ARTICLE

OBC
www.rsc.org/obc

Predicting protein–ligand binding affinities: a low scoring game?†

Philip M. Marsden, Dushyanthan Puvanendrampillai, John B. O. Mitchell* and Robert C. Glen
Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, UK CB2 1EW.
E-mail: jbom1@cam.ac.uk; Tel: +44-(0)1223-762983

Received 24th June 2004, Accepted 3rd September 2004
First published as an Advance Article on the web 27th September 2004

We have investigated the performance of five well known scoring functions in predicting the binding affinities of a diverse set of 205 protein–ligand complexes with known experimental binding constants, and also on subsets of mutually similar complexes. We have found that the overall performance of the scoring functions on the diverse set is disappointing, with none of the functions achieving r^2 values above 0.32 on the whole dataset. Performance on the subsets was mixed, with four of the five functions predicting fairly well the binding affinities of 35 proteinases, but none of the functions producing any useful correlation on a set of 38 aspartic proteinases. We consider two algorithms for producing consensus scoring functions, one based on a linear combination of scores from the five individual functions and the other on averaging the rankings produced by the five functions. We find that both algorithms produce consensus functions that generally perform slightly better than the best individual scoring function on a given dataset.

Introduction

The development of scoring functions to accurately predict experimental binding free energies for protein–ligand complexes is currently a major challenge in structure-based drug design. Virtual screening of large chemical databases often involves pre-filters such as druglikeness, followed by docking of potential ligands into specific known target receptor sites.[1–4] Docking programs search for viable conformations of ligands to reside in target receptor sites and a scoring function, often different from that used within the docking algorithm, ranks the docked conformations in terms of the quality of the fit between the ligand and receptor. As well as ranking different docked 'poses' of a single complex, the scoring function should ideally also be able to make predictions of binding affinity, which allows different candidate molecules to be ranked in terms of their predicted binding to a given target. The ability of docking programs to accurately predict protein–ligand complex geometries has significantly improved in recent years,[5] so it is reasonable to expect that the best methods will increasingly give correctly docked solutions. However, predicting binding affinities is a different problem. We expect that improvements in the scoring functions will be crucial in addressing this.

The formation of a protein–ligand complex involves the steric fit and physicochemical interactions between the two molecules, conformational changes, and changes in interaction with solvent. Methods of evaluating the associated free energy change generally use potential functions, either empirical or knowledge-based, specifically designed for this purpose.[6] Typically, the experimental K_d values are treated as givens and little or no attention is paid to the various experimental methods, including displacement of different ligands, by which they have been measured.

Empirical scoring functions estimate the binding free energy of a protein–ligand complex by partitioning it into a sum of terms identified with various physically distinct contributions.[7–11] These terms are typically identified with concepts such as hydrogen bonding, van der Waals and hydrophobic contacts, and with specific identifiable entropic contributions such as frozen rotations and solvation entropy.

Knowledge-based methods take a different approach, deriving energy-like functions by considering the distributions of interatomic distances from experimental protein–ligand complex structures. These distributions are converted into energy-like functions by assuming Boltzmann-like energetics,[12–16] the process being carried out for each pair of atom types. When a given test structure is assessed, features common in the database score favourably, and less frequent interactions score less favourably. The overall score indicates how much a structure 'resembles' real protein–ligand complexes. Importantly, this approach requires only structural, and not binding, data to derive the parameters. Knowledge-based methods have been shown to give useful correlations between experimental and calculated binding energies.[12,15,16] The topology and connectivity of molecules makes it very hard to assess how much independent information the ensemble of interatomic distances represents; clearly the positions of covalently bonded neighbouring atoms are correlated. Thus an additional empirical parameter, calibrating the knowledge-based 'energy' against binding energies, should greatly improve the accuracy of the predictions in absolute terms. For BLEEP, this has the effect of dividing the raw energies, as calculated by the original methodology,[14,15] by a factor of 11.948.

Empirical scoring functions make the assumption that the energy contribution from each type of interaction present in the complex can be summed to give the free energy change of binding. The empirical scoring function used in DOCK[17–20] ignores solvation, conformation and entropic effects and uses molecular mechanics to estimate the binding energy. The lack of entropic and conformational effects means that DOCK may be expected to give relatively poor correlations for diverse sets of complexes. The ChemScore scoring function, which is also empirical in nature, uses lipophilic interactions, metal–ligand binding, hydrogen bonding and ligand flexibility loss to calculate binding energy.[21] The empirical scoring function used in the docking program GOLD models the hydrogen bond interactions in the protein–ligand complex, including their directionality, van der Waals contacts between protein and ligand, and also the ligand's conformational energy. The GOLD scoring function performs well in cases with predominately polar interactions, but poorly for primarily hydrophobic cavities with few polar interactions.[22] Both the DOCK and GOLD functions were designed for accurate docking rather than affinity prediction.

In a study of four scoring functions using FlexX as the docking program, Stahl and Rarey[5] suggested that some functions are more suited for the evaluation of specific types

† This is one of a number of contributions on the theme of molecular informatics, published to coincide with the RSC Symposium "New Horizons in Molecular Informatics", December 7th 2004, Cambridge UK.

DOI: 10.1039/b409570g

of targets, such as lipophilic and polar receptors. Based on the complex being docked, combinations of two from the four possible scoring functions showed significantly improved results in virtual screening. Their results support the idea that, as yet, no single scoring function is able to work for every protein–ligand complex. Thus, consensus scoring, where the results generated from several different scoring functions are combined, has recently been more widely used to improve the hit rates from docking virtual libraries into target receptor sites. A number of different approaches to consensus scoring are possible, of which averaging the scores from different functions is one. In essence, the mean of the binding energies predicted by a number of scoring functions is likely to be closer to the true value than is the energy from a single function. Scoring a molecule according to the best value from any of the functions is another possible consensus approach; this would aggregate the top few hits from each function into the final list. The study of Wang *et al.* of eleven scoring functions reported that only four gave rank correlation coefficients (R_s) better than 0.5 with experimental binding affinities.[23] Bissantz *et al.*'s evaluation of seven scoring functions showed that they were unable to predict absolute binding free energies.[24] Charifson *et al.* reported that a consensus approach gave a significant reduction in the number of false positive results, compared with individual scoring functions.[25] Some methods have included a feedback loop where the results from the scoring function are used to encourage the docking method to alter the conformation of the ligand molecule in the receptor site, thus improving the subsequent scoring function results.[26]

Here we compare the results of five different scoring functions in calculating binding affinities for 205 protein–ligand complexes with known three-dimensional structures,[27] and also for various subsets of mutually similar complexes. We used the knowledge-based methods PMF[16,28] and BLEEP[14,15] and the empirical scoring functions of GOLD,[22] DOCK[17] and ChemScore.[21,29] All these knowledge-based and empirical scoring functions have been used in molecular design, but no comparative study of their performance on a single test set has yet been published. We also define and examine the performance of two consensus functions, each based on these five scoring functions.

Results

We applied the five scoring functions to a test set totalling 205 different protein–ligand complexes from the Protein Databank (PDB). We also used a number of subsets of these data: five functionally-based subsets (Table 1); subsets based on ChemScore training and test sets[21] (Table 2); and the BLEEP test set[15] (Table 3).

Scores were calculated using BLEEP, PMF, GOLD, DOCK and ChemScore and compared with the experimental logK_d values. We calculated r^2 and R_s between the scores given by the five scoring functions and the experimentally measured logK_d values for the 205 protein–ligand complexes in the dataset (Table 1 and Fig. 1). Of the five scoring functions evaluated, BLEEP gave the best agreement between its calculated binding free energy and experimental logK_d values, with $R_s = 0.59$. GOLD, ChemScore and DOCK had mutually similar R_s values of 0.50, 0.44 and 0.43 respectively, while PMF gave an R_s value of 0.32. All five scoring functions gave modest r^2 values, the highest being the 0.32 given by BLEEP.

We identified five characteristic subsets of our data (see Table 1), each of which represents a series of related protein–ligand complexes; these are similar, but not identical, to those used by Ishchenko and Shakhnovich.[13] Our serine proteinase subset contains 35 complexes, for which BLEEP and DOCK scores gave a good correlation with experimental logK_d, with r^2 values of 0.74 and 0.69 together with R_s values of 0.82 and 0.83 respectively. PMF and GOLD performed reasonably well, but ChemScore was not able to reproduce the trend in binding affinities. For our sample of 25 metalloproteinase

Table 1 Correlations between calculated and experimental logK_d values given by five scoring and two consensus functions[a]

		BLEEP		PMF		GOLD		DOCK		ChemScore		AR		LC	
Dataset	No. of complexes	R_s	r^2	R_s	r^2	R_s	r^2	R_s	r^2	R_s	r^2	R_s	r^2	R_s	r^2
All	205	0.59	0.32	0.32	0.16	0.50	0.20	0.43	0.21	0.44	0.19	0.58	0.34	0.62 ± 0.02	0.36 ± 0.02
Serine proteinases	35	0.82	0.74	0.75	0.51	0.61	0.47	0.83	0.69	0.13	0.00	0.80	0.65	0.81 ± 0.03	0.70 ± 0.04
Metalloproteinases	25	0.72	0.44	0.45	0.33	0.44	0.14	0.42	0.20	0.43	0.17	0.55	0.45	0.67 ± 0.04	0.48 ± 0.06
Carbonic anhydrase II	18	0.53	0.47	0.46	0.32	0.42	0.34	0.54	0.01	0.50	0.28	0.70	0.43	0.35 ± 0.08	0.15 ± 0.08
Sugar binding proteins	30	0.76	0.58	0.45	0.09	0.02	0.00	0.05	0.00	0.29	0.08	0.51	0.29	0.63 ± 0.05	0.48 ± 0.05
Aspartic proteinases	38	0.08	0.01	−0.52	0.25	0.02	0.00	0.08	0.01	−0.08	0.00	−0.10	0.02	0.37 ± 0.04	0.13 ± 0.01

[a] Here and in subsequent tables, r^2 denotes the square of the product moment correlation coefficient and R_s is Spearman's rank correlation coefficient. AR denotes the Average Rank consensus method and LC the Linear Combination of scoring functions consensus method; these are described more fully in the Methods section.

Table 2 Correlations between calculated and experimental $\log K_d$ values for five scoring functions on the ChemScore training and test sets

			BLEEP		PMF		GOLD		DOCK		ChemScore	
Class	Dataset	No. of complexes	R_s	r^2	R_s	r^2	R_s	r^2	R_s	r^2	R_s	r^2
1	Serine proteinases	14	0.84	0.81	0.74	0.68	0.45	0.39	0.48	0.56	0.53	0.40
2	Metalloproteinases	14	0.88	0.77	0.76	0.57	0.79	0.53	0.74	0.71	0.78	0.61
3	Miscellaneous	17	0.41	0.06	0.29	0.02	0.07	0.00	0.01	0.01	0.00	0.00
4	Sugar binding proteins	16	0.60	0.42	0.76	0.21	−0.04	0.06	0.74	0.08	0.21	0.10
5	Aspartic proteinases	17	0.53	0.11	−0.30	0.10	0.13	0.01	0.44	0.07	0.18	0.01
	ChemScore test set 1[21]	20	0.36	0.12	0.15	0.01	0.32	0.07	0.42	0.14	0.82	0.57
	ChemScore test set 2 (endothiapepsins)[21]	10	0.03	0.07	0.31	0.58	0.37	0.43	0.22	0.34	0.10	0.23
	All ChemScore training and test sets	108	0.56	0.31	0.34	0.04	0.31	0.16	0.39	0.23	0.47	0.25
	All other complexes	97	0.60	0.36	0.33	0.19	0.55	0.27	0.36	0.19	0.38	0.16
	Entire dataset	205	0.59	0.32	0.32	0.16	0.50	0.20	0.43	0.21	0.44	0.19

Table 3 Correlations between calculated and experimental $\log K_d$ values for five scoring functions on the BLEEP test set

		BLEEP		PMF		GOLD		DOCK		ChemScore		AR	
Dataset	No. of complexes	R_s	r^2	R_s	r^2	R_s	r^2	R_s	r^2	R_s	r^2	R_s	r^2
BLEEP test set[15]	90	0.77	0.54	0.50	0.34	0.72	0.41	0.69	0.45	0.43	0.17	0.75	0.56
All other complexes	115	0.46	0.18	0.21	0.04	0.32	0.09	0.25	0.06	0.44	0.20	0.43	0.21
BLEEP test set including outliers	96	0.71	0.41	0.51	0.32	0.68	0.32	0.65	0.38	0.43	0.17	0.72	0.48
Entire dataset	205	0.59	0.32	0.32	0.16	0.50	0.20	0.43	0.21	0.44	0.19	0.58	0.34

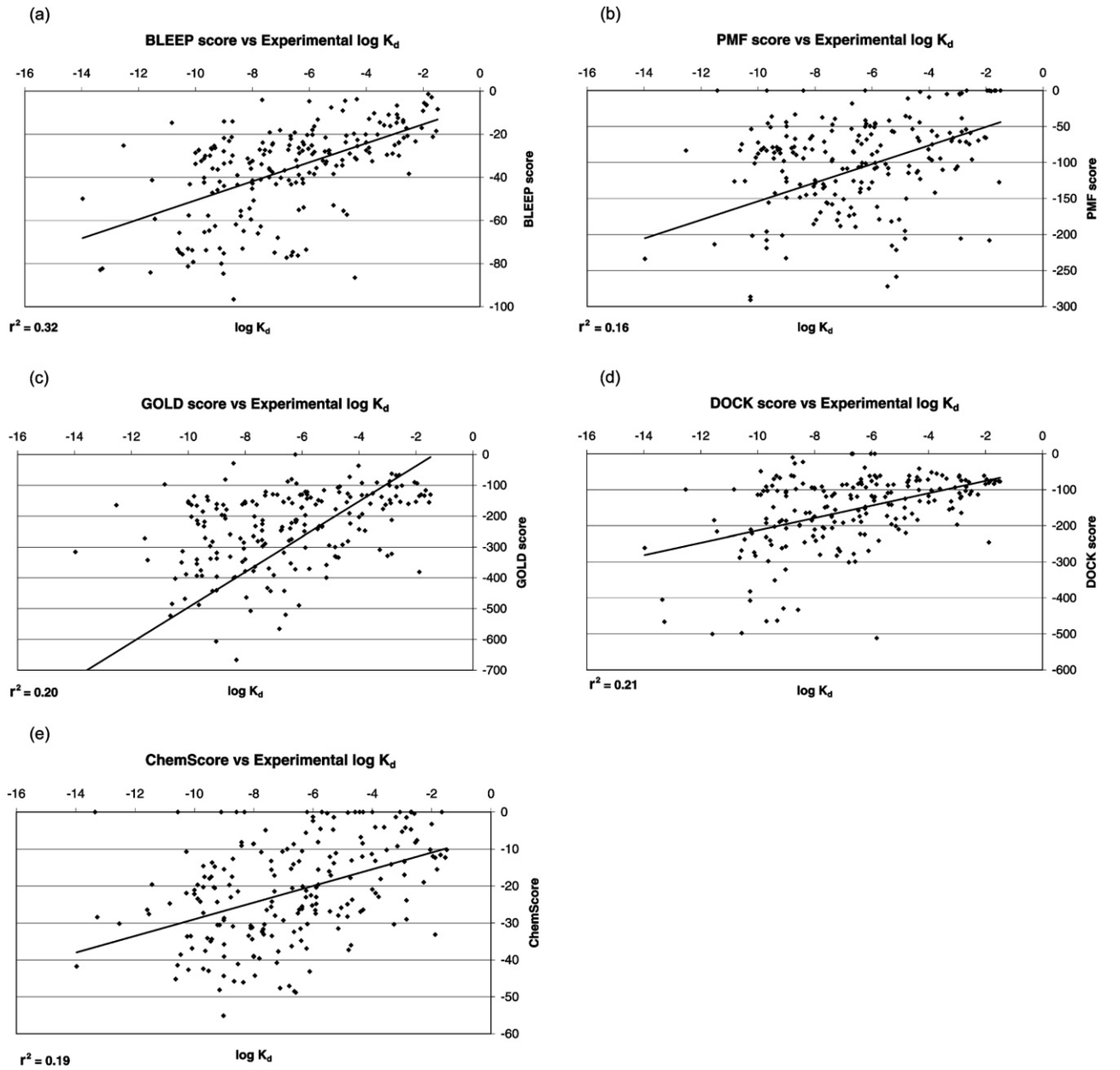

Fig. 1 Experimental logK_d values for 205 complexes plotted against the corresponding scores calculated by (a) BLEEP; (b) PMF; (c) GOLD; (d) DOCK; (e) ChemScore.

complexes, BLEEP gave an r^2 of 0.44 and an R_s value of 0.72. The ChemScore, PMF, GOLD and DOCK scoring functions performed poorly in predicting the binding energies for this subset. All five methods gave R_s values between 0.42 and 0.54 for 18 carbonic anhydrase II complexes. For the subset of 30 sugar binding proteins, four of the scoring functions had a poor correlation with experimental logK_d, though BLEEP did reasonably well with an R_s value of 0.76. For the final subset of 38 aspartic proteinase complexes, the situation was very disappointing, with none of the methods producing any kind of useful correlation. The r^2 value of 0.25 for PMF actually came from a negative correlation coefficient, $r = -0.50$.

A Naïve BLEEP method, where there was no human intervention to fix apparent errors, produced essentially identical results to BLEEP (any difference being less than 0.1 kJ mol^{-1} or 0.0175 log$K$ units) for 171 of the 205 complexes, slightly different results (by between 0.1 and 1 kJ mol^{-1}) for five complexes, moderately different results (by between 1 and 10 kJ mol^{-1}) for eight complexes, and very different results (by more than 10 kJ mol^{-1}) in only two cases. Naïve BLEEP failed to produce a result for 19 complexes. The R_s and r^2 values between Naïve BLEEP and experiment over the 186 complexes were virtually identical to those of BLEEP. These results suggest that BLEEP would fail in about 10% of cases when used as a 'black box', but reassure us that it would not produce a significant number of erroneous but credible scores.

Table 2 gives the correlation coefficients for the various ChemScore test sets. The most clear-cut result here is that ChemScore performs much better on its own 'Test set 1'[21] than it does on the remainder of our dataset. Table 3 shows that the BLEEP test set[15] is a relatively 'easy' set, with BLEEP, PMF, GOLD and DOCK all performing much better on these 90 complexes than on the remaining 115, although ChemScore performs equally well on the two sets.

We now discuss the results obtained using two consensus methods, Linear Combination of scoring functions (LC) and Average Rank (AR), each of which is described in the Methods section. Table 4 shows more details of the performance of the LC function when the dataset was partitioned into training and test sets of 100 and 105 complexes, respectively. LC only modestly outperformed the best of the individual functions (BLEEP) in this test, with $R_s = 0.62 \pm 0.02$ and $r^2 = 0.36 \pm 0.02$. The weights w_i are not reliable indicators of the quality of the individual scoring functions, since some pairs of functions are highly correlated and hence contribute similar information to LC. Looking at our data as a whole (Tables 1, 3 and 4), the

Table 4 Scoring and consensus data[a] for ten partitionings into {100 training} + {105 test}

	R_s	r^2	RMS	σ	w
BLEEP	0.60 ± 0.02	0.33 ± 0.02	3.00 ± 0.04	0.90 ± 0.03	0.42 ± 0.03
PMF	0.29 ± 0.02	0.14 ± 0.01	4.16 ± 0.07	0.057 ± 0.004	0.34 ± 0.02
GOLD	0.48 ± 0.02	0.19 ± 0.01	4.93 ± 0.06	0.016 ± 0.002	−0.07 ± 0.02
DOCK	0.42 ± 0.02	0.21 ± 0.02	3.57 ± 0.04	0.046 ± 0.002	0.06 ± 0.02
ChemScore	0.43 ± 0.02	0.18 ± 0.01	3.65 ± 0.06	1.39 ± 0.06	0.39 ± 0.02
AR	0.56 ± 0.02	0.33 ± 0.02	2.49 ± 0.05		
LC	0.62 ± 0.02	0.36 ± 0.02	2.64 ± 0.04		

[a] Values for σ and w were calculated for each scoring function over the training set of 100 complexes. R_s, r^2 and RMS error (in $\log_{10}K_d$ units) were calculated over the test set of 105 complexes. The values quoted are the means over 10 randomised partitionings of the total dataset, with the standard errors of the mean quoted after the ± symbol. In the current version of BLEEP, energies are now pre-scaled by a factor of (1/11.948) compared with the methodology published[14] in 1999, in order to bring them more closely into line with experimental values. Hence the σ value calculated here for BLEEP is close to 1. The AR predicted $\log K_d$ values used the mean and standard deviation of the distribution of the experimental values in the training set and the ranks of complexes in the test set.

consensus methods AR and LC both generally performed similarly to the best individual scoring function on a given dataset. LC performed poorly on the carbonic anhydrase II set due to a single outlying score for one function (PDB entry 1okm with DOCK), which led to a large RMS error whenever this was found in the small test set of nine complexes. The RMS errors of the predicted K_d values from Table 4 were 2.64 ± 0.04 for LC, 3.00 ± 0.04 for BLEEP and 2.49 ± 0.05 $\log K_d$ units for AR. The relative success of AR on this measure is probably due to the averaging inherent in AR diminishing the effect of outlier scores from individual functions.

Discussion

The inescapable conclusion from these results is that the problem of accurately predicting the binding energies of a large and diverse set of protein–ligand complexes is an extremely difficult one. None of the scoring functions tested here achieved r^2 values above 0.32 when tested on the full 205 complex dataset. This is a modest level of performance, although we should note in defence of the GOLD and DOCK functions that they were designed to identify the correct geometries of bound complexes and were not intended to be applied to the problem of affinity prediction.

In practice, however, it is rarely necessary to deal with such a diverse set of complexes. It is more likely that the scoring functions will be used to evaluate either different docked 'poses' of a single ligand in a given receptor, or comparative binding energies of a series of (possibly related) ligands in a given binding site. Predicting the binding affinities for a set of similar ligands binding related proteins is a commonly used way of validating scoring functions, as the structural data are available in the public domain in the PDB.[38,39] This represented the main way of testing both PMF[16] and ChemScore,[21] and was one of a number of tests used to validate BLEEP.[15] There is substantial overlap between the test sets used in the original PMF[16] and ChemScore[21] publications, some of which have formed the basis of the test sets[27] in the present study (see Tables 1 and 2). Such problems, involving families of related complexes, are expected to be substantially more tractable than tackling a highly diverse set of complexes, but harder than limiting the test set to complexes of a single protein.

The diverse set problem is also quite different from virtual screening, that is the identification of a small number of hits amongst many non-binders. In a diverse set consisting only of binders, binding energy is likely to correlate with molecular weight.[30] Scoring functions typically sum atom–atom contributions, hence tending to give better scores to larger molecules,[31–33] and this non-specific effect will contribute to the apparent success of the function. Virtual screening, however, requires the identification of specific favourable and unfavourable interactions and, in a dataset with many non-binders, molecular weight is not a good predictor of affinity. The approach where a diverse library is computationally docked into a variety of targets is reasonable from a virtual screening point of view, when only a general indication of likely binding is required, but will not accurately predict affinities.

The headline figures for the correlation coefficient given here seem less impressive than in previous work.[15] This is partly due to the choice of dataset. Six outliers were excluded from that previous analysis of BLEEP, which raised the r^2 value from 0.40 to 0.55. Using the current version of BLEEP, the corresponding r^2 values are 0.41 with outliers present and 0.54 without them. Also, as discussed above, the BLEEP test set seems to have been 'easier' than the current set (Table 3). The outliers seem to be potential-specific, so the six 'worst case' complexes for BLEEP degrade its performance much more than they do the performance of other scoring functions (compare the first and third rows of Table 3). The effects of selective outlier removal are probably part of the reason why a given potential often performs better on the test set from its own publication than on other, apparently similar, sets.

It seems to us that the comparison of scoring functions would be greatly facilitated if the community were to adopt standard test sets on which to carry out such experiments. Separate sets would be required to test the performance of these functions on different kinds of problem, such as virtual screening, affinity prediction in diverse sets, affinity prediction in homologous sets, and docking. For our part, we have made available the experimental binding affinity data used in this work.[27] We note also that Wang *et al.* have recently made available a database of 1359 experimental protein–ligand binding affinities.[34]

We have found substantial differences amongst our five characteristic subsets. The serine proteinase subset is 'easiest', with four of the five scoring functions giving R_s values above 0.6. In contrast, none of these functions was able to find meaningful correlations for the aspartic proteinases, where the protocol employed here takes no account of the role played by water in the active site. The other three subsets were of intermediate difficulty. It seems to us that the current state of scoring functions is adequate for only some sets of complexes. In other cases, it may be necessary to design specific functions for specific classes of target, for instance as Verkhivker and Rejto[35] created a function for HIV proteinase complexes.

Our results also suggest that consensus scoring performs somewhat better than any individual scoring function, and is inherently less prone to the effects of potential-specific outliers. We recognise that the use of such methods does significantly increase the computational effort required and hence may not always seem worthwhile in CPU-intensive and relatively low accuracy applications such as virtual screening. Nonetheless, using consensus scoring based on a plurality of functions is likely to increase the diversity of hits, as it has for chemical structure searches.[36]

Methods

Preparation of the dataset

The test set used in this study consists of 205 protein–ligand complexes[27] with experimentally measured K_d values, which have been assembled from scoring function and/or docking evaluation studies.[5,7,23,25,37] Their dissociation constants range from −1.49 to −13.97 in $\log K_d$ units, spanning over 12 orders of magnitude and having been measured by a variety of experimental methods. All ligands bind non-covalently to the proteins. The 205 complexes include the 90 on which BLEEP was originally tested[15] and the six outliers that were excluded from that analysis (but only two of the 188 complexes from which BLEEP was derived[14]), 78 from the ChemScore training sets and 30 from the ChemScore test sets.[21]

Coordinates for the protein–ligand complexes were downloaded from the Protein Databank (PDB).[38,39] For BLEEP, the PDB file provided both protein and ligand molecules, which were automatically identified by the software. For PMF, GOLD, DOCK and ChemScore, the protein and ligand molecules were extracted from the ATOM and HETATM records in the PDB file, and saved in MOL2 format. Where metal ions were present in the protein, the standard metal parameter file was used and the atom types of the metals verified. For multimeric proteins with multiple active sites, a single instance of the ligand was extracted for scoring to maintain consistency with the BLEEP methodology. Hydrogen atoms were added to both protein and ligand molecules using Sybyl®,[40] except for BLEEP where only polar hydrogens were added using HBPLUS.[41] The Gasteiger method was used to assign charges for the Sybyl® implementation of DOCK and protonation states were chosen as those expected at pH 7, given the pK_a values of the ionisable groups. The atom types and bond orders of the protein and ligand molecules were confirmed manually.

BLEEP is stand-alone software available from the Unilever Centre for Molecular Science Informatics. The CScore consensus scoring module within Sybyl® 6.9[40] generates results using the PMF, GOLD, DOCK, FSCORE and ChemScore scoring functions. We made no use of the FSCORE or CScore functions. This work was carried out using an SGI ORIGIN R6000 twin processor workstation.

Scoring functions

The reader is referred to the original papers for full details of the scoring functions used. These include two knowledge-based functions, BLEEP[14,15] and PMF,[16,28] together with the empirical functions from the GOLD[22] and DOCK[20] docking protocols, and also the empirical ChemScore[21] function. BLEEP, PMF and ChemScore are designed explicitly to predict binding affinities, while the GOLD and DOCK functions are not; these functions are intended to give accurate docked geometries. All of the scoring functions are available commercially with the exception of BLEEP, which academic users can obtain from us without charge.

In the few cases where BLEEP failed to produce a result, or where its result seemed surprising, we have investigated on a case-by-case basis and fixed any errors that occurred, most often in the automatic assignment of atom types. In order to illustrate the use of BLEEP as a black box, which is a better model of how high throughput screening would work, we have also recorded a set of 'Naïve BLEEP' results, where no attempt is made to fix errors and failed runs are discarded. For the other scoring functions, we have investigated any scores that seemed suspicious and fixed any obvious errors. Where a positive (repulsive) score has persisted, it has been treated as zero in our analyses.

Consensus by linear combination of scoring functions (LC)

We have investigated a consensus scoring function based on the linear combination of scaled results from the existing functions. We call this LC (Linear Combination). First, we scaled each of the original five scoring functions F_i by a factor σ_i, designed to minimise the root mean square error of the predicted $\log K_d$ values. This calibrated the individual scoring functions against one another. We then used multiple linear regression to minimise the root mean square error, over a suitable training set, of the function

$$\mathrm{LC} = \Sigma_i w_i(\sigma_i F_i) \quad (1)$$

where the w_i are weights to be optimised. The use of the scaling factors σ_i allows meaningful direct comparison of the weights assigned to the individual functions F_i within the overall LC. These weights measure the contributions to LC, but not the performance, of the functions F_i.

The LC procedure was applied by randomly splitting the 205 complexes into a training set of 100 and a test set of 105. The process was carried out ten times, with different random partitionings into training and test sets, the function being trained afresh over each training set. When applied to smaller datasets, LC was retrained on (approximately) half of the set and tested on the other half. This process was carried out, as before, on ten randomised partitionings of each set.

Consensus by average rank (AR)

One simple way to obtain a consensus score is to rank all complexes in the relevant test set according to each scoring function, and then to use the average of these ranks as a measure of the ligand's predicted affinity. In this implementation, which we call AR (Average Rank), each complex c is ranked from 1 (least affinity) upwards using each scoring function i (this index ranging over the five scoring functions), with ranks R_i^c. The AR function is then defined as

$$\mathrm{AR}(c) = \sum_{i=1}^{5} -R_i^c / 5. \quad (2)$$

This definition ensures that the strongest predicted affinities are associated with the largest negative AR values. AR can be used to predict $\log K_d$ by mapping the AR scores onto a normal distribution with the same mean and standard deviation as the experimental $\log K_d$ values in the training set. We chose to pursue AR, having found in preliminary studies that it outperformed the alternative of taking the best rank.

Statistical measures

Evaluating the correlation between calculated binding energies and experimental values provides an indication of the performance of a scoring function and of how well these values can predict the binding affinities of protein–ligand complexes. The Pearson's product moment correlation coefficient, denoted by r, is a measure of the linear association between two variables, here the experimental $\log K_d$ and the score or energy calculated by a particular scoring function. The square of the product moment correlation coefficient, denoted by r^2, is a commonly used measure of correlation, giving a value of between 0 and 1. The correlation between experimental $\log K_d$ and calculated binding energy does not, however, have to be linear, and Spearman's rank correlation coefficient (R_s) is an accurate quantitative measure of the relationship between two sets of rankings.

Conclusions

The scoring functions tested here fared only moderately well in predicting the binding affinities of a diverse set of 205 protein–ligand complexes. Their performance on subsets of mutually similar complexes was very mixed. Several of the scoring functions did well on the set of 35 serine proteinases, but all of them failed to make useful predictions for a set of 38 aspartic

proteinases. This suggests that there is still plenty of room for improvement in scoring functions. One potentially fruitful approach would be to base future knowledge-based functions solely on the structures of high affinity complexes. It is likely that scoring functions focused on specific classes of protein and ligand will sometimes be useful.

We have examined two different consensus scoring methods. One is based on averaging the ranks given by the five functions and the other on a linear combination of their scores. Each of these generally performed as well as, or slightly better than, the best individual scoring function on any given dataset.

Our results suggest that different test datasets vary considerably in difficulty, and this makes it hard to compare the results of different studies in the literature. We believe that the community should adopt standard test datasets for affinity prediction, for docking and for virtual screening.

Acknowledgements

We thank Unilever PLC, the EPSRC (Grant GR/R24753) and the Isaac Newton Trust for their generous sponsorship of this work.

References

1 T. Langer and R. D. Hoffmann, *Curr. Pharm. Des.*, 2001, **7**, 509–527.
2 W. P. Walters, M. T. Stahl and M. A. Murcko, *Drug Discovery Today*, 1998, **3**, 160–178.
3 R. Lahana, *Drug Discovery Today*, 1999, **4**, 447–448.
4 C. S. Ramesha, *Drug Discovery Today*, 2000, **5**, 43–44.
5 M. Stahl and M. Rarey, *J. Med. Chem.*, 2001, **44**, 1035–1042.
6 C. Peréz and A. R. Ortiz, *J. Med. Chem.*, 2001, **44**, 3768–3785.
7 P. R. Andrews, in *3-D QSAR in Drug Design. Volume 1: Theory, Methods and Applications*, H. Kubinyi, ed., Kluwer, Dordrecht, 1993, pp. 13–40.
8 H.-J. Böhm, *J. Comput.-Aided Mol. Des.*, 1994, **8**, 243–256.
9 R. X. Wang, L. Liu, L. H. Lai and Y. Q. Tang, *J. Mol. Model.*, 1998, **4**, 379–394.
10 R. Wang, L. Luhua and S. Wang, *J. Comput.-Aided Mol. Des.*, 2002, **16**, 11–26.
11 H.-J. Böhm, *J. Comput.-Aided Mol. Des.*, 1998, **12**, 309–323.
12 H. Gohlke, M. Hendlich and G. Klebe, *J. Mol. Biol.*, 2000, **295**, 337–356.
13 A. V. Ishchenko and E. I. Shakhnovich, *J. Med. Chem.*, 2002, **45**, 2770–2780.
14 J. B. O. Mitchell, R. A. Laskowski, A. Alex and J. M. Thornton, *J. Comput. Chem.*, 1999, **20**, 1165–1176.
15 J. B. O. Mitchell, R. A. Laskowski, A. Alex, M. J. Forster and J. M. Thornton, *J. Comput. Chem.*, 1999, **20**, 1177–1185.
16 I. Muegge and Y. C. Martin, *J. Med. Chem.*, 1999, **42**, 791–804.
17 I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Landridge and T. E. Ferrin, *J. Mol. Biol.*, 1982, **161**, 269–288.
18 B. K. Shoichet, R. M. Stroud, D. V. Santi, I. D. Kuntz and K. M. Perry, *Science*, 1993, **259**, 1445–1450.
19 I. D. Kuntz, *Science*, 1992, **257**, 1078–1082.
20 R. M. A. Knegtel, I. D. Kuntz and C. M. Oshiro, *J. Mol. Biol.*, 1997, **266**, 424–440.
21 M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee, *J. Comput.-Aided Mol. Des.*, 1997, **11**, 425–445.
22 G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *J. Mol. Biol.*, 1997, **267**, 727–748.
23 R. Wang, Y. Lu and S. Wang, *J. Med.Chem.*, 2003, **46**, 2287–2303.
24 C. Bissantz, G. Folkers and D. Rognan, *J. Med. Chem.*, 2000, **43**, 4759–4767.
25 P. S. Charifson, J. J. Corkery, M. A. Murcko and W. P. Walters, *J. Med. Chem.*, 1999, **42**, 5100–5109.
26 H. Gohlke and G. Klebe, *J. Med. Chem.*, 2002, **45**, 4153–4170.
27 http://www-mitchell.ch.cam.ac.uk/datasets.html.
28 I. Muegge, *J. Comput. Chem.*, 2001, **22**, 418–425.
29 C. A. Baxter, C. W. Murray, D. E. Clark, D. R. Westhead and M. D. Eldridge, *Proteins: Struct., Funct., Genet.*, 1998, **33**, 367–382.
30 I. D. Kuntz, K. Chen, K. A. Sharp and P. A. Kollman, *Proc. Natl. Acad. Sci. USA*, 1999, **96**, 9997–10002.
31 M. D. Miller, S. K. Kearsley, D. J. Underwood and R. P. Sheridan, *J. Comput.-Aided Mol. Des.*, 1994, **8**, 153–174.
32 P. Ferrara, H. Gohlke, D. J. Price, G. Klebe and C. L. Brooks, *J. Med. Chem.*, 2004, **47**, 3032–3047.
33 Y. Pan, N. Huang, S. Cho and A. D. MacKerell, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 267–272.
34 R. Wang, X. Fang, Y. Lu and S. Wang, *J. Med. Chem.*, 2004, **47**, 2977–2980.
35 G. M. Verkhivker and P. A. Rejto, *Proc. Natl. Acad. Sci. USA*, 1996, **93**, 60–64.
36 R. P. Sheridan and S. K. Kearsley, *Drug Discovery Today*, 2002, **7**, 903–911.
37 P. Bamborough and F. E. Cohen, *Curr. Opin. Struct. Biol.*, 1996, **2**, 236–241.
38 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
39 http://www.rcsb.org/pdb/.
40 *SYBYL*, version 6.9, Tripos Inc., St. Louis, MO, USA; http://www.tripos.com/.
41 I. K. McDonald and J. M. Thornton, *J. Mol. Biol.*, 1994, **238**, 777–793.