General and Targeted Statistical Potentials for Protein–Ligand Interactions

Wijnand T. M. Mooij* **and Marcel L. Verdonk**
Astex Therapeutics Ltd., 436 Cambridge Science Park, Milton Road, Cambridge, CB4 0QA United Kingdom

ABSTRACT **We present a novel atom–atom potential derived from a database of protein–ligand complexes. First, we clarify the similarities and differences between two statistical potentials described in the literature, PMF and Drugscore. We highlight shortcomings caused by an important factor unaccounted for in their reference states, and describe a new potential, which we name the Astex Statistical Potential (ASP). ASP's reference state considers the difference in exposure of protein atom types towards ligand binding sites. We show that this new potential predicts binding affinities with an accuracy similar to that of Goldscore and Chemscore. We investigate the influence of the choice of reference state by constructing two additional statistical potentials that differ from ASP only in this respect. The reference states in these two potentials are defined along the lines of Drugscore and PMF. In docking experiments, the potential using the new reference state proposed for ASP gives better success rates than when these literature reference states were used; a success rate similar to the established scoring functions Goldscore and Chemscore is achieved with ASP. This is the case both for a large, general validation set of protein–ligand structures and for small test sets of actives against four pharmaceutically relevant targets. Virtual screening experiments for these targets show less discrimination between the different reference states in terms of enrichment. In addition, we describe how statistical potentials can be used in the construction of targeted scoring functions. Examples are given for cdk2, using four different targeted scoring functions, biased towards increasingly large target-specific databases. Using these targeted scoring functions, docking success rates as well as enrichments are significantly better than for the general ASP scoring function. Results improve with the number of structures used in the construction of the target scoring functions, thus illustrating that these targeted ASP potentials can be continuously improved as new structural data become available. Proteins 2005;61:272–287.** **© 2005 Wiley-Liss, Inc.**

Key words: statistical potentials; protein–ligand interactions; docking; virtual screening; targeted scoring functions

INTRODUCTION

Finding novel leads for drug targets is a crucial first step in drug design. There are various ways to go about this process of hit and lead discovery, employing experimental and computational techniques. If the three-dimensional (3D) structure of the drug target under consideration is known, structure-based design techniques can be used. Predicting binding geometries and affinities for protein–ligand complexes is at the heart of many structure-based drug design methods. A large number of different protein–ligand docking programs have been developed to propose binding modes for (potential) ligands. DOCK,[1] FlexX,[2] AutoDock,[3] and GOLD[4,5] are probably the most widely used, but many more have been described in the literature.[6] With efficient search engines, these docking calculations can be fast enough to be used in a virtual screening set-up, where large databases of compounds are screened against a drug target of interest.

In addition to efficient search algorithms, a method is needed to score the proposed binding modes. From a theoretical point of view, this would require a calculation of the free energy of binding. Unfortunately, methods to predict binding free energy (differences) are computationally expensive. Therefore, in the interest of computational speed, very simplified potentials are often used in protein–ligand docking.[7] Obviously, speed is even more essential in a virtual screening context. Some of these functions are (partly) based on molecular mechanics force fields.[8–10] Other functions are regression based,[11,12] where parameters are derived from a set of experimental binding affinities and structures.

An entirely different approach is to try to exploit the wealth of experimental information available from the Protein Data Bank (PDB).[13] A large number of protein–ligand structures is present in this database, and this number is rising at an ever-increasing rate. All of this structural information must contain information about the underlying molecular recognition process, and over the years a number of methods have been described to derive

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

*Correspondence to: W. Mooij, Astex Therapeutics Ltd., 436 Cambridge Science Park, Milton Road, Cambridge, CB4 0QA United Kingdom. E-mail: w.mooij@astex-therapeutics.com

Received 11 November 2004; Revised 4 March 2005; Accepted 4 March 2005

Published online 16 August 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20588

scoring functions directly from protein–ligand structures. Such scoring functions are known as either knowledge-based potentials, potentials of mean force, or statistical potentials. We will review Drugscore[14] and PMF,[15] two widely used scoring functions of this type, and describe a new statistical potential, the Astex Statistical Potential (ASP), that we have developed. We will investigate the influence of the reference state on the derived statistical potential by constructing two additional statistical potentials that differ only in their choice of reference state but are otherwise identical to ASP. The reference states for those two potentials are defined along the lines of Drugscore and PMF. Since the aim of this study is to compare reference states rather than scoring functions *per se*, these functions are not exact copies of the methods described in literature

In an ideal world, a universal scoring function would suffice for all targets. In practice, however, the success rate of predicting the correct binding mode, as well as enrichments obtained by virtual screening, vary widely with the combination of protein target and scoring function.[16,17] Given the simplicity of the scoring functions, this may not be surprising. In any case, it would be very useful if one could use target-specific data to tailor a scoring function in order to increase its performance for that specific target. Statistical potentials seem to be an ideal starting point for such targeted scoring functions, especially when target-specific data comes in the form of protein–ligand structures. We will describe how to derive targeted statistical potentials based on structural knowledge of protein–ligand complexes for a specific target.

Drugscore and PMF

A number of different statistical potentials have been described in the literature.[14,15,18–26] We will focus the comparison here on Drugscore[14] and PMF,[15] which are the most widely used. The basis for any statistical potential is the comparison of the observed number of contacts between certain atom types to the number of contacts one would expect if there were no interaction between the atoms. The term *reference state* is often used in this context, and is usually thought of as the (hypothetical) state of no interaction. In practice, it is constructed as some kind of average over all observations. The term is, however, rather imprecise. For example, it does not clarify whether any correction terms for excluded volume are part of the reference state, or if they are corrections to be applied to it. We will use the term *reference state* to mean the expected number of contacts if there were no interaction between the atoms, incorporating any corrections.

The choice of reference state is crucial in determining how the raw distributions of observations are transformed into potentials, and PMF and Drugscore differ considerably in how this number of expected contacts is calculated. Atom–atom contacts are analyzed in terms of radial-distribution functions (RDFs), so the number of contacts, n_{obs}, between protein atoms of type i and ligand atoms of type j in the database are tabulated as a function of their separation r. A statistical potential between two atom types i and j is then defined as

$$\mathrm{StatScore}(i,j,r) = -\ln \frac{n_{\mathrm{obs}}(i,j,r)}{\mathrm{n}_{\mathrm{exp}}^{\mathrm{StatScore}}(i,j,r)} \qquad (1)$$

where $n_{\mathrm{exp}}^{\mathrm{StatScore}}(i,j,r)$ is the reference state, or the expected number of contacts between atom types i and j at distance r.

In PMF, the expected number of contacts between atom types i and j is calculated as

$$n_{\mathrm{exp}}^{\mathrm{PMF}}(i,j,r) = \frac{\sum_{r'=0}^{R_{\max}} n_{\mathrm{obs}}(i,j,r')}{(4/3)\pi R_{\max}^{3} \cdot F(j,R_{\max})} \cdot 4\pi r^{2}\Delta r \cdot f(j,r) \qquad (2)$$

where Δr is the bin width used in the RDFs, and $R_{\max}$ is the maximum distance between protein and ligand atoms considered in the RDF. The terms $f(j,r)$ and $F(j,R_{\max})$ are ligand-dependent volume corrections that will be explained in more detail below. Without these volume corrections, the expected number of contacts is the product of the average contact density in the sphere with radius $R_{\max}$ and the volume of a spherical shell at distance r. It is key to notice that without the volume correction the reference state is not distance dependent (apart from the trivial r^2 dependency for the volume of a spherical shell). Also note that in PMF, the score for a specific atom pair is completely independent of the occurrences of contacts for other pairs of atom types.

As stated above, the expected number of contacts contains corrections to account for the volume occupied by the ligand.[27] The reasoning behind this is that not all of the volume around a ligand atom j is actually available to a protein atom i, as protein atoms cannot occupy space already taken by ligand atoms. The effect of excluded volume on RDFs has previously been recognized in a molecular dynamics study of monosaccharides in water.[28] For the calculation of the expected number of contacts in statistical potentials, one should only use the available volume, instead of the full volume of a spherical shell. Hence the multiplication by $f(j,r)$, which is the fraction of the spherical shell at distance r around a ligand atom of type j that is available to protein atoms. The value $f(j,r)$ is a function of the distance: 0 for very short distances, rising to 1 for long distances. The more buried within a molecule a certain atom type tends to be, the lower the fraction will become at shorter distances. In PMF, $f(j,r)$ is calculated for each atom type j by counting the number of ligand atoms of any type around all ligand atoms of type j in the PDB. This number of ligand atoms is translated into a fraction of excluded volume by dividing it by the number of atoms at distance r for a reference sphere completely filled with protein.

A similar correction term, $F(j,R_{\max})$, is included in the calculation of the average contact density. The value of $F(j,R_{\max})$ is the fraction of the complete sphere of radius $R_{\max}$ that is available to protein atoms. This term appears to be less important than $f(j,r)$, as $F(j,r_{\max})$ must be close to 1 for all atom types: the fraction of the volume of a sphere of 12 Å ($R_{\max}$ used in PMF) that is occupied by ligand atoms will be relatively small.

The Drugscore potential is defined as $-\ln g(i,j,r)/g(r)$, where $g(i,j,r)$ is the normalized RDF for atoms i and j, and $g(r)$ is the mean normalized RDF over all atoms types. To highlight the differences from PMF, we reformulate this into the form of eq. (1), which results in

$$n_{\text{exp}}^{\text{Drugscore}}(i,j,r) = \sum_{r'=0}^{R_{\max}} \left(\frac{n_{\text{obs}}(i,j,r')}{4\pi r'^2 \Delta r} \right) \cdot 4\pi r^2 \Delta r \cdot C(r) \tag{3}$$

for the expected number of contacts in Drugscore, with

$$C(r) = \frac{1}{(I \times J)} \cdot \sum_{i'} \sum_{j'} \frac{n_{\text{obs}}(i',j',r)/4\pi r^2 \Delta r}{\sum_{r'=0}^{R_{\max}} (n_{\text{obs}}(i',j',r')/4\pi r'^2 \Delta r)} \tag{4}$$

where I and J are the total number of different atom types for protein and ligand, respectively.

As in the PMF potential, the expected number of contacts for Drugscore is based on the average contact density for atom types i and j. However, instead of multiplying by a ligand volume correction term as in PMF, this average contact density is now multiplied by $C(r)$, an average contact-preference for all atoms types. The value of $C(r)$ is the mean over all atom type combinations of the ratio of the contact density at distance r and its average contact density. Now, at short range all atom pairs will exhibit a lower-than-average contact density, as a result of volume already occupied by bonded and neighboring atoms. So the $C(r)$ term expresses the general tendency of all atom pairs to form fewer interactions at short range, and therefore it can be seen as a 'universal' correction for the reduced available volume at shorter atomic separation. In other words, Drugscore does not use an explicit, ligand atom-type specific volume correction like PMF, but excluded volume effects are not completely ignored, as an average correction is incorporated.

It is worth noting that the Drugscore reference state is not so much a state of no interaction as one of average interaction. Interactions that are common to many protein and ligand atom types, such as van der Waals contacts, will be represented in $C(r)$, and as a result they will be partly 'averaged out' from the final potentials. Another effect of the Drugscore approach is that the expected number of contacts for a specific pair of atom types contains the number of observed contacts for any pair of atom types at that distance. So, observations for an atom pair of types C and D will influence the potential for interaction between atom types A and B via their influence on the reference state.

In addition to the statistical atom–atom pair potentials, Drugscore also contains a solvent-accessible surface (SAS)-dependent singlet term. This potential is constructed by calculating the non-polar surface area on both protein and ligand in both complexed and uncomplexed states. It captures the tendency of protein and ligand atoms to become buried upon formation of a protein–ligand complex, rewarding burial of certain atom types and penalizing the burial of others. The addition of the SAS term to the pair potential has been reported[14] to give only a slight improvement in recognizing experimental binding modes.

Astex Statistical Potential (ASP)

One key thing still appears to be missing from both the Drugscore and the PMF reference states: the difference in accessibility of different protein atoms, due to the non-uniform distribution of atom types in the protein. The propensity of different residues, and hence of different protein atom types, to occur on the SAS of ligand binding sites will vary. As a result, certain atom types will be more likely to be found in the vicinity of a ligand atom, irrespective of any specific interactions between protein and ligand atoms. Also, when two protein atom types are equally likely to occur on the surface of a binding site, they can still differ at short range, as certain atoms will be more shielded by their bonded and neighboring protein atoms than others. For example, around a side-chain lysine NH_3 there tends to be more space available for ligand atoms than around a backbone NH. The expected number of contacts should be corrected for these characteristics of the environment of different atom types in the protein. Another way of looking at it is this: why would we only correct for the volume available to protein atoms around ligand atoms, and not for the volume available to ligand atoms around protein atoms?

One could hope that the SAS-dependent term of Drugscore is somehow capturing such effects, as this term is obviously dependent on available surfaces (and therefore volumes). This term, however, takes the SAS of the uncomplexed protein as a starting point and scores the differences in surface area that occur upon complex formation. In this way, the SAS potential scores the burial of non-polar surface area, it does not score or correct for the tendency of certain atom types to have more or less SAS to start with in the uncomplexed state. Therefore, this score does not capture the effect that fewer contacts can be expected for protein atom types that are more shielded.

To be able to correct for these effects, information on the environment of protein atoms around ligand binding sites is required. This can be obtained by calculating the available volume in spherical shells around each protein atom that is within $R_{\max}$ of any ligand atom. Such a protein volume correction will capture the differences in available volumes around protein atom types, exactly like the ligand volume correction used in PMF does for ligand atoms. In addition, it will also capture the propensity for atom types to occur buried inside the protein, or exposed on the surface; an atom type that generally occurs within the protein interior will have very low available volume, not just at short distances but also at longer distances.

We constructed the novel ASP potential, which incorporates available volume corrections for protein as well as for ligand atoms. Both volume corrections were calculated with the use of a grid, where the points are marked as 'protein,' 'ligand,' or 'solvent.' Co-factors are considered to be part of the protein. The fraction of available volume for a protein atom type i is calculated as the fraction of non-protein grid points at distance r, averaged over all protein atoms of type i that are within 8.0 Å of a ligand atom, in all complexes in the database.

$$f_p(i,r) = \langle(\mathrm{gridpts}_i^{\mathrm{total}}(r) - \mathrm{gridpts}_i^{\mathrm{protein}}(r))/\mathrm{gridpts}_i^{\mathrm{total}}(r)\rangle \tag{5}$$

The ligand volume correction is calculated similarly, averaging over all ligand atoms of type j in all complexes.

$$f_l(j,r) = \langle(\mathrm{gridpts}_j^{\mathrm{total}}(r) - \mathrm{gridpts}_j^{\mathrm{ligand}}(r))/\mathrm{gridpts}_j^{\mathrm{total}}(r)\rangle \tag{6}$$

For ASP, the expected number of contacts for a given atom pair i and j is defined as

$$n_{\mathrm{exp}}^{\mathrm{ASP}}(i,j,r) = \left\langle \frac{n_{\mathrm{obs}}(i,j,r')}{f_p(i,r')f_l(j,r')4\pi r'^2\Delta r} \right\rangle_{r'=6.0}^{r'=8.0} \cdot f_p(i,r) \cdot f_l(j,r) \cdot 4\pi r^2 \Delta r \tag{7}$$

The expected number of contacts is the product of an average contact density with the doubly corrected volume of a sphere at distance r. The average contact density is taken to be the average between 6.0 and 8.0 Å of the corrected RDF, and not the overall average for the sphere with radius R_{max}. At this long range, atoms are thought to be making no specific interactions, and are therefore an appropriate choice to supply the reference contact density. This also ensures that the scores will be close to zero within this distance range.

Targeted Scoring Functions

To derive scoring functions targeted to a specific protein or protein family, one obviously needs some target-specific information that can be used to improve the scoring function. Astex's approach to drug discovery is based on fragment-based screening using high-throughput X-ray crystallography. As a result, large numbers of protein–ligand crystal structures are routinely obtained, even in the early stages of a drug-discovery project. Therefore it seems appropriate to use this wealth of structural data to improve general scoring functions. One could think of many ways to exploit this detailed information on 3D binding modes for series of molecules. In the context of this work, we investigate whether a useful targeted scoring function can be derived along the lines of statistical potentials.

Our targeted-scoring function is based on a general database, the PDB, and a target-specific database of protein–ligand crystal structures. The statistical atom–atom potentials are derived from the target-specific database, the general database, or from a combination of the two. If there is no information in the target-specific database, the PDB-derived potentials are used. If there is some information in the target-specific database, the observations from the two databases are mixed, and if there is sufficient target-specific information, only those data are used.

Defining $c_{\mathrm{obs}}(i,j,r)$, the volume-corrected density of observations at distance r for the atom types i and j is

$$c_{\mathrm{obs}}(i,j,r) = n_{\mathrm{obs}}(i,j,r)/(f_p(i,r) \cdot f_l(j,r) \cdot 4\pi r^2 \Delta r) \tag{8}$$

the ASP score can be rewritten as

$$\mathrm{ASP}(i,j,r) = -\ln \frac{c_{\mathrm{obs}}(i,j,r)}{\langle c_{\mathrm{obs}}(i,j,r')\rangle_{r'=6.0}^{r'=8.0}} \tag{9}$$

For the targeted potentials, the volume-corrected density of observations is defined as a mix between the PDB and the target-specific database using

$$c_{\mathrm{obs}}^{\mathrm{mix}}(i,j,r) = \{1 - w[n_{\mathrm{obs}}^{\mathrm{targ}}(i,j)]\} \frac{c_{\mathrm{obs}}^{\mathrm{PDB}}(i,j,r)}{n_{\mathrm{obs}}^{\mathrm{PDB}}(i,j)} + w(n_{\mathrm{obs}}^{\mathrm{targ}}(i,j)) \cdot \frac{c_{\mathrm{obs}}^{\mathrm{targ}}(i,j,r)}{n_{\mathrm{obs}}^{\mathrm{targ}}(i,j)} \tag{10}$$

where $c_{\mathrm{obs}}^{\mathrm{PDB}}(i,j,r)$ and $c_{\mathrm{obs}}^{\mathrm{targ}}(i,j,r)$ are the corrected densities of observations in the PDB and in the target-specific database respectively, and $n_{\mathrm{obs}}^{\mathrm{PDB}}(i,j)$ and $n_{\mathrm{obs}}^{\mathrm{targ}}(i,j)$ are the total number of observations within R_{max} in the two databases. Normalizing by those totals puts the target-specific data and the general PDB densities on the same scale. We can normalize in this straightforward way because the absolute scale of $c_{\mathrm{obs}}^{\mathrm{mix}}(i,j,r)$ itself is irrelevant, as ASP is finally based on the ratio of this density and its long-range average. The importance given to the specific versus the general data is determined by the weight function $w[n_{\mathrm{obs}}^{\mathrm{targ}}(i,j)]$ as a function of the total number of observations for atom types i and j in the target-specific database. In this study, we used a linear switching function, $w[n_{\mathrm{obs}}^{\mathrm{targ}}(i,j)] = \min[1, n_{\mathrm{obs}}^{\mathrm{targ}}(i,j)/N_{\mathrm{max}}]$, where N_{max} is the chosen number of observations at which the potentials become completely determined by the data in the targeted database.

MATERIALS AND METHODS

Potentials

Statistical potentials were calculated using the ASP reference state, and for comparison we also calculated potentials using reference states along the lines of Drugscore and PMF. For all three potentials, details of atom typing, grid spacing, etc. were identical. This allows us to focus on the influence of the choice of reference state on the potentials and the results, which could otherwise be obscured by the influence of these other implementation details.

For the PMF reference state, we did not attempt to recalculate the ligand-volume correction exactly as described in the literature. Instead we used the ligand-volume correction as calculated for the ASP scaling. We then calculated the potentials fully analogously to ASP, omitting the protein-volume correction. This does not amount to an exact implementation of PMF as described in the literature. However, it does capture the essence of the PMF scaling: potentials are calculated directly from RDFs while applying an available volume correction for the ligand atoms. These potentials are referred to as PMF-scaled, to indicate that the numbers of observations have been transformed into potentials ('scaled') using a reference state along the lines of PMF.

For the Drugscore scaling, the method by which the potentials were calculated from the raw distributions is identical to the one described in literature.[14] However, these Drugscore-scaled potentials will differ from the original Drugscore potentials, most importantly because we used different atom typing. In addition, we based our potential on a larger selection of protein–ligand complexes from the PDB, using the crystal symmetry to expand the ligand-binding sites where necessary. Also, we did not implement the SAS-dependent singlet potentials, as we wished to focus on the influence of the choice of reference state on the potential. These potentials are referred to as Drugscore-scaled, to indicate that the Drugscore reference state has been used.

To derive the pair-potentials, a database of protein–ligand complexes was constructed from the PDB. Ligand bond types were assigned using an in-house program, which uses a combination of rules and algorithms used in BALI[29] and an approach developed by Sayle.[30] Bond types for proteins were assigned based on residue and atom names. Each ligand was classified as normal, covalent, or co-factor. The assignment of *co-factors* was based on the following common residue names: HEM, NAD, FAD, ADP, ATP, NAP, NDP, GDP, and IDP. *Covalent ligands* were assigned based on short distances between protein and ligand. *Normal ligands* were all remaining ligands. Only binding sites for normal ligands with a heavy-atom count between 6 and 60 were added to the database. Crystallographic symmetry was used to expand the binding sites if needed; symmetry-related protein and ligand atoms were generated up to a radius of 15 Å around each ligand. If a ligand occurred more than once in a single PDB entry, and the binding sites were (pseudo) symmetric, only the first occurrence of the ligand was used. This was automatically decided based on the similarity of contacts between ligand and protein. For a complete list of PDB entries used in the derivation of the potentials, see the Supplementary Material.

Atom–atom contact tables, as well as protein and ligand available volumes, were calculated for the atom types given in Table I. Atom types were determined based on the element types of the atoms, combined with the bond types; the presence or absence of hydrogen atoms was not taken into consideration Many atom types employed were similar to the atom types used in PMF; for most carbon atom types, both a non-polar and a polar version were defined. Like the Drugscore authors, we decided not to use hydrogen atoms, and for nitrogen atoms, this means that it can be uncertain if an atom is a hydrogen-bond acceptor or a donor. Additional nitrogen atom types were defined, trying to make sure that nitrogen atoms that can safely be assumed to be donors were not grouped together with those that can safely be assumed to be acceptors, or with those for which the protonation state is uncertain. Also, atom types were chosen to keep the atoms in different functional groups separate, in order to capture differences in hydrogen-bonding behavior (e.g. nitro, sulfonamide and carbonyl groups). Additional atom types were defined to separate backbone nitrogen and oxygen atoms from those in asparagine, glutamine, aspartic acid, and glutamic acid side chains, and to separate serine and threonine hydroxyl oxygen from those in tyrosine.

Atom typing was completed in an identical manner for both protein and ligand molecules, but obviously many atom types only occur in ligand molecules. Only metal-coordinating atoms were typed separately on the protein. This was to prevent short protein-acceptor to ligand-acceptor distances that occur between acceptors that coordinate the same metal ion and lead to general attractive potentials between such acceptor types. Contact tables were generated with 0.1 Å bins, volume corrections with 0.2 Å bins, atom–atom pair potentials with 0.1 Å bins. Some smoothing of the pair potentials was achieved by assigning observations to both neighboring bins with a weight of ⅓ of the weight for the central bin.

Atom–atom potentials were calculated for each atom pair with more than 150 observations within $R_{\max}$. For atom pairs that did not fulfil this criterion, the pair potential was set to zero for all distances. Different values for this cut off were tested, but in general the success rate in reproducing binding modes for the complexes in the Cambridge Crystallographic Data Centre (CCDC)/Astex validation set[31] proved to be insensitive to this. However, the subset of metal complexes proved more sensitive, as the number of metal-acceptor observations can be low. The total number of metal contacts is simply limited by the occurrence of the ligand atom type. This is in contrast to contacts to a protein carbon atom type, for example, in which one occurrence of a ligand atom type will give rise to many observations in different bins, due to different occurrences of that atom type in the protein. Despite this low number of observations for some ligand acceptor types, they do contain useful information. For example, the contacts can nearly exclusively occur below 3 Å (e.g. for the O.nco type), telling us that these ligand atoms very much favor coordinating a metal ion. If a larger cut-off value for the minimum number of observations is used, some metal-acceptor potentials will be set to zero. This proves more harmful than using a rather noisy potential based on a limited number of observations.

At short range, bins will have no observations, and for those bins the potential is set to 10. As mentioned above, some atom-type combinations like metal-acceptor contacts can lack observations at long range. To prevent strong repulsion between those atom types at those distances, the potential is set to zero for bins that contain no observations but are at longer range than the minimal observed distance for that pair.

The statistical potentials were augmented with the Chemscore clash term and the Chemscore internal energy term.[11,32,33] The internal energy term was needed to ensure that only reasonable ligand geometries were considered during the docking. The clash term prevented protein–ligand clashes for atom pairs for which the statistical potential itself appeared to be too soft to provide the required repulsion at short range. It also prevented overlap between atom types for which there were not enough observations to calculate a potential at all. This led to the

TABLE I. Atom Types Used in Statistical Atom–Atom Potentials And Their Observed Frequencies

Type	P[a]	L[b]	Description
C.3	735515	35154	*sp*^{3} Carbon
C.3p	508407	35459	C.3 bonded to polar atom(s)
C.2	-	2818	*sp*^{2} Carbon
C.2p	459918	5697	C.2 bonded to polar atom(s)
C.oo	41935	5270	Carbon in carboxylate group
C.ar	275873	17780	Aromatic carbon, or *sp*^{2} carbon in ring
C.arp	77758	11986	C.ar bonded to polar atom(s)
C.cat	18602	354	Carbon atom in amidinium or guanidinium group
C.1	-	117	*sp* carbon atom
O.2	414745	7397	Carbonyl oxygen atom
O.2n	32877	374	Carbonyl oxygen atom in carbamoyl group
O.3	55034	16359	Hydroxyl oxygen
O.3a	19733	791	O.3 oxygen bonded to C.arp carbon
O.co2	80361	6368	Oxygen in carboxylate group
O.onco	-	219	Oxygen in hydroxamic acid group
O.e	-	6266	Ether oxygen
O.ex	-	3258	Ether oxygen bonded to non-carbon atom(s) (e.g. P-O-P)
O.n	-	461	Oxygen bonded to nitrogen
O.p	-	7889	Oxygen bonded to phosphorus
O.s	-	1202	Oxygen bonded to sulfur
O.crd	4267	-	Protein-only: oxygen coordinating a metal ion
N.ar	23662	5645	Aromatic nitrogen, or nitrogen in ring with C.ar carbons
N.pl3	9546	602	Nitrogen in indole or pyrrole ring
N.plc	55850	867	Nitrogen in amidinium or guanidinium group
N.o	-	119	Nitrogen bonded to oxygen atom
N.am	415404	4112	Amide (not in carbamoyl)
N.c2	33072	1715	Nitrogen bonded to C.2p, or in sulfonamide.
N.c3	16236	1468	Terminal nitrogen bonded to C.3p
N.1	-	80	Nitrogen bonded to exactly one non-hydrogen atom
N.2	-	631	Nitrogen bonded to exactly two non-hydrogen atoms
N.3	-	455	Nitrogen bonded to exactly three non-hydrogen atoms, non planar
N.4	-	281	Nitrogen bonded to exactly four non-hydrogen atoms
N.nda	18597	2301	Planar nitrogen, bonded to exactly three non-hydrogen atoms
N.crd	2489	-	Protein only: nitrogen coordinating a metal ion
S.3	13848	980	Sulfur bonded to exactly two non-hydrogen atoms
S.2	3789	356	Sulfur bonded to exactly one non-hydrogen atom
S.o2	-	428	Sulfur bonded to at least two oxygen atoms
S.crd	763	-	Protein only: sulfur atom coordinating a metal ion
P.3	-	3101	Any phosphorus atom
F	-	391	Fluorine atom
Cl	-	183	Chlorine atom
Br	-	53	Bromine atom
I	-	65	Iodine atom
Metal	2827	-	Metal ion

[a]Total number of protein atoms of this type found within 10 Å of any ligand atom.

[b]Total number of ligand atoms of this type in the database.

following functional form for the statistical potential that was used to drive the docking studies presented:

$$\text{Fitness} = -C_s\sum_p\sum_l \text{StatScore}(p,l,r_{pl}) - E_{\text{int}} - E_{\text{clash}} \quad (11)$$

where E_{int} is the Chemscore ligand internal energy, and E_{clash} is the Chemscore clash energy. The summation is over all combinations of protein atoms, p, and ligand atoms, l, within 6.0 Å, and r_{pl} is the distance between protein atom p and ligand atom l. To speed up scoring and docking, grids were pre-calculated for each atom type using a grid spacing of 0.3 Å. C_s is a scaling factor. After some optimization, we used $C_s = 0.2$ for the ASP potential. To determine the C_s weights for the PMF-scaled and the Drugscore-scaled potentials, we calculated the scores for the experimental binding modes of all complexes in the CCDC/Astex validation set. Linear regression of the statistical scores against the ASP scores resulted in scaling factors that brought these scores onto the same scale as the ASP score, resulting in $C_s = 0.4$ for the PMF-scaled and $C_s = 0.5$ for the Drugscore-scaled function.

Targeted Scoring Functions

We derived four targeted versions of ASP (tASP) for cdk2, based on four different target-specific databases. These four databases were constituted in the following way. The first (tASP-1) consisted of 25 structures from the PDB.[34] The second (tASP-2) added 18 in-house low molecular weight complexes to this database. These ligands all had a molecular weight under 160. The third database (tASP-3) consisted of the 25 PDB structures together with 55 protein–ligand complexes from our in-house database, and the fourth (tASP-4) consisted of the 25 PDB structures together with 116 protein–ligand complexes from our in-house database. These last two selections were just two snap-shots at arbitrary points in time of our in-house cdk2 structure database. The sets consisted of both fragments and more lead-like compounds of molecular weight up to 400. The linear weighting function between these targeted databases and the PDB was used with $N_{\max}$ set to 500 observations.

Docking Test Sets

We employed the CCDC/Astex validation set,[31] and used GOLD[4,5,33] to perform the docking, each time driven by the scoring function under consideration. We used the GOLD Default 4 settings for the genetic algorithm (GA). Hence, for each compound, GOLD performed 10 dockings, each consisting of 10,000 GA operations. GOLD terminates early when the top three dockings are within 1.5 Å of each other. This is a relatively fast search setting. Binding mode predictions can be more accurate if longer settings are used, but here we used a setting that is more suitable for virtual screening. Each docking was followed by a simplex optimization in which all flexible ligand torsions and the position and orientation of the ligand were refined to the nearest local optimum.

In addition to the CCDC/Astex validation set,[31] we investigated the success rate of docking for test sets of active compounds against neuraminidase, ptp1b, cdk2, and the estrogen receptor. The sets of known actives contained 15 X-ray structures for neuraminidase, 5 for ptp1b, 35 for cdk2, and 5 for the estrogen receptor. For cdk2 and the estrogen receptor, the actives were split into two sets to account for the difference in molecular size and the resulting changes in the protein binding site. For each set, one protein structure was used in the dockings for all compounds. To select which protein structure to use, dockings were performed against a variety of structures. The structures that were most promiscuous (i.e. against which most compounds were docked correctly) were used in the dockings. More details on the exact set-up of the binding sites used can be found elsewhere.[17]

Virtual Screening

We also compared the performance of the different statistical potentials in virtual screening experiments against these four targets. Here, sets of known actives were pooled together with a large number of random compounds; all compounds were docked, scored and ranked, and the ranks of the actives were converted into enrichment plots. We have shown that these 'random' compounds need to be chosen carefully, as the choice of 'random' libraries can strongly affect the enrichments obtained.[17] For example, if the known actives are typically much larger than the 'random' compounds, trivial enrichments can easily be obtained. To avoid such bias, all of our validations were run against *focused libraries*, i.e. libraries that contained random compounds with 1D properties, similar to those of the actives. The 1D properties we used here were:

(i) number of hydrogen-bond donors,
(ii) number of hydrogen-bond acceptors, and
(iii) number of non-polar atoms.

Inactive compounds were picked from our in-house database, ATLAS, which contains approximately 3.1 million compounds from high-throughput screening suppliers.

TABLE II. Overview of Six Validation Sets Used

		X-ray			
	N	PDB	Astex	SE[a]	Affinity range (M)
Neuraminidase	15	15		15	10^{-10}-10^{-3}
Ptp1b	25	5		25	10^{-7}->10^{-3}
Cdk2 MW < 250	41	1	17	41	10^{-5}->10^{-3}
Cdk2 MW > 250	23	11	6	23	10^{-8}-10^{-4}
ER agonists	20	3		3	10^{-10}-10^{-7}
ER antagonists	17	2		2	10^{-10}-10^{-7}

[a] Number of actives for which there is structural evidence, either from X-ray crystallography or from direct binding methods, that they bind in the binding site studied.

All virtual screens were performed using our web-based, in-house virtual screening platform.[35] SMILES[36] strings were prepared manually for all the known actives, resulting in six small active compound libraries. Table II lists the sources of the known actives in our test set; they originate from both literature and from our in-house compound collection. All of these active compound libraries and the (random) focussed libraries were virtually screened against their corresponding targets, using the following protocol:

(i) Compound SMILES strings were charged using a fixed set of rules.
(ii) Compound 3D input structures were generated from the SMILES strings using Corina.[37]
(iii) Compounds were docked against the target using GOLD and the Chemscore function; dockings were run on a Linux cluster.
(iv) Compounds were then re-scored with the various statistical potentials.

Each virtual screen was repeated five times to minimize the effect of the stochastic nature of the docking algorithm. Merged libraries were generated for all 25 combinations of the five virtual screening runs of the active compound library and the five runs for the corresponding focussed compound library; each merged library was ranked, and the 25 ranked lists were averaged. More details on the active compound libraries and the focused libraries can be found elsewhere.[17]

RESULTS AND DISCUSSION

Potentials

We constructed statistical potentials from all protein–ligand complexes in the PDB. Our analysis identified 9209 ligands in 5839 different PDB entries. Table I lists the occurrences of the different atom types in proteins and ligands in this database. As can be seen, there is huge variation in the number of occurrences of atom types, and as a result, some combinations of protein and ligand atom types will be more than adequately sampled, whereas other combinations will be poorly sampled. Clearly, potentials involving, for example, C.1, N.1, Br, or I atoms will not contain enough observations to construct reliable potentials for contacts with many protein atom types.

In the end, the potentials are based on the number of observations in the bins, but the total number of observations can be rather meaningless, as it is highly dependent on the radius used in collecting this total. When we collect all the observations within 6 Å, the highest number of observations is for C.3-C.3 with 196,447. A typical hydrogen-bond interaction such as N.plc (protein) to O.co2 (ligand) amounts to 11,936 observations. However, many combinations of atom types feature far fewer interactions. In total, 516 potentials (52%) have fewer than 500 observations within 6 Å, a situation not dissimilar to Drugscore, in which 60% of the potentials have been reported[14] to fail this criterion. Of these 516 potentials, 300 involve the least frequently occurring protein atom types Metal, N.crd, S.crd, O.crd, and S.2, or the least frequently occurring ligand atom types Cl, Br, I, N.o, N.1, N.4, and O.nco. We do not think this is a problem, as many of these interactions will hardly ever be required. And as long as these interactions are not the dominating factor for that protein–ligand complex, we will still be able to score the complex based on the remainder of the molecule.

To investigate the influence of the choice of reference state on the derived statistical potentials, we constructed statistical potentials with the reference states of Drugscore and PMF, in addition to the ASP potentials. These Drugscore-scaled and PMF-scaled potentials are not intended to be exact implementations of the published Drugscore and PMF methods. They are identical to ASP in aspects such as atom typing, grid spacing, and the database of structures from which they are derived, and hence will differ from the published methods. The only difference among the PMF-scaled, the Drugscore-scaled, and the ASP potentials is in the reference state used to convert the observed number of contacts into statistical potentials. This means we can relate any differences to the choice of reference state. Nevertheless, the derived potentials do appear to be very similar to the original reported potentials, especially in cases in which the atom typing allows for a direct comparison. In some other cases, differences among the potentials may occur due to differences in atom-type definitions and the database of structures compared to the original implementations of the methods.

In Figure 1, some of our Drugscore-scaled potentials are given. This figure shows the same selection of potentials as Figure 3 in the original Drugscore publication.[14] Our Drugscore-scaled potentials are very similar to the Drugscore potentials shown there. In Figure 2, some PMF-scaled potentials are given. These potentials can be compared to NCOC, OCNC, cFcF, and cPOC potentials depicted in the original PMF paper.[15] Again, our PMF-scaled potentials are very similar to the PMF potentials. For both methods, differences can be the result of the different selection of structures from which our potentials are derived. And for the PMF-scaled potentials there are further differences, such as in the ligand-volume correction. This correction serves to capture the same effect in the original PMF and our PMF-scaled potentials, but here the ligand-volume correction is derived by explicitly calculating available volumes around all ligand atoms on a grid; the original PMF method employs the counting of ligand atoms in course spherical shells around a central ligand atom.

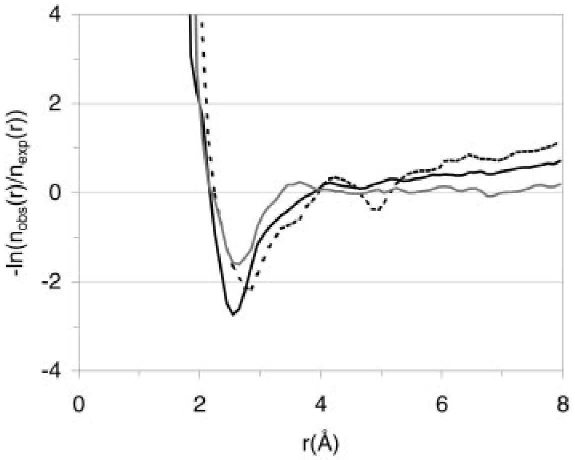

Fig. 1. Drugscore-scaled potentials for N.plc-O.co2 (dotted black line), O.co2-O.3 (solid black), and O.3-O.3 (solid gray line) atom pairs (first atom: protein; second atom: ligand).

To investigate the effect of the reference state, particularly the new protein volume correction present in ASP, we compared the ASP potentials to PMF-scaled potentials. The atom–atom potentials for the N.am/O.2 atom-type pairs are shown in Figure 3. The PMF-scaled potentials [Fig. 3(a)] feature the expected hydrogen-bond well, but both potentials rise up to +1 around 4 Å. Note that this shape of a minimum below 3 Å, followed by a maximum just below 4 Å is again very similar to the NDOA potential shown in the original PMF paper.[15] The corresponding ASP potentials [Fig. 3(b)] are much more flat at long range; they remain close to zero for distances down to around 4 Å; the small second 'well' in the potential, which is more pronounced in the potential where the N.am type is on the protein, can be attributed to secondary contacts to neighboring amides in the protein backbone. Another notable difference for the N.am/O.2 potentials is in their hydrogen-bond well depths. It seems reasonable to expect both amide-carbonyl potentials to be fairly similar, whether the carbonyl is on the protein and the amide is on the ligand or the other way around. However, the PMF-scaled N.am-O.2 potential is much more shallow than O.2-N.am one, whereas in ASP they are of similar depth.

The difference between the PMF-scaled and the ASP potentials is that the latter contains a correction for excluded volume effects on the protein side in its reference state. Indeed, inspection of the available volume graphs (Fig. 4) for the protein atom types involved shows that the anomalous effects in the PMF-scaled potentials can be explained by the reduced amount of available volume around these backbone atoms. For both the N.am and O.2 types, there is a decrease in available volume from 8 to 4 Å. Less available volume translates into fewer observed contacts at those distances. When this is not corrected for, like in the PMF approach, the shortage of observed contacts results in an unfavorable score, hence the rise

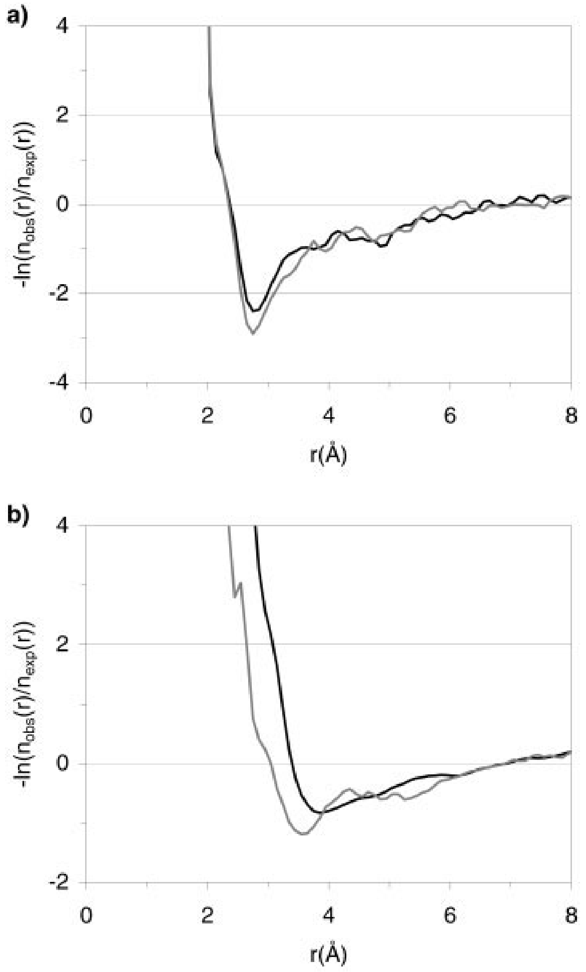

Fig. 2. (**A**) PMF-scaled potentials for N.c3-O.co2 (black line), and O.co2-N.c3 (gray line) atom pairs. (**B**) PMF-scaled potentials for C.ar-C.ar (black line), and C.arp-O.co2 (grey line) atom pairs (first atom: protein; second atom: ligand).

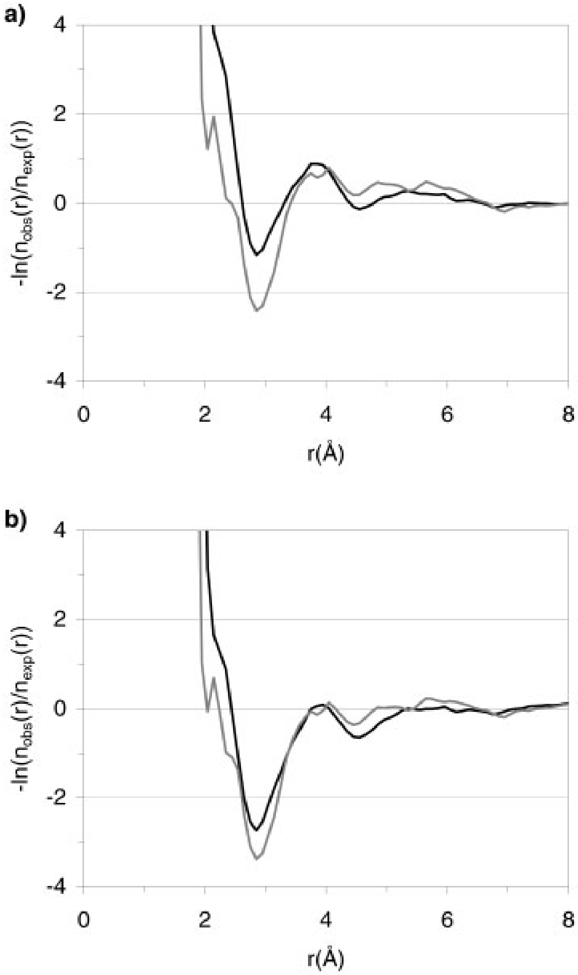

Fig. 3. Potentials for N.am-O.2 (black line), and O.2-N.am (gray line) atom pairs (first atom: protein; second atom: ligand). (**A**) PMF-scaled potentials; (**B**) ASP potentials.

above the baseline, resulting in the maximum just below 4 Å for the PMF-scaled potentials. At distances shorter than 4 Å, the available volume decreases more strongly for N.am than for O.2. Again, less available volume around N.am means fewer observed contacts, which in turn means a less favorable score, which explains that in the PMF-scaled potentials the N.am-O.2 potential is shallower than the O.2-N.am potential. When one corrects for these reduced available volumes, as is done in ASP, both potentials become of similar depth, and the repulsive region just below 4 Å disappears as well.

Both ASP and PMF-scaled potentials contain (the same) correction for the excluded volume around ligand atoms. Available volume graphs for some ligand atom types are given in Figure 5, where significant differences in available volume between atom types can be seen. Such differences again result in differences in the number of contacts that can be expected, which should be corrected for when deriving a statistical potential. Note that the Drugscore scaling, with its general volume correction for all atom pairs, does not correct for such differences between atom types. Also note that protein available volumes (Fig. 4) do not approach 1 for long distances, like ligand available volumes (Fig. 5) do. The height reached by an available volume graph for a protein atom type is an indication of the degree of 'buriedness' of the atom type.

The N.am/O.2 potentials illustrate the importance of ASP's protein volume correction for contacts to main-chain atoms. These protein atom types tend to be strongly shielded, and as a result statistical potentials involving these atoms can look rather strange when one does not correct for the strongly reduced available volumes around them. An example is shown in Figure 6, where ASP, Drugscore-scaled, and PMF-scaled potentials are given for the atom-type combination N.am–N.ar (protein backbone amide to an aromatic nitrogen in the ligand, an important interaction in many kinase inhibitors). All of the potentials feature a hydrogen-bond well, but in both the Drugscore-scaled and the PMF-scaled potentials a general 'repulsion' at short distances masks the hydrogen bond. In fact, the potentials are nearly repulsive at the hydrogen-bond distance. This can again be attributed to the reduced available volume at shorter range around the N.am atoms (see

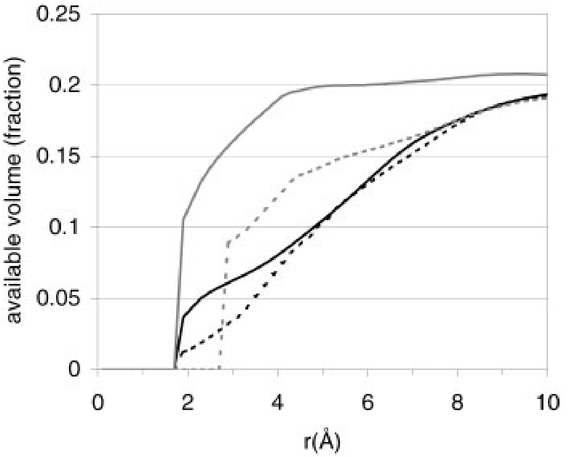

Fig. 4. Available volumes (fractions) for different atom types in proteins. N.am: backbone amide atom (dotted black line), N.c2: side-chain amide in asparagine or glutamine (solid gray line), C.3: sp^3 carbon (dotted gray line), O.2: backbone carbonyl oxygen (solid black line).

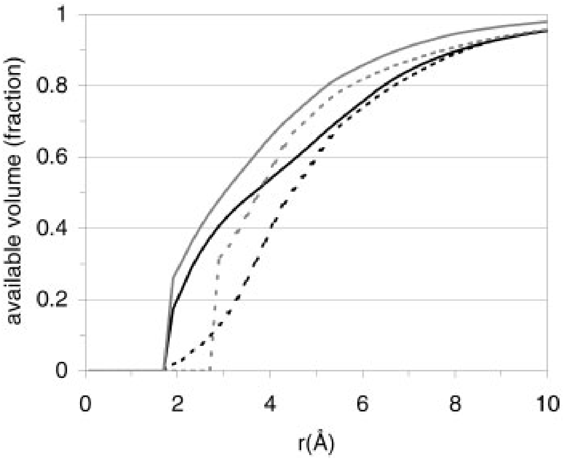

Fig. 5. Available volumes (fractions) for different atom types in ligands. N.am: amide atom (dotted black line), O.3: alcohol oxygen (solid gray line), C.3: sp^3 carbon (dotted gray line), O.2: carbonyl oxygen (solid black line).

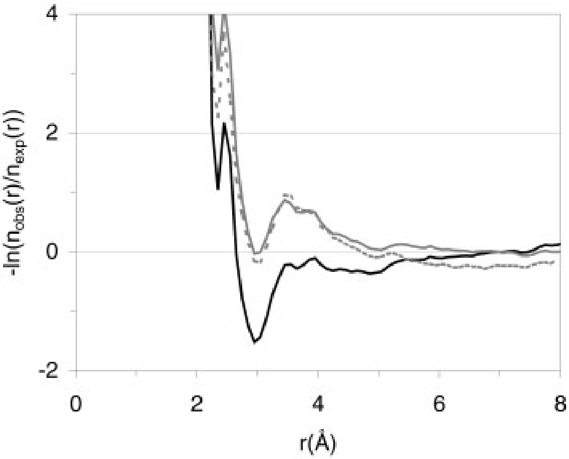

Fig. 6. Potentials for backbone amide to aromatic nitrogen (N.am-N.ar), using the three different methods: ASP (solid black line), Drugscore-scaled (dotted gray line), and PMF-scaled (solid gray line).

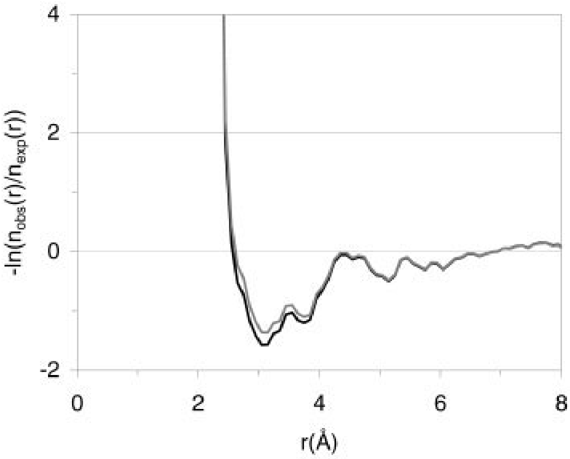

Fig. 7. Potentials for side chain NH_2 to aromatic nitrogen (N.c2-N.ar), using the three different methods: ASP (solid black line) and PMF-scaled (solid gray line).

Fig. 4), which leads to fewer observed contacts at these distances. When one corrects the expected number of contacts for the reduced available volumes, as is done in the ASP method, the underestimation of this hydrogen-bond potential is fully removed.

Such repulsive behavior in the PMF-scaled and the Drugscore-scaled potentials is seen in potentials of various ligand atom types with protein main-chain nitrogen or oxygen atoms. It also occurs, for example, for aromatic carbons in the ligand. This is in line with the observation by Stahl[16] of repulsive PMF potentials for phenyl rings close to protein amide bonds. In corresponding ASP potentials, such repulsive behavior is not observed, due to its protein available volume correction.

This protein volume correction is most important for the shielded atom types, such as backbone atoms. Many side-chain atom types are much less shielded. For exposed atom types, the volume correction on the protein side will be small, and as a result PMF-scaled potentials will look very similar to ASP potentials. An example of such a situation is given in Figure 7. Shown here is the N.c2 (protein) to N.ar (ligand) potential. On the protein, N.c2 occurs as the NH_2 in an asparagine or glutamine side chain. Inspection of the available volume for the N.c2 atom type (Fig. 4) shows a much-increased available volume around this atom compared to the backbone amide (N.am) type. The available volume is basically flat down to 4 Å, below which the available volume starts to decrease somewhat, but overall the available volume is much less restricted than for the N.am atom type. And indeed, in Figure 7 one can see that for N.c2, the small protein volume correction has little effect on the resulting potential. Note that the secondary minimum just below 4 Å is a result of N.ar ligand atoms that hydrogen bond to the carbonyl oxygen of the carbamoyl group instead of the NH_2 group. This situation arises from the fact that the N.ar atom type can be a hydrogen-bond donor as well as an acceptor.

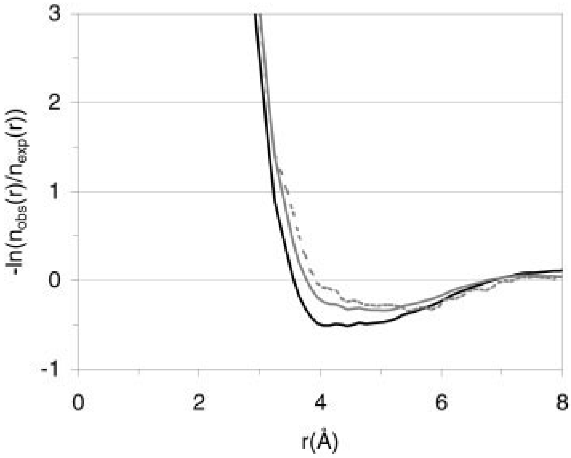

Fig. 8. Potentials for sp^3 carbon interactions (C.3-C.3), using the three different methods: ASP (solid black line), Drugscore-scaled (dotted gray line), and PMF-scaled (solid gray line).

In Figure 8, we compare ASP, PMF-scaled, and Drugscore-scaled potentials for the interaction between two sp^3 carbon atoms. The Drugscore-scaled potential has its minimum only just below 6 Å, whereas it occurs at closer range for the PMF-scaled potential. This difference is caused by Drugscore's reference state, which is really one of 'average interaction,' as explained in the theory section. The C.3-C.3 potential for Drugscore shows the extent to which a van der Waals-type interaction is stronger for C.3 atoms than for the average atom pair. Comparing the Drugscore-scaled and PMF-scaled potentials, this effect of averaging out seems to be small. The ASP potential, on the other hand, gives rise to a deeper potential, and the minimum-energy separation is significantly shorter. Overall, the ASP potential resembles more what one would expect for a van der Waals potential. Again, the difference between the PMF-scaled and the ASP potential is caused by the correction for the reduced available volume around C.3 protein atoms (Fig. 4).

Docking Performance

We tested the performance of the different statistical potentials in generating and recognizing the correct binding mode for a wide variety of protein–ligand complexes. Table III shows the docking success rates of the three statistical potentials against the CCDC/Astex validation set.[31] Comparing the results of the different statistical potentials, the ASP potential significantly outperforms the Drugscore-scaled and PMF-scaled potentials in terms of docking performance. The ASP potential performs very similarly to both Chemscore and Goldscore for the complete list of complexes in this validation set, as well as for the subsets of molecules that are classified as druglike or fragment-like.

In addition to the CCDC/Astex validation set,[31] we investigated the success rate of docking for test sets of active compounds against four pharmaceutical targets. The four targets considered were neuraminidase, ptp1b, cdk2, and the estrogen receptor. Table IV lists the success rates against these targets in terms of their ability to reproduce the binding modes of the actives for which X-ray structures are available. For comparison, the performances of Chemscore and Goldscore have been included in this table. Note that for some targets the numbers of X-ray structures that are available are very low, and as a consequence the uncertainties in the success rates are high.

When we compare ASP to both Chemscore and Goldscore for each target, ASP always performs better than the worst of Chemscore or Goldscore. Sometimes it performs as well as the best of these two, although it never achieves better success rates than the best of the two. The Drugscore-scaled and PMF-scaled potentials, on the other hand, perform satisfactorily for some targets, but very poorly for others.

In Figure 9, the result of a docking using the ASP potential is displayed for one of the cdk2 ligands (PDB code 1fvt). Also displayed are the ASP fields for three atom types. As can be seen, the ligand has been docked correctly into the active site, and the 'hot-spots' for the donor, acceptor and aromatic carbon atoms grossly correspond to the observed ligand atom types. The ASP fields show a preference for the stacking of aromatic rings, as the aromatic carbon hotspots are placed in the vicinity of phenylalanine residues. The sulfonamide end of the ligand is predicted to be slightly displaced from the experimental structure. This seems to be caused by the fact that this is a cross docking against a general cdk2 site, in which the hydrogen-bond acceptor for the sulfonamide NH_2 is in a slightly different position compared to the 1fvt protein structure.

Predicting Binding Affinities

For 60 complexes from the CCDC/Astex validation set, we have direct access to reliable affinity data.[33] For these complexes, we obtained the scores of the experimental binding modes using the different statistical potentials. Binding affinities generally correlate moderately with molecular size, and obviously one would hope that a scoring function outperforms such a simple descriptor as molecular weight; for this set of compounds the R^2 value for the correlation between binding affinity and heavy-atom count is 0.37. Drugscore-scaled (R^2 = 0.53) and ASP (R^2 = 0.51) scoring functions give very similar correlations to both Goldscore (R^2 = 0.55) and Chemscore (R^2 = 0.53). For the PMF-scaled scoring function, a much poorer correlation (R^2 = 0.33) is obtained.

In a previous study,[17] we tested Drugscore for structure-based virtual screening. In that study, we used an implementation of Drugscore that is closer to what is described by Gohlke et al.[14] It did not follow that work exactly, since we did not implement the SAS-dependent singlet potentials, and we based our potentials on a different selection of protein–ligand complexes. The key difference between that previous Drugscore implementation and the current Drugscore-scaled potentials lies in the atom typing. Here we use the atom typing as we used for ASP, whereas in the previous study we used the Sybyl atom types, as was done

TABLE III. Success Rates[a] for Binding Mode Predictions Against CCDC/Astex Validation Set[31] Using Statistical Potentials, and for Comparison *Goldscore* and *Chemscore* Values

	N	Goldscore[b]	Chemscore[b]	ASP	Drugscore-scaled	PMF-scaled
Clean list[c]	224	63.0 (1.8)	60.9 (1.8)	61 (1)	51 (1)	43 (3)
Druglike list[d]	139	73.4 (2.1)	74.8 (2.1)	72 (2)	63 (2)	52 (4)
Fragment-like list[e]	79	76.4 (2.8)	78.6 (3.9)	77 (3)	68 (2)	57 (4)

[a] Percentage of complexes for which the top-ranked GOLD solution is within 2.0 Å RMSD of the experimental binding mode. All values are averages over five runs. Standard deviations are given in parentheses.

[b] Averages over fifty runs.[33]

[c] The 'clean list' is a subset of the CCDC/Astex validation set for which complexes do not exhibit protein–ligand clashes, crystallographic contacts, or unlikely ligand geometries; for closely related complexes, only one representative was kept.[31]

[d] A subset of the 'clean list' for which ligands have 10 or fewer rotatable bonds and a polar surface area equal to or less than 140 $Å^2$[31]

[e] A subset of the 'clean list' for which ligands are not covalently bound to the protein, have five or fewer rotatable bonds and between 7 and 20 non-hydrogen atoms.[31]

TABLE IV. Success Rates[a] for Binding Mode Predictions Against Six Virtual Screening Validation Sets, Using Statistical Potentials, as well as for Goldscore and Chemscore

	N	Goldscore	Chemscore	ASP	Drugscore-scaled	PMF-scaled
Neuraminidase	15	96 (4)	63 (6)	79 (5)	59 (7)	40 (15)
Ptp1b	5	76 (9)	56 (30)	76 (8)	84 (16)	64 (9)
Cdk2 MW $<$ 250	18	22 (7)	48 (6)	32 (4)	10 (5)	18 (2)
Cdk2 MW $>$ 250	17	33 (3)	65 (4)	56 (5)	40 (8)	14 (8)
ER agonists	3	53 (30)	80 (18)	60 (13)	33 (23)	40 (15)
ER antagonists	2	90 (22)	60 (22)	90 (20)	90 (20)	100 (0)

[a] Percentage of complexes for which the top-ranked GOLD solution is within 2.0 Å RMSD of the experimental binding mode. All values are averages over the five runs used in the virtual screening validation experiments (Fig. 10). Standard deviations are given in parentheses.

by Gohlke. It is interesting to note that this appears to have a marked effect on the correlation with the binding affinities. Using the Drugscore implementation closer to the literature version, a much poorer correlation (R^2 = 0.29) is observed than for the current Drugscore-scaled version.

Virtual Screening Performance

We compared the performance of the different statistical potentials in virtual screening experiments against the four pharmaceutical targets also used in the docking experiments. Enrichment plots for these virtual screening test cases are shown in Figure 10. In contrast to the docking experiments, ASP, Drugscore-scaled, and PMF-scaled potentials all perform similarly in these virtual screens. A notable exception is the estrogen receptor. For the agonists, the Drugscore-scaled function gives by far the best enrichments, but for the antagonists the ASP method produces the best results. Considering both agonists and antagonists together, ASP produces the more consistent results. For the other targets, results for the different scaling methods are much more similar. The biggest remaining differences are observed for ptp1b, for which the ASP results are the best. For all other targets, the differences are only a few percent, which is not significant.

As explained in the previous section, we used a different implementation of Drugscore, closer to the literature version, in a previous virtual screening study.[17] Results reported there were in line with those reported by Stahl.[16] Again, it is interesting to note that considerably different results are obtained with the current Drugscore-scaled potentials. The enrichment for neuraminidase is significantly higher. For the estrogen receptor, the previously reported results were better.[17] As was the case for the binding affinities, we can only attribute these differences to the different atom typing used in the two versions.

In the abovementioned study[17] we investigated the use of Chemscore and Goldscore in virtual screening. That work used the same test sets as we used in the present work, so we can also compare the statistical potentials to those scoring functions. Since such a comparison is not the focus of this article, we will not go into too much detail. In short, ASP performs similarly to Goldscore. Like Goldscore, it performs significantly better than Chemscore for neuraminidase, but for cdk2 and the estrogen receptor Chemscore is better. All in all, ASP compares reasonably well to Goldscore and Chemscore but it never performs significantly better than the best of these two scoring functions.

Targeted Scoring Functions

In our approach to targeted scoring, the size and contents of the target-specific database will influence whether the targeted statistical potentials will lead to improved results in docking and virtual screening. To investigate this effect, we used four different target specific databases, all for the same target (cdk2). The first consists of only the target-specific data available from the PDB. These are the kind of data that would be available from the start for a drug-discovery project. The second database adds a number of small fragments to this database. Since Astex uses

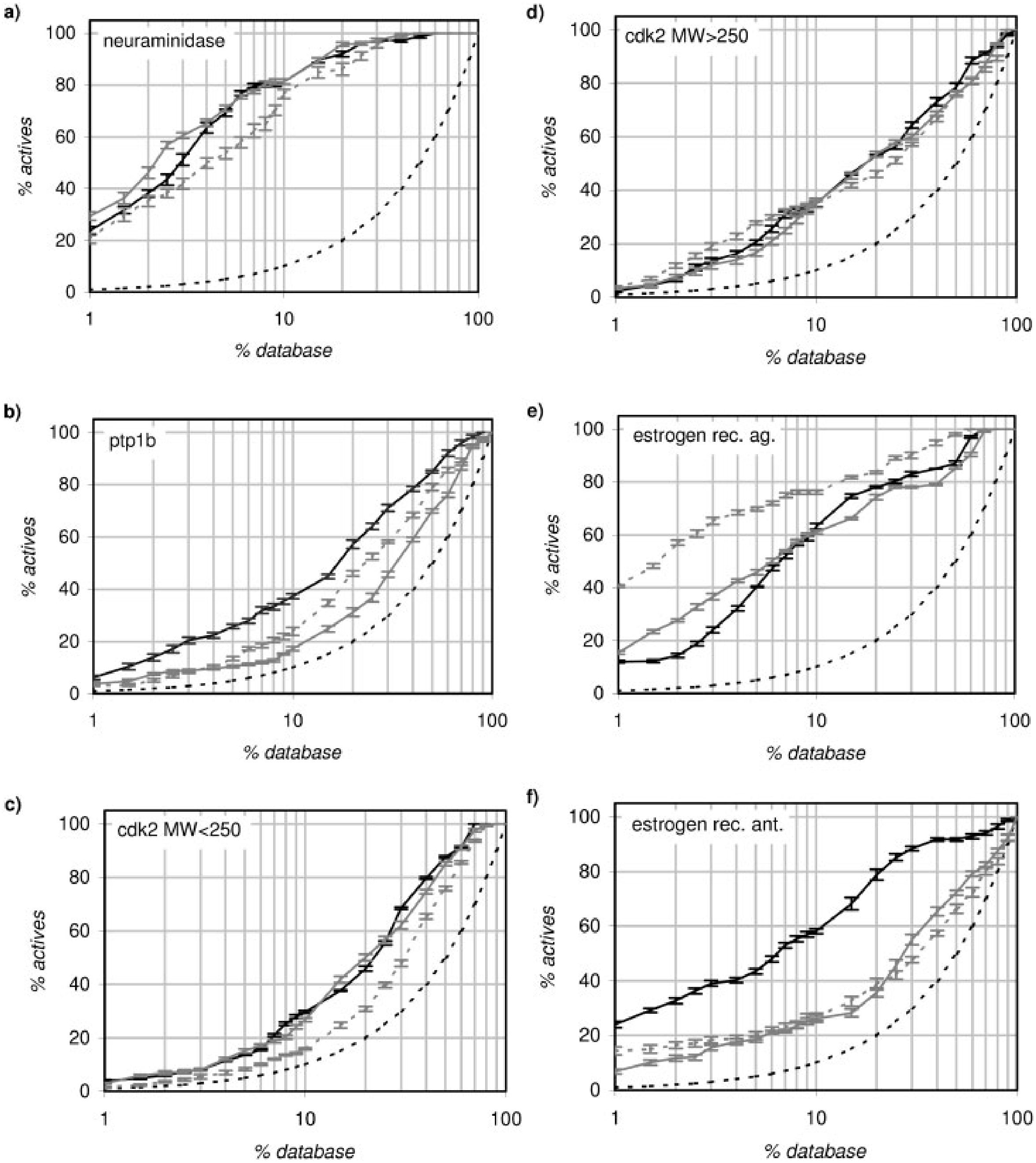

Fig. 10. Enrichment graphs for the six virtual screening validation sets studied. All results are averages over the 25 combinations of the five runs for the actives and the five runs for the focused library. The error bars represent the errors in the mean. The Chemscore function was used to drive the dockings. Results are shown for ranking with ASP (solid black line), PMF-scaled (solid gray line) and Drugscore-scaled (dotted gray line) potentials. The dotted black line represents the fraction of actives expected at random.

X-ray crystallography to screen fragments, these are data that would generally become available during the early stages of a project. The final two target-specific databases resemble the later stages of a project, when complexes varying from fragment hits to lead-like molecules continue to be added to our in-house structure database.

Table V lists the success rates for the different targeted-ASP functions in terms of their ability to reproduce the binding modes of the actives in both of the cdk2 test sets for which X-ray structures are available. For both test sets, the success rate gradually improves upon adding more and more structures to the targeted database. For the high-molecular weight set, the success rate seems to level out at 70% for both tASP-3 and tASP-4 (55 and 116 structures respectively).

Enrichment plots for cdk2 using the targeted scoring functions are depicted in Figure 11. For clarity, we only show the more important top 10% of the database in the

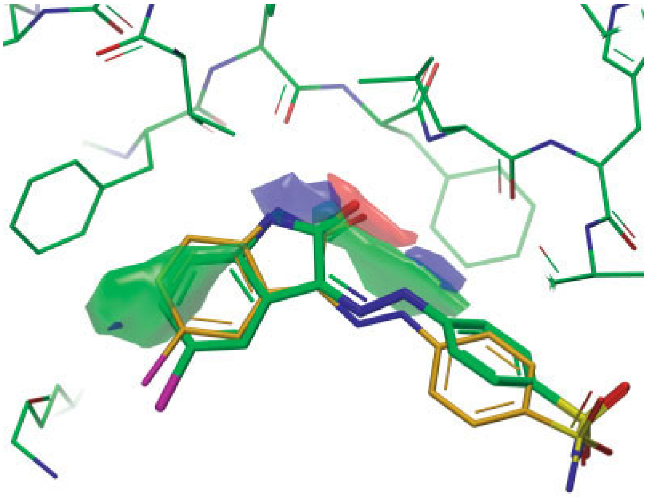

Fig. 9. Predicted binding mode for the cdk2 ligand from PDB entry 1fvt, docked against the 1di8 binding site. This is the top-ranked solution produced by GOLD, using the ASP potential. In orange is the overlaid experimental ligand structure (root mean squared deviation: 0.85 Å). In green is displayed the ASP field for the C.ar atom type, contoured at −9.5, in red the O.2 field (−14.0) and in blue the N.c2 field (−15.0); (Picture created with AstexViewer™.[38])

graph. For the low molecular weight set, modest improvements are obtained for each increase in size of the targeted database. Overall, the number of known actives retrieved from the top 2% of the database increases from 6% for ASP to 17% for tASP-4. For the high molecular weight compounds, bigger increases in enrichment are obtained. Here, the number of known actives retrieved from the top 2% of the database increases from 6% (ASP) to 25% for tASP-1, i.e. just using the cdk2 data from the PDB. Adding just a small number of fragments does not improve on this (23% for tASP-2). However, adding more structures results in significant improvements (37% for tASP-3 and 40% for tASP-4). For both the high and the low molecular weight compounds, the improvements in enrichment seem to level out for tASP-3. Increasing the size from 80 structures (tASP-3) to the 141 structures in tASP-4 has limited effect.

In comparing the statistical potentials to Chemscore, it should be noted that cdk2 had already proven to be a very difficult target for both Goldscore and Chemscore. This prompted the addition of a specific CH · · O hydrogen bond term to Chemscore, which was used for all cdk2 Chemscore dockings in this study. In a previous study,[17] this CH · · O term was seen to result in improved enrichments, as well as docking success rates, compared to the standard Chemscore function. So, one could argue that this version of Chemscore is already somewhat targeted towards cdk2.

Of course, it would have been disappointing if the targeted scoring function had not shown improved results. After all, it is based on detailed information of the binding modes of molecules to this specific target. Nevertheless, these results indicate that the essential information on binding preferences for a specific target does not get lost upon transforming the (3D) binding modes into 1D potentials. An alternative would be to use the experimental binding modes to construct a pharmacophore model and to apply that in docking and virtual screening. For the present example of cdk2, inhibitors invariably form a hydrogen bond with the backbone NH of Leu83. Applying this as a pharmacophore is much more restrictive than the approach presented here: a hydrogen bond to a backbone NH might be more favorable in the targeted scoring function than in general ASP, but it is not enforced and not restricted to any specific amide.

TABLE V. Success Rates[a] for Binding Mode Predictions Against cdk2, Using Targeted ASP Potentials, as well as for General ASP

	N	ASP	tASP-1	tASP-2	tASP-3	tASP-4
Cdk2 MW < 250	18	32 (4)	40 (9)	49 (5)	58 (5)	63 (3)
Cdk2 MW > 250	17	56 (5)	64 (3)	66 (5)	70 (4)	69 (5)

[a] Percentage of complexes for which the top-ranked GOLD solution is within 2.0 Å RMSD of the experimental binding mode. All values are averages over the five runs used in the virtual screening validation experiments (Fig. 4). Standard deviations are given in parentheses.

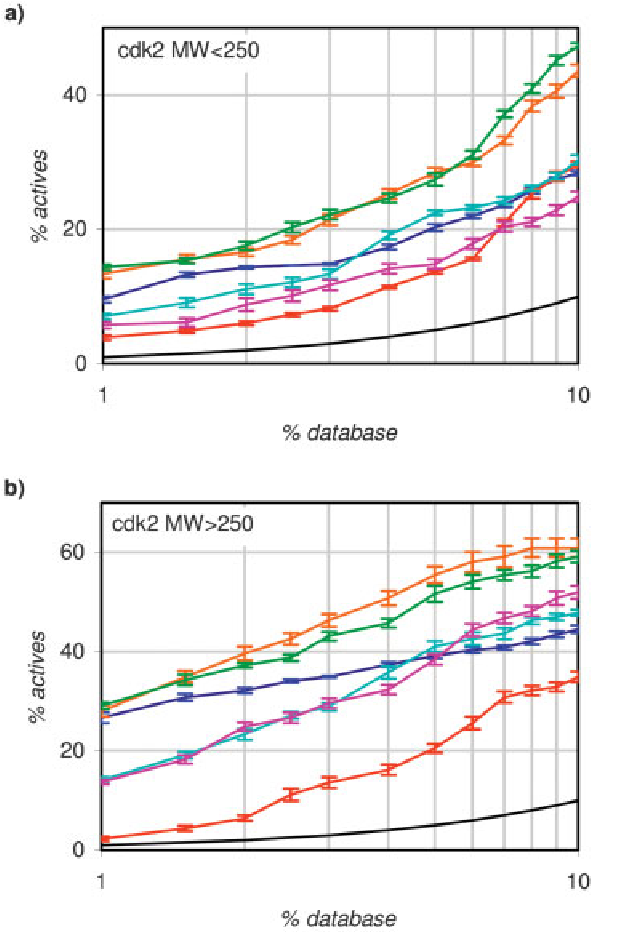

Fig. 11. Enrichment graphs for the two cdk2 virtual screening validation sets. All results are averages over the 25 combinations of the five runs for the actives and the five runs for the focused library. The error bars represent the errors in the mean. The Chemscore function was used to drive the dockings. Results are shown for ranking with ASP (red), tASP-1 (pink), tASP-2 (light-blue), tASP-3 (green), tASP-4 (orange), and Chemscore (blue). The black line represents the fraction of actives expected at random.

A targeted scoring function is much fuzzier than a pharmacophore, and can just as easily be applied in cases where a conserved motif among the known binders is less obvious. Moreover, the method provides a smooth transition from general to target-specific data, as shown by the different targeted-ASP functions that we present. Newly available structural knowledge of known-binders can be fed back into the scoring functions to continuously improve docking and virtual screening results.

CONCLUSIONS

We have described ASP, a statistical potential derived from a database of protein–ligand structures using a novel reference state. We have compared our new potential to otherwise identical potentials that use reference states along the lines of PMF and Drugscore. Inspection of graphs of individual atom–atom potentials has highlighted shortcomings in the potentials that were derived using these literature reference states. These shortcomings are caused by a factor unaccounted for in these two reference states. Neither considers differences in exposure of protein atom types towards ligand binding sites. In ASP, we introduce a volume correction for protein atoms, in addition to one for ligand atoms, to correct for such differences. The new protein volume correction was shown to be especially important for interactions to backbone atom types.

We have shown that the potentials using the ASP reference state give better success rates in docking than when the PMF or Drugscore reference states are used. This is the case both for the CCDC/Astex validation set as well as for a test set of actives against four pharmaceutically relevant targets. In terms of docking and the prediction of binding affinities, ASP achieves results that are roughly similar to those obtained using Goldscore or Chemscore. Enrichment plots for virtual screens against the four test targets showed less discrimination between the different reference states.

In addition, we have described how statistical potentials can be used in the construction of targeted scoring functions. Our approach uses ASP potentials that are biased towards a target-specific database of protein–ligand complexes. As a test case, cdk2 is presented. Starting from the general ASP function, improvements in docking success and virtual screening enrichment rates were seen to be function of the size of the target-specific database: the more protein–ligand complexes of the target under consideration available, the better the result for the derived targeted potential. Clearly, the targeted statistical potential approach cannot be used when only a handful of structures is available, but we have shown that with just publicly available PDB data, the ASP function can already be improved significantly. Also, improvements were seen to level out after adding around 50 in-house structures. In a drug-discovery context, this is a number of structures that can become quickly available within a modern crystallographic screening set-up.

The general, non-targeted ASP method behaves similarly, but is not superior to, well-established scoring functions like Chemscore and Goldscore, which is a satisfactory result in itself for a statistical potential. Moreover, for those scoring functions, it is not clear how increased (structural) knowledge about known binders can be used to tailor the functions to a specific target under consideration. In contrast, for the ASP potential, our targeted ASP approach provides a well-defined approach to gradually move from a general scoring function towards a targeted function as more and more protein–ligand structures become available for the target.

ACKNOWLEDGMENTS

We thank Dr. C.W. Murray for comments and discussions and for carefully reading this manuscript. We also thank Drs. P. Watson and M.J. Hartshorn for their ligand bond-typing program, and Dr. M. O'Reilly for crystallography on our in-house cdk2 compounds.

REFERENCES

1. Makino S, Kuntz ID. Automated flexible ligand docking method and its application for database search. J Comput Chem 1997;18: 1812–1825.
2. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261:470–489.
3. Goodsell DS, Olson AJ. Automated docking of substrates to proteins by simulated annealing. Proteins 1990;8:195–202.
4. Jones G, Willett P, Glen RC. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J Mol Biol 1995;245:43–53.
5. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748.
6. Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. J Comput Aided Mol Des 2002;16:151–166.
7. Gohlke H, Klebe G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed 2002;41:2644–2676.
8. Cornell, WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. J Am Chem Soc 1995;117:5179–5197.
9. Jorgensen WL, Tiradorives J. The OPLS potential functions for proteins - Energy minimizations for crystals of cyclic-peptides and crambin. J Am Chem Soc 1988;110:1657–1666.
10. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM-A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4: 187–217.
11. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions .1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aid Mol Des 1997;11:425–445.
12. Bohm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein ligand complex of known 3-dimensional structure. J Comput Aid Mol Des 1994;8:243–256.
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–242.
14. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 2000;295:337–356.
15. Muegge I, Martin YC. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. J Med Chem 1999;42:791–804.
16. Stahl M, Rarey M. Detailed analysis of scoring functions for virtual screening. J Med Chem 2001;44:1035–1042.
17. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WTM, Murray CW, Taylor RD, Watson P. Virtual screening using protein-ligand

docking: avoiding artificial enrichment. J Chem Inf Comput Sci 1004;44:793–806.

18. Sotriffer CA, Gohlke H, Klebe G. Docking into knowledge-based potential fields: a comparative evaluation of DrugScore. J Med Chem 2002;45:1967–1970.
19. Muegge I, Martin YC, Hajduk PJ, Fesik SW. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. J Med Chem 1999;42:2498–2503.
20. Wallqvist A, Jernigan RL, Covell DG. A preference-based free-energy parameterization of enzyme-inhibitor binding. Application to HIV-1-protease inhibitor design. Protein Sci 1995;4:1881–1903.
21. Wallqvist A, Covell DG. Docking enzyme-inhibitor complexes using a preference-based free-energy surface. Proteins 1995;25: 403–419.
22. Verkhivker GM, Appelt K, Freer ST, Villafranca JE. Empirical free-energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. Protein Eng 1995;8:677–691.
23. Dewitte RS, Shakhnovich EI. SMoG: de Novo design method based on simple, fast, and accurate free energy estimates: 1. Methodology and supporting evidence. J Am Chem Soc 1996;118: 11733–11744.
24. Dewitte RS, Ishchenko AV, Shakhnovich EI. SMoG: De novo design method based on simple, fast, and accurate free energy estimates. 2. Case studies in molecular design. J Am Chem Soc 1997;119:4608–4617.
25. Mitchell JBO, Laskowski RA, Alex A, Thornton JM. Bleep - Potential of mean force describing protein-ligand interactions: I. Generating potential. J Comput Chem 1999;20:1165–1176.
26. Mitchell JBO, Laskowski RA, Alex A, Forster MJ, Thornton JM. Bleep - Potential of mean force describing protein-ligand interactions. II. Calculation of binding energies and comparison with experimental data. J Comput Chem 1999;20:1177–1185.
27. Muegge I. Effect of ligand volume correction on PMF scoring. J Comput Chem 2001;22:418–425.
28. Astley T, Birch GG, Drew MGB, Rodger PM, Wilden RH. Effect of available volumes on radial distribution functions. J Comput Chem 1998;19:363–367.
29. Hendlich M, Rippmann F, Barnickel G. Bali: Automatic assignment of bond and atom types for protein ligands in the Brookhaven protein data bank. J Chem Inf Comput Sci 1997;37:774–778.
30. Sayle R. PDB: Cruft to content (perception of molecular connectivity from 3D coordinates). 2001. Daylight user meeting MUG01 (http://www.daylight.com/meetings/mug01/Sayle/m4xbondage.html).
31. Nissink JW, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R. A new test set for validating predictions of protein-ligand interaction. Proteins 2002;49:457–471.
32. Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD. Flexible docking using Tabu search and an empirical estimate of binding affinity. Proteins 1998;33:367–382.
33. Verdonk ML, Cole JC, Hartshorn M, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins 2003;52: 609–623.
34. PDB refcodes: 1fvv,1qmz,1jsv,1jst,1hck,1h1s,1fq1,1fin,1e9h,1e1x, 1e1v,1b39,1b38,1ke9,1ke8,1ke7,1ke6,1ke5,1jvp,1g5s,1fvt,1dm2, 1di8,1ckp,1aq1.
35. Watson P, Verdonk ML, Hartshorn MJ. A web-based platform for virtual screening. J Mol Graph Model 2003;22:71–82.
36. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 1998;28:31–36.
37. Gasteiger J, Rudolph C, Sadowski J. Automatic generation of 3D-atomic coordinates for organic molecules. Tetrahedron Comput Methodol 1990;3:537–547.
38. Hartshorn MJ. AstexViewer: a visualisation aid for structure-based drug design. J Comput Aid Mol Des 2002;16:871–881.