Protein−Ligand Docking against Non-Native Protein Conformers

Marcel L. Verdonk,* Paul N. Mortenson, Richard J. Hall, Michael J. Hartshorn,§ and Christopher W. Murray

Astex Therapeutics Ltd., 436 Cambridge Science Park, Milton Road, Cambridge CB4 0QA, United Kingdom

Received July 7, 2008

In the validation of protein−ligand docking protocols, performance is mostly measured against native protein conformers, i.e. each ligand is docked into the protein conformation from the structure that contained that ligand. In real-life applications, however, ligands are docked against non-native conformations of the protein, i.e. the apo structure or a structure of a different protein−ligand complex. Here, we have constructed an extensive test set for assessing docking performance against non-native protein conformations. This new test set is built on the Astex Diverse Set (which we recently constructed for assessing native docking performance) and contains 1112 non-native structures for 65 drug targets. Using the protein−ligand docking program GOLD, the Astex Diverse Set and the new Astex Non-native Set, we established that, whereas docking performance (top-ranked solution within 2 Å rmsd of the experimental binding mode) is ∼80% for native docking, this drops to 61% for non-native docking. A similar drop-off is observed for sampling performance (any solution within 2 Å): 91% for native docking vs 72% for non-native docking. No significant differences were observed between docking performance against apo and nonapo structures. We found that, whereas small variations in protein conformation are generally tolerated by our rigid docking protocol, larger protein movements result in a catastrophic drop-off in performance. Some docking performance and nearly all sampling performance can be recovered by considering dockings produced against a small number of non-native structures simultaneously. Docking against non-native structures of complexes containing ligands that are similar to the docked ligand also significantly improves both docking performance and sampling performance.

INTRODUCTION

Advances in technologies like X-ray crystallography and NMR have meant that 3-dimensional (3D) structures are available for an ever-increasing number of drug-targets. Protein−ligand docking takes a central role in most modern structure-based drug discovery projects, where it can be used for virtual screening and interactive design. A range of protein−ligand docking programs has been reported in the literature, and an increasing number of these tools are available to the community, e.g. DOCK,[1] GOLD,[2,3] GLIDE,[4,5] and ICM.[6]

The performance of a docking program is normally measured by its ability to reproduce the ligand binding modes for a test set of protein−ligand complexes from the Protein Data Bank[7] (PDB). A number of such test sets have been constructed over the years, but most suffer from issues around data quality, diversity, and relevance to drug discovery. Recently, we constructed the Astex Diverse Set: a new test set containing 85 protein−ligand complexes that addresses the issues with the existing sets.[8] In brief, this set (i) only includes structures of targets that are relevant to drug discovery or agrochemistry and contain druglike ligands; (ii) contains a diverse set of protein−ligand complexes with no target represented more than once; and (iii) contains only high-quality structures for which structure factors have been deposited in the PDB and for which the electron density accounts for all parts of the experimental binding mode of the ligand.

Although other measures for docking success have been proposed,[9,10] the most commonly used performance indicator for a docking program is the percentage of ligands for which the top-ranked solution produced by the docking program is within 2.0 Å Root-Mean-Square Distance (rmsd) of the experimental binding mode; we will refer to this as the docking performance. The docking performance of good quality docking protocols tested against large test sets of protein−ligand complexes, as reported in the literature is approximately 70−80%.[5,8] From such success rates, one might be inclined to conclude that there is limited scope for further improving docking protocols. However, the success rates quoted are for docking each ligand against its native protein conformation, i.e. the 3D coordinates of the protein are taken from the same structure that contained the ligand. Hence, any protein conformational changes induced by the ligand are reflected in the protein structure that the ligand is docked against (we will refer to this as "native docking"). In real-life applications where the binding modes of newly designed compounds are predicted, the protein structure used to dock against will be that of a complex containing another ligand or that of the apo form of the protein (we will refer to this as "non-native docking"). In reality therefore, when applied to actual drug discovery projects, the success rates of docking programs will be lower than those quoted above. In addition, there is a real danger that docking protocols will

* Corresponding author e-mail: m.verdonk@astex-therapeutics.com.
§ Present address: Dotmatics Ltd., 1-2 Thorley Hall Stables, Bishop's Stortford CM23 4BE, United Kingdom.

10.1021/ci8002254 CCC: $40.75
Published on Web 10/28/2008

be overtrained on validation sets containing only native protein conformations.

Several authors have studied the performance of docking programs against non-native conformations of the targets. However, the numbers of targets investigated in these studies has been limited. Murray and co-workers reported on the performance of the docking program PRO_LEADS[11] against a set of 21 protein−ligand structures for thrombin (6), thermolysin (9), and neuraminidase (6).[12] The authors found that, although success-rate for native docking was 76%, this dropped to 49% for non-native docking. Birch and colleagues also investigated neuraminidase and found that the sensitivity of the docking performance to small induced fits was strongly dependent on the scoring function used.[13] Whereas Goldscore performed well for non-native docking on a test set of 15 'clean' neuraminidase complexes (89% success), Chemscore performed poorly (49% success). A test set of 73 complexes of trypsin (10), thrombin (26), and HIV-1 protease (37) was used by Erickson et al.[14] to assess the effect of ligand-induced fit on the docking performance of CDOCKER.[15] Native docking performance was 67% for trypsin, 36% for thrombin, and 50% for HIV-1 protease, but a significant drop-off in performance was observed when docking was done against the "average" conformation (60%, 27%, and 35%, respectively) or against the apo conformation of the protein (37%, 9%, and 4%, respectively). Recently, the same group reported a study on a wider range of 8 drug targets,[16] in which they showed a drop-off in docking performance from 39% for native dockings to 26% for non-native dockings (results are averages over the 8 targets for CDOCKER). Cavasotto and Abagyan reported on the performance of their docking program ICM[6] against a set of four protein kinases.[17] These authors also observed a significant drop-off in success rates for non-native dockings, when compared to native dockings against these targets. The native docking performance reported is remarkably high: 100% of ligands are docked within 2 Å of the experimental binding mode. Unfortunately, the fraction of non-native dockings for which the ligand is predicted within 2 Å can only be derived from this paper for two targets: 50% for CDK2 and 56% for cAPK. Corbeil et al. tested their docking protocol FITTED against a test set of 33 complexes for five drug targets and observed a drop-off from 67% for native dockings to 49% for non-native dockings.[18] There are a host of other studies in the literature reporting on test sets for non-native docking,[19−22] particularly for the evaluation of the performance of docking protocols taking into account protein flexibility. The most diverse of these test sets was reported recently by Bottegoni et al.,[23] who constructed a test set of 32 complexes for 16 drug targets to test the performance of their flexible protein docking protocol SCARE.

Particularly now an increasing number of docking programs attempt to account for protein flexibility (e.g., FlexE,[24] GLIDE,[21] SCARE/ICM,[23] etc.), there is a real need for a substantial, high-quality, relevant, and diverse test set for non-native docking performance. Here, we report on the construction of a large test set for assessing docking performance against non-native protein conformations. This new set, the Astex Non-native Set, is built on the Astex Diverse Set and contains 1112 structures for 65 drug targets. The Astex Non-native Set has undergone quality checks similar to those used in the construction of the original native-docking Astex Diverse Set and will be made publicly available via the Cambridge Crystallographic Data Centre Web site (http://www.ccdc.cam.ac.uk). Using this new test set, we investigate the effects of ligand-induced fit on rigid-protein docking performance using the protein−ligand docking program GOLD.[2,3] In addition, we investigate the potential of docking protocols taking into account multiple protein conformations simultaneously.

Table 1. Filters Applied to Construct the Astex Non-Native Set[a]

	non-native structures	targets
total number of relevant active sites	2842	83
resolution ≤2.5 Å	2340	81
protein sequence identical for binding site	1541	75
one site per PDB entry (lowest B-factor)	1287	75
filters applied during setup of structures	1142	66
visual inspection	1112	65

[a] The numbers of targets and non-native structures available at each stage are shown.

METHODOLOGY

The Astex Diverse Set was used as a basis for the construction of the Astex Non-native Set. For each entry in the Astex Diverse Set, we set out to find all entries in the PDB where the target is the same, but the structure is either an apo structure or a complex with a different ligand. The non-native docking validation then consisted of docking each ligand from the Astex Diverse Set against all its non-native protein conformers. Hence, the Astex Non-native Set only contains non-native structures of the protein targets in the Astex Diverse Set and does not contain any new ligands. We believe this strategy is useful for the following reasons: (i) The approach allows a direct comparison of non-native docking against the Astex Non-native Set to native docking against the Astex Diverse Set as both sets consist of exactly the same targets and ligands. (ii) It allows us to separate the problem of ligand-induced protein *conformation* from other ligand-dependent factors like protonation states and tautomers in protein residues; although these are clearly important issues, we feel the purpose of a non-native docking set is rather to assess the influence of protein movement on docking performance. (iii) The newly added ligands and their structures would need to be of the same quality, diversity, and relevance as the structures in the Astex Diverse Set itself; only a limited number of targets have additional liganded structures that conform to these criteria.

Selecting the Structures. In order to construct the Astex Diverse Set, we had clustered all protein chains in the PDB into separate targets, based on their sequence homology.[8] We then selected, where possible, one representative from each PDB cluster, where the target is relevant to drug discovery or agrochemistry. This resulted in a test set of 85 complexes for 85 different targets. We reused this clustered version of the PDB (after updating it with structures deposited up to August 2007) in order to construct the Astex Non-native Set: for each entry in the Astex Diverse Set in turn, we retrieved all other structures from the PDB cluster that contained that entry. Only structures determined at resolutions ≤2.5 Å were kept. Table 1 lists the filters applied to the Astex Non-native Set and the number of remaining structures and targets at each stage.

Superimposing the Structures. For each target, every non-native structure was superimposed onto the corresponding structure in the Astex Diverse Set (we will refer to this as the reference structure). This is an important and nontrivial task. When the ligands are docked against non-native conformations of the target, it is critical that the non-native structures have been carefully superimposed onto the reference structure in order to get a reliable measure of the rmsd of the docked ligand against the experimental binding mode of the ligand in the corresponding entry in the Astex Diverse Set (the "reference ligand"). For each superimposition of a non-native structure against its reference structure, our approach was as follows. First each chain in the non-native structure was sequence-aligned against each contact chain in the reference structure (a contact chain being a chain that contains atoms within 4 Å of a non-hydrogen atom in the ligand); chain pairs "match" when they have >75% sequence identity. For each permutation of matching chain pairs, superimposition was attempted. The non-native structure was discarded if no permutation could be found for which all protein residues in the reference structure within 6 Å of the reference ligand had a matching residue in the non-native structure; i.e. all binding site residues need to be present and the same as in the reference structure. A large number of mutated structures were removed by this process. The superimposition of a particular chain permutation started off with a list of residue pairs, consisting of all residues in the reference structure within 6 Å of the ligand, and their corresponding residues in the non-native structure. The actual superimposition was done by the proprietary tool KFIT, using an iterative superimposition algorithm.[25] In the first step of this algorithm, all protein backbone and Cβ atoms for the residues in the superimposition list were superimposed. In the second step, all atom pairs which, after superimposition, were >1.0 Å apart were removed from this list, and the superimposition was repeated with the new list of atom pairs. In subsequent steps, atom pairs were either removed if they were >1.0 Å apart or added from the original list of atom pairs (binding site residues only) if they had been deleted in a previous step, but were now within 1.0 Å. This process was repeated until the list of superimposed atom pairs no longer changed. This procedure ensured that the final superimposition was done on the parts of the binding site that are structurally most conserved. Chain permutations for which the final list of superimposed atom pairs corresponded to fewer than 6 residues were discarded (this only affected 9 structures). Of the remaining chain permutations, the superimposition with the lowest B-factor (averaged over the binding site residues) was selected to be included in the Astex Non-native Set. Only one non-native conformation was kept for each PDB entry.

Identifying Ligands. Ligands were extracted from the PDB files and atom typed using a proprietary protocol as previously described.[8] Only ligands that, after superimposition, were in close proximity to the reference ligand were considered. Next, all ligands were classified as *solvent*, *cofactor*, or *ligand*. The *solvent* class contains small ions, solvents, etc., which were assigned from a list of 443 HET group names (see the Supporting Information). *Cofactors* were assigned from target-specific HET-group lists (see the Supporting Information), depending on whether the site occupied by the reference ligand coincides with the substrate/cofactor site or not (e.g., ADP was regarded a cofactor for the pantothenate kinase entry in the Astex Diverse Set but not for the p38 entry). All remaining ligands were put in the *ligand* class. Binding sites containing no ligands in the *ligand* class were flagged as apo structures.

Setting up the Structures. The previous steps resulted in 1287 non-native structures, each superimposed onto its reference structure from the Astex Diverse Set. Each of these non-native structures now needed to be set up for docking the reference ligand from the corresponding entry in the Astex Diverse Set. A lot of thought and effort has gone into the preparation of the target structures in the Astex Diverse Set in terms of protonation states, tautomers, and side chain flips for histidine, asparagine, and glutamine side chains. Hence, the strategy we adopted here was to—where appropriate—prepare the binding sites of the non-native structures in exactly the same way as we prepared those of the corresponding reference structures. The automated protocol we developed for this process started off by defining the binding site in both the reference and non-native structures as all residues that were within 6 Å of any heavy atom of the reference ligand. The term "residue" here includes cofactors, ions, and any other moieties that were retained in the reference structure in the Astex Diverse Set. Where the list of binding site residues included a metal ion, any additional residues within 3 Å of the ion were also included, and analogously residues within 3.5 Å of a retained water molecule were added.

For each residue in the reference binding site, we attempted to find the matching residue in the non-native structure (not necessarily within the non-native binding site). If no matching residue could be found, or if the residue in the reference structure contained atoms that were absent or different in the non-native structure, then that non-native structure was discarded. For each protein residue in the non-native binding site, the same procedure was applied, i.e. the structure was removed if no appropriate match could be found in the reference structure. For nonprotein residues in the non-native structure (e.g., cofactors, metal ions, waters, etc.), if no matching residue was found in the reference structure, then the unmatched residue was removed, but this was not considered sufficient reason to discard the structure as a whole.

In cases where protein side chains were disordered in the non-native structure, the following approach was taken to select one of the conformers: (i) where possible, the conformer with the highest occupancy was selected; (ii) for side chains with equal occupancy, in cases where one of the conformers clashed with the ligand, the other conformer was selected; and (iii) if no decision could be made based on rules (i) and (ii), then the conformer listed first in the PDB file was selected. A total of 3123 residues in 322 structures across 44 targets showed multiple conformations. Of these, in 1530 cases the conformation was chosen based on occupancy, i.e. rule (i), and in the remaining 1593 cases it was chosen arbitrarily, i.e. by applying rule (iii). Interestingly, although part of the algorithm dealing with selecting the conformers, rule (ii) was never actually used in practice.

Next, the protocol dealt with side chain flips for asparagine, glutamine, and histidine residues. If, for the relevant matching atom pairs in these residues (terminal carboxamide group for asparagine and glutamine ring atoms for histidine),

the superimposition onto the reference structure was good (rmsd <0.7 Å), then the residue was left unchanged. If, however, superimposition of the side chain was poor, but good after flipping it 180°, then the side chain was flipped. Subsequently, protonation states and tautomers were copied from the native structure onto the non-native structure. For example, if a carboxylic acid was protonated on one of its oxygen atoms in the native structure, then the equivalent oxygen atom was protonated in the non-native structure. Which oxygen atom was "equivalent" was decided based on distance where the superimposition of the two carboxylate groups was good and by atom name otherwise. Bond-typing and protonation states for cofactors retained in the binding site were also copied from native structure to non-native structure. Finally, water molecules that were retained in a couple of the non-native structures to define the geometries of metal ions were oriented in the same way as those in the native structures. This process resulted in 1142 non-native structures that—within the binding site—contain exactly the same set of atoms as their corresponding reference structures in the Astex Diverse Set and that mirror the reference structure as closely as possible in terms of protonation states, tautomers, and side-chain flips.

Visual Inspection. At this stage, all remaining 1142 structures were visually inspected. One reason for doing this was to check, and if necessary correct, the automated apo/nonapo assignments for all structures (see above). Furthermore, a number of structures were discarded at this stage for several, more subjective reasons. For example, three Cdk2 structures were removed because we deemed the disorder models that were used in these structures physically and/or theoretically unrealistic. Also, 18 dihydrofolate reductase structures were removed because they contained an additional cofactor, not present in the reference structure, which we believed was responsible for some conformational changes observed in the protein (i.e., the induced changes were not ligand induced).

Docking Protocols. All docking runs were carried out using an in-house version of GOLD. We used the "default 1" search settings for the Genetic Algoritm (GA) and the Goldscore function as a scoring function.[26] All docking runs were also performed using the Chemscore function, but these results are included in the Supporting Information. For native docking runs, each ligand was docked against its native structure in the Astex Diverse Set. For non-native runs, each ligand was docked against all non-native structures available for that target in the Astex Non-native Set. All docking runs were repeated five times, and the success rates reported are the averages over these five runs.

Measuring Performance. Docking performance for native docking runs was simply defined as the percentage of complexes (targets) for which the top-ranked predicted binding mode was within 2 Å of the experimental binding mode, i.e.

$$\text{native docking performance} = 100\% \cdot \frac{N_{correct}}{N} \qquad (1)$$

where N is the number of targets, and $N_{correct}$ is the number of targets for which the top-ranked docking solution is within 2 A rmsd of the reference ligand. However, if we used the same performance measure for non-native docking runs (i.e., the percentage of the 1112 structures for which the ligand was docked correctly), then that success rate would be highly biased toward targets for which large numbers of non-native conformations are available. Hence, we defined the success rate for non-native docking runs as

$$\text{non-native docking performance} = 100\% \cdot \frac{1}{N}\sum_{i=1}^{N} \frac{n_{correct}(i)}{n_{total}(i)} \qquad (2)$$

where $n_{correct}(i)$ is the number of non-native structures of target i for which the top-ranked binding mode of the reference ligand was within 2 Å of its experimental binding mode, and $n_{total}(i)$ is the total number of non-native structures for target i. In total, the Astex Non-native Set contains structures for 65 targets, i.e. $N = 65$. However, for some analyses we will use subsets of the structures (e.g., containing only apo structures) for which $N < 65$. Wherever comparisons are made between native and non-native docking, the performance is calculated over the exact same set of targets.

RESULTS AND DISCUSSION

Composition of the Set. Table 2 shows the composition of the Astex Non-native Set in terms of how many structures the set contains for each target. The set contains at least one non-native structure for 65 out of the 85 targets in the Astex Diverse Set, 39 targets have at least 5 non-native structures, and 27 targets have at least 10 non-native structures. The Astex Non-native Set contains 211 apo structures for 29 targets and 901 nonapo structures for 59 targets. Figure 1 shows the distribution of the numbers of apo and nonapo structures for all of the 65 targets.

Table 2 gives an indication of how flexible the different targets in the set are. For example, nNos and alpha mannosidase II are extremely rigid, with average site RMSDs of 0.4 Å and 0.3 Å, respectively, and, as a result, show hardly any bad clashes with the reference ligand. On the other hand, p38 shows much more active site flexibility, with an average site rmsd of 2.6 Å, resulting in a significant number of bad clashes with the reference ligand. Interestingly, heat shock protein 90 (HSP90) has a similar average site rmsd to p38, but in this case no bad clashes are observed. This is because in the HSP90 entry in the Astex Diverse Set (2bsm) an α-helix containing Gly108 has partially collapsed into the binding site and contacts the reference ligand. This conformation is observed in a number of HSP90 structures, but in other structures the uncollapsed helix form is observed and Gly108 is situated well away from the reference ligand, resulting in a high site rmsd, but no bad clashes.

Figure 2 shows distributions of the site rmsd, the maximum distance any site atom moves, and the number of bad clashes with the reference ligand. It is clear that the Astex Non-native Set contains a variety of protein movements, ranging from hardly any changes in protein conformation, to minor ligand induced fit, to major protein movements. When downloaded from the CCDC Web site, the Astex Non-native Set will be supplemented with a spreadsheet containing information such as site rmsd, maximum atom movement, maximum residue movement, etc., so that a user can select the structures that are most appropriate for the application that is being tested.

Superimposition and Protein Movement. Figure 3 illustrates the importance of the careful superimposition of

Table 2. Overview of the 65 Targets in the Astex Non-native Set, an Indication of Their Flexibility, and the Native and Non-native Docking Performance[a]

							target-specific docking performance	
target	entry[b]	all	apo	nonapo	<site rmsd>[c]	<bad clashes>[d]	native[e]	non-native[f]
HIV-1 protease	1kzk	134	0	134	0.9 (0.2)	0.6 (0.8)	100 (0)	64 (2)
thrombin	1oyt	123	12	111	0.7 (0.5)	0.5 (2.1)	100 (0)	94 (1)
carbonic anhydrase II	1oq5	116	48	68	0.7 (0.1)	1.1 (0.5)	96 (5)	1 (1)
cyclin-dependent kinase 2	1ke5	80	9	71	1.3 (0.5)	2.6 (2.8)	100 (0)	58 (1)
deoxy hemoglobin	1g9v	69	66	3	1.5 (0.9)	3.3 (3.7)	0 (0)	0 (0)
cytochrome P450cam	1p2y	40	2	38	0.6 (0.3)	0.6 (0.6)	58 (18)	76 (2)
p38 kinase	1ywr	36	6	30	2.6 (1.0)	5.4 (4.4)	92 (8)	6 (1)
factor Xa	1lpz	33	0	33	1.0 (0.2)	2.4 (2.1)	100 (0)	82 (4)
aldose reductase	1t40	33	2	31	1.0 (0.5)	5.4 (5.3)	100 (0)	36 (3)
heat shock protein 90	2bsm	32	3	29	2.6 (1.0)	0.0 (0.0)	100 (0)	99 (2)
acetylcholinesterase	1gpk	28	8	20	1.0 (0.4)	0.2 (0.4)	34 (15)	16 (6)
urokinase	1owe	26	0	26	0.6 (0.3)	0.2 (1.2)	8 (13)	9 (3)
beta-lactamase	1l2s	25	2	23	0.7 (0.5)	0.2 (0.6)	8 (8)	0 (0)
neuraminidase	1l7f	25	13	12	0.6 (0.2)	0.6 (0.5)	100 (0)	100 (0)
dipeptidyl peptidase IV	1n1m	19	5	14	0.5 (0.1)	0.2 (0.4)	100 (0)	100 (0)
nNOS	1mmv	19	1	18	0.4 (0.2)	0.1 (0.2)	100 (0)	95 (0)
alpha-mannosidase II	1hww	19	0	19	0.3 (0.0)	0.3 (0.5)	100 (0)	89 (4)
factor VIIa	1ygc	17	0	17	0.9 (0.7)	1.8 (4.3)	2 (4)	26 (9)
TGT	1n2v	16	2	14	1.3 (0.6)	0.9 (2.0)	0 (0)	16 (3)
myosin II	1yv3	16	16	0	2.3 (0.2)	14.0 (0.6)	100 (0)	0 (0)
Chk1	2br1	16	1	15	0.7 (0.2)	0.8 (1.1)	86 (17)	46 (9)
phosphodiesterase 4D	1xoq	13	0	13	0.7 (0.3)	3.5 (1.9)	92 (8)	6 (6)
tryptophan synthase	1k3u	12	0	12	0.5 (0.2)	0.2 (0.6)	100 (0)	92 (6)
adenosine deaminase	1uml	11	1	10	0.9 (0.4)	1.5 (4.8)	4 (5)	38 (4)
neuraminidase	1vcj	11	1	10	1.2 (0.1)	0.8 (0.4)	100 (0)	100 (0)
Chitinase B	1w1p	10	2	8	0.7 (0.4)	0.3 (0.5)	100 (0)	68 (4)
VDR	1s19	10	0	10	0.2 (0.0)	0.0 (0.0)	100 (0)	100 (0)
dihydrofolate reductase	1s3v	9	1	8	1.0 (0.3)	0.1 (0.3)	100 (0)	91 (5)
thymidylate kinase	1w2g	8	0	8	0.3 (0.1)	0.0 (0.0)	100 (0)	100 (0)
FTase	1mzc	7	0	7	0.4 (0.0)	0.1 (0.4)	100 (0)	49 (22)
GSK-3beta	1q41	7	0	7	0.8 (0.2)	1.0 (1.3)	100 (0)	77 (8)
pantothenate synthetase	1n2j	7	0	7	0.2 (0.0)	0.0 (0.0)	36 (18)	14 (0)
PNMT	1hnn	6	0	6	0.7 (0.5)	0.0 (0.0)	100 (0)	100 (0)
c-Abl tyrosine kinase	1opk	6	0	6	1.3 (0.9)	0.2 (0.4)	100 (0)	100 (0)
uridine phosphorylase	1u1c	5	0	5	1.3 (1.2)	1.6 (2.3)	100 (0)	60 (0)
dihydrofolate reductase	1ia1	5	1	4	0.5 (0.3)	0.0 (0.0)	100 (0)	100 (0)
MetAP2	1r58	5	0	5	1.0 (0.3)	1.8 (1.6)	70 (16)	32 (11)
quinone reductase 2	1sg0	5	1	4	0.3 (0.1)	0.0 (0.0)	100 (0)	100 (0)
deoxycytidine kinase	1p62	5	0	5	0.5 (0.1)	0.0 (0.0)	100 (0)	100 (0)
penicillin G acylase	1gm8	4	1	3	0.6 (0.3)	0.5 (0.6)	0 (0)	0 (0)
phosphodiesterase 5A	1xoz	4	0	4	0.8 (0.2)	0.0 (0.0)	100 (0)	35 (14)
c-Jun terminal kinase 3	1pmn	4	0	4	1.1 (0.5)	4.0 (2.8)	30 (7)	50 (0)
thymidine kinase	1of1	3	0	3	0.8 (0.2)	0.7 (1.2)	100 (0)	80 (18)
progesterone receptor	1sqn	3	0	3	0.5 (0.1)	0.0 (0.0)	100 (0)	100 (0)
protein kinase 5	1v0p	3	2	1	1.6 (0.1)	5.0 (1.0)	96 (5)	0 (0)
neprilysin	1r1h	3	0	3	1.0 (0.3)	2.7 (2.1)	100 (0)	0 (0)
estrogen receptor alpha	1sj0	2	0	2	2.0 (0.1)	0.0 (0.0)	100 (0)	100 (0)
metallo beta-lactamase	1jje	2	0	2	1.3 (1.2)	0.0 (0.0)	0 (0)	50 (0)
glutamate receptor 6	1tt1	2	0	2	0.9 (0.2)	4.5 (0.7)	100 (0)	50 (0)
TR beta-1	1n46	2	0	2	1.2 (0.0)	0.0 (0.0)	100 (0)	100 (0)
A-FABP	1tow	2	0	2	0.4 (0.1)	0.0 (0.0)	0 (0)	0 (0)
HIV-1 RT	1jla	1	0	1	1.5 (0.0)	6.0 (0.0)	100 (0)	0 (0)
COX-1	1q4g	1	0	1	0.3 (0.0)	0.0 (0.0)	100 (0)	100 (0)
NS5B polymerase	1yvf	1	1	0	0.8 (0.0)	3.0 (0.0)	6 (9)	0 (0)
HPPK	1hq2	1	0	1	0.1 (0.0)	0.0 (0.0)	100 (0)	100 (0)
androgen receptor	1z95	1	0	1	0.5 (0.0)	0.0 (0.0)	100 (0)	100 (0)
IMPDH	1meh	1	1	0	1.6 (0.0)	0.0 (0.0)	100 (0)	100 (0)
PNP	1q1g	1	0	1	0.7 (0.0)	1.0 (0.0)	100 (0)	100 (0)
IAG-NH	1hp0	1	0	1	1.2 (0.0)	0.0 (0.0)	54 (11)	100 (0)
beta-II tryptase	2bm2	1	0	1	0.9 (0.0)	2.0 (0.0)	84 (11)	20 (45)
auxin-binding protein 1	1lrh	1	1	0	0.6 (0.0)	0.0 (0.0)	98 (4)	100 (0)
carbonic anhydrase XII	1jd0	1	1	0	0.5 (0.0)	1.0 (0.0)	96 (9)	40 (55)
thiamin pyrophosphokinase	1ig3	1	0	1	0.7 (0.0)	0.0 (0.0)	100 (0)	100 (0)
ADAM33	1r55	1	1	0	0.8 (0.0)	0.0 (0.0)	100 (0)	100 (0)
thymidine phosphorylase	1uou	1	0	1	0.7 (0.0)	1.0 (0.0)	100 (0)	100 (0)

[a] Standard deviations are given in parentheses. [b] The PDB entry of the reference structure in the Astex Diverse Set. [c] All atom rmsd over target atoms that are—either in the reference structure or in the non-native structure—within 4 Å of the reference ligand, averaged over all non-native structures for a given target. [d] Number of bad clashes between non-native structure and the reference ligand, averaged over all non-native structures for a given target. The minimum contact distance between a pair of heavy atoms was set to the sum of their van der Waals radii (assigned according to the rules of Tsai et al.[27]), or 2.6 Å for donor−acceptor pairs, or 1.5 Å for metal-acceptor pairs. Any contact closer than this minimum distance was defined as a clash, and any contact shorter than this distance minus 0.9 Å was defined as a bad clash. [e] Native docking performance for individual targets is defined as the percentage of docking runs (out of ten runs) for which the top-ranked GOLD solution has rmsd <2.0 Å. Results shown are averages over five runs. [f] Non-native docking performance for individual targets is defined as the percentage of non-native structures for a given target for which the top-ranked GOLD solution has rmsd <2.0 Å. Results shown are averages over five runs.

the non-native structures onto the reference structure for HSP90 (2bsm). In this target, significant movement occurs around Gly108 (see above). In the reference structure this part of the structure is in the collapsed-helix conformation.

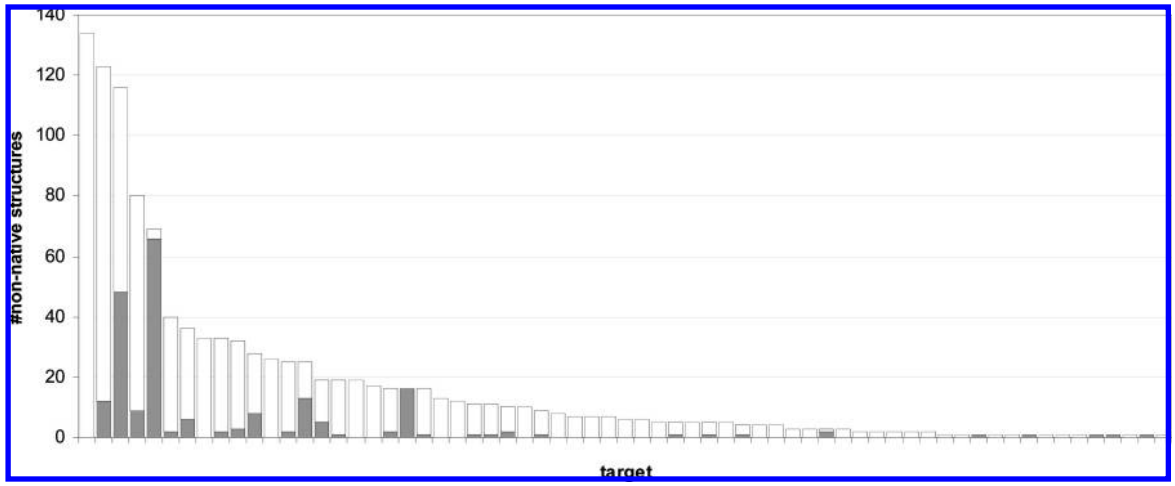

Figure 1. Distribution of the number of non-native structures per target. Numbers are also listed in Table 2. Apo-structures are represented as the gray parts of the bars, whereas nonapo structures are represented by the white sections of the bars.

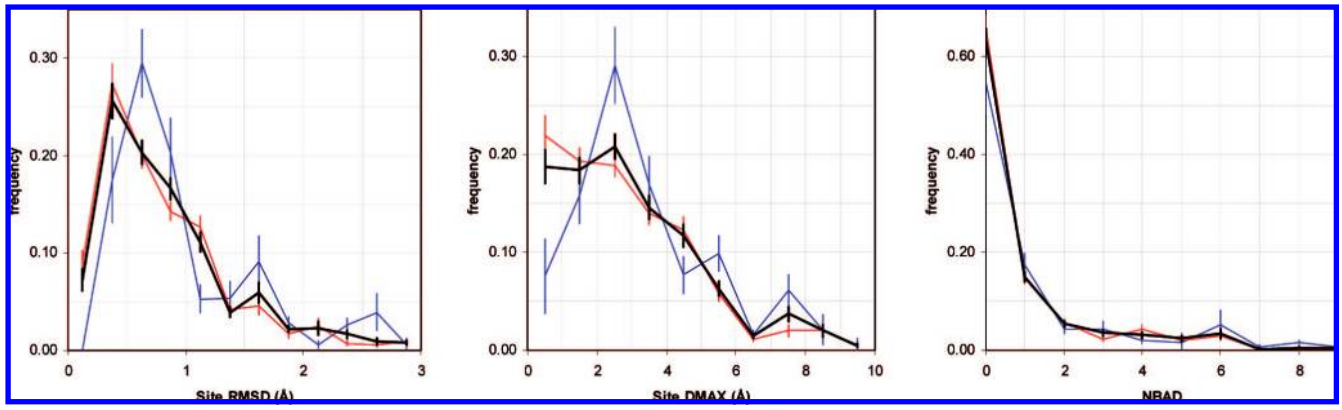

Figure 2. Distributions of site rmsd, maximum site atom movement (DMAX), and number of bad clashes with the reference ligand (NBAD) for all entries in the Astex Non-native Set. Results are shown for all structures (black), apo structures (blue), and nonapo structures (red). Site rmsd and DMAX are over all atoms that—either in reference or non-native structure—are within 4 Å of the reference ligand. Frequencies were calculated in such a manner that each target contributed equally to the distributions.

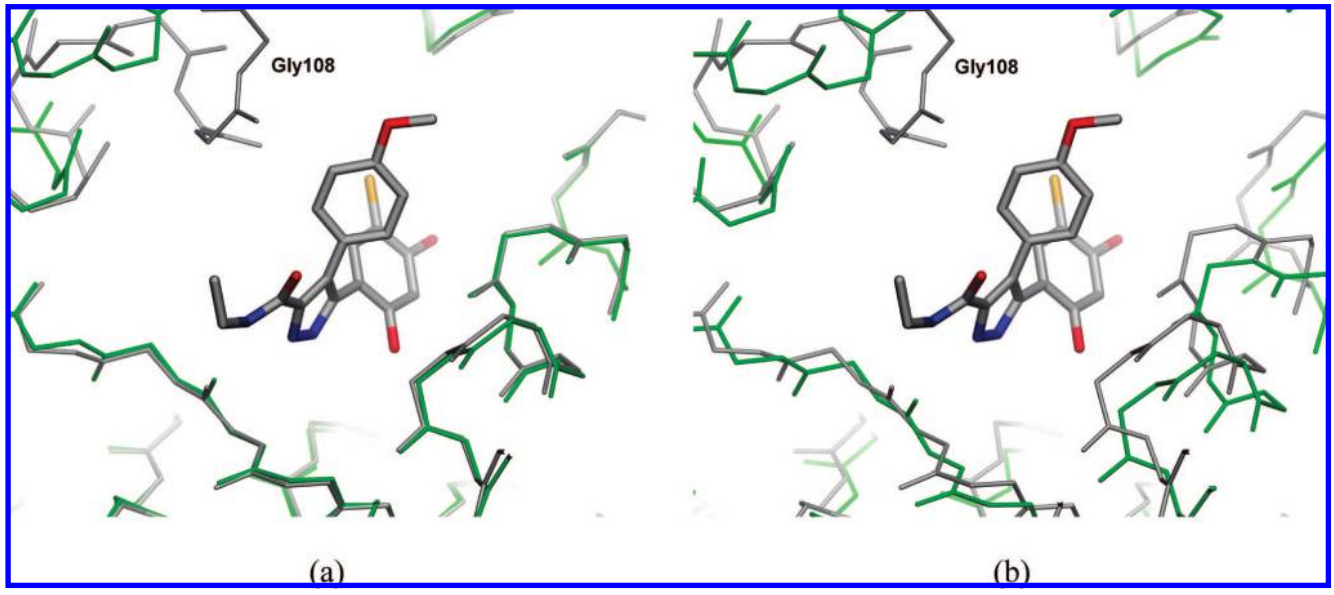

Figure 3. Superimposition of a non-native structure (1uye) for HSP90 onto its reference structure (2bsm), by applying the protocol used in the construction of the Astex Non-native Set (a) and by superimposing all main chain and Cβ atoms within 4 Å of the reference ligand (b). The reference structure is shown in gray; the non-native structure is shown in green; only main chain atoms are shown for the target; the reference ligand is shown in gray (carbon atoms). Pictures were created using AstexViewer 2.[28]

The example in Figure 3 shows the superimposition of a non-native structure for which the region around Gly108 is in the uncollapsed helix conformation (1uye) onto the reference structure. Because our superimposition algorithm focuses in on the structurally conserved areas in the binding sites, the more rigid parts of the binding pocket are nicely

Table 3. Overall Docking Performance, As Calculated Using Eqs 1 and 2[a]

		docking performance (%)	
	targets[b]	non-native	native
all structures	65	61.0 (1.2)	79.7 (1.3)
apo structures	29	55.9 (2.0)	75.9 (0.0)
nonapo structures	59	61.3 (1.3)	79.3 (1.4)

[a] Standard deviations are given in parentheses. They only reflect the nondeterministic nature of the search algorithm and do not include sampling errors, which are related to the size of the test set (see ref 29). [b] Number of targets in each subset. For example, apo structures are only available for 29 targets. Non-native docking performances listed are measured over the targets and structures in each subset. Native docking performances listed are measured over the targets in each subset.

superimposed. If, however, superimposition had been done simply on all main-chain and Cβ atoms within 4 Å of the reference ligand, the large protein movement around Gly108 would have resulted in a poor superimposition of the non-native structure and atoms in key areas around the ligand would have shifted by as much as 0.8−1.2 Å. Critically, this shift has a direct effect on the rmsd values calculated for dockings against the non-native structure.

Overall Docking Results. Table 3 lists the overall docking performance results obtained for native docking (65 entries from the Astex Diverse Set) and non-native docking (1112 structures for the same 65 targets in the Astex Non-native Set), as calculated using eqs 1 and 2, respectively. The native docking results we obtained for the 65 targets (79.7 (1.3)%) are comparable to those obtained against the full Astex Diverse Set (80.5 (0.5)%[8]), indicating that the 65 targets form a representative subset of the original 85 targets. For non-native docking, the docking performance is much lower at 61%. This is clearly a significant drop-off but not as disastrous as was observed by some other authors who reported non-native docking performance*s* below 50%. There are various explanations for this, for example: (i) The sets studied contained different targets and different ligands. (ii) We have only included structures that contained the exact same set of binding site atoms present in the reference structure; it is not clear whether this was the case in studies reported in the literature. (iii) Our superimposition focused in on the structurally conserved parts of the binding sites. (iv) The structures in our set mirror the reference structure as closely as possible in terms of protonation states, tautomers, and side-chain flips.

It is difficult to make comparisons for individual targets with other non-native docking studies presented in the literature, because the docking methodology and the complexes investigated both differ. However, we see some interesting, although strictly anecdotal, similarities between our results and those reported by Sutherland et al. for CDOCKER.[16] For example, the target that gave the most robust performance for non-native docking using CDOCKER was the estrogen receptor alpha (62%). In agreement with this, we find that the ligand from the Astex Diverse Set entry for this target can be docked correctly against either of the two non-native structures. Also, the targets for which Sutherland et al. report the biggest drop-off in docking performance are Cdk2 and p38. Here, we also observe that, whereas docking against the native structures almost invariably produces the correct binding mode for these two targets/structures, docking against non-native structures gives poor docking performance (58% for Cdk2 and 6% for p38). It is worth pointing out here that the p38 reference structure has the so-called DFG-loop in the "in" conformation, whereas 9 out of the 27 non-native structures have the DFG-loop in the "out" conformation (for these 9 structures it would be impossible for a rigid-protein docking protocol to generate the correct binding mode). In agreement with our own previous work on neuraminidase, we again found that Goldscore is a very robust scoring function for docking against non-native conformations of this target. For both strains of neuraminidase represented in the Astex Diverse Set, the ligands are docked correctly against all non-native structures.

In terms of overall docking performance, results using the Chemscore function (see the Supporting Information) are very similar to those obtained for the Goldscore function. For individual targets, however, clear differences are observed. This is not surprising as the performance of scoring functions tends to be highly target dependent.

Apo vs Nonapo Structures. From the results presented in Table 3 it appears that there is, on average, little difference between the docking performances against apo and nonapo structures. This may be a little surprising as it is often thought that apo structures tend to have somewhat collapsed binding sites. This is certainly the case for some targets, e.g. for myosin II all non-native structures in the new set are apo structures, and in all cases the binding site is completely collapsed, making it impossible to correctly predict the binding mode for the reference ligand unless protein flexibility is taken into account in the docking protocol. The Chemscore function appeared to perform slightly better for nonapo structures than for apo structures, but the difference is not statistically significant (see the Supporting Information). From Figure 2, it appears that there are some minor differences between apo and nonapo structures in terms of their distributions of maximum atom movement, and possibly site rmsd, but this does not translate into differences in the number of bad clashes. Hence, it is impossible to conclude whether in general apo structures are more suitable for protein−ligand docking applications than nonapo structures or vice versa.

Influence of Clashes. Rigid protein docking protocols, like the one we are using here, are expected to suffer more when protein movements become more pronounced, in particular when, in the non-native structure, the protein penetrates the volume occupied by the ligand in the reference structure. The results in Table 4 show this effect very clearly for rigid protein docking using GOLD against the Astex Non-native Set. The worse the clashes are between reference ligand and non-native protein structure, the poorer the docking performance is. Interestingly, however, when only non-native structures are considered for which no bad clashes with the reference ligand are observed, there is no drop-off in docking performance. Most of these non-native structures do still exhibit soft clashes with the reference ligand. This means that, in general, the docking tool we have used here (GOLD/Goldscore) is relatively insensitive to small protein movements. Chemscore results follow the same trend (see the Supporting Information), but in this case there still is a 9% drop-off in performance for non-native structures exhibiting

Table 4. Effect of Clashes on Docking Performance[a]

worst clash[b]	targets[c]	docking perfomance (%) non-native	native
$d_{clash} > -0.5$ Å	45	76.8 (0.8)	75.1 (2.4)
-1 Å $< d_{clash} \leq -0.5$ Å	39	65.6 (0.4)	77.9 (1.4)
-2 Å $< d_{clash} \leq -1$ Å	34	39.6 (3.7)	75.9 (1.3)
$d_{clash} \leq -2$ Å	18	10.5 (1.4)	70.0 (3.0)

[a] Standard deviations are give in parentheses. They only reflect the nondeterministic nature of the search algorithm and do not include sampling errors, which are related to the size of the test set (see ref 29). [b] Clash severity is defined as $d_{clash} = d - d_{vdw}$, where d is the distance between protein and ligand atom, and d_{vdw} is the sum of the van der Waals radii of the atoms involved. The worst clash (i.e., most negative d_{clash} value) is stored for each complex. [c] Number of targets for which at least one non-native structure with a "worst clash" value in the specified range exists in the Astex Non-native Set. Non-native docking performances listed are measured over the targets and structures within the specified "worst clash" range. Native docking performances listed are measured over the targets in the specified "worst clash" range.

only soft clashes with the reference ligand; this implies that the Chemscore function is more sensitive to small variations in protein conformation than the Goldscore function. Even using the Goldscore function, however, there are some cases where docking performance is highly sensitive to small changes in protein conformation. For the carbonic anhydrase entry, for example, the reference ligand is always docked correctly against its native structure but virtually never against a non-native conformation of the target, even though the protein movement is minimal. The incorrect solutions all have the pyrazole ring twisted by approximately 90°, causing the trifluoromethyl and 4-methyl-phenyl substituents to occupy alternate pockets. For native docking, the score gap between the top-ranked (correct) solution and the second-ranked (incorrect) solution is very small (only ~0.5 Goldscore units), and hence very small changes in protein conformation can cause these two solutions to be swapped around. Another example of the sensitivity of the docking results to small variations in protein conformation is adenosine deaminase. In this case, docking the reference ligand against its native conformation nearly always results in an incorrectly predicted binding mode (rmsd≈7 Å). However, docking against several of the non-native conformations available for this target produces correct binding modes. One of these non-native structures is PDB entry 1o5r; docking the reference ligand against this structure almost invariably produces binding modes with rmsd≈1Å with respect to the experimental binding mode. Again, the conformational differences between native (1uml) and non-native (1o5r) structures are minimal. In this case, the global score optimum for native docking actually corresponds to the correct binding mode (rmsd=1.0 Å, Goldscore=82.6). However, this solution is hardly ever found by the search algorithm, which could be caused by three factors: (i) the search space is significant for this ligand (11 rotatable bonds); (ii) the correct binding mode could correspond to very sharp optimum in the score landscape that is difficult to find; and (iii) the highest scoring incorrect binding mode has Goldscore=82.3, i.e. like in the case of carbonic anhydrase, the score gap is very small. For docking against non-native structures, the score gap between best-scoring correct solution and best-scoring incorrect solution varies between −1.9 and 11.7. For the 1o5r entry, this score gap is 8.5—much greater than for native docking—which we believe is why it is easier to find the correct binding mode against this non-native structure.

For the example shown in Figure 4 it is much clearer why docking against a number of non-native conformations fails, whereas it is successful against the native structure. When the reference ligand for the aldose reductase entry in the Astex Diverse Set is docked against its native structure, GOLD invariably produces the correct binding mode with an rmsd≈0.6 Å relative to experiment. However, the correct binding mode is only produced against 12 out of the 33 non-native conformations for this target. In aldose reductase a specificity pocket can open and close between Leu300 and Trp111. In the native structure for this target in the Astex Diverse Set (1t40), the specificity pocket is in the "open" form and binds the benzothiazole part of the ligand. When the reference ligand is docked against the 17 non-native structures that are in the "closed" form, none of the binding modes produced are correct. However, the ligand is docked correctly against 12 of the 16 non-native structures in the "open" form. Interestingly, if we consider the dockings against all non-native structures and select the binding mode with the best Goldscore value, then an "open" form structure is selected and a correct docking is produced (rmsd≈0.6 Å).

Influence of Native Docking Performance. We split the entries in the Astex Diverse Set into three categories, according to GOLD's ability to dock the ligand correctly against its native protein structure: (i) entries for which GOLD always produces the correct binding mode; (ii) entries for which more than half—but not all—of the solutions produced by GOLD are correct; and (iii) entries for which less than half of the solutions produced by GOLD are correct. Table 5 shows the non-native and native docking performance for the three different categories. The non-native docking performance clearly follows the native docking performance. What is interesting though is that the drop-off in performance is relatively small for entries that GOLD always predicts well (~15%), whereas it is significantly larger for complexes for which it is harder to find the correct binding mode (~30%). Chemscore results follow the same trend (see the Supporting Information). One possible explanation for this result would be that the targets in category (i) are more rigid than those in category (ii), but this is not the case: the average site RMSDs are 0.70(5) Å and 0.67(3) Å, for categories (i) and (ii), respectively. In our opinion, the observed difference stems from differences in the scoring function landscape for the two categories. When the global scoring function optimum is well-defined and "funnel-shaped" (category (i) entries), then this optimum may be insensitive to small variations in protein conformation. However, when the scoring function landscape is rugged and the global optimum is one of many optima (category (ii) entries), small changes in protein conformation may well shift the global optimum to another solution.

Multiple-Structure Docking. A widely used strategy for taking protein flexibility into account, which is also routinely applied at Astex, is to dock compounds against several conformations of the target (obtained from separate X-ray structures) and then combine the results produced by the docking algorithm. Using the results obtained against the Astex Non-native Set, we can investigate on a large and unbiased test set how useful such approaches are likely to

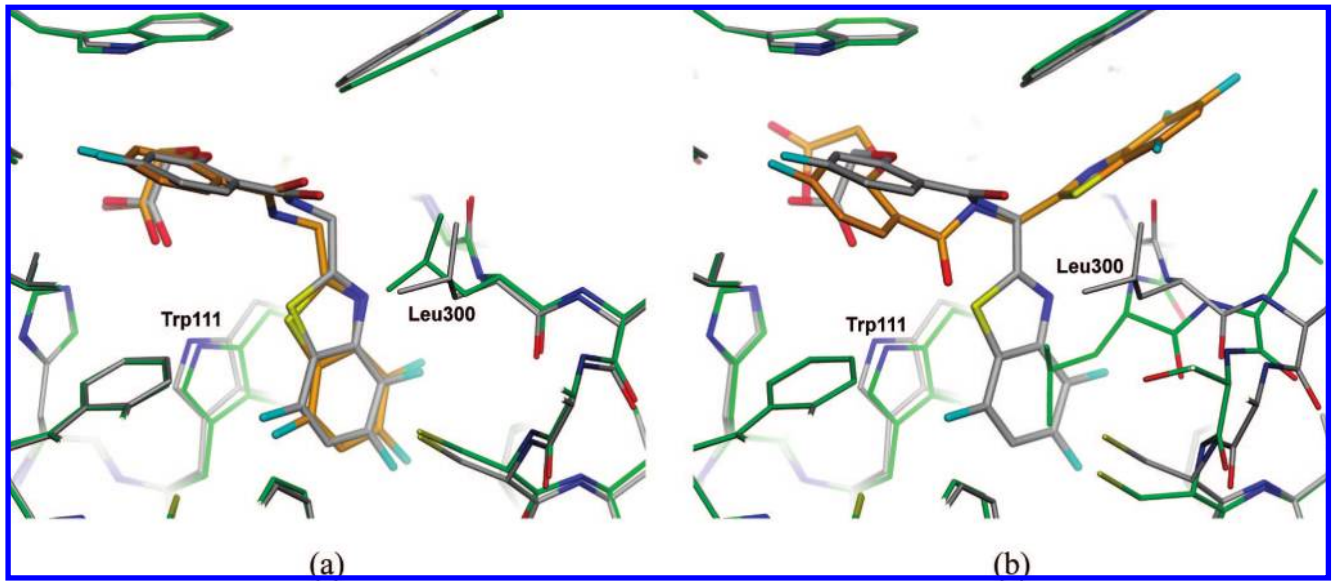

Figure 4. Top-ranked GOLD solutions for docking the aldose reductase reference ligand against non-native structures 1t41 (a) and 1el3 (b). Pictures were created using AstexViewer 2.[28]

Table 5. Effect of Native Docking Performance on Non-native Docking Performance[a]

native docking performance range[b]	targets[c]	docking perfomance (%)	
		non-native	native
100%	29	85 (1)	100 (0)
50−100%	21	60 (3)	92 (4)
<50%	15	17 (4)	23 (4)

[a] Standard deviations are give in parentheses. They only reflect the nondeterministic nature of the search algorithm and do not include sampling errors, which are related to the size of the test set (see ref 29). [b] This is the native docking performance listed for individual targets in Table 2. [c] The number of targets with the specified native docking performance. Native docking performances listed are measured over the targets in each native docking performance range. Non-native docking performances listed are measured over the targets and structures in each native docking performance range.

Table 6. Docking Performance and Sampling Performance for Various Multi-Protein-Conformer Protocols[a]

	docking performance (%)	sampling performance (%)
native docking	79.7 (1.3)	90.8 (1.9)
non-native docking; standard protocol	61.0 (1.2)	71.9 (2.1)
multiconformer; NMAX[b]=2	64.7 (0.7)	80.4 (1.0)
multiconformer; NMAX[b]=3	66.4 (1.5)	83.6 (1.5)
multiconformer; NMAX[b]=5	66.8 (1.2)	86.5 (1.7)
multiconformer; all structures	67.1 (1.8)	90.8 (1.9)

[a] Standard deviations in success rates are given in parentheses. They only reflect the nondeterministic nature of the search algorithm and do not include sampling errors, which are related to the size of the test set (see ref 29). [b] NMAX is the maximum number of non-native structures considered per target. For targets that have more than NMAX non-native structures in the Astex Non-native Set, NMAX structures were selected at random, and docking performance and sampling performance were calculated; this process was repeated five times, and the success rates were averaged.

be. When, for each target, we consider all dockings against all non-native structures and select the docking with the best score, the docking performance increases to 67% (see Table 6). Figure 5 confirms that the non-native conformations that are being selected by this protocol are sensible. When the binding modes are correctly predicted, the selected non-native protein conformations are, compared to the average non-native structure, significantly closer to the native conformation. On the other hand, when binding modes are not correctly predicted, the non-native protein conformation appears to be selected more or less randomly from the available non-native conformers.

A strategy for dealing with small induced-fit effects is to dock the compound(s) using a rigid docking protocol and then, for each of the dockings produced, allow the complex to relax, for example using force-field based methods, hopefully resulting in a better ranking of the binding modes. A proviso for this approach is that the rigid docking protocol must produce the correct binding mode as one of its solutions. At this point it is useful to define "sampling performance" as the percentage of cases for which the GOLD solution nearest to the experimental binding mode has rmsd <2 Å. It is clear from Table 6 that, like the docking performance, the sampling performance is also significantly lower for non-native dockings (71%) than it is for native dockings (91%). Therefore, a rescoring protocol like the one described above can only improve the docking performance to at best 71%, as in the remaining cases, GOLD never produces a solution within 2 Å rmsd. However, if we consider all dockings produced against all non-native structures, the sampling performance improves significantly to 91%.

There are a couple of caveats with the above methodology. First, the multiple-structure protocol allows the GA to sample longer (although an extra parameter is also introduced: the protein conformer), and this may have an effect on the success rates produced. We know from experience, however, that the "Default 1" GA setting is already quite exhaustive and that the improvements obtained by the multiple-structure protocol are unlikely to be a result of increased sampling. Second, the number of protein conformations considered here

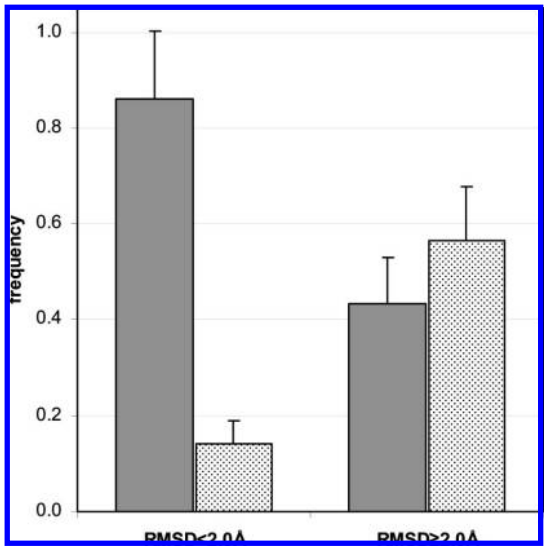

Figure 5. Frequency of selecting "correct" (solid gray) and "incorrect" (dotted) protein conformers using the multiple-structure docking protocol. Results are shown for targets for which the selected binding mode has rmsd <2.0 Å and for targets for which rmsd ≥2.0 Å. In this analysis, we regarded a selected protein conformer "correct" if its site rmsd (sRMSD) was smaller than the median site rmsd for that target. Only targets were considered for which min(sRMSD) < 1.0 Å, max(sRMSD) > 1.0 Å, and max(sRMSD)-min(sRMSD) > 0.5 Å, in order to exclude targets for which no non-native conformer close to the native conformer is available, or for which all non-native conformers are close to the native conformer, or for which all non-native conformers are very similar. Poisson counting statistics was assumed to estimate the standard deviations.

for some targets is significantly higher than we would consider practical for real-life applications (e.g., 134 conformations for HIV-1 protease). Hence, we reduced the maximum number of non-native conformations considered per target in the multiple-structure protocol down to 5, 3, and 2 non-native structures (see Table 6). Interestingly, only a very limited number of non-native structures (3−5) needs to be considered in order to achieve the improved docking performance (% top-ranked) of ∼67%. More importantly, by considering only a small number of non-native conformations the sampling performance increases significantly. By considering a maximum of 5 non-native structures per target, the sampling performance improves to 87%, giving rescoring protocols a much better chance of recovering poorly ranked correct solutions. We believe that even fewer structures would be required if the non-native conformations were selected more carefully. Here we simply selected 5 (or 3 or 2) non-native structures at random, but one could imagine picking a "conformationally diverse" selection of non-native structures instead. The results we obtained here for multiple-structure docking are very similar to what Barril and Morley observed for Cdk2 and HSP90. They also found that a limited number of structures can result in much improved sampling performance. However, they found that for the purpose of identifying actives in a virtual screening exercise against these targets, using multiple structures can be counterproductive as it is more susceptible to false positives. Although not the focus of our current study, the Astex Non-native Set would be an excellent starting point to assess whether this is true in general. Chemscore results (see the Supporting Information) closely follow the trends observed for Goldscore.

Effect of Ligand Similarity. Another generally applied strategy to increase the chances that docking will be successful for a newly designed compound is to dock against X-ray structures of complexes that contain a similar ligand. The hope is that for similar ligands, the induced fit will be similar and hence the protein conformer will be more suitable. Here we calculated the similarity of each ligand from a nonapo structure in the Astex Non-native Set with respect to their reference ligand. Next, we calculated the docking performance and sampling performance for non-native structures containing ligands that are similar to the reference ligand and for structures containing ligands that are dissimilar from the reference ligand (see Table 7). It turns out that, when protein conformations are used from complexes containing compounds which are dissimilar to the ligand that is being docked, both docking performance and sampling performance are poor. However, when protein conformers are used from complexes containing ligands that are similar to the docked ligand, docking performance and sampling performance are as high as for native docking. Chemscore results follow the same trend (see the Supporting Information). Interestingly, for non-native structures containing ligands that are similar to the reference ligand the binding site conformation is also significantly closer to the conformation of the reference structure (see Table 7). In other words, similar ligands result in similar binding site conformations. These results highlight the importance to obtain experimental binding modes regularly as part of the drug discovery cycle.

CONCLUSIONS

We have constructed a test set for assessing docking performance against non-native conformations of protein targets. This set is based on a high-quality and diverse test

Table 7. Docking Performance and Sampling Performance for Docking against Non-native Structures Containing Ligands That Are Similar to the Reference Ligand and against Structures Containing Ligands That Are Dissimilar to the Reference Ligand[a]

			docking performance (%)		sampling performance (%)	
ligand similarity[d]	targets[b]	<site rmsd>[c]	non-native	native	non-native	native
>0.7	29	0.48(6)	75.5 (1.4)	75.2 (1.5)	87.9 (2.4)	86.9 (2.9)
<0.5	54	0.98(8)	54.8 (2.1)	79.3 (1.5)	66.5 (2.6)	91.5 (2.1)

[a] Standard deviations in success rates are given in parentheses. They only reflect the nondeterministic nature of the search algorithm and do not include sampling errors, which are related to the size of the test set (see ref 29). [b] Number of targets for which at least one non-native structure of a complex containing a ligand in the specified similarity range exists in the Astex Non-native Set. Native docking performances and sampling performances listed are measured over the targets within the specified similarity range. Non-native docking performances and sampling performances listed are measured over the targets and structures within the specified similarity range. [c] Average site rmsd with respect to the reference structure for the structures within the specified similarity range. Standard errors in the means are given in parentheses. [d] Tanimoto coefficient for 2048-bit Daylight fingerprints of native and non-native ligand.

set for native protein−ligand docking: the Astex Diverse Set. The new set, the Astex Non-native Set, contains 1112 structures from the PDB for 65 targets relevant to drug discovery. When selecting and setting up the non-native structures we have ensured the following: (i) The resolution of the structures is ≤2.5 Å. (ii) All binding site residues are the same and complete in all structures. (iii) The structures are superimposed on the structurally most conserved parts of the binding site. Next, we used this new test set to establish the docking performance of GOLD against non-native protein conformations. We found that, whereas docking performance against native structures is 80% (top-ranked solution within 2 Å rmsd), it is only 61% against non-native structures. Sampling performance (any solution within 2 Å rmsd) also drops significantly: 91% for native docking and 71% for non-native docking. We observed no significant difference between docking performance against apo and nonapo structures. Small ligand-induced protein movements appear to have little effect on rigid docking performance, but larger conformational changes have a catastrophic effect on performance.

The significant drop-off in docking performance to 61% highlights the importance of docking protocols taking into account protein flexibility. Here, we investigated the simplest method for dealing with protein flexibility, i.e. parallel docking against multiple protein conformers taken directly from separate X-ray structures. When the best-scoring binding mode is selected from docking runs against several non-native structures, docking performance improves to 67%. Considering multiple protein conformers also dramatically improves sampling performance: even considering as few as 3 to 5 non-native conformations, chosen randomly from the available structures, can improve sampling performance to ∼86%. We believe these results are encouraging for docking protocols taking into account multiple non-native structures simultaneously.

Finally, we observed that docking against non-native structures of complexes containing ligands that are similar to the ligand that is being docked results in significantly higher docking performance and sampling performance, highlighting the importance of X-ray structures as a closely integrated part of structure-based design cycles.

ACKNOWLEDGMENT

The authors thank our collaborators from the Cambridge Crystallographic Data Centre, Dr. Suzanne Brewerton and Dr. Gianni Chessari from Astex for useful discussions and feedback during the construction of the Astex Non-native Set, and Dr. Ian Tickle from Astex for developing and customizing KFIT.

Supporting Information Available: Lists of HET groups for solvents and cofactors and docking results using the Chemscore function. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.

(2) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43–53.

(3) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(4) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(5) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.

(6) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM-A new method for protein modeling and design - applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488–506.

(7) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(8) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–741.

(9) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Giordanetto, F.; Cotesta, S.; Mcmartin, C.; Kihlen, M.; Stouten, P. F. Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871–881.

(10) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

(11) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **1998**, *33*, 367–382.

(12) Murray, C. W.; Baxter, C. A.; Frenkel, A. D. The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 547–562.

(13) Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I. J.; Verdonk, M. L. Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. *J Comput.-Aided Mol. Des* **2002**, *16*, 855–869.

(14) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47*, 45–55.

(15) Wu, G.; Robertson, D. H.; Brooks, C. L.; Vieth, M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm. *J. Comput. Chem.* **2003**, *24*, 1549–1562.

(16) Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J. Chem. Inf. Model.* **2007**, *47*, 2293–2302.

(17) Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337*, 209–225.

(18) Corbeil, C. R.; Englebienne, P.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model.* **2007**, *47*, 435–449.

(19) Meiler, J.; Baker, D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* **2006**, *65*, 538–548.

(20) Nabuurs, S. B.; Wagener, M.; de, V. J. A flexible approach to induced fit docking. *J. Med. Chem.* **2007**, *50*, 6507–6518.

(21) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49*, 534–553.

(22) Zhao, Y.; Sanner, M. F. FLIPDock: docking flexible ligands into flexible receptors. *Proteins* **2007**, *68*, 726–737.

(23) Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). *J. Comput.-Aided Mol. Des* **2008**, *22*, 311–325.

(24) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377–395.

(25) Kearsley, S. K. On the orthogonal transformation used for structural comparisons. *Acta. Crystallogr. A* **1989**, *45*, 208–210.
(26) Verdonk, M. L.; Cole, J. C.; Hartshorn, M.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **2003**, *52*, 609–623.
(27) Tsai, J.; Taylor, R.; Chothia, C.; Gerstein, M. The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* **1999**, *290*, 253–266.
(28) Hartshorn, M. J. AstexViewer: a visualisation aid for structure-based drug design. *J. Comput.-Aided Mol. Des* **2002**, *16*, 871–881.
(29) Nissink, J. W. M.; Murray, C. W.; Hartshorn, M. J.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins* **2002**, *49*, 457–471.

CI8002254