Multiple Active Site Corrections for Docking and Virtual Screening

Guy P. A. Vigers* and James P. Rizzi

Array BioPharma Inc., 3200 Walnut St, Boulder, Colorado 80301

Received April 8, 2003

Several docking programs are now available that can reproduce the bound conformation of a ligand in an active site, for a wide variety of experimentally determined complexes. However, these programs generally perform less well at ranking multiple possible ligands in one site. Since accurate identification of potential ligands is a prerequisite for many aspects of structure-based drug design, this is a serious limitation. We have tested the ability of two docking programs, FlexX and Gold, to match ligands and active sites for multiple complexes. We show that none of the docking scores from either program are able to match consistently ligands and active sites in our tests. We propose a simple statistical correction, the multiple active site correction (MASC), which greatly ameliorates this problem. We have also tested the correction method against an extended set of 63 cocrystals and in a virtual screening experiment. In all cases, MASC significantly improves the results of the docking experiments.

Introduction

One of the central problems in computational chemistry is the following: Given the structure of a protein active site and a list of potential small molecule ligands, predict the binding mode and estimate the binding affinity for each ligand. This is a complex and multifaceted problem and has been the subject of intensive research, since it has immediate and fundamental impact on many aspects of structure-based drug design and virtual screening. The problem is usually broken down into three phases: First, employ computational methods that can reproduce the bound conformation of a ligand in a high-resolution X-ray crystal structure. Second, use the method to rank a series of ligands, either directly from their docking scores or after rescoring using some other method. Third, use the docked conformations of the top-ranked ligands to generate more accurate estimates of ligand affinity using more computationally intensive techniques.

The first step in this process is usually referred to as "docking"[1]. Multiple programs have been written to address this area. Programs that we have tested in-house include AutoDock,[2] Dock,[1] FlexX,[3] Fred,[4] Glide,[5] Gold,[6] and QXP/Flo.[7] A comprehensive review of these packages is beyond the scope of this paper, but the reader is referred to refs 8 and 9 for further information. In general, the programs operate as follows: A large number of conformations are generated for the small molecule, either prior to docking (e.g., Dock, Fred) or internally in the docking program (e.g., FlexX, Gold). Each conformation is positioned in the active site in a variety of orientations, the combination of conformation and orientation being known as a "pose". Many poses are selected and ranked on some scoring function, to determine the best overall pose. To make the docking problem tractable, the programs usually assume that the active site is rigid (possibly including tightly bound water molecules) and that only the small molecule can move (but see also refs 9−12 for other possible methods). Also, the scoring functions developed to date usually use either a classical molecular mechanics force field (e.g., Dock, Gold), an empirical function describing terms such as hydrogen-bonding and lipophilicity (e.g., FlexX), or a "knowledge-based" potential derived from analysis of protein−ligand complexes (e.g., PMF[13,14]).

In our experience, and in common with many other authors, we have found that the best of these packages now give excellent agreement between predicted and observed binding conformations for a wide variety of ligands. In this paper we have worked with both FlexX and Gold, two programs that operate by very different methods: FlexX builds up the conformation of the molecule in the active site by docking and assembling fragments, while Gold uses a genetic algorithm on the whole molecule (and on H-bond donors in the protein) to determine the conformation and orientation of the ligand. FlexX can be combined with CScore in Sybyl,[15] so that the top-ranked poses can be evaluated with multiple different scoring functions.[16,17] Gold has its own internal scoring function based on the Sybyl force field.[6] Overall, while this is an area of intensive and ongoing development, we will assume for the moment that the docking problem may now be considered tractable, provided that the protein active site (including any essential water molecules) is in the correct conformation and that all ligands under consideration will bind optimally to this conformation of the active site.

The second step of the procedure, ranking multiple ligands against a single site, is more problematic. The scoring functions used in docking programs must of necessity contain many approximations, since they have to be fast enough to evaluate extremely large numbers of possible poses. For instance, they do not usually take explicit account of entropic changes due to displacement of waters from the active site or of reduction in degrees of freedom in the ligand. Furthermore, a scoring function that can select the correct pose for one ligand will not necessarily rank a series of potential ligands in the correct order. For instance, a scoring function that is

* Corresponding author. Phone: 303-381-6649. Fax: 303-386-1129. E-mail: gvigers@arraybiopharma.com.

10.1021/jm030161o CCC: $27.50 © 2004 American Chemical Society
Published on Web 12/06/2003

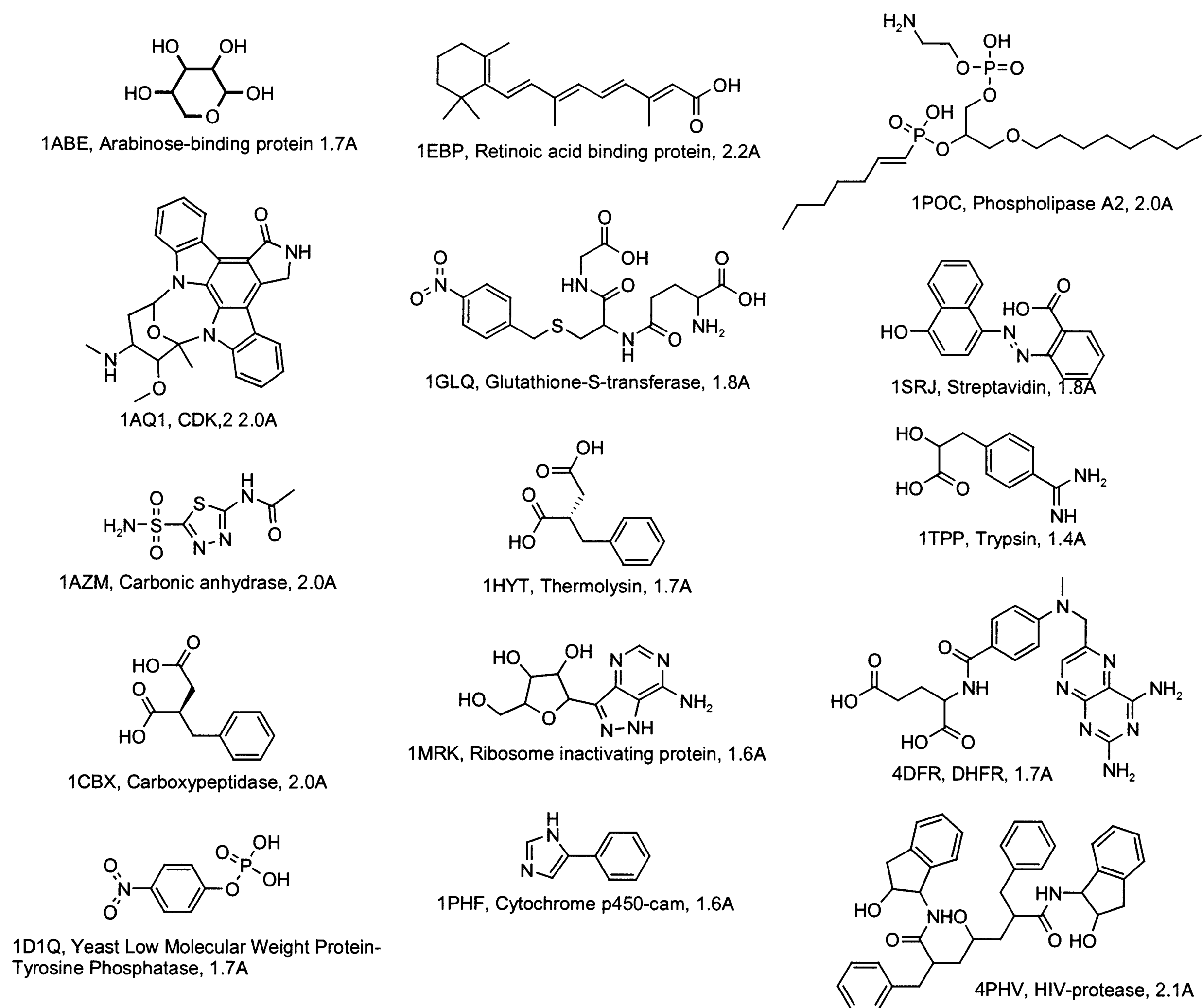

Figure 1. Ligand–protein complexes selected from the PDB for test set 1.

strongly correlated to, say, molecular weight (MW) may give the correct poses for a whole series of ligands but nonetheless predict their relative binding affinities completely incorrectly. It is with this part of the procedure that this paper is concerned.

The third step, more accurate estimation of binding energy, is beyond the scope of this paper. It can be addressed by free energy perturbation theory,[18] MM/PBSA calculations,[19] the OWFEG free-energy grid method,[20] or other techniques. Since all of these calculations are fairly computer-intensive, however, it is clear that we desire as accurate as possible a prior rank-ordering of compounds, to ensure that the best ligands are taken through to this level.

In this paper we have selected 15 diverse cocrystal structures from the Protein Data Bank (PDB).[21] We show that both FlexX and Gold can reproduce the bound conformations of each of the endogenous ligands with good accuracy. We have then taken all 15 ligands and docked them into each of the 15 active sites. We show that this step is highly problematic, since some ligands score well against all active sites, while many of the ligands score poorly against even their cognate active sites. We propose a simple correction for this problem, which we term multiple active site correction (MASC) scoring. We show that MASC scoring greatly improves the rank-ordering of the docking scores for the 15 test cases. We then examine a larger test set of 63 cocrystal complexes, to ensure that the correction is not a statistical artifact and to determine how many control sites are required for adequate correction. Final, we apply the correction method to a set of p38 inhibitors and Ptp-1b inhibitors docked into the active site of p38α, to see whether MASC scoring improves the retrieval of know inhibitors from a population of decoy molecules.

Methods

Selection of Test Set 1. Fifteen cocrystal structures were selected from the PDB, with the requirements that they be solved to ≤2.0 Å resolution (2.2 Å for retinoic acid binding protein 1EBP and 2.1 Å for HIV protease 4PHV) and that they represent diverse proteins and ligands. Most of the structures selected were taken from the lists used by the developers to test Gold and FlexX[3].[6] Ligands and active sites were prepared as described below. All complexes not reproduced by FlexX with a reasonable degree of accuracy were rejected. The final list (test set 1) is shown in Figure 1. The ligands and active sites appear to offer a range of diverse characteristics, as shown in Tables 1 and 2 of the Supporting Information.

Selection of Test Set 2. We also downloaded the CCDC/Astex test set of 92 "clean" cocrystal test structures.[22] From this set, we selected 63 complexes that had well-typed ligands. Complexes were prepared as described below. The full list of the 63 structures (test set 2) is shown in Table 3 of the Supporting Information.

Selection of Test Set 3. An additional 15 kinase complexes were chosen from the PDB. The list of 15 kinases is shown in

Table 1. Ligand Rmsd between Observed and Docked Orientations for the 15 Ligands of Test Set 1 in Their Cognate Active Sites[a]

ligand	run 1	run 2	run 3	ligand	run 1	run 2	run 3
abe_lig	1.20	0.83	0.83	hyt_lig	1.93	2.04	1.84
aq_lig	0.32	0.38	0.35	mrk_lig	1.31	1.34	1.43
azm_lig	2.23	2.34	2.17	phf_lig	0.66	1.26	1.28
cbx_lig	0.92	1.14	1.22	phv_lig	1.86	1.86	1.24
dfr_lig	2.18	1.49	4.71	poc_lig	0.81	1.05	1.80
dq_lig	1.32	2.06	1.90	srj_lig	0.72	0.72	0.72
epb_lig	1.88	2.06	2.05	tpp_lig	0.80	1.22	1.25
glq_lig	4.53	4.38	5.61	mean	1.51	1.61	1.89

[a] All results were from Gold. Run 1 used a 6/12 internal VdW potential and the default speed settings. Run 2 used a 4/8 internal VdW potential and default speed settings. Run 3 used the 4/8 internal VdW potential and the 3× speedup settings.

Table 4 of the Supporting Information. Active sites were prepared as for test set 1 (below). Thirty Ptp-1b inhibitors and 30 p38 inhibitors, all with measured $K_i < 1\ \mu$M, were selected from in-house databases. Representative compounds are shown in Figure 1 in the Supporting Information. In addition, a "decoy" set of small molecules was constructed from the MDL Drug Data Report[23] (MDDR) database of biologically relevant compounds: peptides and salts were removed and 600 molecules were randomly selected from 6654 remaining. We refer to the percentage of the database that must be screened to retrieve 90% of the know ligands as S_{90}. We have solved X-ray structures of all the p38 inhibitors cocrystallized with p38α. All of the inhibitors bind in the ATP-binding site of p38α, and none of them utilize the novel allosteric binding site recently described for certain urea-containing molecules.[24]

Preparation of Proteins and Ligands for Docking. To prepare test set 1, extraneous polypeptide chains and heteroatoms were deleted, and water molecules were deleted unless they had a B-factor of <15.0, implying that they were tightly bound. Atom and bond types were manually corrected and hydrogens were added to all ligands. For Gold, the ionization state and atom-typing of the ligands was adjusted in accordance with the Gold manual, but for FlexX the ligands were left in their default Concord/Sybyl states. Before docking, the ligand was extracted from the complex and the position and geometry of the small molecule was effectively randomized by running it through Concord[25] and minimizing for up to 1000 steps. For Gold runs, all hydrogens were added to the protein and the complex was minimized in Sybyl for 100 steps under standard conditions. This has the effect of removing major distortions in geometry but had minimal effect on the geometry of the active site. For FlexX, the original PDB file was used for docking, hydrogens were not added, and no minimization was done. Each individual ligand was docked into its cognate active site, to check that the true conformation could indeed be reproduced with reasonable accuracy and that no egregious errors had been made in preparation of the active site or small molecule (Table 1). In practice, it was found that any complex that could be correctly reproduced with FlexX could also be reproduced by Gold, but that Gold could correctly reproduce complexes that FlexX could not.

For test set 2, hydrogens were already present in the CCDC/Astex mol2 files. No waters were added to the structures and the protein was not minimized. Atom-typing and ionization states were checked for all ligands. Ligand conformations were effectively randomized by 1000 steps of simulated annealing at 300 °C in vacuo, followed by up to 1000 steps of minimization.

Test set 3 was treated similarly to test set 1: Waters with B-factors of ≤15 were included and the complexes were minimized as above. The endogenous ligands were not used for any docking studies, but the kinase active sites were used to generate control scores for the MASC studies of p38α described below.

Docking Trials. All FlexX runs were performed with FlexX version 1.9, using the default parameters. Initial Gold trials were performed using the default parameters (1× speedup, 6/12 internal potential, no flipping of amide bonds, etc). These are referred to as parameter set 1. Later trials were performed with a 4/8 internal VdW potential, 3× speedup, and allowing flipping of amides and rings (see the Gold documentation for detailed explanation of these parameters). These are referred to as parameter set 2. No significant differences were seen between versions 1.2 and 2.0 of Gold, and the 3× speedup was found to produce generally excellent docking results (Table 1). The internal 4/8 VdW potential was chosen after docking studies showed that it was essential for accurate docking of some molecules. All docking trials for both FlexX and Gold were performed with an active site radius of 10.0 Å for all proteins. In all docking runs, only the top-scoring conformation was selected for each molecule.

Calculation of Mean Ranking Error (MRE). We used two measures to assess the accuracy of the docking scores: The first measure is simply the number of ligands correctly matched with their active sites. The second measure is the mean ranking error, which was calculated as follows: Given scores S_{ij} for a series of ligands (i) in an active site (j), one can assess an error function:

$$\mathrm{Err}_j = (S_{j\mathrm{best}} - S_{jj})/(S_{j\mathrm{best}} - S_{j\mathrm{worst}})$$

where S_{jj} is the score for the correct (endogenous) ligand (j) in site (j), $S_{j\mathrm{best}}$ is the best score for any ligand in site (j), and $S_{j\mathrm{worst}}$ is the worst score for any ligand. ($S_{j\mathrm{best}} - S_{j\mathrm{worst}}$) is then the range of scores for all the ligands in the active site, and ($S_{j\mathrm{best}} - S_{jj}$) is the difference between the correct ligand score and the best ligand score. Thus, $\mathrm{Err}_j = 0$ if the correct ligand scores best of all the tested ligands, and $\mathrm{Err}_j = 1$ if it scores worst of all. The mean ranking error (MRE) can then be calculated by taking the mean value of Err_j across all active sites (j). A value of MRE = 0.0 would indicate perfect scoring, while MRE = 0.5 would indicate approximately random scoring.

Calculation of MASC Scores. For a ligand (i) docked across multiple active sites (j) one can calculate standard statistics as follows:

$$\mu_i = \sum\nolimits_j (S_{ij})/N \quad j = 1, N$$

$$(\sigma_i)^2 = \sum\nolimits_j (S_{ij} - \mu_i)^2/(N-1) \quad j = 1, N$$

where μ_i and σ_i are the mean and standard deviation of the scores for compound i across all active sites j.

The docking scores can then be corrected as follows:

$$S_{ij}' = (S_{ij} - \mu_i)/\sigma_i \quad \text{the MASC score}$$

where S_{ij}' is the modified score for compound i in active site j and S_{ij} is the original score. We have termed this corrected score the multiple active site correction score or MASC score. As shown below, this simple statistical correction greatly increases the accuracy of ligand scoring and reduces the MRE of docking studies.

Transformation of Gold Scores. The MASC described above requires that the scores are normally distributed. This was found to be true for the FlexX scores (data not shown), but the distribution of Gold scores showed a long negative tail (Figure 2, left). To place the Gold scores on a normal distribution, an empirical transformation was applied as follows:

$$S_{ij}' = \exp(S_{ij}/50.0)$$

where S_{ij} is the original score and S_{ij}' is the modified score. This transformation places the Gold scores in an approximately Gaussian distribution (Figure 2, right). The exact values of the correction function did not appear to be critical: The divisor could range from 25 to 75 with good results, and any exponent could be used. It is important to note that this transformation does not change the rank-ordering of scores, but simply places them on a normal distribution. Other

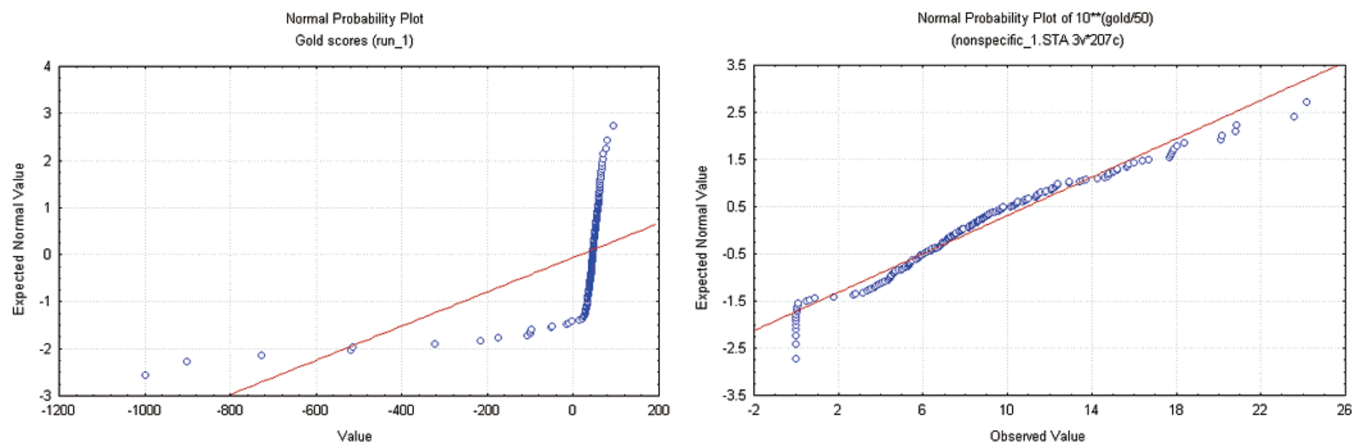

Figure 2. (left) Normal-probability plot of Gold scores for all 15 ligands in the 15 active sites shown in Figure 1. The scores are far from normally distributed and have a long negative tail. (right) Normal-probability plot of transformed scores (see Methods). The scores now show an approximately normal distribution.

Table 2. FlexX/Chemscore Scores for All 15 Ligands against All 15 Active Sites[a]

	abe_ site_1	aq_ site_1	azm_ site_1	cbx_ site_1	dfr_ site_1	dq_ site_1	epb_ site_1	glq_ site_1	hyt_ site_1	mrk_ site_1	phf_ site_1	phv_ site_1	poc_ site_1	srj_ site_1	tpp_ site_1
abe_lig	-24.5	-17.1	-13.2	-25.0	-14.4	-19.2	-10.6	-14.7	-21.0	-14.9	-16.6	-15.8	-13.4	-15.7	-11.9
aq_lig	-22.6	-41.4	-21.2	-22.4	-25.6	-16.1	-31.6	-36.8	-22.5	-18.9	-18.6	-24.4	-25.2	-24.8	-23.7
azm_lig	-20.9	-15.4	-30.3	-30.7	-21.0	-25.8	-18.1	-25.5	-38.1	-22.5	-20.7	-24.0	-26.6	-22.0	-20.1
cbx_lig	-19.6	-20.5	-31.8	-42.0	-27.9	-30.5	-20.1	-22.4	-42.3	-20.4	-26.3	-20.7	-23.5	-23.3	-17.0
dfr_lig	-22.4	-27.0	-23.6	-38.4	-32.1	-29.1	-29.5	-26.3	-43.5	-33.5	-18.9	-22.6	-26.3	-27.3	-22.4
dq_lig	-13.4	-18.9	-22.9	-33.3	-21.4	-34.0	-16.7	-19.4	-24.2	-23.2	-16.7	-19.1	-21.8	-16.5	-14.6
epb_lig	-25.0	-36.2	-36.9	-26.6	-25.6	-36.0	-46.8	-33.8	-36.6	-35.1	-30.2	-26.5	-36.0	-30.8	-25.1
glq_lig	-15.4	-21.9	-25.9	-27.3	-29.6	-26.5	-19.5	-29.0	-42.1	-26.0	-23.8	-27.5	-34.3	-26.0	-25.6
hyt_lig	-17.0	-19.5	-30.6	-37.8	-23.0	-28.0	-21.8	-24.0	-42.2	-18.5	-28.6	-22.5	-24.5	-20.9	-15.7
mrk_lig	-20.0	-17.5	-21.3	-23.9	-18.6	-22.7	-20.6	-20.1	-27.1	-22.0	-21.4	-18.8	-17.2	-17.6	-15.4
phf_lig	-16.5	-29.9	-25.8	-27.8	-21.9	-25.9	-23.1	-21.9	-29.0	-25.0	-28.4	-19.0	-20.0	-28.5	-16.4
phv_lig	-39.3	-46.7	-35.0	-46.1	-41.3	-22.8	-57.4	-35.6	-59.2	-46.5	-28.9	-57.0	-35.9	-37.9	-28.5
poc_lig	-21.8	-30.3	-36.3	-27.6	-33.5	-19.6	-22.9	-32.5	-40.7	-31.1	-25.1	-29.1	-39.6	-21.7	-28.7
srj_lig	-26.9	-29.2	-23.0	-32.0	-28.1	-26.1	-41.7	-29.0	-37.9	-34.0	-28.7	-29.8	-31.6	-42.6	-26.8
tpp_lig	-14.1	-23.1	-24.8	-25.5	-20.3	-24.9	-18.5	-21.3	-28.6	-22.3	-19.7	-25.4	-19.4	-25.2	-25.6

[a] Active sites (abe_site, aq_site etc.) run across the page and ligands (abe_lig, aq_lig etc) run down the page. The best-scoring compound is highlighted for each active site (i.e. for each column). Correctly identified endogenous ligands are colored green, while incorrect ones are colored red. All ligands were docked with FlexX and scored using the Sybyl Chemscore. More negative scores are better.

transformations could probably be used equally well, as long as they produce an approximately Gaussian distribution of scores and do not change the rank-ordering of the compounds. All Gold scores reported in this paper have been transformed in this fashion. They are referred to as “raw” gold scores in order to distinguish them from the MASC-corrected scores. No transformation was necessary for FlexX scores.

Results

Reproduction of 15 Cocrystal Structures. Fifteen cocrystal structures were selected from the PDB and prepared as described in Methods. Each of the 15 ligands from test set 1 was docked into its cognate active site, to ensure that the docking programs were able to reproduce the bound conformations correctly. As expected, both FlexX and Gold were able to reproduce the bound conformations of these 15 ligands with good accuracy. Representative rms deviations between predicted and observed orientations are shown in Table 1. The results are generally excellent, with mean rmsd values of 1.5−1.9 Å. The worst results were for ligand glq_lig (4.5−5.6 Å), where a nitrophenyl group was consistently docked into an alternate pocket.

Docking of all 15 Ligands against all 15 Active Sites. All 15 ligands were then docked into each of the 15 active sites, using FlexX and the Chemscore scoring function. The results are shown in Table 2. Since the ligands and active sites are so diverse (Figure 1 and Supporting Information), it is unlikely that any ligand will bind to any except its cognate active site. Furthermore, we have already checked that each endogenous ligand can be docked with good fidelity. It should therefore be easy for the docking scores to match each active site with its endogenous ligand. However, it can be seen from Table 2 that this is not the case: Only three out of 15 ligands score as the best ligands for their cognate active sites, while the mean ranking error (MRE; see Methods) is 25%. Certain compounds are found to score very well against multiple sites. For instance, phv_lig (an HIV-protease inhibitor, Figure 1) is ranked as the best ligand for eight out of the 15 active sites. This is clearly not an adequate success rate for virtual screening. One explanation for the poor success of the docking score is that some ligands are evaluated as good binders regardless of the active site, for instance,

Table 3. FlexX Scores after Correction with MASC Scoring

	abe_site_1	aq_site_1	azm_site_1	cbx_site_1	dfr_site_1	dq_site_1	epb_site_1	glq_site_1	hyt_site_1	mrk_site_1	phf_site_1	phv_site_1	poc_site_1	srj_site_1	tpp_site_1
abe_lig	-1.88	-0.14	0.79	-1.99	0.51	-0.62	1.40	0.42	-1.05	0.39	-0.02	0.17	0.74	0.19	1.10
aq_lig	0.37	-2.41	0.57	0.38	-0.08	1.32	-0.96	-1.73	0.38	0.90	0.95	0.09	-0.03	0.04	0.20
azm_lig	0.56	1.52	-1.08	-1.15	0.55	-0.30	1.05	-0.24	-2.46	0.28	0.60	0.02	-0.43	0.38	0.70
cbx_lig	0.81	0.69	-0.76	-2.07	-0.25	-0.59	0.74	0.44	-2.10	0.70	-0.05	0.66	0.31	0.34	1.14
dfr_lig	0.89	0.18	0.71	-1.56	-0.59	-0.14	-0.21	0.28	-2.34	-0.82	1.43	0.85	0.29	0.13	0.89
dq_lig	1.27	0.36	-0.31	-2.03	-0.05	-2.16	0.73	0.27	-0.52	-0.35	0.73	0.33	-0.12	0.77	1.07
epb_lig	1.23	-0.60	-0.72	0.96	1.12	-0.58	-2.34	-0.22	-0.68	-0.43	0.37	0.97	-0.57	0.27	1.20
glq_lig	1.85	0.78	0.13	-0.10	-0.48	0.03	1.18	-0.38	-2.51	0.12	0.48	-0.14	-1.25	0.11	0.17
hyt_lig	1.07	0.74	-0.76	-1.72	0.26	-0.41	0.43	0.13	-2.32	0.86	-0.48	0.34	0.07	0.55	1.24
mrk_lig	0.09	0.94	-0.35	-1.22	0.56	-0.81	-0.12	0.06	-2.28	-0.59	-0.36	0.51	1.03	0.89	1.66
phf_lig	1.65	-1.32	-0.41	-0.85	0.46	-0.44	0.17	0.45	-1.11	-0.22	-0.99	1.09	0.88	-1.01	1.66
phv_lig	0.17	-0.50	0.57	-0.45	-0.01	1.68	-1.48	0.52	-1.64	-0.48	1.12	-1.44	0.48	0.30	1.16
poc_lig	1.17	-0.14	-1.07	0.28	-0.63	1.50	1.00	-0.49	-1.75	-0.27	0.65	0.05	-1.58	1.19	0.09
srj_lig	0.75	0.34	1.43	-0.15	0.54	0.90	-1.85	0.37	-1.19	-0.49	0.44	0.23	-0.07	-2.02	0.76
tpp_lig	2.29	-0.15	-0.59	-0.80	0.61	-0.62	1.09	0.33	-1.63	0.07	0.78	-0.74	0.87	-0.70	-0.82

[a] Active sites (abe_site, aq_site etc.) run across the page and ligands (abe_lig, aq_lig etc) run down the page. The best-scoring compound is highlighted for each active site (i.e. for each column). Correctly identified endogenous ligands are colored green, while incorrect ones are colored red. More negative scores are better.

because they are large (e.g. aq_lig) or hydrophobic (e.g. epb_lig). Some component of this "stickiness" may in fact represent true biologically relevant nonspecificity,[26] but most of it is more likely to be due to ligand-dependent biases in the docking algorithm's scoring function. If the function is less than perfect, the programs will tend to overestimate the binding affinity of some ligands for all active sites, regardless of whether the ligands will in fact bind to them or not.

Multiple Active Site Correction of Docking Scores. To reduce the problems of ligand-dependent bias, we constructed a modified score for each molecule, as described in the Methods section:

$$S_{ij}' = (S_{ij} - \mu_i)/\sigma_i$$

where S_{ij} is the score for molecule i in site j, S_{ij}' is the modified score, and μ_i and σ_i are the mean and standard deviation of S_{ij} for molecule i against all active sites j. We have called this corrected score the multiple active site correction score or MASC score. The physical significance of the MASC score is discussed more fully below, but it can be thought of either as a statistical measure of ligand specificity or as a correction for ligand-related bias in the scoring function. The modified scores for the 15 ligands in test set 1 can readily be calculated from the data shown in Table 2. The resulting MASC scores are shown in Table 3. The number of correctly identified ligands has now improved from 3/15 to 11/15, while the MRE has dropped from 25% to 5.5%. The same ligands and active sites were also docked with Gold, using either parameter set 1 or 2 (see Methods). In both cases, the raw docking scores performed poorly at identifying the correct ligand for each active site, but the MASC scores did significantly better. The results for Gold with parameter Set 2 are shown in Tables 4 and 5: Before correction, 4/15 ligands are correctly identified, with an MRE of 33%. After correction, 11/15 ligands are correctly identified, with an MRE of 9%. Interestingly, the most promiscuous ligands identified by FlexX and Gold are different: in Table 2 the two highest scoring ligands are phv_lig and epb_lig, while in Table 4 they are poc_lig and glq_lig. This implies that the ligand-dependent biases for FlexX/Chemscore and Gold/parameter set 2 are somewhat different. However, the MASC-scoring technique works equally well in both cases.

The MASC was found to improve the performance of all the scoring functions in FlexX/CScore, as shown in Table 6. Before correction the best CScore function was Chemscore (3/15 correct, MRE = 25%) and the worst was Dock (3/15 correct, MRE = 42%). After MASC, the best score was again Chemscore (11/15 correct, MRE = 5.5%) and the worst was again Dock (7/15 correct, MRE = 18%). However, even the worst score after MASC was better than the best score before correction. Similarly, using Gold and parameter set 2, MASC scoring improved the results from 4/15 correct (MRE = 32%) to 11/15 correct (MRE = 9%, Table 6). While both FlexX and Gold worked well on the complexes in test set 1, additional work showed that Gold is able to handle a wider variety of problems, especially with parameter set 2 (see Methods). All further work described here was therefore performed with Gold and parameter set 2.

Testing MASC Scoring on a Larger Test Set. An important question, left unanswered by the work described so far, is "How many sites are required to calculate a reliable score?" Also, might the MASC only work when the same active sites are used for the calculation of the MASC as for the calculation of the MRE? In the worst case, is the apparent improvement a coincidence, dependent on the composition of the test set? To address these questions, we generated a second test set of 63 cocrystals from the CCDC/Astex test set, as described in Methods. Gold docking was performed on all 63 structures, using all 63 ligands as before. With the raw Gold scores, the MRE for all compounds against all active sites was 42.0%. After MASC, the MRE fell to 17.5% (Table 6). These results are somewhat less accurate than for test set 1, both before and after MASC scoring. This may be because less care was taken

Table 4. Gold Scores before MASC Scoring[a,b]

	abe_ site_1	aq_ site_1	azm_ site_1	cbx_ site_1	dfr_ site_1	dq_ site_1	epb_ site_1	glq_ site_1	hyt_ site_1	mrk_ site_1	phf_ site_1	phv_ site_1	poc_ site_1	srj_ site_1	tpp_ site_1
abe_lig	3.08	2.02	2.34	2.27	1.92	2.42	1.95	1.93	2.30	2.27	2.32	2.19	2.45	2.30	1.98
aq_lig	0.00	6.16	2.17	2.52	3.01	3.15	0.72	2.72	2.51	4.19	0.01	2.97	3.12	3.05	2.87
azm_lig	1.56	2.32	2.51	2.64	2.34	2.95	2.38	2.32	2.54	2.59	2.72	2.29	2.71	2.56	2.26
cbx_lig	2.56	2.99	2.73	5.08	3.14	3.67	2.88	2.74	3.16	3.50	3.46	2.76	3.22	3.21	2.53
dfr_lig	0.01	4.33	3.67	5.19	5.35	4.15	4.02	4.17	5.83	4.58	0.46	3.98	3.80	4.85	3.66
dq_lig	2.28	2.59	2.92	4.26	2.82	4.35	2.47	2.52	2.53	3.20	3.42	2.33	2.89	3.13	2.60
epb_lig	0.09	3.34	2.60	2.75	3.31	3.48	3.78	2.73	2.58	2.67	0.20	2.84	2.82	3.13	2.98
glq_lig	0.02	4.36	4.55	4.49	5.47	5.26	4.51	4.35	4.85	5.13	1.74	5.13	6.32	5.65	4.04
hyt_lig	2.68	2.99	2.89	4.08	2.97	3.64	2.86	2.73	3.36	3.37	3.17	2.70	3.56	3.15	2.65
mrk_lig	0.73	2.44	2.42	2.42	2.64	2.90	2.53	2.58	2.52	3.21	2.29	2.69	2.89	3.09	2.29
phf_lig	1.76	2.16	1.88	1.96	2.11	2.32	2.02	1.99	1.97	2.21	2.46	1.96	2.08	2.16	2.26
phv_lig	0.00	5.48	3.61	3.40	5.26	5.13	5.24	4.09	5.48	4.24	0.00	8.30	5.46	4.69	5.50
poc_lig	0.00	4.67	3.96	6.41	5.34	3.52	4.79	5.23	4.40	5.81	0.14	4.70	8.78	6.26	4.02
srj_lig	0.08	3.31	2.63	3.97	3.32	2.60	3.49	2.64	2.70	3.45	1.18	2.89	3.11	4.95	2.80
tpp_lig	0.94	2.81	2.71	4.10	2.82	3.65	2.63	2.91	2.60	3.26	3.05	3.16	3.14	3.01	2.54

[a] Gold was run with parameter set 2, as described in Methods. More positive scores are better. [b] Active 0sites (abe_site, aq_site etc.) run across the page and ligands (abe_lig, aq_lig etc) run down the page. The best-scoring compound is highlighted for each active site (i.e. for each column). Correctly identified endogenous ligands are colored green, while incorrect ones are colored red.

Table 5. Gold Scores after MASC Scoring[a]

	abe_ site_1	aq_ site_1	azm_ site_1	cbx_ site_1	dfr_ site_1	dq_ site_1	epb_ site_1	glq_ site_1	hyt_ site_1	mrk_ site_1	phf_ site_1	phv_ site_1	poc_ site_1	srj_ site_1	tpp_ site_1
abe_lig	2.82	-0.78	0.30	0.07	-1.13	0.59	-1.02	-1.09	0.19	0.08	0.24	-0.21	0.68	0.18	-0.91
aq_lig	-1.69	2.29	-0.29	-0.06	0.26	0.35	-1.22	0.07	-0.07	1.02	-1.68	0.23	0.33	0.28	0.17
azm_lig	-2.82	-0.41	0.20	0.61	-0.35	1.60	-0.21	-0.40	0.30	0.46	0.88	-0.49	0.86	0.37	-0.60
cbx_lig	-0.98	-0.30	-0.70	3.03	-0.06	0.78	-0.47	-0.69	-0.02	0.51	0.45	-0.66	0.07	0.06	-1.02
dfr_lig	-2.40	0.29	-0.12	0.82	0.92	0.17	0.09	0.19	1.22	0.44	-2.12	0.07	-0.04	0.61	-0.13
dq_lig	-1.06	-0.57	-0.06	2.05	-0.22	2.19	-0.76	-0.68	-0.67	0.39	0.73	-0.97	-0.10	0.27	-0.55
epb_lig	-2.38	0.68	-0.02	0.13	0.64	0.81	1.09	0.10	-0.04	0.05	-2.27	0.21	0.19	0.48	0.34
glq_lig	-2.77	-0.02	0.10	0.06	0.68	0.55	0.08	-0.03	0.29	0.47	-1.68	0.47	1.22	0.80	-0.22
hyt_lig	-1.06	-0.31	-0.56	2.31	-0.37	1.25	-0.63	-0.94	0.58	0.61	0.12	-1.01	1.06	0.07	-1.13
mrk_lig	-3.15	-0.11	-0.16	-0.16	0.23	0.68	0.04	0.13	0.02	1.25	-0.39	0.32	0.68	1.03	-0.39
phf_lig	-1.83	0.38	-1.12	-0.69	0.15	1.30	-0.34	-0.55	-0.66	0.67	2.05	-0.71	-0.01	0.43	0.95
phv_lig	-2.08	0.51	-0.37	-0.47	0.41	0.35	0.40	-0.14	0.51	-0.07	-2.08	1.85	0.50	0.14	0.53
poc_lig	-2.04	0.06	-0.26	0.84	0.36	-0.46	0.11	0.31	-0.06	0.58	-1.98	0.07	1.91	0.78	-0.23
srj_lig	-2.50	0.39	-0.22	0.98	0.40	-0.24	0.55	-0.21	-0.16	0.51	-1.51	0.02	0.21	1.85	-0.07
tpp_lig	-2.87	-0.11	-0.26	1.78	-0.10	1.12	-0.39	0.04	-0.42	0.55	0.24	0.40	0.37	0.17	-0.52

[a] Active sites (abe_site, aq_site etc.) run across the page and ligands (abe_lig, aq_lig etc) run down the page. The best-scoring compound is highlighted for each active site (i.e. for each column). Correctly identified endogenous ligands are colored green, while incorrect ones are colored red. More positive scores are better.

preparing active sites and ligands for the larger test set, the complexes were not minimized before docking (see Methods), or the larger test set included more difficult cases. However, once again MASC scoring greatly improves the accuracy of ligand ranking.

All 63 ligands were then docked against the 15 active sites from test set 1. Using these sites for MASC gave a MRE of 18.7%, which is almost as accurate as the full set of test set 2. Thus, a set of 15 independent active sites can be used to correct all 63 ligands across all 63 active sites. To find out how few active sites are required, we recalculated the MASC scores using subsets of sites from test set 1, selected in several different ways (alphabetically, reverse-alphabetically, or in three different random orders). The results are shown in Figure 3. In all cases, as additional active sites are added to the correction set, the MRE drops from an uncorrected value of 42% to a value of approximately 18%. In general it appears that any seven to nine control sites will provide near-optimal MASC.

Application of MASC Scores to Database Enrichment Studies. From the above results, it appears that the MASC scoring works well when the correction is calculated from a series of diverse control sites. However, it is also possible that the technique would work better if the correction was calculated from a set of sites more closely related to the target site, rather than from a diverse set of control sites. To address this question, and to test the method further, we performed a virtual-screening experiment as follows: We selected 30 inhibitors of p38α MAP kinase[27−29] and 30 inhibitors of Ptp-1b protein-tyrosine phosphatase[30−33] from our internal datasets, and mixed them with 600 druglike molecules from the MDDR (see Methods). We also

Table 6. Summary of the Docking Results from Both FlexX and Gold, against Test Sets 1 and 2[a]

			raw score		MASC score	
test set	program	scoring function	correct/ total	MRE	correct/ total	MRE
1	FlexX	Chemscore	3/15	0.25	11/15	0.06
1	FlexX	Dock score	3/15	0.42	7/15	0.18
1	FlexX	FlexX score	6/15	0.26	8/15	0.18
1	FlexX	Gold score	3/15	0.36	9/15	0.14
1	FlexX	PMF score	4/15	0.32	5/15	0.19
1	Gold/ps 1	Gold	6/15	0.18	10/15	0.12
1	Gold/ps 2	Gold	4/15	0.32	11/15	0.09
2	Gold/ps 2	Gold	5/63	0.42	26/63	0.18
2[b]	Gold/ps 2	Gold	5/63	0.42	20/63	0.19

[a] FlexX results were scored using the five different scoring functions implemented in FlexX/CScore. Gold results were scored with the Gold scoring function in all cases, but different parameter sets (ps) were used for docking (see Methods). [b] Corrected with test set 1.

selected 15 kinase cocrystals from the PDB to serve as a convergent set of control sites as described in Methods (test set 3). The 660 potential ligands were docked into the active sites of p38α and Ptp-1b, as well as into the control sites from test sets 1 and 3 (i.e. the diverse control sites and the kinase controls, respectively). Compounds were then ranked by raw Gold score or by MASC score.

The known Ptp-1b ligands were readily retrieved as the best ligands for the Ptp-1b active site, before or after MASC scoring (data not shown). However, the situation was more interesting for the p38α active site: As shown in Figure 4, the raw Gold scores yield a slight enrichment of the p38 ligands over the MDDR compounds, with 90% of the inhibitors found in the first 60% of the database (Figure 4, top). That is, $S_{90} = 60\%$ for the p38 ligands (see Methods). However, Ptp-1b inhibitors score better than p38 inhibitors and are preferentially selected. In fact, $S_{90} = 13\%$ for the Ptp-1b inhibitors, even though we are screening against the p38α active site! However, this anomaly is largely corrected by MASC scoring (Figure 4, middle): After MASC scoring, $S_{90} = 38\%$ for the p38 inhibitors, while $S_{90} = 79\%$ for the Ptp-1b inhibitors. Thus, p38 inhibitors are now preferentially selected over both Ptp-1b inhibitors and over compounds from the MDDR. Moreover, the 30 kinases of test set 3 appear to give a weaker correction than the diverse set (Figure 4, bottom), with $S_{90} = 62\%$ for the p38 inhibitors, and $S_{90} = 67\%$ for the Ptp-1b inhibitors. Therefore, MASC scoring improves the results of database-enrichment studies, and a set of diverse control sites appears to give better correction than a set of sites related to the target.

Discussion and Conclusions

In this paper we have shown that, while the best available docking programs are now capable of docking a single ligand into a protein active site with a high probability of success, they tend to perform poorly when comparing multiple potential ligands. We have shown that this problem is not because the docking programs fail to reproduce the bound conformation of the ligand, but is rather because of ligand-dependent biases in the scoring function. Some ligands are scored as binding tightly to many active sites, whereas others are rarely evaluated as good. We have described a simple correction, the MASC score, which greatly ameliorates the problem by calculating docking scores for each ligand against a series of control sites. We have shown that this correction works well with two different docking programs (FlexX and Gold) and with multiple scoring functions and docking parameters. We have also shown that seven to nine diverse control sites appear to be sufficient for optimal correction, regardless of the target site of interest. The MASC is different for each molecule, and is also specific for the docking program and scoring function used. However, once it has been calculated for a molecule, it can be used to correct the docking score for that molecule against any target site of interest. Also, the MASC-scoring methodology remains conceptually the same, whatever docking program and scoring function are used.

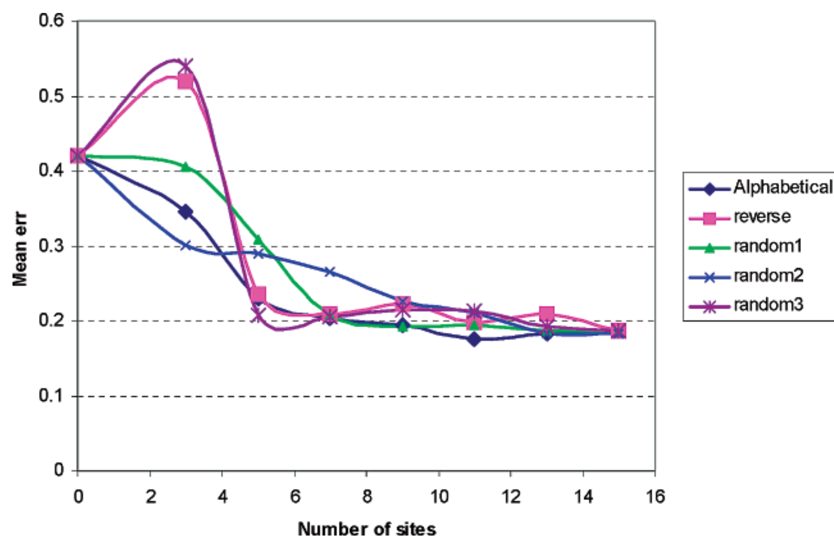

Figure 3. Variation of mean ranking error (MRE) with number of sites. Increasing numbers of control sites from test set 1 were selected alphabetically, reverse-alphabetically, or in three different random orders. The scores from these sites were then used to calculate values of μ_i and σ_i for use in MASC scoring. In every case, the MRE fell from the uncorrected value of 42% to approximately 18% as the number of correcting sites increased. At three active sites the correction sometimes made the MRE worse, presumeably because three sites are insufficient to give a reliable value of σ_i. In most cases, good MASC was seen with seven or more active sites.

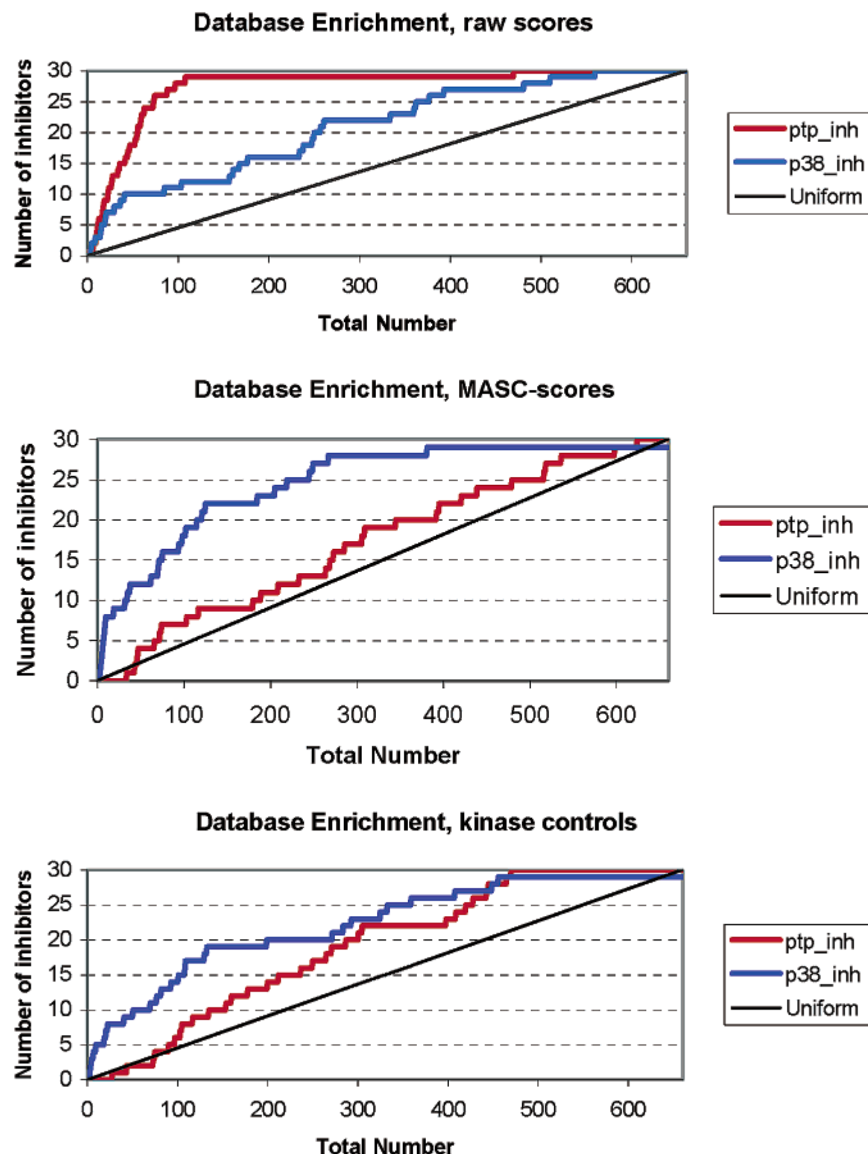

Figure 4. Database enrichment study. Thirty p38 inhibitors, 30 Ptp-1b inhibitors, and 600 compounds from the MDDR were docked into the active site of p38α and then ranked by score before or after MASC scoring. All panels show the number of p38 or Ptp-1b inhibitors returned against the total number of compounds selected. The top panel uses the raw Gold scores, the middle panel uses the MASC score corrected against the 15 control site of test set 1, and the lower panel uses the MASC scores corrected against the 15 kinases of test set 3.

The MASC is seemingly straightforward but, to the best of our knowledge, has not previously been applied to docking and virtual screening. Possibly this is because the technique is computationally somewhat expensive. However, with the continuing decline in computing cost, this is not a significant problem. We have also set up a relational database using the open-source MySQL software[34] to store the docking scores. The system is employed as follows: First, all compounds of interest are prescreened against a set of control sites and the docking scores are stored in the database. Since the docked conformations themselves do not need to be stored, this leads to a quite compact database representation. These control scores are used to calculate values of μ_i and σ_i for use in the MASC . Each screening target can then be run against the database as it becomes available, and the MASC can be applied from the stored data.

The MASC score can be thought of in several ways: If, in the best possible case, the initial docking scores were accurate evaluations of the binding free energy across all molecules in all sites, then the MASC score would be equivalent to a measure of specificity for each compound against the target site, in units of standard deviations about the mean. However, if the docking scores have ligand-dependent biases, then the correction will function more simply to reduce these biases. For instance, if the score for molecule i in site j, S_{ij}, was related to the true free energy of binding, ΔG_{ij}, by the equation

$$S_{ij} = A_i \Delta G_{ij} + C_i$$

where A_i and C_i are molecule-dependent scaling factors, then the mean score μ_i would be proportional to C_i, and σ_i would be proportional to A_i. Thus the MASC score would be independent of both these terms. Of course, if the value of S_{ij} has a more complex relationship to ΔG_{ij}, then the MASC may fail. However, the experiments described above indicate that, with current docking and

scoring methods, the MASC scoring does indeed significantly improve the results. This implies that ligand-dependent biases in the scoring functions are indeed a significant limitation on the use of docking scores for virtual screening.

We have shown that the MASC improves the retrieval of true inhibitors and decreases the retrieval of false inhibitors, when applied to database enrichment studies. One danger of database enrichment studies is shown by the fact that the raw Gold scores give excellent enrichment for Ptp-1b inhibitors over decoy compounds, even when docking in the active site of p38α, a clearly erroneous result. However, MASC scoring does greatly reduce the erroneous selection of Ptp-1b inhibitors and improves the selection of true p38 inhibitors. We have also shown that a set of diverse control sites provides better correction than a set of convergent controls. This result may seem counterintuitive, since a convergent set of controls might be expected to provide finer discrimination between potential ligands. However, the result can be understood as follows: ligands that score well against p38α may also score well against several other kinases. They will then have a high mean score for the control (kinase) sites and their MASC scores will revert to near zero. Conversely, ligands that score poorly against p38α may also score poorly against the other control kinases, and their MASC scores will also revert to near zero. Therefore, discrimination will actually be reduced if the control sites are chosen to be similar to the target site.

Other authors[35] have suggested alternative empirical corrections for the scoring functions, such as dividing by $\sqrt{N}$, where N is the number of non-hydrogen atoms in the molecule, or $\sqrt{\text{MW}}$. In our hands this does indeed also help, reducing the MRE for test set 2 from 42% to 30% (compared to 18% for MASC scoring). Linear regression analysis of the results from test set 2 indicated that the raw Gold scores appeared to correlate most closely with the number of rotatable bonds, with a second component due to the number of hydrogen-bond donors (data not shown). However, corrections based on these terms did not improve on the correction achieved by $\sqrt{N}$. This correction is certainly much simpler than the full MASC scoring and can be applied to all scores before calculation of the MASC score. Notice, however, that such corrections will cancel out on calculation of the MASC score: As described above, factors that are added to a molecule's score will modify the value of the mean, while multiplicative factors will modify the mean and σ. These factors will then cancel out on calculation of the MASC score. In fact, it is likely that the $\sqrt{N}$ correction is a specific example of one of the many components that are incorporated in the MASC. It is also possible that a more exhaustive search could derive a single-active site correction for a particular docking score that would perform better than the $\sqrt{N}$ term, but we have not as yet been able to find one. Also, such a correction is likely to be specific to a given docking method and scoring function, whereas the method for the MASC scoring remains the same whatever docking and scoring methods are used.

Another fruitful suggestion is to employ multiple sequential filters on potential ligands: One can prefilter the molecular databases for druglike compounds[36] and the docking results can be reexamined by human experts in order to prioritize the resulting hits.[36] One can also use consensus scoring to filter the docking results through a series of different scoring methods, to remove those poses that are only compatible with a few of the scoring algorithms.[8,9,37] The filtering can be done to retain only those poses that score well in all scoring functions[16,17] or by using a linear combination of the scoring functions.[38] One of the benefits of sequential filtering may be that it helps to correct for systematic errors in the scoring functions, though in a less explicit way than the MASC score. We have not yet attempted to combine these methods with the MASC, though there is no reason this should not be done. For instance, one could generate a linear combination of several scoring functions and then apply a MASC to the result or even use MASC on each of the individual scores before combining them. Such approaches, however, are beyond the scope of this paper.

This study was originally undertaken to test the concept of differential docking, where a set of compounds is docked against a target (e.g. p38α) and the top hits from this screen are then tested for specificity against other target family members (e.g. p38γ[39]). However, our results show that the scoring functions currently available fall far short of the accuracy necessary for this kind of docking. Presumably, as the scoring functions improve (e.g. refs 40−42) and the docking programs become still better, MASC will become steadily more similar to differential docking. In the meantime, we hope that multiple active site correction will serve to improve the reliability of docking and virtual screening studies and aid in structure-based drug design efforts.

Acknowledgment. We thank the other scientists at Array BioPharma for stimulating discussions. We particularly thank Barb Brandhuber and Ken Brameld for critical reading of the manuscript, Jim Blake for providing the cleaned-up MDDR database, Ellen Laird for the 15 kinase cocrystals, and Kevin Condroski for setting up and maintaining our Linux cluster.

Supporting Information Available: Tables of ligand properties, active site properties, cocrystals selected for test set 2, and kinases selected for test set 3. Also figure showing three representatives each of the p38 and Ptp1b inhibitors. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kuntz, I. D.; et al. A Geometric Approach to Macromolecule−Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(2) Morris, G. M.; et al. Automated docking using a lamarckian genetic algoritm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(3) Rarey, M.; et al. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−89.
(4) *FRED*, Openeye Software Inc., Santa Fe, NM.
(5) *Glide*, Schrodinger Inc.
(6) Jones, G.; et al. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(7) McMartin, C.; Bohacek, R. Powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333−344.
(8) Bissantz, C., Folkers, G.; Rognan, D. Protein-based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(9) Toledo-Sherman, L. M.; Chen, D. High-throughput virtual screening for drug discovery in parallel. *Curr. Opin. Drug Discovery Dev.* **2002**, *5* (3), 414−421.

(10) Broughton, H. B. A method for including protein flexibility in protein–ligand docking: Improving tools for database minimg and virtual screening. *J. Mol. Graphics* **2000**, *18*, 247–304.
(11) Clausen, H.; et al. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377–395.
(12) Waszkowycz, B. Structure-based approaches to drug design and virtual screening. *Curr. Opin. Drug Discovery Dev.* **2002**, *5* (3), 407–413.
(13) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein–ligand interactions: A simplified approach. *J. Med. Chem.* **1999**, *42*, 791–804.
(14) Muegge, I.; et al. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J. Med. Chem.* **1999**, *42*, 2498–2503.
(15) *Sybyl*, Tripos Inc.
(16) Charifson, P. S.; et al. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100–5109.
(17) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.
(18) Kollman, P. A. Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules. *Acc. Chem. Res.* **1996**, *29*, 461–469.
(19) Huo, S.; et al. Molecular dynamics and free energy analyses of Cathepsin D-inhibitor interactions: Insight into structure-based ligand design. *J. Med. Chem.* **2002**, *45*, 1412–1419.
(20) Pearlman, D. A.; Charifson, P. S. Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP Kinase protein system. *J. Med. Chem.* **2001**, *44*, 3417–3423.
(21) Berman, H. M.; et al. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(22) CCDC, CCDC/Astex dataset.
(23) *MDDR*, MDL/Prous Scientific Inc.
(24) Pargellis, C. A.; et al. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nature Struct. Biol.* **2002**, *9* (4), 268–272.
(25) Pearlman, R. S. *Concord user's manual*; Tripos Inc., St. Louis, MO.
(26) McGovern, S. L.; et al. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **2002**, *45*, 1712–1722.
(27) Wilson, K. P.; et al. Crystal structure of p38 mitogen-activated protein kinase. *J. Biol. Chem.* **1996**, *271* (44), 27696–27700.
(28) Tong, L.; et al. A highly specific inhibitor of human p38 MAP kinase binds in the ATP pocket. *Nature Struct. Biol.* **1997**, *4* (4), 311–316.
(29) Wilson, K. P.; et al. The structural basis for the specificity of pyridinylimidazole inhibitors of p38 MAP kinase. *Chem. Biol.* **1997**, *4*, 423–431.
(30) Pannifer, A. D. B.; et al. Visualization of the cysteinyl-phosphate intermediate of a protein-tyrosine phosphatase by X-ray crystallography. *J. Biol. Chem.* **1998**, *273*, 10454–10462.
(31) Elchebly, M.; et al. Increased insulin sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase-1b gene. *Science* **1999**, *283*, 1544–1548.
(32) Andersen, H. S.; et al. 2-(oxalylamino)-benzoic acid is a general, competitive inhibitor of protein-tyrosine phosphatases. *J. Biol. Chem.* **2000**, *275* (10), 7101–7108.
(33) Iversen, L. F.; et al. Structure-based design of a low molecular weight, nonphosphorous, nonpeptide, and highly selective inhibitor of protein-tyrosine phosphatase 1b. *J. Biol. Chem.* **2000**, *275* (14), 10300–10307.
(34) *MySQL*, MySQL Inc.
(35) Pan, Y.; et al. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267–272.
(36) Blake, J. F.; Laird, E. R. Recent advances in virtual ligand screening. *Annu. Rep. Med. Chem.* **2003**.
(37) Clark, R. D.; et al. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics* **2002**, *20*, 281–295.
(38) Terp, G. E.; et al. A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein–ligand binding affinities. *J. Med. Chem.* **2001**, *44*, 2333–2343.
(39) Bellon, S.; et al. The structure of phosphorylated p38gamma is monomeric and reveals a conserved activation-loop conformation. *Structure* **1999**. *7*, 1057–1065.
(40) Eldridge, M. D.; et al. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.
(41) Stahl, M.; Bohm, H.-J. Development of filter functions for protein–ligand docking. *J. Mol. Graphics* **1999**, *16*, 121–132.
(42) Nissink, J. W. M.; Verdonk, M. L.; Klebe, G. Simple knowledge-based descriptors to predict protein–ligand interactions. Methodology and validation. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 787–803.

JM030161O