Journal of Computer-Aided Molecular Design **17:** 861–880, 2003.

Protein-ligand binding free energy estimation using molecular mechanics and continuum electrostatics. Application to HIV-1 protease inhibitors

V. Zoete[a], O. Michielin[b] & M. Karplus[a,c]

[a]*Laboratoire de Chimie Biophysique, Institut de Science et d'Ingénierie Supramoléculaires, Université Louis Pasteur, 8 allée Gaspard Monge, BP 70028 Strasbourg Cedex, France (Phone: +33 (0)3 90 24 51 23, Fax: +33 (0)3 90 24 51 24, e-mail: marci@isis-ulp.org);* [b]*Ludwig Institute for Cancer Research, Lausanne Branch, Chemin des Boveresses, 155, CH-1066 Epalinges, Switzerland;* [c]*Department of Chemistry and Biological Chemistry, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA (Phone: +1 617-495-4018, Fax: +1 617-496-3204, e-mail: marci@tammy.harvard.edu)*

Received 4 July 2003; accepted in revised form 12 December 2003

Key words: binding free energy estimation, continuum electrostatics, drug design, HIV-1 protease, LIE method, ligand binding

Summary

A method is proposed for the estimation of absolute binding free energy of interaction between proteins and ligands. Conformational sampling of the protein-ligand complex is performed by molecular dynamics (MD) *in vacuo* and the solvent effect is calculated *a posteriori* by solving the Poisson or the Poisson–Boltzmann equation for selected frames of the trajectory. The binding free energy is written as a linear combination of the buried surface upon complexation, SAS_{bur}, the electrostatic interaction energy between the ligand and the protein, E_{elec}, and the difference of the solvation free energies of the complex and the isolated ligand and protein, ΔG_{solv}. The method uses the buried surface upon complexation to account for the non-polar contribution to the binding free energy because it is less sensitive to the details of the structure than the van der Waals interaction energy. The parameters of the method are developed for a training set of 16 HIV-1 protease-inhibitor complexes of known 3D structure. A correlation coefficient of 0.91 was obtained with an unsigned mean error of 0.8 kcal/mol. When applied to a set of 25 HIV-1 protease-inhibitor complexes of unknown 3D structures, the method provides a satisfactory correlation between the calculated binding free energy and the experimental pIC_{50} without reparametrization.

Introduction

An approach to structure-based computer-aided ligand design consists of three steps [1, 2]. In the first, positions and orientations for the binding of a series of molecular fragments are determined in the known structure of the macromolecular target. A procedure such as the multiple copy simultaneous search method (MCSS) [3] or the solvation energy for exhaustive docking (SEED) method [4] can be employed for this purpose. In the second step, selected molecular fragments are connected to form putative ligands, using methods such as HOOK [5], DLD [6] or CCLD [7]. At this stage, hundreds to many thousands candidate ligands can be obtained. In the final step, these ligand candidates are examined for synthetic feasibility and their free energies of binding are estimated to determine which of them are likely to have the strongest affinities for the target. The binding free energy calculation can be time consuming because it requires an accurate determination of the interactions and a reliable treatment of solvent effects. Finding methods for evaluation of the ligand binding affinity that are fast enough to treat thousands of candidate ligands in a reasonable time is a major challenge in drug design.

A variety of methods have been proposed for estimating the binding affinity [8, 9]. They range from the very fast QSAR type approach [10–16] to the most theoretically correct free energy simulation [17–21]. They include first principle approaches [1,

22–25], regression-based scoring functions [26, 27], knowledge-based potentials [28–32] and the promising linear interaction energy approach (LIE) [33]. An efficient first principle approach is the molecular mechanics – Poisson–Boltzmann surface area (MM-PBSA) method that has been developed to describe the free energy of macromolecular systems [34]. In this approach, the binding free energy is calculated as the unweighted sum of gas-phase energies, solvation free energies and entropic contributions, averaged over a series of snapshots from MD trajectories. The gas phase terms are obtained by molecular mechanics calculations, the electrostatic contribution to the solvation term is calculated by solving the Poisson–Boltzmann equation [35] and the non polar contribution to the solvation free energy is assumed to be proportional to the solvent accessible surface area. This approach has been applied to estimate the binding energy for protein/protein [36, 37] or protein/ligand systems [38–40] and also to rank possible binding positions of small molecular fragments within a macromolecule binding site cleft [41].

In the present study, we introduce a method that is related most closely to the LIE as implemented by Åqvist et al. [33] and modified by Jorgensen and coworkers [42, 43], although the final formulation is quite different. The LIE approach makes use of linear response theory and requires simulations only for the free ligand in solution and for the bound ligand system. The free energy of binding is estimated from a linear combination of the differences of the average inhibitor/inhibitor-environment electrostatic and van der Waals interaction energies, using the following expression:

$$\Delta G_{bind} \cong \alpha(\langle V_{l-s}^{vdW}\rangle_{bound} - \langle V_{l-s}^{vdW}\rangle_{free}) + \beta(\langle V_{l-s}^{elec}\rangle_{bound} - \langle V_{l-s}^{elec}\rangle_{free}) \quad (1)$$

The subscript l designates the ligand and s the 'solvent'. It should be noted that the 'solvent' is the environment of the ligand; i.e., water in the free state and the water plus the protein in the bound state. In the first study, Åqvist et al. [33] fixed the value of β to 0.5 in accord with linear response theory and the value of α was fitted ($\alpha = 0.161$) to the calculated and experimental binding free energies of four endothiapepsin inhibitors. A mean unsigned error of 0.4 kcal/mol was obtained for this small set of inhibitors. We note that estimates of absolute binding free energies are calculated by this approach, in contrast to most applications of free energy simulations, which focus on the difference between binding free energies of two similar ligands.

When the method was applied to other systems, Hansson et al. [44] found that it was necessary to use different values for the β coefficient, between 0.5 and 0.33, depending on the charge and the chemical nature of the ligand. It was also shown that the value of the α coefficient could be adjusted to obtain more accurate predictions on other systems [44, 45]. In a related study, Wang et al. [46] found that the weight of the van der Waals interaction term can be expressed approximately as a function of the active site hydrophobicity for the systems studied; values in the range 0.11 to 0.87 were obtained. To reproduce the experimental binding free energies of a set of 9 thrombin inhibitors, Equation 1 was modified by Hertzog et al. [42] by adding a third term proportional to the buried solvent accessible surface area, $\gamma\langle\Delta SASA\rangle$; their fit led to coefficients $\alpha = \beta = 0.131$ defined as in Equation 1. However, as pointed out by the authors [42] and by Hansson et al. [44], the buried surface area and the $\langle V_{l-s}^{vdW}\rangle$ term are strongly correlated so that this modification is the same as the addition of a constant term. Also, charged ligands are found to require a special treatment [47]. It was suggested that the use of explicit water molecules and a finite spherical system for a charged ligand could introduce errors primarily for two reasons. First, the electrostatic interaction energy of the ligand with its environment does not include the contribution of the medium outside the sphere. Second, overpolarization of the solvent introduces an overestimation of the solvation free energy of the charge in the free state, relative to the bound state [47]. These problems were reduced by neutralizing the protein (so that the charge of the ligand environment is the same in the free and bound state) and by using the reaction field method to calculate the long-range electrostatic interactions [47].

Although the LIE method of Åqvist et al. is much faster than a full free energy simulation, it is still too slow for screening a large number of ligands. Recently, Zhou et al. [43] described a modified LIE-type method based on a continuum treatment of the solvent. Instead of using explicit solvent, the two molecular dynamics simulations required for Equation 1 are performed with the surface generalized Born model [48]. The terms taken into account in this LIE model lead to an expression of the form

$$\Delta G_{bind} \cong \alpha(\langle V_{l-s}^{vdW}\rangle_{bound} - \langle V_{l-s}^{vdW}\rangle_{free}) + \beta(\langle V_{l-s}^{elec}\rangle_{bound} - \langle V_{l-s}^{elec}\rangle_{free})$$

$$+ \gamma(\langle V^{cav}\rangle_{bound} - \langle V^{cav}\rangle_{free}) \qquad (2)$$

The first two terms are of the same form as in Equation 1, except that the electrostatic contribution includes the reaction field in addition to the unshielded Coulomb interaction energy. Since the simulations are performed without explicit solvent molecules, $\langle V_{l-s}^{vdW}\rangle_{free}$ is equal to zero. A term for the cavity energy difference between the bound and free ligand, V^{cav}, has been added. It is taken to be proportional to the solvent accessible surface area and implicitly includes the van der Waals energy between the solute and the solvent. The authors applied this method to two sets of 20 HIV-1RT ligands and 7 thrombin inhibitors; in both systems the compounds studied were structurally related. The weights of the terms in Equation 2 are dependent on the system; i.e., they use $\alpha = 0.178$, $\beta = 0.320$ and $\gamma = 1.74$ for the HIV-1RT ligands and $\alpha = -0.224$, $\beta = 0.0271$ and $\gamma = 11.6$ for the thrombin ligands. They found good results, with standard deviations of 1.0 and 0.9 kcal/mol, respectively, between the calculated and experimental binding free energies of the two systems. The precision is similar to that obtained with explicit solvent. However, the replacement of the explicit solvent by a continuum solvent model achieved a 12-fold increase in speed.

We have developed a simplified method for the estimation of absolute binding free energies inspired by, though different from, the LIE approach. The aim is the same as that of Zhou et al. [43]; i.e. ideally to obtain a method for estimating binding free energies that is fast, accurate and general. In our approach, the conformational sampling of the bound ligand system is performed *in vacuo* with restraints on the target. The solvation contribution to the binding energy is calculated *a posteriori* by solving the Poisson equation (or the Poisson-Boltzmann equation if the ionic strength is known) for selected frames of the trajectory. No conformational sampling of the free ligand is done. The van der Waals interaction energy term is replaced by the surface area buried upon complexation.

The parameters of the proposed approach were determined by applying it to a training set of 16 ligand/HIV-1 protease complexes of known 3D structure. Since the first structural characterizations of the protease [49, 50] about 150 liganded or unliganded HIV-1 protease structures have been published [51]. The large amount of structural data, combined with binding constants for many inhibitors, have made the HIV-1 protease a test case for structure assisted drug design and molecular modeling methods [52, 53]. The HIV-1 protease in this training set is the wild type enzyme (i.e. without any mutation). The resulting equation was then applied to a test set of 25 additional ligands developed by Enanta Pharmaceuticals. Since no crystal structures were available for the test set, the calculated binding free energies were derived from estimated binding modes. This test set was used because it corresponds to a true drug design process where a large number of structurally quite similar compounds are synthesized and tested but where little information on the 3D structure of the complexes is available.

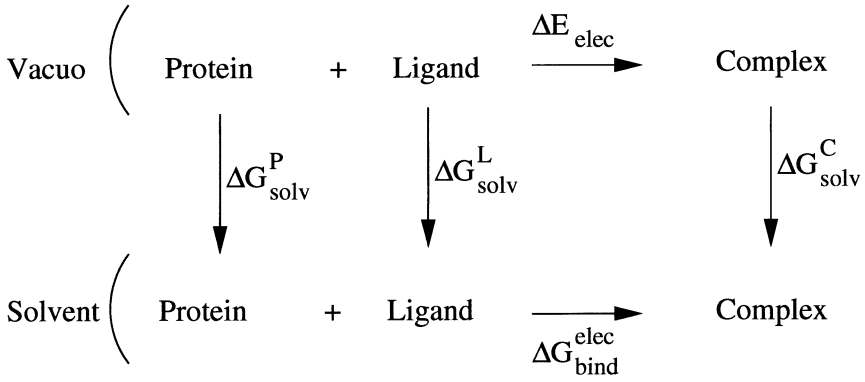
The next section presents the methodology, including the development of the approximate free energy equation and the details of the computations. Subsequently results are described.

Methodology

Formulation of the free energy equation

The method described here is based on the Linear Interaction Energy approach first developed by Åqvist et al. [33]. Several modifications have been introduced to simplify the procedure and to reduce the computation time. This is important if one wishes to employ such a method for screening a large number of compounds. First, the conformational sampling by molecular dynamics is performed *in vacuo*, rather than in explicit solvent. To prevent nonphysical conformational modifications, harmonic restraints are applied to the protein, and a distance dependent dielectric constant is employed for the interactions. No restraints are applied to the ligand.

In the original LIE method (Equation 1), the total electrostatic contribution to the binding free energy is estimated by calculating the difference in the electrostatic interaction energies between the ligand and its surroundings (solvent and protein); that is the quantity calculated is $\langle V_{l-s}^{elec}\rangle_{bound} - \langle V_{l-s}^{elec}\rangle_{free}$, where the brackets indicate simulation averages. This estimate is justified by the Linear Response Approximation (LRA) [33, 54, 55], in which the total electrostatic solvation energy is equal to half of this difference, though in the actual applications the factor is used as a parameter, as described in the Introduction. Thus, the LIE approach treats both the electrostatic solvation energy and the electrostatic interaction energy in a single term. As already mentioned in the Introduction, Zhou et al. [43] described a modified LIE method.



$$\Delta G^{elec}_{bind} = \Delta E_{elec} + \Delta G^{C}_{solv} - (\Delta G^{P}_{solv} + \Delta G^{L}_{solv})$$

$$\Delta G_{solv} = \Delta G^{C}_{solv} - (\Delta G^{P}_{solv} + \Delta G^{L}_{solv})$$

Figure 1. Definition of the electrostatic solvation free energy contribution, ΔG_{solv}, in Equation 3. ΔG^{elec}_{bind} is the electrostatic contribution to the binding free energy. ΔG^{C}_{solv}, ΔG^{P}_{solv} and ΔG^{C}_{solv} are respectively the electrostatic solvation free energies of the complex, the protein and the ligand. ΔE_{elec} is the Coulomb electrostatic interaction between the ligand and the protein.

The authors still described the electrostatic contribution to the binding free energy with the same LRA term $\langle V^{elec}_{l-s} \rangle_{bound} - \langle V^{elec}_{l-s} \rangle_{free}$, but now calculated it with the generalized Born model, rather than in the presence of explicit solvent.

In the approach described here, the effect of the solvent on the binding free energy is determined by use of a Poisson continuum model. Formally, we can write the binding free energy as:

$$\Delta G_{bind} = \Delta E_{vdW} + \Delta E_{elec} + \Delta G_{solv} + \Delta G^{complex}_{np} + \Delta G_{vib,conf} \quad (3)$$

where ΔE_{vdW} and ΔE_{elec} are, respectively, the van der Waals and electrostatic Coulomb interaction energies between the ligand and the protein. ΔG_{solv} is the electrostatic solvation free energy contribution to the binding free energy. It is equal to the difference of the solvation free energies of the complex and the isolated ligand and protein, $\Delta G^{C}_{solv} - (\Delta G^{P}_{solv} + \Delta G^{L}_{solv})$. As shown in the thermodynamic cycle of Figure 1, $\Delta E_{elec} + \Delta G_{solv}$ gives the total electrostatic contribution to the binding free energy. $\Delta G^{complex}_{np}$ is the nonpolar solvent (hydrophobic) contribution to the free energy of binding; i.e. we separate the van der Waals interaction energy and the solvation term. In the present method, as in the LIE approach, we modify Equation 3 by weighting the different contributions with adjustable parameters. Moreover, the values of these energy terms are averaged over a molecular dynamics simulation.

In contrast to the original LIE method (Equation 1), we separate the influence of the electrostatic solvation energy, ΔG_{solv}, and the electrostatic interaction energy, ΔE_{elec}, between the ligand and the protein. These terms, referred to as $\langle \Delta G_{solv} \rangle$ and $\langle E_{elec} \rangle$, are obtained by averaging ΔG_{solv} and ΔE_{elec} over a set of structures obtained from the MD trajectory performed for the protein-ligand complex. Thus, no MD simulation for the ligand alone is required. This assumes that the MD of the complex samples the appropriate ligand conformations, which is an approximation. Further, as already mentioned, the MD is done *in vacuo*. The solvent effect, $\langle \Delta G_{solv} \rangle$, is calculated *a posteriori*, for selected frames of the trajectory, by solving the finite difference approximation to the Poisson equation [56]; if the ionic strength of the solution is known, the Poisson–Boltzmann equation can be used. The use of an explicit solvation energy term and a Coulomb interaction term is expected to provide a better description of the role of each of these terms. In first-principle-based approaches for estimating binding free energies [8, 22, 41], these two terms are generally treated separately. This is important because they make contributions of opposite sign [1, 22, 41] and largely cancel.

$\Delta G_{np}^{complex}$ is assumed to be proportional to the buried solvent accessible surface upon complexation (SAS_{bur}); i.e., $\Delta G_{np}^{complex} = \sigma \times SAS_{bur}$, where $SAS_{bur} = SAS_{lig} + SAS_{prot} - SAS_{comp}$, where SAS_{lig}, SAS_{prot} and SAS_{comp} are, respectively, the solvent accessible surfaces of the isolated ligand, the system composed of the protein and the active site water molecule, and the complex.

The van der Waals and Coulomb interaction energies between the ligand and the protein/active site water molecule, as well as the buried surface upon complexation, are calculated for each frame of the MD trajectory using the CHARMM program. Since the van der Waals interaction energy and the buried surface are correlated [42, 44], only one of these two terms is used in the model equation, and a constant term, k, is added [44, 57].

In the original LIE approach and the present model, no terms to account for other types of contributions, such as the change in entropy on binding, are included. However, such effects can be included implicitly, at least in part, by the calibration of the method (i.e. adjustment of prefactors) and the use of a constant term, both evaluated by use of a training set. This calibration also accounts for errors in the force field. Based on Equation 3 and the above discussion, we write the binding free energy as

$$\Delta G_{bind} = k + \alpha\langle E_{vdW}\rangle + \beta\langle E_{elec}\rangle + \gamma\langle \Delta G_{solv}\rangle \quad (4)$$

or

$$\Delta G_{bind} = k' + \alpha'\langle SAS_{bur}\rangle + \beta'\langle E_{elec}\rangle + \gamma'\langle \Delta G_{solv}\rangle \quad (5)$$

depending on whether we use $\langle E_{vdW}\rangle$ or $\langle SAS_{bur}\rangle$ to account for the non-polar contribution. In Equations 4 and 5, the values of the constants k, k', α, α', β, β' and γ, γ' are determined by use of the training set.

To study the importance of the conformational sampling, the results obtained using this approach were compared to the ones calculated without any conformational sampling; i.e. with only one structure obtained from the crystal structure by minimization.

Computational method

All the molecular mechanics calculations reported in this work were performed with the CHARMM program [58], using the CHARMM22 [59] all-hydrogens force field, unless explicitly stated otherwise.

The training and test sets

Table 1 gives the name and chemical nature of the selected ligands, their binding free energies and the PDB code of the crystal structures of the complexes. Figure 2 gives the chemical structure of the ligands present in the training set. As can be seen, the ligands belong to several different chemical families (hydroxyethylene, difluoroketone, cyclic urea or sulfoamide, penicillin-derived, allophenylnorstatine, reduced amide and phosphinate). However, it should be noted that all the ligands are neutral when bound to the protein. Applying the equation described to charged ligands (U89360E and L739622) did not provide an accurate estimate of the binding free energy (data not shown), suggesting that charged ligands have to be addressed separately.

The test set of 25 ligands was developed by Enanta Pharmaceuticals, as mentioned in the Introduction. The chemical structures of these ligands are given in Figure 3. No experimental binding free energies are available for these ligands, but pIC_{50} were measured [60] and these are reported in Figure 3.

Topology and parameters for the ligand and molecular fragments

The partial atomic charges for the HIV-1 protease inhibitors were chosen to reproduce the quantum mechanical electrostatic potential at points surrounding the molecule, subject to the constraint that the sum of the charges should be equal to the net charge of the molecule. Each compound was minimized at the AM1 level and the minimized structure was used to calculate the Hartree Fock 6.31G* electrostatic potential. Atomic partial charges were obtained by fitting point charges to the electrostatic potential calculated with a geodesic point selection scheme [61], using the ELPOT and PDC modules of GAMESS [62]. Charges of symmetrically equivalent atoms were required to be equal.

The all-hydrogens parameters for van der Waals and bonded energy terms for the compounds were developed to be used in conjunction with the CHARMM22 all-hydrogens parameter set. Atomic radii and missing force field parameters were extrapolated from similar parameters taken from CHARMM22 or Quanta (Accelrys, San Diego, CA). The non standard parameters are available from the authors.

Table 1. Structure files, ligands names and chemical types of the ligands of the selected complexes.

Structure files[a]	Ligands	Prot.[b]	Chemical type	ΔG_{bind}^{exp} [c]	Ref.
1HVL	A76889	25′	C2 symmetry-based diol isostere	−14.16	[80]
1HVK	A76928	25	C2 symmetry-based diol isostere	−15.60	[80]
1HVI	A77003	25	C2 symmetry-based diol isostere	−15.54	[80]
1HVJ	A78791	25	C2 symmetry-based diol isostere	−16.22	[80]
1DIF	A79285	25	Difluoroketone isostere	−15.17	[81]
1OHR	AG1343	25	Hydroxyethylene isostere	−12.38	[82]
1AJX	AHA001	25′	Cyclic urea	−11.26	[74]
1AJV	AHA006	25′	Cyclic sulfoamide	−10.98	[74]
1HTF	GR126045	25′	Penicillin-derived	−9.82	[83]
1HPX	KNI272	25	Allophenylnorstatine analog	−16.02	[63]
1HSG	L735524 (Indinavir)	25′	Hydroxyethylene isostere	−13.21	[84, 85, 86]
2BPV	L738317	25′	Hydroxyethylene isostere	−10.53	[84]
1HBV	SB203238	25′	Reduced amide isostere	−9.06	[87]
1HOS	SB204144	25	Symmetric phosphinate	−12.17	[88]
1HPS	SB206343	25	Hydroxyethylene isostere	−13.12	[89]
1HPV	VX478	25′	Hydroxyethylene isostere	−13.12	[69]

[a]Code in the PDB.
[b]Protonation state used for the calculations with the corresponding complex. 25 means that ASP25 was protonated whereas ASP25′ was not. 25′ means that ASP25′ was protonated whereas ASP25 was not.
[c]Experimental binding free energy, expressed in kcal/mol.

Determination of the protease ASP25 protonation state

There is evidence that either ASP25 or ASP25′ can be protonated in a complex, depending on the inhibitor [63–66]. The protonation states of the ASP residues were determined using the coordinates in the PDB file and protonating either ASP25 or ASP25′. Both complexes were then minimized with a distance dependent dielectric constant, $\epsilon = 2r$, using the Adopted Basis Newton-Raphson (ABNR) minimization algorithm, until the root mean square (RMS) of the energy gradient reached a value of 0.01 kcal/mol Å. The protonation state of the complex having the lowest total potential energy was retained. This method gives results identical to those obtained in the work of Tawa et al. [67] for A77003, KNI272 and L735524; they determined the protonation states with semi-empirical quantum chemical (MNDO) calculations. The proposed protonation states used in the work for the training set are given in Table 1. As explained below, ASP25′ was always protonated in the case of the Enanta compounds complexes.

Docking of the test set compounds

All the molecules in the test set are derivatives of the Enanta Pharmaceuticals lead compound EP-000987 (see Figure 3). We calculated a reliable position for EP-000987, which served as the docking template for the other compounds of the test set; i.e. the coordinates of EP-000987 were used as a starting point for the minimization of the other complexes.

Given the structural similarity between EP-000987 and Bila2450 (Figure 4), the experimental binding mode of the latter was used as a starting point for docking the former. EP-000987 was designed as a HIV-1 protease inhibitor. However the inhibitor Bila2450 was only crystallized in complex with the HIV-2 protease inhibitor (1IDB in PDB [68]). Thus it was first necessary to determine the binding mode of Bila2450 in the wild-type HIV-1 protease. Due to the structural similarity between the central parts of EP-000987, Bila2450 and VX478, the experimental 3D structure of the latter, in complex with the wild-type HIV-1 protease (1HPV in PDB [69]), was used as a template. First, the experimental structures of the Bila2450/HIV-2 protease and VX478/HIV-1 protease complexes were superposed using the backbone atoms of the protein chains. The RMSD between the two protease backbones is 1.1 Å. The inhibitor VX478, as well as the HIV-2 protease and the corresponding active site water molecule were then removed. The resulting structure corresponds to the estimated binding mode of Bila2450 in the wild-type HIV-1 protease, including the corresponding active site water molecule.

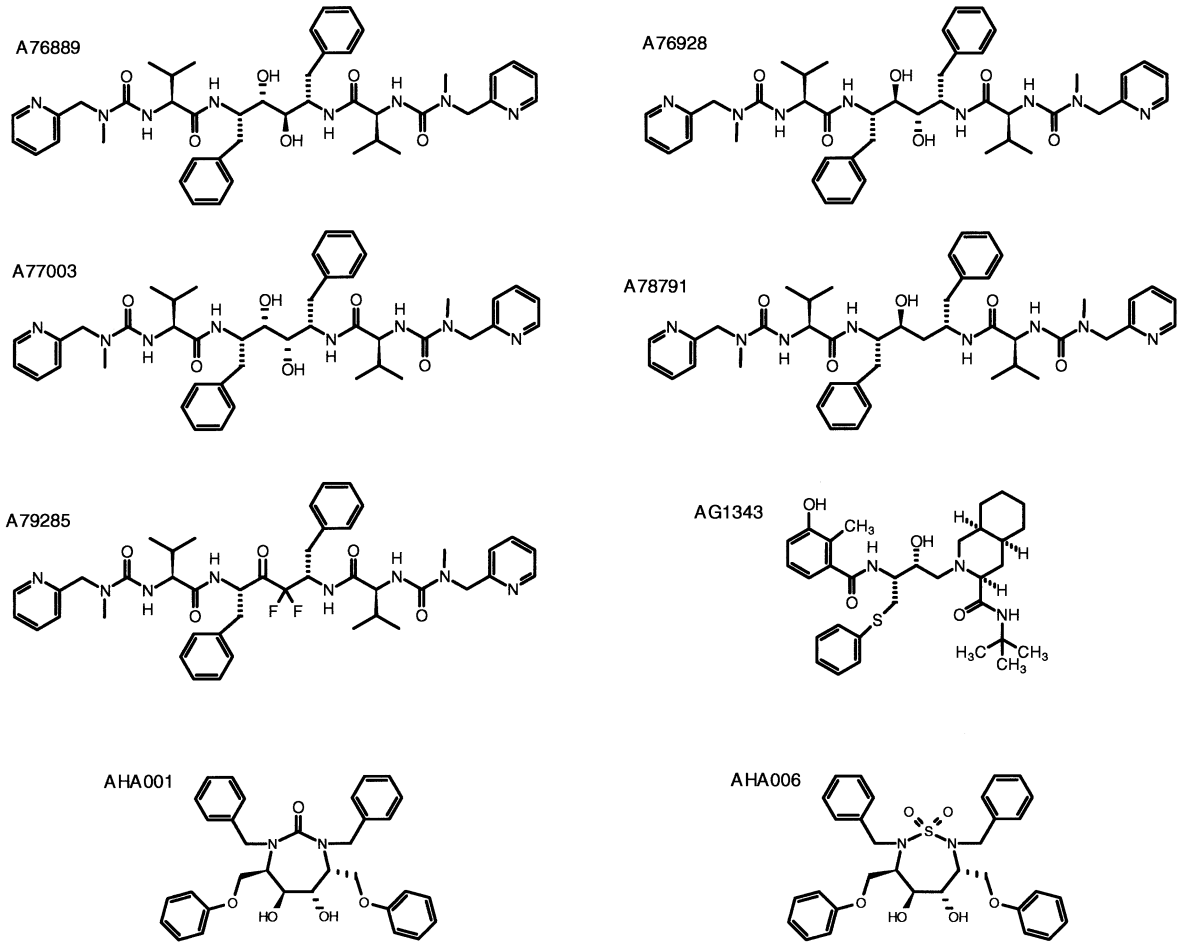


Figure 2. Ligands of the training set.

The atomic coordinates of Bila2450 in complex with the HIV-1 protease were then used as a template to dock EP-000987 into the enzyme active site. Starting from this position, EP-000987 was fully minimized in the fixed protease using 200 steps of Steepest Descent (SD) followed by 2000 steps of Adopted Basis Newton Raphson minimization, with a distance dependent dielectric constant, $\epsilon = 2r$. A MD run was then made at 500 K for 100 ps to sample the conformation space of the ligand, with $\epsilon = 2r$. The protease was kept fixed during the MD run and a 50 kcal/mol.Å^2 harmonic constant was applied to the oxygen atom of the active site water molecule. Fifty frames were extracted from the trajectory (one every 2 ps) and fully minimized. The 50 structures were very similar; the mean RMSD between the docked positions is only 0.25 Å. The left-hand side of the ligand showed very little variation; the most mobile region of the ligand is the right-hand side t-butylbenzamide.

The coordinates of the calculated docking position of EP-000987 served as the starting coordinates for the corresponding atoms of the other EP molecules. Each of the ligands was first minimized with 500 steps of ABNR using a distance dependent dielectric constant, $\epsilon = 4r$, in the fixed protease. Then, the complex was submitted to conformational sampling and post-processing was done as described below. For all the EP ligands, ASP25′ was protonated: the OH group of the ASP25′ carboxylic acid function donates a hydrogen bond to the central hydroxyl (OH) function of the EP compounds. The active site water molecule was always present.

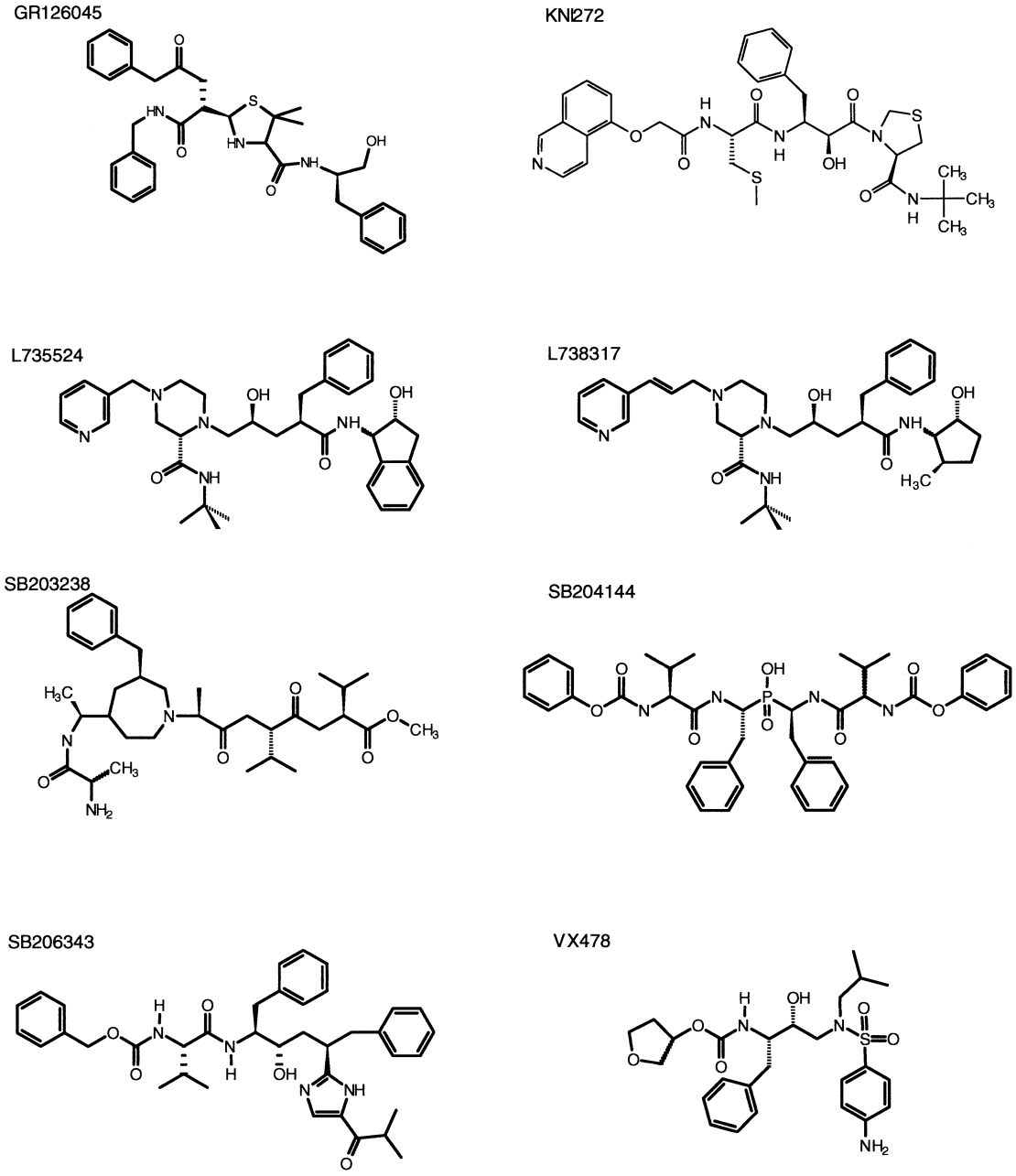
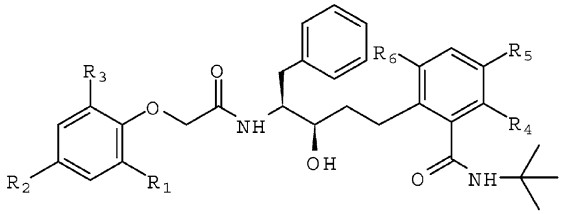


Figure 2. (Continued).

Minimization and conformational sampling

Conformation sampling of the complexes between the ligands and the HIV-1 protease was performed by using a molecular dynamics run. To prevent nonphysical conformational change arising from the absence of explicit solvent or the use of a distance dependent dielectric, harmonic restraints were applied to the HIV-1 protease. Harmonic restraints of 10, 15 and 30 kcal/mol.Å^2 were applied to the protein atoms that were less than 6 Å, between 6 and 11 Å, or more than 11 Å from any ligand atom, respectively. The active site water molecule, when present, was restrained by a 5 kcal/mol.Å^2 constraint. No constraints were applied to the ligand atoms. The length of the bonds involving a hydrogen atom were held fixed using the SHAKE algorithm [70].



Name	R_1	R_2	R_3	R_4	R_5	R_6	pIC_{50}
EP-000244							6.54
EP-000987	$-CH_3$		$-CH_3$				8.48
EP-001186		$-CH_3$					6.05
EP-001214	$-F$		$-F$				7.31
EP-001218	$-CH_3$	$-Br$	$-CH_3$				7.53
EP-001224	$-Cl$	$-Cl$	$-Cl$				6.25
EP-001225	$-Br$	$-CH_3$	$-Br$				6.43
EP-001237	$-CH_3$		$-CH_3$		$-CH_3$		8.52
EP-001238	$-CH_3$		$-CH_3$			$-CH_3$	7.68
EP-001239	$-CH_3$		$-CH_3$	$-F$			8.48
EP-001249	$-CH_3$		$-CH_3$	$-CH_3$			8.10
EP-001294	$-CH_3$		$-CH_3$		$-OCH_2COOCH_3$		4.77

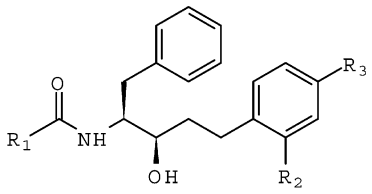
Figure 3. Ligands of the test set. IC_{50} are expressed in mol/l.

Starting from the crystal coordinates, the complex was initially minimized until the energy gradient reaches a value of 0.01 kcal/mol.Å. Then 50 ps of MD equilibration were performed at 300 K, using a distance dependent dielectric constant, $\epsilon = 4r$. Shifted electrostatic and switched van der Waals interactions, with a 12 Å cutoff, were used. A 1 fs step was used to integrate the equation of motion, and the coordinates were saved every 0.1 ps. The equilibration run was followed by a 100 ps production run, with the same conditions.

Post-processing of the MD trajectory

After the MD simulation was completed, the trajectory was processed to calculate the binding free energy following Equations 4 or 5; the van der Waals and electrostatic interaction energies between the ligand and its environment (HIV-1 protease and active site water molecule if present) were calculated for 1000 frames of the production run (every 0.1 ps). Shifted electrostatic and switched van der Waals interactions, with a 12 Å cutoff and $\epsilon = 1$, were used. The CHARMM van der Waals and electrostatic interaction energies values were averaged over the trajectory and the mean values were used in Equations 4 or 5. To calculate the buried surface upon complexation for each of the 1000 frames, the solvent accessible surfaces of the complex (SAS_{comp}), the isolated ligand (SAS_{lig}) and the system composed of the protein and the active site water molecule (SAS_{prot}) were calculated using the analytic algorithm in CHARMM and a probe radius of 1.4 Å. It also was averaged over the 1000 frames.

The electrostatic solvation energies were calculated by solving the linearized Poisson equation. For this purpose, the UHBD program [71] with an interface to CHARMM [72] was utilized in conjunction with the CHARMM all-hydrogen parameter set; i.e. the atom partial charges and radii from CHARMM22 were employed. The electrostatic solvation energy was calculated with a two-step procedure. A first calculation was made using a 3.0 Å grid spacing and a solvent layer (i.e. space between the structure and the closest



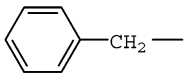
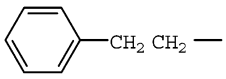
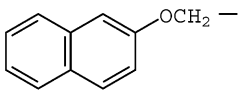
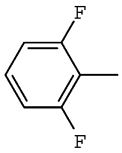
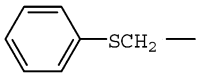
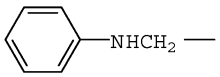
Name	R_1	R_2	R_3	pIC_{50}
EP-000158	tBuO—	—CONHiBu		7.01
EP-000373	EtO—	—CONHtBu		7.51
EP-000760		—CONHtBu		5.03
EP-000763		—CONHtBu		5.99
EP-000770	EtO—	—CONHiPent		6.05
EP-000771	EtNH—	—CONHtBu		6.03
EP-000774	tBuO—	—NHCOtBu		6.80
EP-000782		—CONHtBu		5.63
EP-000964	tBuO—	—CONHtBu	—OH	4.80
EP-000969		—CONHtBu		6.56
EP-001190		—CONHtBu		5.48
EP-001192		—CONHtBu		5.33
EP-001231		—CONHtBu		5.33

Figure 3. (Continued).

border of the grid) of about 20 Å around the complex. This provided the boundary potential for a second 'focusing' calculation. The latter utilized a grid spacing of 0.50 Å and a 5 Å borderspace. The protein dielectric constant was set to 1, and a dielectric constant of 80 was used for the exterior to represent an aqueous environment. A probe sphere radius of 1.4 Å was used. The ionic strength was set to 0. For calculation of the solvent-inaccessible volume in UHBD, the atomic surface was described by 3500 points. The electrostatic solvation free energy, ΔG_{solv}, was calculated for 50 frames (every 2 ps of the 100 ps production run) and

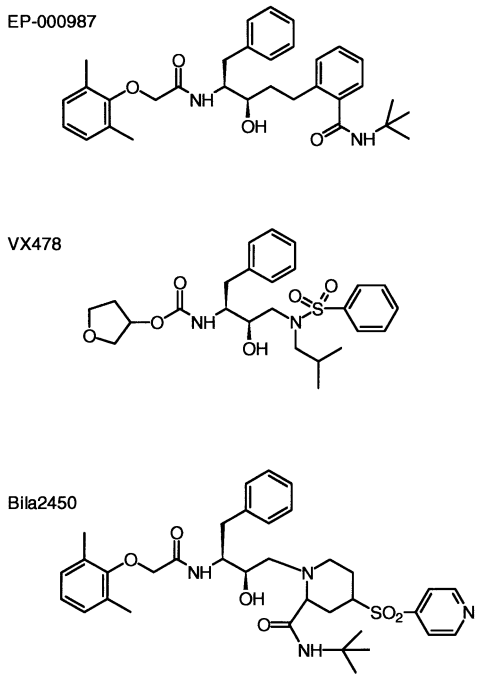


Figure 4. Structures of EP-000987, VX478 and Bila2450.

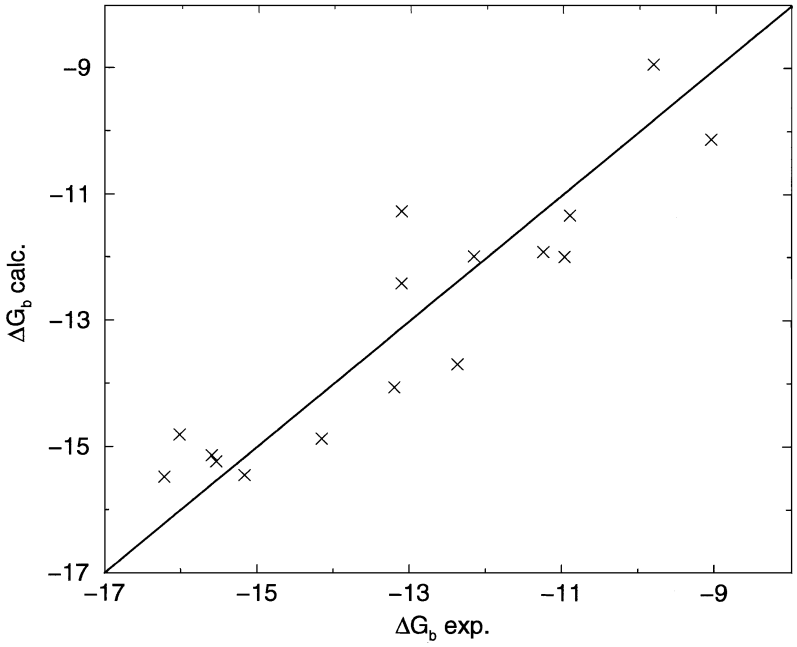


Figure 5. Calculated binding free energy using Equation 6 versus experimental binding free energy, for the 16 HIV-1 protease inhibitors of the training set. The line corresponds to the diagonal. The correlation between the estimated and experimental binding free energies is good if the points are close to the diagonal.

averaged. Only 50 frames were used because of the computational cost of the calculation. A comparison with the results obtained using 200 frames for several complexes showed that the use of 50 frames was satisfactory.

Results

As outlined in the Introduction, we compare the results obtained with conformational averaging to those obtained with a single minimized structure. The training set analysis is described first, followed by that for the test set.

Training set

With conformational sampling

Using the protocol described in the Computational Details, the following results were found for the training set of 16 HIV-1 protease inhibitors:

$$\begin{aligned}\Delta G_{bind} &= 0.45 + 0.0400 \times \langle E_{elec}\rangle \\ &\quad + 0.214 \times \langle \Delta G_{solv}\rangle \\ &\quad - 0.0237 \times \langle SAS_{bur}\rangle \end{aligned} \tag{6}$$

$$\mathrm{R} = 0.9128 \text{ and } \mathrm{Q} = 0.8403$$

$$\mathrm{RMSD} = 0.90,\ \langle|\text{error}|\rangle = 0.79 \text{ and range of errors} = 0.18 \text{ to } 1.85 \text{ kcal/mol}$$

and

$$\begin{aligned}\Delta G_{bind} &= -3.57 + 0.0065 \times \langle E_{elec}\rangle \\ &\quad + 0.089 \times \langle \Delta G_{solv}\rangle \\ &\quad + 0.254 \times \langle E_{vdW}\rangle \end{aligned} \tag{7}$$

$$\mathrm{R} = 0.8118 \text{ and } \mathrm{Q} = 0.6522$$

$$\mathrm{RMSD} = 1.42,\ \langle|\text{error}|\rangle = 0.95 \text{ and range of errors} = 0.08 \text{ to } 3.18 \text{ kcal/mol}$$

where R is the correlation coefficient, Q the leave-one-out method correlation coefficient, and RMSD and $\langle|\text{error}|\rangle$ represent, respectively, the root mean square deviation between the calculated and experimental binding free energy and the mean unsigned error. As can be seen, the mean unsigned error is only 0.79 kcal/mol for the 16 inhibitors with Equation 6, which is very encouraging (see Figure 5). This is to be compared with the difference in the binding free energies, for a given protein-ligand complex, measured in different laboratories, which can be as large as 1 kcal/mol. For Equation 6 the Student's t-test values are 1.9($p<0.03$), 4.7($p<0.001$) and 7.4($p<0.001$) for $\langle E_{elec}\rangle$, $\langle \Delta G_{solv}\rangle$ and $\langle SAS_{bur}\rangle$, respectively. For Equation 7 the Student's t-test values are 0.2, 1.8($p<0.05$) and 4.5($p<0.001$) for $\langle E_{elec}\rangle$, $\langle \Delta G_{solv}\rangle$ and $\langle E_{vdW}\rangle$, respectively. This indicates that all the energy terms are statistically significant in Equation 6, while $\langle E_{elec}\rangle$ makes no significant contribution in Equation 7. After removing $\langle E_{elec}\rangle$ from Equation 7, the rest of the equation and the statistical parameters are nearly unchanged. We also used the well known randomization test to validate Equation 6. The binding free energies of the 16 molecules were randomized and Equation 5 was fitted to the resulting activities without modifying the calculated energy terms ($\langle E_{elec}\rangle$, $\langle \Delta G_{solv}\rangle$ and $\langle SAS_{bur}\rangle$). The significance of Equation 6 would be decreased if it was possible to find a correlation between the real energy terms and randomized binding free energies. Figure 6 gives the Leave-one-out method correlation coefficient, Q, versus the correlation coefficient, R, for the correlation corresponding to the real binding free energies and 500 correlations corresponding to randomized binding free energies [73]. As can be seen, the real relation is significantly separated from those resulting from the randomization of the binding free energies. This confirms the significance of Equation 6. Table 2 reports the energy terms, buried solvent accessible surface and the calculated and experimental binding free energies for the 16 HIV-1 protease inhibitors.

As expected, the mean van der Waals interaction energy between the ligand and its environment, $\langle E_{vdW}\rangle$, is strongly correlated with the buried surface upon complexation $\langle SAS_{bur}\rangle$ ($\mathrm{R} = 0.90$). The value of the buried surface upon complexation is less sensitive to small conformational modifications, and thus shows less variation during a MD trajectory than the van der Waals interaction energy. The RMSD of E_{vdW} for a given trajectory is about 3 to 4.5% of $\langle E_{vdW}\rangle$, whereas the RMSD of SAS_{bur} is only 0.5 to 1.5% of $\langle SAS_{bur}\rangle$. Moreover, the use of $\langle SAS_{bur}\rangle$ provides better agreement ($\mathrm{R} = 0.8118$ when using $\langle E_{vdW}\rangle$ instead of $\mathrm{R} = 0.9128$ when using $\langle SAS_{bur}\rangle$) Therefore, Equation 6, instead of Equation 7, was used to estimate the binding free energies of the test set molecules.

In Equation 6, both $\langle E_{elec}\rangle$ and $\langle SAS_{bur}\rangle$ make favorable contributions to the binding free energy ($\langle E_{elec}\rangle$ is negative whereas $\langle SAS_{bur}\rangle$ is positive, see Table 2). By contrast, the electrostatic solvation energy, $\langle \Delta G_{solv}\rangle$, is always positive (see Table 2) and makes an unfavorable contribution to the binding free energy. Given the values of $\langle SAS_{bur}\rangle$ and $\langle \Delta G_{solv}\rangle$

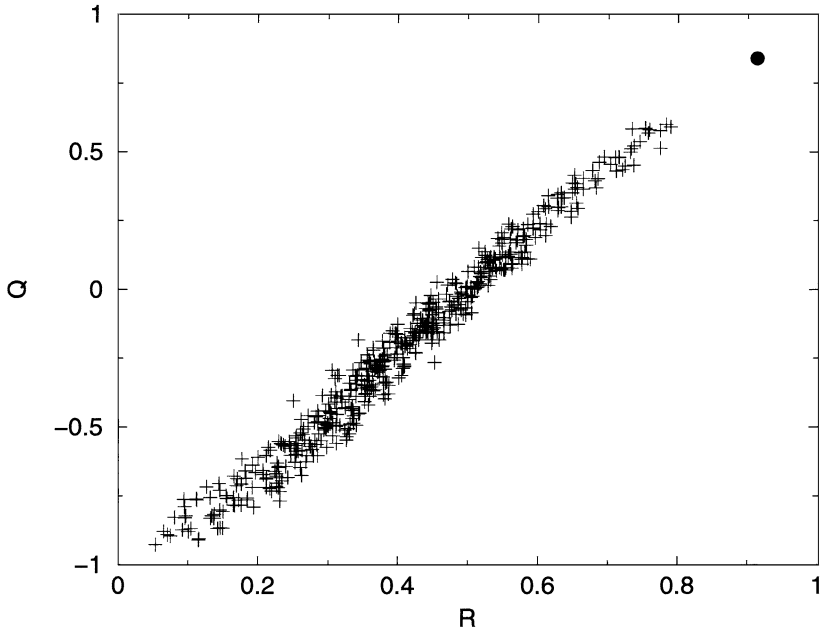


Figure 6. Randomization test for Equation 6. Leave-one-out method correlation coefficient, Q, versus the correlation coefficient, R, for the correlation corresponding to the real binding free energies (filled circle) and the 500 ones resulting from the randomized binding free energies (+).

Table 2. Interaction energies, solvation energies and buried surface upon complexation calculated for the training set compounds, as well as calculated and experimental binding free energies. All energies are in kcal/mol and surfaces in Å^2.

	Minimization only					Molecular dynamics					
Ligand	E_{vdW}	E_{elec}	ΔG_{solv}	SAS_{bur}	ΔG^{calc}_{bind}	$\langle E_{vdW}\rangle$	$\langle E_{elec}\rangle$	$\langle \Delta G_{solv}\rangle$	$\langle SAS_{bur}\rangle$	ΔG^{calc}_{bind}	ΔG^{exp}_{bind}
A76889	−79.03	−41.71	100.73	1498	−14.41	−72.38	−35.69	100.75	1498	−14.87	−14.16
A76928	−86.86	−46.07	100.85	1508	−14.78	−80.86	−32.70	100.72	1514	−15.14	−15.60
A77003	−86.55	−77.03	107.99	1512	−15.29	−81.87	−65.07	106.53	1516	−15.24	−15.54
A78791	−79.40	−46.32	99.67	1496	−14.72	−73.51	−39.38	99.50	1506	−15.47	−16.22
A79285	−88.66	−76.12	106.63	1517	−15.42	−81.85	−62.97	106.77	1531	−15.46	−15.17
AG1343	−68.62	−57.36	81.57	1247	−12.95	−65.04	−49.55	82.75	1262	−13.69	−12.38
AHA001	−60.63	−64.65	80.21	1154	−11.65	−56.13	−62.50	81.20	1151	−11.91	−11.26
AHA006	−75.10	−53.81	81.52	1203	−11.95	−69.18	−40.50	81.49	1194	−11.99	−10.98
GR126045	−55.35	−51.71	80.62	1083	−9.64	−50.66	−50.60	85.55	1085	−8.94	−9.82
KNI272	−75.29	−44.42	83.33	1332	−13.80	−70.31	−37.92	83.89	1339	−14.81	−16.02
L735524	−76.26	−62.13	94.18	1375	−13.84	−71.24	−53.95	93.81	1370	−14.06	−13.21
L738317	−69.10	−57.86	92.63	1361	−13.60	−66.56	−47.03	103.45	1354	−11.33	−10.91
SB203238	−57.84	−2.43	87.53	1255	−9.86	−57.31	−20.83	94.02	1262	−10.13	−9.06
SB204144	−80.27	−46.85	99.27	1450	−13.90	−77.44	−45.88	112.58	1466	−11.98	−12.17
SB206343	−72.43	−55.49	105.62	1363	−11.66	−72.54	−37.65	101.65	1399	−12.41	−13.12
VX478	−64.61	−54.97	79.15	1148	−11.26	−61.18	−48.43	81.23	1148	−11.27	−13.12

in Table 2 and the value of the corresponding coefficient in Equation 6, we see that these two terms make contributions to the variation of the binding free energy of the same magnitude but of opposite sign for the system under investigation. $\langle SAS_{bur}\rangle$ varies from 1085 to 1531 Å^2, and thus its contribution to ΔG_{bind} ranges from −25.7 to −36.3 kcal/mol, while $\langle \Delta G_{solv}\rangle$ varies from 81.20 to 112.58 kcal/mol

and its contribution to ΔG_{bind} ranges from +17.4 to +24.1 kcal/mol. $\langle E_{elec} \rangle$ ranges from −20.83 to −65.07 kcal/mol but because of its very small coefficient, the contribution to ΔG_{bind} varies only between −0.8 and −2.6 kcal/mol. $\langle E_{elec} \rangle$ thus has a secondary, but non-negligible role in determining ΔG_{bind}. If $\langle E_{elec} \rangle$ is omitted, the statistical parameters R, RMSD and $\langle |\text{error}| \rangle$ of the new correlation become, respectively, 0.8882, 1.08 and 0.88. It is interesting to note that the coefficient of $\langle \Delta G_{solv} \rangle$ in Equation 6 (i.e. 0.214) is of the same order as the one determined by Åqvist et al. for $(\langle V_{l-s}^{elec} \rangle_{bound} - \langle V_{l-s}^{elec} \rangle_{free})$ for different systems with explicit solvent (0.33 to 0.5); i.e. in a typical LIE application to estimate the binding free energies of mannitol derived HIV-1 protease inhibitors, $(\langle V_{l-s}^{elec} \rangle_{bound} - \langle V_{l-s}^{elec} \rangle_{free})$ ranges from +0.32 to −14.32 kcal/mol and its contribution to ΔG_{bind} ranges from +0.16 to −7.16 kcal/mol ($\beta = 0.5$) [74]. However, in the present study, the contributions of the electrostatic solvation free energy and the Coulomb interaction energy between the ligand and its environment are treated independently, whereas these contributions are combined in the LIE equation.

The conformational sampling makes possible an estimate of one contribution to the uncertainties in the calculated ΔG_{bind}. The standard deviation of $\langle E_{elec} \rangle$, $\langle \Delta G_{solv} \rangle$ and $\langle SAS_{bur} \rangle$, averaged over the 16 protein-ligand complexes are, respectively, 6 kcal/mol, 3 kcal/mol and 10 Å^2. Given the corresponding coefficients in Equation 6, the actual contributions of each of these energy terms have uncertainties of 0.2, 0.6 and 0.2 kcal/mol, respectively. Thus, based on this one contribution, the total uncertainty in the calculated ΔG_{bind} is about 1.0 kcal/mol, close to the RMSD of 0.9 kcal/mol observed between the calculated and experimental binding free energies for the 16 protein-ligand complexes. The largest contribution to the uncertainty comes from the electrostatic solvation energy.

Minimization alone

Because the PB free energy calculation is sensitive to the details of the structure, the MD search of the thermally allowed conformational space is expected to provide better estimates of the PB contribution to the binding energy [75]. To test the importance of the effect in the present study, the correlation found for Equation 6 and the experimental binding free energies was calculated without performing conformational sampling. The energy terms were calculated for a single structure obtained by minimization of the complex as described in the Computational Details.

The equation fitted to the data is

$$\begin{aligned} \Delta G_{bind} &= 2.11 + 0.0434 \times E_{elec} \\ &\quad + 0.145 \times \Delta G_{solv} \\ &\quad - 0.0196 \times SAS_{bur} \end{aligned} \tag{8}$$

$$\text{R} = 0.8109 \text{ and } \text{Q} = 0.6088$$

$$\text{RMSD} = 1.28, \langle |\text{error}| \rangle = 1.04 \text{ and range of errors} = 0.25 \text{ to } 2.69 \text{ kcal/mol}$$

and the individual results are also given in Table 2.

As can be seen, the results of the fit are noticeably worse than in Equation 6. Two compounds have calculated binding free energies in the minimization protocol with particularly large errors; they are SB204144 and L738317. The electrostatic solvation energies calculated for these two complexes are underestimated by about 14 kcal/mol and 11 kcal/mol, respectively, relative to the values calculated with conformational sampling. An important contribution arises from the fact that the positions of the left and right end phenyl of SB204144 and of the left end of L738317, in the minimized conformation, are about 2 Å more distant from ARG8 and ARG8′ than in the average structure from the MD run.

The terms E_{elec}, ΔG_{solv} and SAS_{bur} were also calculated, for each complex, for a single average structure from the 100 ps MD run. The average structures were relaxed by 200 steps of steepest descent minimization before the energy terms were calculated. The correlation was re-calculated and provided estimated binding free energies with a precision similar to that of Equation 6; i.e. R = 0.9183, Q = 0.8606, RMSD = 0.96, $\langle |\text{error}| \rangle$ = 0.77, and range of errors = 0.08 to 1.44 kcal/mol. This suggests, as has been pointed out earlier [76], that it is not the conformational averages per se, but rather the improved average structure, which is the essential result.

Docking of EP-000987

The largest difference between the docking modes of Bila2450 in the HIV-2 protease and EP-000987 in the HIV-1 protease involves the left-hand aryl group (see Figure 7). The left-hand phenyl of EP-000987 seems to be less deeply buried in the S2 pocket of HIV-1 protease than is the same cycle of Bila2450 in the S2 pocket of HIV-2 protease. Actually, the V32I mutation between the HIV-1 and HIV-2 proteases is

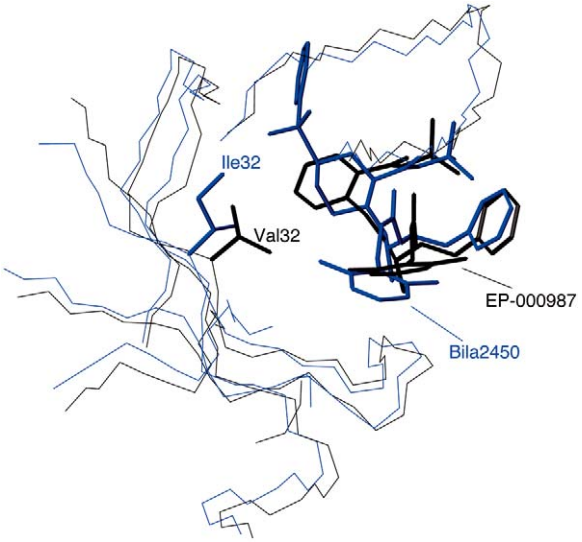


Figure 7. Superposed structures of the HIV-1 protease in complex with EP-000987 and HIV-2 protease in complex with Bila2450 (1IDB). Removal of the Ile32 side chain of HIV-2 protease from the active site cleft according to the Val32 side chain of HIV-1 protease. For clarity, monomer 2 of the protease is not shown. The HIV-2 protease/Bila2450 complex is in blue and the HIV-1 protease/EP-000987 complex is in black. Val32 and Ile32 side chains as well as EP-000987 and Bila2450 are in thick lines and monomers 1 are in thin lines.

characterized by the fact that the Ile32 side chain of HIV-2 protease is more buried in the protein core than the Val32 of HIV-1 protease (Figure 7). The removal is about 1.9 Å. Thus the S2 pocket of HIV-1 protease is actually shallower than the S2 pocket of HIV-2 protease, and the left-hand aryl group of EP-000987 is repelled towards the center of the active site cleft. Since the docking of the ligand was studied in the fixed protease, an additional minimization of the entire complex, without any constraint on the protease and the active site water molecule, was made to assess the possibility that Val32 could adapt to the perturbation provided by EP-000987. The Val32 residue did not show any large movement in the presence of EP-000987 during this minimization.

The RMSD between the two ligands is about 1.7 Å for the left-hand aryl group, 0.3 Å for the P1-phenyl ring, 0.8 Å for the t-butyl group and 0.8 Å for the hydroxyl function making H-bonds with the ASP25 and ASP25′. The two carbonyl oxygen atoms of EP-000987 are well placed to make the so-called H-bonded bridge with the active site water molecule and the two Ile50 and Ile50′ backbone NH atoms.

Test set

Equation 6 was applied to the test set of 25 EP-000987 derivatives provided by Enanta Pharmaceuticals. The average energy terms and the calculated binding free energy compared to the experimental pIC_{50} are given in Table 3. It should be noted that Equation 6 was applied without any refinement of the coefficients for the new set of compounds. Table 3 and Figure 8 show that there is a good correlation between the calculated binding free energy and the experimental activity. As can be seen in Figure 8, the deviation between the points and the correlation line is generally within the uncertainty of Equation 6. However, the correlation coefficient (0.64) is significantly lower than in the case of the training set. Moreover, the actual slope of the regression line (0.4) is lower than what is expected (1.3). Actually, since all the pIC_{50} were calculated under the same conditions, the ratios of two IC_{50} for two different ligands with the same enzyme should be equivalent to the K_i ratio, according to the Cheng–Prusoff equation [77]. This equation gives $K_i = IC_{50}/(1 + (D/K_d))$, where D is the concentration of the radioligand and K_d is the radioligand binding constant to the protein under the same experimental conditions. Thus, the slope of the ΔG_{bind} vs. pIC_{50} line is $RTln10 \sim 1.3$.

Different causes could be involved in this behavior. First, we should not forget that the binding free energies were calculated for estimated binding modes and not for experimentally determined ones. The correction in the docking position that can be expected from the thermodynamic sampling may be insufficient in some cases, and the method could lead to errors in the estimation of the affinity of badly placed ligands. Moreover, the range of calculated binding free energies of the given molecules is only about 3 kcal/mol. This value is significantly less than the 7 kcal/mol variation observed for the 16 compounds of the training set. In both cases the points are regularly distributed over the ΔG_{bind} and K_i or pIC_{50} ranges and the standard deviation of the calculated ΔG_{bind} is around 0.9 kcal/mol. Thus, it is not surprising that a smaller correlation coefficient is obtained for the test set. The correlation between the experimental pIC_{50} and the calculated ΔG_{bind} shows that this method can be fruitful in determining *a priori* the range of activity of different putative ligands once docked into the enzyme. Moreover, it provides an understanding of the influence of the different energy terms on the activity,

Table 3. Interaction energies, solvation energies and buried surface upon complexation calculated for the test set compounds, as well as calculated and experimental binding free energies. All energies are in kcal/mol and surfaces in Å^2.

Ligand	$\langle E_{vdW} \rangle$	$\langle E_{elec} \rangle$	$\langle \Delta G_{solv} \rangle$	$\langle Bur. \rangle$	ΔG_{bind}^{cal}	pIC_{50}
EP-001237	−67.15	−35.10	82.74	1240	−12.64	8.52
EP-000987	−63.25	−26.99	77.49	1196	−12.39	8.48
EP-001239	−67.32	−28.27	78.09	1216	−12.79	8.48
EP-001249	−68.54	−29.58	80.21	1228	−12.67	8.10
EP-001238	−67.75	−28.87	82.36	1242	−12.52	7.68
EP-001218	−68.48	−36.50	88.13	1300	−12.96	7.53
EP-000373	−57.48	−27.78	72.41	1111	−11.50	7.51
EP-001214	−63.56	−34.08	81.11	1201	−12.02	7.31
EP-000158	−56.91	−29.24	73.18	1099	−11.11	7.01
EP-000774	−57.70	−23.98	70.22	1100	−11.55	6.80
EP-000969	−61.63	−26.38	76.26	1136	−11.21	6.56
EP-000244	−63.70	−31.38	79.01	1196	−12.24	6.54
EP-001225	−68.57	−37.84	84.63	1246	−12.48	6.43
EP-001224	−68.47	−37.68	88.13	1246	−11.73	6.25
EP-000770	−54.88	−32.51	74.15	1137	−11.93	6.05
EP-001186	−64.33	−29.87	84.63	1221	−11.57	6.05
EP-000763	−64.12	−32.36	80.85	1208	−12.17	5.99
EP-000771	−54.78	−28.71	77.29	1100	−10.24	6.03
EP-000782	−59.37	−19.14	78.85	1163	−11.00	5.63
EP-001190	−65.20	−29.71	81.93	1208	−11.83	5.48
EP-001231	−61.95	−34.16	78.19	1177	−12.08	5.33
EP-001192	−67.50	−31.12	85.77	1225	−11.47	5.33
EP-000760	−61.81	−27.64	89.72	1239	−10.82	5.03
EP-000964	−57.69	−44.69	82.44	1124	−10.33	4.80
EP-001294	−70.76	−55.12	101.53	1336	−11.69	4.77

which could be useful in a QSAR or other improved design approach.

Time estimates

Table 4 gives the CPU requirement for explicit solvent LIE calculations and our approach. The CPU timing was obtained on a SGI R14000/500 Mhz Origin 3800 cluster. Since for the calculations used in this work we collected the data during 100 ps, following a 50 ps equilibration, we chose to carry out MD simulations with explicit solvent of the same length to estimate the time required by the standard LIE approach. MD simulations of the complex with explicit water molecules, for the LIE approach, were carried out using stochastic boundary conditions [17, 78]. Since the cutoff on the non-bonded interactions was 12 Å and the distance between ASP25′ OD1 and the farthest ligand atom was about 11 Å, a 25 Å water sphere centered on the ASP25′ OD1 was used. As can be seen, the MD simulations (50 ps equilibration followed by 100 ps data collection) with stochastic boundary conditions for the bound and free states necessary for the LIE approach, required 76.8 h. Our approach only needs an *in vacuo* MD simulation of the bound state, requiring 9.6 h, followed by a postprocessing requiring 5.4 more hours, for a total of 15 h. Therefore, the present approach is approximately 5 times faster than the standard LIE approach.

Concluding discussion

A method has been developed for the fast estimation of absolute binding free energies of protein-ligand complexes. The conformational sampling of the complex is performed by MD *in vacuo* and the solvent effect is calculated *a posteriori* for selected frames of the trajectory; no MD run of the ligand alone is necessary.

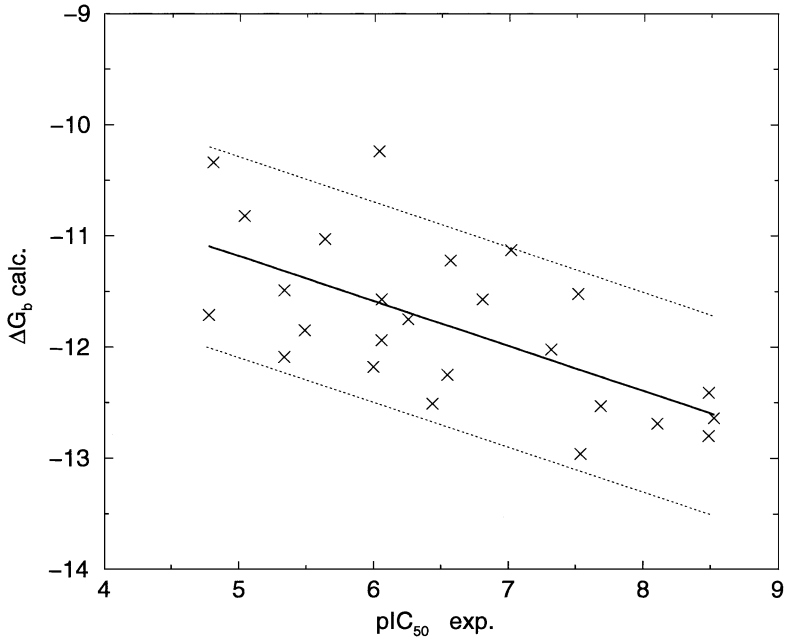


Figure 8. Calculated binding free energy using Equation 6 versus experimental pIC_{50}, for the 25 HIV-1 protease inhibitors of the test set. The thick line corresponds to the correlation and the dotted lines correspond to the correlation ± the RMSD seen in Equation 6; i.e. 0.9 kcal/mol.

The method has been shown to be applicable to ligands with very different chemical structures.

The method presented here uses terms that are included in the MM-PBSA approach; e.g. Poisson–Boltzmann electrostatics together with gas phase energy and surface area terms. Also, as in the MM-PBSA approach, these terms are averaged over a series of snapshots from MD trajectories. However, the new approach differs from the MM-PBSA approach since the energy terms are weighted and the weighting terms are fitted to experimental binding free energies. Also, there is no explicit entropic term in the new method; the entropic contribution to the binding is taken into account implicitly through the constant term and the weighting coefficients. Also, to our knowledge, whereas the MM-PBSA method gives satisfactory results for the ranking of structurally similar ligands [40] or for the ranking of putative binding modes of a given small molecule on a protein surface [41], it appears not to be sufficiently precise [79] to provide the refined estimates of the absolute binding free energies of interest in the present study.

Although conceptually related to the LIE method, the formulation introduced in the present study is significantly different. In the LIE method, the desolvation energy of the ligand is estimated by the difference in electrostatic interaction between the ligand and its environment for the free and bound state. Here, the electrostatic term is replaced by two terms treated independently: the electrostatic solvation energy term calculated by solving the Poisson equation, and a Coulomb interaction energy between the ligand and the enzyme in the bound state. The solvation energy term and the Coulomb interaction energy are, respectively, unfavorable and favorable to the binding affinity, the first term being the most important. Finally, the difference in van der Waals interaction energy between the

Table 4. CPU requirement comparison between the explicit solvent LIE approach and the present approach. Times are given in hours.

Model	Calculation	Total CPU
LIE	Unbound, equilibration (50 ps)	12.8H
	Unbound, data collection (100 ps)	25.6H
	Bound, equilibration (50 ps)	12.8H
	Bound, data collection (100 ps)	25.6H
New	Unbound, equilibration	0
	Unbound, data collection	0
	Bound, equilibration (50 ps)	3.2H
	Bound, data collection (100 ps)	6.4H
	Postprocessing[a]	5.4H

[a]Calculation of SAS_{bur}, E_{vdW} and E_{elec} for 1000 frames and ΔG_{solv} for 50 frames.

ligand and its environment for the free and bound state present in the LIE is replaced by the buried surface upon complexation. This last term provides a better correlation between the calculated and experimental binding free energies and shows less variation during the conformational sampling.

The method has been shown to give good results for a training set of 16 HIV-1 protease inhibitors for which the experimental 3D structure of the complex with the enzyme is known. A correlation coefficient of 0.91 was obtained, with a RMSD of 0.9 kcal/mol and a mean unsigned error of only 0.8 kcal/mol. When applied to a set of 25 HIV-1 protease inhibitors of unknown 3D structures, without any reparametrization, the method provides a satisfying correlation that could be used in order to estimate *a priori* the range of activity of putative ligands. The present approach could be efficient in a Drug Design process to discriminate between a large number of putative inhibitors for a given enzyme, given the fact that it would first be necessary to calibrate the equation coefficient to the system under investigation.

The binding free energy estimation is noticeably worse if no search of the thermally allowed conformational space is carried out. In this case, the results of the present method are similar to the approach of Holloway et al. [26], who estimated the pIC_{50} for a training set of 33 congeneric hydroxyethylene isosteres against the HIV-1 protease using a regression based score involving the total intermolecular interaction energies between the ligands and the protease, without conformational sampling. They used only the electrostatic and van der Waals interaction energies. They found a correlation coefficient of 0.89 and a cross-validated correlation coefficient of 0.87 for their training set between the calculated and experimental activities. These results are better than those obtained using Equation 8. However, it should be noted that the present study uses structurally unrelated compounds, while Holloway et al. used very similar compounds.

Acknowledgements

We are grateful to Dr. Roland Stote and Prof. Markus Meuwly for fruitful discussions. We thank Dr. Michael Schaeffer for providing the CHARMM-UHBD interface. V.Z. was supported in part by Enanta Pharmaceuticals Inc. The research at Harvard was supported in part by the National Institute of Health. Calculations were performed at the Centre Informatique National de l'Enseignement Supérieur (CINES, Montpellier, France) and the Centre d'Etude du Calcul Parallèle de Strasbourg (Illkirch, France). We thank Dr. Jens Eckstein and Dr. Steven Riesinger of Enanta Pharmaceuticals Inc. for providing the activities and chemical structures of the test set compounds.

References

1. Caflisch, A. and Karplus, M., Perspect. Drug Discov. Design, 3 (1995) 51.
2. Caflisch, A., Walchi, R. and Ehrhardt, C., News Physiol. Sci., 13 (1998) 182.
3. Miranker, A. and Karplus, M., Proteins: Struct. Funct. Genet., 11 (1991) 29.
4. Majeux, N., Scarsi, M., Apostokalis, J., Ehrhardt, C. and Caflisch, A., Proteins: Struct. Funct. Genet., 37 (1999) 88.
5. Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Proteins: Struct. Funct. Genet., 19 (1994) 199.
6. Miranker, A. and Karplus, M., Proteins: Struct. Funct. Genet., 23 (1995) 472.
7. Caflisch, A., J. Comput.-Aided Mol. Design, 10 (1996) 371.
8. Ajay and Murcko, M.A., J. Med. Chem., 38 (1995) 4953.
9. Gohlke, H. and Klebe, G., Curr. Opin. Struct. Biol., 11 (2001) 231.
10. Gupta, S.P., Babu, M.S. and Sowmya, S., Bioorg. Med. Chem., 6 (1998) 2185.
11. Garg, R., Gupta, S.P., Gao, H., Babu, M.S., Debnath, A.K. and Hansch, C., Chem. Rev., 99 (1999) 3525.
12. Waller, C.L., Oprea, T.I., Giolitti, A. and Marshall, G.R., J. Med. Chem., 36 (1993) 4152.
13. Debnath, A.K., J. Med. Chem., 42 (1999) 249.
14. So, S.S. and Karplus, M., J. Med. Chem., 40 (1997) 4347.
15. So, S.S. and Karplus, M., J. Med. Chem., 40 (1997) 4360.
16. So, S.S. and Karplus, M., J. Comput.-Aided Mol. Design, 13 (1999) 243.
17. Brooks III C.L., Karplus, M. and Pettitt, B.M., Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics, Vol. 71 of Advances in Chemical Physics, Wiley-Interscience, New York, 1988.
18. Archontis, G. and Karplus, M., J. Chem. Phys., 105 (1996) 11246.
19. Zeng, J., Treutlein, H.R. and Simonson, T., Proteins: Struct. Funct. Genet., 31 (1998) 186.
20. Rao, B.G., Tilton, R.F. and Singh, U.C., J. Am. Chem. Soc., 114 (1992) 4447.
21. Rao, B.G. and Murcko, M.A., Protein Eng., 9 (1996) 767.
22. Caflisch, A., Schramm, H. J. and Karplus, M., J. Comput.-Aided Mol. Design, 14 (2000) 161.
23. Stultz, C.M. and Karplus, M., Proteins: Struct. Funct. Genet., 40 (2000) 258.
24. Schapira, M., Totrov, M. and Abagyan, R., J. Mol. Recognit., 12 (1999) 177.
25. Bardi, J. S., Luque, I. and Freire, E., Biochemistry, 36, (1997) 6588.
26. Holloway, M.K. et al., J. Med. Chem., 38 (1995) 305.
27. Perez, C., Pastor, M., Ortiz, A.R. and Gago, F., J. Med. Chem., 41 (1998) 836.
28. Böhm, H.J., J. Comput.-Aided Mol. Design, 8 (1994) 243.
29. Böhm, H.J., J. Comput.-Aided Mol. Design, 12 (1998) 309.

30. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., J. Am. Chem. Soc., 118 (1996) 3959.
31. Muegge, I. and Martin, Y., J. Med. Chem., 42 (1999) 791.
32. Muegge, I., Perspect. Drug Discov. Design, 20 (2000) 99.
33. Åqvist, J., Medina, C. and Samuelson, J.-E., Protein Eng., 7 (1994) 385.
34. Srinivasan, J., Cheatham III, T.E., Cieplak, P., Kollman, P.A. and Case, D.A., J. Am. Chem. Soc., 120 (1998) 9401.
35. Gilson, M.K. and Honig, B.H., Proteins: Struct. Funct. Genet., 3 (1988) 32.
36. Wang, W. and Kollman, P.A., J. Mol. Biol., 303 (2000) 567.
37. Gohlke, H., Kiel, C. and Case, D.A., J. Mol. Biol., 330 (2003) 891.
38. Kuhn, B. and Kollman, P.A., J. Am. Chem. Soc., 122 (2000) 3909.
39. Lee, T.S. and Kollman, P.A., J. Am. Chem. Soc., 122 (2000) 4385.
40. Wang, J., Morin, P., Wang, W. and Kollman, P.A., J. Am. Chem. Soc., 123 (2001) 5221.
41. Sirockin, F., Sich, C., Improta, S., Schaefer, M., Saudek, V., Froloff, N., Karplus, M. and Dejaegere, A., J. Am. Chem. Soc., 124 (2002) 11073.
42. Jones-Hertzog, D.K. and Jorgensen, W.L., J. Med. Chem., 40 (1997) 1539.
43. Zhou, R., Friesner, R.A., Ghosh, A., Rizzo, R., Jorgensen, W.L. and Levy, R.M., J. Chem. Phys. B, 105 (2001) 10388.
44. Hansson, T., Marelius, J. and Åqvist, J., J. Comput.-Aided Mol. Design, 12 (1998) 27.
45. Paulsen, M.D. and Ornstein, R.L., Protein Eng., 9 (1996) 567.
46. Wang, W., Wang, J. and Kollman, P.A., Proteins: Struct. Funct. Genet., 34 (1999) 395.
47. Åqvist, J., J. Comput. Chem., 17 (1996) 1587.
48. Ghosh, A., Rapp, C.S. and Friesner, R.A., J. Phys. Chem. B, 102 (1998) 10983.
49. Navia, M.A., Fitzgerald, P.M., McKeever, B.M., Leu, C.T., Heimbach, J.C., Herber, W.K., Sigal, I.S., Darke, P.L. and Springer, J.P., Nature, 337 (1989) 615.
50. Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B.K., Baldwin, E., Weber, I.T., Selk, L.M., Clawson, L., Schneider, J. and Kent, S.B., Science, 245 (1989) 616.
51. Vondrasek, J. and Wlodawer, A., Science, 272 (1996) 337.
52. Vondrasek, J. and Wlodawer, A., Proteins: Struct. Funct. Genet., 49 (2002) 429.
53. Zoete, V., Michielin, O. and Karplus, M., J. Mol. Biol., 315 (2002) 21.
54. Marelius, J., Graffner-Nordberg, M., Hansson, T., Hallberg, A. and Åqvist, J., J. Comput.-Aided Mol. Design, 12 (1998) 119.
55. Sham, Y.Y., Chu, Z.T., Tao, H. and Warshel, A., Proteins: Struct. Funct. Genet., 39 (2000) 393.
56. Warwicker, J. and Watson, H.C., J. Mol. Biol., 157 (1982) 671.
57. Ljungberg, K.B., Marelius, J., Musil, D., Svensson, P., Norden, B. and Åqvist, J., Eur. J. Pharm. Sci., 12 (2001) 441.
58. Brooks, B.R., Bruccoleri, R., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4 (1983) 187.
59. MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. and Karplus, M., J. Phys. Chem. B, 102 (1998) 3586.
60. Eckstein, J. and Riesinger, S., private communication.
61. Spackman, M.A., J. Comput. Chem., 17 (1996) 1.
62. Schmidt, M.W., Baldridge, K.K., Boatz, J.A., Elbert, S.T., Gordon, M.S., Jensen, J.J., Koseki, S., Matsunaga, N., Nguyen, K.A., Su, S., Windus, T.L., Dupuis, M. and Montgomery, J.A., J. Comput. Chem., 14 (1993) 1347.
63. Baldwin, E.T., Bhat, T.N., Gulnik, S., Liu, B., Topol, I.A., Kiso, Y., Mimoto, T., Mitsuya, H. and Erickson, J.W., Structure, 3 (1995) 581.
64. Harte, W.E. and Beveridge, D.L., J. Am. Chem. Soc., 115 (1993) 1231.
65. Geller, M., Miller, M., Swanson, S. M. and Maizel, J., Proteins: Struct. Funct. Genet., 27 (1997) 184.
66. Wang, Y.X., Freedberg, D.I., Yamazaki, T., Wingfield, P.T., Sthal, S.J., Kaufman, J.D., Kiso, Y. and Torchia, D.A., Biochemistry, 35 (1996) 9945.
67. Tawa, G.J., Topol, I.A., Burt, S.K. and Erickson, J.W., J. Am. Chem. Soc., 120 (1998) 8856.
68. Tong, L., Pav, S., Mui, S., Lamarre, D., Yoakim, C., Beaulieu, P. and Anderson, P.C., Structure, 3 (1995) 33.
69. Kim, E.E., Baker, C.T., Dwyer, M.D., Murcko, M.A., Rao, B.G., Tung, R.D. and Navia, M.A., J. Am. Chem. Soc., 117 (1995) 1181.
70. Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J.C., J. Comput. Phys., 23 (1977) 327.
71. Davis, M.E., Madura, J.D., Luty, B.A. and McCammon, J.A., Comput. Phys. Commun., 62 (1991) 187.
72. Schaefer, M., private communication.
73. So, S.S. and Karplus, M., J. Med. Chem., 39 (1996) 5246.
74. Hulten, J., Bonham, N.M., Nillroth, U., Hansson, T., Zuccarello, G., Bouzide, A., Åqvist, J., Classon, B., Danielson, U.H., Karlen, A., Kvarnstrom, I., Samuelsson, B. and Hallberg, A., J. Med. Chem., 40 (1997) 885.
75. Archontis, G., Simonson, T. and Karplus, M., J. Mol. Biol., 306 (2001) 307.
76. van Vlijmen H.W., Schaefer, M. and Karplus, M., Proteins: Struct. Funct. Genet., 33 (1998) 145.
77. Cheng, Y. and Prusoff, W.H., Biochem. Pharmacol., 22 (1973) 3099.
78. Brooks III C.L. and Karplus, M., J. Mol. Biol., 208 (1989) 159.
79. Kuhn, B. and Kollman, P.A., J. Med. Chem., 43 (2000) 3786.
80. Hosur, M.V., Bhat, T.N., Kempf, D.J., Baldwin, E.T., Liu, B., Gulnik, S., Wideburg, N.E., Norbeck, D.W., Applet, K. and Erickson, J.W., J. Am. Chem. Soc., 116 (1994) 847.
81. Silva, A.A.M., Cachau, R.E., Sham, H.L. and Erickson, J.W., J. Mol. Biol., 255 (1996) 321.
82. Kaldor, S.W., Kalish, V.J., Davies, F.F., Shetty, B.V., Fritz, J.E., Appelt, K., Burgess, J.A., Campanale, K.M., Chirgadze, N.Y., Clawson, D.K., Dressman, B.A., Hatch, S.D., Halil, D.A., Kosa, M.B., Lubbehusen, P.P., Muesing, M.A., Patick, A.K., Reich, S.H., Su, K.S. and Tatlock, J.H., J. Med. Chem., 40 (1997) 3979.
83. Johti, H., Singh, O.M.P., Weir, M.P., Cooke, R., Murray-Rust, P. and Wonacott, A., Biochemistry, 33 (1994) 8417.
84. Munshi, S., Chen, Z., Li, Y., Olsen, D.B., M.E. Fraley, R.W. Hungate and Kuo, L.C., Acta Crystallogr., D54 (1998) 1053.
85. Dorsey, B.D., Levin, R.B., McDaniel, S.L., Vacca, J.P., Guare, J.P., Darke, P.L., Zugay, J.A., Emini, E.A., Schleif, W.A., Quintero, J.C. et al., J. Med. Chem., 37 (1994) 3443.
86. Chen, Z., Li, Y., Chen, E., Hall, D.L., Darke, P.L., Culberson, C., Shafer, J.A. and Kuo, L.C., J. Biol. Chem., 269 (1994) 26344.

87. Hoog, S.S., Zhao, B., Winborne, E., Fisher, S., Green, D.W., DesJarlais, R.L., Newlander, K.A., Callahan, J.F., Moore, M.L., Huffman, W.F. et al., J. Med. Chem., 38 (1995) 3246.
88. Abdel-Meguid, S.S., Zhao, B., Murthy, K.H., Winborne, E., Choi, J.K., DesJarlais, R.L., Minnich, M.D., Culp, J.S., Debouck, C., Tomaszek, T.A., Meek, T.D. and Dreyer, G.B., Biochemistry, 32 (1993) 7972.
89. Thompson, S.K., Murthy, K.H., Zhao, B., Winborne, E., Green, D.W., Fisher, S.M., DesJarlais, R.L., Tomaszek, T.A., Meek, T.D., Gleason, J.G. and Abdel-Meguid, S.S., J. Med. Chem., 37 (1994) 3100.
90. Koradi, R., Billeter, M. and Wüthrich, K., J. Mol. Graphics, 14 (1996) 51.