J Comput Aided Mol Des (2008) 22:169–178
DOI 10.1007/s10822-007-9167-2

Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection?

Andrew C. Good · Tudor I. Oprea

Received: 30 October 2007 / Accepted: 19 December 2007 / Published online: 9 January 2008

Abstract Over the last few years many articles have been published in an attempt to provide performance benchmarks for virtual screening tools. While this research has imparted useful insights, the myriad variables controlling said studies place significant limits on results interpretability. Here we investigate the effects of these variables, including analysis of calculation setup variation, the effect of target choice, active/decoy set selection (with particular emphasis on the effect of analogue bias) and enrichment data interpretation. In addition the optimization of the publicly available DUD benchmark sets through analogue bias removal is discussed, as is their augmentation through the addition of large diverse data sets collated using WOMBAT.

Keywords Virtual screening · Enrichment · Validation · Analogue bias · Chemotypes · DUD · WOMBAT

Introduction

Over the years structure-based virtual screening (SVS) has been applied successfully to a wide array of targets [1–7]. Nevertheless, the numerous reviews dedicated to SVS testify to its continuing technological evolution, most notably in the field of scoring function development [8–15]. As a consequence the computational chemist is presented with a plethora of techniques from which to choose, rendering optimal technique selection a significant challenge. In an attempt to provide insights into the relative merits of SVS methodology, numerous retrospective docking tool comparisons have been undertaken [16–23]. Such research generally utilizes the docking methodology available in-house to search for known binders (actives) to a number of targets seeded into a decoy database of (presumed) inactive compounds. The ability of said tools to enrich their ranked compound lists with actives is then quantified using metrics relating the ranking of the actives to the decoys.

While the aim of these enrichment studies is to provide quantitative insight into docking performance, the myriad variables that underpin SVS calculations have the potential to confound results interpretation. Despite their importance, the effects of said variables in the context of virtual screening comparisons have received relatively scant attention in the literature. Here we investigate these shortcomings through systematic analysis of the enrichment studies of Cummings et al. [19]. Input parameter selection, starting conformation, docking technique version, enrichment cutoff selection and data set choice and construction are all explored for their potential to effect study outcome.

An additional problem endemic to SVS enrichment studies is the issue of public data set availability. Historically many docking tool comparisons have utilized proprietary data sets to work with compound collections of significant size. This has rendered direct comparisons between studies a tricky task at best. Where public data sets have been used, the sizes of said sets have typically been too limited to provide statistically meaningful insight. Recently Huang et al. [24, 25] went some way to rectifying this issue with the release of the DUD data set. DUD

A. C. Good (✉)
5 Research Parkway, Wallingford, CT 06492, USA
e-mail: andrew.good@bms.com

T. I. Oprea
Division of Biocomputing, Department of Biochemistry and Molecular Biology, MSC11 6145, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA

provides a wide variety of target data sets (~ 3,000 ligands covering 40 targets) together with a carefully curated set of decoys. As such DUD provides a repository of publicly accessible SVS data that can form the foundation for future baseline enrichment calculations. There are, however, still a number of issues to be addressed, particularly with respect to target selection and analogue bias. With this in mind we have clustered the DUD data roughly according to chemotype. The resulting structure subset is designed to render enrichment results more relevant to scaffold hopping ability. In addition we have mined the WOMBAT database [26–27] to provide a series of publicly available large diverse data sets [28], significantly augmenting those already collated for DUD. The results of these efforts are detailed below.

Methods

Effects of docking tool version selection and data set up

A common issue for enrichment test results interpretation is primarily a function of the scientific process. Practitioners of enrichment studies are typically limited to the docking tools they have in-house, and the comparisons often take place over a protracted time span. Delays can be further exacerbated when the time from article acceptance to publication is significant. As such enrichment studies are rarely exhaustive in the tools compared. Further, the rapid evolution of docking techniques mean that the numbers described may no longer be reflective of tool performance.

Another problem to consider in the context of enrichment interpretation is that of tool set up. It is often argued that the default set up is the one to choose for comparisons. While this may provide a measure of baseline performance it is hardly reflective of real world SVS, where target knowledge often plays a central role in search query construction. Further, for many tools defaults are less indicative of best practice and more a reflection of the latest research focus of the designers.

An additional consideration in tool set up is that of database construction. Choices regarding techniques used in 3D model building and issues such as tautomer definitions and protonation states can all potentially effect enrichment.

With these potential problems in mind a number of experiments have been undertaken using the Cummings Thrombin and PTP1B data sets to highlight their effects in a more quantitative manner:

(i) Data sets have been rerun using GLIDE version 4.0207 in SP mode [29] with the 3D models built using LIGPREP version 2.0113 [29]. These have been compared with the results in the original paper where CORINA [30]/CONCORD [31] were applied to prepare the 3D models for GLIDE comparisons using version 2.5. All LIGPREP calculations were undertaken with tautomeric and protonation states fixed to match those of the original data set provided by Cummings and co-workers. Note that for all calculations, the enrichment (EF) at any given percentage cutoff of the database (x%) has been calculated according to formula in Eq. 1:

$$EF^{x\%} = \frac{Hits_{sampled}}{N_{samples}} \frac{N_{total}}{Hits_{total}} \quad (1)$$

where $Hits_{sampled}$ is the number of hits found at x% of the database screened, $N_{sampled}$ is the number of compounds screened at x% of the database, $Hits_{total}$ is the number of actives in entire database, and N_{total} is the number of compounds in the entire database.

(ii) To highlight the potential effect of tool set up, DOCK 4 calculations have been rerun using a calculation setup constrained to exploit data provided by the ligand template used in a ROCS search on the same data [32] by Hawkins et al. (DOCK ROCS-like). Hawkins and co-workers used the constituent ligands from the target proteins of this and other data sets, comparing ROCS ligand-based screening enrichments with their protein docking counterparts. Comparison with Cummings DOCK data is of little utility, however, since the search spaces diverge (ligand only versus active site region carved out by DOCK's SPHGEN program) as are the ligand conformers interrogated during the search (DOCK generated versus Omega generated). With this in mind DOCK site points have been constrained to map key sub site interactions using only the template ligand atoms applied in the ROCS searches. In this way the search space is restricted to that used in the ROCS calculation (i.e. the active site around the bound ligand), allowing for a more relevant comparison. Energy grids have been softened to a 4/8 potential and OMEGA [32] derived conformations generated for the actives and decoys, again to improve consistency between calculations. The DOCK site point set ups are highlighted in Fig. 1. In addition calculations were run with and without electrostatics included in the scoring, and results were also normalized by heavy atom count to weight against larger molecules. The search variations are highlighted to illustrate the variation possible in results for real world screen settings. A full listing of the primary input parameters used for these DOCK calculations can be found in the dock.in file in supplementary information.

(iii) Given the importance of acidic interactions for PTP1B activity, the selected protonation states of acidic functionality in the Cummings decoy data set have been investigated. A visual analysis of the decoys showed that, while phosphates, phosphonates, sulfonates and carboxylates had been deprotonated, a number of the more esoteric

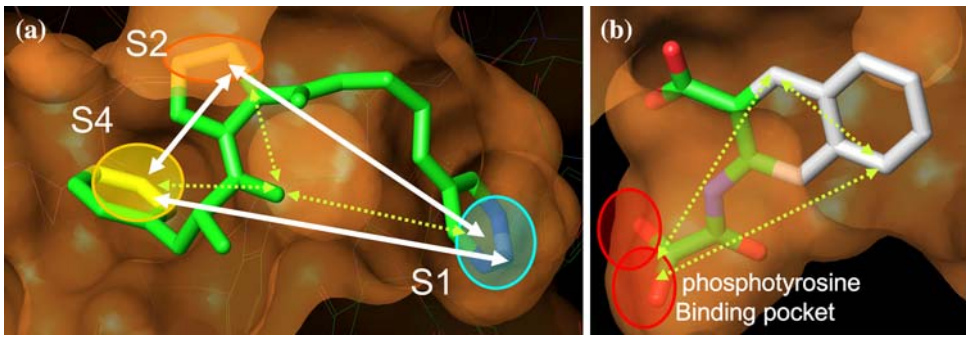

Fig. 1 Site point set ups for DOCK 4 calculations on (**a**) Thrombin and (**b**) PTP1B. Thrombin search parameters: four node search (one sample clique highlighted, thick lines indicate clique distance pair mappings between critical regions), three critical regions (S1 donor or base, S2/S4 hydrophobic, all remaining atom types set to generic. PTP1B search parameters: Three node search (one sample clique highlighted), one critical region (phophotyrosine pocket acidic center, perm one from two), remaining atoms mapped to relevant atoms type of ligand template as defined in DOCK chem.defn file

acidic groups had not. With this in mind an OECHEM toolkit [32] python script designed to allow SMARTS string interrogations of data sets was employed to determine the number of decoys potentially affected by this. The SMARTS queries employed are shown in Table 1.

Effects of target and data set selection

Target, active compound and decoy selection can all have significant impact on enrichment results. Huang et al. [24] provide an excellent example of this, highlighting significant enrichment reduction when comparing standard decoy sets with those selected as part of the DUD collection. Here we delve more deeply into the target and data set selections highlighting issues relating to analogue bias and target utility.

(i) The analogue bias intrinsic to the sets is explored though visual analysis and comparison with 2D descriptor enrichment performance. FCFP2 Scitegix fingerprints [33] have been employed in this study. The ligand templates used are the same as those employed by Hawkins et al. [34] in their ROCS calculations on the Cummings data. In addition the ligand extracted from the neuraminidase template protein (2qwg) together with the actives and decoys used in the studies by Pham and Jain [35] have also been studied this way. Note that the original analysis involved a protein docking approach to enrichment. This has been modified to allow ligand-based screening on the data, which was chosen as the screening set was available for download [36]. The authors consider that the data is representative of the sets typically seen for this target [22, 23, 34, 37].

(ii) The analysis of data sets has been extended to include the DUD set, with particular emphasis on the problem of analogue bias and virtual screening compatibility. To this end all DUD ligand sets have been systematically filtered using the lead-like criteria of Oprea et al. [38] (molecular weight <450 and AlogP [39] <4.5 (5.5 for nuclear hormone receptor targets to reflect their preference for hydrophobic moieties)) and reduced graphs [40, 41] (see Fig. 2), through the application of a PIPELINE PILOT script [33].

All ligands passing the lead-like filters and exhibiting the same graph have been assigned to a single cluster, with the smallest molecule of each cluster defined as the parent. Note that the selection of reduced graphs is the result of extensive analyses of clustering using a variety of 2D descriptors. These investigations centered on their ability to group according to chemotype in a manner consistent with selections made using visual analysis. In general small changes in atom types within a framework or substitution pattern were sufficient to confound most descriptors. Of those tested only reduced graphs were found to produce clusters that were visually satisfactory in the majority (80–90%) of cases.

Note that, while useful, reduced graphs are still not a perfect technique for chemotype partition. Overall the technique provides an effective conservative method for automated chemotype assignment, with the following limitations. First as one might imagine it is less

Table 1 Acidic SMARTS searches of Cummings decoy set, with number of protonated examples found for each type

SMARTS string definition	Acidic moieties covered	Decoy set examples
C(=O)[NH]S(=O)(=O)	Acyl sulfonamide	6
c[NH]S(=O)(=O)	Aryl sulfonamide	16
C(=O)[NH1;R]C(=O)	Barbiturates/thiazolidinediones...	20
C(=O)C([OH1])=C	Vinylogous carboxylic acids	3
Cnnn[nH1]	Tetrazoles	26

In all 71 decoys were found to contain protonated potentially acidic functionality using these queries

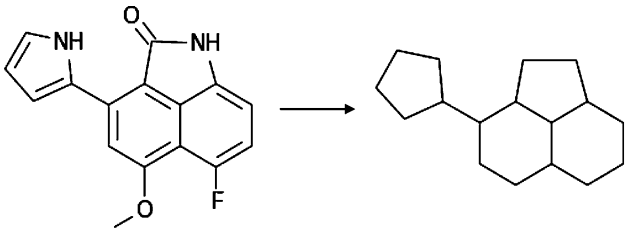

Fig. 2 Example of a reduced graph representation

discriminating with simple systems, particularly those containing only one ring. As such the method will tend to produce less fragment-like chemotypes than expected with targets that contain uniquely small ligands (e.g. DUD Comt data set [24]). Secondly, the technique will still differentiate small changes in ring structure (e.g. phenyl for thiophene) as rendering a new chemotype when really said change creates a simple analogue. Similarly the addition of a small carbocycle e.g. cyclopropyl in place of a small alkyl will also produce a new chemotype. Data users should feel free to contextualize their results from the perspective of these limitations. It is important to point out, however, that the method still produces clusters that in the majority of cases make sense. It is particularly adept at clustering the analogues that abound for chemotypes popular with medicinal chemists.

(iii) Based on the analogue bias seen in many of the DUD sets, the need for additional publicly available screening data is still a clear priority. To this end 13 targets have been abstracted from the WOMBAT database using the same filtering and clustering procedure applied to the DUD data. In each case the activity data and original reference associated with each compound has been included to allow further insight, and allow data sub setting (e.g. restriction to compounds with lead-like rather than drug-like potency).

As with all data sets there are a number of caveats in their construction that should be noted. No attempt has been made to separate human data from other species for any of the targets. Typically the sequence identity for the target chosen is high and often the alternative species has been chosen as a surrogate for human data. Further, other than for aldose reductase the number of non-human data points is typically fairly limited. Nevertheless this may have some effect on the ability to interpret relative activities in some instances. Activity data also needs to be interpreted with the realization it has been extracted from multiple sources. This can have significant consequences. For example for CDK-2 the cyclin variant chosen varies from assay to assay. Further the concentration of ATP is not known for each data point. Variations in ATP assay concentration are known to produce significant alterations in activity values returned. In addition, it is not known if each inhibitor class hits its given kinase in the same state (activated versus inactivated). Also the relative loop positions for each chemotype are unknown (e.g. P38 DFG loop in versus loop out). All this must be borne in mind during results interpretation, with users who find subtle issues based on these or other factors being encouraged to report them for future data annotation and refinement.

For HIV reverse transcriptase data, WOMBAT does not differentiate NNRTIs from NRTIS. AlogP (>1) and LogS (<−2) constraints have been used to differentiate the two classes, and substructure searches on the primary NNRTI chemotypes reveal no hits, suggesting good differentiation. It is still possible that an NRTI or two still lurks in the data, however. Along the same lines for acetylcholinesterase no differentiation is made between catalytic/peripheral/dual binding inhibitor classes, so both active sites should be considered for searches involving this target.

For two NHR target selections and the D2 data the biological effect field has been used to only keep molecules designated as antagonists. This definition is to some extent defined by the nature of the assay applied, however, so additional filters removing steroid and dopamine specific substructures from the lists have been used to further refine the data. Some ambiguity regarding the antagonist definition remains, however.

Results

(i) An example comparison of LIGPREP versus CORINA 3D models is shown in Fig. 3. Enrichment comparisons for different virtual screening runs using LIGPREP model and GLIDE version combinations are shown in Fig. 4.

(ii) Enrichment comparisons between ROCS and DOCK 4 running in ROCS-like mode for the Cummings et al. Thrombin and PTP1B data are shown in Fig. 5.

(iii) The number of protonated acidic decoys found in the Cummings decoy set is listed in Table 1.

(iv) Results for the analogue bias enrichment experiments involving the data of Cummings et al. [19] are shown in Fig. 6. For the Pham and Jain comparison [35], the template and active compounds are shown in Fig. 7 to highlight the obvious analogue bias. Enrichment tests using FCFP2 fingerprints resulted in all hits appearing in the top 13 of the hit list.

(v) The results produced upon application of the lead-like and reduced graph filters to the DUD data sets are shown in Table 2. An example of a large reduced graph cluster extracted from the Cox-2 DUD data set is shown in Fig. 8.

(vi) Results for the WOMBAT clustering exercise are given in Table 3.

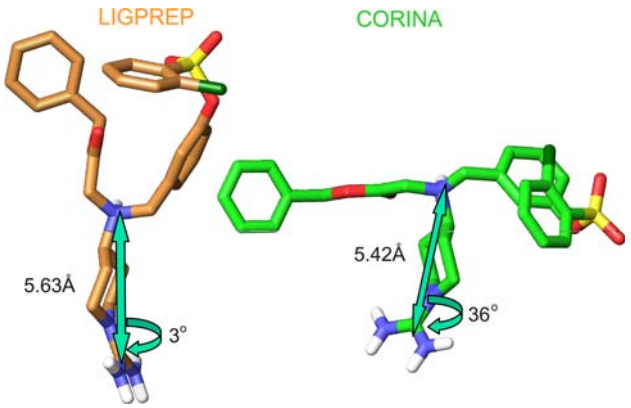

Fig. 3 Example of differences between LIGPREP and CORINA structures for Cummings data set molecule Thrombin hit 3

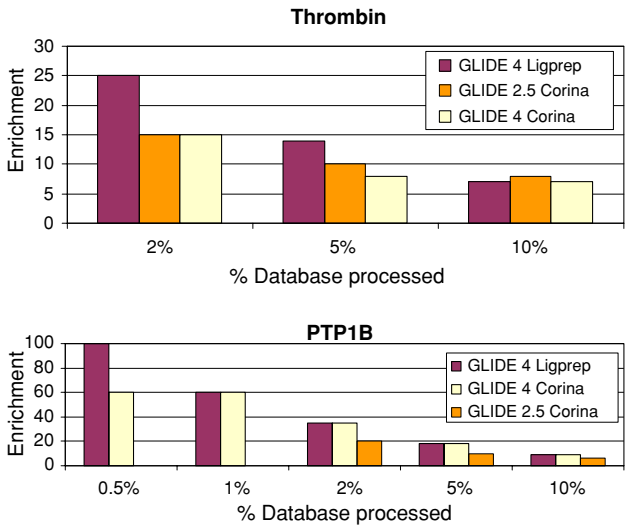

Fig. 4 Enrichment comparisons using different versions of GLIDE and alternate 3D model sources for the Thrombin and PTP1B data of Cummings et al. [19]

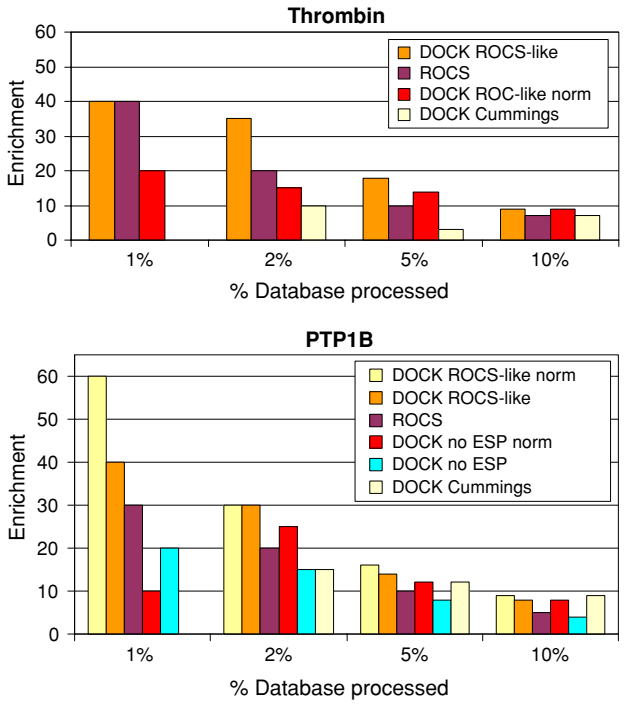

Fig. 5 Thrombin and PTP1B enrichment comparisons between ROCS and DOCK running in ROCS-like mode with and without electrostatics (no ESP) and normalized by heavy atom count (norm)

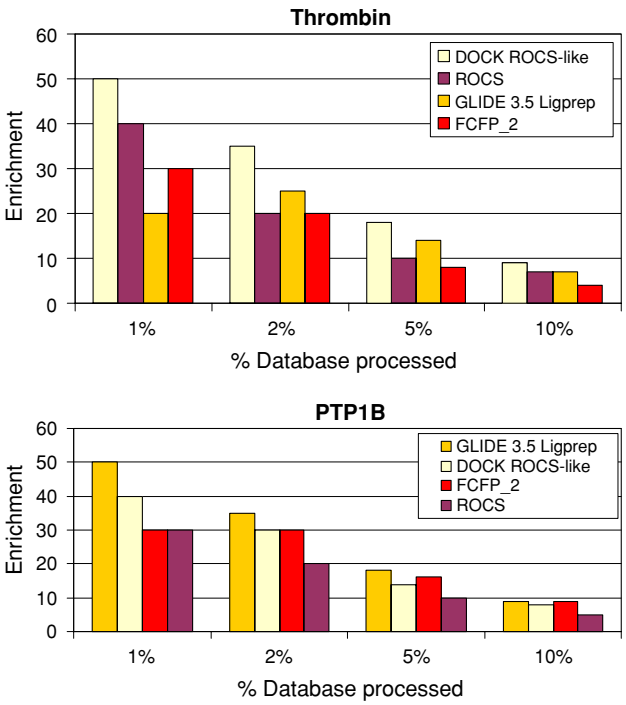

Fig. 6 Enrichments using simple 2D FCFP2 fingerprints compared with the 3D methods used on the Thrombin and PTP1B data

Discussion

It is important to note at the outset of this discussion that Cummings and coworkers [19] are to be lauded for placing their data in the public domain for study. Without this it would not have been possible to make the comparisons documented here. Many of the choices made by Cummings et al. are typical for this type of study. As such the issues raised are general to the field and not a specific critique of this paper.

Figure 3 highlights an example of differences that can occur when different 3D model builders are applied. The primary change shown for this molecule is in the torsion of the guanidine group. This of potential importance for the Thrombin searches due the critical nature of the group for binding. Note also the difference shown for the guanidine carbon to amine distance. This is in part due to systematic

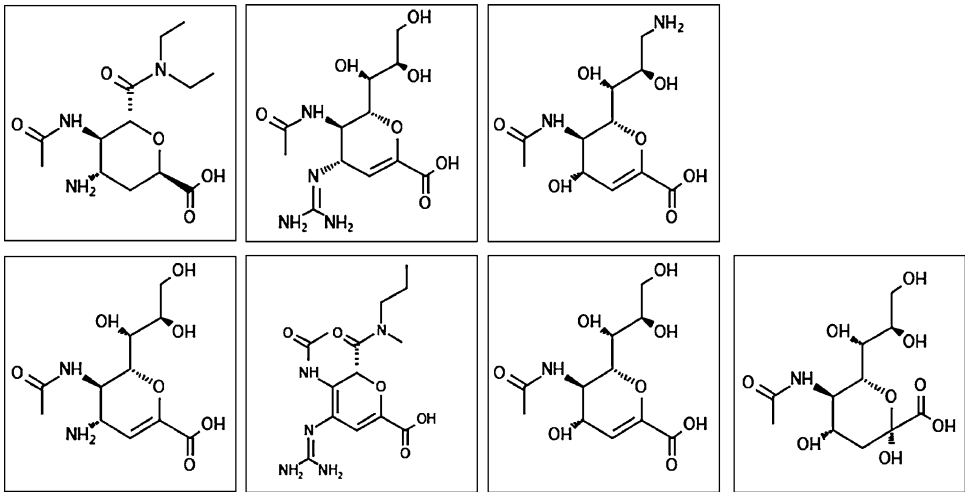

Fig. 7 Neuraminidase actives data set used by Pham and Jain [35]

disparities in bond lengths between LIGPREP and CORINA of around 0.01–0.02 Å per bond. While this change is small, when combined with other parameter differences it can manifest significant changes in conformation. For docking techniques such as GLIDE and DOCK that apply hard repulsion terms this has the potential to produce concomitant alterations in hit list ranking. These issues are highlighted by the changes in enrichment seen between CORINA and LIGPREP searches shown in Fig. 4. The data also illustrate the potential effects of applying different versions of a rapidly evolving program such as GLIDE to the same problem. For Thrombin, selection of GLIDE version 4 alone produces a slight decrease in performance, suggesting that this data set has not figured in scoring function calibration (a potential issue touched on elsewhere [42]). The version change in conjunction with the application of LIGPREP models produces defined benefits across the board, however. In the case of PTP1B, GLIDE version 4 produces significant increases in enrichment, with the addition of LIGPREP models only improving results at the top of the hit list (0.5%). These results illustrate another important point for the readers consideration, that of enrichment cutoffs. For their comparisons Cummings and coworkers analyzed enrichments at database cutoffs >2%. It is likely that this choice has been made as there studies involve relatively small data set sizes (5–10 actives and 1,000 decoys). The translation of these cutoffs to a real world virtual screen is problematic, however, since such screens typically involve databases in excess of 10^6 molecules in size. A 2% cutoff in this context thus represents over 20,000 compounds, which is already at the outer edge of search space for most virtual screens. It is debatable whether cutoffs greater than this have relevance unless extremely impressive cherry picking facilities are available. Nevertheless cutoffs as high as 10% have been applied elsewhere [21]. With this in mind smaller cutoffs at 0.5–1% as shown in Fig. 4 have been applied for PTP1B to highlight differences in regions of ranking space more relevant to a typical screen.

Figure 5 highlights what can happen as the input parameters of a docking study are varied. For DOCK ROCS-like mode, the screen has been modified to restrict DOCK search space to the region of the template ligand. In both the Thrombin and PTP1B cases, the resultant searches (see methods) produce enrichments superior to both the ROCS searches and those originally undertaken by Cummings and coworkers. For programs like DOCK which contain many parameters that can fundamentally alter results, the importance of this problem can not be overstated. It also highlights some of the potential dangers in comparing ligand-based searches with those that are protein-based without careful reference to set up. Good et al. highlighted the results variation possible in DOCK in their studies focused on pharmacophore constraint performance [43]. DOCK has been applied in a wide variety of ways in other studies focused on enrichment. For example, Bissantz et al. [17] used the geometry searching with core fragment placement for conformational searching, while Charifson et al. [16] applied pharmacophore constraints with pregenerated conformations for their studies. The work of Charifson is more reflective of the search customization used in real world screens, while Bissantz essentially applied the program in the default mode provided by the software.

The improved enrichments seen in DOCK ROCS-like search results speak to the value of customized application, highlighting the potential dangers in interpreting default search results. This argument could be further extended to make the assertion that DOCK is outperforming ROCS when applied in a manner more consistent with the ligand constraints intrinsic to a ROCS search. Such an assertion is not without risk, however, given that both target studied involve a key salt bridge interaction. Electrostatic potential (ESP) is highly sensitive to such interactions which favors

Table 2 DUD clustering data

Target	Total ligands in DUD sets	Ligands passing "lead-like" filter	"Lead-like" filtered reduced graph clusters
COX-2	412	250	44
EGFR	458	379	40
CDK2	58	55	32
VEGFR2	78	49	31
INHA	86	58	23
PDE5	76	34	22
PDGFRB	169	136	22
SRC	159	102	21
P38	353	219	20
FXA	146		19
ACE	49	46	18
ACHE	106	101	18
HIVRT	41	35	17
ALR2	26	26	14
DHFR	407	387	14
THROMBIN	68	26	14
FGFR1	170	73	12
COX-1	24	23	11
AR	74	63	10
ER agonist	67	63	10
GPB	52	52	10
ADA	37	37	8
ER antagonist	39	18	8
NA	49	49	7
PARP	35	33	7
TK	22	22	7
TRYPSIN	46	10	7
AMPC	21	21	6
PPAR gamma	82	7	6
GART	31	13	5
HMGA	35	25	4
HSP90	25	24	4
PNP	30	30	4
PR	27	22	4
HIVPR	62	6	3
RXR alpha	20	18	3
COMT	11	11	2
GR	78	9	2
MR	15	13	2
SAHH	33	33	2
Average	94	66	13

DOCK, since by default it includes an ESP term in its scoring function. Further, the data in Table 1 highlights the fact that around 7% of the decoys used in these enrichment studies while potentially acidic, have been left protonated. This has implications for PTP1B enrichment, since the lack of a charge will lower the ranking of these molecules when ESP is used, potentially reducing decoy noise. These results illustrate the effect decoy database set up can have on enrichment outcomes. With this in mind, Fig. 5 further highlights the variation possible in result depending on other related aspects of set up. DOCK scores excluding ESP on PTP1B significantly reduce enrichment. Such calculations are often undertaken when tight pharmacophore constraints have been applied instead to produce the required hydrogen bonding interactions. This is not the case for these searches where the constraints applied are both generic and loose. As such this search is also an unfair comparison since electrostatic acts as DOCK's color force field equivalent. Nevertheless one can argue that the result highlights the upper bound of the noise effect. Normalizing the search result produces significant improvements in PTP1B enrichment both in ESP and non-ESP mode. This method is often used in screens where smaller (high efficiency) leads are preferred. Correcting for binding efficiency in this way works well as many of the PTP1B actives are small. It is no universal panacea, however, as can be seen in Thrombin, where normalization decreases enrichment as the actives are much larger. These results are not designed to highlight the best methods or set up. Rather they point to the difficulty of applying the one size fits all approach to parameterization typically used in enrichment search set ups. Further they point to the problems intrinsic in data interpretation due to variation in primary target site interactions and ligand property distributions.

The PTP1B data has other issues associated with it, including the fact that a number of the actives are close analogues of each other. The difficulty of interpreting results from such data sets has already been expounded upon [44], but it is an issue of sufficient importance to warrant further analysis. The utility of 2D searches as enrichment controls has been mentioned elsewhere (Christopher Murray Astex, personal communications 2005). Figure 6 highlights this concept, comparing 3D enrichments with those of the 2D FCFP2 descriptors. It can be seen that not all 3D methods outperform the 2D technique, highlighting potential analogue bias in the data as much as any shortcomings in screening technique. This is not a reflection on the relative merits of 2D versus 3D methods, both of which have significant utility and only limited information overlap [45]. Rather, the results highlight potential issues in interpreting said merits in the context of these target data.

The analogue bias issue is perhaps most clearly illustrated by FCFP2 searches against the neuraminidase structures shown in Fig. 7. This data comes from the work of Pham and Jain [35], and the search against this target was repeated using the ligand abstracted from the template protein and decoy set detailed in the article. Note that the

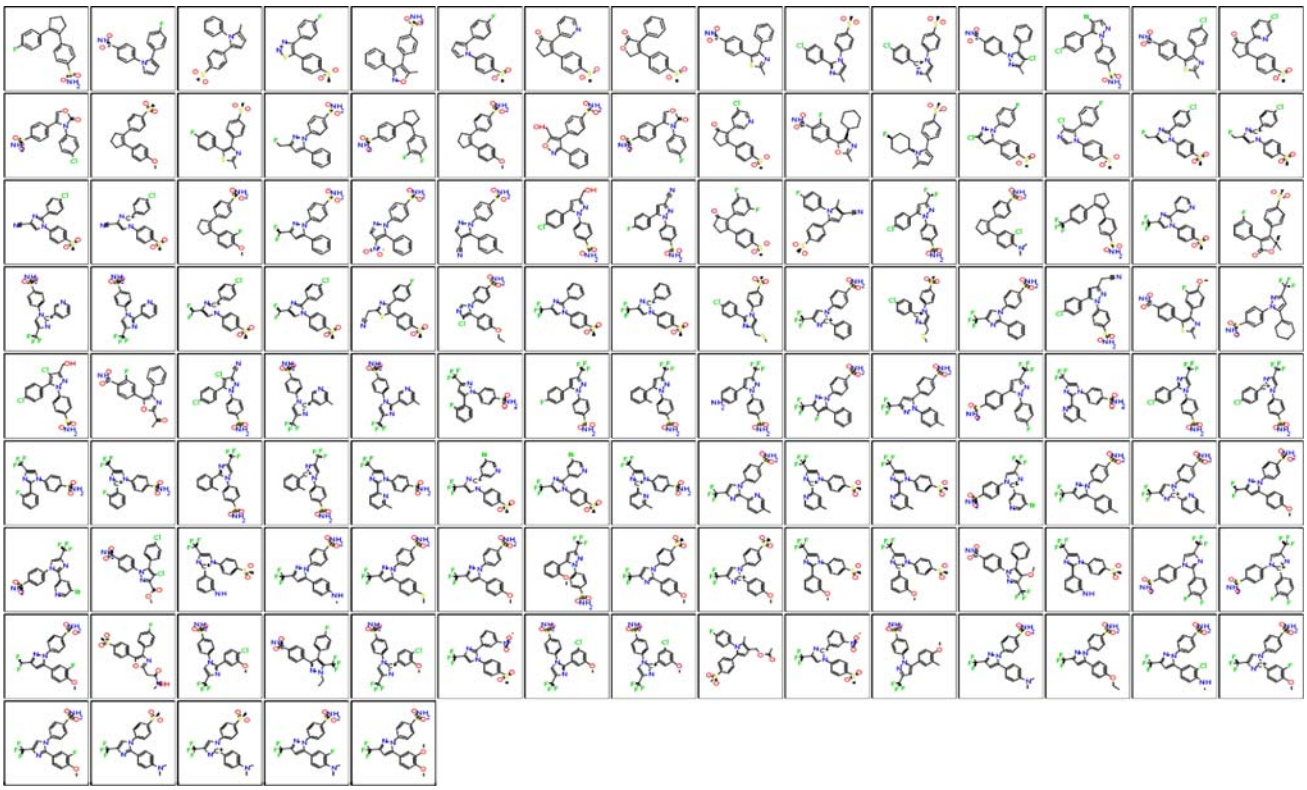

Fig. 8 DUD Cox-2 cluster compounds sharing the Vioxx/Celebrex reduced graph (~25% of the Cox-2 set)

Table 3 WOMBAT cluster sizes after application of lead-like and reduced graph filters

Target	Wombat	DUD
CDK-2	152	32
EGFR	74	40
P38	59	20
AR	36	10
PPARG	27	6
ER alpha	64	8
COX-2	76	44
ALR2	42	14
PDE5	88	22
HIV-RT	99	17
FXA	107	19
IMPDH	49	
D2 antagonists	323	
Average	75	21

DUD cluster sizes for equivalent targets given where available. Note that the Averages are given for data sets where both WOMBAT and DUD data set information are available

neuraminidase data in this study was used in a docking experiment. Nevertheless it nicely highlights the magnitude of effect possible with ligand-based retrieval methods when analog bias is present. For this search 6 of the top 8 and all 7 of the top 13 structures were found to be active, enrichment rates that would be all but impossible with most 3D methods. Neuraminidase has been used in a variety of enrichment studies [22, 23, 34, 37], and one could contend that the data sets chosen in these studies are potentially more relevant to virtual screening analysis. The authors would argue, however, that the nature of the target and its binding requirements (small active site and highly engineered substrate mimics) does not lend itself to the construction of diverse chemotypes and by extension useful virtual screening tests.

The issues raised with regard to neuraminidase highlight that it is not only analogue bias that needs to be considered when determining the suitability of a given target for enrichment studies. Readers may have noticed that no analysis has been made of the HIV protease results produced by Cummings and coworkers. The target is a flagship of structure-based design with many published complexes, and is thus popular choice for enrichment comparisons. It is, however, unsuitable for such studies for a variety of reasons. First the active compounds are generally not docking friendly. This can be seen in the Cummings HIV data set, where one molecule contains a flexible seven member ring with three substituents, two of which form part of a chiral center. Such a molecule requires myriad ring conformation starting points if there is

to be any real hope of successful docking, but only one was constructed for the docking studies. Another of the actives contains nineteen flexible bonds. The resulting combinatorial explosion renders the conformational search of this molecule at best highly problematic. In addition the physical properties of the actives place them on the outer bounds of space relative to the noise molecules. As such simple descriptors such as molecular weight are better able to distinguish the actives from the decoys than any 2D or 3D method. The source for the majority of these issues is intrinsic to the peptide-mimetic ancestry central most HIV protease inhibitors. Thus, as with neuraminidase, such molecules are generally best avoided if for no other reason than that there is very little chance that they would ever be discovered in a real world screen.

It is important to note that protein docking studies also have the potential to be effected by such bias, although its effects are more subtle. For example GLIDE protein preparation minimizes the active site with respect to its cognate ligand. As such the active site will be optimized to become a cast for fitting close analogues of said structure. Further, should a particular program be optimized to abstract a dominant chemotype from a given data set, its headline hit rate will look good even though the number of chemotypes retrieved may be limited. With this in mind, the problem of potential analogue bias in the DUD sets has been examined (Table 2). The application of a lead-like filter only reduces average ligand data set size by a third. When combined with the reduced graph cluster filter, however, a further five fold reduction is observed, with a cluster count average of 13 versus the original data set count of 94. The kinase data sets hold up best, reflecting a diverse chemotype membership. PDE5 and COX-2 are also well represented, although Fig. 8 highlights that large clusters of related compounds still abound within said sets. The serine protease class is also reasonably well stocked with chemotypes, though users are cautioned that these are dominated by arginine mimics (this is also the case with the WOMBAT data). Said mimics dominate binding and make it easy for certain scoring functions that weight such salt bridge interactions heavily. Issues such as these are not uncommon when data sets are investigated in depth. As such those using DUD/WOMBAT sets are encouraged to include supplemental data in their publication highlighting the molecules and orientations found, and inspect said results carefully for unforeseen artifacts.

Thirteen targets have been culled for data from the WOMBAT database. The targets represent nine classes, with both kinases and nuclear hormone receptors represented in triplicate. This has been done both to allow tests for target class scoring function bias and to reflect the general focus of CADD projects. A larger set of dopamine D2 antagonists has also been included to permit GPCR homology model testing and ligand-based virtual screening on a very large and diverse set of chemotypes. As can be seen from Table 3, where overlap exists with the DUD data, WOMBAT sets are on average over three times larger in size, and as such represent a potentially useful resource for more extensive enrichment testing. These data sets also include original reference and activity data to allow more in depth analysis of enrichment results. Note that the data sets have been published on the web to allow easy public access [28].

Conclusions

The studies detailed above highlight a number of ways in which the myriad variables intrinsic to a virtual screen can affect screening enrichment. Parameters as diverse as docking tool set up, active compound collections, decoy set construction choices and target selections all have the ability to confound results interpretability. Central to these problems is the difficulty of direct comparison due to the relative dearth of diverse publicly available data sets. The DUD repository goes someway to alleviating this, though it too is not without shortcomings. With this in mind we have created clustered versions of DUD to permit cleaner enrichment comparisons. In addition new larger sets have been collated from WOMBAT to supplement the DUD data. It is hoped that these sets will form the baseline tests for subsequent enrichment studies. This will allow repeated application using different techniques and set ups, by extension permitting the ongoing comparisons critical to objective assessment of enrichment performance.

Acknowledgments This work was supported in part by support the New Mexico Tobacco Settlement fund (TIO). Thanks go to Andrei Leitão of UNM Biocomputing for his help in differentiating WOMBAT HIV NNRTI/NRTI compounds.

References

1. Gruneberg S, Wendt B, Klebe G (2001) Angew Chem Int Ed Engl 40:389
2. Wu JH, Batist G (2001) Anticancer Drug Design 16:129
3. Paiva AM, Vanderwall DE, Blanchard JS, Kozarich JW, Williamson JM, Kelly TM (2001) Biochim Biophys Acta 1545:67
4. Perola E, Xu K, Kollmeyer TM, Kaufmann SH, Prendergast FG, Pang Y-P (2000) J Med Chem 43:401
5. Evers A, Klebe G (2004) J Med Chem 47:5381
6. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT, Shoichet BK (2002) J Med Chem 45:2213
7. Brenk R, Naerum L, Gradler U, Gerber H-D, Garcia GA, Reuter K, Stubbs MT, Klebe G (2003) J Med Chem 46:1133
8. Kitchen DB, Deconrez H, Furr JR, Bajorath J (2004) Nat Rev Drug Discov 3:935
9. Alvarez JC (2004) Curr Opin Chem Biol 8:365

10. Shoichet BK (2004) Nature 432:862
11. Barril X, Hubbard RE, Morley SD (2004) Mini Rev Med Chem 4:779
12. Jansen JM, Martin EJ (2004) Curr Opin Chem Biol 8:359
13. Krovat EM, Steindl T, Langer T (2005) Curr Comput Aided Drug Des 1:93
14. Mohan V, Gibbs AC, Cummings MD, Jaeger EP, DesJarlais RL (2005) Curr Pharm Des 11:323
15. Rajamani R, Good AC (2007) Curr Opin Drug Discov Devel 10:308
16. Charifsen PS, Corkery JJ, Murcko MA, Patrick Walters W (1999) J Med Chem 42:5100
17. Bissantz C, Folkers G, Rognan D (2000) J Med Chem 43:4759
18. Miteva MA, Lee WH, Montes MO, Villoutriex BO (2005) J Med Chem 48:6012
19. Cummings MD, DesJarlais RL, Gibbs AC, Mohan V, Jaeger EP (2005) J Med Chem 48:962
20. Evers A, Hessler G, Matter H, Klabunde T (2005) J Med Chem 48:5448
21. Chen H, Lyne PD, Giordanetto F, Lovell T, Li J (2006) J Chem Inf Model 46:401
22. Warren GL, Andrews CW, Cappelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) J Med Chem 49:5912
23. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon J-F, Cornell WD (2007) J Chem Inf and Model 47:1504
24. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49:6789
25. http://www.dud.docking.org/ 10/07
26. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI (2004) In: Oprea TI (ed) WOMBAT: world of molecular bioactivity, in chemoinformatics in drug discovery. Wiley-VCH, New York, pp 223–239
27. http://www.sunsetmolecular.com/products/?id=4 10/07
28. http://www.dud.docking.org/wombat/accessed 10/07
29. LIGPREP, developed and distributed by Schrodinger Inc.: http://www.schrodinger.com/ProductDescription.php?mID=6&sID=7&cID=0 10/07
30. Sadowski J, Gasteiger J (1993) J Chem Rev 7:2567
31. Pearlman RS (1987) Chem Des Autom News 2:5
32. ROCS, OMEGA and the OECHEM toolkit are all developed and distributed by Openeye Scientific Software Inc: http://www.eyesopen.com/products 10/07
33. Pipeline Pilot, developed by Scitegic: http://www.scitegic.com/ 10/07
34. Hawkins PCD, Skillman AG, Nicholls A (2007) J Med Chem 50:74
35. Pham TA, Jain AN (2006) J Med Chem 49:5856
36. http://www.biopharmics.com/downloads.html accessed 11/07
37. Steindl T, Langer T (2004) J Chem Inf Comput Sci 44:1849
38. Oprea TI, Davis AM, Teague SJ, Leeson PD (2001) J Chem Inf Comput Sci 41:1308
39. Ghose AK, Pritchett A, Crippen GM (1988) J Comput Chem 9:80
40. Barnard JM (1993) J Chem Inf Comput Sci 33:532
41. Gillet VJ, Willett P, Bradshaw J (2003) J Chem Inf Comput Sci 43:346
42. Jain AN (2007) 234th ACS National Meeting, Boston, USA, Aug 19–23, COMP-147
43. Smith R, Hubbard RE, Gschwend DA, Leach AR, Good AC (2003) J Mol Graph Model 22:41
44. Good AC, Hermsmeier MA, Hindle SA (2004) J Comput Aided Mol Des 18:529
45. Oprea TI (2002) J Braz Chem Soc 13:811