An Iterative Knowledge-Based Scoring Function to Predict Protein–Ligand Interactions: I. Derivation of Interaction Potentials

SHENG-YOU HUANG, XIAOQIN ZOU
Department of Biochemistry, Dalton Cardiovascular Research Center, University of Missouri, Columbia, Missouri 65211

Received 16 March 2006; Accepted 23 May 2006
DOI 10.1002/jcc.20504
Published online 18 September 2006 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Using a novel iterative method, we have developed a knowledge-based scoring function (ITScore) to predict protein–ligand interactions. The pair potentials for ITScore were derived from a training set of 786 protein–ligand complex structures in the Protein Data Bank. Twenty-six atom types were used based on the atom type category of the SYBYL software. The iterative method circumvents the long-standing reference state problem in the derivation of knowledge-based scoring functions. The basic idea is to improve pair potentials by iteration until they correctly discriminate experimentally determined binding modes from decoy ligand poses for the ligand-protein complexes in the training set. The iterative method is efficient and normally converges within 20 iterative steps. The scoring function based on the derived potentials was tested on a diverse set of 140 protein–ligand complexes for affinity prediction, yielding a high correlation coefficient of 0.74. Because ITScore uses SYBYL-defined atom types, this scoring function is easy to use for molecular files prepared by SYBYL or converted by software such as BABEL.

J Comput Chem 27: 1866–1875, 2006

Key words: scoring function; protein–ligand interactions; ligand binding; knowledge-based; statistical potentials

Introduction

A great challenge in molecular docking and structure-based drug design is the development of an accurate and efficient scoring function to assess free energies of ligand binding. Numerous efforts have been made to overcome this barrier. The most rigorous and elaborate method includes free energy perturbation and thermodynamic integration, which treat water molecules explicitly (see ref. 1 for review). These methods, together with their simplified approaches such as LIE, PROFEC, and OWFEG (see ref. 1 and references therein), are computationally expensive and therefore impracticable for virtual database screening. A second type of approach treats water as a continuum dielectric medium, varying from a simple distant-dependent dielectric model[2] to more rigorous models such as the Poisson–Boltzmann model[3–6] and generalized Born model.[7–9] A third type of scoring functions uses an empirical approach to obtain a set of potential parameters by reproducing the experimentally measured binding affinities of a training set of protein–ligand complexes with known structures.[10–16]

A fourth type of scoring functions is based on a statistical potential or knowledge-based approach, which are derived from pairing frequencies of protein–ligand atom pairs observed in a database such as the Protein Data Bank (PDB).[17] In contrast to the earlier-mentioned empirical scoring functions, the knowledge-based scoring functions convert structural information into free energies without any knowledge of binding affinities, and thus are expected to be general because of its large training set of available structures. The pairwise-based potentials also make knowledge-based scoring functions as fast as empirical scoring functions. The idea of statistical potentials originates from the field of protein folding and structure prediction.[18–21] Recently, knowledge-based approaches have extensively been used to develop the scoring functions for protein–ligand interactions.[22–31]

The theory behind the statistical potential approach is the assumption of Boltzmann-like energetics. Namely, the Boltzmann distribution law for a single closed system held at fixed temperature is applicable to a database of structures (e.g. PDB).[32–34] Then, by applying an inverse Boltzmann relation

$$u(r) = -k_B T \ln(\rho(r)/\rho^*(r)) \tag{1}$$

Correspondence to: X. Zou; e-mail: zoux@missouri.edu

Contract/grant sponsor: Tripos (St. Louis, MO)
Contract/grant sponsor: NIH; contract/grant number: DK61529
Contract/grant sponsor: AHA; contract/grant number: 0265293Z
Contract/grant sponsor: OpenEye (Santa Fe, NM)

one can extract the interaction potentials from structural information. Here, k_B is the Boltzmann constant, and T is the absolute temperature. $\rho(r)$ is the number density of the protein–ligand atom pair at distance r, and $\rho^*(r)$ is the atom pair density in a "reference" state where the interatomic interactions are zero.

Despite its efficiency, there exists an inherent limitation for the knowledge-based scoring function because it involves calculation of a reference state. As pointed out by Thomas and Dill,[32] ideal reference states are not achievable, and the current methods to construct reference states are normally based on randomizing disconnected atoms and implicitly ignore excluded volume, sequences, and connectivity. Therefore, the extracted potentials by these methods are not equal to the true potentials. In their seminal work, Muegge and Martin carefully introduced a ligand volume correction factor in order to obtain an appropriate reference state.[24] Recently, Zhou and co-workers proposed an analytical approximation of a distance-scaled, finite, ideal-gas reference state.[31] Still, it is a great challenge to calculate the ideal reference state $\rho^*(r)$.

In the elegant work of Thomas and Dill,[33] the long-standing reference state problem was circumvented by an iterative method in extracting the effective interaction potentials. The potentials were obtained by iteration until they correctly discriminate a set of known protein folds from decoy conformations. Similar iterative methods have also been used in physics to solve the inverse problem, i.e. to derive Hamiltonian (or effective interaction potentials) of a given system from the information on its microscopic structures, e.g., known pair distribution functions.[35,36]

Despite successes in some simple systems,[33,36] the earlier-mentioned iterative methods cannot be directly applied to real protein-related systems with complicated interactions. To date, the reference state problem is still a big hurdle in the development of knowledge-based scoring functions. In this work, we presented a new, efficient iterative method to extract effective interaction potentials from a database of protein–ligand complex structures, which circumvents the calculation of the reference state. The basic idea of the method is to improve trial potentials iteratively through comparison of calculated and experimental pair distribution functions until native complex structures in the training database can correctly be discriminated from decoy structures. The iterative method converges fast, usually within 20 iterative steps. The scoring function based on the derived potentials, referred to as ITScore, was then used to predict binding affinities of a diverse set of 140 protein–ligand complexes. The results were compared with the experimental affinity data.

Materials and Methods

An Iterative Method to Extract Potentials from Complex Structures

As pointed out in ref. 32, an accurate reference state is not achievable during the derivation of pair potentials. Here, we circumvented the reference state calculation by using an iterative method. The overall idea of the method is to adjust pair potentials by iteration until they correctly discriminate native binding modes from decoy ligand poses using a training set of protein–ligand complex structures in the Protein Data Bank (PDB). We followed the assumption of pairwise additivity of atomic interactions. We also considered only the intermolecular interactions, namely, treating the molecules as rigid bodies.

Specifically, we started with a set of initial values for the pair potentials, $u_{ij}^{(0)}(r)$, where i and j stand for a protein atom type and a ligand atom type, and r is the atom pair distance. We will describe how to choose initial potentials $u_{ij}^{(0)}(r)$ in the Section Derivation of the initial potentials. At each iterative step, we calculated the binding score of every ligand orientation for each protein–ligand complex in the training database by summing over all interatomic (denoted as P–L) interactions using the current pair potentials, say, $u_{ij}^{(n)}(r)$ for the n-th iterative step

$$\text{energy score} = \sum_{\text{P-L atom pair}} u_{ij}^{(n)}(r). \tag{2}$$

We then identified the best-scored ligand orientation for each ligand-protein complex, which is the predicted binding mode using the current pair potentials. We next checked the following convergence criterion:

$$\eta \equiv \frac{1}{M} \sum_{m}^{\text{rmsd}_m<2} 1 > \eta_0 \tag{3}$$

where M is the number of protein–ligand complexes in the training database, and rmsd_m is the root mean square deviation between the best-scored ligand orientation and the experimentally determined orientation ("native binding mode") for the m-th complex. The convergence parameter η represents the success rate of identifying native-like binding modes with the criterion of rmsd < 2 Å. In the present work, the convergence threshold η_0 was set to be 99%.

In general, the initial potentials $u_{ij}^{(0)}(r)$ are unlikely to pass the convergence criterion (eq. (3)). Therefore after each step we constructed new potentials for the next cycle as[36]

$$u_{ij}^{(n+1)}(r) = u_{ij}^{(n)}(r) + \lambda k_B T \left(g_{ij}^{(n)}(r) - g_{ij}^{\text{obs}}(r)\right) \tag{4}$$

where k_B is the Boltzmann constant, and T denotes the system temperature. Without loss of generality, $k_B T$ was set to unit 1 in our work. λ is a parameter to control the speed of convergence and was set to 0.5.[36] $g_{ij}^{\text{obs}}(r)$ is the experimentally observed pair distribution function for the native binding modes of the training set, and $g_{ij}^{(n)}(r)$ is the predicted pair distribution function for the ligand ensemble structures calculated at the n-th iterative step. The details of the calculations on pair distribution functions are given in the next subsection.

Using the new pair potentials, we re-checked the convergence parameter η with eq. (3), and repeated the cycle until the convergence criterion is satisfied. We thus obtained a final set of pair potentials that discriminates native ligand poses from decoy poses.

Calculations of the Pair Distribution Functions

The experimentally observed pair distribution functions $g_{ij}^{\text{obs}}(r)$ were calculated by using the following formula:

$$g_{ij}^{\text{obs}}(r) = \rho_{ij}^{\text{obs}}(r) / \rho_{ij,\text{bulk}}^{\text{obs}} \tag{5}$$

where $\rho_{ij}^{\text{obs}}(r)$ and $\rho_{ij,\text{bulk}}^{\text{obs}}$ are the number densities of atom pair *ij* occurring in a spherical shell of radius from $r - dr/2$ to $r + dr/2$ and in a reference sphere of radius $R_{\max}$, respectively. In the present work, the bin size *dr* was set to 0.1 Å and the radius of the reference sphere, $R_{\max}$, was set to 10 Å.[37] $\rho_{ij}^{\text{obs}}(r)$ and $\rho_{ij,\text{bulk}}^{\text{obs}}$ were calculated as

$$\rho_{ij}^{\text{obs}}(r) = \frac{1}{M}\sum_{m}^{M}\frac{n_{ij}^{m}(r)}{4\pi r^2 dr} \quad \text{and} \quad \rho_{ij,\text{bulk}}^{\text{obs}} = \frac{1}{M}\sum_{m}^{M}\frac{N_{ij}^{m}}{V(R_{\max})} \tag{6}$$

where $n_{ij}^{m}(r)$ and N_{ij}^{m} are the numbers of atom pair *ij* in the spherical shell and the reference sphere for the *m*-th native complex structure, respectively. $V(R_{\max}) = 4\pi R_{\max}^3/3$ is the volume of the reference sphere. Obviously, $N_{ij}^{m} = \sum_{r=0}^{r=R_{\max}} n_{ij}^{m}(r)$. Again, *M* is the number of protein–ligand complexes in the training database.

Calculations of $g_{ij}^{(n)}(r)$, i.e. the predicted pair distribution functions at the *n*-th iterative step, are more complicated. For simple systems such as a monatomic system with a few hundred atoms, at each cycle, one can use Monte Carlo simulations to generate an ensemble of structures using the updated pair potentials and calculate the pair distribution functions from the ensemble structures.[36] However, this method is impractical for our system, which contains tens of atom types, thousands of atoms, and hundreds of complexes. Performing Monte Carlo simulations at every iterative cycle to generate a set of ligand ensemble poses for every binding site using the updated pair potentials is beyond today's computer power. Therefore we used a Boltzmann-weighted averaging method similar to that used by Thomas and Dill.[33] Our method is described in the following paragraph.

Before running the iterations, we generated a series of putative ligand binding orientations (decoys) for each protein–ligand complex, which covered the space in the binding pocket as much as possible and served as an approximation for an ensemble of structures. This is a one-time computation; the decoy orientations were saved to be used for the whole iterative procedure. Then, at each iterative step, we calculated the pair distribution function $g_{ij}^{(n)}(r)$ as

$$g_{ij}^{(n)}(r) = \rho_{ij}^{(n)}(r)/\rho_{ij,\text{bulk}}^{(n)} \tag{7}$$

where $\rho_{ij}^{(n)}(r)$ and $\rho_{ij,\text{bulk}}^{(n)}$ are the number densities of atom pair *ij* occurring in a spherical shell of radius from $r - dr/2$ to $r + dr/2$ and in a reference sphere of radius $R_{\max}$ at the *n*-th iterative cycle for the decoy complex structures, respectively. Here, $\rho_{ij}^{(n)}(r)$ and $\rho_{ij,\text{bulk}}^{(n)}$ were calculated as the Boltzmann-weighted pair frequencies over different decoy structures:

$$\rho_{ij}^{(n)}(r) = \frac{1}{ML}\sum_{m}^{M}\sum_{l}^{L}\frac{n_{ij}^{ml}(r)\,e^{-\beta U_{ml}}}{4\pi r^2 dr} \quad \text{and}$$

$$\rho_{ij,\text{bulk}}^{(n)} = \frac{1}{ML}\sum_{m}^{M}\sum_{l}^{L}\frac{N_{ij}^{ml}\,e^{-\beta U_{ml}}}{V(R_{\max})} \tag{8}$$

where $\beta = 1/k_{\text{B}}T$ and was set to 1 as mentioned already. $n_{ij}^{ml}(r)$ and N_{ij}^{ml} are the numbers of atom pair *ij* in the spherical shell and the reference sphere for the *l*-th decoy ligand orientation of the *m*-th complex, respectively. U_{ml} is the energy score of this ligand orientation, defined as the sum over all interatomic interactions (see eq. (2)). *L* is the total number of putative ligand orientations generated for each complex (including the native binding mode).

In the present work, the putative ligand orientations for each complex were generated with the molecular docking program DOCK 4.0.[38] Specifically, around 50 sphere points were selected for each protein in the training set. The grid spacing was set to 0.2 Å. Only van der Waals (VDW) interactions between non hydrogen atoms were considered. The maximum number of ligand orientations for each complex was set to 200.

Derivation of the Initial Potentials

In principle, one can start with any initial guess for the pair potentials, $u_{ij}^{(0)}(r)$. Given the rugged landscape in the multi-dimensional parameter space, a good set of initial potentials makes potential searching much more efficient.

The initial potential function used in the present work is a weighted combination of two types of energy functions. The first energy function is the potential of mean force, $w_{ij}(r)$, defined as

$$w_{ij}(r) \equiv -k_{\text{B}}T \ln g_{ij}^{\text{obs}}(r). \tag{9}$$

Because low occurrences of atom type pairs may result in a high uncertainty in the potential derivation, we ignored the contributions from those pair types that were insufficient in statistics. Specifically, we kept only the atom type pairs whose observed occurrences in the reference sphere were larger than 500; otherwise the corresponding interaction potentials, including $w_{ij}(r)$ and $u_{ij}(r)$, were set to zero. If no atom pair occurred in a particular spherical shell, the corresponding pair potential at the distance of the shell was arbitrarily set to 3 kcal/mol,[24] representing an unfavorable interaction.

It is well known that potentials of mean force, $w_{ij}(r)$, are not true potentials and lack an effective short-distant repulsive component. It is therefore necessary to incorporate a second VDW energy function to avoid steric clash.[39] Here, the Lennard-Jones 6–12 potentials, $v_{ij}(r)$, were used. The initial potentials $u_{ij}^{(0)}(r)$ were then calculated as

$$u_{ij}^{(0)}(r) = \begin{cases} w_{ij}(r) & \text{for hydrogen-bond pairs} \\ \frac{v_{ij}(r)\,e^{-v_{ij}(r)} + w_{ij}(r)\,e^{-w_{ij}(r)}}{e^{-v_{ij}(r)} + e^{-w_{ij}(r)}} & \text{otherwise} \end{cases}. \tag{10}$$

The VDW radii in the present work were taken from the AMBER force field.[2,38,40,41] To ensure that $v_{ij}(r)$ and $w_{ij}(r)$ yield similar potential minima for typical hydrophobic interactions such as C3F–C3F interactions, the well depths in $v_{ij}(r)$ were set to three times of the corresponding values given in the AMBER force field.

To remove the fluctuations of the initial potentials at large distances, which inherently result from the statistical uncertainties, we smoothed the initial potentials obtained from eq. (10) using a smooth function similar to that used in ref. 42:

$$F(r) = \begin{cases} \kappa(r_{\min} - r)^s - \epsilon_0 & r < r_{\min} \\ \epsilon_0(e^{-\alpha(r - r_{\min})} - 1)^2 - \epsilon_0 & r_{\min} \le r \le r_{\text{c}} \\ 0 & r > r_{\text{c}} \end{cases} \tag{11}$$

where ϵ_0, $r_{\min}$, and r_c denote the well depth, position of the potential minimum, and cutoff radius for $u_{ij}^{(0)}(r)$ calculated from eq. (10), respectively. ϵ_0 and $r_{\min}$ were directly obtained from the curve of $u_{ij}^{(0)}(r)$. The cutoff radius r_c was set to 6 Å in the present work.[24] The atom type pair (*ij*)-dependent parameters κ, s, and α were determined by least-square fitting between $u_{ij}^{(0)}(r)$ (eq. (10)) and its substitute, $F(r)$ (eq. (11)). The first expression in eq. (11) is an approximate of the VDW repulsion. The second expression stands for a Morse potential.

Flowchart

Figure 1 shows a flowchart for our iterative method and is summarized as follows:

1. Download protein–ligand complex structures from the PDB and prepare the database of native structures.
2. Calculate the experimentally observed pair distribution functions $g_{ij}^{\text{obs}}(r)$ of the protein–ligand atom pairs for the native binding modes using eqs. (5) and (6). Derive initial potentials $u_{ij}^{(0)}(r) = F_{ij}(r)$ using eqs. (9), (10), and (11).
3. For each protein–ligand complex, generate a series of putative ligand orientations around the binding site by using a molecular docking program such as DOCK 4.0.[38] These ligand orientations will serve as structure ensembles for iterative calculations.
4. Set the iterative step $n = 0$ and start iteration.
5. Calculate the energy scores of different ligand orientations for each complex using eq. (2) with the current interaction potentials $u_{ij}^{(n)}(r)$. Identify the best-scored binding mode.
6. Calculate the convergence parameter (or success rate) η according to eq. (3). If the convergence criterion is satisfied, skip to step 9; otherwise, continue.
7. Calculate the current pair distribution functions $g_{ij}^{(n)}(r)$ of the structure ensembles by using a Boltzmann-weighted averaging method (eqs. (7) and (8)).
8. Modify the current potentials $u_{ij}^{(n)}(r)$ according to eq. (4) and obtain a set of improved potentials $u_{ij}^{(n+1)}(r)$. Increase the iterative step n by one. Return to step 5 for the next cycle.
9. The iterative procedure is finished. Write out the final potentials.
10. Smooth the extracted potentials by using a smoothing algorithm described by Mitchell et al.[26] to account for inherent errors in experimental data (e.g., inaccuracies of atom positions can be as large as 0.4 Å for a resolution of 2.5 Å[43]). That is, the value of a potential in the *i*-th bin is set to the weighted average of 1 : 2 : 4 : 2 : 1 of the potentials from bins $(i-2)$ to $(i+2)$.

Our iterative method is very efficient, and usually converges within 20 cycles.

Preparation of the Training Database

With the initial potentials and the iterative method, one can extract effective pair potentials from any database of protein–ligand complex structures. In the present work, we chose known complex structures available in the PDB.[17] Specifically, we used only the experimentally determined X-ray complex structures with a resolution better than 2.5 Å. We also discarded the PDB entries complexed with RNA, DNA, covalently bound ligands, peptide inhibitors or ligands with less than 5 or more than 66 heavy atoms. We further

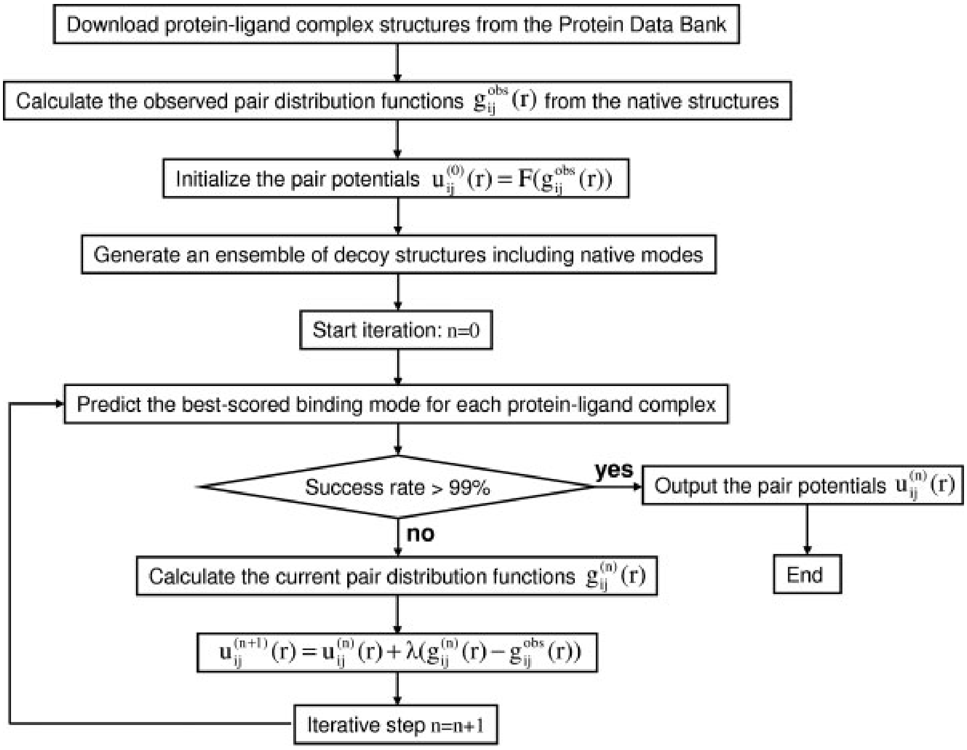
Download protein-ligand complex structures from the Protein Data Bank
Calculate the observed pair distribution functions $g_{ij}^{obs}(r)$ from the native structures
Initialize the pair potentials $u_{ij}^{(0)}(r) = F(g_{ij}^{obs}(r))$
Generate an ensemble of decoy structures including native modes
Start iteration: n=0
Predict the best-scored binding mode for each protein-ligand complex
Success rate > 99%
yes
Output the pair potentials $u_{ij}^{(n)}(r)$
End
no
Calculate the current pair distribution functions $g_{ij}^{(n)}(r)$
$u_{ij}^{(n+1)}(r) = u_{ij}^{(n)}(r) + \lambda(g_{ij}^{(n)}(r) - g_{ij}^{obs}(r))$
Iterative step n=n+1

Figure 1. The flowchart of the iterative method.

excluded the complexes crystalized at abnormal pH conditions (lower than 6.5 or higher than 7.5). As reported in the previous works,[25,31] the inclusion of metal ions did not give a significant improvement for knowledge-based scoring functions. Therefore, considering the complication of ion-mediated interactions,[44,45] we ignored the entries complexed with metal ions other than Na^+ and K^+ near the binding sites of interest and left the challenging issue for the future study. To avoid redundancy, we chose only one representative for the complexes with the same second and third letter of the four-letter identification code of the PDB entry because they may share similar protein structures, unless the bound ligands are very different. We treated a complex bound with multiple ligands as separate entries, and the cofactor ligand as part of the protein. Finally, we removed the complexes that contain severe steric clashes (i.e., the distance between a protein atom and a ligand atom is shorter than 1.75 Å), because experimentally determined structural data with large errors will significantly affect the development or evaluation of a reliable scoring

Table 1. List of Protein and Ligand Atom Types Based on the SYBYL Definition.

sp/sp^2 Carbon (C.1, C.2, C.ar and C.cat)	
C2+	Carbon bonded to a positively charged nitrogen
C2−	Carbon bonded to a negatively charged oxygen
C2N	Carbon in amide groups
C2O	Carbon bonded to O.2, but not belonging to C2+, C2−, or C2N
C2X	Other sp/sp^2 carbon
sp^3 Carbon (C.3)	
C3F	Carbon only bonded to carbon or hydrogen
C3X	Carbon other than C3F
sp^2 Nitrogen (N.2, N.ar, N.am, and N.pl3)	
N2N	Nitrogen in amide groups
N2+[a]/NC	Positively charged nitrogen
N21	Nitrogen bonded to one non hydrogen atom
N22	Nitrogen bonded to two non hydrogen atoms
N2X	Nitrogen except N2N, N2+, N21, and N22
sp Nitrogen (N.1)	
N1	All sp nitrogen
sp^3 Nitrogen (N.3 and N.4)	
N3+[a]/NC	N.4 or nitrogen bonded to one or two non-hydrogen atoms
N3X	sp^3 nitrogen except N3+
sp^2 Oxygen (O.2)	
O2	All sp^2 oxygen
sp^3 Oxygen (O.3)	
O31	Oxygen bonded to one non hydrogen atom
O32	Oxygen bonded to two non hydrogen atoms
Negatively charged oxygen (O.co2)	
OC	All negatively charged oxygen
Sulfur (S.2, S.3, S.O, S.O2, etc.)	
S1	Sulfur single-bonded to one non hydrogen atom
SO	Sulfur bonded to sp^2 oxygen
SX	Sulfur except S1 and SO
Phosphorus (P.3)	
P	All phosphorus
Halogan (F, Cl, Br, and I)	
F	All fluorine
Cl	All chlorine
Br[b]	All bromine
I[b]	All iodine
Metal ions	
MET	Metal ions (MG, ZN, CA, etc.)

[a] The atom types "N2+" and "N3+" are grouped as the charged nitrogen "NC" because they normally carry a positive charge.

[b] The atom types "Br" and "I" are grouped as "Br" because of their low occurrences. Therefore, ITScore uses a total of 26 atom types.

function.[46,47] A total of 786 complex structures were obtained (see Appendix).

In our prepared protein–ligand structures, the water molecules were removed from the complexes. The hydrogen atoms were not considered explicitly. A total of 26 atom types were used, based on the definitions provided by SYBYL software (Tripos, Inc.). A summary of the atom types is listed in Table 1. Our scoring function is therefore readily applicable to molecular files prepared by SYBYL (or converted by BABEL[48]) and other widely-used molecular docking programs such as DOCK.[38]

Results and Discussion

Extracted Pair Potentials

The number of atom pair occurrences depends on the atom types. The largest number of atom type pairs (165, 366) was found for the C3F–C2X pair. We retained only the interaction potentials of the atom type pairs whose occurrences were above 500 to ensure a good statistics. Thus, there were 236 pairs of effective interaction

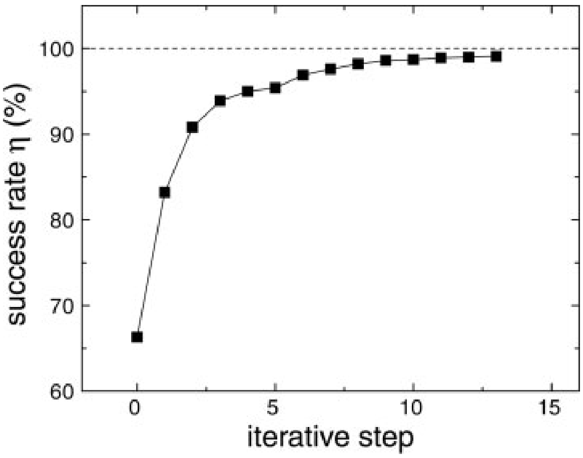

Figure 2. The convergence parameter η, representing the success rate of identifying near-native binding modes (rmsd < 2 Å), as a function of the iterative step. The dashed line ($y = 100\%$) is plotted for reference.

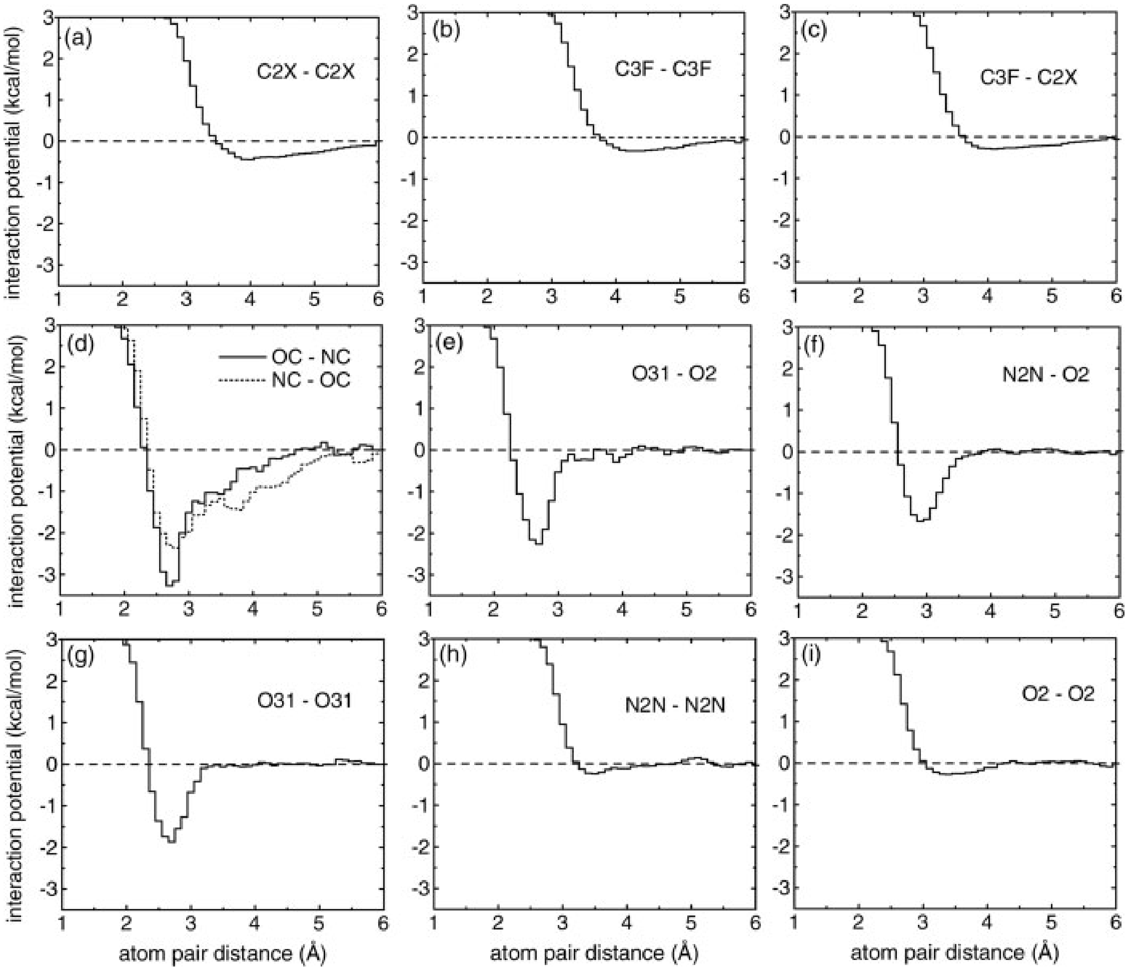

Figure 3. Twelve selected pair potentials of ITScore. The first atom-type label refers to a protein atom, and the second to a ligand atom. The dashed line ($y = 0$) is plotted for reference.

potentials out of 676 possible pairs. Figure 2 shows the convergence parameter η, defined in eq. (3), as a function of the iterative step. It is seen from this figure that the convergence parameter (or success rate) rapidly approaches 100% with the increase of the iterative step, indicating that almost all the native binding modes of the protein–ligand complexes in the training database were found at the end of the iteration. The fast convergence of the iterative procedure is a sign of the efficacy of our method in extracting effective interaction potentials.

Figure 3 shows a selected set of derived pair potentials. Several characteristic features can be seen in the figure, which are consistent with experimental findings. For the C2X–C2X, C3F–C3F and C3F–C2X potentials, there exists a minimum around 4 Å, corresponding to hydrophobic interactions between these atom types. The C2X–C2X interaction is slightly stronger and occurs at a slightly shorter distance than the other two types of hydrophobic interactions. The reason is that the atom type C2X mainly consists of aromatic carbons (C.ar), which are known to have stronger and entropically more favorable interactions.[49] For the OC–NC (or NC–OC), O31–O2, and N2N–O2 interactions, there exists a valley between 2.7 and 2.9 Å, which is consistent with hydrogen bond interactions between these atom types. It is also seen that the OC–NC interaction is stronger and wider than the other two interactions. This is due to the involvement of an additional favorable salt bridge, as OC and NC are oppositely charged. The O31–O31 potential also exhibits a minimum at a distance of 2.7 Å because of a polar charge-assisted interaction.[50] The interactions are weak for N2N–N2N and O2–O2 pairs, because they are either hydrogen bond donor-donor pairs or acceptor-acceptor pairs and thus cannot form hydrogen bonds. The repulsive electrostatic interactions due to the same type of partial charges also weaken the pair interactions.

Table 2. Pearson Correlation Coefficients of the Binding Affinity Predictions on Four Diverse Test Sets Calculated with ITScore.

No.	Test set	Ref.	No. of complexes	Correlation coefficients
1	Muegge and Martin's test set	24	77	0.81
2	Eldridge et al.'s training set	16	79	0.75
3	Böhm's training set	11	73	0.80
4	Sets 1–3		140	0.74

To examine the sensitivity of ITScore to the initial decoy structures, we also used a different set of decoy structures generated using DOCK 4.0 with a different "random seed" parameter. The derived effective pair potentials are not significantly different from the potentials derived from the previous set of decoy structures. Thus, the pair potentials we derived are quite robust.

Binding Affinity Predictions on Four Diverse Test Sets

Based on the derived pair potentials $u_{ij}(r)$, our scoring function (referred to as ITScore) is calculated as

$$\text{energy score} = \sum_{\text{P-L atom pair}} u_{ij}(r). \tag{12}$$

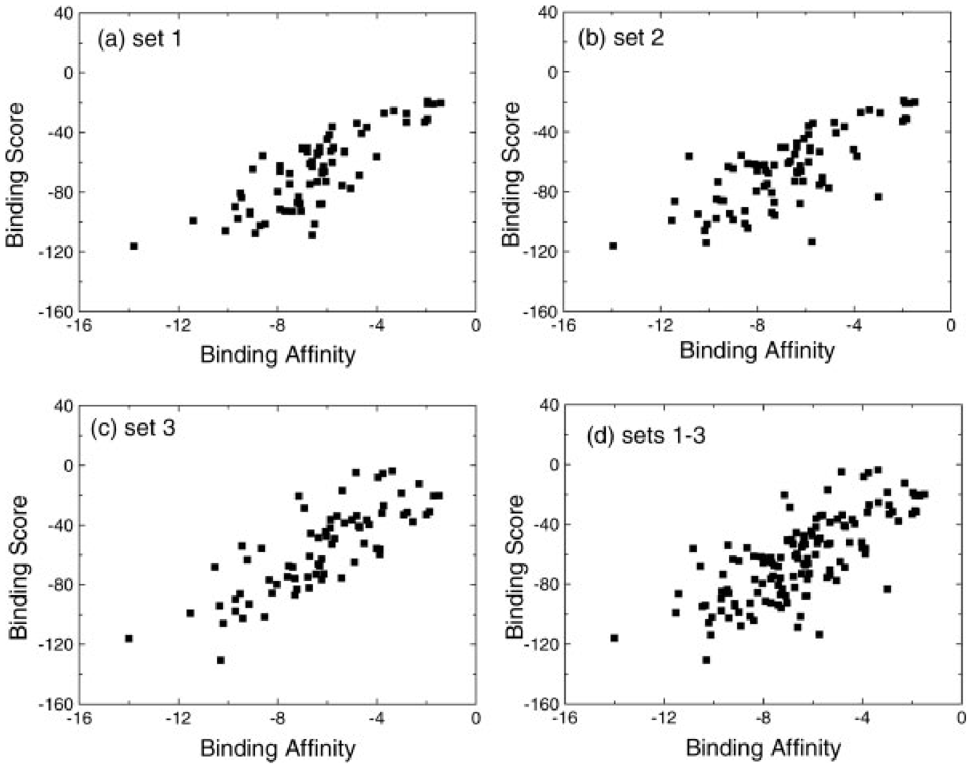

Figure 4. Correlations between the experimentally determined binding affinities and the calculated binding scores with ITScore on the four test sets given in Table 2.

The scoring function was tested on binding affinity predictions for four diverse data sets of protein–ligand complexes with known crystal structures and binding affinities. The results are shown in Table 2 and Figure 4. The first set was the test set for the PMF scoring function constructed by Muegge and Martin, which includes a total of 77 protein–ligand complexes.[24] The second set was obtained from the training set of the empirical scoring function, ChemScore, used by Eldridge et al.[16] (79 complexes). The third set was taken from the training set of another empirical scoring function, LUDI, prepared by Böhm[11] (73 complexes). One note about the last two sets is that we included every complex in the original references with available 3D structure in the PDB except the biotin-streptavidin complex (PDB code: 1stp). As pointed out by Muegge and Martin,[24] streptavidin is a tetramer, but 1stp contains only a single subunit bound with a biotin molecule. The inclusion of other streptavidin subunits may dramatically enhance binding of biotin, so 1stp was removed from our test sets. To further evaluate the robustness of our scoring function on diverse complexes, we combined these three test sets to a large set of 140 protein–ligand complexes as listed in Table 3, which served as the fourth test set. Table 2 shows that ITScore yields a good correlation in affinity prediction for every test set, suggesting the robustness of ITScore predictions. The Pearson's correlation coefficient is 0.74 for the total 140 complexes (i.e., Test Set 4). More detailed validation of the scoring function is given in the accompanying paper.[51]

Table 3. Test set of 140 protein—ligand complexes.[a]

1aaq	1abe	1abf	1acj	1adb	1add	1apb	1apt	1apu	1bap
1bra	1bzm	1cbx	1cil	1cps	1ctt	1dih	1dog	1dwb	1dwc
1dwd	1ebg	1eed	1ela	1elc	1epo	1epp	1etr	1ets	1ett
1fkf	1hbv	1hpv	1hsl	1htf	1htg	1hvi	1hvj	1hvk	1hvl
1hvr	1l82	1l83	1l86	1l87	1ldm	1lyb	1mbi	1mfe	1mnc
1nsc	1nsd	1pgp	1phe	1phf	1phg	1ppc	1pph	1ppk	1pso
1r09	1rbp	1rne	1sbp	1sre	1tlp	1tmn	1tmt	1tng	1tnh
1tni	1tnj	1tnk	1tnl	1ulb	2cgr	2cpp	2ctc	2er0	2er6
2er7	2er9	2gbp	2gpb	2ifb	2phh	2r04	2tmn	2tsc	2xis
2ypi	3cpa	3dfr	3er3	3ptb	3tmn	3tpi	4dfr	4er1	4er4
4fab	4gr1	4hmg	4hvp	4phv	4tln	4tmn	4ts1	5abp	5cna
5cpp	5er2	5hvp	5tim	5tln	5tmn	6abp	6acn	6cpa	6rsa
6tmn	7abp	7cpa	7cpp	7dfr	7hvp	8abp	8cpa	9aat	9abp
9hvp									

[a] Among the listed PDB codes, nine complex structures, each containing two different ligand conformations, were treated separately: 1abe, 1abf, 5abp, 6abp, 7abp, 8abp, 9abp, 1apb, and 1bap.

Conclusion

We have developed an efficient knowledge-based scoring function (ITScore) to predict protein–ligand interactions by using a novel iterative method, based on 786 structures of protein–ligand complexes and 26 atom types. The basic idea of the iterative method is to circumvent the inaccessible reference state problem by improving the pair potentials until they are able to discriminate native binding modes from decoy ligand structures. The iterative procedure normally converges within 20 iterative steps. The derived pair potentials show consistent behaviors with experimental findings. Test of ITScore on binding affinity predictions yields good correlations for four data sets consisting of a total of 140 diverse ligand-protein complexes with known affinities.

Notice that ITScore requires no prior knowledge on measured binding affinities to derive the pair potentials. Furthermore, the training set we used in the present work is large and diverse, and contains statistically sufficient atom pairs for extraction of the pair potentials. Therefore, the good correlation we obtained on affinity predictions with ITScore shows the robustness of this scoring function on evaluating protein–ligand interactions. The validation and application of ITScore will be described in detail in part II of this work.[51]

Appendix

The 786 PDB entries used for derivation of ITScore. The labels (a, b, c, and d) refer to the complexes in which one, two, three, and four different ligands were used, respectively.

```
a 11as 12as 1a28 1a2c 1a4w 1a5x 1a61 1a78 1a80 1a82 1a8g 1a8i
1a8r 1a9u 1aax 1ae8 1afe 1afq 1agw 1alw 1amw 1aqb 1aqw 1aqx 1arg
1asu 1au4 1ax0 1ax2 1axr 1aym 1b0o 1b0u 1b16 1b1c 1b3d 1b7y 1b9t
1bb0 1bht 1biw 1bj5 1bji 1bju 1bjv 1bky 1bl5 1bl6 1bm7 1bmk 1bs1
1bso 1bsv 1btn 1bu5 1bv7 1bvy 1bwa 1bwb 1bwc 1bwu 1c23 1c4q 1c4u
1c4v 1c50 1c5o 1c7o 1c80 1c81 1c83 1c88 1c8k 1c8u 1c9h 1ca8 1cet
1cg6 1cgk 1cgz 1chw 1cjw 1ckm 1ckp 1cly 1cpu 1cqf 1cqq 1cyd 1cza
1czi 1d0h 1d1q 1d2a 1d2s 1d5z 1d6s 1d6w 1d7u 1d9i 1db4 1dbt 1dbv
1dek 1dg9 1dgl 1dht 1di8 1diw 1djr 1dkf 1dkp 1dll 1dm2 1dmk 1dnl
1dp2 1dqp 1dv2 1dvj 1dvs 1dvt 1dvu 1dvx 1dvz 1dxr 1dy4 1dz8 1e1x
1e20 1e2i 1e2j 1e3r 1e3v 1e40 1e4i 1e4n 1e56 1e5j 1e6e 1e6r 1e7a
1e7b 1e7f 1e7v 1e8w 1e8z 1e9h 1ecs 1ecv 1egh 1eix 1ek5 1el5 1eqc
1eqy 1ewk 1exa 1ey3 1ez1 1ez9 1f17 1f3a 1f3b 1f3t 1f4e 1f4l 1f5v
1f7k 1f7p 1f8g 1f9d 1fae 1fao 1fbo 1fbw 1fby 1fcx 1fd0 1ffq 1fgi
1fj4 1fjj 1fk8 1fkn 1flm 1fm7 1fmj 1fqa 1fqo 1frz 1fs4 1ftk 1ftq
1fu4 1fv0 1fvt 1fwu 1fxs 1fy7 1g0n 1g0o 1g13 1g2n 1g56 1g5y 1g6n
1g74 1g7f 1g86 1g8i 1g98 1g9q 1g9v 1ga1 1gan 1gbn 1gck 1gfz 1gg2
1gg6 1ghw 1gi8 1gii 1gj7 1gj8 1gjb 1gjd 1gkl 1gm8 1goo 1goy 1gp2
1gsa 1gx8 1gxa 1gz8 1h01 1h06 1h08 1h0s 1h0v 1h1r 1h3n 1h46 1h5u
1h6c 1h78 1h94 1h9z 1ha2 1hdo 1hg4 1hg5 1hi3 1hj6 1hk4 1hlf 1hm2
```

1hmu 1ho4 1hop 1hox 1hp1 1hqp 1hv6 1hw8 1hzz 1i37 1i7p 1i9h 1ia9
1iah 1iep 1iex 1ig3 1iin 1iiu 1ik4 1ikg 1inf 1iri 1is4 1is6 1iuc
1iwe 1iwh 1ix7 1iz2 1j0d 1j0i 1j16 1j17 1j1g 1j4h 1j6z 1j84 1j8a
1j8v 1jbw 1jc9 1jdt 1jdx 1jep 1jg6 1jil 1jjv 1jkl 1jkx 1jlx 1jmo
1js3 1jsh 1jsv 1jtv 1ju4 1jzs 1k06 1k0y 1k12 1k3l 1k3t 1k5s 1k6x
1k9j 1k9s 1k9t 1ka0 1kak 1kav 1kdk 1kdt 1ke8 1ke9 1ki3 1kj1 1kkp
1kl1 1kly 1km3 1km6 1kn2 1koj 1kpm 1kpv 1kqr 1kqz 1kr0 1kti 1kuj
1kuy 1kv1 1kwc 1kxh 1kz8 1kzk 1l2t 1l3i 1l8g 1lbb 1lbl 1lgt 1lh0
1lhu 1lkd 1llq 1lor 1los 1lot 1lp6 1lti 1lua 1lw5 1lwj 1lwn 1lwo
1lyx 1m0u 1m1d 1m26 1m3u 1m48 1m4d 1m5b 1m5w 1m7q 1m9n 1mai 1me3
1mfi 1mgp 1mj7 1mja 1mjl 1ml3 1ml6 1mly 1mo9 1mp0 1mq5 1mqe 1mrx
1mrz 1ms8 1ms9 1msm 1mtv 1mvt 1mwe 1mxu 1my2 1mzs 1n0u 1n1e 1n1g
1n1t 1n3o 1n5t 1n6b 1n71 1n83 1ndi 1ndj 1ne7 1nep 1ney 1nf0 1nhu
1nl9 1nli 1nm6 1nmd 1nmk 1nny 1no6 1npl 1nq2 1nq5 1nr6 1nt1 1nuq
1nut 1nux 1nvq 1nwl 1nz7 1nzd 1o0s 1o2g 1o3p 1o3w 1o57 1o72 1o94
1o9b 1obh 1oc5 1ofs 1ogx 1oh0 1oiq 1oir 1oit 1oja 1oni 1ony 1onz
1opk 1oss 1oth 1ouk 1ouy 1ove 1ox5 1oyn 1oz0 1p1q 1p2d 1p5e 1p5r
1p60 1p61 1p6x 1p72 1p84 1pa9 1pbq 1pds 1pgt 1ph0 1pj6 1pjc 1pkd
1pme 1pmn 1pmq 1pmv 1pq6 1pwz 1px0 1pxh 1pyn 1pz4 1q3d 1q3w 1q41
1q6m 1q6p 1q9d 1q9m 1qa0 1qb6 1qbn 1qbo 1qcf 1qdc 1qfv 1qhf 1qhi
1qi0 1qi5 1qkj 1qkq 1qkt 1ql9 1qmg 1qnf 1qpc 1qpe 1qpk 1qwn 1qxk
1r0p 1r14 1ra8 1ra9 1rbo 1rc4 1rf7 1rg7 1rx5 1say 1sfc 1swn 1tmk
1tox 1tsl 1uca 1ugx 1uja 1uom 1v39 1vp9 1yef 1zeg 2ans 2arc 2bkj
2chb 2cmk 2dpm 2dub 2gpa 2ki5 2lbd 2man 2pgt 2skc 2sli 2src 2vp3
3cpu 3ert 3mag 3man 3mct 3std 3upj 4dcg 4fiv 6atj 7taa

[b] 1a27 1aoe 1aua 1b4d 1bf3 1bgn 1bhx 1bkw 1boz 1bq1 1c0l 1c14
1c3v 1c5w 1cml 1cnq 1d8a 1del 1dfo 1doh 1dtl 1e1k 1e1y 1e5q 1e6z
1e7y 1eef 1eqa 1esv 1f06 1f0y 1g6o 1g76 1gp6 1gw2 1h16 1h61 1he3
1hnn 1hwi 1i00 1i0z 1i10 1i2c 1i5r 1i7g 1ia1 1ib0 1icq 1ipf 1iut
1ixn 1iyk 1j4r 1ja9 1jdj 1jdv 1je1 1jqi 1k0j 1k4m 1k97 1kbi 1ki6
1lev 1lf9 1lx6 1m67 1m78 1m7y 1mfp 1mrq 1mxh 1n5q 1naa 1nlm 1ocn
1od8 1ofd 1oj4 1p7c 1pq9 1pt8 1pty 1q4s 1qha 1qrr 1r31 1rb3 1rh3
1rx2 1ueh 2ae2 2udp

[c] 1dqa 1n8u 1p9l 1qz6

[d] 1gth

Acknowledgments

We thank Prof. Irwin D. Kuntz and Dr. Hao-Yang Liu for helpful discussions and Prof. Kevin Gillis for valuable comments on the manuscript.

References

1. Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. Annu Rev Biophys Biomol Struct 2001, 30, 211.
2. Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. J Comput Chem 1992, 13, 505.
3. Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. J Comput Chem 2002, 23, 128.
4. Grant, J. A.; Pickup, B. T.; Nicholls, A. J Comput Chem 2001, 22, 608.
5. Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Proc Natl Acad Sci USA 2001, 98, 10037.
6. Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. J Mol Biol 2002, 322, 339.
7. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. J Am Chem Soc 1990, 112, 6127.
8. Zou, X.; Sun, Y.; Kuntz, I. D. J Am Chem Soc 1999, 121, 8033.
9. Liu, H.-Y.; Kuntz, I. D.; Zou, X. J Phys Chem B 2004, 108, 5453.
10. Böhm, H. J. J Comput Aided Mol Des 1994, 8, 243.
11. Böhm, H. J. J Comput Aided Mol Des 1998, 12, 309.
12. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Freer, S. T. Chem Biol 1995, 2, 317.
13. Gehlhaar, D. K.; Bouzida, D.; Rejto, P. A. In Rational Drug Design: Novel Methodology and Practical Applications; Parrill, L.; Reddy, M. R.; Eds.; American Chemical Society: Washington, DC, 1999, p. 292.
14. Jain, A. N. J Comput Aided Mol Des 1996, 10, 427.
15. Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. J Am Chem Soc 1996, 118, 3959.
16. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. J Comput Aided Mol Des 1997, 11, 425.
17. 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, 235.
18. Tanaka, S.; Scheraga, H. A. Macromolecules 1976, 9, 945.
19. Miyazawa, S.; Jernigan, R. L. Macromolecules 1985, 18, 534.
20. Sippl, M. J. J Mol Biol 1990, 213, 859.
21. Vajda, S.; Sippl, M.; Novotny, J. Curr Opin Struct Biol 1997, 7, 222.
22. Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. Protein Eng 1995, 8, 677.
23. DeWitte, R. S.; Shakhnovich, E. I. J Am Chem Soc 1996, 118, 11733.
24. Muegge, I.; Martin, Y. C. J Med Chem 1999, 42, 791.

25. Muegge, I. J Med Chem 2005, http://pubs.acs.org/cgi-bin/abstract.cgi/jmcmar/asap/abs/jm050038s.
26. Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. J Comput Chem 1999, 20, 1165.
27. Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M. J Comput Chem 1999, 20, 1177.
28. Gohlke, H.; Hendlich, M.; Klebe, G. J Mol Biol 2000, 295, 337.
29. Ishchenko, A. V.; Shakhnovich, E. I. J Med Chem 2002, 45, 2770.
30. Velec, H. F. G.; Gohlke, H.; Klebe, G. J Med Chem 2005, 48, 6296.
31. Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. J Med Chem 2005, 48, 2325.
32. Thomas, P. D.; Dill, K. A. J Mol Biol 1996, 257, 457.
33. Thomas, P. D.; Dill, K. A. Proc Natl Acad Sci USA 1996, 93, 11628.
34. Koppensteiner, W. A.; Sippl, M. J. Biochemistry (Moscow) 1998, 63, 247.
35. Soper, A. K. Chem Phys 1996, 202, 295.
36. Almarza, N. G.; Lomba, E. Phys Rev E 2003, 68, 011202.
37. Muegge, I. Perspect Drug Discov Des 2000, 20, 99.
38. Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. J Comput Aided Mol Des 2001, 15, 411.
39. Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik, S. W. J Med Chem 1999, 42, 2498.
40. Weiner, S. J.; Kollman, P. A.; Case, D. A. J Am Chem Soc 1984, 106, 765.
41. Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. J Comput Chem 1986, 7, 230.
42. Muryshev, A. E.; Tarasov, D. N.; Butygin, A. V.; Butygina, O. Y.; Aleksandrov, A. B.; Nikitin, S. M. J Comput Aided Mol Des 2003, 17, 597.
43. Kossiakoff, A. A.; Randal, M.; Guenot, J.; Eigenbrot, C. Proteins 1992, 14, 65.
44. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. J Med Chem 2004, 47, 3032.
45. Raha, K.; Merz, K. M., Jr. J Am Chem Soc 2004, 126, 1020.
46. Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Code, J. C.; Taylor, R. Proteins 2002, 49, 457.
47. Davis, A. M.; Teague, S. J.; Kleywegt, G. J Angew Chem, Int Ed 2003, 42, 2718.
48. Walters, P.; Stahl, M. BABEL, version 1.6 © 1992–1996, University of Arizona.
49. Burley, S. K.; Petsko, G. A. Science 1985, 229, 23.
50. Davis, A. M.; Teague, S. J Angew Chem, Int Ed 1999, 38, 736.
51. Huang, S.-Y.; Zou, X. J Comput Chem 2006, 27, 1876.