Analysis of knowledge-based protein–ligand potentials using a self-consistent method

J. SHIMADA, A.V. ISHCENKO, AND E.I. SHAKHNOVICH

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138

(RECEIVED August 16, 1999; FINAL REVISION January 26, 2000; ACCEPTED January 27, 2000)

Abstract

We propose a self-consistent approach to analyze knowledge-based atom–atom potentials used to calculate protein–ligand binding energies. Ligands complexed to actual protein structures were first built using the SMOG growth procedure (DeWitte & Shakhnovich, 1996) with a chosen input potential. These model protein–ligand complexes were used to construct databases from which knowledge-based protein–ligand potentials were derived. We then tested several different modifications to such potentials and evaluated their performance on their ability to reconstruct the input potential using the statistical information available from a database composed of model complexes. Our data indicate that the most significant improvement resulted from properly accounting for the following key issues when estimating the reference state: (1) the presence of significant nonenergetic effects that influence the contact frequencies and (2) the presence of correlations in contact patterns due to chemical structure. The most successful procedure was applied to derive an atom–atom potential for real protein–ligand complexes. Despite the simplicity of the model (pairwise contact potential with a single interaction distance), the derived binding free energies showed a statistically significant correlation (~ 0.65) with experimental binding scores for a diverse set of complexes.

Keywords: computational ligand design; knowledge-based potentials; protein–ligand interactions; self-consistent method

Accurate prediction of protein–ligand binding free energies remains one of the most important unsolved problems of computational biophysics. For years, it has been an elusive goal with only limited success reported in the literature (Ajay & Murcko, 1995; Martin, 1997). Even with structural data where the problem of predicting binding conformations is eliminated, many attempts have failed to consistently correlate predicted binding free energies with experimental data for a diverse set of protein–ligand complexes. The difficulty lies precisely in that no single physical interaction or entropic factor plays a dominant role in the complex thermodynamics of binding.

Many of the computational approaches used for the prediction of binding strength are based on force-field calculations, such as free energy perturbation methods (Kollman, 1993). They employ all-atom energy functions derived from first principles and parameterized for structure refinement or molecular dynamic calculations. Although some of these methods have led to successful prediction (Rao et al., 1992), the main problem associated with such approaches has been their computational inefficiency. A method

that is moderately successful, but computationally costly, has limited applicability for design applications where rapid screening techniques are often necessary to identify lead compounds.

To overcome the computational inefficiencies of force-field methods, fast empirical scoring functions were developed (Böhm, 1994, 1998; Head et al., 1996; Eldridge et al., 1997). Typically, the binding free energy is arbitrarily decomposed into various enthalpic and entropic features that are described by simple functional forms incorporating free parameters. The parameters are obtained by optimization procedures maximizing the correlation between the computed and experimental binding free energies of a set of complexes with known structures and binding constants. Although relatively high correlations were obtained for diverse training sets, it is unclear how accurately these approaches predict the affinities of ligands not in these sets.

An alternative approach to develop fast scoring functions is based on the idea of using knowledge-based methods employing coarse-grained models (Chaikin & Lubensky, 1995) of protein–ligand interactions. Knowledge-based potentials have a key advantage over empirical scoring functions. There are only a small number of complexes for which both structure and binding constant are publicly available. Since many of these complexes feature similar ligands, it is unclear whether there is sufficient diversity to adequately represent all features of the binding process when param-

Reprint requests to: E.I. Shakhnovich, Harvard University, Department of Chemistry, 12 Oxford Street, Cambridge, Massachusetts 02138; e-mail: eugene@belok.harvard.edu.

etrizing an empirical scoring function. In contrast, only structural information is necessary for deriving knowledge-based parameters, meaning that a bigger and more diverse set of complexes can be utilized. For this reason, knowledge-based potentials are expected to be more transferable than empirical scoring functions to complexes not found in the training database.

Knowledge-based methods have been widely used for protein structure prediction with a certain degree of success (reviewed in Jernigan & Bahar, 1996). In most knowledge-based potentials for protein folding and threading, each amino acid is represented by one or two interacting centers. The existing approaches to derive such potentials can be grouped roughly into two categories. The first one is a statistical approach, based on the “quasi-chemical approximation” (Miyazawa & Jernigan, 1985; Sippl, 1990; DeWitte & Shakhnovich, 1994; Hinds & Levitt, 1994; Finkelstein et al., 1995) that attempts to interpret frequencies of observed contacts between different amino acids in terms of residue–residue interaction energies. The second approach is based on the optimization of parameters to achieve the best discrimination of the native structure from a set of decoys (Goldstein et al., 1992a, 1992b; Mirny & Shakhnovich, 1996). A systematic comparison of different approaches for deriving a protein folding potential is given in Mirny and Shakhnovich (1996).

Because a small molecule’s binding free energy may change dramatically through the replacement of a few atoms, the potential function for ligand binding must be able to reflect atom-scale changes. It should therefore be taken in an all-atom representation. To achieve computational tractability for ligand design applications, a certain level of coarse-graining can be introduced. The first attempt to develop a coarse-grained knowledge-based atom–atom potential is due to Wallquist et al. (1995). Another early coarse-grained method was developed by DeWitte and Shakhnovich (1996) and featured in the ligand design application SMOG (Small Molecule Growth). This approach assumed that every protein–ligand complex can be decomposed into a set of essential structural features that each independently contribute to the binding free energy. Since the simplest and natural choice for an unit of interaction was the protein atom–ligand atom contact pair, the binding free energy (or “interaction score”) was then determined according to

$$G = \sum_p \sum_l B(\sigma_p, \sigma_l) \Delta(p, l) \quad (1)$$

where σ_p and σ_l are the atom types of protein atom p and ligand atom l , respectively, and $\Delta(p, l)$ is one if p and l are in contact and zero otherwise. Two atoms were classified to be of the same type if they had on average similar chemical properties. This method was used in several illustrative ligand design examples (DeWitte et al., 1997).

Despite the simplicity of this potential energy function (the “SMoG 96” potential), high correlations were achieved for three sets of protein–ligand complexes (purine nucleoside phosphorylase (PNP), HIV protease, and the SH3 domain) taken separately. However, the correlation was poor when the three sets of complexes were taken together. Qualitatively, it was noticed that the performance of the SMOG 96 potential was better for ligands with significant hydrophobic character than for ones stabilized by polar interactions. It is thus believed that the good correlations were explained by the fact that the SMOG 96 potential estimated ligand hydrophobicity within each set fairly well. Poor estimates of hydrogen bonding and electrostatic energies, in contrast, appear to

have prevented good correlation when the PNP and HIV sets were taken together.

Several distance-dependent knowledge-based functions were also developed (Verkhivker et al., 1995; Mitchell et al., 1999; Muegge & Martin, 1999). In a recent work (Muegge & Martin, 1999), significant correlation was obtained for a testing set of 77 complexes ($R^2 = 0.61$) and for some smaller protein family-based sets. This demonstrates that a knowledge-based potential function of significant predictive power exists. However, a more systematic study of knowledge-based methods is necessary to determine which features are most important for successful prediction. Such a study is presented in the current work.

The derivation of SMOG 96, as well as most other knowledge-based parameter sets, is based on a number of assumptions that are not rigorously justified or tested. For example, the assumptions behind the quasi-chemical approximation were criticized by several authors (Thomas & Dill, 1996; Betancourt & Thirumalai, 1999; Micheletti et al., 1999), although in some limiting cases this approximation was shown to be effective for deriving potentials (Finkelstein et al., 1995; Mirny & Shakhnovich, 1996) that yielded meaningful free energy predictions (Kocher et al., 1994). A systematic approach to analyze methods for deriving folding potentials was proposed by Mirny and Shakhnovich (1996). In this study, we extend their self-consistent approach to knowledge-based protein–ligand atom–atom potentials. The main idea of this approach is as follows. Using the SMOG growth algorithm, ligands are designed for a diverse set of proteins according to an arbitrarily chosen potential. Together with the proteins, these ligands comprise a database of model protein–ligand complexes. Within this model system, the binding strength of a ligand to a protein is given by the interaction score as computed by the chosen potential. A knowledge-based method is next used to derive parameters from the model database. The success of this method is then measured on how well the interaction scores computed with the derived parameters correlates with those computed with the chosen potential.

Theory

The Boltzmann hypothesis

In original knowledge-based approaches for model protein folding studies, formulations of the potentials were based on a heuristic connection between the frequency of observing a particular structural feature X in the database and its energy:

$$\text{frequency} \propto \exp(-\beta E(X)). \quad (2)$$

Naively, this may appear to be nothing more than a statement of the Boltzmann distribution, which describes the system (in this case, the database of structures) attaining canonical equilibrium. However, this cannot be the correct interpretation. Database structures have conformations frozen in their global free energy minima. A database of structures therefore cannot represent a proper statistical ensemble of many conformational states, which would be a prerequisite for establishing thermodynamic equilibrium. A more plausible explanation of this Boltzmann hypothesis is given by Finkelstein et al. (1995). They argued that in an ensemble of random sequences that stabilize a particular fold, a Boltzmann-like distribution arises because low-energy structural features exponentially increase the number of sequences that contain those features.

The database from which knowledge-based parameters are derived must therefore resemble an ensemble of random stable sequences.

A self-consistent method for testing knowledge-based potentials

If one accepts the Boltzmann hypothesis, the success of a knowledge-based potential is dependent on two factors. To begin, the physical model, which specifies the free energy function under coarse-graining, must be realistic. In addition, the statistical procedure that calculates the required frequencies from the structural data must be accurate. If the probabilities in Equation 2 are not correctly computed, then the parameters derived for even a good physical model will be problematic.

Every knowledge-based potential must ultimately be judged on the correlation between its predictions and experimental data. Although such a test can be used to assess a given potential's success or failure, it does not allow one to rigorously identify and analyze the weaknesses of the potential.

The testing method we are presenting enables one to analyze the statistical procedure used to derive a potential *independently* of the details of the physical model. This is accomplished by creating an fictional world of protein–ligand complexes, for which the physical model—namely, the functional form of the potential used to create the complexes—is completely known. We can then test a statistical procedure by seeing whether the correct parameterization can be derived from a database of these “toy” complexes given the functional form of the potential.

We can create a large number of toy complexes using SMOG's fragment-based Monte Carlo (Binder & Heerman, 1992) growth procedure provided that we have the crystal structure of the target. Since the potential is used only when the energy is calculated during the evaluation of the Metropolis criterion (Metropolis et al., 1953), any potential can be used with this procedure. The Metropolis criterion will ensure that interaction scores of the grown ligands will be canonically distributed according to the selected input potential. Because SMOG growth is fragment-based, the resulting fragment distribution should be canonical as well. If an infinitely diverse fragment library were to be used, then all possible combinations of fragments would be explored. From this random ensemble of protein–ligand complexes, only the most stable (i.e., complexes with low SMOG interaction scores) are chosen. Therefore, this situation is analogous to that described by Finkelstein et al. (1995) for protein folding, where stable sequences (stable fragment-based ligands) are selected from an ensemble of random sequences (random fragment-based ligands). The Boltzmann hypothesis should then be applicable to structural features of a resolution finer than fragments, such as pairwise atomic contacts. In other words, the existence of low energy contacts will exponentially increase the number of fragment-based ligands that allow such contacts, and the Boltzmann distribution should apply to pairwise contacts in any large ensemble generated by SMOG. A toy database can then be constructed from the generated protein–ligand complexes, with the scores given by the “true” input potential indicating the strength of the binding interaction. These true interaction scores should be understood as the experimental data of the toy world.

A knowledge-based method can now be evaluated on its ability to recover the true parameters from the toy database. More specifically, the knowledge-based parameters derived from a group of toy complexes—the training set—should yield interaction scores for another group of toy complexes—the testing set—that corre-

late well with their true interaction scores. In the case when the true and the derived potentials share the same functional form (i.e., an atom–atom pairwise potential using the same atom typing scheme), the physical assumptions of the model concerning the form of the free energy function are not explicitly tested. Rather, the focus is on the formulation by which the statistics observed in the database is related to the potential parameters via the Boltzmann hypothesis. Failure to pass this self-consistent test must indicate a fundamental procedural flaw in the knowledge-based method, which cannot be eliminated by improving the physical description being used.

The SMOG 96 potential

DeWitte and Shakhnovich (1996) adapted the Boltzmann hypothesis to protein–ligand complexes by proposing that a measure of the frequency of (σ_p, σ_l) contacts in a database of high-resolution crystal structures is given by

$$p(\sigma_p, \sigma_l) = \frac{N(\sigma_p, \sigma_l)}{\sqrt{N(\sigma_p)}\sqrt{N(\sigma_l)}} \quad (3)$$

where $N(\sigma_p, \sigma_l)$ is the number of (σ_p, σ_l) contacts, and $N(\sigma_l)$ and $N(\sigma_p)$ are the number of σ_l and σ_p atoms, respectively, making at least one contact. Importantly, two atoms were defined to be in contact if they were within 5 Å of each other. This length scale was chosen to correspond to the correlation length of solvent ordering, which implicitly accounts for desolvation effects resulting from a (σ_p, σ_l) contact pair formation. The potential parameters $B(\sigma_p, \sigma_l)$ were then obtained by invoking the Boltzmann hypothesis

$$p(\sigma_p, \sigma_l) = p^{\text{ref}} \exp(-B(\sigma_p, \sigma_l)) \quad (4)$$

or equivalently,

$$B(\sigma_p, \sigma_l) = -\ln \left[\frac{p(\sigma_p, \sigma_l)}{p^{\text{ref}}} \right]. \quad (5)$$

Because structural reorganizations of the protein, ligand, and solvent around a particular contact (σ_p, σ_l) are averaged, the potential $B(\sigma_p, \sigma_l)$ should be understood as a potential of mean force. It is important to note that this last statement is based on two main assumptions common to most knowledge-based potentials (e.g., Sippl et al., 1996). First, it is assumed that the binding strength of a protein–ligand complex is correlated to the multiparticle potential of mean force values $[B(\mathbf{r}_1, \dots, \mathbf{r}_n)]$, where n is the number of particles] that are calculated from fixed, crystal atomic positions. The resulting multiparticle potential of mean force is a function of the protein and ligand atomic coordinates alone, since the knowledge-based approach assumes that a sufficiently rigorous statistical analysis from a large number of crystal structures leads to the averaging out of the solvent and intramolecular (such as the protein's) degrees of freedom. Second, it is generally costly to directly calculate $B(\mathbf{r}_1, \dots, \mathbf{r}_n)$ for a series of complexes. The superposition approximation (Hansen & McDonald, 1996) is therefore used to simplify $B(\mathbf{r}_1, \dots, \mathbf{r}_n)$ by writing it as the sum of pairwise atomic potentials of mean force $B(\mathbf{r}_i, \mathbf{r}_j)$, that is,

$$B(\mathbf{r}_1, \dots, \mathbf{r}_n) \approx \sum_{i < j} B(\mathbf{r}_i, \mathbf{r}_j).$$

In doing so, multiparticle correlations that make substantial contributions to the system's partition function are neglected. Although it has not yet been rigorously justified that these two assumptions, which were originally developed for a theory modeling dilute liquids, are applicable to an ensemble of crystal structures, past successes in protein folding and protein–ligand binding energy prediction warrant further investigations of the knowledge-based approach.

It is clear from Equation 3 that the $p(\sigma_p, \sigma_l)$ values are not normalized and consequently should not be understood strictly as probabilities of contact formation. They are instead statistical measures of the likelihood of observing protein–ligand contacts in the database relative to the corresponding values in the reference state (p^{ref}). In our definition, the hypothetical reference state is a purely random mixture of connected protein and connected ligand atoms that do not interact, i.e., $B(\sigma_p, \sigma_l) = 0$ for all (σ_p, σ_l) contact pairs, subject to two constraints: (1) the total number of atoms of a given type is constant, and (2) there is a fixed boundary between the protein and the ligand. DeWitte and Shakhnovich (1996) chose the simple approximation $p^{\text{ref}} = \langle p(\sigma_p, \sigma_l) \rangle_{(\sigma_p, \sigma_l)}$ for the reference state, where $\langle \rangle_{(\sigma_p, \sigma_l)}$ denotes an average over all contact pair types.

A generalization of the SMOG 96 potential

Our generalization of the SMOG 96 potential centers around a reformulation of the reference state probabilities for protein–ligand complexes. We present the following connection between probabilities and the Boltzmann factor:

$$P(\sigma_p, \sigma_l) = C\Omega(\sigma_p, \sigma_l)\exp(-B(\sigma_p, \sigma_l)) \quad (6)$$

where C is the normalization constant and $\Omega(\sigma_p, \sigma_l)$ is the nonenergetic contribution to the observed probabilities. The normalization constant C is then obtained by

$$\ln C = \left\langle \ln \frac{P(\sigma_p, \sigma_l)}{\Omega(\sigma_p, \sigma_l)} \right\rangle_{(\sigma_p, \sigma_l)} + \langle B(\sigma_p, \sigma_l) \rangle_{(\sigma_p, \sigma_l)} \quad (7)$$

with the average contact free energy $\langle B \rangle$ being specified as a parameter. The probabilities $P(\sigma_p, \sigma_l)$ are computed according to

$$P(\sigma_p, \sigma_l) = \frac{N(\sigma_p, \sigma_l)}{\sum_p \sum_l N(\sigma_p, \sigma_l)} \quad (8)$$

and are thus properly normalized. In this formulation, the reference state probability distribution is $P^{\text{ref}} = C\Omega(\sigma_p, \sigma_l)$.

The role of $\Omega(\sigma_p, \sigma_l)$ is to account for statistical effects present in the database that are *independent* of energetic effects. For example, probabilities collected from a database of proteins with deep, hydrophobic pockets will possess an inherent bias toward carbon contacts. In this situation, the low probabilities observed for noncarbon contacts are not a direct reflection of their energetics (e.g., some hydrophilic contacts, such as a hydrogen bonding pair in certain environments, can be very energetically favorable), but instead mirror a distribution of σ_p that is highly biased toward carbon contacts. Our description can decouple such effects from energetic considerations by allowing $\Omega(\sigma_p, \sigma_l)$ to scale as the num-

ber of σ_p atoms in the database. A lower $\Omega(\sigma_p, \sigma_l)$ will imply a reduced $P(\sigma_p, \sigma_l)$, but not necessarily a more unfavorable $B(\sigma_p, \sigma_l)$. We can rewrite the above expression as

$$P(\sigma_p, \sigma_l) = C\exp(-\tilde{B}(\sigma_p, \sigma_l)) \quad (9)$$

with $\tilde{B}(\sigma_p, \sigma_l) = -\ln \Omega(\sigma_p, \sigma_l) + B(\sigma_p, \sigma_l)$. The term $\ln \Omega(\sigma_p, \sigma_l)$ can therefore be viewed as the entropic contribution to the binding free energy parameter that measures the effective number of ways of forming a (σ_p, σ_l) contact pair.

It is clear that the term $\Omega(\sigma_p, \sigma_l)$ must be some function of both $N(\sigma_l)$ and $N(\sigma_p)$, since these numbers will influence the frequency of contacts. To a first approximation, we can assume $\Omega(\sigma_p, \sigma_l)$ to have the functional form

$$\Omega(\sigma_p, \sigma_l) = N(\sigma_l)^\alpha N(\sigma_p)^\beta. \quad (10)$$

The physically allowed values for α and β are from 0 to 1, since $\Omega(\alpha = 1, \beta = 1)$ would equal the number of contacts formed if every protein atom made a contact with every ligand atom of the given types.

We can relate this new formulation to the SMOG 96 approach using proper normalization, which is accomplished by combining Equations 3 and 8 to get

$$P(\sigma_p, \sigma_l) = \frac{P(\sigma_p, \sigma_l)\sqrt{N(\sigma_p)}\sqrt{N(\sigma_l)}}{\sum_p \sum_l N(\sigma_p, \sigma_l)}. \quad (11)$$

The reference state probabilities for the SMOG 96 potential are hence

$$P(\sigma_p, \sigma_l) = Cp^{\text{ref}}\sqrt{N(\sigma_p)}\sqrt{N(\sigma_l)} \quad (12)$$

$$= C'\Omega(\alpha = \frac{1}{2}, \beta = \frac{1}{2}), \quad (13)$$

where C' is a constant.

Results

The removal of compositional effects is critical for successful parameter derivation

Parameters were derived from training sets constructed using two methods (see Materials and methods). Under the *ab initio* method, the training set was composed of randomly selected complexes. In the other, Metropolis Monte Carlo was used to optimize the training set such that the correlation of a testing set was maximized. When the training set was optimized, the parameters were tested on a set completely independent from the optimization procedure (the “independent” training set). In both cases, the success of the parameters was evaluated on a testing set that was not used to selected training set members.

In Figure 1A, we present the correlation between true interaction scores and those computed by the SMOG 96 procedure as a function of the training set size. The procedure did not recover the true parameter set under both derivation methods.

Under a good derivation procedure, the correlation should stabilize or increase with training set size as statistical fluctuations of the contact frequencies decrease. However, as the training set size

Fig. 1. A: The correlation of the true and SMOG 96 testing set binding scores computed with parameters derived from the *ab initio* and optimization procedures. Only well-designed complexes were used in the selection of training sets. For the *ab initio* derivation procedure, the correlation represents an average over 500 runs for each training set size. For the optimization procedure, 100 runs were completed for each set size with 50 trypsin complexes serving as the testing set within the optimization loop. The average correlation of the independent testing set (arabinose-binding protein, acetylcholinesterase, carboxypeptidase, PNP, and streptavidin) is shown. **B:** The free energy parameters derived from training set optimization using the SMOG 96 method. The $\langle C, \sigma_l \rangle$ parameters are shown by thick lines.

was increased, the correlation remained poor (>0.2). A detailed analysis of the parameters derived through training set optimization revealed that this knowledge-based procedure assigned significantly lower parameter values to carbon contacts than others (Fig. 1B). For very small sizes, moderately high correlation was attained only by selecting the few training set complexes with an unusually high number of (N, N) and (O, O) contacts, so that their derived parameter values were lower than that of (C, O) . However, as the training set size increased, complexes with less unusual contact compositions were added. The optimization procedure was then unable to prevent the carbon parameters from becoming the lowest in the potential, contributing to the rapid decline in the correlation.

Since carbon contacts in the true parameter set are only weakly attractive, Equation 4 appeared to overemphasize the strength of carbon interactions. As stated earlier, the observed frequency of carbon contacts in the toy database can be influenced by factors other than energetic ones, whose effects can be carried out of the Boltzmann factor by use of the modified SMOG function (Equation 9). We plotted the dependence of the correlation on α and β in Figure 2A. The maximal correlation was achieved when the exponents $\alpha = \beta = 1$. This confirmed that the true parameter set was not recovered by Equation 4, which used $\alpha = \beta = \frac{1}{2}$, because compositional factors were not adequately addressed.

Figure 2B shows that the parameter $\langle B \rangle$ should be chosen so as to coincide with the actual average of the true parameter set (-0.09).

Fig. 2. A: The average correlation resulting from applying the modified SMOG potential (Equations 6, 9) to randomly selected training sets of 100 complexes for various values of α and β . There is a sharp peak at $\alpha = \beta = 1$. **B:** The effect of the parameter $\langle B \rangle$ on the average correlation for randomly selected training sets from a pool of well-designed complexes. Here, $\alpha = \beta = 1$.

When a value lower than -0.09 was chosen, the correlation did improve with smaller training sets, but declined with larger ones. Because the correlation must not decline with increasing set size, this was an undesirable trend.

How does the degree of design in the toy database affect the correlation?

The correlation was also strongly dependent on the binding strength of ligands selected for the training set (Fig. 3). When poorly designed (i.e., higher true interaction score) ligands were chosen for the training set, the correlation was lower by 0.25 than when well designed (i.e., lower true interaction score) ligands were used. Because of the Metropolis criterion imposed by the SMOG growth procedure, the poorly designed complexes feature either a large number of poorly interacting contacts or a small number of strongly interacting ones. Poor design will therefore increase the statistical noise in the database.

Fragment renormalization improves the correlation

Initially, we believed that family-based averaging should have the effect of leveling off biases caused by the different representation of protein families in the database. The structures available in the Brookhaven Protein Data Bank (PDB) (Bernstein et al., 1997) are often either highly redundant (many ligands bound to a particular protein) or nonredundant (one ligand bound to a single protein). On the one hand, one would like to avoid choosing too many structures involving the same protein, since it would obviously bias the statistics toward interactions characteristic for that protein. On the other, one would like to create a training set with the maximal number of structures to reduce the statistical errors. If there were a sufficiently large number of families, it would seem that family-based averaging would allow one to include all the structures from one family, while minimizing the biases caused by structural redundancies. As shown in Figure 4A, this averaging

Fig. 3. The effect of the training set design on the average correlation for randomly selected training sets with $\alpha = \beta = 1$. The circles and crosses represent training sets drawn from a pool of poorly and well-designed complexes, respectively.

procedure actually worsened the correlation by ~ 0.15 (compare with Fig. 3). We suspect this was due to the fact that there was an insufficiently large number of families to ensure convergence of the B values (see Equation 15).

In contrast, fragment renormalization, where contact statistics were averaged over chemical structure, improved the correlation (Fig. 4B). Although this procedure did not make a difference when working with well-designed training sets, the improvement was substantial for poorly designed sets. Furthermore, the correlation curves do not appear to be leveling off with set size, and it is plausible that for large enough sets, fragment renormalization will yield the best correlation of all the methods investigated for both well and poorly designed databases. Because the contact distribution in ligands grown by SMOG is based on the addition or removal of fragments from a library of a *finite* size, there is a discrepancy between the established atom-atom distribution and the one predicted by the Boltzmann hypothesis. This averaging procedure does appear to account for these biases arising out of fragment-based growth, and a more sophisticated averaging method than the one we outlined here may even further improve the correlation.

Training set optimization leads to perfect recovery of the true parameters

Thus far, we have shown that a simple functional form for $\Omega(\sigma_p, \sigma_l)$ is sufficient to attain a correlation of 0.8 under the *ab initio* derivation method if the proper values are assigned for the exponents α and β (Fig. 3). Using $\alpha = \beta = 1$, it was possible to achieve *near perfect* correlation for the independent testing set when we optimized the training set (Fig. 5). As shown by the SMOG 96 data (Fig. 1), it is important to note that the proper determination of $\Omega(\sigma_p, \sigma_l)$ was critical for the success of this optimization method.

A preliminary test on real proteins

To anticipate future studies, we applied our procedure with the training set optimization method to real proteins. Although the potential is considerably simpler than any recently reported in the literature (pairwise potential with a single fixed interaction distance), the correlation for a diverse independent testing set attained 0.65 and is quite stable with respect to the training set size (Fig. 6A). It is clear that the binding scores obtained with our simple scoring function are qualitatively correlated with experimental binding constants for the independent testing set (Fig. 6B). We are aware that this potential is too simple for real proteins, and we therefore intend to investigate more sophisticated functional forms and atom typing schemes. We are optimistic that higher correlations for real proteins will be obtained in the near future.

Discussion

We have shown that with the use of a self-consistent method, it is possible to test the validity of procedures by which knowledge-based parameters are derived from a database of structures. Using a revised formulation of the Boltzmann hypothesis (Equation 9), we were able to derive parameters from a toy database that strongly correlated with the true parameters. In the process we identified the factors that are critical for improving the SMOG 96 potential. This self-consistent technique is general enough that many existing knowledge-based methods can be studied along these lines, which

Fig. 4. Correlation obtained from using (A) family-based averaging and (B) fragment renormalization for both randomly selected well designed (×) and poorly designed (○) training sets.

Fig. 5. Application of the training set optimization method to the modified SMOG potential. The training set was well designed. Family-based averaging and fragment renormalization were not used.

may assist in understanding what effects contribute to the observed probabilities $P(\sigma_p, \sigma_l)$.

The key modification to the SMOG 96 procedure presented here concerns the treatment of the reference state. Observed contact probabilities are not the result of only energetic effects, as one would believe from a straightforward application of the canonical distribution. Certain atom types (mainly carbon) may exhibit an unusually high contact frequency compared to what one would expect from energetic considerations alone. Overrepresentation can be caused by atoms that form links between fragments that are most important for binding. Atoms may also appear in ligands simply to facilitate chemical synthesis. Such atoms can appear near the protein–ligand interface, and hence contribute to the observed contact distribution. All knowledge-based methods must account

for such nonenergetic influences on contact frequencies when estimating the reference state. Our data suggests that this problem with atom type composition can be addressed by having the reference state probabilities scale as $N(\sigma_p) \cdot N(\sigma_l)$. The major problem with the reference state used in SMOG 96 (see Equation 13) appears to have been the small values of the exponents α and β .

Another problem associated with the reference state concerns correlations in the contact statistics resulting from chemical structure. As different sets of atoms belong to chemically stable fragments, they cannot be viewed as independent species. Hence, in an ensemble of ligands generated from a finite number of fragments, their contact patterns may be strongly correlated. This leads to the difficulty of properly handling such correlations when estimating the state of no interactions. Based on our results, it appears that fragment renormalization can help reduce such effects.

The treatment of the reference state presented here is focused on atom–atom potentials. The estimation of the reference state is expected to be more difficult for atom–atom potentials than for traditional residue–residue protein folding potentials, as the problems mentioned above are expected to be more severe with a lesser degree of coarse graining. Although the solutions presented here may not be directly applicable to residue–residue folding potentials, the self-consistent methodology may be adapted to investigate reference state issues in the context of protein folding (Godzik et al., 1995).

Using random training sets, we were not able to achieve perfect correlation between the true and computed interaction scores. However, when training set optimization was applied, near perfect correlation was obtained. The significance of these results follows from the fact that all knowledge-based potentials are dependent on what complexes are chosen for the training set. A potential may therefore have poor predictive power because the complexes in database were not carefully selected. The optimization method allows one to test potentials under the condition of having the *best possible* database given the available pool of complexes. As we have seen, some potentials do not show improvement with optimization, indicating that they are fundamentally flawed. If a potential does not show near perfect correlation with training set

Fig. 6. **A:** The highest independent testing set correlation attained out of 100 optimization runs for a given training set size. The correlation is between computed binding scores and experimental binding constants for real proteins. **B:** Comparison of the computed binding scores and experimental binding constants for real protein complexes. The parameters were derived using the training set (size = 30) that yielded the highest correlation in **A**. The light symbols denote testing set complexes, and bold symbols independent testing set complexes.

optimization for a self-consistent system, the potential will most likely have little predictive power when applied to real proteins. This should be viewed as a necessary but not sufficient criterion for determining whether the potential have high predictive power. With real proteins, since one can never perfectly know the functional form of nature's potential, one may need to carefully select the training set complexes to attain the *practical* goal of binding free energy prediction. For this reason, it is important to have a potential that is sensitive to training set optimization.

Furthermore, the real protein–ligand complexes available for a training set will inevitably be small in number and limited in diversity. Simply increasing the size of training set without careful consideration may actually worsen the performance of a potential, as statistical biases may be enforced. Rather, a smaller database featuring a diverse set of interactions may yield better results. In a situation where there are restriction on the size of a database, careful consideration of the database composition may be important in obtaining a potential useful for pharmaceutical applications.

As with any optimization procedure, care must be taken in not finding a potential that has limited transferability outside of the training set. We emphasize that the attainment of near perfect correlation with the training set optimization does not necessarily indicate a potential with high predictive power. If the testing set is not sufficiently diverse (e.g., contains only hydrophobic ligands), then the optimization procedure will pick complexes of similar types for the training set, giving rise to a solution that is optimal with respect to only one type of interactions. The parameters obtained will not be suitable to predict binding constants for ligands stabilized by other types of interactions (e.g., hydrogen bonds, polar interactions, etc.). When the training set is optimized for such a testing set, it is also the case that modifications to the potential that improve (or worsen) the correlation under the training set optimization procedure may not improve (or worsen) the predictive power of the potential. For this reason, the predictive power of a potential can only be correctly measured using a testing set that is sufficiently diverse.

Having stated these limitations on the training set optimization method, it is important to note that if properly applied, the method may be useful in troubleshooting a knowledge-based potential. When we applied training set optimization, we always examined the correlations observed in the *independent* testing sets (see Figs. 5, 6A), which were not involved in the optimization loop. If high correlations are observed in the independent testing set and the independent testing set is sufficiently different from the testing set used for training set optimization, it is likely that the potential has high predictive power. On the other hand, if the opposite is true, it is likely that one or more of the following possibilities was true: the testing set and/or the training set pool (from which the optimized training set was chosen) were not sufficiently diverse, or the potential function was flawed.

Although optimization was used to derive our potential for real proteins, we believe the 0.65 correlation we obtained (Fig. 6A) is an accurate indication of its predictive power for the following two reasons. First, the correlation improves or stabilizes upon adding more complexes to the training set. More importantly, the parameters were applied to an independent testing set that has little similarity to the testing set used in optimization. Given the limited number of protein–ligand complexes publicly available, it was necessary to use our intuition in deciding which proteins to include in the pool of training set complexes (from which the optimized training sets were chosen) and the nonindependent testing set. Since maximal diversity is desired in these two sets, we made our selection to maximize the types of pairwise interactions represented in these two sets. It is conceivable that if all selection possibilities are considered a correlation higher than the one we reported will be observed.

We want to emphasize, however, that any kind of optimization is against the spirit of knowledge-based potentials. In principle, knowledge-based methods should not involve any optimization. The fact that (near) perfect correlation was attained only through optimization indicates that the parameter extraction method used was imperfect. We intend to vigorously pursue improvements to

this method so that training set optimization will not be required for achieving better correlation.

Finally, the present analysis was not used to determine the most appropriate functional form of a protein–ligand potential. In this regard, our preliminary results for real proteins are worth noting. Despite the simplicity of the potential, a statistically significant correlation was obtained. Further improvements to the functional form of the potential—such as different length scales for different types of interactions (Muegge & Martin, 1999), environmental context contributions, and greater number of atom types—are likely to lead to greater predictive power. Steady progress in this direction will hopefully provide a solution of the problem of computational ligand design.

Materials and methods

Creation of model databases

The true potential consisted of a 5×4 matrix of the contact parameters $B(\sigma_p, \sigma_l)$, in which the protein and ligand atoms were classified into five (C, O, N, S, MEtal) and four types (C, O, N, S), respectively. The length scale of interaction was set at 3.5 Å. It is important to note that to avoid having the low free energy ligands composed entirely of carbon atoms, the true parameter set had to be chosen such that the strongest interactions involved noncarbon types. Our true parameter set therefore featured sufficiently low values for (O, O) and (N, N), but only slightly attractive values for (C, σ_l) and (σ_p , C) contacts (Table 1). Using a fragment library of 137 common organic fragments, including rings and heterocycles, ligands were grown under the true potential for 14 proteins. After SMOG growth, the ligand interaction scores for each protein in the database was canonically distributed (Fig. 7). The resulting protein–ligand complexes were divided into two separate nonintersecting groups. The first group (alpha-lytic protease, carbonic anhydrase, chymotrypsin, endothiapepsin, lysozyme, matrix metalloproteinase, papain, and SH3 domain) formed the pool of complexes from which the members of the training set were selected. The second group (arabinose-binding protein, acetylcholinesterase, carboxypeptidase, PNP, streptavidin, and trypsin) formed the pool of available testing set members.

Two procedures for deriving and testing parameter sets

Once the parameters were derived from the training set, they were used to compute the interaction scores of the testing set complexes.

Table 1. The true parameter set used to generate model ligands^a

		σ_l			
		C	O	N	S
σ_p	C	−0.035	−0.01	−0.01	−0.03
	O	−0.02	−0.21	−0.03	−0.11
	N	−0.02	−0.03	−0.26	−0.11
	S	−0.5	0.15	0.05	0.1
	ME	−0.4	−1.00	0.3	0.3

^a σ_p and σ_l refer to protein and ligand types, respectively.

Fig. 7. The distribution of true binding free energies for ligands grown by SMOG for five proteins: CA, carbonic anhydrase; CHY, chymotrypsin; PNP, purine nucleoside phosphorylase; SH3, SH3 domain; STP, streptavidin.

The correlation between the computed and true interaction scores for the testing set complexes was used to evaluate the success of the parameter derivation method. Under this general scheme, two methods for deriving potentials were explored:

1. *Ab initio derivation* (Fig. 8A). In this situation, the parameters were calculated from randomly selected training sets and immediately applied to the testing set.
2. *Derivation by training set optimization* (Fig. 8B). The training set was optimized to maximize the correlation between the computed and true testing set interaction scores. The optimization was initiated by selecting a random training set of a given size. At each step, a randomly selected member of the training set was swapped with another complex from the pool of available training set members. The move was accepted with a probability determined by the Metropolis criterion:

$$p = \begin{cases} 1 & \text{if } \Delta C \geq 0 \\ \exp(\Delta C/T) & \text{otherwise} \end{cases} \quad (14)$$

where ΔC is the change in the correlation after the move and T is the simulation temperature. The correlation for a second “independent” testing set was tracked during the optimization, but was not used in the optimization procedure.

Calculation of knowledge-based parameters

The probabilities $P(\sigma_p, \sigma_l)$, the $\Omega(\sigma_p, \sigma_l)$ factors, and the normalization factor C were computed according to Equations 8, 10, and 7, respectively.

Two averaging strategies were tested in conjunction with the potential:

1. *Family-based averaging.* The proteins in the training set were divided into families. For our model database, all complexes

Fig. 8. Parameter sets are derived and tested in two ways. **A:** Ab initio derivation. **B:** Training set optimization. The dotted box represents the optimization loop from which the parameters are derived. Note that information involving the independent testing set is not used in the derivation.

sharing a common protein were grouped into a single family. The parameters were determined as an average over families:

$$B(\sigma_p, \sigma_l) = \langle B^\alpha(\sigma_p, \sigma_l) \rangle_\alpha \quad (15)$$

where $B^\alpha(\sigma_p, \sigma_l)$ is the parameter value calculated solely from the database of structures belonging to the family α .

2. *Fragment renormalization.* Contact frequencies of atoms within the same fragment in the ligand were scaled by $1/N_f$, where N_f is the size of the fragment. Atoms were said to belong to the same fragment if they were connected by nonrotatable bonds (see Fig. 9 for an illustration).

Application to real proteins

Training set optimization was used to obtain free energy parameters for real proteins. The pool of protein–ligand complexes from

which the optimized training set was obtained, consisted of structures obtained from the Brookhaven PDB or contributions from other labs (PNP). The following sets of structures were selected for the pool: carbonic anhydrase (26 structures), carboxypeptidase (9), chymotrypsin (19), elastase (22), HIV protease (12), penicillopepsin (11), serine protease (8), and thermolysin (13). Three proteins comprised the testing set: purine nucleoside phosphorylase (5), endothiapepsin (7), and a selection (6 in total: FKBP, retinol binding protein, streptavidin, intestinal fatty acid binding protein, p-hydroxybenzoate hydroxylase, triose phosphate isomerase) from the training set described in Böhm (1994). And, finally, the independent testing set was made of various complexes of sugar binding proteins (15), matrix metalloproteinases (7), and alpha-lytic proteases (6). The binding constants for all complexes are taken from the literature (Babine & Bender, 1997; Eldridge et al., 1997; Boriack-Sjodin et al., 1998) or provided by the contributors. The extended set of atom types was used for the real complexes. Eleven

Fig. 9. An example of the fragment renormalization procedure. Fragments were identified by a graph theoretical method. Once the ligand's bonding structure was transformed into a graph, articulation points, which are points if removed would disconnect the graph into two separate graphs, were identified (Sedgewick, 1990). Bonds spanning two articulation points were classified as rotatable if the chemical hybridization of at least one of the points is greater than sp^2 .

protein types (*C3*, *C2*, *CP*, *CA*, *CC*, *OD*, *OA*, *ND*, *NM*, *S*, *ME*) and eight ligand types (*C3*, *C2*, *CP*, *OD*, *OA*, *ND*, *NA*, *S*) were defined, where *C3* refers to sp^3 nonpolar carbon; *C2*, sp^2 nonpolar carbon; *CP*, polar carbon; *CC*, carbonyl carbon; *CA*, protein alpha backbone carbon; *OD*, oxygen hydrogen bond donor; *OA*, oxygen hydrogen bond acceptor; *ND*, nitrogen hydrogen bond donor; *NA*, nitrogen hydrogen bond acceptor; *NM*, protein backbone nitrogen; *S*, sulfur; *ME*, metal atom. $\langle B \rangle$ was chosen to be zero since it yielded the highest correlation upon performing an investigation similar to that described in Figure 2B.

Acknowledgments

We would like to thank Professor Steve Elick of Cornell University for generously providing us with unpublished structures and binding constants for several PNP-inhibitor complexes.

References

- Ajay, Murcko MA. 1995. Computational methods to predict binding free energy in ligand-receptor complexes. *J Med Chem* 38:4953–4967.
- Babine RE, Bender SL. 1997. Molecular recognition of protein–ligand complexes: Applications to drug design. *Chem Rev* 97:1359–1472.
- Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1997. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol* 112:535–542.
- Betancourt MR, Thirumalai D. 1999. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci* 8:361–369.
- Binder K, Heerman DW. 1992. *Monte Carlo simulations in statistical physics*. Berlin: Springer-Verlag.
- Böhm H. 1994. The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. *J Comput-Aided Mol Des* 8:243–256.
- Böhm H. 1998. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J Comput-Aided Mol Des* 12:309–323.
- Boriack-Sjodin PA, Zeitlin S, Chen H, Crenshaw L, Gross S, Dantanarayana A, Delgado P, May JA, Dean T, Christianson DW. 1998. Structural analysis of inhibitor binding to human carbonic anhydrase II. *Protein Sci* 7:2483–2489.
- Chaikin PM, Lubensky TC. 1995. *Principles of condensed matter physics*. Cambridge, UK: Cambridge University Press.
- DeWitte RS, Ishchenko AV, Shakhnovich EI. 1997. SMOG: De novo design method based on simple, fast and accurate free energy estimates. 2. Case studies in molecular design. *J Am Chem Soc* 119:4608–4617.
- DeWitte RS, Shakhnovich EI. 1994. Pseudodihedrals: Simplified protein backbone representation with knowledge-based energy. *Protein Sci* 3:1570–1581.
- DeWitte RS, Shakhnovich EI. 1996. SMOG: De novo design method based on simple, fast and accurate free energy estimates. 1. Methodology and supporting evidence. *J Am Chem Soc* 118:11733–11744.
- Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. 1997. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput-Aided Mol Des* 11:425–445.
- Finkelstein AV, Badretdinov AY, Gutin AM. 1995. Why do protein architecture have Boltzmann-like statistics? *Proteins Struct Funct Genet* 23:142–150.
- Godzik A, Kolinski A, Skolnick J. 1995. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci* 4:2107–2117.
- Goldstein RA, Luthey-Schulten ZA, Wolynes PG. 1992a. Optimal protein-folding codes from spin-glass theory. *Proc Natl Acad Sci USA* 89:4918–4922.
- Goldstein RA, Luthey-Schulten ZA, Wolynes PG. 1992b. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc Natl Acad Sci USA* 89:9029–9033.
- Hansen J-P, McDonald IR. 1996. *Theory of simple liquids*. London: Academic Press.
- Head RD, Smythe ML, Oprea TI, Waller CL, Green SM, Marshall GR. 1996. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J Am Chem Soc* 118:3959–3969.
- Hinds DA, Levitt M. 1994. Exploring conformational space with a simple lattice model for protein structure. *J Mol Biol* 243:668–682.
- Jernigan R, Bahar I. 1996. Structure-derived potentials for folding simulations. *Curr Opin Struct Biol* 6:195–209.
- Kocher JA, Rooman MJ, Wodak SJ. 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence–structure matches. *J Mol Biol* 235:1598–1613.
- Kollman PA. 1993. Free energy calculations—Applications to chemical and biological phenomena. *Chem Rev* 7:2395–2417.
- Martin YC. 1997. Challenges and prospects for computational aids to molecular diversity. *Perspect Drug Discovery Des* 7/8:159–172.
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys* 21:1087–1092.
- Micheletti C, Maritan A, Banavar JR. 1999. A comparative study of existing and new design techniques for protein models. *J Chem Phys* 110:9730–9738.
- Mirny LA, Shakhnovich EI. 1996. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol* 264:1164–1179.
- Mitchell JBO, Laskowski RA, Alex A, Thornton JM. 1999. BLEEP—Potential of mean force describing protein–ligand interactions: I. Generating potential. *J Comput Chem* 20:1165–1176.
- Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* 18:534–552.
- Muegge I, Martin YC. 1999. A general and fast scoring function for protein–ligand interactions: A simplified potential approach. *J Med Chem* 42:791–804.
- Rao BG, Tilton RF, Singh UC. 1992. Free energy perturbation studies on inhibitor binding to HIV-1 proteinase. *J Am Chem Soc* 114:4447–4452.
- Sedgewick R. 1990. *Algorithms in C*. Reading, MA: Addison-Wesley.
- Sippl MJ. 1990. Calculation of conformational ensemble from potential of mean force. An approach to knowledge-based prediction of local structures in globular proteins. *J Mol Biol* 213:859–883.
- Sippl MJ, Ortner M, Jaritz M, Lackner P. 1996. Helmholtz free energies of atom pair interactions in proteins. *Fold Des* 2:289–298.
- Thomas PD, Dill KA. 1996. Statistical potentials extracted from protein structures: How accurate are they? *J Mol Biol* 257:457–469.
- Verkhivker G, Appelt K, Freer ST, Villafranca JE. 1995. Empirical free energy calculations of ligand–protein crystallographic complexes. I. Knowledge-based ligand–protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng* 8:667–691.
- Wallquist A, Jernigan R, Covell D. 1995. A preference-based free-energy parametrization of enzyme–inhibitor binding–application to HIV-1 protease inhibitor design. *Protein Sci* 4:1881–1903.