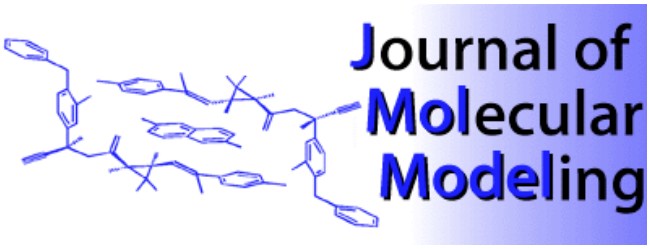
J. Mol. Model. **1998**, *4*, 379 – 394



FULL PAPER

© Springer-Verlag 1998

SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-Ligand Complex

Renxiao Wang, Liang Liu, Luhua Lai, and Youqi Tang

Institute of Physical Chemistry, Peking University, Beijing 100871, P.R.China. Tel: +86-10-62751490; Fax: +86-10-62751725. E-mail: lai@ipc.pku.edu.cn

Received: 29 September 1998 / Accepted: 15 October 1998 / Published: 1 December 1998

Abstract A new method is presented to estimate the binding affinity of a protein-ligand complex with known three-dimensional structure. The method, SCORE, uses an empirical scoring function to describe the binding free energy, which includes terms to account for van der Waals contact, metal-ligand bonding, hydrogen bonding, desolvation effect, and deformation penalty upon the binding process. The coefficients of each term are obtained by multivariate regressional analysis of a diverse training set of 170 protein-ligand complexes. The final scoring function reproduces the binding free energies of the whole training set with a cross-validated deviation of 6.3 kJ/mol. The predictive ability of the function is further tested by a set of 11 endothiapepsin complexes and the internal consistency of the function is demonstrated in a stepwise procedure named Evolutionary Test. A major innovation of this method is the introduction of an atomic binding score which allows the researcher to inspect and optimize the lead compound rationally in a structure-based drug design scheme.

Keywords Protein-ligand complex, Binding affinity, Empirical scoring function, Structure-based drug design

Introduction

Three-dimensional structures of proteins, provided by either X-ray crystallography or NMR spectroscopy, are becoming increasingly important in the design of novel drugs. They have enabled medicinal chemists directly to inspect those structural properties of a target protein that are essential for interacting with a ligand. This in principle allows for the rationalisation of the design process which is referred to as structure-based drug design [1-12]. In such a process, new leads may originate from three-dimensional database searching or so-called *de novo* methods. However, all these approaches are limited by the accuracy with which the affinity of proposed ligands can be gauged. Since correct ranking of putative ligands for synthesis is a prerequisite to a useful strategy for drug design, there is a clear need for an objective method that is able to predict the binding affinity of a protein-ligand complex in a quantitative way.

Predicting the binding affinity of a ligand to its target protein is still a scientific challenge at present. A comprehensive review of this area has been provided by Ajay and Murcko [13]. With respect to the rigorous calculation of relative binding energies, substantial progress has been made with free energy perturbation (FEP) [14] which is currently

Correspondence to: L. Lai

the only method that attempts to deal seriously with calculating ensemble averages and considers solvent molecules explicitly. However, despite various approximations geared towards performance enhancement, this method is computationally intensive and restricted to small molecular systems. Therefore, it is less practical in drug design where the synthetic chemists require fast feedback from the modeling department.

Another popular method for assessing protein-ligand binding is molecular mechanics. Following the pioneering idea of Goodford [15], the interaction energy between the protein and its ligand is calculated by a simplified, often grid-based force field. Basic components may include steric and electrostatic energies, sometimes supplemented by other terms accounting for hydrogen bonding and solvation effects [16-20]. The purely molecular mechanics-based method has been applied widely to molecular docking studies which aim at finding the bound conformation of the ligand. But for estimating binding affinities, the success of such approaches depends on the ability to define the set of ligands on which predictions will be applied carefully.

More recently, empirical schemes have met with significant interest. The basic assumption underlined in such approaches is that the overall binding free energy can be decomposed into components. This can be written out conceptually by the following equation.

$$\Delta G_{binding} = \Delta G_{motion} + \Delta G_{interaction} + \Delta G_{solvent} + \Delta G_{configuration}$$

The parameters in the equation are often determined from binding data in a statistical manner. This kind of approach is also referred to as "Master Equation" [13]. At the very beginning, such approaches were also applied only to congeneric series [21-25]. As a breakthrough, Böhm was the first who developed a general-purposed empirical function to describe the binding energy [26]. The free energy of binding was written as the sum of terms including a constant representing overall rotational and translational entropy loss, a sum over all hydrogen bonds formed, a sum over all ionic interactions, the loss of lipophilic surface area upon binding, and the number of torsions that are frozen. A linear equation was obtained through regressional analysis of 45 protein-ligand complexes and the equation seemed to have reasonable predictive ability in the example tested. A more complicated procedure was reported by Head et al. who had examined 51 protein-ligand complexes using partial least squares regression and a neural network [27]. They developed a hybrid model combining energetic considerations from molecular mechanics and calculated molecular properties related to desolvation and entropy loss upon the binding process. Similar empirical schemes are also reported by Gschwend et al. [28] and Eldridge et al. [29] who have obtained scoring functions by using much larger data sets. These approaches are of course simplistic methods, but they could capture the essential physics of protein-ligand binding at modest computational cost. They have been proved valuable in screening database hits and scoring molecules generated by *de novo* design programs.

In this paper, we present a new general-purposed empirical method, SCORE, for estimating the absolute binding affinity of a protein-ligand complex with known three-dimensional structure. We try to accomplish two goals in this study: (i) developing a fast, accurate, and robust scoring function for structure-based drug design. We have used a linear empirical scoring function to describe the binding free energy in which new terms and parameters are used. The final model was obtained by regressional analysis upon a training set, which is the largest one yet reported, composed of 170 complex structures. (ii) providing a practical tool to interpret the interaction between the protein and its ligand. According to our method, the binding affinity of the ligand can be decomposed to the contribution of individual atoms. Each ligand atom gets a score, the called atomic binding score, indicating its role in the binding process. The introduction of the atomic binding score allows the designer to inspect and optimize the lead compound structure in a more rational way.

Methods and computation results

Training set

The training set used in this study comprises 170 protein-ligand complexes (see Table 1). All the complexes were taken from the Protein Data Bank (PDB) [30]. Since our interests are concentrated on small non-covalently bound ligands, those complexes containing covalently-bound ligands, complex ligands (such as heme), and macromolecular ligands were stripped out of the data set. More than seventy different kinds of proteins are represented in this training set and all the structures are of high resolution (better than 3.2 Å). The experimentally determined binding data were cited from the literature [26-29] and expressed in the negative logarithms of dissociation equilibrium constants, i.e. pK_d, for convenience. The pK_d values in this set range from 1.54 to 13.96, covering over 12 orders of magnitude. We have not checked the binding data for differences in temperature or salt concentrations during measurement.

Each complex in the training set was processed with the SYBYL software [31] as follows. First, the ligand was extracted from the complex structure, assigned proper atom and bond types, and written out as a separate file in MOL2 format. The remainding part of the complex, i.e. the protein, was then written out to another file in PDB format. Water molecules, metal ions, and other cofactors were left with the protein and treated as part of it. No further structure minimisation was performed on either the ligand or the protein. A special note should be addressed here is that hydrogen atoms are unnecessary in the structure because our algorithm, as described below, considers heavy atoms only.

Table 1 *Training set used in SCORE*

PDB	pK_d	Resl.	Protein/ligand
1aaq	8.40	2.5	HIV-1 protease/hydroxyethylene isostere
1abe	6.52	1.7	L-arabinose binding protein/L-arabinose
1abf	5.42	1.9	L-arabinose binding protein/D-fucose
1adb	8.40	2.4	alcohol dehydrogenase/CNAD
1adf	4.58	2.9	alcohol dehydrogenase/TAD
1apb	5.82	1.76	L-arabinose binding protein P254G/D-fucose
1apt	9.4	1.8	penicillopepsin/pepstatin analogue
1apu	7.49	1.8	penicillopepsin/IvaValValSta-OEt
1apv	9.00	1.8	penicillopepsin/IvaValVal(H)Dfo-N-methylamide
1apw	8.00	1.8	penicillopepsin/IvaValValDfo-N-methylamide
1bap	6.85	1.75	L-arabinose binding protein P254G/L-arabinose
1bra	1.82	2.2	trypsin mutant/benzamidine
1cbx	6.35	2.0	carboxypeptidase A/L-benzylsuccinate
1cla	5.28	2.34	chloramphenicol acetyltransferase/chloramphenicol
1cps	6.66	2.25	carboxypeptidase A/CPM
1csc	7.10	1.7	citrate synthase/carboxymethyl coenzyme A
1csc	1.62	1.7	citrate synthase/L-malate
1dbb	9.00	2.7	DB3/progesterone
1dbj	7.68	2.7	DB3/aetiocholanolone
1dbk	8.09	3.0	DB3/5-b-androstanedione
1dbm	9.44	2.7	DB3/progesterone analogue
1dhf	7.4	2.3	DHFR/folate
1dih	5.74	2.2	dihydrodipicolinate R/NADPH
1dr1	5.57	2.2	dihydrofolate reductase/biopterin
1drf	7.44	2.0	dihydrofolate reductase/folate
1dwb	2.90	3.16	thrombin/benzamidine
1dwc	7.41	3.0	thrombin/MD-805
1dwd	8.18	3.0	thrombin/NAPAP
1ebg	10.82	2.1	enolase/phosphonoacetohydroxamate
1etr	7.41	2.2	thrombin/MQPA
1ets	8.22	2.3	thrombin/NAPAP
1ett	6.19	2.5	thrombin/4-TAPAP
1fbc	6.26	2.6	fructose-1,6-bisphosphatase/2,5-anhydroglucitol-1,6-bisphosphate
1fbf	6.00	2.7	fructose-1,6-bisphosphatase/2,5-anhydromannitol-1,6-bisphosphate
1fbp	4.82	2.5	fructose-1,6-bisphosphatase/AMP
1fkb	9.70	1.7	FK506 binding protein/rapamycin
1fkf	8.77	1.7	FK506 binding protein/FK506
1gst	4.68	2.2	glutathione S-transferase/glutathione
1hbv	6.37	2.3	HIV-1 protease/SB-203238
1hpv	9.22	1.9	HIV-1 protease/VX-478
1hsl	7.30	1.89	histidine binding protein/Histidine
1htf	8.09	2.2	HIV-1 protease/GR-126045
1htg	9.68	2.0	HIV-1 protease/GR-137615
1hvi	10.07	1.8	HIV-1 protease/A-77003
1hvj	10.45	2.0	HIV-1 protease/A-78791
1hvk	10.11	1.8	HIV-1 protease/A-76928
1hvl	9.00	1.8	HIV-1 protease/A-76889
1hvs	10.08	2.25	HIV-1 protease V82A/A-77003
1l83	3.40	1.70	lysozyme/benzene
1ldm	5.44	2.1	M4 lactate dehydrogenase/NAD
1lgr	3.07	2.8	glutamine synthetase/AMP
1lyb	11.42	2.5	cathepsin D/pepstatin
1mcb	4.84	2.7	immunoglobulin/peptide

Table 1 *Training set used in SCORE (continued)*

PDB	pK_d	Resl.	Protein/ligand
1mcf	5.15	2.7	immunoglobulin/peptide
1mch	5.15	2.7	immunoglobulin/peptide
1mcj	3.78	2.7	immunoglobulin/peptide
1mcs	4.84	2.7	immunoglobulin/peptide
1mdq	5.10	1.9	maltose binding protein A301GS/maltose
1mfe	5.31	2.0	immunoglobulin/D-gal-D-abe-D-man
1mnc	9.00	2.1	neutrophil collagenase/hydroxamate
1nnb	5.30	2.8	neuraminidase/DANA
1phh	7.35	2.3	p-hydroxylbenzoate hydroxylase/FAD
1pgp	5.70	2.5	6-PGDH/6-phosphogluconic acid
1ppc	6.16	1.8	trypsin/NAPAP
1pph	6.22	1.9	trypsin/3-TAPAP
1ppk	7.66	1.8	penicillopepsin/phospho analogue
1ppl	8.55	1.7	penicillopepsin/IVA-VAl-VAl-LEU-P-(O)PHE-OME
1ppm	5.80	1.7	penicillopepsin/CBZ-ALA-ALA-LEU-P-(O)PHE-OME
1rbp	6.72	2.0	retinol binding protein/retinol
1rne	8.70	2.4	renin/CGP-38560
1rnt	5.18	1.9	ribonuclease T1/2'-GMP
1rus	3.08	2.9	rubisco/3-phosphoglycerate
1snc	6.70	1.65	staphylococcal nuclease/deoxythymidine 3',5'-bisphosphate
1tha	5.35	2.0	transthyretin/3,3'-diiodo-L-thyronine
1tlp	7.56	2.3	thermolysin/phosphoramidon
1tmn	7.47	1.9	thermolysin/N-(1-carboxy-3-phenyl)-L-LeuTrp
1tmt	6.24	2.2	thrombin/D-Phe-Pro-Arg
1tng	2.93	1.8	trypsin/aminomethylcyclohexane
1tnh	3.37	1.8	trypsin/4-fluorobenzylamine
1tni	1.70	1.9	trypsin/4-phenylbutylamine
1tnj	1.96	1.8	trypsin/2-phenylethylamine
1tnk	1.49	1.8	trypsin/3-phenylpropylamine
1tnl	1.88	1.9	trypsin/t-2-phenylcyclopropylamine
1ulb	4.40	2.75	PNP/guanine
1xli	1.48	2.5	D-xylose isomerase/5-thio-alpha-D-glucose
2ak3	3.86	1.9	adenylate kinase isoenzyme-3/AMP
2cgr	7.27	2.2	immunoglobulin/GAS
2csc	3.36	1.7	citrate synthase/D-malate
2ctc	3.89	1.4	carboxypeptidase A/L-phenyl lactate
2dbl	8.70	2.9	DB3/pregnane analogue
2dri	6.52	1.6	D-ribose binding protein/b-D-ribose
2gbp	7.40	1.9	galactose binding protein/galactose
2ifb	5.44	2.0	fatty acid binding protein/C15COOH
2ldb	4.15	3.0	L-lactate dehydrogenase/NAD+
2mcp	4.70	3.1	immunoglobulin/phosphocholine
2phh	3.36	2.7	PHBH/ADP ribose
2phh	4.60	2.7	PHBH/p-hydroxybenzoic acid
2pk4	4.32	2.25	plasminogen kringle 4/aminocaproic acid
2r04	6.22	3.0	virus coat protein/compound 4
2rnt	3.78	1.8	ribonuclease T1 K25/guanylyl-2',5'-guanosine
2sns	6.70	1.5	staphylococcal nuclease/2'-deoxy-3',5'-diphosphothymidine
2tmn	5.89	1.6	thermolysin/N-phosphory-L-leucinamide
2xim	2.28	2.3	D-xylose isomerase K253R/xylitol
2xis	5.82	1.71	xylose isomerase/xylitol
2ypi	4.82	2.5	TIM/phosphoglycolic acid
3cla	4.94	1.75	chloramphenicol acetyltransferase/chloramphenicol

Table 1 *Training set used in SCORE (continued)*

PDB	pK_d	Resl.	Protein/ligand
3cpa	4.00	2.0	carboxypeptidase A/GT
3csc	5.15	1.9	citrate synthase/acetyl coenzyme A
3csc	2.64	1.9	citrate synthase/L-malate
3fx2	9.3	1.9	flavodoxin/FMN
3gap	5.00	2.5	catabolite gene activator protein/cAMP
3pgm	3.19	2.8	phosphoglycerate mutase/phosphoglycerate
3ptb	4.50	1.7	trypsin/benzamidine
3tmn	5.90	1.7	thermolysin/ValTrp
4cla	5.47	2.0	chloramphenicol acetyltransferase/chloramphenicol
4dfr	8.62	1.7	DHFR/methotrexate
4fab	8.05	2.7	IgG kappa Fab 4-4-20/fluorescein dianion
4gr1	2.20	2.4	glutathione reductase/retro-GSSG
4hvp	6.11	2.3	HIV-1 protease/MVT-101
4mdh	3.23	2.5	cytoplasmic malate/NAD+
4phv	9.17	2.10	HIV-1 protease/L-700417
4sga	3.27	1.8	proteinase A/Ace-Pro-Ala-Pro-Phe
4tim	2.16	2.4	triosephosphate isomerase/2-phosphoglycerate
4tln	3.72	2.3	thermolysin/Leu-NHOH
4tmn	10.17	1.7	thermolysin/ZFpLA
4ts1	5.61	2.5	Tyrosyl-transfer RNA synthetase/tyrosine
4xia	1.54	2.3	D-xylose isomerase/D-sorbitol
5abp	6.64	1.8	ABP/D-galactose
5acn	2.80	2.1	aconitase/tricarballylic acid
5cna	2.00	2.0	concanavalin A/a-Me-D-mannopyranoside
5enl	3.8	2.2	enolase/2-phospho-D-glycerate
5hvp	7.46	2.0	HIV-1 protease/acetylpepstatin
5icd	5.29	2.5	isocitrate dehydrogenase/isocitrate
5ldh	2.82	2.7	lactate dehydrogenase/isocitrate
5p21	5.32	1.35	ras p21 protein/GPPNP
5sga	2.85	1.8	proteinase A/Ace-Pro-Ala-Pro-Tyr
5tim	2.30	1.83	triosephosphate isomerase/DTT
5tln	6.37	2.3	thermolysin/INA
5tmn	8.04	1.6	thermolysin/ZGp(NH)LL
5xia	2.60	2.5	D-xylose isomerase/xylitol
6abp	6.36	1.67	L-arabinose binding protein M108L/L-arabinose
6apr	7.77	2.5	rhizopuspepsin/pepstatin
6cpa	11.52	2.0	carboxypeptidase A/ZAAp(O)F
6enl	3.0	2.2	enolase/phosphoglycolic acid
6rnt	2.37	1.8	ribonuclease T1/2'-AMP
6tim	3.21	2.2	triosephosphate isomerase/glycerol-3-phosphate
6tmn	5.05	1.6	thermolysin/ZGp(O)LL
7abp	6.46	1.67	L-arabinose binding protein M108L/D-fucose
7acn	4.31	2.0	aconitase/isocitrate
7cat	8.00	2.5	catalase/NADPH
7cpa	13.96	2.0	carboxypeptidase A/BZ-FVP(O)F
7dfr	4.96	2.5	DHFR/folate
7dfr	6.10	2.5	DHFR/NADP+
7est	7.60	1.8	elastase/TFAP
7hvp	9.62	2.4	HIV-1 protease/JG-365
7tim	5.40	1.9	triosephosphate isomerase/phosphoglycolohydroxamate
7tln	2.47	2.3	thermolysin/CH2CO-Leu-OCH3
8abp	6.60	1.49	L-arabinose binding protein M108L/D-galactose
8acn	7.14	2.0	aconitase/nitroisocitrate

Table 1 *Training set used in SCORE (continued)*

PDB	pK_d	Resl.	Protein/ligand
8atc	7.57	2.5	aspartate carbamoyltransferase/PALA
8cpa	9.15	2.0	carboxypeptidase A/BZ-AGP(O)F
8hvp	9.00	2.5	HIV-1 protease/U-85548E
8icd	3.02	2.5	isocitrate dehydrogenase/isocitrate
8xia	2.95	1.9	D-xylose isomerase/D-xylose
9aat	8.22	2.2	aspartate aminotransferase/pyridoxal-5'-phosphate
9abp	8.00	1.97	L-arabinose binding protein P254G/D-galactose
9hvp	8.35	2.8	HIV-1 protease/A-74704
9ldt	5.43	2.0	lactate dehydrogenase/NADH
9ldt	4.74	2.0	lactate dehydrogenase/oxamate
9rub	4.70	2.6	rubisco/ribulose-1,5-bisphosphate

Scoring function

We assume that the free energy change in the protein-ligand binding process can be dissected into basic components. Our scoring function takes the following form.

$$pK_d = K_0 + K_{vdw} + K_{metal} + K_{hbond} + K_{desolvation} + K_{deformation} \quad (1)$$

Here, K_{vdw} represents the contribution of van der Waals interaction between the protein and its ligand, K_{metal} the contribution of metal-ligand bonding, K_{hbond} the contribution of hydrogen bonding, $K_{desolvation}$ the contribution of desolvation effect, and $K_{deformation}$ the contribution of deformation. K_0 is the regression constant which may contain the translational and rotational entropy loss upon the binding process.

(1) Van der Waals (VDW) interaction. This kind of interaction is a balance between attractive dispersion force and short-range repulsion. Although it is well accepted that van der Waals interactions play a fundamental role in the binding of the protein and its ligand, arguments exist in how to represent it in calculating the binding affinity. Some researchers assume that protein-ligand, protein-solvent, and ligand-solvent interfaces are well packed and hence neglect any change in the VDW interactions upon binding. Some others assume that VDW interactions are better in a complex and therefore explicitly include them. By analyzing the training set, we believe that they all tell only part of the story. In general, one will find a closely packed interaction interface in a protein-ligand complex where many atom pairs are in a distance much shorter than the sum of their VDW radii, i.e. they form VDW bumps. Not all of these bumps come from hydrogen bonding or other strong interactions. Some of them are just the result of the tight binding between other parts of the protein and its ligand. It is not reasonable to assume that such a situation can also be found on the protein-solvent or ligand-solvent interface where the water molecules are mobile. Thus, our conclusion is that the VDW attraction between the protein and its ligand can be neglected due to the competitive interaction with water in the unbound state while the VDW repulsion cannot.

In our algorithm, the term for VDW interaction is simply a pairwise counting of VDW bumps between the protein and the ligand,

$$K_{vdw} = \sum_i \sum_j VB(d_{ij}) \quad (2)$$

$$VB(d_{ij}) = 1 \quad d_{ij} < r_i + r_j - 0.60\text{Å}$$
$$= 0 \quad d_{ij} \geq r_i + r_j - 0.60\text{Å}$$

where r_i is the VDW radius of ligand atom *i* and r_j is the VDW radius of protein atom *j*; d_{ij} is the distance between atom *i* and *j*.

(2) Metal-ligand bonding. A variety of proteins have metal cations in their active sites, such as Mg^{2+}, Ca^{2+}, Mn^{2+}, and Zn^{2+}. In such cases, coordinate bonding between the metal and the ligand can often be important for the stability of the complex. In our algorithm, the metal cation in the active site is treated as part of the protein and metal-ligand bonding is distinguished from hydrogen bonding. By browsing the *International Tables for Crystallography* [32], we find that, in common coordinate compounds, most of Mg/Ca/Mn/Zn ... O/N bonds lengthen between 1.9Å to 2.2Å. As an approximation, the ideal bond length of a metal ... O/N bond is set to 2.0Å in our algorithm and a distance function is used to account for the deviation from the ideal value,

$$MB(d) = 1.0 \quad d < 2.0\text{Å}$$
$$= 3.0 - d \quad 2.0\text{Å} \leq d < 3.0\text{Å}$$
$$= 0 \quad 3.0\text{Å} \leq d$$

where d is the metal ... O/N bond length. The cutoff of 3.0Å comes from the observation that there is no metal-ligand bond longer than this in the entire training set.

The term for metal-ligand bonding in our algorithm is the sum over all metal-O/N bonds,

$$K_{metal} = \sum_i \sum_j MB\left(d_{ij}\right) \quad (3)$$

where d_{ij} refers to the distance between ligand atom *i* and metal *j*. We do not differentiate the types of metal ... O/N bonds so that no weight factor is needed.

(3) Hydrogen bonding. Hydrogen bonding is no doubt one of the key features for a specific binding process. Such interaction may happen when two atoms get close enough and form a donor-acceptor pair. In our algorithm, a hydrogen bond donor is defined as a nitrogen or oxygen atom with hydrogen attached; while an acceptor is defined as a nitrogen, oxygen, or fluorine atom with at least one vacant valence to accept a hydrogen atom. Accordingly, all the atoms on the protein and the ligand are labeled as either donor (D), acceptor (A), donor/acceptor (DA), or none (N).

The geometry of a hydrogen bond is characterized by two parameters: the bond length, i.e. the distance between D and A, and the bond angle, i.e. the angle among D-H …A. The calculation of the former is straightforward. We define that a hydrogen bond is possible only when the bond length is shorter than the sum of VDW radii of D and A. However, the calculation of the bond angle is of some difficulty since we avoid the explicit use of hydrogen atoms in the structure. To circumvent this problem, we use two other angles involving only heavy atoms instead. They are computed among X-D…A and D…A-X, where X represents the adjacent heavy atom or, if there are more than one adjacent atom, their geometric centres. We have investigated the distribution of these two angles among all typical kinds of hydrogen bond and found that they are not likely to be lower than 70 degrees for a plausible hydrogen bond. Thus, an angle cutoff is set in our algorithm for defining a hydrogen bond: if either of these two angles is lower than 70 degrees, the geometry of the donor-acceptor pair under investigation is poor and therefore overlooked. In the case of water-involved hydrogen bond in which water has no adjacent heavy atom, only the possible angle is used in judgement.

In other empirical methods [26], the distance dependence of hydrogen bonding strength is gauged typically by using a linear distance function which decreases from 1 to 0 in a certain range. Such definition is rather subjective since hydrogen bonding need not behave in such a simple and ideal manner. In our algorithm, a step function is used instead. We define,

SHB (d)	= 1	$d < d_0$ - 0.60Å
	= 0	otherwise
MHB (d)	= 1	d_0 - 0.60Å ≤ $d < d_0$ – 0.30Å
	= 0	otherwise
WHB (d)	= 1	d_0 – 0.30Å ≤ $d < d_0$
	= 0	otherwise

where d represents the distance between D and A; d_0 represents the sum of VDW radii of D and A. SHB, MHB, and WHB are indicators for strong, moderate, and weak hydrogen bond respectively. In addition, because of the specificity of water-involved hydrogen bond, we use another three indicators defined in the same way

SWH (d)	= 1	$d < d_0$ - 0.60Å
	= 0	otherwise
MWH (d)	= 1	d_0 - 0.60Å ≤ $d < d_0$ – 0.30Å
	= 0	otherwise
WWH (d)	= 1	d_0 – 0.30Å ≤ $d < d_0$
	= 0	otherwise

to represent strong, moderate, and weak water-involved hydrogen bond respectively. The contribution of the six terms above will be determined by regression and therefore the use of an imagined distance function is avoided.

The angular dependence of hydrogen bonding strength proved to be another problem. In our method, we do not use a function to account for this for two reasons. First, it is difficult to design such a function. Many researchers, including us, have investigated the distribution of hydrogen bond angle by analyzing various databases of small molecules or macromolecules. The general conclusion is that, although some kinds of hydrogen bonds tend to favour certain orientations, there is usually a wide distribution in hydrogen bond angles. Thus, it seems quite unpractical to using one angular function to deal with all kinds of hydrogen bonds. Second, it may not be necessary to design such a function. In fact, there is not enough experimental evidence to explain how a hydrogen bond acts if its angle deviates from the "ideality". Therefore, it is not surprising that there is no standard method to gauge the angular dependence of hydrogen bonding strength at present. We have tried some simple-formed angular functions while developing our model. However, such attempts were proved to help little.

In our method, the contribution of hydrogen bonding is the sum over all hydrogen bonds formed between the protein and its ligand.

$$\begin{aligned} K_{hbond} &= K_{SHB} + K_{MHB} + K_{WHB} + K_{SWH} + K_{MWH} + K_{WWH} \\ &= \sum_i \sum_j SHB\left(d_{ij}\right) + \sum_i \sum_j MHB\left(d_{ij}\right) \\ &+ \sum_i \sum_j WHB\left(d_{ij}\right) + \sum_i \sum_j SWH\left(d_{ij}\right) \\ &+ \sum_i \sum_j MWH\left(d_{ij}\right) + \sum_i \sum_j WWH\left(d_{ij}\right) \end{aligned} \quad (4)$$

Here, we do not differentiate the types of hydrogen bonds so that no weight factor is needed.

(4) Desolvation effect. Since both the protein and its ligand are solvated before complexation, the protein-ligand binding process is accompanied by desolvation, which undergoes changes in entropy as well as in enthalpy. This kind of effect

is very difficult to characterize accurately. Both the long-range "hydrophobic effect" and specific hydrogen bondings of water could be important in elucidation. These features make it unamenable to simple additive pairwise interactions. So far this effect is typically measured by calculating the buried hydrophobic surface areas during the binding process [26-29]. However, several defects lie in such approaches: (i) polar and non-polar atoms are differentiated by very coarse definitions; (ii) the choice of solvent accessible surface or molecular surface seems to depend totally on the researcher's intuition; (iii) it is not always clear which buried surface should be calculated, (a) only the protein, (b) only the ligand, or (c) both the protein and the ligand; (iv) accurate, analytical algorithm for calculating the surface is impossible. Numerical integration has to be used as an approximation.

We have adopted a different method to measure the desolvation effect, which is simple and explicit. First, each atom is assigned a quantitative scale to represent its hydrophobicity. In a previous paper [33], we have reported a method, XLOGP, to calculate the octanol/water partition coefficient for an organic compound. The basic idea was that the logP value of the whole molecule could be expressed as the summation of atomic contributions. The contributions of different atom types were derived from the regression analysis of a large set of compounds. These parameters are therefore transplanted into this study as the atomic hydrophobic scales: the more positive the value, the more hydrophobic is the atom; and the more negative the value, the more hydrophilic is the atom. In our algorithm, an atom is considered as hydrophobic when its hydrophobic scale is larger than 0.20 units. Second, the "environment" of a ligand atom is defined as the assembly of all the neighbouring protein atoms within 5Å. The hydrophobicity of the environment is expressed by the sum of the hydrophobic scales of all the protein atoms forming the environment. If the sum is positive, the ligand atom is considered to be in a hydrophobic environment, otherwise it is considered to be in a hydrophilic environment. Therefore, in principle there will be four situations: a hydrophilic ligand atom in a hydrophilic environment, a hydrophilic ligand atom in a hydrophobic environment, a hydrophobic ligand atom in a hydrophilic environment, and a hydrophobic ligand atom in a hydrophobic environment. In the first three situations, the lose of solvation shell (desolvation) during binding is compensated more or less by the interaction with a hydrophilic counterpart. Hence, a significant change in the overall binding free energy is not expected. However, in the last situation, perfect hydrophobic matching forms and thus makes a favourable contribution to the protein-ligand binding process.

In our algorithm, the term for desolvation effect is a sum over all hydrophobic matchings between the protein and its ligand,

$$K_{HM} = \sum_i F_i \times HM_i \tag{5}$$

where HM_i is an indicator of hydrophobic matching. It is set to 1 if ligand atom *i* is hydrophobic and placed in a hydrophobic environment; otherwise it is set to 0. F_i is the atomic hydrophobic scale of ligand atom *i*. It is used as a weight factor here to meet the expectation that a more hydrophobic atom may contribute more to the hydrophobic effect.

(5) Deformation effect. Deformation refers to the conformational changes during the binding process. On one hand, it causes adverse entropic changes due to the freezing of internal rotations of both the protein and its ligand; on the other hand, it causes adverse enthalpic change due to the strain energy exerted during binding. Based on the principles of statistical thermodynamics, the entropic change is usually estimated by using a constant value per rotatable bond that is frozen. However, the enthalpic change is more difficult to elucidate.

We have noticed that Nicklaus et al., in an informative approach [34], found that the deformational enthalpy of the ligand upon the binding process also correlates well with the number of rotatable bonds. Therefore, as a simplification, we use the number of rotatable bonds, i.e. rotors, as a double-purposed parameter to estimate both the entropic and enthalpic change in deformation. In our algorithm, the term for the deformation effect is simply the number of rotors in the ligand. If a rotor is split into halves and assigned onto the two atoms involved, this term can also be written as,

$$K_{RT} = \sum_i 0.5 \times RT_i \tag{6}$$

where RT_i is the number of rotors in which ligand atom *i* is involved. Rotor is defined as acyclic sp^3-sp^3 and sp^3-sp^2 single bond. Rotations of terminal $-CH_3$, $-NH_2$, or -OH, whose rotation do not produce any new conformation of heavy atoms are not taken into account. The flexibility of cyclic portions of the ligand is ignored.

We have also tried to incorporate the deformation effect of the protein into computation by simply counting the rotors of the active site residues or using other protein side-chain entropic scales [35]. But such attempts simply did not help to improve the result. One possible reason is that, even in the unbound state, the side chains of the active site residues are generally immobilized due to the stack of neighbouring residues. Therefore in most cases, unlike the ligand, the protein changes very little to accommodate the ligand. Another reason may be that empirical parameters are too rough to model the deformation of the protein. Such behaviour needs to be modeled by more rigorous and exhaustive dynamic simulations, which is definitely unpractical for a fast empirical method. Thus, as an simplification, we neglect the deformation effect of the protein during the binding process in our method.

At this point, a summary of our scoring function should be given. We compute the dissociation constant of a protein-ligand complex by summing all the terms described above.

$$pK_d = K_0 + c_1 \times K_{VB} + c_2 \times K_{MB} + c_3 \times K_{SHB} + c_4 \times K_{MHB} + c_5 \times K_{WHB} + c_6 \times K_{SWH} + c_7 \times K_{MWH} + c_8 \times K_{WWH} + c_9 \times K_{HM} + c_{10} \times K_{RT} \quad (7)$$

There are a total of 11 adjustable parameters in this scoring function. They will be given by the regressional analysis of the entire training set.

According to our algorithm, all the terms in the scoring function can be computed on the sum over the contribution of ligand atoms. After simple linear transformation, the function can be rewritten as following,

$$\begin{aligned} pK_d &= K_0 + c_1 \times K_{VB} + \ldots + c_{10} K_{RT} \\ &= K_0 + c_1 \times \sum_i K_{VB,i} + \ldots + c_{10} \times \sum_i K_{RT,i} \\ &= K_0 + \sum_i \left(c_1 \times K_{VB,i} + \ldots + c_{10} \times K_{RT,i} \right) \\ &= K_0 + \sum_i K_i \end{aligned} \quad (8)$$

in which the binding affinity of the whole ligand is expressed as the addition of the contributions of each ligand atom. We call K_i the atomic binding score. It characterizes the role of an individual ligand atom during the binding process in a quantitative way. Its potential application will be discussed later in this paper.

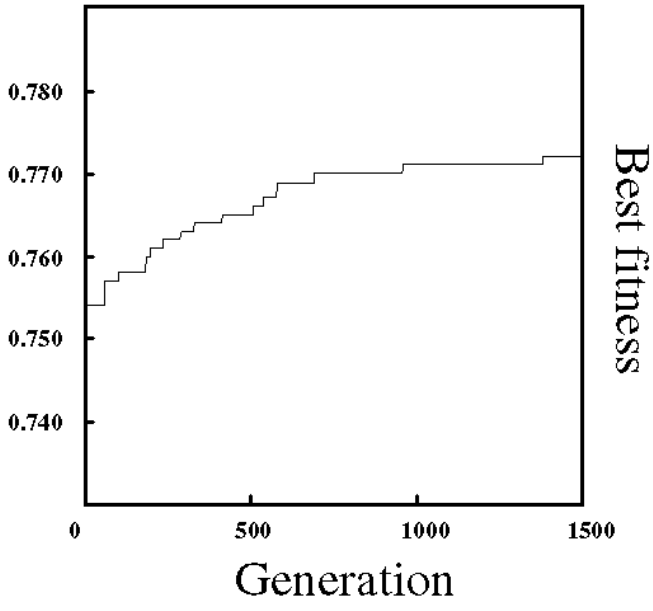


Figure 1 *The best fitness observed among the whole population along a GA procedure*

Table 2 *Atom types and VDW radii used in SCORE*

Symbol	Description	radius (Å)
C.3	sp3 hybridized carbon	1.94
C.2	sp2 hybridized carbon	1.90
C.1	sp hybridized carbon	1.90
C.ar	aromatic carbon	1.85
O.3	sp3 hybridized oxygen	1.74
O.2	sp2 hybridized oxygen	1.66
O.w	water oxygen	1.77
N.3	sp3 hybridized nitrogen	1.87
N.2	sp2 hybridized nitrogen	1.86
N.1	sp hybridized nitrogen	1.86
N.ar	aromatic nitrogen	1.86
N.am	amide nitrogen	1.83
N.pl3	trigonal planar nitrogen	1.86
S.3	sp3 hybridized sulfur	2.09
S.2	sp2 hybridized sulfur	2.01
S.o	sulfoxide sulfur	2.01
S.o2	sulfone sulfur	2.01
F	fluorine	1.77
Cl	chlorine	2.00
Br	bromine	2.22
I	iodine	2.42
P	phosphor	2.03

VDW radius set

As described above, we use VDW radii in the calculation of VDW bump and hydrogen bonding. Since each force field has its own set of VDW radii, making a choice among them is rather a subjective issue. When developing our scoring function, we originally adopted the VDW radius set from the AMBER force field [36] as it is well established for modeling macromolecules. By using this set of VDW radii in the scoring function, we obtained promising results in the regressional analysis of the training set. But since the AMBER force field is parameterized to reproduce internal properties, such as conformation, dipole moment, and heat of formation, we believe that some optimisation on this VDW radius set is necessary for the purpose of binding affinity estimation. We optimized it by applying a Genetic Algorithm (GA) [37].

In a GA procedure, potential solutions to the problem being studied are represented as data structures called chromosomes. For our problem, a real-value string chromosome is used. We adopt 22 atom types defined in the Tripos force field to classify carbon, oxygen, nitrogen, sulfur, phosphor, and halogens (see Table 2). A chromosome thus contains 22 elements to represent the VDW radii for all the atom types. For each chromosome, regressional fitting of the whole training set is done by using Equation 7 and the VDW radius set the chromosome represents. The fitness of the chromosome is given the value of the squared correlation coefficient there-

Table 3 *Coefficients of each term in the final scoring function*

Term	description	Coefficient [a]	
(1)	VDW bump (VB)	-0.168	(±0.110)
(2)	Metal-ligand bonding (MB)	0.916	(±0.580)
(3)	Strong hydrogen bonding (SHB)	0.593	(±0.198)
(4)	Moderate hydrogen bonding (MHB)	0.216	(±0.170)
(5)	Weak hydrogen bonding (WHB)	0.141	(±0.125)
(6)	Strong water-involved H-bonding (SWH)	0.291	(±0.259)
(7)	Moderate water-involved H-bonding (MWH)	-0.708	(±0.313)
(8)	Weak water-involved H-bonding (WWH)	0.327	(±0.258)
(9)	Hydrophobic matching (HM)	1.178	(±0.253)
(10)	Rotor (RT)	-0.169	(±0.081)
Regression constant		2.254	(±0.914)

[a] The value in brackets is 95% confidence interval

fore obtained in the regression. Thus, an ideal chromosome will have the fitness of 1.00. In this way, each chromosome is directly evaluated according to its ability to reproduce the binding affinities of the training set. The GA operators used in our study include mutation and crossover. Here, mutation is a single-point mutation, i.e. only one randomly selected element in the chromosome is changed by a certain amount. The amount is designed to be a random number in the Gaussian distribution with zero mean and a variance of 0.02Å. Mutation requires one parent chromosome and produces only one child. Crossover is also a single-point crossover, i.e. a position along the chromosome is selected at random and all the elements subsequent to the chosen position are then swapped over between the two chosen chromosomes. Crossover requires two parent chromosomes and produces two children.

An initial population of chromosomes is generated by mutation on the original AMBER VDW radius set. After generating the initial population, the GA then runs in cycles. Roulette-wheel selection is used to choose parents for producing new members for the next generation. It works by giving each member of the population a slice of the wheel, the size of the slice being proportional to the fitness of the member. In this way, when the wheel is spun, a fitter member will be more likely to be chosen than a less fit member. We adopt steady-state-with-no-duplicates strategy in our GA procedure. In each cycle, a new chromosome is produced either by mutation or crossover on the selected parent. After duplicate check, it is compared with the worst member of the existing population. If the new one is better, it becomes a member of the population and the original worst one is discarded; if not, the new one is discarded and GA goes into next generation with the population unchanged. This process is repeated until a pre-set limit of generation is reached.

We ran the above GA procedure with a population size of 100, a generation limit of 1500, a mutation rate of 0.50, and a crossover rate of 0.50. During the procedure, we recorded the average and the best fitness of the whole population. Because of the nature of GA, the procedure has been run a number of times to find the best solution to the problem. A typical run is shown in Figure 1 in which the best fitness of the population has been optimized from 0.754 to 0.777. In fact all the runs could optimize the average and the best fitness of the population to approximately the same level. The best set of VDW radii found among all solutions is listed in Table 2. All the values in this set seem to be reasonable and, since hydrogen is bypassed in the computation, they can be considered as the radii for united atoms. This VDW radius set is adopted in following computations.

Regression

Using Equation 7, standard multivariate regression was performed on the whole training set. It yielded a squared correlation coefficient (r^2) of 0.777, a standard deviation (s) of 1.16 log units, which corresponds to 6.6 kJ/mol in binding

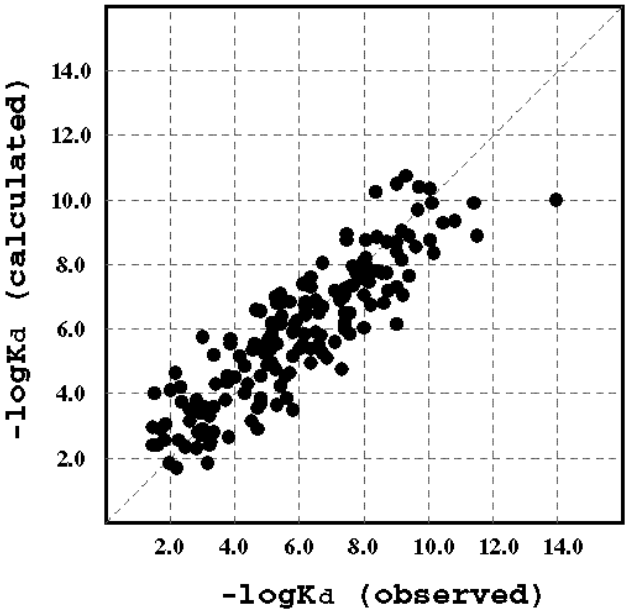


Figure 2 *The correlation between the experimental and calculated pK_d values of 170 complexes in the training set*

Table 4 *Protein-ligand complexes in the test set*

PDB entry	Resl. (Å)	Protein/ligand	Exp.[a]	Pred.[b]
1eed	2.0	endothiapepsin/PD-125754	4.90	6.15
1epo	2.0	endothiapepsin/CP-81282	7.96	8.84
1epp	1.9	endothiapepsin/PD-130693	7.16	6.58
2er0	3.0	endothiapepsin/L-364099	6.40	7.86
2er6	2.0	endothiapepsin/H-256	7.22	6.99
2er7	1.6	endothiapepsin/H-261	9.00	8.67
2er9	2.2	endothiapepsin/L-363564	7.80	7.83
3er3	2.0	endothiapepsin/CP-71362	7.10	6.90
4er1	2.0	endothiapepsin/PD-125967	6.62	7.69
4er2	2.0	endothiapepsin/pepstatin	9.30	9.27
4er4	2.1	endothiapepsin/H-142	6.80	7.00

[a] Experimentally determined pK_d values.
[b] Predicted pK_d values given by SCORE.

free energy at room temperature, and a Fisher value F(10,159) of 57.8. The coefficients of each term in the scoring function are listed in Table 3. The correlation between the experimental and calculated pK_d values of the training set is shown in Figure 2.

A leave-one-out cross-validation was also performed on the training set and yielded a squared correlation coefficient (q^2) of 0.743 and a standard deviation (s_{PRESS}) of 1.10 log units, which corresponds to 6.3 kJ/mol in binding free energy at room temperature. Here, q^2 and s_{PRESS} are defined as

$$q^2 = \frac{1 - \sum \left(y_{pred} - y_{obs}\right)^2}{\sum \left(y_{obs} - y_{mean}\right)^2} \tag{9}$$

and

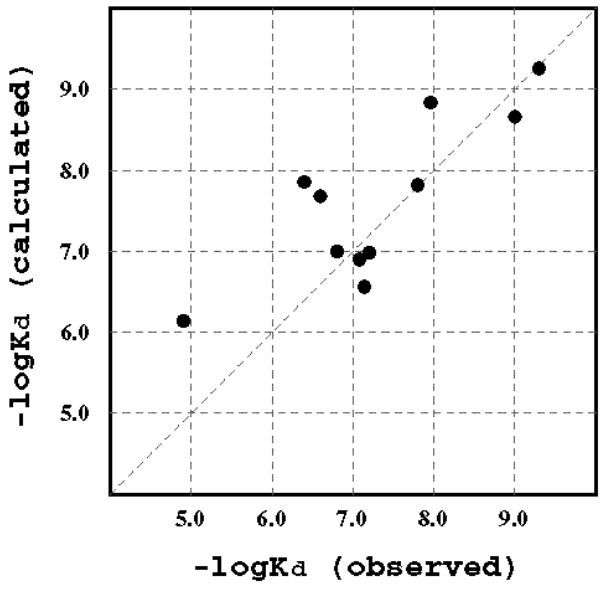


Figure 3 *The correlation between the experimental and calculated pK_d values of 11 endothiapepsin complexes in the test set*

$$s_{PRESS} = \sqrt{\frac{\sum \left(y_{pred} - y_{obs}\right)^2}{\left(N - k - 1\right)}} \tag{10}$$

where N represents the number of samples in the training set and k represents the number of terms in the scoring function.

Test set

The true value of any empirical model lies in its predictive ability. In this study, we have used 11 endothiapepsin complexes as the test set (see Table 4). This set was chosen for two reasons: first, there is no other endothiapepsin complex in the training set; second, the ligands in these complexes are peptides, which are generally larger and more flexible than the ligands in the training set. Therefore, this test set tends to be a real challenge.

Applying our scoring function to the test set yielded a predictive squared correlation coefficient (r^2_{pred}) of 0.654 and a standard deviation (s_{pred}) of 0.55 log units (3.2kJ/mol at 298K). The correlation between the experimental and predicted pK_d values of the test set is shown in Figure 3.

Evolutionary test

We have designed a stepwise procedure, Evolutionary Test, to confirm the robustness and the internal consistency of the scoring function. This procedure was started from constructing a data set of 30 complexes. The complexes were randomly selected from the training set without duplicate. Then, standard regression and leave-one-out cross-validation were performed on this data set by using the scoring function. The model thus obtained was also applied to the test set. To minimize the coincidence in such analysis, the above process was repeated 20 times and the average values of all the results, including r^2, s, q^2, s_{PRESS}, r^2_{pred} , s_{pred}, and the coefficients of each term in the scoring function were recorded. The second step of the procedure was constructing a larger data set of 40

Table 5 *Coefficients of each term in the scoring function in the Evolutionary Test*

Size[a]	VB	MB	SHB	MHB	WHB	SWH	MWH	WWH	HM	RT	Const.
30	-0.171	1.135	0.668	0.138	0.185	0.297	-0.655	0.163	1.221	-0.154	1.865
40	-0.152	0.916	0.612	0.219	0.177	0.244	-0.619	0.264	1.225	-0.180	1.952
50	-0.179	0.998	0.646	0.184	0.152	0.310	-0.713	0.319	1.183	-0.155	2.152
60	-0.174	0.924	0.615	0.198	0.169	0.258	-0.629	0.325	1.213	-0.170	2.071
70	-0.175	0.962	0.613	0.197	0.168	0.255	-0.678	0.317	1.201	-0.164	2.109
80	-0.159	0.977	0.579	0.202	0.158	0.282	-0.690	0.335	1.227	-0.166	2.075
90	-0.174	0.901	0.612	0.180	0.155	0.268	-0.697	0.333	1.153	-0.152	2.268
100	-0.159	0.866	0.584	0.210	0.160	0.272	-0.682	0.290	1.165	-0.153	2.150
110	-0.164	0.865	0.597	0.198	0.149	0.235	-0.675	0.333	1.169	-0.154	2.227
120	-0.167	0.954	0.602	0.197	0.147	0.264	-0.718	0.323	1.165	-0.153	2.238
130	-0.170	0.927	0.599	0.186	0.157	0.274	-0.712	0.326	1.189	-0.160	2.228
140	-0.177	0.930	0.622	0.184	0.150	0.275	-0.687	0.321	1.177	-0.155	2.206
150	-0.171	0.932	0.606	0.187	0.154	0.278	-0.686	0.325	1.180	-0.158	2.207
160	-0.178	0.936	0.615	0.193	0.155	0.272	-0.685	0.336	1.192	-0.162	2.178

[a] Size of the data set.

complexes and repeating all the statistical analysis. Then, the size of the data set was increased to 50, 60, …, until to the original training set itself. The r^2, q^2, and r^2_{pred} obtained along this procedure are shown in Figure 4a; while the s, s_{PRESS}, and s_{pred} are shown in Figure 4b. The coefficients in the scoring function obtained along this procedure are listed in Table 5.

Program description

Based on the final scoring function obtained, we have written a program, SCORE, in the C++ language. All the necessary inputs to perform computation include a file storing the protein in PDB format and another file storing the corresponding ligand in MOL2 format. The program will read in the structures, assign atom types and parameters, do the calcula-

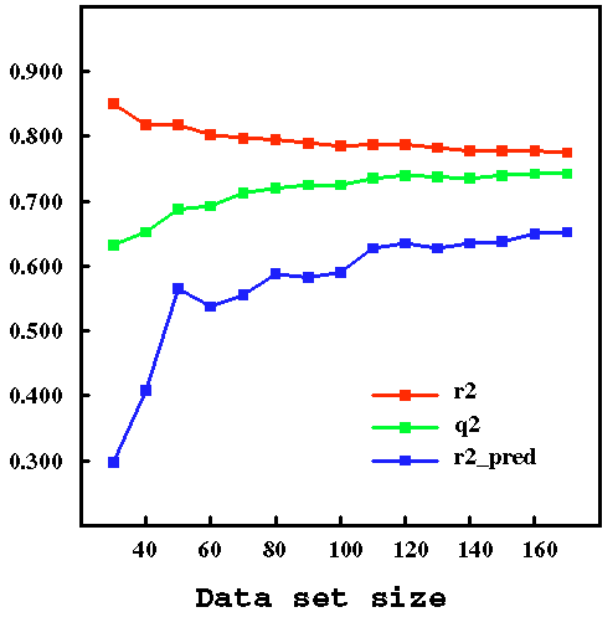


Figure 4a *The squared correlation coefficients observed in the Evolutionary Test*

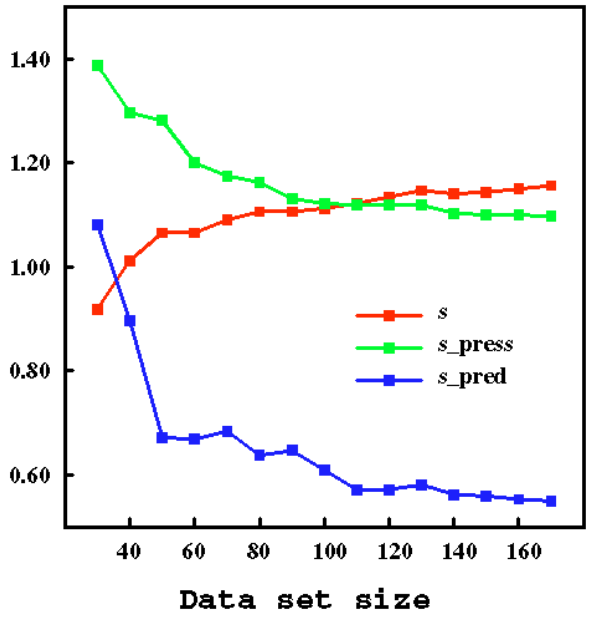


Figure 4b *The standard deviations (in log units) observed in the Evolutionary Test*

Table 6 *Comparison of SCORE with other similar approaches*

Approach	Bohm [a]	Head [b]	Gschwend [c]	Eldridge [d]	SCORE
Samples [e]	45	51	103	82	170
Terms [f]	5	13	8	5	11
r^2 [g]	0.762	0.85	0.745	0.710	0.777
s [h]	7.9	5.8	7.2	8.0	6.6
F [i]	32.1	17.8	39.6	–	57.8
q^2 [j]	0.696	0.78	0.701	0.658	0.743
s_{PRESS} [k]	9.3	6.5	–	8.7	6.3

[a] Ref. 26
[b] Ref. 27
[c] Ref. 28
[d] Ref. 29
[e] Number of complexes used in the training set.
[f] Number of terms in the scoring function
[g] Squared correlation coefficient given by regressional fitting.
[h] Standard deviation in regressional fitting, kJ/mol.
[i] Fisher significant ratio.
[j] Squared correlation coefficient given by leave-one-out cross-validation
[k] Standard deviation in leave-one-out cross-validation, kJ/mol

tion, and then give the dissociation constant of the given protein-ligand complex. The whole computing process for one complex is typically within a second on a SGI O2/R10000 workstation. The computational results are output into a text file in which the detailed information of each ligand atom, including the atomic binding score, is tabulated. Atomic binding scores are also written into the MOL2 file which stores the ligand structure so that the user can observe them directly in the SYBYL graphic interface. In addition, we classify ligand atoms into different built-in atom sets in the MOL2 file in terms of their atomic binding scores: a ligand atom with a binding score higher than 0.10 units is defined as "GOOD"; a ligand atom with a binding score lower than –0.10 units is defined as "BAD"; while a ligand atom otherwise is defined as "NEUTRAL". These atom sets can be viewed in different colours in SYBYL (see Figure 5). This makes the interpretation of protein-ligand interaction possible in a more straightforward manner.

Discussion

A state-of-the-art solution

We have presented a new empirical method for estimating the binding affinity of a protein-ligand complex with known three-dimensional structure. It uses a simple scoring function to capture the essential physics of the binding process. The final model achieved a standard deviation of 6.3 kJ/mol in the leave-one-out cross-validation analysis of 170 complex structures. When it was applied to the test set, which contains 11 endothiapepsin complexes, the standard error of prediction was only 3.2 kJ/mol. It will be naive to expect our simple empirical scoring function to supplant other theoretically rigorous and computational exhaustive methods like FEP. But estimating the absolute dissociation constant of a protein-ligand complex with an average error of approximately one magnitude is already a significant improvement for structure-based drug design schemes. Since our scoring function can be calculated on the fly, it could be incorporated into database searching or *de novo* design programs in which usually thousands of candidate compounds have to be ranked in a relatively short time. After necessary modification, it may also has potential application to other studies, such as molecular docking, host-guest chemistry, and protein engineering.

Our original motive of this study is to develop a scoring function for use specifically in protein-ligand binding affinity estimation rather than borrow a functional description from another branch of computational chemistry. There are several features of our method which deviate from molecular mechanics-based functions. First, we bypass the need for hydrogen atoms in computation. It is because exhaustive minimisation of the crystalline structure is usually needed to release the internal steric repulsion after adding hydrogen atoms and the placement of hydrogen atom on rotatable terminal group, such as hydroxyl group, is still a head-aching problem. Therefore we consider hydrogen atoms implicitly in the heavy atoms to which they attach and the computational results demonstrate that this strategy works. Second, we choose to use "soft" potentials to avoid the over-sensitivity to precise atomic positions. Therefore our scoring func-

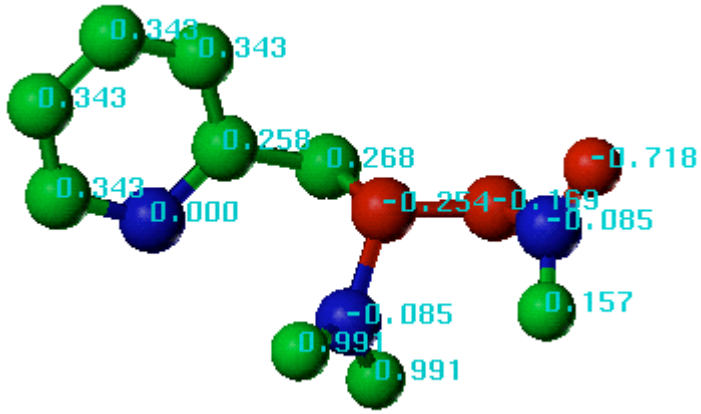


Figure 5 *Illustration of atomic binding score: good atoms in green; bad atoms in red; while neutral atoms in blue. (L-benzylsuccinate bound to carboxypeptidase A, PDB entry 1CBX)*

tion can tolerate small uncertainties in the coordinates due to the experimental determination or molecular modeling. This feature may be helpful for a practical structure-based drug design procedure. Third, although we used to include a term to calculate the "electrostatic interaction" by using traditional Coulomb equation, we found that it is simply unnecessary for *our* scoring function. Thus we skip problematic issues such as selection of partial charge set and dielectric behaviour, both of which remain subjective. Fourth, we have used a new VDW radius set which is probably the first one developed specially for binding affinity estimation. The VDW radius set from AMBER has served as a good start point in the development of our scoring function. But since it is parameterized against an endpoint different from ours, we expect that some optimisation on this set will help to improve the result. As our computational results have shown, the improvement is not marginal.

Compared to other similar approaches [26-29], we have obtained better statistical results. As shown in Table 6, our scoring function achieves the best regressional significance (F value) and the smallest standard deviation in cross-validation (s_{PRESS}). This may be the result of using a larger training set (as discussed in next section) as well as adopting new methods to calculate hydrogen bonding and hydrophobic effect during the protein-ligand binding process. While giving better results, our scoring function still maintains its conciseness. No structure minimisation or additional treatment is needed before calculating the binding affinity of the complex. This feature lends much convenience to SCORE's user.

Evolutionary test of an empirical model

All empirical methods have to use a training set and therefore they suffer from it: the content of the training set will influence the final model. The influence could be vital especially when the training set is not large and diverse enough. However, how can one know whether the training set is large enough for the model he is studying? And, how can one know whether his model will be stable if he could use a larger training set? These questions actually trouble every researcher who works on an empirical model and, unfortunately, we have not found any validation method reported yet to answer these questions.

As described in the *Methods* section, we have proposed a new procedure in an attempt to answer these questions. In such a procedure, the same model is tested by performing regressions on randomly constructed data sets. When the size of the data set increases continuously, conclusions can be drawn from the tendencies observed in the regressional results obtained along this procedure. This procedure simulates the process in which a researcher is able to use larger and larger data sets to train his model. That is the reason why we call this procedure "Evolutionary Test". The idea embedded in such a procedure is: we do not know what it will be; but we can make a reasonable prediction by checking what it has been. An ideal empirical model will have such features in an Evolutionary Test: first, its predictive ability should increase with the increase in the size of the data set; second, the coefficients in the model should converge to certain values instead of fluctuating randomly all the time. If a point of convergence is observed after which the predictive ability of the model under investigation does not improve significantly any more, it probably means that the data set at that point is already large enough for the model under investigation and thus the attempt of using a larger training set is not necessary.

Our scoring function behaves in the Evolutionary Test just as we have expected. As one can see in Figure 4a and Figure 4b, the regressional fitting of our scoring function decreases with the increase in the size of the data set. This is not surprising considering that the function is the same while the diversity of the data set is increasing. However, the predictive ability of the function, which is tested by cross-validation and the test set, keeps increasing with the increase in the size of data set. It indicates that the scoring function does get better trained by using larger data sets and therefore the use of the current training set in our study is absolutely necessary. In addition, the coefficients of each term in our scoring

function are generally coherent and tend to converge to certain values along this procedure (see Table 5). Based on these results, we can come to the conclusion that our scoring function is a robust, self-consistent empirical model. But we just cannot neglect that, although the size of data set has less significant influence after it exceeds 100-120, a convincible convergence has not achieved yet. According to the tendency observed so far, a training set consisting of approximately 200-250 protein-ligand complexes will be ideal for our scoring function.

Atomic binding score

Predicting the absolute binding affinity of a protein-ligand complex at a reasonable level is certainly very useful, but it may not be enough. In a drug design procedure, medicinal chemists also care for how to optimize a known lead compound rather than merely know whether it is good or not. The optimisation of the lead compound is usually intended to enhance the favourable interaction with the receptor and diminish the unfavourable. This needs to identify the relationships between the binding affinity and the chemical structure of the ligand and preferably evaluate them in a quantitative way.

Maybe the most attractive feature of our method is the ability of decomposing the binding affinity of the ligand to its target protein into the contribution of its component atoms. As having been described in the *Methods* section, we accomplish this by basing our algorithm entirely on atom-addition. By checking the atomic binding score, one can get a clear idea of whether a ligand atom is favourable to the binding process and how much it affects (see Figure 5). Furthermore, one can figure out why it is so by checking each energy term in the atomic binding score. This feature allows the direct relationship between the binding affinity and the structure of the ligand and is especially valuable for lead optimisation in structure-based drug design. By using the program SCORE, we can propose the following multi-step procedure for rational optimisation of lead compounds.

(1) Get the complex structure of the known lead compound, either by experimental determination or molecular modeling.

(2) Subject the complex structure to SCORE to compute the binding affinity.

(3) Analyze SCORE's output and figure out the favourable and the unfavourable parts of the lead compound.

(4) Design derivatives by enhancing the favourable parts, changing the unfavourable parts, or adding new functional groups to gain additional interactions.

(5) Model the complex structures of the derivatives by molecular docking and energy minimisation.

(6) Repeat step (2)(3)(4)(5) until the predicted binding affinities of the designed compounds reach a satisfactory level.

(7) Organic synthesis and bioassay.

(8) With the feedback from experiments, go to step (1) to start a new round of lead optimisation.

In fact, step (2) to step (6) form a cycle of virtual lead optimisation on the computer. This will help to save labour and cost in the following experiment step and thus improve the overall efficiency of the drug discovery process. A computer program that can automatically perform the virtual lead optimisation is currently under development in our lab.

Conclusions

In this paper, we have presented the development of an empirical scoring function for use in structure-based drug design schemes with emphasis on robustness over structural diversity, accuracy in absolute binding affinity estimation, and speed of computation. Terms used in the scoring function are devised to capture the essential energetics of the protein-ligand binding process. The final model is derived by regression of a large training set and yields promising results for both the training set and the test set. The robustness and internal consistency of the scoring function are demonstrated by a new validation procedure called Evolutionary Test. Compared with other similar approaches, our method has improved the quality of binding affinity prediction. Furthermore, with the introduction of atomic binding score, our method provides a practical tool for rational optimisation of lead compounds in a drug discovery process.

Supplementary material available statement The source code of the program, SCORE, is available by contacting the authors.

Acknowledgments This work is financially supported by the Science and Technology Ministry of China. We also thank Dr. Daniel A. Gschwend for his generous offering of data.

References

1. Kuntz, I.D. *Science* **1992**, *257*, 1078-1082.
2. Kuntz, I.D.; Meng, E.C.; Schoichet B.K. *Acc. Chem. Res.* **1994**, *27*, 117-123.
3. Verlinde, C.L.; Hol, W.G. *Structure* **1994**, *2*, 577-587.
4. Meng, E.C.; Shoichet, B.K.; Kuntz, I.D. *J. Comp. Chem.* **1992**, *13*, 505-524.
5. Lauri, G.; Bartlett, P.A. *J. Comput.-Aided Mol. Des*. **1994**, *8*, 51-66.
6. Miller, M.D.; Kearsley, S.K.; Underwood, D.J. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 153-174.
7. Lawrence, M.C.; Davis, P.C. *Proteins.* **1992**, *12*, 31-41.
8. Nishibata, Y.; Itai, A. *Tetrahedron* **1991**, *47*, 8985-8990.
9. Böhm, H.J. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61-78.
10. Gillet, V.J.; Johnson, A.P.; Mata, P. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 127-153.
11. DeWitte, R.S.; Shakhnovich, E.I. *J. Am. Chem. Soc*. **1996**, *118*, 11733-11744.
12. Luo, Z.; Wang, R.; Lai, L. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1187-1194.

13. Ajay; Murcko, M.A. *J. Med. Chem.* **1995**, *38*, 4953-4967.
14. Kollman, P. *Chem. Rev.* **1993**, *93*, 2395-2417.
15. Goodford, P.J. *J. Med. Chem.* **1985**, *28*, 849-857.
16. Luty, B.A.; Wasserman, Z.R.; Stouten, P.F.W. *J. Comp. Chem.* **1995**, *16*, 454-464.
17. Sansom, C.E.; Wu, J.; Weber, I.T. *Protein Eng.* **1992**, *5*, 659-667.
18. Holloway, M.K.; Wei, J.M. *J. Med. Chem.* **1995**, *38*, 305-317.
19. Ortiz, A.R.; Pisabarro, M.T.; Gago, F. *J. Med. Chem.* **1995**, *38*, 2681-2691.
20. Grootenhuis, P.D.J.; Galen, P.J.M.V. *Acta. Cryst.* **1995**, *D51*, 560-566.
21. Horton, N.; Lewis, M.L. *Protein Sci.* **1992**, *1*, 169-181.
22. Bohacek, R.S.; McMartin C. *J.Med.Chem.* **1992**, *35*, 1671-1684.
23. Krystek, S.; Stouch, T.; Navotny, J. *J.Mol.Biol.* **1993**, *234*, 661-679.
24. Williams, D.H.; Searle, M.S.; Mackay, J.P. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 1172-1178.
25. Weng, Z.; Vajda, S.; Delisi, C. *Protein Sci.* **1996**, *5*, 614-626.
26. Böhm, H.J. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243-256.
27. Head, R.D.; Smythe, M.L.; Oprea, T.I. *J. Am. Chem. Soc.* **1996**, *118*, 3959-3969.
28. Gschwend, D.A.; Good, A.C.; Kuntz, I.D. *J. Mol. Recog.* **1996**, *9*, 175-186.
29. Eldridge, M.D.; Murray, C.W.; Auton, T.R.; Paolini, G.V.; Mee, R.P. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425-445.
30. Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.B. *J. Mol. Biol.* **1977**, *112*, 535-542.
31. SYBYL 6.3, Tripos Associates, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144, U.S.A, 1996.
32. Wilson, A., Ed. *International Tables for Crystallography*, Vol.C; KLUWER: London, 1992.
33. Wang, R.; Fu, Y.; Lai, L. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615-621.
34. Nicklaus, M.C.; Wang, S.; Driscoll, J.S. *Bioorgan. Med. Chem.* **1995**, *3*, 411-428.
35. Sternberg, M.J.E.; Chickos, J.S. *Protein Eng.* **1994**, *7*, 149-155,
36. Cornell, W.D.; Cieplak, P.; Bayly, C.I. *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.
37. Davis, L. *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York, NY, U.S.A., 1991.