# Information Theory-Based Scoring Function for the Structure-Based Prediction of Protein−Ligand Binding Affinity

Mahesh Kulharia,[†,‡] Roger S. Goody,[†] and Richard M. Jackson*[,‡]

Department of Physical Biochemistry, Max Planck Institute of Molecular Physiology, Otto Hahn Strasse 11, Dortmund, Germany 44227, and Institute of Molecular and Cellular Biology, University of Leeds, Leeds, U.K. LS2 9JT

Received April 10, 2008

The development and validation of a new knowledge based scoring function (SIScoreJE) to predict binding energy between proteins and ligands is presented. SIScoreJE efficiently predicts the binding energy between a small molecule and its protein receptor. Protein−ligand atomic contact information was derived from a Non-Redundant Data set (NRD) of over 3000 X-ray crystal structures of protein−ligand complexes. This information was classified for individual "atom contact pairs" (ACP) which is used to calculate the atomic contact preferences. In addition to the two schemes generated in this study we have assessed a number of other common atom-type classification schemes. The preferences were calculated using an information theoretic relationship of joint entropy. Among 18 different atom-type classification schemes "ScoreJE Atom Type set2" (SATs2) was found to be the most suitable for our approach. To test the sensitivity of the method to the inclusion of solvent, Single-body Solvation Potentials (SSP) were also derived from the atomic contacts between the protein atom types and water molecules modeled using AQUARIUS2. Validation was carried out using an evaluation data set of 100 protein−ligand complexes with known binding energies to test the ability of the scoring functions to reproduce known binding affinities. In summary, it was found that a combined SSP/ScoreJE (SIScoreJE) performed significantly better than ScoreJE alone, and SIScoreJE and ScoreJE performed better than GOLD::GoldScore, GOLD::ChemScore, and XScore.

## INTRODUCTION

The success of *in silico* approaches for structure-based drug design depend on the timely application of the principles governing the dynamics of ligand−protein interactions.[1] Current docking approaches involve generating favorable ligand orientations in the protein binding site, by sampling conformational space, followed by scoring these by their predicted interaction energy.[2] The limitation in the scoring step stems from the time needed to score each solution and the level of accuracy required for the calculation of the interaction energy or, at the very least, the correct discrimination of active from inactive compounds. A number of simplified scoring functions have been developed which are fast and easy to apply but provide only moderate levels of accuracy, hence there is still much room for improvement.

Current scoring functions used to estimate ligand−protein affinity can be classified into three categories: first-principle methods, knowledge-based methods, and, finally, regression-based scoring functions.[3] Knowledge-based scoring functions are derived from the quantification of frequencies of interacting atomic pairs observed in protein−ligand complexes.[4] The process of atomic-pair-interaction-frequency quantification has been based on a number of mathematical relationships. The earliest example of such a function was in the field of protein folding where Boltzmann's law was used to derive the potential of mean force for interacting residues.[5,6] Later, similar functions were developed for scoring ligand−protein interactions. Wallqvist et al.[7] studied a data set of 38 complexes, calculating the frequencies of atomic interactions at the protein−protein interface and converted these into an atom-atom preference score, using the ratio of the fraction of the total interface area contributed by each pair to the product of the fraction of their respective contributions to the protein surface. Verkhivker et al.[8] developed a related approach using a set of 30 protease-inhibitor complexes and the inverse Boltzmann law to develop distance-dependent pair potentials from interacting atoms in combination with terms to model conformational entropy[9] and the hydrophobic effect.[10] Using this scoring function they could estimate the affinity of HIV-1 proteases for several different inhibitors. Subsequent potentials utilized yet more structural information. SMoG-Score was developed from 109 crystal structures using statistical mechanics.[11] Potentials of mean force were derived by Muegge et al. using the inverse Boltzmann law for a data set of 697 protein−ligand complexes.[12] Mitchell et al. developed BLEEP using a data set of 820 protein−ligand complexes with hydrogen atoms added (using HBPlus[13]) and used the inverse Boltzmann law.[14] A semiempirical pair-potential for Ne−Ne was used as a reference state. They further derived BLEEP-II by including interactions of protein and ligand with water molecules (explicitly added using AQUARIUS2[15]). Gohlke et al.[16] derived DrugScore using distance-dependent pair-potentials from a data set of 6026 protein−ligand complexes and incorporated solvent accessible surface area based solvation potentials from a database

* Corresponding author phone: +44(0)113 343 2592; fax: +44(0)113 343 3167; e-mail: r.m.jackson@leeds.ac.uk.
[†] Max Planck Institute of Molecular Physiology.
[‡] University of Leeds.

10.1021/ci800125k CCC: $40.75
Published on Web 09/04/2008

of 1376 protein−ligand complexes. Cline et al.[17] used an information theoretic relationship of mutual information to quantify information in amino-acid contact potentials for protein structure prediction. They studied the contribution of amino-acid character in terms of hydropathy, charge, disulfide bonding, and residue burial to the mutual information.

The Boltzmann law is very useful for determining the interaction energy values from a database using the observed frequencies of joint occurrences. It has been commented that the variation in temperature factors for the protein−ligand atoms[18] give rise to heterogeneity in the interaction database which complicates the application of the inverse Boltzmann law. However, even though knowledge-based methods are susceptible to the artifacts in data collection they have performed surprisingly well, in some cases better than force-field based scoring functions.[19,20]

Here we present a novel knowledge-based scoring function: ScoreJE−derived from ligand−protein atomic contacts. Our approach differs from the previous scoring functions in two important aspects. First, it uses over 3000 structurally nonredundant protein−ligand complexes which is more than used in the development of the previous knowledge based scoring functions, the only exception being DrugScore, which uses a 30% sequence identity cutoff for the creation of the nonredundant data set of complexes. Second in using the mathematical relationship of joint entropy for deriving the atomic contact preferences it bypasses the problems implicit in the application of the inverse Boltzmann law by eliminating the need for a reference state. These preferences are derived for describing the energetics of short-range atomic interactions. A Single-body Solvation Potential (SSP) is developed using the joint entropy of protein−water atom contact probabilities and is combined with ScoreJE to obtain SIScoreJE (SSP included ScoreJE). These functions were tested for their ability to predict the binding energies of test data sets containing 100 protein−ligand complexes.

The overall aim was to develop a novel knowledge-based scoring function for predicting protein−ligand interaction energies. The main objective was to calculate a set of atomic contact preferences for the protein−ligand and protein−water interactions. A secondary aim was to evaluate the potential of using information theory and a new atom type classification scheme (alongside popular atom-type classification schemes currently in use) to optimally describe protein−ligand interactions.

## METHODS

**Construction of an Atom Pair Contact Database.** Nearly 30,000 protein−ligand complexes present in PDBSUM[21−23] with structural information were extracted from the PDB.[24] From these, only protein−ligand complexes having experimentally determined X-ray crystal structures with a resolution better than 2.5 Å were retained. The average resolution of the protein−ligand complexes thus retained was 1.95 Å. In addition complexes having either a covalently bound ligand or involving cosolvents or metallic ions or not having a classification in SCOP[25] (version 1.63) were removed. A ligand was classified as noncovalently bound to the protein if none of its atoms were within the covalent bond interaction distance. The covalent bond interaction distance for a specific pair of protein and ligand atom was the sum of their atomic radii (see Table 1, Supporting Information) plus a 10% tolerance limit.

**Table 1.** Tabulation of Different Atom-Type Classification Schemes and Their Resultant Alphabet Sizes

| classification type | size of classification | classification type | size of classification |
|---|---|---|---|
| AMBER[a] | 56 | MENG[a] | 52 |
| Broto, P. et al.[a] | 245 | MM2[a] | 64 |
| BLEEP[a] | 32 | MM3[a] | 120 |
| CFF91[a] | 87 | MM+[a] | 40 |
| CHARMM[a] | 79 | MMFF[a] | 58 |
| Crippen et al.[a] | 82 | SATs1[b] | 16 |
| CVFF[a] | 66 | SATs2[b] | 24 |
| GRID[a] | 62 | TRIPOS[a] | 32 |
| H-bond (Vega template)[a] | 9 | Universal[a] | 47 |

[a] For these atom-types Atom type Descriptive Language (ATDL) templates were used as provided in the Vega Software. [b] These atom type classification systems were designed in Atom type Descriptive language.

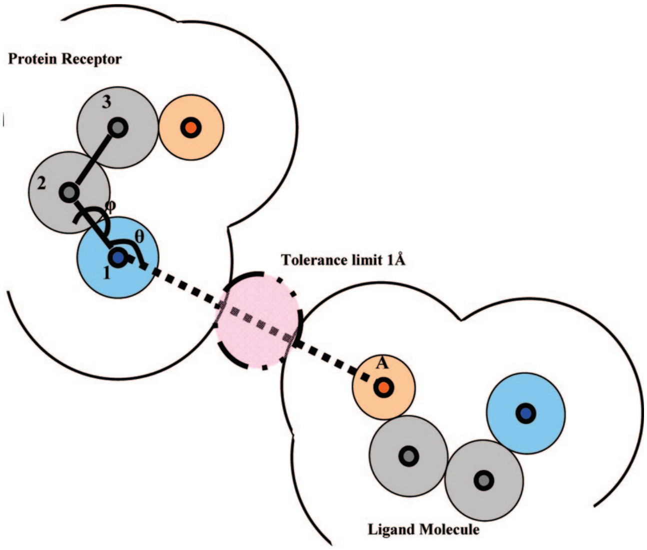

**Figure 1.** Ligand atom A is in contact with protein atom 1. Protein atom 1 is bonded with non-hydrogen protein atoms 2 and 3. The quantified interactions are as follows: interatomic A1 distance, planar angle (<A12), and dihedral (A12−123).

Only the best resolution complex with a unique ligand name and a unique SCOP superfamily was retained. These comprised a nonredundant data set (NRD) with only one ligand representative for each SCOP superfamily. The data set was not further processed (i.e., to retain only druglike molecules), as the intention was to create a scoring function that is applicable to small molecule-protein interactions in general, including cognate ligands such as biological substrates and carbohydrates. Despite this 93% of the ligands in the data set have >5 and <50 heavy atoms (see Figure 1, Supporting Information). Therefore the data set has a similar size composition to those of typical druglike data sets where these size limits have been used as a filter.[16] Two atoms belonging to two different molecules were considered to be in contact if the intervening distance between the atoms was less than the sum of their van der Waals radii plus 1 Å tolerance limit.[26−28] This is an atomic contact method, most similar in its definition to SMoG since an interaction is either present or absent based on the cutoff. Hydrogen atoms were added to ligand and protein using the molcharge utility of

**Table 2.** Increase in Interacting Atom-Atom Combinations[a]

| atom type sets | Combination1 | Combination2 | Combination3 | Combination4 |
|---|---|---|---|---|
| AMBER | 1867 | 16463 | 15460 | 142434 |
| BLEEP | 344 | 5103 | 4274 | 106358 |
| BROTO | 3323 | 20560 | 19841 | 131204 |
| CFF91 | 2785 | 21683 | 20450 | 156505 |
| CHARMM | 1876 | 15332 | 13978 | 142591 |
| CRIPPEN | 3665 | 28626 | 28238 | 179601 |
| CVFF | 2196 | 18391 | 17190 | 148097 |
| GRID | 1862 | 15060 | 14575 | 132136 |
| HBOND | 51 | 1028 | 691 | 46294 |
| MENG | 918 | 8617 | 7523 | 131492 |
| MM2 | 1213 | 11093 | 9932 | 126356 |
| MM3 | 1599 | 13086 | 11884 | 125786 |
| MMFF | 1410 | 13395 | 12440 | 136650 |
| MM PLUS | 1040 | 10012 | 8614 | 129607 |
| SATs1 | 138 | 2697 | 2053 | 101157 |
| SATs2 | 233 | 4568 | 3573 | 129565 |
| TRIPOS | 380 | 5300 | 4341 | 113238 |
| UNIV | 404 | 5656 | 4758 | 108785 |

[a] Combination1 - when only atomic character is considered, Combination2 - when interatomic A-1 distance and atomic character are considered, Combination3 - when dihedral (A12-123) and atomic character are considered, Combination4 - when interatomic A-1 distance, planar angle (<A12), dihedral (A12-123), and atomic character are considered.

QuACPAC (OpenEye Scientific Software). QuACPAC has been widely used for assignment of protonation states and partial atomic charges.[29] The output format of the file was converted from mol2 to pdb using a Perl script. The atom-names in the parent PDB file were discarded, and the tripos atom-type (as assigned by molcharge) was used. Manual checking of many of the more complex ligands was undertaken to confirm the correct assignments. For every protein−ligand complex, the interaction information involving protein atoms in contact with atoms of a single molecule of a specific ligand was extracted and placed in the Pair Contact Database (PCD). In total there were 1.1 million atomic contacts. For each interacting atomic contact pair information about atomic orientation was also stored. This includes the distance (A1), angular (<A12), and dihedral (A12−123) relationships between the terminal ligand atom (A) and the ultimate (1), penultimate (2), and antepenultimate (3) atoms of the protein (Figure 1).

The program Vega[30,31] was used to convert the atom types in the atom contact pair database from Tripos format to those listed in Table 1. Vega uses ATDL (atom type descriptive language) for classifying the typing atoms on the basis of their connectivity (see Supporting Information - Vega templates). In addition to the atom type conversion templates for the commonly used force fields we created our own classification schemes (see Supporting Information - SATs1 and SATs2 templates). The use of multiple atom-type classification schemes allows their comparative study and thus identification of the most suitable atom-type classification for the information theoretic approach used here (Table 2).

**Calculation of Atomic Contact Preferences.** Mutual information and joint entropy are indicative of the extent to which the distributions of two variables are related. While mutual information is a measure of mutual dependence of two variables,[32] joint entropy is the amount of uncertainty associated with two variables.[33] Mutual information is defined by

$$\mathrm{I(X,Y)} = \sum_{x}\sum_{y}\left(\mathrm{P}(x,y)\log\left(\frac{\mathrm{P}(x,y)}{\mathrm{P}(x)\mathrm{P}(y)}\right)\right) \qquad (1)$$

where P(*x*,*y*) is the joint probability distribution function of X and Y, and P(*x*) and P(*y*) are the marginal probability distribution functions of X and Y, respectively (where *x* and *y* are the ligand and protein atom types, respectively, in an interacting atom-atom pair). Joint entropy is defined by

$$\mathrm{H(X,Y)} = -\sum_{x}\sum_{y}\mathrm{P}(x,y)\log(\mathrm{P}(x,y)) \qquad (2)$$

For a given pair of interacting atoms the value of "P(*x*,*y*)log [(P(*x*,*y*)/(P(*x*)P(*y*))]" and "−P(*x*,*y*)log (P(*x*,*y*))" were respectively considered as the contribution of an individual pair of atoms toward the obtainable amount of mutual information and joint entropy. A complete set of these pairwise contributions (termed as MI-coefficients and JE-coefficients) for all the atom-atom contact pairs form the ensemble of atomic contact preferences between a protein and a ligand in the complexed state. MI-coefficients and JE-coefficients for an atom-atom pair were respectively defined as

$$\mathrm{MIcoeff(X,Y)} = \mathrm{P}(x,y)\log\left(\frac{\mathrm{P}(x,y)}{\mathrm{P}(x)\mathrm{P}(y)}\right) \qquad (3)$$

$$\mathrm{JEcoeff(X,Y)} = -\,\mathrm{P}(x,y)\log(\mathrm{P}(x,y)) \qquad (4)$$

The coefficients are applied to the set of atomic contacts between a specific protein−ligand complex to obtain the amount of mutual information or joint entropy for that complex. For a protein−ligand complex the sum of coefficients associated with all atom contact pairs was considered as the ScoreMI and ScoreJE, respectively:

$$\mathrm{ScoreMI(P{:}L)} = \sum_{x}\sum_{y}\left(\mathrm{P}(x,y)\log\left(\frac{\mathrm{P}(x,y)}{\mathrm{P}(x)\mathrm{P}(y)}\right)\right) \qquad (5)$$

$$\mathrm{ScoreJE(P{:}L)} = -\sum_{x}\sum_{y}\mathrm{P}(x,y)\log(\mathrm{P}(x,y)) \qquad (6)$$

**Generation of a Protein−Water Contact Database and Atomic Solvation-Desolvation Measures.** The preference of protein atoms to make contact with water atoms was calculated using the same approach as outlined above. Coordinates of water molecules were obtained by modeling water molecules on the protein surface using the AQUARIUS2 software.[15] For a data set of 999 proteins hydration shells were generated around each protein. Molcharge was used for the addition of protons, and Vega was used to convert the atom-types to the correct format from which protein atom-water contact pairs were extracted. These contacts were used to derive SSPs using the following relationship:

$$\mathrm{SSPcoeff(X)} = -\mathrm{P}(x,\mathrm{H_2O})\log[\mathrm{P}(x,\mathrm{H_2O})] \qquad (7)$$

SSP coefficients for protein atom-types were added to ScoreJE coefficients to obtain SIScoreJE

$$\mathrm{SIScoeff(P:L)} = \sum_{x}\sum_{y}(\mathrm{JEcoeff(X,Y)} + \mathrm{SSPcoeff(X)}) \qquad (8)$$

where SSP(X) is the Single-body solvation potential for X protein atom type. SIScoreJE comprised of JEcoeff(X,Y) and SSPcoeff(X) values.

**Protein−Ligand Test Set.** The performance of ScoreJE and SIScoreJE to predict the binding energy was evaluated on a data set of 100 protein−ligand complexes. None of these was a member of our training data set. Some of the complexes were obtained from a data set of 205 protein−ligand complexes;[34] others were taken from SCORPIO[35] and bindDB.[36] As the scores developed here consider only protein−ligand complexes, nucleic acid−ligand complexes were also excluded from the test set. The ligands in this data set were considerably diverse in terms of number of rotatable bonds (0−24), molecular mass (71−824 amu), number of heavy atoms (7−62), and number of aromatic rings (0−4). This displays a similar level of diversity to the training set in terms of the number of heavy atoms involved (see Figure 1, Supporting Information).

Since the training data set has a single SCOP superfamily representative for a specific ligand in complex with a protein, some of the SCOP superfamilies have more members than others. Intuitively the SCOP superfamilies having a larger number of member proteins could introduce an element of overtraining; however, since the ligand population has no repetition, the effect is minimal. Moreover, in order to eliminate all possible effects due to SCOP superfamily representation in the training set a "tailor made" training data set was created for each ligand in the test set. For each test set member the contacts were derived from the NRD (see above) by removing all those proteins belonging to the same SCOP superfamily as the test data set member. These were then used to derive atomic contact preferences which were unique for each test set member, removing any possible bias due to protein evolutionary relatedness (as defined by SCOP) during cross-validation.

**Ability To Identify Near-Native Configurations in Docking.** A data set of 50 protein−ligand complexes was used to determine the efficiency of the ScoreJE in identification of near-native configurations produced during docking. Only those protein−ligand complexes that did not have the presence of a cofactor or metallic ions in the ligand binding site were considered. The protonation states of ligand and the protein molecules were determined by OpenEye software (molcharge). For each of the test set members 100 docking solutions were generated using the default parameters of the GOLD[34] docking program. These docking solutions were ranked according to their ScoreJE and GoldScore values. The performance of the scoring function can be assessed from its success in ranking the near-native configurations of the ligand highly.

**Virtual Ligand Screening.** The discrimination of known inhibitors out of a large database of compounds is an important evaluation criterion for a scoring function. The rate of database enrichment with ligands indicates the proficiency of the scoring function to identify the true positives from a collection of physically similar but noninteracting molecules. The performance of ScoreJE and SIScoreJE for virtual ligand screening was evaluated using two databases from the DUD[37] collection. The DUD collection differs from similar collections[38] in their focus on decoys. The decoy molecules had similar physical properties (such as hydrophobicity, number of H-bond donor, number of rotatable bonds and acceptors, and molecular mass). The two databases used in ScoreJE evaluation were adenosine demaninase (ADA consisting of 23 known active ligands and 753 decoys) and phosphoribosylglycinamide formyltransferase (GART with 21 known active ligands and 821 decoys). For each protein (ADA and GART) the active site was defined as the residues surrounding the known ligand molecule (provided in the DUD data set) by using Surflex.[39] The respective small molecule collections were docked in the defined active sites under the premin and remin options. The former minimizes the energy of the small molecule before docking, whereas the latter performs all-atom active site minimization after docking. Top 20 docked poses with a minimum rmsd of at least 0.5 Å were rescored and ranked using ScoreJE (versions 1−4).

## RESULTS

**Choice of Scoring Function.** Scoring matrices for ScoreMI and ScoreJE from the Pair Contact Databases (PCDs) of 18 different atom-type classification schemes were calculated according to eqs 3 and 4, respectively (see the Methods section). For the comparative evaluation of the two scoring functions (ScoreMI and ScoreJE) 100 protein−ligand complexes were used, for which the experimental binding energies were known (A complete list is given in the Supporting Information). For each member of the test set, full cross-validation was performed eliminating any possible bias due to protein evolutionary relatedness in the training set. The ScoreMI and ScoreJE for protein−ligand complexes were calculated by summing up the MIcoeff and JEcoeff assigned for each atomic contact pair in the interaction database (eqs 5 and 6). The correlation coefficients ($R^2$ values) between scores were thus calculated (ScoreMI and ScoreJE), and the experimental binding energies were obtained via linear regression for all 18 atom-type classification schemes (Figure 2).

ScoreJE performs either better or equal to ScoreMI in all but two cases and gives the better correlation between calculated score and binding energies. In the subsequent work ScoreJE was adopted as the basal scoring function of choice.

**Choice of Scoring Parameters.** The best possible combination of atom-type set and orientation index was determined. The orientation indices include the distance between the interacting atoms, the angle of <A12, and the dihedrals of A12−123 (see the Methods section). These indices are continuous for protein−ligand interacting atom pairs; therefore, the orientations were binned into discrete values with distances rounded to one decimal place, and planar angles and dihedrals were binned in intervals of 10°. ScoreJEcoefficients were then calculated for each of orientation index for each of the 18 atom type data sets and the test set scored. The correlation coefficient ($R^2$) values between the calculated score and experimental binding energies are given in Figure 3.

As can be seen, the ScoreJE calculated for atom-type pair alone for SATs2 performed best. Also, in most of the cases atom-type pair alone performed better than the combinations of atom-type, intervening distances, angles, and dihedrals. Even though the number of descriptors for each atom-type classification scheme increases for the atom-type orientation index combinations of B, C, and D, the performance in terms of the correlation coefficient between calculated scores and experimental binding energies remains almost the same relative to atom-atom pairs alone. To test if ScoreJE with atom-type pair alone was sensitive to small changes in ligand conformation the minimization of 50 protein−ligand com-

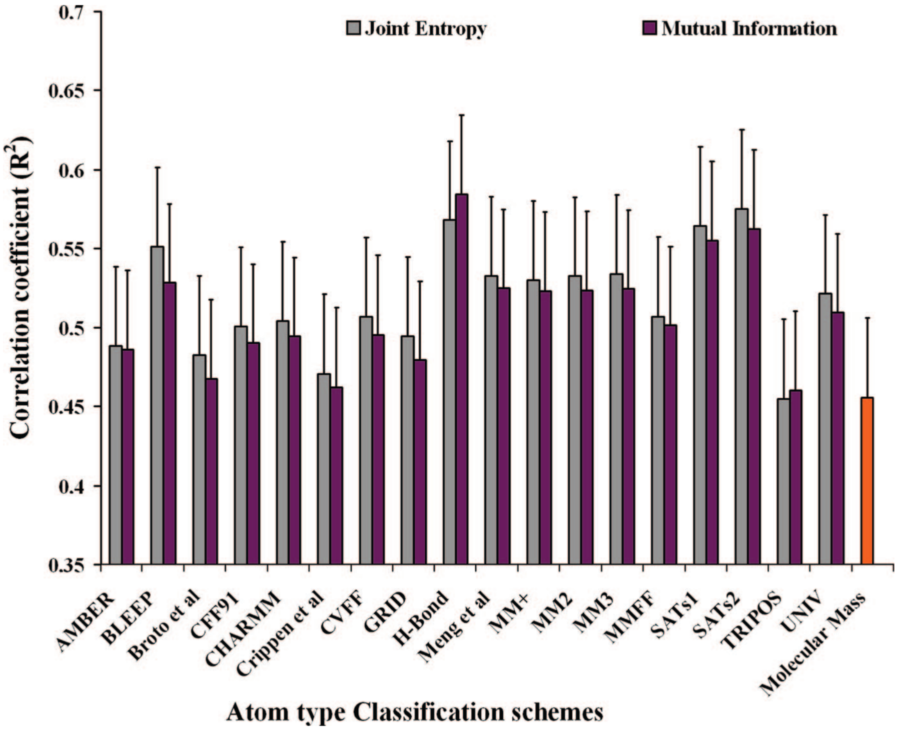

**Figure 2.** Comparative study of correlation coefficients ($R^2$) between experimental binding energies and the scores calculated using ScoreMI (Mutual Information) and ScoreJE (Joint Entropy) during cross-validation for 18 different atom-type schemes.

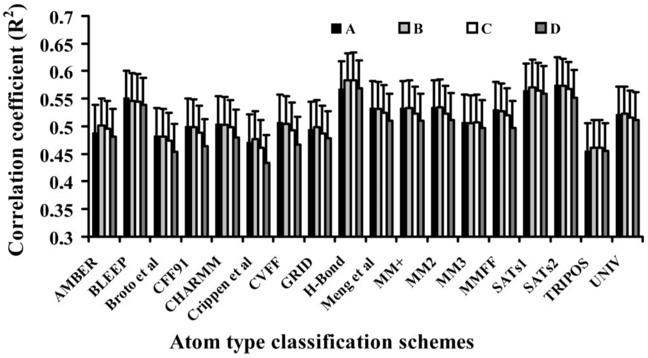

**Figure 3.** Comparative study of correlation between experimental binding energies and the scores calculated using ScoreJE (Joint Entropy) when A - only atom typecontacts are considered; B - atom typecontacts and interatomic distance are considered (A-1 from Figure 1); C - atom typeand dihedral angle (A12−123) are considered; D - atom type, interatomic distance (A-1), planar angle (A12), and dihedral angle (A12−123) are considered.

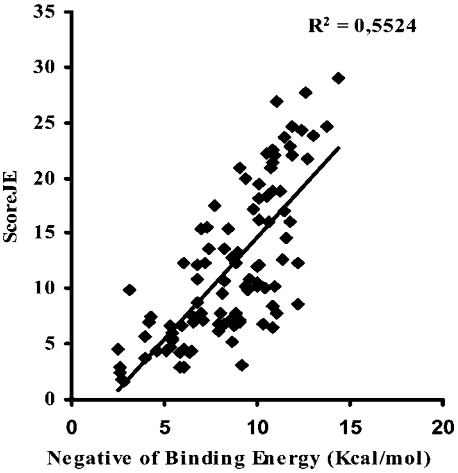

**Figure 4.** ScoreJE vs binding energy for 100 protein−ligand complexes.

plexes was undertaken, and the scores were compared with those of the crystallographic structures. It was found that both the scores and the number of scored contacts remain similar. The number of contacts varied at most by 5% (see the Supporting Information).

**Inclusion of Solvation Effects and SIScoreJE.** ScoreJE and SIScoreJE were obtained for atomic contact pairs for the SATs2 atom-type classification. The calculated scores for the 100 protein−ligand complexes of the test data set are plotted against the experimentally determined binding energies in Figures 4 and 5 for ScoreJE and SIScoreJE, respectively.

ScoreJE includes only protein−ligand direct interactions, whereas the SIScoreJE also includes the indirect interactions that take place with solvent molecules modeled using AQUARIUS2 (see the Methods section). The overall SIScoreJE scores correlated slightly better with the experimental binding energy than those with ScoreJE. In order to understand the influence and utility of SIScoreJE for various functional classes of proteins the above data set is further subdivided into acid proteases (31), serine proteases (12), carbohydrate binding proteins (16), and miscellaneous groups (41). A summary of results is given in Figure 6.

Predicting the binding energies of carbohydrates to their cognate binding proteins has previously been reported to be very problematic;[40] however, it is here that SIScoreJE performs significantly better than ScoreJE perhaps because carbohydrate binding sites are generally well-solvated. Indeed, a study which compared a series of different scoring functions (BLEEP, PMF, GOLD, DOCK, ChemScore) to predict the experimental logKd for 30 sugar binding proteins found that all but one method (BLEEP) gave poor correlation with experimental values.[41] For the serine proteases ScoreJE performs slightly better than SIScoreJE. A majority of the

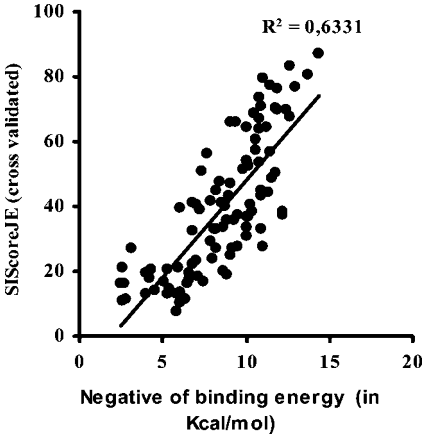

**Figure 5.** SIScoreJE vs binding energy for 100 protein−ligand complexes.

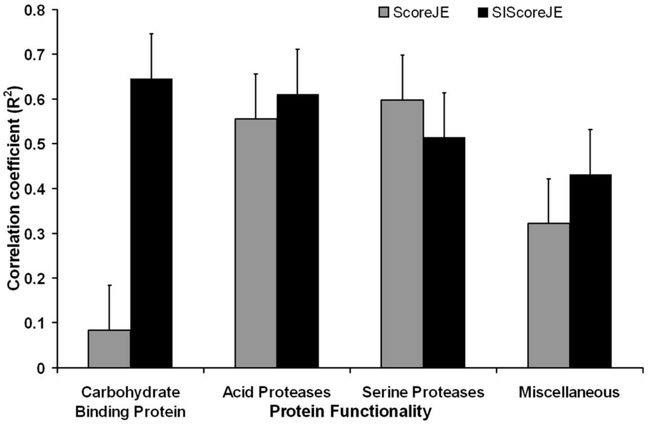

**Figure 6.** The correlation coefficient between experimental binding energies and ScoreJE (gray) and experimental binding energies and SIScoreJE (black) for different functional classes of protein−ligand complexes.

protease binding sites in the test data set have a predominance of hydrophobic character. Understandably, SIScoreJE is therefore less likely to have a significant influence in these complexes.

**Comparison of Different Scoring Functions.** The ability of ScoreJE and SIScoreJE to correlate with binding energy was compared against GoldScore, ChemScore, and X-Score.[20] The test set of 100 protein−ligand complexes were rescored using GOLD::GoldScore, GOLD::ChemScore, and X-Score. While GOLD::GoldScore provides a measure of fitness for the ligand−protein complex, it was not found to be very effective in predicting the binding energies. The GOLD::ChemScore estimates the $\Delta G$ of interaction in addition to the fitness score. X-Score gives three different scores. A consensus score (an average of the three as suggested by Wang et al.[20]) was taken as a measure for the binding energy. The degree of correlation between the calculated scores and experimental binding energy for the scoring functions is given in Figure 7.

X-Score performed better than GoldScore and ChemScore which is consistent with the results obtained by Wang et al.[20] However ScoreJE and SIScoreJE have the best correlation between predicted score and experimental binding energy.

**Ability To Identify Near-Native Configurations in Docking.** The ability of different scoring functions to successfully predict ligand binding affinity is one of the tests by which they can be compared. In addition, we have also analyzed the performance of the four versions of ScoreJE scoring functions to correctly identify near native ligand−protein conformations generated during docking runs. The four functions include different levels of information about atomic orientation according to the rules discussed above (Figures 1 and 3). The mutual interaction preferences were calculated by considering the interacting atom-type identities along with information about the interatomic distances (version2); dihedral angle (version3); and a combination of distances, planar angle, and dihedral angle (version4). The docking program GOLD[34] was used to generate 100 docking solutions for each of the 50 protein−ligand complexes in the docking test set (see methods). These poses were ranked according to the rmsd of the docking solution from the native ligand conformation present in the crystal structure. Conformations which had a rmsd less than 2 Å were considered as positives and the rest as negatives. The poses were evaluated using our SATs scoring functions as well as several other scoring functions available in GOLD, and receiver operating characteristic (ROC) curves were plotted in Figure 8. Ranking of the docked poses according to the various scoring functions indicate that GoldScore[34] performs best followed closely by ScoreJEversion4.

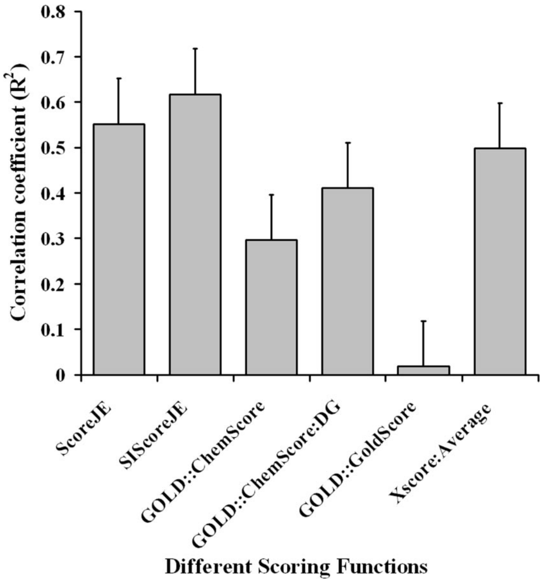

**Figure 7.** Comparative analysis of the scoring functions.

**Virtual Ligand Sceening.** A further test of the scoring function is its ability to predict biologically active ligands in virtual screening studies via its ability to retrieve biologically active ligands among a data set of physically similar decoys. We have performed enrichment studies comparing the ability of the four versions of ScoreJE to correctly identify known actives. The enrichment studies used the DUD collection for ADA and GART (see methods). The plot of the enrichment curves are shown in Figure 9. The proportion of the bioactive ligands found in the top 1 or top 10% of the ranked database is always higher than that identified by random selection (black curve) using all versions of ScoreJE. The level of applicability of a scoring function for virtual screening is directly proportional to this difference.[37]

**Phosphoribosylglycinamide Formyltransferase (GART).** GART is an essential enzyme in de novo purine biosynthesis, and hence it is important as a chemotherapeutic target. The

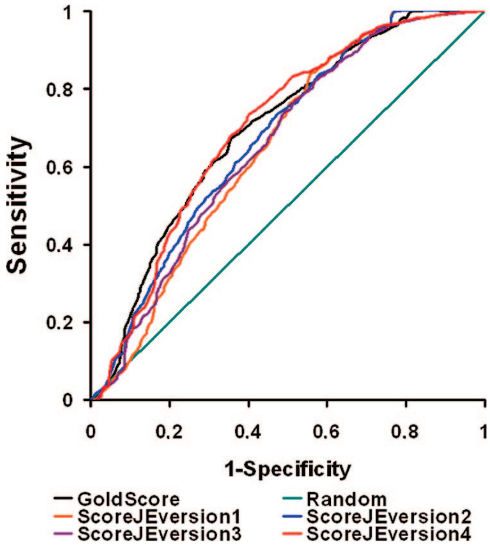

**Figure 8.** Receiver operating characteristic curves for comparative study of efficiency of GoldScore and various versions of ScoreJE in identifying near-native docking solutions for a data set of 50 protein−ligand complexes. ScoreJE when version1 - only atom typecontacts are considered, version2 - atom typecontacts and interatomic distance are considered (A-1 from Figure 3), version3 - atomtypes and dihedral angle (A12−123 from Figure 3) are considered, version4 - atom types, distances, angles, and dihedrals are included in the calculation of ScoreJE.

proportion of known actives in the top 1% of the ranked database is high at approximately 20−28% (Figure 9a). The performance of various versions of ScoreJE is similar. This is in contrast to the better performance of ScoreJE version4 during the identification of binding orientation. The predicted binding orientation of the top ranked ligand closely matches that of the crystallographic ligand (Figure 10a).

**Adenosine Deaminase (ADA).** ADA is a metalloenzyme which deaminates the adenosine. This protein has been targeted in the past for the treatment of cancer, rheumatoid arthritis, and psoriasis. The active site is large and contains a $Zn^{2+}$ coordination core consisting of 3 histidines. The proportion of known actives in the top 1% of the ranked database is 4−8%; however, in the top 10% it is 30−35%, a useful performance if not nearly as good as GART (with 75−80% of actives in the top 10%). In this case too the binding orientation of the top ranked ligand is similar to the cystallographic ligand in the active site (Figure 10b).

## DISCUSSION

Mutual information is a widely used statistic in several fields. Having first appeared in Shannon's paper[33] it has gained widespread acceptance in the applications of information theory. Whereas joint entropy[32] measures the amount of uncertainty or entropy associated with two random variables, mutual information measures the information. The protein−ligand interaction information obtained from the crystallographically determined structures was converted into protein−ligand atomic contact preferences. These were combined with predicted solvent interactions to create solvation potentials in SIScoreJE.

In our study joint entropy and mutual information coefficients were considered as scores representing the atomic contact preference scores. The ScoreJE for an interaction depends on the joint probability of occurrence of the protein−ligand interacting atom pair, and ScoreMI depends on the marginal probabilities of the individual atom types. The success of joint entropy over mutual information is evident for almost all atom-type definitions (Figure 2). This could be attributed to the nature of these two quantities. Mutual information is the amount of information one can calculate for the occurrence of an event on the basis of knowledge about the occurrence of another related event. In protein−ligand interaction terms: mutual information reduces the choice of a ligand atom identity that can form an interaction with a given protein atom in a particular ligand binding pocket. While this is useful it does not provide the measure of the amount of information the system will gain once the interaction takes place. Joint entropy, on the other hand, is a measure of uncertainty associated with two random variables. As information decreases in uncertainty, the joint entropy provides a more accurate estimate about the information the system gains once the ligand atom (with highest joint probability of occurrence) forms an interaction with the protein atom. Only when the joint probability of occurrence of the two entities is absolute does the mutual information become equal to the joint entropy. Perhaps localized regions on protein surface with high cumulative joint probability of occurrence have greater ligand binding potential. Such regions might be considered as information "hotspots" on the protein surface.[16]

In order to quantify the effects of orientation of interacting atoms on the efficiency of ScoreJE additional parameters (see the Methods section) were included in the calculation of the scoring function. Two interesting trends can be seen in Figure 3. First the basal scoring function (based on atom-type identity alone) performed best. Addition of orientation indices to the basal scoring function (calculated on the basis of atom-type identities) provides an enhanced ability to distinguish the stereochemically unfavorable interactions. However, as the probability of occurrence of stereochemically unfavorable contacts in high-resolution crystal structures is very low, the increased ability of the scoring function to identify unfavorable contacts remains unutilized. To test the hypothesis ligands were docked into their cognate receptor site using the GOLD program, and 100 poses were generated. Since the docking of a ligand to a receptor creates a number of stereochemically unfavorable atomic interactions, the data set of docked poses was used to study whether the inclusion of other orientational information has any effect on near-native pose identification. The ScoreJEversion4 (calculated by including the interatomic distances, planar angles and dihedrals) performed almost as well as GoldScore. The basal scoring function did not perform as well. This demonstrates the differential ability of the ScoreJEversion4 to distinguish between the stereochemically unfavorable and favorable contacts in a docking context. The performances of the other scoring functions were comparable to the basal scoring function. The number of interacting atom-type combinations increased dramatically (Table 2) on inclusion of additional orientation information. As the number of atomic contacts in the interaction database was constant the average amount of the information available per interacting atom-type combination was reduced during the calculation of ScoreJEversion2, ScoreJEversion3, and ScoreJEversion4. However, the effect of this reduced information did not affect

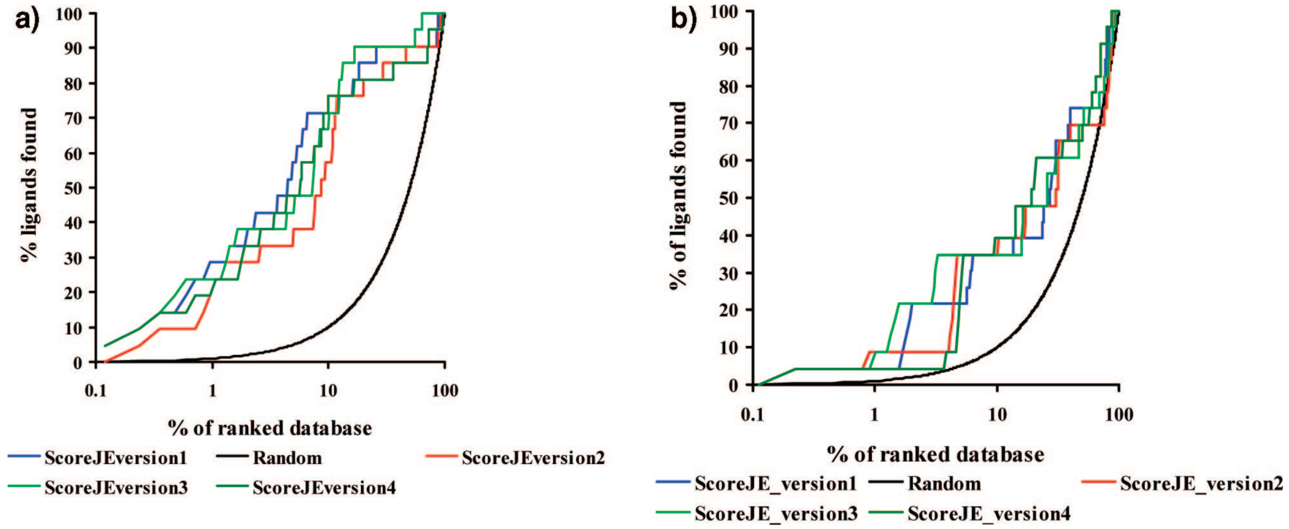

**Figure 9.** Docking enrichment curves for phosphoribosylglycinamide formyltransferase (GART) (a) and adenosine deaminase (ADA) (b). The docked and ranked database (*x*-axis) is plotted against the percentage of known bioactive molecules found in the database. The black curve represents the random selection of ligands.

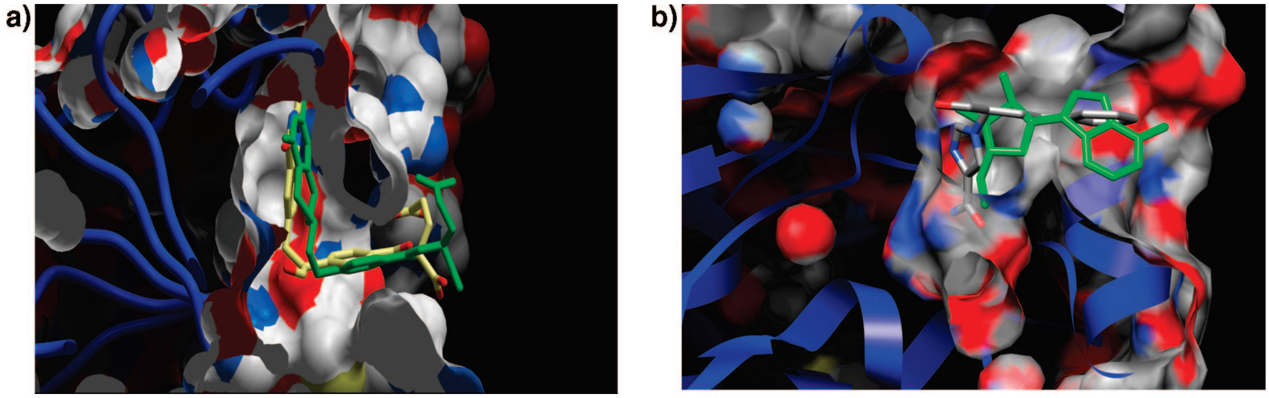

**Figure 10.** Binding orientation for the highest ranked small molecule (green) superimposed on the crystallographic ligand structure: (a) GART and (b) ADA. Chimera was used to generate the images.

the ability of the scoring function to predict the binding energies (Figure 3).

The development of the Single-body Solvation Potential was based on the same principle of joint entropy. As the number of water molecules in the crystal structures were inadequate to generate the solvation potentials the interaction between the protein atoms and the modeled water molecules was used. In order to make the SSP free from bias 1000 proteins were used for modeling the water molecules using AQUARIUS2 (see the Methods section). A large number of interactions occur between the ligand and protein atoms during complex formation as a consequence of the hydrophobic effect.[42] Inclusion of SSPs in ScoreJE improved the degree of correlation between the predicted scores and the experimental binding energies (Figures 4 and 5). However this improvement was mostly as a result of the carbohydrate binding proteins where water mediated interactions are more common (Figure 5). This leaves room for the development of better alternative solvation models with ScoreJE, creating a more efficient scoring function.

Evaluation of the accuracy of ScoreJE, SIScoreJE, GOLD: GoldScore, GOLD:ChemScore, and X-Score to predict protein−ligand binding affinity was carried out, and SIScoreJE and ScoreJE were seen to perform better than the rest. XScore is an empirical scoring function and has been seen to perform better than most scoring functions currently in use.[34] Similarly, ChemScore is an empirical scoring function that is widely applied (available in Sybyl and GOLD) in docking.

The enrichment of the database for the bioactive ligands is another check for the performance of a scoring function, and in this regard it is encouraging to see that ScoreJE performs well. However the performance of any scoring function for enriching the ranked database is also dependent on the binding pose generation algorithms. Hence the accuracy with which we have docked the ligands and the currently accepted definition for a correctly predicted pose (i.e., less than or equal to 2 Å rmsd from the crystallographic pose) also influence the quality of the results presented here.

## CONCLUSIONS

This paper describes the development of a novel, knowledge based scoring function designed to estimate the protein−ligand interaction energy. The ScoreJE was tested on a set of 100 protein−ligand complexes. The ability of the scoring function in ranking the protein−ligand docking solutions has been investigated. The ScoreJE scoring function which included the information of orientation along with the identities of the interacting atoms performs at the same level as GOLD:GoldScore in its ability to identify near-native configurations in docking while at the same time giving a higher degree of correlation between calculated and experimental binding affinity.

## ACKNOWLEDGMENT

This work was funded by the IMPRS-CB in the form of a studentship for M.K. and was carried out by M.K. while working as a visiting research student at the University of Leeds.

**Supporting Information Available:** Atom type definition, binding energy evaluation test set, atomic and van der Waals radii values, composition of ligand population, and comparision of the number of ligand−protein atomic contact in X-ray crystal structures. This information is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Rauh, D.; Klebe, G.; Stubbs, M. T. Understanding protein-ligand interactions: the price of protein flexibility. *J. Mol. Biol.* **2004**, *335* (5), 1325–1341.

(2) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11* (13), 580–594.

(3) Zentgraf, M.; Steuber, H.; Koch, C.; La Motta, C.; Sartini, S.; Sotriffer, C. A.; Klebe, G. How reliable are current docking approaches for structure-based drug design? Lessons from aldose reductase. *Angew. Chem., Int. Ed.* **2007**, *46* (19), 3575–3578.

(4) Gohlke, H.; Klebe, G. Statistical potentials and scoring functions applied to protein-ligand binding. *Curr. Opin. Struct. Biol.* **2001**, *11* (2), 231–235.

(5) Hendlich, M.; Lackner, P.; Weitckus, S.; Floeckner, H.; Froschauer, R.; Gottsbacher, K.; Casari, G.; Sippl, M. J. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **1990**, *216* (1), 167–180.

(6) Tanaka, S.; Scheraga, H. A. Statistical mechanical treatment of protein conformation. I. Conformational properties of amino acids in proteins. *Macromolecules* **1976**, *9* (1), 142–159.

(7) Wallqvist, A.; Jernigan, R. L.; Covell, D. G. A preference-based free-energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. *Protein Sci.* **1995**, *4* (9), 1881–1903.

(8) Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. 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* (7), 677–691.

(9) Pickett, S. D.; Sternberg, M. J. Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol.* **1993**, *231* (3), 825–839.

(10) Sharp, K. A.; Nicholls, A.; Fine, R. F.; Honig, B. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science* **1991**, *252* (5002), 106–109.

(11) DeWitte, R. S.; Shakhnovich, E. I. 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* (47), 11733–11744.

(12) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42* (5), 791–804.

(13) McDonald, I. K.; Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994**, *238* (5), 777–793.

(14) Mitchell, J. B. O.; Laskowski, R. A.; Alexander, A.; Forster, M. J.; Thornton, J. M. BLEEP - potential of mean force describing protein-ligand interactions II: Calculation of binding energies and comparison with experimental data. *J. Comput. Chem.* **1999**, *20* (11), 1177–1185.

(15) Pitt, W. R.; Goodfellow, J. M. Modelling of solvent positions around polar groups in proteins. *Protein Eng.* **1991**, *4* (5), 531–537.

(16) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295* (2), 337–356.

(17) Cline, S. M.; Karplus, K.; Lathrop, R.; Smith, T. F., Jr.; Haussler, D. Information-theoretic dissection of pairwise contact potentials. *Proteins* **2002**, *49* (1), 7–14.

(18) Finkelstein, A. V.; Gutin, A. M.; Badretdinov, A. Y. Perfect temperature for protein structure prediction and folding. *Proteins* **1995**, *23* (2), 151–162.

(19) Sternberg, M. J.; Bates, P. A.; Kelley, L. A.; MacCallum, R. M. Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.* **1999**, *9* (3), 368–373.

(20) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 2114–2125.

(21) Laskowski, R. A. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* **2001**, *29* (1), 221–222.

(22) Laskowski, R. A.; Hutchinson, E. G.; Michie, A. D.; Wallace, A. C.; Jones, M. L.; Thornton, J. M. PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.* **1997**, *22* (12), 488–490.

(23) Laskowski, R. A.; Chistyakov, V. V.; Thornton, J. M. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.* **2005**, *33* (database issue), D266−D268.

(24) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(25) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247* (4), 536–540.

(26) Allen, F. H.; Baalham, C. A.; Lommerse, J. P. M.; Raithby, P. R. Carbonyl-Carbonyl Interactions can be Competitive with Hydrogen Bonds. *Acta Crystallogr., Sect. B: Struct. Sci.* **1998**, *54* (3), 320–329.

(27) Diago, L. A.; Morell, P.; Aguilera, L.; Moreno, E. Setting up a large set of protein-ligand PDB complexes for the development and validation of knowledge-based docking algorithms. *BMC Bioinfo.* **2007**, 8, 310.

(28) Moreno, E.; León, K. Geometric and chemical patterns of interaction in protein-ligand complexes and their application in docking. *Proteins* **2002**, *47* (1), 1–13.

(29) Schnecke, V.; Kuhn, L. A. Virtual screening with solvation and ligand-induced complementarity. *Perspect. Drug Discovery Des.* **2004**, *20* (1), 171–190.

(30) Pedretti, A.; Villa, L.; Vistoli, G. VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs. *J. Mol. Graphics Model.* **2002**, *21* (1), 47–49.

(31) Pedretti, A.; Villa, L.; Vistoli, G. VEGA-an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming. *J. Comput.-Aided. Mol. Des.* **2004**, *18* (3), 167–173.

(32) Reza, F. Memoryless Finite Schemes. In *An Introduction to Information Theory*, 1st ed.; Dover Publications: Mineola, NY, 1994; Vol. 1, pp 104−106.

(33) Shannon, C. E. A Mathematical Theory of Communication. *The Bell System Tech. J.* **1948**, *27* (4), 623–656.

(34) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **2003**, *52* (4), 609–623.

(35) Ladbury, J. E. Scorpio Database. http://www.biochem.ucl.ac.uk/scorpio/scorpio.html(accessed July 26, 2007).

(36) Bader, G. D.; Donaldson, I.; Wolting, C.; Ouellette, B. F.; Pawson, T.; Hogue, C. W. BIND-The Biomolecular Interaction Network Database. *Nucleic Acids Res.* **2001**, *29* (1), 242–245.

(37) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801.

(38) Zhang, Q. M. I. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. *J. Med. Chem.* **2006**, *49* (5), 1536–1548.

(39) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, *46* (4), 499–511.

(40) Taylor, J. S.; Burnett, R. M. DARWIN: a program for docking flexible molecules. *Proteins* **2000**, *41* (2), 173–191.

(41) Marsden, P. M.; Puvanendrampillai, D.; Mitchell, J. B. O.; Glen, R. C. Predicting protein-ligand binding affinities: a low scoring game. *Org. Biomol. Chem.* **2004**, *2*, 3267–3273.

(42) Williams, D. H.; Bardsley, B. Estimating binding constants - The hydrophobic effect and cooperativity. *Perspect. Drug Discovery Des.* **1999**, *17* (1), 43–59.

CI800125K