Journal of Computer-Aided Molecular Design **18:** 261–270, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

PLASS: Protein-ligand affinity statistical score – a knowledge-based force-field model of interaction derived from the PDB

V.D. Ozrin, M.V. Subbotin & S.M. Nikitin
Algodign LLC, Bolshaya Sadovaya 8, Moscow 123379, Russian Federation

Received 15 December 2003; accepted in revised form 7 June 2004

Key words: knowledge-based potentials, ligand-to-protein binding affinity, scoring function

Summary

We have developed PLASS (Protein-Ligand Affinity Statistical Score), a pair-wise potential of mean-force for rapid estimation of the binding affinity of a ligand molecule to a protein active site. This scoring function is derived from the frequency of occurrence of atom-type pairs in crystallographic complexes taken from the Protein Data Bank (PDB). Statistical distributions are converted into distance-dependent contributions to the Gibbs free interaction energy for 10 atomic types using the Boltzmann hypothesis, with only one adjustable parameter. For a representative set of 72 protein-ligand structures, PLASS scores correlate well with the experimentally measured dissociation constants: a correlation coefficient R of 0.82 and RMS error of 2.0 kcal/mol. Such high accuracy results from our novel treatment of the volume correction term, which takes into account the inhomogeneous properties of the protein-ligand complexes. PLASS is able to rank reliably the affinity of complexes which have as much diversity as in the PDB.

Introduction

The key problem of computer-aided drug design is fast and sufficiently accurate evaluation of binding affinity of ligand-protein complexes based on information about the chemical and 3D structures of both molecules. Because the quantum and statistical mechanic treatment of the problem is currently impossible, several approximations have been proposed that are referred to as scoring functions or force field potentials [1]. Scoring functions have different prediction accuracies depending on the assumptions made. They consume computer time from milliseconds to several hours per complex. There are basically two types of fast scoring function widely used in structural drug design: empirical and knowledge-based scores. The first approach uses the set of experimentally measured inhibition and dissociation constants for complexes with known 3D structures as a training set to calculate the set of free parameters of the model by fitting the experimental values. The typical number of free parameters of such models varies from several to several tens. Usually empirical scorings provide free energy prediction accuracy of about 1.5-2.0 kcal/mol, calculated as the RMSD for typical test sets of several tens of complexes with known dissociation or inhibition constants. The main problem with empirical scoring functions is their inapplicability for the complexes which are not included in the training set, i.e. the transferability problem. The knowledge-based (KB) approach, on the other hand, is based on knowledge of the 3D structures of molecular complexes instead of dissociation constants. This method is also referred to as potential of mean force or statistical potential. Usually KB models have just 1 or 2 free parameters introduced to rescale scoring to metric units. With progress in the high-resolution X-ray crystallography of macromolecules, it became possible to build statistically verified force field models based on the counting of atom type spatial correlations. The first KB potential was suggested to predict protein folding by Sippl et al. [2]. But recently encouraging results on prediction of binding free energy of complexes have

*To whom correspondence should be addressed. E-mail: Max.Subbotin@Algodign.com; Vladimir.Ozrin@Algodign.com

been demonstrated by several KB force field models (FFMs) such as BLEEP [3], PMF [4], SMoG [5] and others. These models use almost the same basic assumptions but differ slightly in atom type classification system, correlation counting procedure, and probability interpretation. They demonstrate acceptable correlation with experimental data on complex dissociation constants and may be applied to docking algorithms [6, 7].

The purpose of the investigation described in this paper is to develop a KB FFM that ranks protein-ligand complexes well by predicted Gibbs free energy calculated from atomic coordinates. The paper is organized as follows: In the **Methods** section we describe the method of building the training set and the principles of atom classification. The physical model assumptions used and the counting procedure are presented in the **Calculation of potentials** subsection. Analysis of the efficiency of the constructed model for prediction of dissociation constants is included in the **Results and discussion** section. Final comments on the results obtained are presented in **Conclusions**.

Methods

Obtaining PDB ligands

At the first step we apply a parser of files of PDB format to extract ligands themselves and omit all undesired complexes. We identify the ligand description in the file analyzing all HETATM records, chain identifiers and the PDB header. Also, the atom connectivity of the ligand molecule and between ligand and protein are checked by the interatomic distance criteria. Those give us the ability to correctly extract peptide ligands and classify all ligands by the following criteria:

(a) Resolution of the PDB structure.
(b) Number of heavy atoms of the ligand.
(c) Presence of a covalent bond with the protein.
(d) Presence of ions in the ligand vicinity.
(e) Presence of metal atoms in the ligand molecule.

This allows the elimination of small trivial entities, covalently bonded ligands and molecules with obvious errors of structure.

Assignment of atom types

We use a classification system of 10 types for atoms C, O, N, S, P and halogens to describe both protein and ligand molecules. The definitions of atom types are based on element type, hybridization state and presence of an atom in a functional group, for instance, aromatic ring. The entire classification is presented in Table 1.

As mentioned in the literature, the number of atom types is a compromise of two contradictory requirements. The first one wants the number of types to be maximized to describe the diversity of interaction properties of molecules adequately. But on the other hand, all statistically derived relations have to be proved by a large enough number of counts. In particular, calculation of spatially dependent correlation functions for all possible kinds of atomic pairs requires at least hundreds of counts per distance bin of 0.1 Å to provide the free energy prediction accuracy of 1–2 kcal/mol. Considering the atom classification system as an external model parameter that could be easily changed, we started with the relatively rough classification above. As it turned out, such a simple classification is enough for adequate estimation of the binding energy.

The typesetting algorithm we used is based on the information about chemical structure of hetero groups provided by 'HET DICTIONARY' [8]. A ligand is considered as several hetero or peptide groups connected together to form an entire molecule. For each atom of a ligand molecule the type was assigned using the definition of Table 1 and information about its chemical environment.

Calculation of potentials

For modeling the force field interaction in protein-ligand complexes, we used approaches developed in the statistical theory of multi-component dense disordered media such as liquid. Generally speaking, the applicability of the liquid theory formalism to the intermolecular interaction does not seem evident. In particular, the assumptions of the homogeneity of the medium and its infinite size *a priori* could break down for two interacting molecules of finite size. In practice, however, the positive results of KB models [3–5] show the ability of such a model to describe the basic features of the intermolecular interaction well enough. We suppose that one of the important reasons why the published KB FFMs work satisfactorily are the special precautions introduced by their authors to diminish the influence of the fact that protein and ligand molecules have certain finite volumes separated from one another by some effective interface.

Table 1. Atom classification used in PLASS.

Atom type	Description
C3	sp^3-carbon 5-valence nitrogen S, P atoms halogens
C2	sp^2-carbon
CA	carbon of aromatic structure
OA	sp^2-oxygen
OH	oxygen of hydroxyl group
OP	sp^2-oxygen of carboxyl, phosphonic and sulphonic groups
OE	sp^3-oxygen bonded to two heavy neighbors
ND	nitrogen bonded to one or more hydrogen atoms
NA	nitrogen of aromatic structure
NI	nitrogen of indole group

The scheme of calculation of the basic model variables is close to that used by Muege and Martin [4], although it contains a number of important differences of volume correction calculation procedure described in detail in the Appendix. It is assumed that the database consists of a set of selected protein–ligand (PL) complexes enumerated by index p, $p = 1,2,\dots M$. For an arbitrary PL complex, pairs of atoms (i,j) of the type $t_i = a$ from protein and of the type $t_j = b$ from ligand are considered. All r-dependent functions are defined at the equidistant grid points $R_k = k\Delta$ with $\Delta \sim 0.1$ Å ranging from 0 to $\mathrm{R}_{\max} = 12$ Å. The basic function for the considered approach, $\Delta N_{ab}(r)$, is the total number of ab type pairs with inter-atomic distance $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$ in the interval $s_r = (r - \Delta, r)$, calculated by summation of the pair occurrences over the database.

It is generally accepted that the ratio $\Delta f_{ab} = \Delta N_{ab}(r)/N$ estimates the probability of appearance of the ab-type contact in an arbitrary PL complex at distances within the interval $s_r = (r - \Delta, r)$. Here N is the total number of arbitrary contacts within the sphere $R_{\max}$. Then, following the approximation of a many-component liquid as a disordered infinite medium, we introduce pair correlation functions $g_{ab}(r)$:

$$\Delta N_{ab}(r) \equiv p_a l_b g_{ab}(r) \frac{N}{V} \Delta V \qquad (1)$$

where $V = \frac{4}{3}\pi R^3_{\max}$ is the volume of the interacting sphere and $\Delta V(r) = \frac{4}{3}\pi \left(r_i^3 - r_{i-1}^3\right)$, and p_a and l_b are the probabilities of an arbitrary contact of a protein a-type atom and a ligand b-type atom within the $R_{\max}$ sphere respectively, defined in terms of $\Delta N_{ab}(r)$.

On the other hand, following Sippl [2], we believe that the spatial distribution of the atom type pair is described by the Boltzmann law, $g_{ab}(r) \sim \exp(-\Delta G_{ab}(r)/RT)$, where $\Delta G_{ab}(r)$ is the contribution of the ab atom pair interaction to the total change ΔG of the Gibbs free energy due to a complex formation at some effective temperature T set equal to 298 K. This contribution is defined by

$$\Delta G_{ab}(r) = -RT \ln\left(f_b(r)\, g_{ab}(r)\right) \qquad (2)$$

where $f_a(r)$ is a set of the ligand atom type dependent functions introduced to eliminate the influence of the ligand molecule volume on the score for the molecular complexes to be described. The necessity of such a so-called *volume correction term* originates from the fact that a ligand has a certain volume which has to be excluded when one calculates probabilities for $g_{ab}(r)$ using Equations 1 and 2. The detailed description of the calculation procedure used for definition of $g_{ab}(r)$ and $f_a(r)$ is presented in the Appendix.

The scoring function estimating the binding free energy of the complex formation is defined as the sum over all interatomic interactions,

$$PLASS = \sum_{\substack{i,\, j;\ r_{ij} < r_{cut-off} \\ t_i = a,\ t_j = b}} G_{ab}(r_{ij}) \qquad (3)$$

where the summation is over all atomic pairs (i,j) of the complex with the interatom distance within the cut-off radius, $r_{cut\text{-}off}$, t_i and t_j designate types of atoms 'i' and 'j', respectively. The calculated PLASS score is related to the binding free energy by

$$\Delta G = \frac{1}{\varepsilon} PLASS \qquad (4)$$

where ε is the only free parameter of the model used to scale PLASS values to compare them with experimental data on dissociation constants.

Results and discussion

We have used our program to filter the entire PDB of 19500 entries (November 2002) and extract complexes satisfying the criteria described above. Thus, we obtained a database of about 2000 files with 2500 ligands. Then the program for typesetting processed the selected set of files for 10 min. We registered about 10^8 interacting atom pairs distributed over 12 Å of interatomic distance. The rarest atom type pair distributions contain about 100 counts per spatial bin of 0.1 Å at the distances of most interest. At the next stage the values defined by Equations 1, 2 and the volume correction function $f_a(r)$ were calculated.

A typical example of the correlation functions we have obtained, for the pair of ND-OP atoms, is shown in Figure 1.

As was expected, the correlation function looks similar to one measured for liquids. The maxima about 2.8 Å distance can be interpreted as the favorable relative position of the donor and acceptor of hydrogen bonds. The several additional extrema correspond to the second coordination radii, or so-called multiparticle correlation. To illustrate this particular phenomenon, one can consider the common relative positions of carboxyl and amidino groups which usually form two H-bonds between two O-N pairs as shown in Figure 1. The N-H-O distances are about 3 Å. But also the two N-O pairs are registered by the counting algorithm of $g_{NO}(r)$ at the distance of 3.5 Å – 'cross-correlations distance'. That can lead to the appearance of the additional peak of dependence of Figure 1 at the 3.5 Å point. Analogously, additional extrema appear in the correlation functions of other atom type pairs. This circumstance makes the direct implementation of KB FFM for docking of a ligand molecule to a protein active site a non-trivial task.

Moreover, there are some obvious features of the correlation function behavior related to the specificity of the system with respect to infinite size homogeneous liquid. First of all, the curve evidently slopes down on average at distances larger than 6 Å and generally becomes lower than 1 at distances larger than 9 Å. That is clear evidence of the limited size of a protein molecule, i.e., the fact that averaged over the database, protein atom density slowly decreases to zero with distance from any fixed point. This circumstance forced us to introduce the corrections for the correlation functions at large distances discussed in detail in the Appendix.

It is worth emphasizing that we used 12 Å spheres to collect all necessary statistical data on interatomic contacts. Any KB model needs to take into account the large influence of volume to calculate the average density of protein atoms of a certain type at non-interaction range from any ligand atom correctly, which is to estimate the so-called 'reference state' corresponding to a fully disordered uniform atom type mixture. Also, there are some unclear arguments in the literature that a long potential cutoff allows to estimate the entropic contribution to free energy more accurately.

Another problem we have faced is the specific influence of so-called (in the literature) 'excluded ligand volume correction'. The issue is that in the theory of liquids the homogeneity of the media is assumed, i.e. liquid consists of particles (molecules) of several types and correlations between them in space are the result of their specific interaction. In the case of protein-ligand complexes, we consider the interaction of atoms of a predefined number of types, but actually all atoms in a complex are separated into two more general classes: protein atom and ligand atom. Atoms of both classes are bonded into the two separated molecules. We are trying to describe the interaction of the molecules at an atom-atom level. Both of the molecules occupy their separated volumes and their interactions mostly take place at the interface. This circumstance forced the authors of Ref. 4 to introduce a set of distance dependent corrections to potentials ascribing to ligand atom types. Based on a close idea, to take into account the excluded ligand volume, we have defined the volume correction factors as the radial distribution of relative protein atom density in the vicinity of an active site averaged over the database. We used these functions to compensate for the leakage of counts of a protein atom at a short distance from a ligand atom of certain type. As was expected, these distributions appear to be sufficiently non-uniform so the influence of the phenomenon on the score is large. We have chosen the piece-wise linear functions $f_a(r)$ to approximate the curves calculated over the database volume correction functions for simplicity. Each of the 10 functions is determined by 3 parameters and defined at the distance interval from 0 to 12 Å. Namely, these are amplitudes A_a for the value of $f_a(r)$ in the interval of (0, R_1) and two distances R_1 and

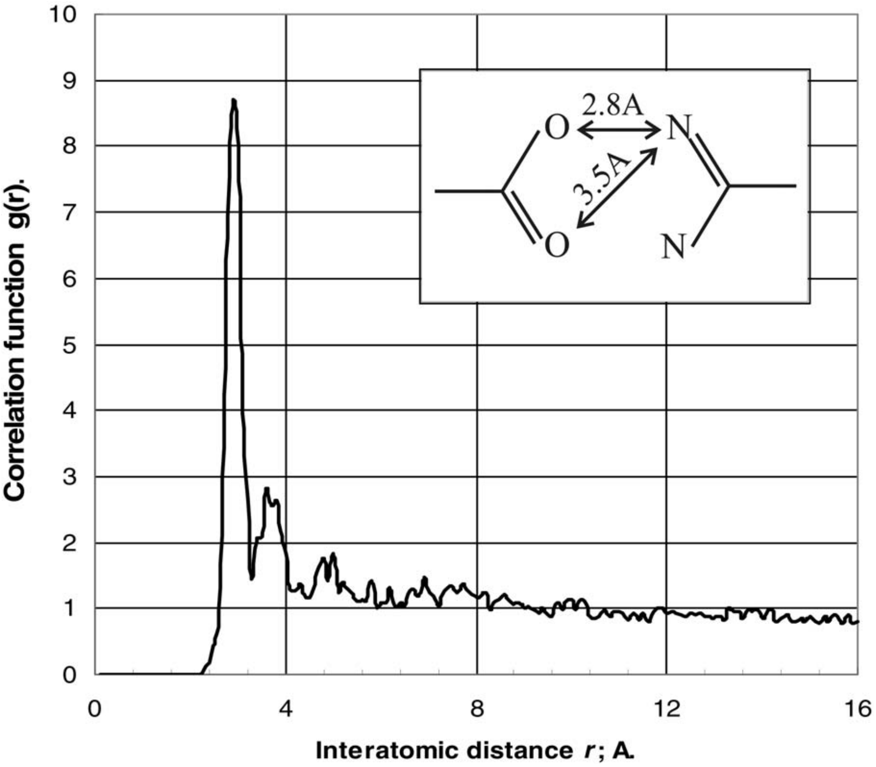

Figure 1. Two-particle correlation function for nitrogen as H-donor (ND type) and polar oxygen as H-acceptor (OP type) demonstrates the evident maxima of the probability curve at distances of about 2.8 Å and 3.5 Å. The first peak could be interpreted as the favorable relative position of N and O atoms for H-bond formation. The second one could be explained by artificial spatial correlation of atoms of carboxyl and amidine groups frequently appearing in protein-ligand complexes.

R_2 between which $f_a(r)$ goes to zero as shown in Figure 2. For all functions the parameters R_1 and R_2 take the same values, $R_1 = 4$ Å and $R_2 = 5$ Å, while their amplitudes are the averaged value of the calculated values over the interval (0, R_1). A more detailed description of the volume correction functions calculation procedure is presented in the Appendix.

To estimate the ability of PLASS to predict the dissociation constant of protein-ligand complexes, we used a procedure of comparison of the score with experimentally measured values of dissociation constants K_d. We have chosen a set of 72 K_d values for complexes presented in the PDB and published in the literature and satisfying the conditions of the database formation. The list of PDB codes of the test set complexes and dissociation constant values is presented in Table 2. Almost all of the K_d values were taken from original papers following references [5, 3, 9]. Note that all of the complexes in Table 2 are included in the test sets of SMOG [5] and/or BLEEP [3].

Examining the original papers about K_d measurements, we have estimated the average reliability of the experimental data to be ±1.5 kcal/mol. Although the instrumental error of experimental techniques used for ΔG measurements is declared to be about 0.1–0.25 kcal/mol, i.e. 10–20% [1], we believe the overall reliability is likely about 1–1.5 kcal/mol. In our opinion, this is due to the absence of knowledge of what particular way of association reaction takes place, keeping in mind a lot of additional chemical and physical factors of great importance. It is not uncommon that published K_d values for the same protein-ligand complex differ by an order of magnitude, and therefore, the difference between the corresponding ΔG values can be several kcal/mol.

The plot of comparison of PLASS prediction for dissociation constants versus experimental values is shown in Figure 4. One can see a sufficiently accurate correlation of the data.

However, a number of outliers with large prediction error are observed. In particular, 6 endothiapepsin complexes are clearly combined in a group with large positive deviation from the prediction line. This fact can be easily explained when we recall that K_d values

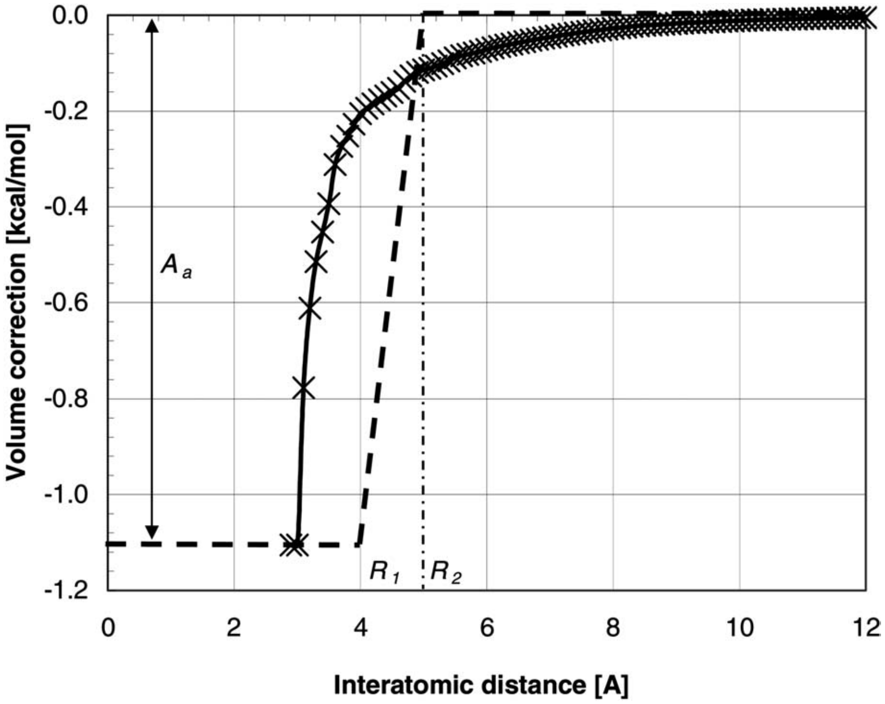

Figure 2. Example of volume correction function calculated over the model database. The correction is for each of the 10 ligand atom types to eliminate the influence of the inhomogeneous atom type density distribution near the protein-ligand interface. The plot presents the volume correction for CA aromatic carbon ligand atom type in energy units, i.e. – T·ln[$f_a(r)$]. Also, the piece-wise linear approximation used to replace $f_a(r)$ in the PLASS dependences is presented (shown by dashed line).

Table 2. PDB codes of 72 protein-ligand complexes used as the PLASS test set.

1abe	1dbb	1ets	1hvk	1ppl	1tnj	2er9	4cla	5tim
1abf	1dbj	1ett	1hvl	1ppm	1tnk	2gbp	4er1	6abp
1apb	1dbk	1fkb	1l83	1pro	1tnl	2ifb	4hmg	6rnt
1apt	1dbm	1fkf	1nnb	1rbp	2ak3	2r04	4phv	6tim
1apu	1eed	1fkg	1nsd	1rnt	2dbl	2rnt	4sga	7abp
1apv	1ela	1fkl	1ppc	1tng	2dri	3cla	5abp	8abp
1apv	1ela	1fkl	1ppc	1tng	2dri	3cla	5abp	8abp
1bap	1etr	1hvj	1ppk	1tni	2er7	3ptb	5sga	9hvp

for complexes of this protein family are usually measured at a pH of about 3. Because most complexes of the database were investigated at pH values of about 7, the endothiapepsin ones do not have typical binding parameters. So, PLASS, like any other knowledge-based model, does not work well for items having a parameter poorly presented in the database. For this reason, we have excluded endothiapepsin complexes 1eed, 2er6, 2er9, 3er3, 5er2 and 4er1 from the test set. Under that assumption, we obtain the correlation coefficient R = 0.82 and RMSD = 2.0 kcal/mol between the scored and experimental values. To our knowledge, this is the best result published for a KB FFM having only one free parameter.

Indeed, the point of great importance is how PLASS predicts dissociation constants for complexes of different protein families having different active sites size, lipophility, flexibility, etc. However, the available amount of experimental data and their quality for each particular protein family seems to be not enough for reliable model performance estimation.

At this point, the question of the influence of systematic and random errors of experimental K$_d$ measurement on the evaluation of the predictive reliability

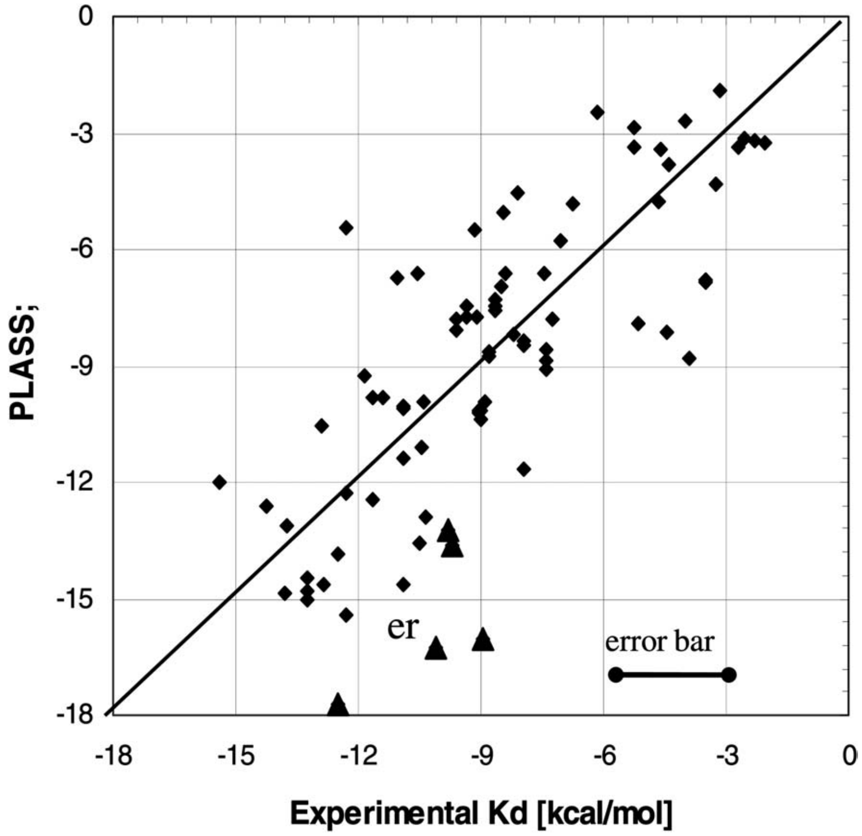

Figure 3. The correlation of the PLASS predictions with experimental data on dissociation constants K_d of test set complexes (see Table 2) is presented. The averaged experimental uncertainty of the K_d measurements is estimated to be 1.5 kcal/mol (half of the error bar shown in the plot). About 90% of the data points lay inside of the error bar interval. PLASS evidently groups the endothiapepsin complexes (data points presented by triangles) and strongly overestimates their affinity usually measured at pH $\approx$ 3–3.5 that is very far from the average of 7 for the database. Like any statistical knowledge-based model, PLASS does not estimate correctly the complexes poorly presented in the training set i.e. in the database. Thus, the selectivity of PLASS could be declared.

of any FFM should be discussed. The correlation coefficient R and RMSD are the model quality criteria. It is clear that there are certain lower limits for both R and RMSD for a fixed set of experimental data points. These limits are due to the uncertainty of the experimental data to be matched by the model. To reveal how experimental data errors affect the uncertainty of R, we have calculated the histogram for R values supposing all dissociation constant values are distributed normally with an experimental error of ±1.5 kcal/mol to simulate the stochastic part of experimental errors. The distribution histogram for R is shown in Figure 4.

One can see that PLASS gives values for a representative test set of complexes which correlates with empirical K_d with a most probable value of the correlation coefficient of 0.8 and an uncertainty of ±0.07. The uncertainty is the consequence only of the quality of the experimental data but not the FFM being tested. In other words, there is no substantial difference between prediction quality of two models when their correlation coefficients with measured values differ by less than about 0.1. Similarly, there is no way to demonstrate further improvement of a KB FFM if its RMSD relative to experimental values is comparable with the average errors of those values.

Also, the quality of the X-ray crystallographic structures of complexes influences the prediction error of any model developed using the atomic coordinates. Because the average uncertainty of atomic coordinates in the database is estimated to be about 0.2 Å, and the averaged derivative of $G_{ab}(r)$ is of the order of 5 kcal/mol/Å at the region of interest, the uncertainty of PLASS predictions is at least 0.2 Å × 5 kcal/mol/Å = 1 kcal/mol. This value gives the approximate theoretically possible level for accuracy of any KB scoring derived from the PDB.

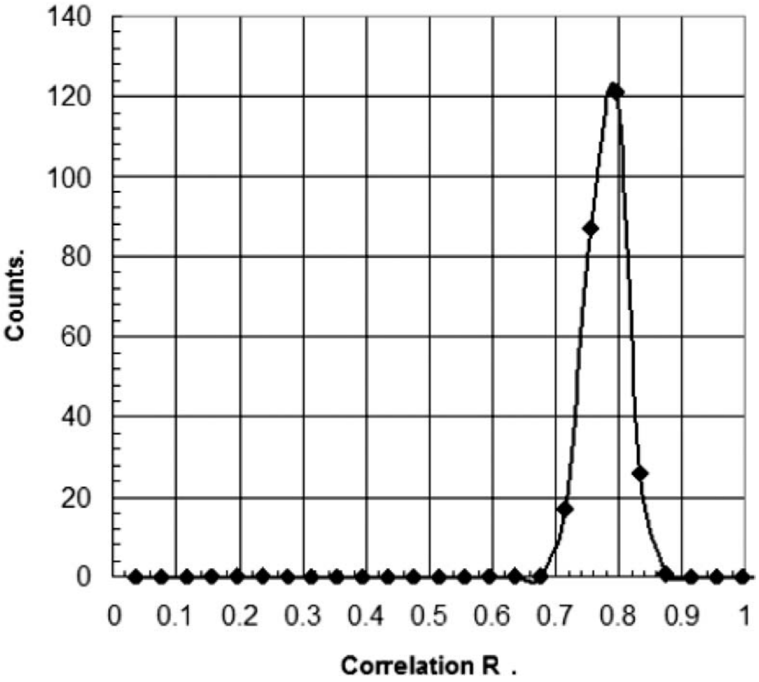

Figure 4. The distribution of the correlation coefficient R of the PLASS prediction with the test set of 72 experimentally measured dissociation constants K_d. To simulate the stochastic part of the experimental error of K_d measurement, several hundreds of values for each test set were taken with a normal distribution. The means of distributions equals the experimental K_d values, and the dispersion corresponds to the common experimental accuracy of K_d measurements of ±1.5 kcal/mol as estimated by experimentalists. The histogram has been calculated supposing R is a function of 72 independent stochastic variables of K_d while the PLASS predicted values are constant.

Conclusions

This paper has described the development of a statistically derived scoring function, PLASS, designed to estimate the free energy of binding for a protein-ligand complex when the 3D structure of the complex is known. The predictions of the dissociation constants for the test set of 72 complexes correlate with experimental data with the value of 0.82 ± 0.07. The uncertainty of the correlation coefficient is caused by the errors of experimental measurements of dissociation constant estimated to be ±1.5 kcal/mol on average. The RMSD of the predicted relative to experimental values is 2.0 kcal/mol. We believe the high correlation of PLASS with experimental data is due to properly taking into account bulk features of protein-ligand complexes by the novel procedure of ligand volume counting and usage of spatially dependent protein atom density. The model is quite adequate for description of basic features of ligand binding at a protein active site, and PLASS can be used to reliably rank ligands by their binding free energies.

Acknowledgements

The authors thank N. Zaitzeva, V. Tarasov, I. Bodrenko and E. Tkalya for fruitful discussions and C. Queen for careful review of the manuscript.

Appendix

For statistical calculation of the pair correlation function $g_{ab}(\boldsymbol{r})$ for a protein–ligand atom pair ab we introduce the function

$$\Delta N_{ij}^{(p)}(r) = \begin{cases} 1 & \text{for} \quad \mathrm{r} < |\mathbf{r}_i - \mathbf{r}_j| < r + \Delta \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where the subscripts i and j enumerate protein and ligand atoms of the types t_i and t_j of the complex p, respectively. The summation of $\Delta N_{ij}^{(p)}(\mathrm{r})$ over all protein atoms of type a and ligand atoms of type b in the complex, and then over all complexes in the database yields

$$\begin{aligned} \Delta N_{ab}(r) &= \sum_p \Delta N_{ab}^{(p)}(r)\,, \\ \Delta N_{ab}^{(p)}(r) &= \sum_{i,\, t_i=a}\ \sum_{j,\, t_j=b} \Delta N_{ij}^{(p)}(r) \end{aligned} \tag{6}$$

which is the total number of *ab* pairs in the spherical shell, $s_r = (r, r + \Delta)$.

The summation of $\Delta N_{ab}(r)$ over all r in the range of (0, R_{max}=12 Å) gives the number of interacting atom pair of types a and b:

$$N_{ab} = \sum_{r} \Delta N_{ab}^{(p)}(r) \tag{7}$$

The total number of contacts of protein atoms of type a and ligand atoms of type b and the total number of pair contacts within the sphere R_{max} are defined, respectively, as

$$P_a = \sum_{b} N_{ab}, \qquad L_b = \sum_{a} N_{ab},$$

$$N = \sum_{a} P_a = \sum_{b} L_b. \tag{8}$$

It is generally accepted that the ratio

$$\Delta f_{ab}(r) = \frac{\Delta N_{ab}(r)}{N} \tag{9}$$

is a good approximation for probability of appearance of the *ab*-type contact in an arbitrary PL complex at distances within the interval $s_r = (r - \Delta, r)$. The values

$$p_a = \frac{P_a}{N}, \qquad l_b = \frac{L_b}{N} \tag{10}$$

estimate probability of contact of protein a-type atom, and probability of contact of ligand b-type atom in the database within the $R_{\max}$ sphere, respectively. Following the approach of the statistical theory of disordered infinite multi-component medium we can define the pair correlation functions $g_{ab}(r)$ by

$$\Delta N_{ab}(r) \equiv p_a l_b g_{ab}(r) \frac{N}{V} \Delta V \tag{11}$$

where $V = \frac{4}{3}\pi R_{\max}^3$ is the volume of the interacting sphere and $\Delta V(r) = \frac{4}{3}\pi \left(r_i^3 - r_{i-1}^3\right)$.

As discussed above, there are two significant differences between the standard subject of statistical physics and the system in hand which force us to introduce two types of corrections for the definitions 8–11. First, this is the limited size of protein–ligand complexes and inhomogeneity of the protein atoms space distribution. To diminish the influence of the last long-range effect to atom type pair correlations we apply the radius-dependent normalization factor defined as

$$N_{ab}(r) = \begin{cases} \sum\limits_{r' \leq 6A} \Delta N_{ab}^{(p)}(r') & \text{for } r < 6\,\text{Å} \\ \sum\limits_{r' \leq r} \Delta N_{ab}^{(p)}(r') & \text{for } r > 6\,\text{Å} \end{cases} \tag{12}$$

The functions $N_{ab}(r)$ are used instead of constants N_{ab} in definitions of the factors P_a, L_b, N, the probabilities p_a, l_b, and the correlation function $g_{ab}(r)$ in Equations 8–11.

The other important feature of the considered system is that atoms are separated into two general classes: protein atom and ligand atom, which are bonded into the two separated molecules occupying their separated volume and interacting mostly by the interface. The volume correction factors are used to compensate for the leakage of counts number of protein atoms at a short distance from ligand atoms. We introduce the average density of protein atoms in the vicinity of an arbitrary ligand atom of type b calculated by summation over the database

$$n_b^P(r) = \frac{1}{M_b \Delta V} \sum_{p} \frac{1}{l_b^{(p)}} \sum_{a} \Delta N_{ab}^{(p)}(r) \tag{13}$$

where $l_b^{(p)}$ is the number of ligand atoms of type b in the complex p, and M_b is the number of complexes including such ligand atoms in the database. Note that for calculations of the average densities (13), two values for radius bin Δr were used: $\Delta r \sim 2$ Å for $r < 4$ Å, and $\Delta r \sim 0.5$ Å for $r > 4$ Å. The volume correction factor is defined as the ratio

$$f_b(r) = \frac{n_b^{Tot}(r)}{n_b^P(r)} \tag{14}$$

where $n_b^{\mathrm{Tot}}(r)$ is the density of both protein and ligand atoms in the vicinity of an arbitrary ligand atom of type b defined similar to Equation 13. Thus, the volume element ΔV in definition (11) of the correlation function is replaced by the ratio $\Delta V/f_b(r)$. The functions $f_b(r)$ are significantly greater than 1 for short distances, r < 4–5 Å, and rapidly tends to 1 at long distances. In fact, the difference $(f_b(r) - 1)$ is negligibly small for r > 6–7 Å for all types b, while the maximum $f_b(r)$ is not greater than 3.

Note that the approach discussed above differs significantly from that developed by Muegge and Martin [4]. According to [4], the volume correction factors (see Equations 8–14 of [4]) can be interpreted as the inverse ratios of relative deviation of local densities of protein and ligand atoms and the deviation of these for average (bulk) densities calculated for total system volume. In contrast with the above, Equation 11 takes into account only the local decrease in the volume allowed for protein atoms and provides a more moderate dependence of volume correction amplitude on ligand atom type. We suppose that this improvement, along with the distance dependent normalization factors introduced in Equation 12, leads to a more correct estimation of the probability of a certain atom-type pair occurrence and increases the accuracy of the model.

References

1. Murcko, A. and Murcko, M.A., J. Med. Chem., 38 (1995) 4953.
2. Sippl, M.J., Ortner, M., Jaritz, M., Lckner, P. and Flockner, H., Folding Design, 1 (1996) 289.
3. Mitchell, J.B.O., Laskowski, R.A., Alex, A. and Thornton, J.M., J. Comput. Chem., 20 (1999) 1165.
4. Mitchell, J.B.O., Laskowski, R.A., Alex, A., Forster, M.J. and Thornton, J.M., J. Comput. Chem., 20 (1999) 1177.
5. Muegge, I. and Martin, Y.C., J. Med. Chem., 42 (1999) 791.
6. Ishchenko, A.V. and Shakhnovich, E.I., J. Med. Chem., 45 (2002) 2770.
7. Gohlke, H., Hendlich, M. and Klebe, G., J. Mol. Biol., 295 (2000) 337.
8. Muegge, I., Martin, Y.C., Hajduk P.J. and Fesik, S.W., J. Med. Chem., 42 (1999) 2498.
9. Het Group Dictionary. Available from Brookhaven National Laboratory WWW address ftp://ftp.ebi.ac.uk/pub/databases/pdb/pub/resources/hetgroups/het_dictionary.txt
10. Wang, R., Liu, L., Lai, L. and Tang, Y., J. Mol. Model., 4 (1998) 379.