9. There are many CDK2 active compounds in the MDDR that have parts resembling the 1aq1 target (staurosporine) and nothing in the MCDIB resembling it. Here we can be a little more general in our explanation, which has to do with the likelihood that an analog of the target will appear in the database. Our target ligands are all from the PDB. Some analog of the target is likely to be in the MDDR, since MDDR is compiled over the patents of many companies. Individual companies like Merck usually concentrate on particular series of compounds for a given therapeutic area, and there may be pressure to avoid series already in the literature (or the PDB) for intellectual property reasons. Of course, the exception would be if the PDB compounds are actually from Merck. There is some support for this explanation based on how many actives in the top 1% of the database are analogs to the target. Let us consider a compound with a similarity to the target >0.65 by AP Dice an analog. There are 29 protein/ligand pairs in common between the MDDR and MCIDB in Table 1. In 10 out of 29 cases, the MDDR has more analogous actives in the top 1%. In 5 cases, the opposite is true. Of those 5 cases, three ligands (THROMBIN/1ta2, HIV_pr/1hsh, and ER/1sj0) are Merck compounds. This is not to say that topological similarities depend on having analogs in the database to get a good EF, only that the more analogs, the more likely there are to be actives that to some extent resemble the target more than inactives resemble the target. This argument about analogs clearly applies to 2D similarity, but to a lesser extent to 3D similarity, with SQW being more sensitive than ROCS-color.

meanRank tells roughly the same story as meanEF, remembering that lower meanRank is better. Here, however, ROCS-color is clearly the best method for both databases, especially for MCIDB.

If we use BEDROC instead of EF, plots very similar to those in Fig. 6 are produced, and a similar story unfolds, again not surprisingly because both BEDROC and EF are "early recognition" metrics. Interestingly, if one uses ROCAUC, TOPOSIM-AP and TOPOSIM-TTDT are the best methods for both MDDR and MCIDB, with ROCS-color being a close third.

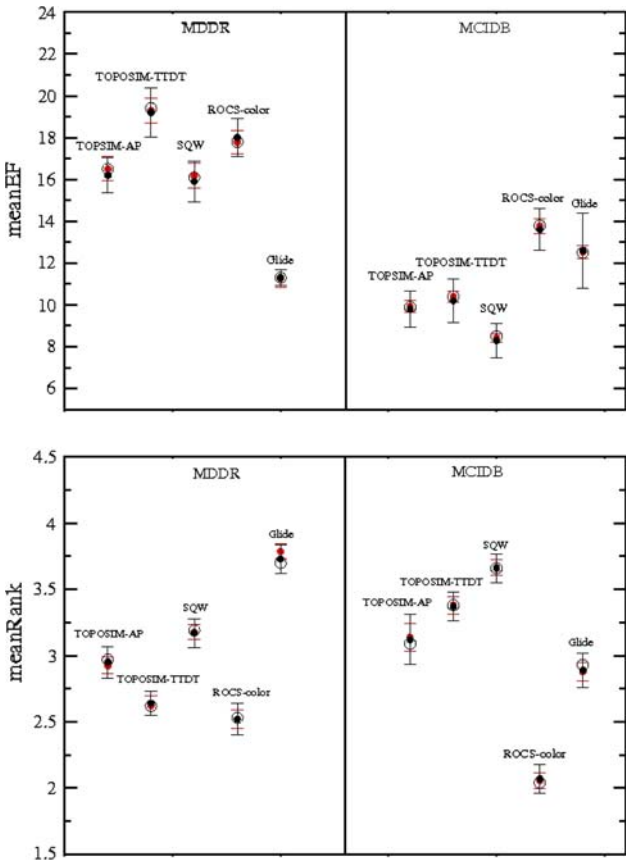

Fig. 6 Two global goodness metrics for two databases. The unperturbed metrics are shown as open circles. In red are shown the mean and ±1 standard deviation global goodness metrics for 20 trials of "database perturbation" where 20% of the database was randomly omitted. In black are shown the mean ±1 standard deviation global goodness metrics for 20 trials of "target set perturbation" where 20% of the targets were randomly omitted. For meanEF higher values are better, while for meanRank, lower values are better

Discussion

All of these results enforce the depressing observation that VS results are so dependent on the exact details of the query and the contents of the database, not to mention the enrichment metric, that it is nearly impossible to compare methods between different VS studies when none of those things are in common, and when only enrichment metrics are published. Comparisons might be easier if more information, like accumulation curves for all target/method combinations, is routinely distributed.

We would like to emphasize the following points:

1. Since results of docking are dependent on the exact crystal structure, referring to the target as "HIV protease", for instance, is not good enough. One has to at least specify the PDB dataset. It has been noted in the past [17–21] that docking enrichments are very sensitive to the particular crystal structure one uses, whether one

is using a homology model etc., not to mention the exact docking parameters. There is some argument for using holo- rather than apo- crystal structures [20]. However, since it has been noted that the best enrichments sometimes occur with homology models rather than the original protein structures [17], getting the protein structure "exactly right" for any particular cocrystallized ligand does not seem to help when trying to dock a large number of diverse actives. Perhaps it is not surprising, then, that enrichments do not depend the resolution of the crystal structure of the protein.

2. The goodness of 2D similarity methods are more sensitive than we would have suspected to the match of target ligands to the set of actives. In our case, the MDDR, which contains compounds from many sources, is more likely to contain analogs to PDB targets than our in-house database. Others, for example Muegge and Enyedy [21], have noticed strong effects of databases on docking results.
3. 3D similarity methods will not necessarily escape this "analog" issue. In particular our in-house method SQW seems to be as sensitive as TOPOSIM-AP to the differences between the MDDR and MCIDB. In contrast, ROCS-color seems less sensitive to the absence of analogs and does very well on both databases.

Acknowledgement The authors thank Christopher Bayly for useful discussions.

References

1. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas CK, Lindsley S, Maiorov V, Truchon J-F, Cornell WD (2007) Comparison of topological, shape, and docking methods in virtual screening. J Chem Inf Model 47:1504–1519
2. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci 25:64–73
3. MDL Drug Data Report licensed by Molecular Design Ltd., San Leandro, CA. http://www.mdli.com
4. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK (2003) Gaussian docking functions. Biopolymers 68:76–90
5. Hawkins PCD (2006) A comparison of structure-based and shape-based tools for virtual screening. Abstracts of Papers, 231st ACS National Meeting, Atlanta, GA, United States, March 26–30, 2006
6. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47:1750–1759
7. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47:1739–1749
8. Nilakantan R, Bauman N, Dixon JS, Venkataraghavan R (1987) Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. J Chem Inf Comput Sci 27:82–85
9. Kearsley SK, Sallamack S, Fluder EM, Andose JD, Mosley RT, Sheridan RP (1996) Chemical similarity using physiochemical property descriptors. J Chem Inf Comput Sci 36:118–127
10. Miller MD, Sheridan RP, Kearsley SK (1999) SQ: a program for rapidly producing pharmacophorically relevant molecular superpositions. J Med Chem 42:1505–1514
11. Edgar SJ, Holliday JD, Willett P (2000) Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. J Mol Graph Model 18:343–357
12. Sheridan RP, Singh SB, Fluder EM, Kearsley SK (2001) Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. J Chem Inf Comput Sci 41:1395–1406
13. Triballeau N, Archer F, Brabet I, Pin J-P, Bertrand H-O (2005) 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 48:2534–2547
14. Seifert MHJ (2006) Assessing the discriminatory power of scoring functions for virtual screening. J Chem Inf Model 46:1456–1465
15. Truchon J-F, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the early recognition problem. J Chem Inf Model 47:488–508
16. Sheridan RP Alternative global goodness metrics and sensitivity analysis: heuristics to check the robustness of conclusions from studies comparing virtual screening methods. J Chem Inf Model (in press)
17. Kairys V, Fernandes MX, Gilson MK (2006) Screening drug-like compounds by docking to homology models: a systematic study. J Chem Inf Model 46:365–379
18. Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M (2004) Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. J Med Chem 47:45–55
19. Andersson CD, Thysell E, Lindstrom A, Bylesjo M, Raubacher F, Linusson A (2007) A multivariate approach to investigate docking parameters' effects on docking performance. J Chem Inf Model 47:1673–1687
20. McGovern SL, Shoichet BK (2003) Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J Med Chem 46:2895–2907
21. Muegge I, Enyedy IJ (2004) Virtual screening for kinase targets. Curr Med Chem 11:693–707