J Comput Aided Mol Des (2012) 26:897–906
DOI 10.1007/s10822-012-9584-8

FRED and HYBRID docking performance on standardized datasets

Mark McGann

Received: 30 April 2012 / Accepted: 16 May 2012 / Published online: 5 June 2012

Abstract The docking performance of the FRED and HYBRID programs are evaluated on two standardized datasets from the Docking and Scoring Symposium of the ACS Spring 2011 national meeting. The evaluation includes cognate docking and virtual screening performance. FRED docks 70 % of the structures to within 2 Å in the cognate docking test. In the virtual screening test, FRED is found to have a mean AUC of 0.75. The HYBRID program uses a modified version of FRED's algorithm that uses both ligand- and structure-based information to dock molecules, which increases its mean AUC to 0.78. HYBRID can also implicitly account for protein flexibility by making use of multiple crystal structures. Using multiple crystal structures improves HYBRID's performance (mean AUC 0.80) with a negligible increase in docking time ($\sim$15 %).

Keywords Docking · Virtual screening · FRED · HYBRID · DUD · Protein flexibility

Introduction

Docking programs are useful in the drug discovery process for both virtual screening and lead optimization [1] Virtual screening is the process of identifying new potential hit molecules or binders from a large database of compounds. Docking programs traditionally use the structure of the target protein to do virtual screening by examining the interactions between ligands in the compound database and the protein receptor site. The probability of a molecule being a binder is determined with a scoring function that converts these ligand–protein interactions into a single numerical score that (hopefully) correlates with potency. In lead optimization, a potential drug candidate is known but is deficient with respect to certain properties, commonly potency or ADME properties. Docking assists in the lead optimization process by giving insight into how modifications to the lead compound will change how the compound interacts with the receptor and thus how the potency will be altered by these changes.

Performance is an important criterion when selecting a docking program. While other factors, such as cost, speed and ease of use are also of practical concern, a docking program must perform well to be seriously considered. For virtual screening, this means that the docking program should rank active compounds higher than inactive compounds. Lead optimization's emphasis, however, is not on always correctly ranking compounds by potency but on correctly posing the compounds in the receptor site, as the docked pose will be closely examined to determine how best to modify it to improve potency. Correct posing is also important for virtual screening, as an incorrectly docked pose will presumably not be scored and ranked correctly.

One of the difficulties in assessing the performance of docking programs is that while they often perform reasonably well on average, the performance of a given program varies a great deal from target to target [2, 3] This is especially problematic when comparing the performance of two docking programs that use different test systems, as the choice of test system can have a significant impact on the

M. McGann (✉)
OpenEye Scientific Software, 9 Bisbee Court Suite D, Santa Fe, NM 87508, USA
e-mail: mcgann@eyesopen.com

Present Address:
M. McGann
OpenEye Scientific Software, 222 3rd Street Suite 3120, Cambridge, MA 02142, USA

results [4]. To combat this, test systems for docking tend to have large numbers of protein–ligand complexes, since a larger test system improves confidence in the mean performance. Using the same test dataset for all docking programs improves the reliability of performance comparisons, as the performance of docking programs is generally positively correlated—that is, a particular target tends to be easy for most programs or difficult for most programs [5].

Underlying the discussion of the performance of docking programs is the assumption that the mean performance is what matters. One might argue, however, that the consistency or reliability of a program is also an important measure of performance. For instance, a program that has an average performance that is 10 % worse than another program might be desirable if the program is extremely consistent: i.e., it performs the same on all targets. An analogy from the world of finance would be comparing the investment return of a certificate of deposit (CD) to investing in the stock market. On average, the stock market will perform better than a CD, but the CD's performance is guaranteed and may be preferable to a risk-adverse investor. While this concept is important, the results presented herein deal primarily with mean performance.

This paper presents the results from the latest versions of the FRED and HYBRID docking programs from OpenEye's OEDocking suite on a standardized set of test systems from the Spring ACS 2011 Docking Symposium. While these standardized test systems have some issues, as discussed later, the value of having results that can more reliably be compared with those of other docking programs overrides these concerns. FRED is a true docking program that uses only the structure of the target protein to pose and score molecules. HYBRID uses both the structure of the target protein and the structure of the ligand bound to the active site to pose and score ligands. The structure of a bound ligand is required to use HYBRID; this information is generally available as proteins are commonly crystallized in the presence of a known binding ligand.

HYBRID is also capable of using multiple conformations of the target protein if more than one crystallographic structure is available. To explore this capability, we have also extended the docking symposium's virtual screening dataset by adding additional protein structures for each target. The results of this extended test dataset are presented in addition to the results of the standardized set for both FRED and HYBRID.

Theory

FRED and HYBRID use an exhaustive search algorithm to dock molecules. Both programs treat ligand conformers as rigid during the docking process, although ligand flexibility is implicitly included by docking multiple conformers of each ligand. The protein structure is also treated as rigid during the docking process for both FRED and HYBRID; however, HYBRID is capable of using multiple conformers of the target protein and therefore can account for protein flexibility as well.

The details of FRED's docking algorithm have been described previously [2]; briefly, FRED takes as input the protein structure and a multi-conformer representation of the ligand to be docked. In the exhaustive search, each ligand conformation is systematically rotated and translated within the active site at a resolution of 1 Å. Every pose that passes a bump check is scored. The top scoring poses, across all conformers of the ligand, are refined by testing nearby rotations and translations at a resolution of 0.5 Å.

The current version of FRED, 3.0, uses the Chemgauss4 scoring function for the final refinement and scoring of molecules and the Chemgauss3 scoring function to score molecules during the initial exhaustive search. Chemgauss3 has been described previously [3]. Chemgauss4 is an evolution of the Chemgauss3 scoring function that has improved recognition of hydrogen bond geometry and also recognizes hydrogen bond networks. Both scoring functions account for hydrogen bond interactions, metal-chelator interactions, desolvation effects and the shape complementarity of the ligand to the active site.

HYBRID's docking and scoring algorithm is identical to FRED's, except that the scoring function FRED uses during the exhaustive search, Chemgauss3, is replaced by a ligand-based scoring function, the Chemical Gaussian Overlay (CGO). CGO scores based on how well the docked molecule matches the shape and 3D arrangement of chemical features of the crystallographic ligand bound to the active site, rather than how well the docked molecule complements the active site. The exhaustive search is therefore strongly biased towards docked poses with a similar binding mode to the crystallographic ligand. Once the exhaustive search is complete, the remainder of HYBRID's docking process is identical to FRED's: each pose is refined by testing nearby rotations and translations at a resolution of 0.5 Å and scored using the Chemgauss4 scoring function.

HYBRID also has the ability to accept multiple structures of the target protein (each must have a bound ligand structure). When supplied with multiple structures of the target protein, each ligand is docked to only one protein structure. The protein structure to which a given molecule is docked is determined by comparing the shape and chemistry of each molecule to the crystallographic ligand of each protein. The molecule is then docked to the protein structure with the most similar crystallographic ligand, scored by shape and chemical similarity in 3D. The

underlying assumption of this method is that the protein conformation most appropriate for docking the ligand is the protein with the most similar ligand bound to it [6]. This method also has the advantage of not significantly increasing the run time of the docking, as each molecule is only docked once. The only additional computational cost is the comparison of the docking ligand to the crystallographic ligands, which is about 20-fold faster than the docking itself.

Experimental

FRED and HYBRID setup

The same setup is used for the cognate docking and virtual screening tests for both FRED and HYBRID. OMEGA 2.4 is used with all default parameters to conformationally expand the docking molecules [7, 8].

The PDB (Protein Data Bank) structures are converted into receptor format using the pdb2receptor program that is included in OEDocking 3.0 distribution with FRED and HYBRID. Both FRED and HYBRID are run with their default settings.

MACCS

For a performance baseline, this work also includes virtual screening results using MACCS keys, a ligand graph similarity measure [9]. MACCS keys are 166 bit structural key descriptors in which each bit is associated with a SMARTS pattern. Each ligand's score is the Tanimoto coefficient between the docked molecule and crystallographic ligand MACCS key descriptor. The MACCS key descriptors and Tanimoto coefficients were calculated with a Python script using the February 2012 release of the OpenEye GraphSim Toolkit [10].

Cognate docking

The tests system cognate docking results for this work is a set of 85 ligand receptor crystal structures used in the Spring 2011 ACS Docking and Scoring Symposium. These structures are the same structures found in the Astex structure reproduction dataset [11]. About 25 % of these structures were noted by the organizers to be questionable in terms of bond ordering, electron density and other factors. These structures are used as provided since the goal is to present results on standard datasets to facilitate comparison, and any fixes/modifications to the dataset will likely not be the same for different groups also performing this analysis.

Cognate docking performance is measured using RMSD between the docked pose and the crystallographic bound ligand. The crystallographic ligand is not optimized versus the scoring function (or any function). To do so would introduce a favorable bias into the results that could not be used in real-world applications where the correct structure is not known beforehand [12].

Virtual screening test sets

Virtual screening performance is tested using the DUD dataset, which consists of 40 protein targets (each with one structure of the target protein), a set of molecules known to be active against the target and a set of decoy molecules selected to match the actives in terms of a number of physio-chemical properties [13]. These decoy molecules are presumed to be inactive, although their activity against the target is not known. A crystallographic ligand is present in the active site of each target protein. The crystallographic ligand is saved as part of the receptor file generated by the pdb2receptor and used by the HYBRID program to enhance docking performance (as described in the Theory section above). FRED does not require the crystallographic ligand information; if it is present in the receptor file, FRED ignores it.

The version of the DUD dataset supplied by the symposium organizers also included an alternate set of actives for 10 of the targets that are drawn from the Wombat database (the same decoys are used) [14]. Results using these actives were not requested for the symposium, but have been requested here for this special issue. This dataset is referred to as DUD-Wombat in the results and discussion.

In addition to the Wombat actives variation of the DUD dataset, we have also included results from a variation of the DUD dataset that has several different protein structures for each target. The additional protein structures for each protein target were drawn from the PDB using the following rules:

1. Between four and seven structures were chosen per target;
2. Only structures that included a crystallographic ligand were selected;
3. Structures with better resolution were favored;
4. Structures were chosen to provide a diversity of ligand structures if possible (similarity estimated visually from the ligand depiction);
5. If structures were available from several research groups, selecting structures from multiple groups was preferred.

Once the structures were chosen with these rules, the total number of targets in the dataset was 38. One target,

pdgfrb, has no crystal structures available in the PDB (the structure in the DUD dataset is a homology model). The second target, Estrogen, is present twice in the DUD dataset, once in agonist form and once in antagonist form, and was combined into a single target containing both the agonist and antagonist structures, actives and decoys. The multi-protein structure variant of the DUD dataset is referred to as MDUD and MDUD-Wombat, respectively, when using the standard DUD actives and Wombat actives. The standard DUD decoys are used for both MDUD and MDUD-Wombat.

Virtual screening performance is measured using AUC (Area Under the Receiver Operator Characteristic) and early enrichment at 2, 1 and 0.1 %. AUC is a global enrichment metric [5]. The early enrichment metric used by the symposium is the fraction of the area under the ROC (Receiver Operator Characteristic) curve from 0 % decoys recovered to early enrichment value (e.g., 2, 1 or 0.1 %). We also normalize this value by the expected fraction of actives recovered if the actives are evenly distributed (i.e., no enrichment), leading to the following formula for fractional Area Under the Curve (fAUC).

$$\mathrm{f\,AUC(x)} = 2\frac{\int_0^{\mathrm{x}} \mathrm{ROC(x')dx'}}{\mathrm{x}^2} \quad (1)$$

where x is the fraction of decoys recovered (e.g., 2, 1 or 0.1 % herein).

Early enrichment metrics have less statistical power and are much more fragile than global enrichment metrics like AUC [2]. The 0.1 % early enrichment metric is particular suspect in this testing, as many of the virtual screening targets have less than 1,000 compounds in their dataset and hence the top 0.1 % of the hit list has less than one compound. Nevertheless, in the interest of reporting standardized results, we include results using both early and global enrichment metrics.

Results and discussion

Cognate docking

FRED's top scoring pose is within 2, 1.5, 1.0 and 0.5 Å of the crystallographic pose for 70, 66, 50 and 9 % of the cognate docking test systems, respectively (see Fig. 1). The mean and median RMSD are 1.9 and 1.1 Å, respectively. The mean and median values differ significantly because the distribution of RMSDs is highly non-normal. Of the two, median RMSD is the more trustworthy number. Mean RMSD is of limited value since it is influenced highly by large RMSDs that have little meaning (e.g., 6 and 10 Å are both incorrect docking; 10 Å really isn't any more wrong than 6 Å, but the difference will have a significant effect on the mean RMSD).

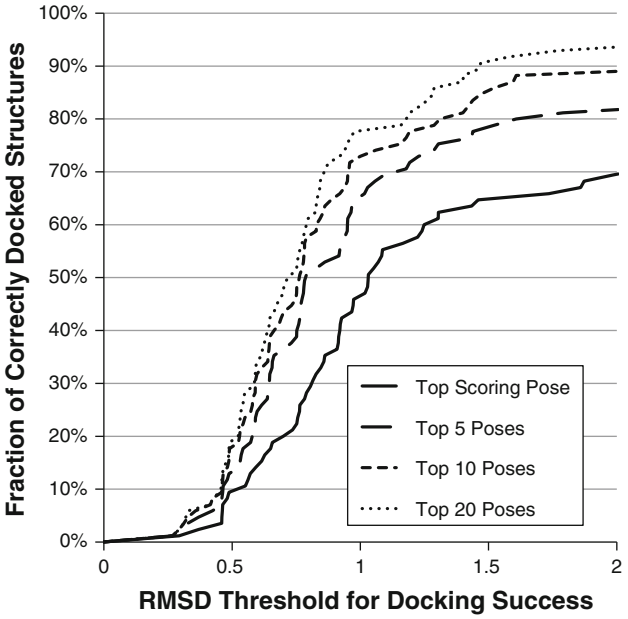

Fig. 1 Pose reproduction success rates for the top 20 (*dots*), 10 (*short dashes*), five (*long dashes*) and top scoring (*solid*) poses. Success rates are given as the fraction of co-crystal complexes (y-axis) in which the docked ligand structure is within a given RMSD (x-axis) of the crystallographic ligand structure

The success rate at 0.5 Å (9 %) is sharply lower than the other success rates (70, 66 and 50 %) because 0.5 Å is below the resolution of the Omega/FRED docking algorithm. FRED docks molecules using an exhaustive search at 1 Å resolution followed by a local optimization at 0.5 Å resolution. Additionally, the input conformers generated by Omega that FRED rigidly docks do not exactly match the X-ray conformation (the median best overlay RMSD of the Omega conformers is 0.4 Å for this dataset). Combining these two factors, the effective resolution of the docking algorithm is closer to 1 Å, where the overall success rate is a more reasonable 50 %. It is possible to increase the resolution of the docking algorithm; however, this is extremely costly in terms of CPU time (it varies by $1/N^6$, where N is the resolution). In addition, in real-world docking scenarios where ligands are cross docked, there will be errors in the protein coordinates due to both the rigid protein approximation and experimental error from crystallography (for the 38 DUD protein structures coordinate error is >0.5 Å for 18 of the structures for which sufficient data exists to calculate the coordinate error—unpublished data). As such docking to resolutions below 0.5 Å is an unrealistic use case.

A pose within 2 Å of the crystallographic pose exists in the top five, 10 and 20 scoring poses for 82, 89 and 94 % of the test systems, respectively (see Fig. 1). Of the six systems where a pose within 2 Å was not found in the top 20 poses, five of those systems (1t40, 1hvy, 1lpz, 1sj0 and

1y6b) had no Omega conformer within 1 Å RMSD overlaid. Of the other 79 systems, there were only two where the best Omega conformer was greater than 1 Å. Thus, having a good Omega conformer (i.e., less than 1 Å) correlates highly with successful docking.

The one system where a pose within 2 Å was not found in the top 20 poses and there was a good Omega conformer was 1gm8. In this system, the ligand's phenyl ring is buried in a hydrophobic pocket while a polar tail sticks out into solvent (see Fig. 2a). FRED correctly places the phenyl ring in this system; however, it prefers to have the polar end of the ligand make several hydrogen bonding interactions with the protein just outside the hydrophobic pocket (see Fig. 2b). FRED's Chemgauss4 scoring function does account for both protein and ligand desolvation; in this particular case, however, the desolvation terms are not strong enough to overcome the hydrogen bonding potential in the ligand tail; therefore no pose in the top 20 is correctly docked. While close examination shows that there are several correctly docked structures in the top 100 poses, these poses tend to score poorly.

Docking failures fall into two categories: search failures and score failures. The case of 1gm8, described above, is an example of a scoring failure. A correctly docked structure was examined by the docking algorithm but not recognized as correct because it scored poorly relative to several incorrectly docked poses. In contrast, a search failure occurs when the docking algorithm never examines

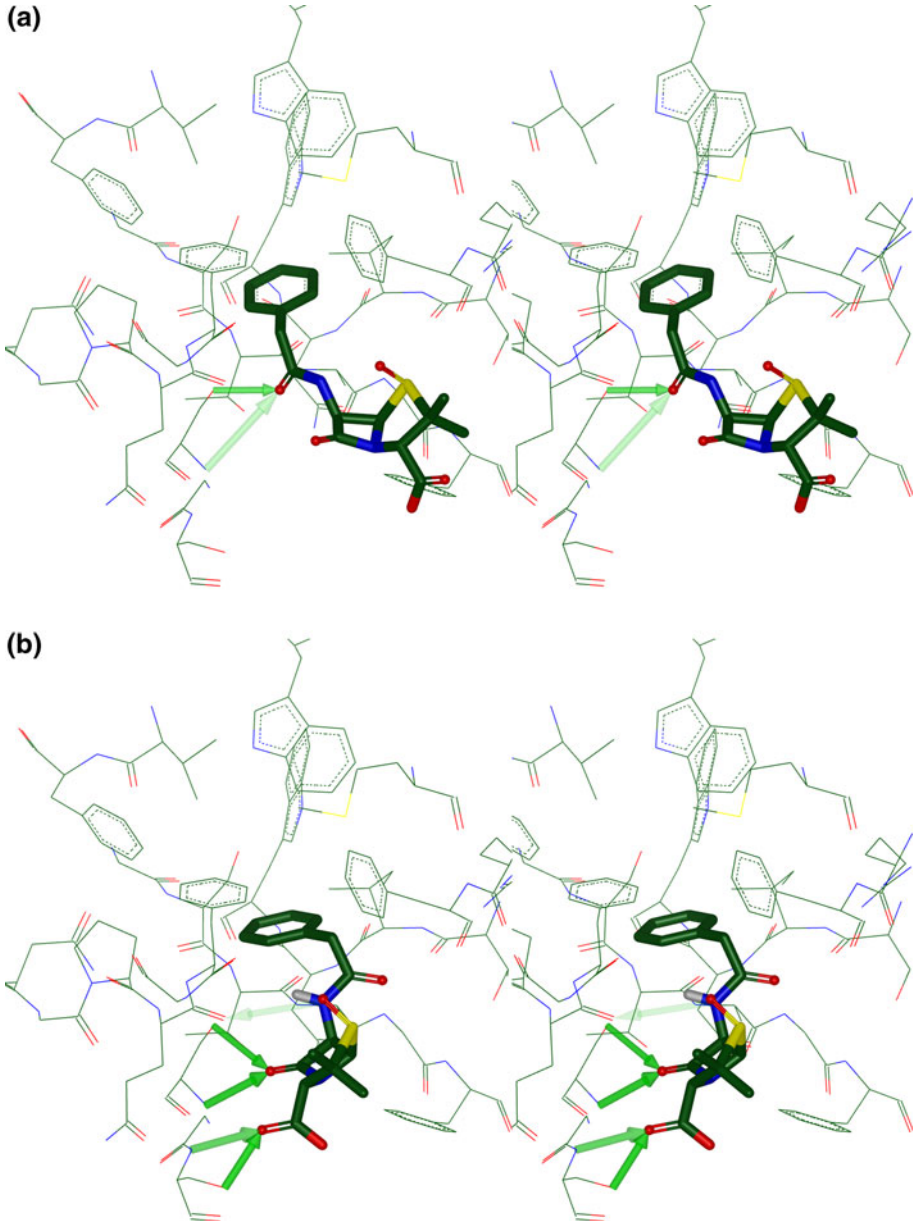

Fig. 2 Structure of pose reproduction target 1gm8 with: **a** The crystallographic pose of the ligand. **b** The pose of the top scoring ligand docked by FRED

the correctly docked structure. An example of a search failure is the 1y6b system. The best Omega conformer of the 1y6b ligand is 2.2 Å RMSD when overlaid onto the crystallographic pose. Since FRED docks by rigidly rotating and translating the conformers in the active site, it is impossible for FRED to generate a pose that is less than 2.2 Å RMSD to the crystallographic pose; thus, docking fails due to a search failure for this system. For the purposes of this paper, we classify a failure as a search failure if the top scoring docked pose is incorrect but does not score as well as the crystallographic ligand and a score failure if it scores better than the crystallographic ligand [3].

Docking success, search failures and score failures are shown in Fig. 3. The score failure rate at 2.0 Å is 10 % while the search failure rate is 20 %. As the RMSD threshold for success decreases, both search and score failures increase, although search failures increase at a greater rate. This is especially evident at the 0.5 Å RMSD threshold, where 70 % of dockings result in search failures compared to 20 % score failures. The high rate of search failures at 0.5 Å RMSD threshold is due to the fact that, as described above, 0.5 Å is below the resolution of FRED's docking algorithm.

No docking program is perfect, and it is useful to be able to predict a priori which systems are likely to be difficult docking targets. One reasonable hypothesis is that docking systems with larger search spaces will be more difficult docking targets. The number of rotatable bonds in the ligand and the size of the active site are both easy to measure beforehand and have a direct effect on the size of the docking search space. Figure 4 shows the results (i.e., success, search failure or score failure) of docking to each target in the test set as a function of the size of the active site and the number of ligand rotatable bonds. Systems with large highly flexible ligands and large active sites are the most difficult to dock to, although there is still a fair number of docking success. Notably, there are very few scoring failures in these systems with large search spaces. Search failures can sometimes mask score failures (if the docking algorithm doesn't examine a high scoring but incorrect pose), but it is still somewhat odd to find no score failures in systems with active sites larger than 1,750 Å or ligands with more than seven rotatable bonds. We hypothesize that these systems tend to have large ligands that have many interactions. The large number of interactions tends to minimize the effects of any incorrectly scored individual interactions.

Virtual screening

HYBRID outperforms FRED on the standard DUD dataset. The mean AUC of HYBRID is 0.78 while FRED's is 0.75 (see Fig. 5a). While relatively small, this difference is

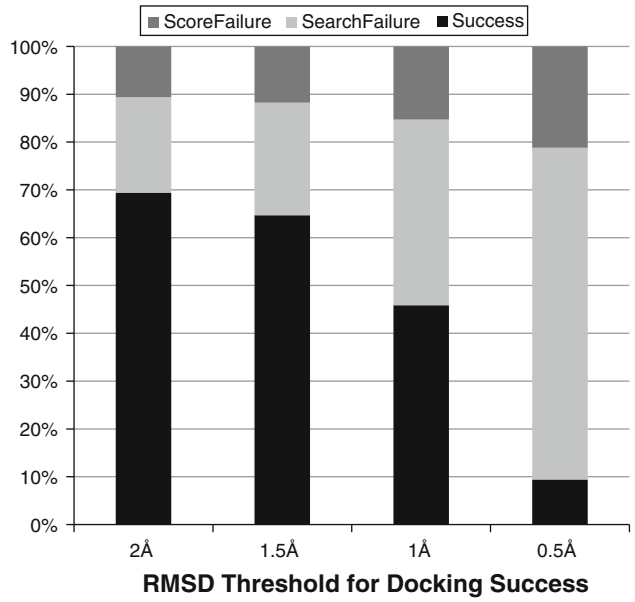

Fig. 3 Pose reproduction failure analysis for docking success thresholds of 2, 1.5, 1.0 and 0.5 Å RMSD. The *black portion of the bars* shows the docking success rates on the test dataset. *Light grey* represents search failures, where the top scoring docked pose is incorrect, but does not score as well as the crystallographic ligand pose. Score failures, where the top scoring docked pose is incorrect and scores better than the crystallographic ligand pose, are shown in *dark grey*

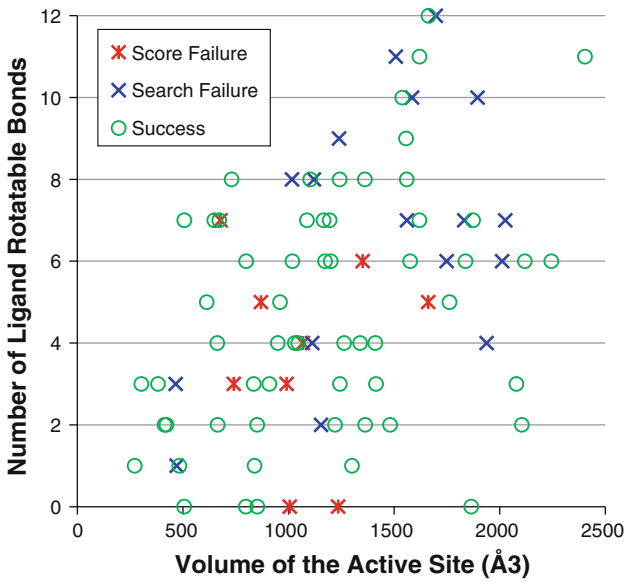

Fig. 4 Pose reproduction results as a function of the size of the active site and the number of rotatable bonds of the ligand. Each *point* is a co-crystal structure in the test set. Systems where FRED correctly docked the ligand (RMSD < 2 Å) are shown as circles. Systems where FRED failed to dock correctly are shown as either *crosses* (score failures) or *stars* (search failures)

statistically significant to 93 % (i.e., there is a 93 % chance that HYBRID will be better than FRED on average). For complete discussion on calculating statistical significance see paper by McGann [3]. The early enrichment results (shown in Fig. 5b–d) show the same relative performance of HYBRID and FRED. The error bars for the early enrichment results are larger than those for AUC, because there is more variance in performance from target to target using an early enrichment metric. The target variance, however, is highly correlated (correlation coefficients are 0.69, 0.82, 0.84 and 0.89 for AUC, fAUC (2 %), fAUC (1 %) and fAUC (0.1 %) respectively—see paper by McGann for correlation calculation details [3]) between FRED and HYBRID (i.e., they tend to do well on the same targets and poorly on the same targets). Thus the confidence in the difference in performance between FRED and HYBRID is still relatively high: 95, 99, 96 and 91 % for AUC, fAUC (2 %), fAUC (1 %) and fAUC (0.1 %), respectively.

The DUD-Wombat dataset is significantly more challenging for both FRED and HYBRID (see Fig. 5). The mean AUC for both is 0.7 and the early enrichment results are similarly degraded. On the subset of 10 targets in DUD for which there are Wombat actives, the average AUC is 0.76 for FRED and 0.77 for HYBRID using the standard DUD actives. Thus the reduced performance on DUD-Wombat relative to DUD is a result of the difficulty of the Wombat actives, not because the particular subset of Wombat targets is difficult.

The difference in performance between FRED and HYBRID on DUD-Wombat is not statistically significant for any of the four metrics, in contrast to the standard DUD results. The reason for this is that the standard DUD actives tend to be analogues of the crystallographic ligand while the DUD-Wombat actives are generally not. Evidence for this can be seen in the MACCS results (which are based on graph similarity to the crystallographic ligand) in Fig. 5 that are good for standard DUD (AUC 0.72) and very poor for DUD-Wombat (AUC 0.48). Since HYBRID's underlying assumption is that ligands will dock in a similar binding mode to the crystallographic ligand, it makes sense that HYBRID will perform better on test systems where the docking ligands are close analogues of the crystallographic ligand, such as standard DUD. HYBRID's performance on the DUD-Wombat dataset shows that even in cases where the active ligands are not analogous to the crystallographic ligand, the performance is not worse than using standard docking with FRED. Thus this data shows that HYBRID is

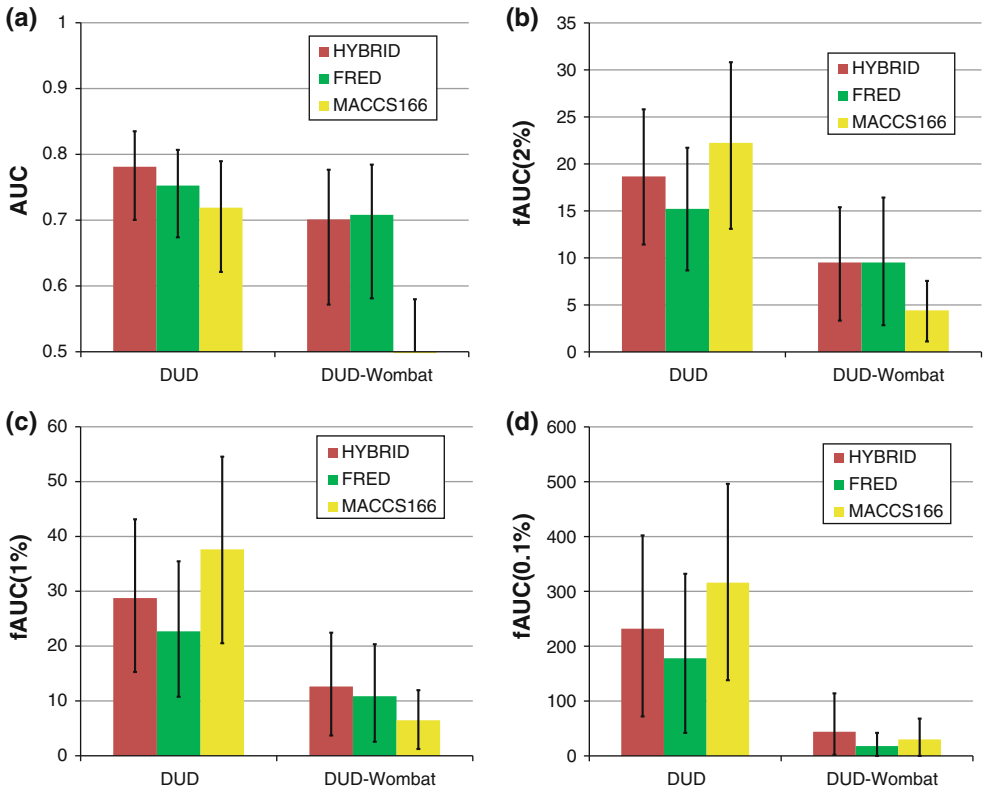

Fig. 5 Mean virtual screening results on the DUD and DUD-Wombat dataset using **a** AUC, **b** fAUC (2 %), **c** fAUC (1 %) and **d** fAUC (0.1 %). AUC is the area under the receiver operator characteristic curve. fAUC is the area under a fraction of receiver operator characteristic curve from 0 % decoys recovered to the given percent decoys recovered (e.g., 2, 1 or 0.1 %) normalized. *Error bars* are the 95 % confidence intervals in the mean results. *Bar's* in each group from left to right are HYBRID, FRED and MACS166

equivalent or superior to FRED for all types of active molecules.

Virtual screening—multiple crystal structures

The results for FRED and HYBRID described above use the rigid protein docking approximation, as do the majority of docking programs. Specifically, this means that the protein structure does not adapt its conformation as each ligand is docked, but rather remains in its crystallographic conformation no matter what ligand is docked to it. The rigid protein approximation is reasonable if the protein itself happens to be very inflexible but is suspect otherwise. In reality, flexible proteins will generally adopt different conformations to bind different ligands. If, however, the protein structure being used for docking has a ligand that is similar to the molecule being docked (when similarity is defined in a relevant manner), the rigid docking approximation may still be valid since similar ligands will tend to have similar binding modes and hence similar protein conformations. This line of reasoning is used by HYBRID to approximate protein flexibility during the docking process.

HYBRID can implicitly account for protein flexibility by using multiple crystallographic structures of the target protein. Results for the MDUD dataset, which contains multiple crystal structures for each target, are shown in Fig. 6 for FRED and HYBRID. Since FRED can only utilize one protein structure at a time, a separate FRED run was performed for each. The FRED results for all the protein structures of a given target were then averaged to get a target result. The results for all targets were then averaged to obtain the mean metric values shown in Fig. 6. Conceptually, the FRED results shown in Fig. 6 are the expected performance if a protein structure is chosen at random. HYBRID can accept either one or multiple protein structures. The results labeled HYBRID-S in Fig. 6 are from HYBRID using only one crystal structure and are calculated the same way as the FRED results. The results referred to as HYBRID-M are for the HYBRID program when it is given multiple crystal structures and allowed to select the best protein structure for each docking ligand.

HYBRID-M outperforms all other tested methods, with a mean AUC of 0.80 on the MDUD dataset. By comparison, the mean AUC on MDUD for HYBRID-S is 0.75 and for FRED is 0.71. Clearly, HYBRID-M is effectively using the additional information present in multiple crystal

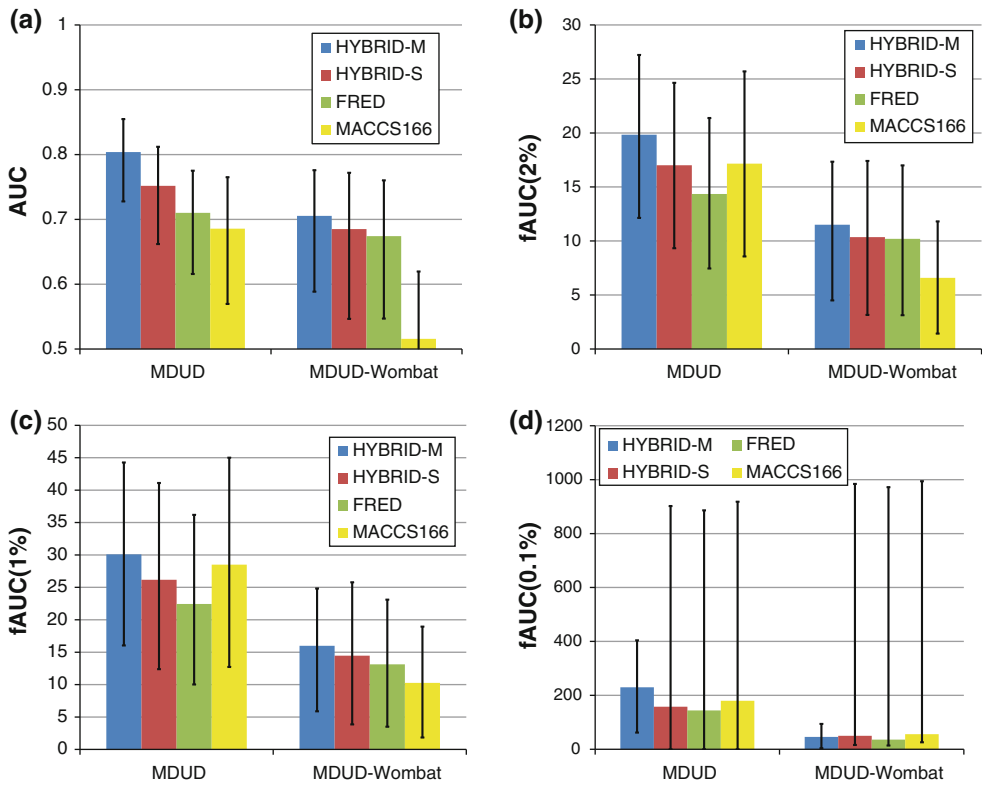

Fig. 6 Mean virtual screening results on the MDUD and MDUD-Wombat dataset using **a** AUC, **b** fAUC (2 %), **c** fAUC (1 %) and **d** fAUC (0.1 %). *Error bars* are the 95 % confidence intervals in the mean results. The FRED and HYBRID-S results used a single crystal structure in each run and the results for each target were averaged across all crystal structures of the target. The HYBRID-M results use all the crystal structures of a target for each run. *Bar's* in each group from left to right are HYBRID-M HYBRID-S, FRED and MACCS166

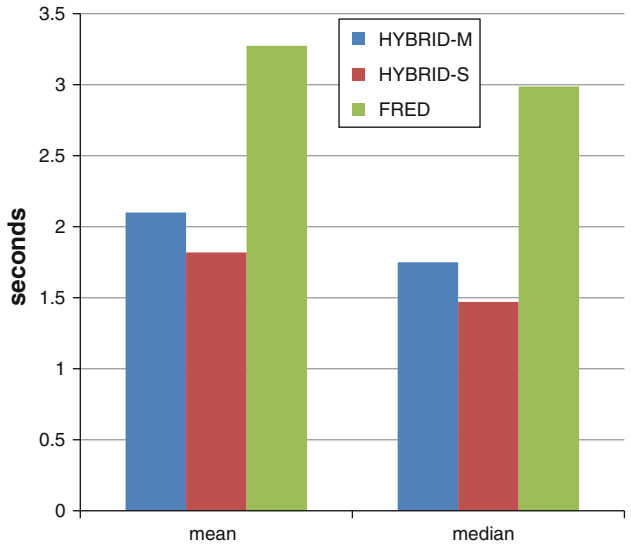

Fig. 7 Average docking time for the MDUD dataset using a single core of a 2.4 GHz Intel Xeon. *Bar's* in each group from left to right are HYBRID-M HYBRID-S and FRED

structures to enhance docking performance. MACCS results, using the crystallographic ligands from each protein target as queries, are competitive for such a simple method but, like the standard DUD results, have poor AUC (0.65) relative to docking methods. These differences are all statistically significant to greater than 95 % confidence. The results on the MDUD-Wombat dataset are poorer overall relative to the MDUD dataset, exactly as those on the DUD-Wombat dataset are poorer than the standard DUD results.

HYBRID with multiple crystal structures not only provides the best virtual screening performance, it also has a negligible impact on docking speed (see Fig. 7). The mean docking time for HYBRID-S is 1.8 s/ligand compared to 2.1 s/ligand for HYBRID-M. The increase in docking time using multiple crystals is minimal because each ligand is still only docked once. The only additional overhead for HYBRID using multiple crystal structures is comparing the ligand being docked to each crystallographic ligand to determine the most appropriate receptor to dock to, and this process is quite rapid compared to the docking. FRED is slower than HYBRID in either mode (3.3 s/ligand) because it uses a structure-based scoring function during the exhaustive search that is more complex (and hence computationally expensive) than the ligand-based scoring function HYBRID uses during the exhaustive search.

Conclusions

FRED has a reasonable docking success rate of 70 % at a 2 Å RMSD threshold. Approximately two-thirds of the docking failures are attributable to the docking algorithm not examining the correct pose, while the remaining third are due to the scoring function not recognizing the correct pose. The success rate drops modestly for 1.5 and 1.0 Å RMSD thresholds (65 and 50 % respectively), and sharply for 0.5 Å RMSD thresholds (9 %) because 0.5 Å is below the resolution of FRED's docking algorithm.

HYBRID outperforms FRED for virtual screening on the standard DUD dataset with a mean AUC of 0.78 compared to FRED's 0.75. The DUD-Wombat dataset proved more challenging for both FRED and HYBRID with a mean AUC of 0.7 for both. The ligands in the DUD-Wombat dataset are quite dissimilar to the ligand each protein structure was crystallized with, and thus one would expect a partially ligand-based method to perform relatively poorly. However, HYBRID still manages to perform equivalently to FRED on the DUD-Wombat dataset while outperforming FRED on the standard DUD dataset. We therefore conclude that HYBRID should be preferred over FRED for any system where the protein is crystallized in the presence of a ligand.

HYBRID with multiple crystal structures for each target improves virtual screening relative to docking with a single crystal structure for each target using either FRED or HYBRID. The mean AUC on the MDUD-Dataset for HYBRID using multiple crystal structures is 0.8, compared to 0.75 and 0.71 using FRED and HYBRID with one crystal, respectively. Using HYBRID with multiple, rather than a single, crystal structures increases docking time by only $\sim$15 %. Thus, using multiple crystal structures with HYBRID is always recommended when multiple protein–ligand complexes have been crystallized.

References

1. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C (2009) Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. J Chem Inf Model 49:1455–1474
2. Warren GL, Andrews CW, Capelli A, Clarke B, LaLonde J, Lambert ML, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Pieshoff CE, Head MS (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49:5912–5913
3. McGann MR (2011) FRED pose prediction and virtual screening accuracy. J Chem Inf Model 51:578–596
4. Cummings MD, DesJarlais RL, Gibbs AC, Mohan V, Jaegar EP (2005) Comparison of automated docking programs as virtual screening tools. J Med Chem 48:962–976
5. Nicholls A (2008) What do we know and when do we know it? J Comput Aided Mol Des 22:133–139
6. Tuccinardi T, Botta M, Giordano A, Martinelli A (2010) Protein kinases: docking and homology modeling reliability. J Chem Inf Model 50:1432–1441
7. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010) Conformer generation with OMEGA: algorithm and

validation using high quality structures from the protein databank and Cambridge structural database. J Chem Inf Model 50: 572–584
8. Perola E, Charifson PS (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. J Med Chem 47:2499–2510
9. Brown RD, Martin YC (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. J Chem Inf Model 37:1–9
10. http://www.eyesopen.com/graphsim-tk. Accessed 30 April 2012
11. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, Murray CW (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem 50:726–741
12. Jain A (2008) Bias, reporting and sharing: computational evaluations of docking methods. J Comput Aided Mol Des 22:201–212
13. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49:6789–6801
14. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI (2004) Chemo informatics in drug discovery. In: Oprea TI (ed) WOMBAT: world of molecular bioactivity. Wiley-VCH, New York