An all atom energy based computational protocol for predicting binding affinities of protein–ligand complexes

Tarun Jain, B. Jayaram*

Department of Chemistry and Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology, Hauz Khas, New Delhi 110016, India

Received 1 September 2005; revised 10 October 2005; accepted 17 October 2005

Available online 4 November 2005

Edited by Robert B. Russell

Abstract We report here a computationally fast protocol for predicting binding affinities of non-metallo protein–ligand complexes. The protocol builds in an all atom energy based empirical scoring function comprising electrostatics, van der Waals, hydrophobicity and loss of conformational entropy of protein side chains upon ligand binding. The method is designed to ensure transferability across diverse systems and has been validated on a heterogenous dataset of 161 complexes consisting of 55 unique protein targets. The scoring function trained on a dataset of 61 complexes yielded a correlation of $r = 0.92$ for the predicted binding free energies against the experimental binding affinities. Model validation and parameter analysis studies ensure the predictive ability of the scoring function. When tested on the remaining 100 protein–ligand complexes a correlation of $r = 0.92$ was recovered. The high correlation obtained underscores the potential applicability of the methodology in drug design endeavors. The scoring function has been web enabled at www.scfbio-iitd.res.in/software/drugdesign/bappl.jsp as binding affinity prediction of protein–ligand (BAPPL) server.

Keywords: Scoring function; Binding affinity; Protein–ligand complexes; Structure-based drug design; Drug design in silico

1. Introduction

Two aspects determine the success of computer-aided structure-based ligand design [1–3] endeavors targeted to proteins: the generation of reasonable ligand-binding modes through sampling the vast conformational space, namely the docking problem [4] and the identification of those binding modes that correspond best to the experimentally given situation based on reasonable estimates of binding affinities namely the scoring problem [5].

Computational approaches which utilize the receptor structure information for estimating binding affinities [6–8] can be classified into four major classes with respect to their methodological background: (1) molecular simulation based approaches [9,10], (2) empirical/force field/additivity based approaches [11–13], (3) knowledge based approaches [14] and (4) hybrid approaches. The basic idea behind molecular simulation based approaches derives from statistical mechanics [15]. Explicit atomic level consideration of solvent molecules, ions and flexibility of both the receptor and the ligand makes these approaches compute intensive and limits the usage of simulation strategies in screening large numbers of ligands against a protein target. Additivity based approaches have given birth to the field of scoring functions. The various scoring functions have been summarized in Table 1. Force field based scoring functions approximate the binding free energy of protein–ligand complexes by a sum of van der Waals, electrostatics and other contributions. The basic assumption underlying empirical/force field/additivity based approaches is that different contributions to free energy of binding can be calculated separately and that they are additive [16]. Knowledge based approaches draw upon statistical analyses of a large number of protein–ligand complexes present in the structural repositories [17]. Based upon the current trend in virtual screening methodologies some requirements for a good scoring function are: structure prediction, affinity prediction, virtual screening and speed [18]. To circumvent certain imperfections of current scoring functions, consensus scoring [19] has been introduced which combines information from different scoring functions to balance errors in single scores. A combination of molecular simulation and additivity approximation based approaches called hybrid methods are also in vogue to obtain the free energy estimates of protein–ligand association. These mainly include the linear interaction energy method (LIE) [20], MMPBSA [21] and MMGBSA [22,23]. These approaches have been developed to estimate binding free energies rather quickly but with reasonable accuracy. Databases like LPDB [24] and PLD [25] providing experimental binding affinities have proven to be extremely valuable for the development and validation of scoring functions.

Comparative evaluations of different docking programs in combination with various scoring functions for their applications in virtual screening have been carried out recently [26,27] and results show that many of the popular scoring functions are able to select out correct docked from misdocked structures, but correlation with the experimental binding affinity still remains a major limiting factor in virtual screening for drug discovery.

In this study, we report a computational protocol which incorporates an all atom energy based empirical scoring function for the prediction of binding affinities of non-metallo

*Corresponding author. Fax: +91 11 2658 2037.
E-mail address: bjayaram@chemistry.iitd.ac.in (B. Jayaram).

URL: www.scfbio-iitd.res.in

0014-5793/$30.00 © 2005 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2005.10.031

Table 1
Some popular scoring functions for estimating binding affinities of protein–ligand complexes

S. no.	Scoring function	Method	Training set	Test set	r
1	DOCK [51]	Force field	–	–	–
2	EUDOC [52]	Force field	–	–	–
3	CHARMm [53]	Force field	–	–	–
4	AutoDock [54]	Force field	–	–	–
5	DrugScore [55]	Knowledge	–	–	–
6	SMoG [56]	Knowledge	–	36	0.79
7	BLEEP [57]	Knowledge	–	90	0.74
8	PMF [58]	Knowledge	–	77	0.78
9	DFIRE [59]	Knowledge	–	100	0.63
10	SCORE [60]	Empirical	170	11	0.81
11	GOLD [61]	Empirical	–	–	–
12	LUDI [62]	Empirical	82	12	0.83
13	FlexX [63]	Empirical	–	–	–
14	ChemScore [64]	Empirical	82	20	0.84
15	VALIDATE [65]	Empirical	51	14	0.90
16	Ligscore [66]	Empirical	50	32	0.87
17	X-CSCORE [67]	Empirical (consensus)	200	30	0.77
18	GLIDE [68]	Force field/empirical	–	–	–
19	Present work	Force field/empirical	61	100	0.92

The method used, the number of complexes considered in the dataset for training and testing and the final correlation coefficient (r) obtained on the test set against experimental binding affinities.

protein–ligand complexes. The scoring function presented is validated on a heterogenous dataset of 161 protein–ligand complexes comprising 55 unique proteins and is fast enough to be used in virtual screening protocols.

2. Theory and methodology

The scoring function employed considers the non-bonded energy of a protein–ligand complex as a sum of three energy terms – electrostatic, van der Waals and hydrophobic, termed here as Model I

$$E = \sum E_{\text{el}} + E_{\text{vdw}} + E_{\text{hpb}}. \tag{I}$$

Here, E is the total non-bonded energy, E_{el} is the electrostatic contribution to the energy, E_{vdw} is the van der Waal term, E_{hpb} is the hydrophobic contribution and the summation runs over all the atoms of the protein–ligand complex. Details of the function and individual terms are provided elsewhere ([28–32] and references therein). In a nutshell, the electrostatic contribution to the interaction energy is computed via Coulomb's law with a sigmoidal dielectric function. The van der Waals interactions are modeled using a (12, 6) Lennard-Jones potential between the atoms of the protein and ligand. The hydrophobic interactions are captured via the Gurney parameter approach, which provides a computationally simple means for treating desolvation effects. The energy function described above enables evaluation of the total non-bonded interaction energy of a protein–ligand complex in aqueous environment from the Cartesian coordinates of all the atoms. We have previously examined and found this scoring function to yield satisfactory energetics on base pairs of DNA [28], alpha helices [29], ion atmosphere around DNA [33]. Recently, we have used the same function in protein structure prediction studies where the function is able to distinguish native from the decoys [34]. In the DNA-drug studies, the function has shown an excellent correlation ($r^2 = 0.95$) with the experimental ΔT_{m} values [35].

2.1. Dataset description

There are about 3500 proteins, complexed with ligands, substrate, prosthetic groups and metal ions in the protein databank (RCSB) [16]. For the present study, we focused on non-metallo protein–ligand complexes and prepared a dataset of 161 complexes (Table I(A) and (B) Supplementary information) as described in the dataset preparation section. The experimental binding free energies for these complexes are available in the public domain databases like LPDB [24] and PLD [25]. A description of the dataset with observed limits of the various descriptors/physicochemical properties is given in Table 2. The dataset contains 55 unique proteins like trypsin, HIV-I protease, alpha thrombin, DHFR, etc., bracketing a variety of forms and functions. Table 2 shows that the dataset in consideration is heterogeneous enough with respect to the ligand, protein and complex descriptors/physicochemical properties to facilitate a rigorous evaluation of the performance of the proposed protocol and its extensions to other systems.

Table 2
Some physicochemical properties with their observed limits in the 161 protein–ligand complex dataset

S. no.	Descriptor/physicochemical property	Limits
Ligand		
1	Number of rotatable bonds	0–32
2	Hydrogen bond donors	0–18
3	Hydrogen bond acceptors	0–26
4	Ligand net charge	(−)5–(+)1
5	$C\log P$ [69]	(−)11–(+)10
6	Molecular weight	95–800
7	Number of heavy atoms	5–62
Protein		
8	Number of unique proteins	55
9	Number of residues	105–833
Complex		
10	Experimental binding affinity (kcal/mol)	(−)15.57–(−)2.03
11	Net charge on the complex	(−)28–(+)11
12	Resolution (Å)	1.25–3.16

2.2. Dataset preparation

Fig. 1 describes a general protocol for the preparation of a non-metallo protein–ligand complex in a force field compatible manner. The protocol is divided into the following steps:

1. *Selection of the complex:* The X-ray coordinates of the complex are extracted from the RCSB and crystallographic water molecules are removed.
2. *Parameterization of the ligand:* The ligand coordinates are extracted from the protein–ligand complex. Hydrogen atoms are added keeping the same ionization states of the atoms as given in the corresponding literature for each complex. The ligand is then AM1 geometry optimized followed by HF/6-31G* ab initio level calculations to obtain the electrostatic potential of the ligand using GAMESS [36]. RESP fitting [37] is then applied on the electrostatic potentials to derive the equivalent partial atomic charges for the ligand. GAFF force field [38] is used to assign the atom types, bond, angle, dihedral and van der Waals parameters for the ligand.
3. *Parameterization of the protein:* Hydrogen atoms are added and the protonation states of the charged residues inside the active site of the protein are adapted as mentioned in the literature for each complex. Assignment of force field parameters for protein atoms is done using the Cornell et al. [39] force field.
4. *Energy minimization of the complex:* The protein–ligand complex is energy minimized in vacuum and with explicit solvent molecules separately using AMBER [40] to remove any clashes from the structure. For vacuum minimizations, 1000 steps of steepest descent and 1500 steps of conjugate gradient are carried out. Water minimization is performed using a truncated octahedron type solvate box with an 8.0 Å cutoff. Minimization with explicit solvent is performed with first restraining the solute and minimizing only the waters so as to relax any kind of gaps present in them. The minimization here involves 500 steps of steepest descent and 500 steps of conjugate gradient. After the solvent is relaxed, an all atom 2500 steps minimization similar to the vacuum minimization is performed.

A protein–ligand complex prepared in the above manner acts as an input for the binding affinity estimates.

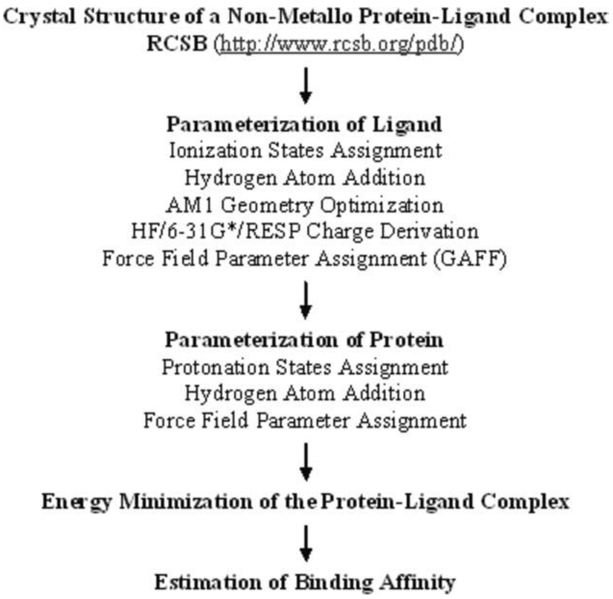
Crystal Structure of a Non-Metallo Protein-Ligand Complex
RCSB (http://www.rcsb.org/pdb/)
Parameterization of Ligand
Ionization States Assignment
Hydrogen Atom Addition
AM1 Geometry Optimization
HF/6-31G*/RESP Charge Derivation
Force Field Parameter Assignment (GAFF)
Parameterization of Protein
Protonation States Assignment
Hydrogen Atom Addition
Force Field Parameter Assignment
Energy Minimization of the Protein-Ligand Complex
Estimation of Binding Affinity

Fig. 1. A computational flowchart adopted for computing binding affinities of protein–ligand complexes.

3. Results and discussion

The calculated protein–ligand interaction energies using Model I for all the 161 complexes (vacuum minimized) are correlated with the experimental binding free energies as shown (Fig. 2). The correlation coefficient r is 0.85 and the RMS error is ±1.71 kcal/mol. Use of explicit solvent during energy minimization gives an $r = 0.84$ for the 161 dataset. Inclusion of explicit solvent does not appear to affect the overall correlation. However, it greatly increases the computational time involved in minimization restricting its applicability in virtual screening programs in structure-based drug design.

Within the framework of the protocol proposed, we attempted to improve the correlation by adopting a more detailed solvation treatment using the Eisenberg–Mclachlan approach [41]. Eisenberg–Mclachlan model has only a limited set of five basic atom types found in proteins, whereas small drug molecules have a variety of atom types and defining them with a limited set would not account for their diversity. Also, the atomic solvation parameters (ASP) were derived using water/octanol partition coefficients of 20 amino acids, which poses a very limited potential for the transferability of these parameters in calculating ligand binding free energies. To circumvent the abovementioned problems, we made two modifications to the approach. In the first step, we have combined the atom types in Cornell et al. [39] force field for proteins/nucleic acids with the atom types in GAFF [38] force field for small molecules. This gives us a common set of 22 atom types (Table 3) with the advantage that, any atom of protein or ligand can be defined using this set ensuring transferability of derived parameters for organic and biological molecules. The second modification involves considering the loss in surface area of individual atoms upon binding instead of taking their surface areas, reflecting the changes in binding process. The solvent accessible surface area of the protein, ligand and the complex is calculated using the Lee and Richard's algorithm [42] with a probe radius of 1.4 Å and is further divided into the surface

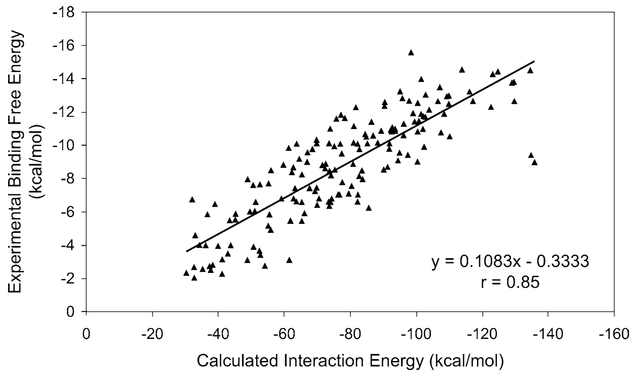

Fig. 2. Correlation between the calculated interaction energies (Model I) and experimental binding free energies for 161 protein–ligand complexes.

Table 3
A description of the 22 derived atom types with their atomic desolvation parameters kcal/mol/Å^2 (ADP)

S. no.	Atom type Symbol	Description	Parameters
1	C1	sp^2 carbonyl	0.1209
2	C2	sp carbon	0.2522
3	C3	sp^2 carbon aliphatic	−0.0283
4	C4	sp^2 carbon aromatic	0.0141
5	C5	sp^3 carbon	0.1276
6	HL	Halogens (Fl, Cl, Br, I)	0.0081
7	H1	Hydrogen bonded to aliphatic carbon	0.0005
8	H2	Hydrogen bonded to aromatic carbon	−0.0040
9	H3	Hydrogen bonded to nitrogen	0.0051
10	H4	Hydroxyl group	0.0013
11	H5	Hydrogen bonded to sulfur	0.0595
12	N1	sp^2 nitrogen in amide groups	−0.0232
13	N2	sp^2 nitrogen in aliphatic systems	−0.0311
14	N3	sp^2 nitrogen in aromatic systems	−0.0111
15	N4	sp nitrogen	0.0037
16	N5	sp^3 nitrogen	−0.0478
17	N6	Amine nitrogen connected to one or more aromatic rings	0.0077
18	O1	Oxygen with one connected atom	−0.0074
19	O2	Oxygen in hydroxyl group	−0.0094
20	O3	Ether and ester oxygen	−0.0147
21	P	Phosphate	0.7097
22	S	Sulfur	0.0109
23	α	Empirical coefficient for electrostatics	0.1049
24	β	Empirical coefficient for van der Waals	0.1281
25	λ	Empirical coefficient for conformational entropy	−0.5385
26	δ	Constant (Intercept)	0.2060

Empirical coefficients for electrostatics (α), van der Waals (β), conformational entropy (λ) and the regression constant (δ) are also given.

areas of individual atoms. van der Waals radii of each atom has been adjusted according to the modified Generalized Born model [43], which is consistent with the Cornell et al. force field. Total surface area of an atom type is obtained by summing up all the contributions from that atom type. The net loss in surface area of an atom type upon binding is computed as

$$\Delta A_{\mathrm{LSA}} = \sum A_{\mathrm{complex}} - \sum A_{\mathrm{protein}} - \sum A_{\mathrm{ligand}},$$

where ΔA_{LSA} is the net loss in surface area of an atom type A. A_{complex}, A_{protein} and A_{ligand} are the total surface areas of atom type A in complex, protein and ligand, respectively.

There is an entropic cost associated with any bimolecular interaction that is a consequence of degrees of freedom of motion lost when two molecules are rigidly constrained within a complex [44]. An important contributor to the energetics of protein folding and protein–ligand binding is the loss of conformational entropy (ΔS_{CR}) of the protein side chains [45]. We have utilized here an empirical scale of side chain conformational entropy developed by Pickett and Sternberg [46]. In this procedure relative accessibility (RA) is used as a measure to determine whether an amino acid side chain is sampling different rotamers or is buried in the folded state. If the RA > 60% then the side chain is free to sample all the conformations and there is no loss of conformational entropy upon folding. If the RA < 60% then the side chain is buried in the folded state and there is an entropic penalty upon folding, where

$$\mathrm{RA}_{\mathrm{folding}} = \frac{\text{calculated accessible surface area of side chain (folded)}}{\text{surface area of that side chain in extended state (unfolded)}}.$$

We have employed RA as a measure to determine the loss of conformational entropy of protein side chains in protein–ligand binding and defined it as

$$\mathrm{RA}_{\mathrm{binding}} = \frac{\text{calculated accessible surface area of side chain in bound form}}{\text{calculated accessible surface area of side chain in unbound form}}.$$

A folded protein is equivalent to an unbound form. Side chains with $\mathrm{RA}_{\mathrm{folding}}$ > 60% and $\mathrm{RA}_{\mathrm{binding}}$ < 60% are considered to have a loss of conformational entropy. The values from the empirical scale [46] for all such residues are added to get a final estimate of the conformational entropy (ΔS_{CR}) loss upon binding.

Following this approach, our empirical free energy function takes the following form (Model II):

$$\Delta G = \alpha(E_{\mathrm{el}}) + \beta(E_{\mathrm{vdw}}) + \sum_{A=1}^{22} \sigma_A \Delta A_{\mathrm{LSA}} + \lambda(\Delta S_{\mathrm{CR}}) + \delta, \qquad \text{(II)}$$

where ΔG is the binding free energy in kcal/mol, E_{el} and E_{vdw} have been defined previously. ΔA_{LSA} is the loss in surface area of the atom type A. We define σ_A as the atomic desolvation parameter (ADP) in kcal/mol/Å^2 for an atom type A. ΔS_{CR} is the loss in conformational entropy of protein side chains upon binding. α, β and λ are the empirical coefficients for electrostatics, van der Waals and conformational entropy respectively and δ is a constant. Model fitting is performed using multiple linear regression to obtain the empirical parameters for Model II. E_{el}, E_{vdw}, ΔA_{LSA} and ΔS_{CR} serve as independent variables and experimental binding free energies (ΔG) serve as dependent variables.

3.1. Model validation

Model validation is a crucial aspect of any model development technique and establishes the predictive power of the model. Recent studies [47,48] have shown that, in addition to leave-one-out (LOO) cross-validation (q^2) procedure, validation of the model using an external test set of compounds is necessary. For a robust validation, the training and test sets must have a uniform distribution of the representative points in the multidimensional descriptor space. In addition the model should also satisfy the following conditions:

1. $q^2 > 0.5$;
2. $R^2 > 0.6$;
3. $\frac{(R^2-R_0^2)}{R^2} < 0.1$ and $0.85 \leqslant K \leqslant 1.15$;
4. $\frac{(R^2-R_0'^2)}{R^2} < 0.1$ and $0.85 \leqslant K' \leqslant 1.15$;
5. $|R_0^2 - R_0'^2| < 0.3$.

All the above terms have been explained in Table II of the Supplementary information.

Keeping these issues in consideration, we started with the leave-one-out cross-validation procedure to make the training and test sets. We used the experimental binding free energy of the complexes as a descriptor for their uniform distribution across multidimensional descriptor space in the training and test sets. A training set of 61 protein–ligand complexes was obtained giving a correlation coefficient $r = 0.92$ for the predicted binding affinities against the experimental binding affinities (Fig. 3). A graphical residual analysis plot (Fig. S(I) of the Supplementary information) of the standardized residuals against the predicted binding affinities for the training set shows a uniform distribution of the points above and below the base line, suggesting that the model fits the data well. The five statistical tests defined above in addition to S_{PRESS} and RMS error were then performed on the training set. The results shown in Table 4 indicate that the model passes all the validation tests. The final validation was performed on the external test set of 100 protein–ligand complexes using the parameters obtained from the training set (Table 3). A correlation coefficient of $r = 0.92$ was obtained on the test set (Fig. 4) between the experimental binding free energies and predicted binding free energies, indicating the robustness of the model and the parameters obtained in predicting the binding affinities of protein–ligand complexes. We further tested the ability of the scoring function in the prediction of relative binding affinities of a series of ligands against the same protein target. From the 100 test set, we selected Alpha Thrombin and HIV-I protease which have more than six distinct ligands. Individual correlation studies on these groups of complexes (Fig. 5A and B) show an average correlation coefficient of $r = 0.86$. The training and the test set PDB IDs of the complexes along with their experimental and predicted binding free energies and component-wise separation of the energetics are provided in Table I(A) and (B) of the Supplementary information.

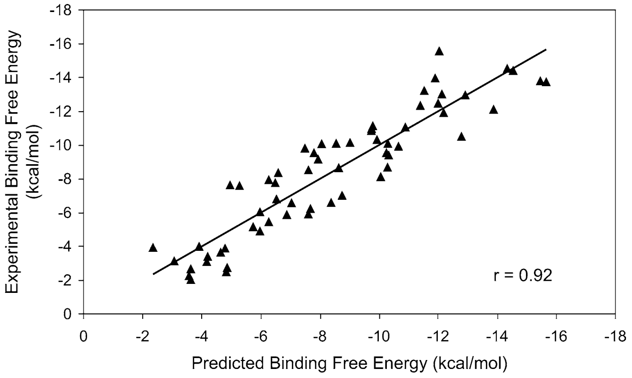

Fig. 3. Correlation between the predicted binding free energies (Model II) and experimental binding free energy for 61 protein–ligand complex training set.

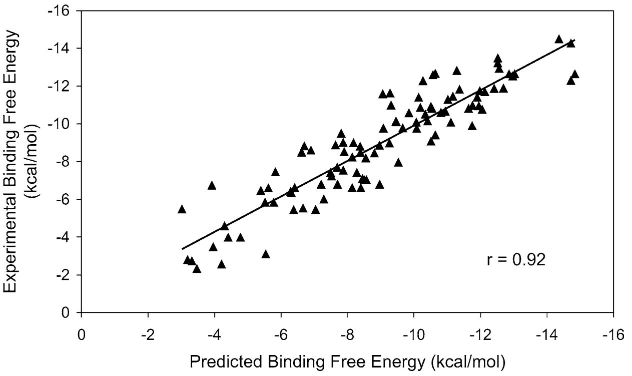

Fig. 4. Correlation between the predicted binding free energies (Model II) and experimental binding free energy for the 100 protein–ligand complex test set.

3.2. Parameter analysis

The empirical scoring function proposed in Model II has 25 independent variables (electrostatics, van der Waals, loss in conformational entropy and 22 atom types for hydrophobicity corresponding to a combined GAFF [38] and AMBER force field [39]) and therefore 25 empirical parameters (Table 3). Fig. S(II) (Supplementary information) gives a percentage wise occurrence of each variable in the 161 dataset. The figure shows that every complex has a net favorable electrostatic and van der Waals contribution towards binding. 24% of the complexes show a loss in conformational entropy of protein side chains upon ligand binding. Of the 22 atom types C1, C4, C5, H1, H2, H3, H4, N1, O1 and O2 occur in more than 90% of the complexes. S, O3, N5, N3 and N4 are present in less than 50% of the complexes. C2, halogens (F, Cl) and P are present in very few complexes (less than 10%).

To assess the effect of each empirical parameter (Table 3) on the scoring function (Model II), we performed a sensitivity analysis of the 25 empirical parameters. Based upon one-factor-at-a-time (OAT) methods of local sensitivity analysis [49], we varied all the parameters one at a time in the range of −0.8 to +0.8, with an increment of 0.0001, keeping the rest of the parameters fixed. The correlation (r) between the experimental binding free energies against the predicted free energies is calculated for all the 161 complexes in the dataset with each increment (Fig. S(III) Supplementary information). Using Fig. S(III), we classify the parameters into three categories;

Table 4
Statistical tests and their respective values for the training set

Statistical test	q^2	R^2	$\frac{(R^2-R_0^2)}{R^2}$	$\frac{(R^2-R_0'^2)}{R^2}$	K	K'	$\lvert R_0^2 - R_0'^2\rvert$	S_{PRESS} (kcal/mol)	RMS error (kcal/mol)
Value	0.85	0.85	−0.18	−0.17	1.00	0.98	0.0038	1.88	±1.43

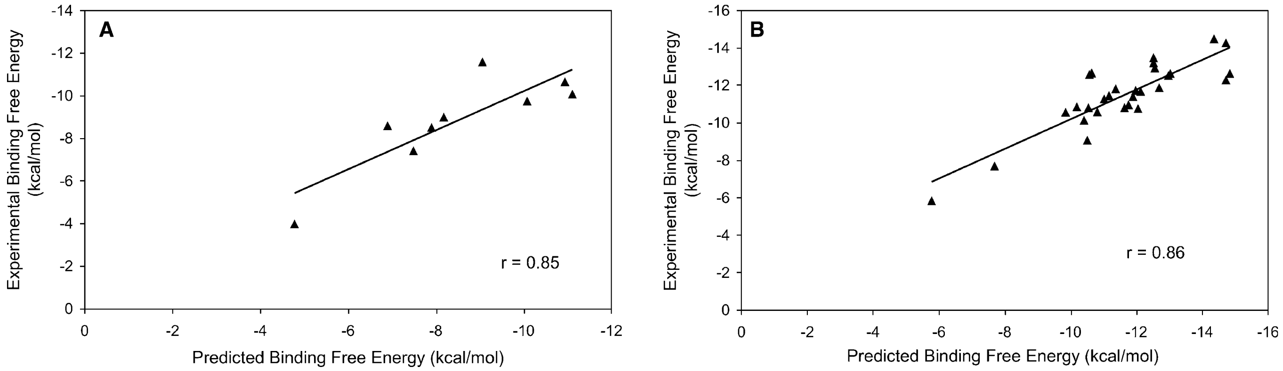

Fig. 5. Correlation for the relative binding affinities of series of ligands against (A) Alpha Thrombin and (B) HIV-I Protease.

(1) highly sensitive: parameters which have a high impact on correlation; α, β, C4, HL, H1, H2, H3, H4, N3, N6, O1, O2, S (2) sensitive: parameters which have a moderate effect on the correlation; C1, C3, N1, N2, N4, N5, O3 and (3) less sensitive: parameters which have a very less effect on the correlation; λ, C2, C5, H5, P.

Although the parameters are empirical (Table 3), the model is phenomenological and is in accord with the thermodynamics of protein–ligand binding. E_{el} and E_{vdw} (namely the electrostatics and van der Waals components of interaction energy) have negative signs and their empirical parameters α and β have positive signs, respectively, demonstrating a net favorable contribution towards binding. ΔS_{CR} has positive sign and its parameter has negative, reflecting that the net loss in conformational entropy of protein side chains is an unfavorable component towards binding. The loss in surface area of all the 22 atom types has a negative sign, indicating that the net loss in surface area is favorable for binding. However, the atomic desolvation parameters for these atom types have different contributions. Carbons, sulfur and phosphorous have positive desolvation parameters, which shows that desolvation of non-polar atoms is favorable for binding. Oxygens and nitrogens have a negative sign on the empirical parameters, indicating that desolvation of polar atoms is unfavorable for binding. Desolvation parameters for halogens and hydrogens have a positive sign, suggesting that their desolvation is favorable for binding.

We performed a component-wise analysis of the binding free energy for the 161 complex dataset (Fig. S(IV) Supplementary information). The figure shows the contribution of each component and their additive sums and their effect on the correlation. van der Waals turns out to be the largest contribution for protein–ligand binding, contributing 0.79 to correlation. This suggests that structural complementarity/packing in particular is an absolute prerequisite for specific binding. Adding electrostatics contribution to van der Waals component further increases the correlation to 0.86, suggesting the importance of hydrogen bonding/ionic interactions in providing specificity to the complex formation. It also suggests the importance of assigning accurate charges for the ligand and protein atoms. Adding hydrophobicity contribution to this further increases the correlation to 0.91, reflecting the importance of solvent in protein–ligand binding. Adding the loss in conformational entropy increments the correlation to 0.92 reflecting the contribution of loss of protein side chain conformation upon ligand binding.

In this study, a computationally tractable protocol using an empirical all atom energy based scoring function (Model II) is presented and its performance in predicting the binding affinities of protein–ligand complexes is apprised. The empirical free energy function comprises contributions from electrostatics with a sigmoidal dielectric function, van der Waals, hydrophobic and loss in conformational entropy of protein side chains. Model validation results prove that the proposed empirical energy function can be easily used for prediction studies. The methodology is sufficiently fast for usage in virtual screening protocols. The results suggest that, partial atomic charges for ligand, correct protonation states for ligand and protein residues in the active site, compatibility between the parameters obtained from GAFF force field for ligand and AMBER force field for proteins, the dielectric function employed, the desolvation parameters for each atom type and energy minimization protocol are some of the important issues which have strengthened the empirical scoring function in obtaining a good correlation between the experimental and the predicted binding affinities of protein–ligand complexes from "single-point" calculations. Heterogenity of the dataset on which the protocol has been validated and parameters obtained promises transferability to protein–ligand systems from different families of proteins, with different active sites and a variety of ligand architectures. A high correlation coefficient ($r = 0.92$) in comparison with other scoring functions in Table 1 suggests that the protocol possesses reasonable accuracy in the prediction of ligand binding affinities against protein targets. An average correlation coefficient of $r = 0.86$ for different ligands against the same targets indicates the ability of the protocol and scoring function to predict relative binding affinities of ligands. The method could be trained for a specific target with improved correlation but at the expense of transferability to other targets. While these results may be by far the best obtained on a large dataset with an atomic level energy based scoring function not customized to any particular system, further improvements are essential for keeping the errors low in the estimated binding affinities. A closer examination of the missing components in the scoring function in relation to binding free energies is the role of explicit waters in the active site besides thermal averaging. Future work would involve extensions of the protocol for predicting binding affinities involving metallo-proteins, where charges of the atoms around the ion in the active site play a critical role.

The empirical energy based scoring function (Model II) has been web enabled at www.scfbio-iitd.res.in/software/drugdesign/bappl.jsp as binding affinity prediction of protein-ligand (BAPPL) server. The server provides two methods as options.

In Method 1, the input is an energy-minimized protein–ligand complex with hydrogens added, protonation states, partial atomic charges and van der Waals parameters assigned. The application then computes the binding free energy of the complex using the specified parameters. In Method 2, the input is an energy-minimized protein–ligand complex with hydrogens added and protonation states assigned. The net charge on the ligand needs to be specified. The application derives the partial atomic charges of the ligand using the AM1-BCC procedure [50] and GAFF force field [38] for van der Waals parameters. Cornell et al. [39] force field is used to assign the force field parameters for proteins. Binding free energy is estimated as in Model II and reported. Although the empirical scoring function has been calibrated using the HF/6-31G*/RESP equivalent partial atomic charges, we have provided the AM1-BCC procedure [49] for deriving partial atomic charges of ligands for Method 2 because this procedure is fast and yields a correlation of $r = 0.91$ on the 161 complex dataset. The coordinates along with all the parameters for binding affinity estimates prepared as described in Fig. 1 are also made accessible at the website at www.scfbio-iitd.res.in/software/drugdesign/proteinliganddataset.htm.

Acknowledgments: Funding from the Department of Biotechnology, Government of India is acknowledged. Help received from Anuj Gupta, Vidhu S. Pandey, Saher Afshan Sheikh and Praveen Aggarwal in preparing the manuscript and the web tool is acknowledged.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2005.10.031.

References

[1] Kuntz, I.D. (1992) Structure-based strategies for drug design and discovery. Science 257, 1073–1082.

[2] Latha, N. and Jayaram, B. (2005) A binding affinity based computational pathway for active-site directed lead design software. Drug Des. Rev. Online 2, 145–165.

[3] Jorgensen, W.L. (2004) The many roles of computation in drug discovery. Science 303, 1813–1818.

[4] Brooijmans, N. and Kuntz, I.D. (2003) Molecular recognition and docking algorithms. Ann. Rev. Biophys. Biomol. Struct. 32, 335–373.

[5] Kitchen, D.B., Decornez, H., Furr, J.R. and Bajorath, J. (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nature Rev. Drug Discov. 3, 935–949.

[6] Ajay and Murcko, M.A. (1995) Computational methods to predict binding free energy in ligand-receptor complexes. J. Med. Chem. 38, 4953–4967.

[7] Gohlke, H. and Klebe, G. (2002) Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptor. Angew. Chem., Int. Ed. 41, 2644–2676.

[8] Reddy, M.R. and Erion, M.D. (2001) Free Energy Calculations in Rational Drug Design, Springer, Germany.

[9] Brandsdal, B.O., Osterberg, F., Almlof, M., Feierberg, I., Luzhkov, V.B. and Åqvist, J. (2003) Free energy calculations and ligand binding. Adv. Protein Chem. 66, 123–158.

[10] Wang, W., Donini, O., Reyes, C.M. and Kollman, P.A. (2001) Biomolecular simulations: recent developments in force fields, simulation of enzyme catalysis, protein–ligand, protein–protein and protein–nucleic acid noncovalent interactions. Annu. Rev. Biophys. Biomol. Struct. 30, 211–243.

[11] Williams, D.H., Cox, J.P.L., Doig, A.J., Gardner, M. and Gerhard, U., et al. (1991) Toward the semiquantitative estimation of binding constants. Guides for peptide–peptide binding in aqueous solution. J. Am. Chem. Soc. 113, 7020–7030.

[12] Novotny, J., Bruccoleri, R.E. and Saul, F.A. (1989) On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. Biochemistry 28, 4735–4749.

[13] Vajda, S., Weng, Z., Rosenfeld, R. and DeLisi, C. (1994) Effect of conformational flexibility and solvation on receptor-ligand binding free energies. Biochemistry 33, 13977–13988.

[14] Gohlke, H. and Klebe, G. (2001) Statistical potential and scoring functions applied to protein–ligand binding. Curr. Opin. Struct. Biol. 11, 231–235.

[15] Gilson, M.K., Given, J.A., Bush, B.L. and McCammon, J.A. (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review. Biophys. J. 72, 1047–1069.

[16] Dill, K.A. (1997) Additivity principles in biochemistry. J. Biol. Chem. 272, 701–704.

[17] Berman, H.M., Westbrook, J., Feng, Z., Gilliand, G. and Bhat, T.N., et al. (2000) The protein data bank. Nucleic Acids Res. 28, 235–242.

[18] Schulz-Gasch, T. and Stahl, M. (2004) Scoring functions for protein–ligand interactions: a critical perspective. Drug Discov. Today: Technol. 1, 231–239.

[19] Charifson, P.S., Corkery, J.J., Murko, M.A. and Walters, W.P. (1999) Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J. Med. Chem. 42, 5100–5109.

[20] Åqvist, J., Luzhkov, V.B. and Brandsdal, B.O. (2002) Ligand binding affinities from MD simulations. Acc. Chem. Res. 35, 358–365.

[21] Simonson, T., Archontis, G. and Karplus, M. (2002) Free energy simulations come of age: protein–ligand recognition. Acc. Chem. Res. 35, 430–437.

[22] Still, W.C., Tempczyk, A., Hawley, R.C. and Hendrickson, T. (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc. 112, 6127–6129.

[23] Kalra, P., Reddy, T.V. and Jayaram, B. (2001) Free energy component analysis for drug design: a case study of HIV-I protease inhibitor binding. J. Med. Chem. 44, 4325–4338.

[24] Roche, O., Kiyama, R. and Brooks III, C.L. (2001) Ligand–protein database: linking protein–ligand complex structures to binding data. J. Med. Chem. 44, 3592–3598.

[25] Puvanendrampillai, D. and Mitchell, J.B.O. (2003) Protein ligand database (PLD): additional understanding of the nature and specificity of protein–ligand complexes. Bioinformatics 19, 1856–1857.

[26] Wang, R., Lu, Y. and Wang, S. (2003) Comparative evaluation of 11 scoring functions for molecular docking. J. Med. Chem. 46, 2287–2303.

[27] Ferrara, P., Gohlke, H., Price, D.J., Klebe, G. and Brooks III, C.L. (2004) Assessing scoring functions for protein–ligand interactions. J. Med. Chem. 47, 3032–3047.

[28] Arora, N. and Jayaram, B. (1998) Energetics of base pairs in B-DNA in solution: an appraisal of potential functions and dielectric treatments. J. Phys. Chem. 102, 6139–6144.

[29] Arora, N. and Jayaram, B. (1997) Strength of hydrogen bonds in α helices. J. Comp. Chem. 18, 1245–1252.

[30] Jayaram, B. and Beveridge, D.L. (1990) Free energy of an arbitrary charge distribution imbedded in coaxial cylindrical dielectric continua: application to conformational preferences of DNA in aqueous solutions. J. Phys. Chem. 94, 4666–4671.

[31] Hodes, Z.I., Nemethy, G. and Scheraga, H.A. (1979) Model for the conformational analysis of hydrated peptides. Effect of hydration on the conformational stability of the terminally blocked residues of the 20 naturally occurring amino acids. Biopolymers 18, 1565–1610.

[32] Hopfinger, A.J. (1971) Polymer–solvent interactions for homopolypeptides in aqueous solution. Macromolecules 4, 731–737.

[33] Young, M.A., Jayaram, B. and Beveridge, D.L. (1998) Local dielectric environment of B-DNA in solution: results from a 14 ns molecular dynamics trajectory. J. Phys. Chem. 102, 7666–7669.

[34] Narang, P., Bhushan, K., Bose, S. and Jayaram, B. (submitted for publication) Protein structure evaluation using an all atom energy based empirical scoring function. J. Biomol. Struct. Dynam.

[35] Shaikh, S.A. and Jayaram, B. (submitted for publication) A computational tool for predicting DNA-drug interaction energy. Chem. Commun.

[36] Schmidt, M.W., Baldridge, K.K., Boatz, J.A., Elbert, S.T. and Gordon, M.S., et al. (1993) General atomic and molecular electronic structure system. J. Comput. Chem. 14, 1347–1363.

[37] Bayly, C.I., Cieplak, P., Cornell, W. and Kollman, P.A. (1993) A well behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J. Phys. Chem. 97, 10269–10280.

[38] Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A. and Case, D.A. (2004) Development and testing of a general amber force field. J. Comput. Chem. 25, 1157–1174.

[39] Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R. and Merz, K.M., et al. (1995) A second generation force field for the simulation of proteins, nucleic acids and organic molecules. J. Am. Chem. Soc. 117, 5179–5197.

[40] Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S. and Cheathem III, J.E., et al. (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comput. Phys. Commun. 91, 1–41.

[41] Eisenberg, D. and McLachlan, A.D. (1986) Solvation energy in protein folding and binding. Nature 319, 199–203.

[42] Lee, B.K. and Richards, F.M. (1971) The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. 55, 379–400.

[43] Jayaram, B., Sprous, D. and Beveridge, D.L. (1998) Solvation free energy of biomacromolecules: parameters for a modified generalized born model consistent with the AMBER force field. J. Phys. Chem. B 102, 9571–9576.

[44] Finkelstein, A.V. and Janin, J. (1989) The price of lost freedom: entropy of bimolecular complex formation. Protein Eng. 3, 1–3.

[45] Doig, A.J. and Sternberg, M.J.E. (1995) Side-chain conformational entropy in protein folding. Protein Sci. 4, 2247–2251.

[46] Pickett, S.D. and Sternberg, M.J.E. (1993) Empirical scale of side-chain conformational entropy in protein folding. J. Mol. Biol. 231, 825–839.

[47] Golbraikh, A. and Tropsha, A. (2002) Beware of q^2!. J. Mol. Graph. Model 20, 269–276.

[48] Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.D., Lee, K.H. and Tropsha, A. (2003) Rational selection of training and test sets for the development of validated QSAR models. J. Comput. Aided Mol. Des. 17, 241–253.

[49] Saltelli, A., Ratto, M., Tarantola, S. and Campolongo, F. (2005) Sensitivity analysis for chemical models. Chem. Rev. 105, 2811–2828.

[50] Jakalian, A., Bush, B.L., Jack, D.B. and Bayaly, C.I. (2000) Fast, efficient generation of high-quality atomic charges. J. Comput. Chem. 21, 132–146.

[51] Ewing, T.J.A., Makino, S., Skillman, A.G. and Kuntz, I.D. (2001) DOCK 4.0: Search strategies for automated molecular docking of flexible molecule database. J. Comput. Aided Mol. Des. 15, 411–428.

[52] Pang, Y.P., Perola, E., Xu, K. and Prendergast, F.G. (2001) EUDOC: A computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. J. Comp. Chem. 22, 1750–1771.

[53] Momany, F.A. and Rone, R. (1992) Validation of a general purpose QUANTA®3.2/CHARMm® force field. J. Comp. Chem. 13, 888–900.

[54] Morris, G.M. et al. (1998) Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. J. Comp. Chem. 19, 1639–1662.

[55] Gohlke, H., Hendlich, M. and Klebe, G. (2000) Knowledge-based scoring function to predict protein–ligand interactions. J. Mol. Biol. 295, 337–356.

[56] DeWitte, R.S. and Shakhnovich, E.I. (1996) SMoG: de novo design method based on simple, fast and accurate free energy estimates. 1. Methodology and supporting evidence. J. Am. Chem. Soc. 118, 11733–11744.

[57] Mitchell, J.B.O., Laskowski, R.A., Alex, A. and Thornton, J.M. (1999) BLEEP: potential of mean force describing protein–ligand interactions: II. Calculation of binding energies and comparison with experimental data. J. Comp. Chem. 20, 1177–1185.

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

[59] Zhang, C., Liu, S., Zhu, Q. and Zhou, Y. (2005) A knowledge-based energy function for protein–ligand, protein–protein and protein–DNA complexes. J. Med. Chem. 48, 2325–2335.

[60] Wang, R., Liu, L., Lai, L. and Tang, Y. (1998) SCORE: A new empirical method for estimating the binding affinity of a protein–ligand complex. J. Mol. Model 4, 379–394.

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

[62] Bohm, H.J. (1998) Prediction of binding constants of protein–ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. J. Comput. Aided Mol. Des. 12, 309–323.

[63] Rarey, M., Kramer, B., Lengauer, T. and Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. 261, 470–489.

[64] Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.P. (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J. Comput. Aided Mol. Des. 11, 425–445.

[65] Head, R.D. et al. (1996) VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. J. Am. Chem. Soc. 118, 3959–3969.

[66] Krammer, A., Kirchhoff, P.D., Jiang, X., Venkatachalam, C.M. and Waldman, M. (2005) LigScore: A novel scoring function for predicting binding affinities. J. Mol. Graph. Model 23, 395–407.

[67] Wang, R., Lai, L. and Wang, S. (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J. Comput. Aided Mol. Des. 16, 11–26.

[68] Friesner, R.A. et al. (2004) Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. 47, 1739–1749.

[69] Wildman, S.A. and Crippen, G.M. (1999) Prediction of physicochemical parameters by atomic contributions. J. Chem. Inform. Comput. Sci. 39, 868–873.