Journal of Computer-Aided Molecular Design, **12:** 27–35, 1998.
KLUWER/ESCOM

Ligand binding affinity prediction by linear interaction energy methods

Tomas Hansson, John Marelius & Johan Åqvist*
Department of Molecular Biology , Uppsala University, Biomedical Centre, Box 590, S-751 24 Uppsala, Sweden

Received 20 April 1997; Accepted 16 August 1997

Key words: binding energy, linear response, molecular dynamics simulations, structure-based ligand design

Summary

A recent method for estimating ligand binding affinities is extended. This method employs averages of interaction potential energy terms from molecular dynamics simulations or other thermal conformational sampling techniques. Incorporation of systematic deviations from electrostatic linear response, derived from free energy perturbation studies, into the absolute binding free energy expression significantly enhances the accuracy of the approach. This type of method may be useful for computational prediction of ligand binding strengths, e.g., in drug design applications.

Introduction

A key problem in computational structure-based ligand design is the estimation of binding affinities for putative receptor-ligand complexes. Several approaches to this problem have been developed, ranging from quick estimates using empirical scoring functions [1–10] to those based on force field calculations. The latter, in turn, span from relatively fast scoring of energy minimized single conformations [11–15] to more complex calculations of free energies involving thermal averaging, such as rigorous free energy perturbation (FEP) techniques [16–19], or simplified versions thereof [20, 21].

It has recently been demonstrated that fairly accurate binding free energy estimates can be obtained from molecular dynamics simulation energy averages, in the context of a linear response assumption for electrostatic interactions together with an empirical expression for nonpolar effects [22–27]. The characteristics of this approach are that (1) thermal averaging over conformations is carried out; (2) in contrast to FEP, no transformation processes involving unphysical states are required; (3) the method is not restricted to relative free energy calculations but is based on estimating the absolute binding energies; and (4) it is considerably slower than scoring of single conformations, but faster than rigorous FEP calculations.

Our implementation of the above approach, which we will refer to as the linear interaction energy (LIE) approximation, is based on evaluation of ligand interaction energetics in the bound and free states. The main trick, deriving from linear response considerations, is that only (convergent) averages of the interaction energies between the ligand and its surroundings need to be evaluated. Accordingly, two MD simulations are carried out, one with the ligand free in solution and one where it is bound to the solvated receptor. The binding free energy is obtained from these averages as

$$\begin{aligned}\Delta G_{\text{bind}} &\simeq \Delta G^{\text{nonpolar}} + \Delta G^{\text{polar}} \\ &= \alpha(\langle V_{l-s}^{\text{vdw}}\rangle_{\text{bound}} - \langle V_{l-s}^{\text{vdw}}\rangle_{\text{free}}) \\ &\quad + \tfrac{1}{2}(\langle V_{l-s}^{\text{el}}\rangle_{\text{bound}} - \langle V_{l-s}^{\text{el}}\rangle_{\text{free}}),\end{aligned} \tag{1}$$

where the subscript $l-s$ denotes ligand–surroundings interactions, while $l-l$ and $s-s$ terms do not enter into the formula. Note that 'surroundings' here denotes all parts of the system except the ligand, such as protein, solvent, co-factors, ions etc. The superscripts vdw and el designate the corresponding terms of the full potential energy function

$$V_{\text{tot}} = V_{l-l} + V_{s-s} + V_{l-s}^{\text{vdw}} + V_{l-s}^{\text{el}}, \tag{2}$$

* To whom correspondence should be addressed.

where the first two terms include both bonded and non-bonded interactions, while the latter two represent the (non-bonded) interactions between the ligand and its surroundings, i.e. Lennard-Jones (12–6) and electrostatic pair potentials. The coefficient $\frac{1}{2}$ in Equation (1) is a consequence of assuming a linear response of the surroundings to electrostatic fields, whereas α is an empirical constant that can be fitted to experimental binding free energies [22]. We have recently examined the range of validity of the electrostatic linear response approximation in more detail [28], as will be discussed below. Equation (1) is, however, based on the assumption that it is exactly valid.

The physical rationale for estimating a nonpolar (hydrophobic and van der Waals) binding contribution by a simple scaling of Lennard-Jones energy averages is as follows. Solvation free energies of nonpolar molecules depend to a good approximation linearly on measures of the molecular size, i.e. chain length, surface area or molecular volume as has been discussed extensively [29–31]. Furthermore, simulation averages of the $l - s$ Lennard-Jones term have also been found to scale linearly with such size measures for various solutes [22, 32]. Thus it seems reasonable to expect a hydrophobic contribution to ΔG according to Equation (1). The empirical coefficient α was initially obtained by fitting to experimental data on a small set of endothiapepsin inhibitors, and found to be 0.161 [22] using a version [22, 33] of the GROMOS force field [34].

The reason why surroundings–surroundings and intramolecular ligand–ligand interactions do not enter in Equation (1) has been discussed elsewhere [22, 28]. It should be noted here that there is no reason to expect a linear response of the surroundings to the intra-ligand interactions, in general. For instance, the fixed covalent bond structure of the ligand often gives rise to nearly constant intramolecular contributions, with a magnitude which will depend on the non-bonded exclusion logic of the force field. On the other hand, the appearance of only $\langle V^{\mathrm{el}}_{l-s} \rangle$ in Equation (1) corresponds to an implicit assumption that the intramolecular contributions themselves respond linearly to the ligand–surroundings interactions V^{el}_{l-s}.

The model of Equation (1), with its initial parameterization [22], has been applied to several different proteins and ligands [23–25, 27] with encouraging results (see Figure 1). Interestingly, the same value of α has proven to yield reasonable accuracy for this whole range of diverse systems. Here, we will discuss how the original model can be improved by incorporation of results from studies of electrostatic response in polar solvents, which indicate systematic deviations from linear response for certain types of ligands. Also, the increased amount of simulation data now available allows for a more reliable parameterization of the model.

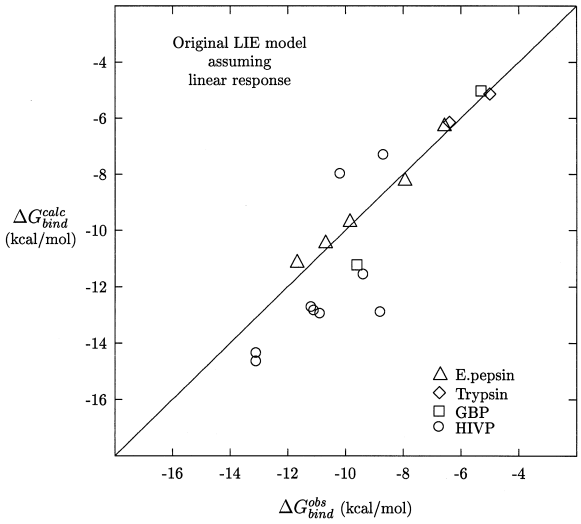

Figure 1. Calculated vs observed absolute free energies of binding for the 18 compounds in Table 1, obtained with the original model (model 0) [22].

Evaluating model extensions

Addition of parameters to a computational model, in order to fit experimental data, should be done carefully to avoid overinterpretation of the data. Also, a model based on physical principles should not be abandoned for a model solely based on statistical fitting. With this in mind, we have examined the performance of a set of different optimized models utilizing results (cf. Table 1) from simulations of endothiapepsin [22], HIV proteinase [23, 24], glucose binding protein [25] and trypsin [27].

The idea that water and protein could have different electrostatic response properties, suggested by recent investigations [28], was pursued by introducing different electrostatic scaling coefficients β_{wat} and β_{prot} for the free and bound states, respectively. Similarly, different nonpolar scaling coefficients α_{wat} and α_{prot} could conceivably allow more accurate modelling of the respective properties of the water and protein environments. In addition, a constant free energy term γ could arise as a non-zero difference

Table 1. Simulation energy averages and experimental binding data (kcal/mol)

Protein–ligand	$\langle V^{\mathrm{vdw}}\rangle_{\mathrm{bound}}$	$\langle V^{\mathrm{vdw}}\rangle_{\mathrm{free}}$	$\langle V^{\mathrm{el}}\rangle_{\mathrm{bound}}$	$\langle V^{\mathrm{el}}\rangle_{\mathrm{free}}$	$\mathrm{Corr}_{\mathrm{el}}$	Class	$\Delta G^{\mathrm{obs}}_{\mathrm{bind}}$	Ref.
EP–I1	−99.42	−59.36	−49.75	−41.72	0.00	1	−10.69	22
EP–I2	−95.48	−58.90	−49.07	−44.38	0.00	1	−7.93	22
EP–I3	−99.05	−58.16	−54.73	−45.61	0.00	1	−11.67	22
EP–I4	−72.17	−40.60	−51.25	−48.84	0.00	2	−6.57	22
EP–I5	−104.77	−62.91	−63.26	−57.33	0.00	1	−9.84	22
Trp–Bza	−20.87	−7.94	−125.38	−115.18	+1.00	crg	−6.4	27
Trp–BzaCOOEt	−29.46	−15.86	−133.62	−125.67	+1.00	crg	−5.0	27
GBP–glc	−25.7	−8.9	−87.0	−69.9	0.00	5	−9.6	25
GBP–meg	−32.0	−12.4	−72.1	−68.3	0.00	4	−5.3	25
HIVP–AcPep	−89.13	−47.29	−255.84	−255.19	−0.90	crg	−10.2	23
HIVP–Pep	−100.26	−55.08	−233.01	−234.99	−1.00	crg	−8.7	23
HIVP–DMP	−86.57	−42.07	−82.22	−67.89	0.00	4	−13.1	24
HIVP–1a	−85.76	−46.66	−50.99	−38.19	0.00	2	−11.2	24
HIVP–2	−86.32	−48.42	−45.17	−34.29	0.00	2	−9.4	24
HIVP–3	−90.37	−52.59	−50.57	−37.00	0.00	2	−8.8	24
HIVP–4	−88.85	−50.91	−49.70	−36.09	0.00	2	−10.9	24
HIVP–5	−88.42	−49.11	−48.97	−36.01	0.00	2	−11.1	24
HIVP–7	−90.86	−44.13	−84.98	−70.78	0.00	2	−13.1	24

EP: endothiapepsin, Trp: β-trypsin, GBP: *S. typhimurium* glucose binding protein and HIVP: HIV-1 proteinase. I1 to I5 are peptide analog inhibitors [22]. Bza: benzamidine, BzaCOOEt: *p*-carbethoxy-benzamidine, glc: β-D-glucose, meg: methyl-β-D-galactoside. AcPep and Pep are the peptide analogs acetylpepstatin and pepstatin, while DMP and 1a to 7 are cyclic urea inhibitors [23, 24]. $\langle V^{\mathrm{vdw}}\rangle_{\mathrm{bound}}$, $\langle V^{\mathrm{vdw}}\rangle_{\mathrm{free}}$, $\langle V^{el}\rangle_{\mathrm{bound}}$ and $\langle V^{el}\rangle_{\mathrm{free}}$ are simulation averages of ligand–surrounding components of the non-bonded potential energy, where *bound* and *free* denote simulations of the ligand inside the solvated receptor and free in solution, respectively, while vdw refers to nonpolar interactions modeled by a Lennard-Jones potential, and el denotes electrostatic interactions. To correct for excluded protein charges distant from the ligand in the simulations, a free energy correction $\mathrm{corr}_{\mathrm{el}}$ (from Coulumb's law with a dielectric constant of 80) is to be added to the computed values for charged ligands [23, 27]. The class of a ligand with a net charge is 'crg', while the class is denoted by the number of hydroxyl groups for uncharged ligands. $\Delta G^{\mathrm{obs}}_{\mathrm{bind}}$ are experimental binding data.

$\gamma_{\mathrm{prot}} - \gamma_{\mathrm{wat}}$ between constant terms in corresponding linear expressions for the solvation energies [22]. Such a constant term in the solvation energy has been discussed by Ben-Naim and co-workers [29, 30]. Several empirical scoring functions for estimating binding free energies from structural data also include a constant term, sometimes proposed to represent entropic contributions to the free energy of binding [1–10]. Jorgensen and co-workers [32, 35] have instead added an accessible surface area dependent term to Equation (1), but this amounts to the same thing as the addition of a constant given that $\langle V^{\mathrm{vdw}}_{l-s}\rangle$ depends linearly on surface area, as results in Refs. 22 and 32 suggest.

A generalized version of Equation (1) can then have the following form

$$\Delta G_{\mathrm{bind}} = \alpha_{\mathrm{prot}}\langle V^{\mathrm{vdw}}_{l-s}\rangle_{\mathrm{bound}} - \alpha_{\mathrm{wat}}\langle V^{\mathrm{vdw}}_{l-s}\rangle_{\mathrm{free}} + \beta_{\mathrm{prot}}\langle V^{el}_{l-s}\rangle_{\mathrm{bound}} - \beta_{\mathrm{wat}}\langle V^{\mathrm{el}}_{l-s}\rangle_{\mathrm{free}} + \gamma \qquad (3)$$

where the α and β parameters are van der Waals and electrostatic scaling factors, respectively, that may now have different values in water and protein environments, while γ is a constant term as described above.

MD simulation data and experimental binding free energies for the 18 receptor–ligand complexes in Table 1 were used to determine optimal coefficients, α_i, β_i and γ, for different versions of Equation (3) by minimization of the residual square sum Q (i.e., least-squares optimization) between model results and experimental values. Minimization of the sum of absolute residuals gave very similar results (not shown). All simulations were performed using a version of the ENZYMIX program [36] and of the GROMOS force field [22, 33, 34].

Addition of adjustable parameters always lowers Q by incorporating idiosyncracies of the data into the added degrees of freedom. We therefore employed

a statistical 'figure of improvement' η to gauge the decrease in Q between two models with different numbers of parameters while penalizing parameter addition and rewarding the existence of many data points lending statistical support to the difference in Q between the models.

$$\eta_{\text{new,old}} = \frac{Q_{\text{old}} - Q_{\text{new}}}{Q_{\text{new}}} \frac{N - n_{\text{new}}}{n_{\text{new}} - n_{\text{old}}} \qquad (4)$$

where Q_{old} and Q_{new} are residual square sums in the two models, n_{old} and n_{new} are the number of parameters in the models and N is the number of data points used in the analysis. Under certain conditions η can be shown to be f-distributed, $\eta \in f(n_{\text{new}} - n_{\text{old}}, N - n_{\text{new}})$, allowing us to estimate whether a model with more coefficients is significantly better than one with fewer coefficients by comparing the η value with the 95% quantile of the f distribution.

The results of these reparameterizations are summarized in Table 2, which also shows some common statistical figures of merit for linear regression models [37], to aid in comparing the different models discussed here: the multiple correlation coefficient r^2 (also known as the 'goodness-of-fit'), the leave-one-out cross-validated correlation coefficient q^2_{LOO} and the leave-one-out cross-validated standard deviation $s^{\text{LOO}}_{\text{PRESS}}$.

The original model (model 0), with Equation (1) parameterized according to Ref. 22, gives an rms deviation of 1.57 kcal/mol for the entire data set of Table 1. Not surprisingly, this value is significantly higher than that obtained for the initial calibration set of endothiapepsin inhibitors [22], which only encompassed a few ligands. Nevertheless, the magnitude of this error still supports the basic assumptions underlying Equation (1), namely that binding free energies can be estimated from the ligand–surrounding energy averages only. A new calibration of the α coefficient of Equation (1) using all the data of Table 1 (model 1) gives a slightly different value of $\alpha = 0.144$ with an rms deviation of 1.45 kcal/mol. This rather small improvement demonstrates that the main source of discrepancy is not the size of the calibration data set, but that the form of the equation needs to be reconsidered in order to achieve more accurate predictions. This is also in agreement with our earlier findings [28].

The main conclusion from the work in Ref. 28 is that deviations from electrostatic linear response may be particularly important for uncharged ligands with certain dipolar groups, suggesting that $\beta < \frac{1}{2}$ should be considered. The more general equation (Equation 3), with $\alpha_{\text{prot}} = \alpha_{\text{wat}}$, $\beta_{\text{prot}} = \beta_{\text{wat}}$ and $\gamma = 0$ (model 2), was found to give a significant improvement over model 1 with $\eta_{2,1} = 19.7 > 4.49 = f_{0.05}(1, 16)$. That is to say that the addition of the new variable parameter is well supported by the data. The resulting values of α and β are 0.186 and 0.320, with an rms deviation of 0.97 kcal/mol. Since most ligands in Table 1 are uncharged, the value of beta obtained for model 2 is in accordance with the results of Ref. 28, where electroneutral compounds were shown to be associated with β-factors of around 0.3–0.4.

Since water and protein are believed to have different dielectric properties, one should also consider the possibility that the coefficients employed in Equation (3) could be different in the two environments. This issue was addressed for α and β separately (models 3 and 4). The introduction of two different α's is clearly not supported by the data, yielding $\eta_{3,2} = 1.9 < 4.54 = f_{0.05}(1, 15)$. That is to say that model 3 offers no improvement over model 2, and, indeed, the values obtained for α_{prot} and α_{wat} are quite similar. The use of two different β's (model 4) does show an improvement over model 2, with the coefficients $\alpha = 0.161$, $\beta_{\text{prot}} = 0.351$ and $\beta_{\text{wat}} = 0.343$ yielding an rms deviation of 0.81 kcal/mol. These values of the β coefficients are, however, very close to one another. In fact, if the minimized residual square sum is plotted as a function of β_{prot} and β_{wat}, an optimal valley can be seen along the diagonal $\beta_{\text{prot}} = \beta_{\text{wat}}$. This indicates that the major characteristics of the electrostatic response are virtually the same in the different environments. However, the decrease in rms deviation for this model compared to model 2 is rather modest (as is also shown by the corresponding value for $\eta_{4,2}$ of 6.9). It may be considered quite remarkable that the electrostatic coefficients, which reflect the relationships between solvation energy and interaction energy in the respective media, are found to be the same in water and protein. This has implications for, e.g. the validity of linear free energy relationships in enzymes [38–41].

The addition of a constant term in expressions for binding free energy estimation has been proposed earlier, in different contexts and on various grounds [1–15]. As far as solvation energies are concerned, a constant term in the corresponding expression is well motivated both by experimental and theoretical studies [29, 30, 42]. The question here is whether such constant terms cancel in binding calculations of the present type, which formally deal with differences in solvation energies. Model 5 considers a residual con-

Table 2. Summary of free energy estimation models

No.	α	β	γ	Rms	⟨|err|⟩	r^2	q^2_{LOO}	$s^{\mathrm{LOO}}_{\mathrm{PRESS}}$
0	0.161	$\frac{1}{2}$	0	1.57	1.20	0.797	0.659	1.88
1	0.144	$\frac{1}{2}$	0	1.45	1.17	0.805	0.723	1.65
2	0.186	0.320	0	0.97	0.72	0.882	0.829	1.20
3	$\alpha_{\mathrm{prot}} = 0.244$ $\alpha_{\mathrm{wat}} = 0.286$	0.299	0	0.92	0.71	0.893	0.831	1.23
4	0.161	$\beta_{\mathrm{prot}} = 0.351$ $\beta_{\mathrm{wat}} = 0.343$	0	0.81	0.54	0.898	0.840	1.08
5	0.147	0.274	−1.87	0.85	0.65	0.870	0.798	1.12
6	0.181	β_{FEP} (class)	0	0.84	0.58	0.890	0.860	0.94
7	0.180	β_{FEP} (class)	−0.03	0.84	0.58	0.889	0.851	1.00

Optimal coefficients from least-squares optimization of the different models described. Root mean square (rms) errors and average absolute errors ⟨|err|⟩ refer to deviations, in kcal/mol, between the optimized models and the experimental values for the 18 systems in Table 1. Models 6 and 7 involve the electrostatic coefficients β_{FEP} (class), which are $\beta_{\mathrm{FEP}}(\mathrm{crg}) = 0.50$, $\beta_{\mathrm{FEP}}(0) = 0.43$, $\beta_{\mathrm{FEP}}(1) = 0.37$ and $\beta_{\mathrm{FEP}}(\geq 2) = 0.33$. These were taken from FEP calculations of the electrostatic response of an aqueous solution to model compounds. The multiple correlation coefficient r^2 of a model measures what proportion of the variance observed in the experimental binding data is 'explained' as model response to the simulation input data x, $r^2 = \mathrm{SSR}/(\mathrm{SSR} + \mathrm{SSE})$, where SSR is the square sum of deviations explained, $\mathrm{SSR} = \sum_x (\Delta G_{\mathrm{model}}(x) - \langle \Delta G_{\mathrm{expt}} \rangle)^2$, and SSE is the residual unexplained square sum of errors $\mathrm{SSE} = \sum_x (\Delta G_{\mathrm{expt}}(x) - \Delta G_{\mathrm{model}}(x))^2$. An 'ideal' model would explain all variation observed, $r^2 = 1.000$. Data overfitting was assessed by leave-one-out cross-validation: optimal parameters were found using each of the 18 data sets missing one of the compounds. Each resulting model parameterization was used to predict the left out ΔG from the left out simulation data, and the square sum of deviations of these predictions from experimental values is the 'Predictive REsidual Sum of Squares' (PRESS). The leave-one-out cross-validated correlation coefficient is then $q^2_{\mathrm{LOO}} = 1 - (\mathrm{PRESS}/\mathrm{SSR})$ and the cross-validated standard deviation is $s^{\mathrm{LOO}}_{\mathrm{PRESS}} = \sqrt{\mathrm{PRESS}/(n - p - 1)}$, where $n = 18$ is the data set size, and p is the number of optimized parameters in the model. γ and $s^{\mathrm{LOO}}_{\mathrm{PRESS}}$ are in kcal/mol.

stant γ, but was found not to improve the agreement with experiment significantly.

The above discussion shows that the introduction of additional model parameters yields relatively little improvement, after adjustment of β away from $\frac{1}{2}$, reflecting deviations from electrostatic linear response, in accordance with Ref. 28. This conclusion is supported by the statistical figures of merit in Table 2.

Incorporating FEP results on deviations from linear response

Our recent investigation of the validity of the linear response assumption in polar solvents [28] showed that the electrostatic scaling factor β, in fact, varies with the chemical composition of the solute. These results can be directly incorporated into Equation (3) by assigning specific values of β to different classes of solutes (ligands). Free energy perturbation calculations showed that electrostatic linear response is obeyed to a good approximation for solvated ions, where accordingly the ratio between ΔG^{el} and $\langle V^{\mathrm{el}}_{l-s} \rangle$ is close to $\frac{1}{2}$. For dipolar compounds the free energy of charging was generally found to be slightly less than $\frac{1}{2}\langle V^{\mathrm{el}}_{l-s} \rangle$ [22, 28, 32]. Furthermore, a trend was observed for the quotient β to vary with the number of hydroxyl groups, so that compounds with more hydroxyl groups are associated with lower values of β. This peculiar behaviour was found to derive from interactions with the hydrogen bonding network of the solvent [28].

The above considerations suggest a model where these chemical composition dependent deviations from linear response are taken into account. A simple implementation of this scheme (model 6) divides compounds into four classes: charged, dipolar with no hydroxyl groups, dipolar with one hydroxyl group and dipolar with two or more hydroxyl groups, and assigns a different β parameter value to each class. These values of β were taken directly from simulations of typical compounds of the different classes (cf. Table 1 of Ref. 28). Examples of such typical compounds are sodium ion, acetone, ethanol and ethylene glycol, respectively

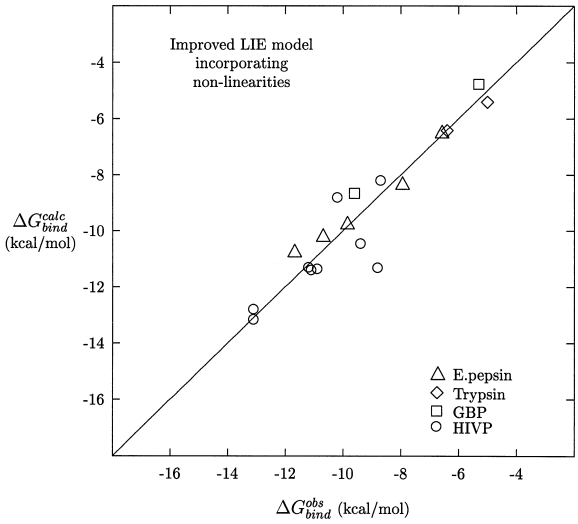

Figure 2. Calculated vs observed absolute free energies of binding for the 18 compounds in Table 1, obtained with the improved model (model 6). This model incorporates specific deviations from electrostatic linear response, derived from FEP calculations [28].

[28]. Model 6 thus corresponds to the approximation

$$\Delta G_{\text{bind}} = \alpha(\langle V^{\text{vdw}}_{l-s}\rangle_{\text{bound}} - \langle V^{\text{vdw}}_{l-s}\rangle_{\text{free}}) + \beta_{\text{FEP}}\,(\langle V^{\text{el}}_{l-s}\rangle_{\text{bound}} - \langle V^{\text{el}}_{l-s}\rangle_{\text{free}}), \quad (5)$$

where

$$\beta_{\text{FEP}}(\text{class}) = \begin{cases} 0.50 \text{ for the 'charged' class,} & (q \neq 0), \\ 0.43 \text{ if class} = 0 & (q = 0, n_{\text{OH}} = 0), \\ 0.37 \text{ if class} = 1 & (q = 0, n_{\text{OH}} = 1), \\ 0.33 \text{ if class} \geq 2 & (q = 0, n_{\text{OH}} \geq 2) \end{cases}$$

It should be emphasized here that this model effectively has only one free parameter, namely α. Based on the results described earlier, the same value of β is used for the bound and free states. Actually, no representative of the dipolar class without hydroxyl groups is present in the data set in Table 1, but for completeness the full details of this model are described here.

The results obtained with the above FEP-derived model (model 6 of Table 2) are found to be in excellent agreement with experimental values (Figure 2). We obtain an rms deviation of 0.84 with an α of 0.181. The result of adding a constant γ to the new model, yielding model 7, was also investigated. Remarkably, the optimal value for such a γ is −0.03 kcal/mol, i.e., virtually zero. We thus find, also for this model, that no additional constant term need be employed for accurate binding free energy estimation. It is interesting to note that the result with model 6 also supports the notion (see above) that the electrostatic response is similar in water and in proteins [38–41]. We find it very encouraging that a model with only one free parameter can reproduce experimental binding data for 18 rather different complexes with an average absolute error of as little as 0.58 kcal/mol.

Discussion

We have examined extensions of our original linear response method [22] for estimating binding free energies. The accuracy of the improved LIE approximation, defined by model 6, seems to compare favourably with other current methods for estimating binding energies. Indeed, the rms deviations in the new model are on the same level as the experimental uncertainties in the calibration set. Therefore, further significant improvements would require larger data sets.

Although model 4 and model 6 exhibit similar rms deviations, we strongly favour the latter, because model 6 has only one free parameter, α. Furthermore, model 6 is derived from physical, rather than statistical considerations, since the results of earlier FEP calculations are incorporated into the β coefficients for different ligand classes [28]. There is also a technical advantage in employing the same coefficients in the bound and free simulations, as in model 6, since errors in the parameterization of the ligand can then be expected to give rise to cancelling errors in the energy expression. However, the different ways of assigning electrostatic scaling parameters in the two models (as β_{prot} and β_{wat}, or as β_{FEP}) must of course have similar net effects, since they give comparable rms deviations for the same data, and the average values of β are similar. For the present data set, model 4 and model 6 both attribute some additional negative free energy to ligands with a net charge compared to uncharged ligands. In model 6 this is done on purpose, to reproduce known differences in electrostatic response properties. In model 4, the fact that ligands with a net charge generally have large absolute values of the electrostatic ligand–surrounding interaction energy averages, makes the effect of the slight difference in β coefficients more pronounced for such compounds. The cross-validated figures of merit in Table 2 support our preference for model 6. Especially, the leave-one-out cross-validated standard deviation $s^{\text{LOO}}_{\text{PRESS}}$ is only

0.94 kcal/mol for model 6, while it is 1.08 kcal/mol for model 4.

The use of the number of hydroxyl groups in a neutral ligand to find the proper value of β is based on our studies of how aqueous solutions respond to the presence of certain small molecules [28]. 'The number of hydroxyl groups' may seem like a somewhat *ad hoc* indicator variable that could hide some modelling error. However, referring to Ref. 28 for details, it appears to be a genuine property of hydroxyl groups, that they more easily than other neutral polar groups can interact with the hydrogen bonding network of an aqueous environment to lower the value of β.

It is interesting to note that the above results (models 6 and 7) suggest that no constant term seems to be required in the present type of scheme for good representation of binding energies. This may seem to be at variance with some recent proposals in the literature. For example, Janin has recently argued that a large contribution from loss of overall translational and rotational entropy accompanies most binding processes [43, 44]. His estimates of the translational contribution are, however, based on the Sackur–Tetrode equation, the applicability of which may be questioned in the context of molecular association in solution [45]. Murphy et al. have, in fact, argued that the configurational volumes associated with translational motion are not appreciably different for a ligand molecule in a complex compared to the ligand in solution. Their argument is based on estimates from crystallographic B-factors and mean free paths. Murphy et al. instead suggest that the so called cratic entropy is best described by the simple entropy of mixing [45]. This leads to an estimated entropy decrease of 8 cal $\mathrm{mol}^{-1}\,\mathrm{K}^{-1}$ (2.4 kcal/mol at 300 K) for bimolecular association. This contribution originates from reduction of the number of particles in solution. In the empirical scoring function of Böhm, a constant free energy term of $+1.3$ kcal/mol is included and designed to represent overall translational and rotational entropy, although it is cautioned that its physical interpretation is not obvious [7].

In the context of overall translational and rotational entropy contributions, the fact that release of 'bound' water molecules often accompanies the binding process is sometimes overlooked. Such release of water molecules would affect both mixing entropy and pure translational and rotational contributions. In fact, Dunitz has estimated from thermodynamic data on crystalline salts, that each released water molecule can contribute up to 2 kcal/mol, where this limit represents a protein site where the water molecules are as rigidly bound as in a typical salt crystal [46]. It is our impression that the somewhat crude nature of current cratic entropy estimates overshadows their usefulness in actual binding energy predictions. Another issue that merits further study in this context is the accuracy in treating internal entropy changes deriving from, e.g., torsional degrees of freedom.

While the interpretation of the polar part of Equations (1), (3) and (5) is rather straightforward, namely that it represents the free energy arising from electrostatic interactions between the ligand and its surroundings, the nonpolar term ($\Delta G^{\mathrm{nonpolar}} = \alpha\{\Delta\langle V_{l-s}^{\mathrm{vdw}}\rangle\}$) is less obviously interpreted. Clearly, it is not the actual 'van der Waals energy' itself that is the physical origin of this free energy contribution. Instead, we find that $\langle V_{l-s}^{\mathrm{vdw}}\rangle$ measures two very important quantities, namely (1) the *size of the solute* and (2) the heavy atom *number density* (ρ) of the surroundings. The latter quantity may at first seem rather irrelevant to the binding energetics. However, since water and protein have significantly different number densities ($\rho_{\mathrm{wat}} = 0.033$ Å^{-3} and $\rho_{\mathrm{prot}} \simeq 0.058$ Å^{-3}) [47], the quantity $\langle V_{l-s}^{\mathrm{vdw}}\rangle$ measures not only packing effects, but also the fraction of protein around the ligand. This fraction can be expected to correlate with hydrophobicity, since proteins are on average more hydrophobic than water. Hydrophobicity combined with ligand size, in turn, determines the magnitude of the hydrophobic effect, which probably constitutes the major part of the $\Delta G^{\mathrm{nonpolar}}$ term.

For a given solvent (surrounding) number density one finds that $\langle V_{l-s}^{\mathrm{vdw}}\rangle$ depends linearly on measures of solute size, such as molecular surface area (see [22, 32]). It is also well known that solvation energies for nonpolar compounds depend linearly on, e.g., surface area. Values for the coefficients relating $\langle V_{l-s}^{\mathrm{vdw}}\rangle$ to ΔG_{sol} for nonpolar solutes in different solvents can then, in principle, be estimated by comparison of these quantities. For instance, the data presented by Abraham [42] (also reviewed in Ref. 31) show that $\Delta G_{\mathrm{sol}}^{\mathrm{water}}$ (kcal/mol) $\simeq -0.014 A_s$ (Å^2) + const. for a representative set of molecules comprising noble gases, n-alkanes, and cyclic and branched hydrocarbons. Utilizing the relationship between $\langle V_{l-s}^{\mathrm{vdw}}\rangle$ and molecular surface area in Figure 2 of Ref. 22 would then yield $\Delta G_{\mathrm{sol}}^{\mathrm{water}} \simeq 0.09\langle V_{l-s}^{\mathrm{vdw}}\rangle + \mathrm{const.}$ (with a new constant). It is worth noting that this value of 'α_{wat}' is not so different from the α of, for example, model 6. A similar estimation of an α for nonpolar solvation in a protein

environment is, of course, more difficult since accurate solvation energies are not available. Even so, it may be of some interest to consider a less polar solvent than water to get an idea of the possible range of α. Again utilizing the solvation data of Ref. 42, we find from simulations of a few n-alkanes in acetone that the value of α for acetone solvation of nonpolar molecules is around 0.33 (i.e., $\Delta G_{\mathrm{sol}}^{acetone} \simeq 0.33\langle V_{l-s}^{\mathrm{vdw}}\rangle + \mathrm{const.}$). Although acetone is not a good protein model (it has, for example, roughly the same heavy atom number density as water), it still seems suggestive that our $\alpha = 0.18$ of model 6 lies between these two estimates.

The main reason why $\langle V_{l-s}^{\mathrm{vdw}}\rangle$ is a good measure of the number density ρ, is that the van der Waals potential well depths are roughly the same for typically relevant pairs of atom types (C, N, O) in the force field used [22, 33, 34]. This might have some implications for the transferability of our approach between different force fields. It has already been pointed out that the actual value of α is bound to be force field dependent [22, 23, 24, 27]. This is also likely to be the case for the β values [28, 32]. While we believe the present approach to have a sound physical basis, one should remember that calibration using a certain force field will yield a parameterization implicitly containing any errors present in this force field or in the simulation strategy (related, e.g., to the treatment of long range electrostatics [23, 27]). A possible example of this could be that the parameterizations of our models all include the GROMOS extended-atom phenyl rings, that are known to be too hydrophobic (J. Marelius, unpublished; see also Ref. 48). This does not necessarily lead to corresponding errors in the binding energies, precisely because these errors have been included in the parameterization. However, obvious imbalances or inconsistencies in the force field are likely to influence binding free energy values.

Conclusions

In summary, it would be valuable, both in basic biochemical research and in pharmaceutically oriented ligand design applications, to have a reliable method for evaluating absolute binding free energies for ligands to macromolecular receptors of known structure. Thermal conformational sampling methods together with linear response approximations show some promise in this respect and seem to be capable of yielding quantitative results in molecular systems of different nature. The accuracy of the reparameterized LIE method (model 6) matches free energy perturbation methods for calculating free energies of binding, with less computational effort, and without the restriction of only calculating relative binding affinities. The accuracy also compares favourably to that of current empirical scoring functions [1–10], but with considerably fewer parameters. The main disadvantage with the present method remains the computational cost of conformational sampling. It seems, however, that in most cases accurate results can be obtained with a reasonable computational effort.

Acknowledgements

Support from the Swedish Research Council for Engineering Sciences (TFR), from the Swedish Natural Science Research Council (NFR) and from the Swedish Foundation for Strategic Research (SSF) is gratefully acknowledged.

References

1. Goodford, P.J., J. Med. Chem., 28 (1985) 849.
2. Tomioka, N., Itai, A. and Iitaka, Y., J. Comput.-Aided Mol. Design, 1 (1987) 197.
3. Novotny, J., Bruccoleri, R.E. and Saul, F.A., Biochemistry, 28 (1989) 4735.
4. Meng, E.C., Shoichet, B.K. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 505.
5. Krystek, S., Stouch, T. and Novotny, J., J. Mol. Biol., 234 (1993) 661.
6. Rotstein, S.H. and Murcko, M.A., J. Med. Chem., 36 (1993) 1700.
7. Böhm, H.-J., J. Comput.-Aided Mol. Design, 8 (1994) 243.
8. Wallqvist, A., Jernigan, R.L. and Covell, D.G., Protein Sci., 4 (1995) 1881.
9. Verkhivker, G., Appelt, K., Freer, S.T. and Villafranca, J.E., Protein Eng., 8 (1995) 677.
10. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Greene, S. and Marshall, G.R., J. Am. Chem. Soc., 118 (1996) 3959.
11. Vajda, S., Weng, Z., Rosenfeld, R. and DeLisi, C., Biochemistry, 33 (1994) 13977.
12. Kurinov, I.V. and Harrison, R.W., Nat. Struct. Biol., 1 (1994) 735.
13. Holloway, M.K., Wai, J.M., Halgren, T.A., Fitzgerald, P.M.D., Vacca, J.P., Dorsey, B.D., Levin, R.B., Thompson, W.J., Chen, L.J., deSolms, S.J., Gaffin, N., Ghosh, A.K., Giuliani, E.A., Graham, S.L., Guare, J.P., Hungate, R.W., Lyle, T.A., Sanders, W.M., Tucker, T.J., Wiggins, M., Wiscount, C.M., Woltersdorf, O.W., Young, S.D., Darke, P.L. and Zugay, J.A., J. Med. Chem., 38 (1995) 305.
14. Ortiz, A.R., Pisabarro, M.T., Gago, F. and Wade, R.C., J. Med. Chem., 38 (1995) 2681.
15. Viswanadhan, V.N., Reddy, M.R., Wlodawer, A., Varney, M.D. and Weinstein, J.N., J. Med. Chem., 39 (1996) 705.
16. Beveridge, D.L. and DiCapua, F.M., Annu. Rev. Biophys. Biophys. Chem., 18 (1989) 431.

17. Jorgensen, W.L., Acc. Chem. Res., 22 (1989) 184.
18. Straatsma, T.P. and McCammon, J.A., Annu. Rev. Phys. Chem., 43 (1992) 407.
19. Kollman, P., Chem. Rev., 93 (1993) 2395.
20. Lee, F.S., Chu, Z.-T., Bolger, M.B. and Warshel, A., Protein Eng., 5 (1992) 215.
21. Gerber, P.R., Mark, A.E. and van Gunsteren, W.F., J. Comput.-Aided Mol. Design, 7 (1993) 305.
22. Åqvist, J., Medina, C. and Samuelsson, J.-E., Protein Eng., 7 (1994) 385.
23. Hansson, T. and Åqvist, J., Protein Eng., 8 (1995) 1137.
24. Hultén, J., Bonham, N.M., Nillroth, U., Hansson, T., Zuccarello, G., Bouzide, A., Åqvist, J., Classon, B., Danielson, H., Karlén, A., Kvarnström, I., Samuelsson, B. and Hallberg, A., J. Med. Chem., 40 (1997) 885.
25. Åqvist, J. and Mowbray, S.L., J. Biol. Chem., 270 (1995) 9978.
26. Paulsen, M.D. and Ornstein, R.L., Protein Eng., 9 (1995) 567.
27. Åqvist, J., J. Comput. Chem., 17 (1996) 1587.
28. Åqvist, J. and Hansson, T., J. Phys. Chem., 100 (1996) 9512.
29. Ben-Naim, A. and Marcus, Y., J. Chem. Phys., 81 (1984) 2016.
30. Ben-Naim, A. and Mazo, R.M., J. Phys. Chem., 97 (1993) 10829.
31. Blokzijl, W. and Engberts, J.B.F.N., Angew. Chem., Int. Ed. Engl., 32 (1993) 1545.
32. Carlson, H.A. and Jorgensen, W.L., J. Phys. Chem., 99 (1995) 10667.
33. Åqvist, J., Fothergill, M. and Warshel, A., J. Am. Chem. Soc., 115 (1993) 631.
34. van Gunsteren, W.F. and Berendsen, H.J.C., Groningen Molecular Simulation (GROMOS) Library Manual, Biomos B.V., Nijenborgh 16, Groningen, The Netherlands, 1987.
35. Jones-Hertzog, D.K. and Jorgensen, W.L., J. Med. Chem., 40 (1997) 1539.
36. Warshel, A. and Creighton, S., In van Gunsteren, W.F. and Weiner, P.K. (Eds.), Computer Simulation of Biomolecular Systems, Vol. 1, ESCOM, Leiden, The Netherlands, 1989, p. 120.
37. Wold, S., Quant. Struct.-Act. Relatsh., 10 (1991) 191.
38. Fersht, A.R., Leatherbarrow, R.J. and Wells, T.N.C., Biochemistry, 26 (1987) 6030.
39. Yadav, A., Jackson, R.M., Holbrook, J.J. and Warshel, A., J. Am. Chem. Soc., 113 (1991) 4800.
40. Warshel, A., Hwang, J.K. and Åqvist, J., Faraday Discuss., 93 (1992) 225.
41. Warshel, A., Schweins, T. and Fothergill, M., J. Am. Chem. Soc., 116 (1994) 8437.
42. Abraham, M.H., J. Am. Chem. Soc., 104 (1982) 2085.
43. Finkelstein, A.V. and Janin, J., Protein Eng., 3 (1989) 1.
44. Janin, J., Proteins, 24 (1996) i–ii.
45. Murphy, K.P., Xie, D., Thompson, K.S., Amzel, L.M. and Freire, E., Proteins, 18 (1994) 63.
46. Dunitz, J.D., Science, 264 (1994) 670.
47. Creighton, T. E., Proteins: Structures and Molecular Properties, Freeman, New York, NY, U.S.A., 1984.
48. Jorgensen, W.L. and Severance, D.L., J. Am. Chem. Soc., 112 (1990) 4768.