DOI 10.1007/s10822-007-9116-0

Calculations of protein-ligand binding entropy of relative and overall molecular motions

Anatoly M. Ruvinsky

Received: 7 January 2007 / Accepted: 27 March 2007 / Published online: 15 May 2007
© Springer Science+Business Media B.V. 2007

Abstract In the context of virtual database screening, calculations of protein-ligand binding entropy of relative and overall molecular motions are challenging, owing to the inherent structural complexity of the ligand binding well in the energy landscape of protein-ligand interactions and computing time limitations. We describe a fast statistical thermodynamic method for estimation the binding entropy to address the challenges. The method is based on the integration of the configurational integral over clusters obtained from multiple docked positions. We apply the method in conjunction with 11 popular scoring functions (AutoDock, ChemScore, DrugScore, D-Score, F-Score, G-Score, LigScore, LUDI, PLP, PMF, X-Score) to evaluate the binding entropy of 100 protein-ligand complexes. The averaged values of binding entropy contribution vary from 6.2 to 9.1 kcal/mol, showing good agreement with literature. We calculate positional sizes and the angular volume of the native ligand wells. The averaged geometric mean of positional sizes in principal directions varies from 0.8 to 1.4 Å. The calculated range of angular volumes is 3.3–11.8 rad^2. Then we demonstrate that the averaged six-dimensional volume of the native well is larger than the volume of the most populated non-native well in energy landscapes described by all of 11 scoring functions.

Keywords Protein-ligand docking · Binding affinity · Entropy · Scoring function · Clustering

A. M. Ruvinsky (✉)
Center for Bioinformatics, The University of Kansas,
2030 Becker Drive, Lawrence, KS 66047, USA
e-mail: ruvinsky@ku.edu

Introduction

The accurate determination of absolute binding free energy (binding affinity) is a key element of computer-aided drug design [1–9] and of statistical thermodynamic description of binding/recognition processes [10–13]. Many of the computational approaches to calculate protein-ligand binding affinity are based on thermodynamic integration [12–14], free energy perturbation [12, 15], the generalized Born and surface area models [16, 17], and the Poisson-Boltzmann method [18–20]. These methods are computationally expensive and therefore cannot be practically applied to virtual screening for bioactive molecules. Alternative approaches for fast screening are based on use of knowledge-based, empirical and force-field scoring functions [21–32]. The differences in the various methods of scoring functions are in the parametrization and functional forms of intermolecular potentials and entropy terms. Despite the considerable progress achieved in virtual screening and computer-aided drug design [1–9], the reliable and fast prediction of binding affinity is still a scientific challenge [2, 9, 10, 33–38]. Included in this challenge are calculations of the binding entropy [10, 39]. Considering strengths and weaknesses of three docking and seven scoring approaches, Bissantz et al. noted that ‘‘poor treatment of entropic components to the binding free energy’’ is a one of ‘‘the well-known scoring function deficiencies’’ [33]. The key importance of entropy calculations for fragment-based drug discovery approaches has been discussed recently by Murray and Verdonk [40, 41].

Entropy contributions arise from the rearrangement of water surface layers around solutes during binding (the nonpolar solvation contribution, accounting for the hydrophobic effect) [42, 43], from the freezing or reduction of translational and rotational degrees of freedom of solutes

[14, 19, 44–50], from the changes of normal modes of solutes during binding [19, 51–53], from the change of the number of conformations (rotamers) [54, 55], and from protonation/ deprotonation events [56, 57].

Much uncertainty and controversy exist regarding the estimation of protein-ligand binding entropy of relative and overall molecular motions within either computationally expensive conformational sampling-based methods [10–13, 58] or the method of scoring function [26, 39, 40]. Often, computationally expensive methods to calculate the entropy are based on normal mode analysis [52, 53], or on slow-growth thermodynamic integration and molecular dynamic (MD) simulations [14], MD simulations and the quasiharmonic model [19, 20], or on evaluation of configurational integrals with the Mining Minima Method [11]. The commonly used approach in virtual screening is to account for the entropy with a regression constant term in empirical scoring functions [26–28, 40]. Recently, we have suggested a nonparametric method to calculate the binding entropy and reported that the method substantially improves predicting of native ligand positions [49, 50]. The method is based on the estimate of the configurational integral through the sizes of clusters obtained from multiple docked positions. It modifies a commonly used form of scoring functions with a term logarithmically dependent on the configurational integral. Since the entropy term has a simple analytical form, the method keeps the important ability of fast ligand screening.

In this paper, we focus on the estimation of the binding entropy of relative and overall molecular motions and characterization of protein-ligand energy landscapes described by 11 popular scoring functions (AutoDock [21], G-Score [22], D-Score [23], LigScore [24], PLP [25], LUDI [26], F-Score [27], ChemScore [28], X-Score [29], PMF [30], DrugScore [31]) and the conformational ensembles of 100 PDB protein-ligand complexes developed by Wang et al. [35]. The binding entropy depends on the configurational integral over six degrees of freedom of relative translations and rotations. The amplitudes of relative motions are restricted by the sizes of the ligand binding well in the energy landscape of protein-ligand interactions. While many calculations use molecular dynamics to investigate positional and angular sizes of the binding wells [14, 18–20, 59, 60] we use scored ensembles of docked ligand positions and a clustering procedure. The sizes are estimated as the intervals of variation of center of ligand mass and Euler angles in the top-scored cluster of ligand positions located within 2 Å of the native position. The computed sizes are in agreement with previously reported assessments of the binding well dimensions. Then we characterize the most populated non-native wells and demonstrate that these wells are narrower than the native wells. The results of entropy calculations show that the method applied in conjunction with 11 scoring functions estimates binding entropy quantitatively similar to molecular dynamics-based methods.

Materials and methods: theory

Before binding with a protein, a solute molecule has three translational and three rotational degrees of freedom (two rotational degrees for a linear molecule) associated with the molecular motion as the whole, and 3 × (number of atoms)–6 vibrational degrees of freedom. During binding the ligand becomes localized in the protein active site. Thus ligand degrees of freedom associated with the overall molecular motions are reduced into the vibrational degrees of freedom of the internal relative motions in the protein-ligand complex. These vibrational modes and the related change of vibrational density of states were observed recently in experimental investigations [61, 62] of far infrared spectrum of protein-ligand and DNA-ligand complexes. Considering only relative protein-ligand motions we can write the protein-ligand binding free energy in the form [10, 11, 14, 20, 49]

$$\Delta G = E_{pl} - E_p - E_l - k_B T \ln\left(\frac{\sigma_l \sigma_p}{\sigma_{pl}} c_o N_a \frac{V_B}{8\pi^2}\right), \quad (1)$$

where

$$V_B = \int\limits_{\Gamma} \exp\left(-\frac{U_{pl}(\mathbf{r}, \theta, \varphi, \psi) - E_{pl}}{k_B T}\right) d\mathbf{r}\ \sin\theta\, d\theta\, d\varphi\, d\psi \quad (2)$$

is the configurational integral of the complex; where $E_{p,l,pl}$ are the ground energies of protein (p), ligand (l) and protein-ligand complex (pl) in solution; N_a is the Avogadro number; c_o = 1 *mol*/*l*; $\sigma_{l,p,pl}$ are the orders of symmetry of ligand, protein and protein-ligand complex (for a nonsymmetrical molecule σ = 1; if a molecule has 2-fold axis of symmetry σ = 2, etc.); $U_{pl}(\mathbf{r}, \theta, \varphi, \psi)$ is the energy of the protein-ligand complex in solution; **r** is the vector of relative translational motions in the complex; $(\theta, \varphi, \psi) = \mathbf{\Omega}$ are Euler angles of relative orientational motions; Γ is the the region of integration in the 6-dimensional space of **r** and (θ,ϕ,ψ); E_{pl} is the minimum of $U_{pl}(\mathbf{r}, \theta, \varphi, \psi)$ in the region Γ; k_B is the Boltzmann constant. The derivation of the sine of the polar angle in Eq. (2) is provided in Appendix. Our interest is the logarithmic term in Eq. (1) accounting for the entropy of relative and overall motions.

To estimate V_B, we first partition all docked positions generated by Wang et al. [35] into non-overlapping clusters in such a way that every cluster contains ligand positions with RMSD less than a definite value, the clustering-RMSD, relative to the ligand position having minimal energy in the cluster (Fig. 1). We use optimal values of the clustering-RMSD that resulted in a maximal success rate of docking [50]. The clustering procedure starts with the ordering of all docked ligand positions according to their energies. Then, starting from the lowest-energy position (the origin of the first cluster), we compute RMSDs of this position with respect to all positions with higher energies. If the RMSD is lower than the clustering-RMSD, then we assign the corresponding high energy position to the first cluster. The second cluster grows up from the lowest-energy position which is not covered by the first cluster. Now we compute RMSDs of this position with respect to all unassign positions with higher energies. The algorithm repeats until all docked positions are assigned to clusters. We would like to note that the number of clusters and their composition depend on a scoring function used to order docked positions. Now we can consider the clusters as the possible ligand binding modes. Further, we designate the docked ligand position having minimal energy in the cluster as the representative position in the cluster. The result of the clustering procedure is a list of representative positions that have RMSDs between one another greater than the clustering-RMSD. All other docked positions are assigned to the cluster of the nearest representative position.

All ligand positions in the cluster numbered i can be considered as snapshots of the ligand motion near the representative docked position $(\mathbf{r}_i, \boldsymbol{\Omega}_i)$. The variation intervals of $(\mathbf{r}, \boldsymbol{\Omega})$ in the cluster give the estimate of the configurational integral as

$$\begin{aligned} V_B(\mathbf{r}_i, \boldsymbol{\Omega}_i) \approx \Gamma_i = {} & [-\cos(\max\theta_i) + \cos(\min\theta_i)] \\ & [\max(\varphi_i) - \min(\varphi_i)][\max(\psi_i) - \min(\psi_i)] \\ & [\max(x_i) - \min(x_i)][\max(y_i) - \min(y_i)] \\ & [\max(z_i) - \min(z_i)], \end{aligned} \tag{3}$$

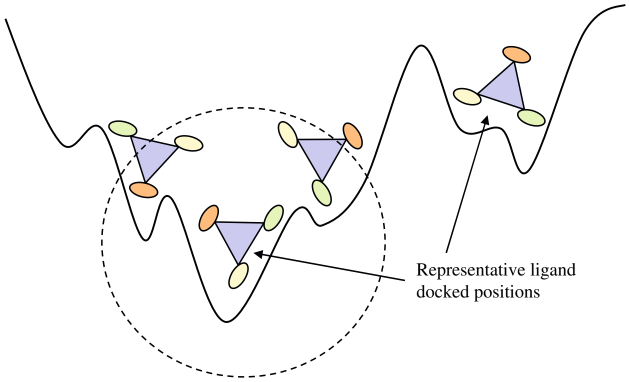

Fig. 1 The protein-ligand binding energy landscape. The large dashed circle shows a multiconformational cluster of docked ligand positions

where $\max(\mathbf{r}_i, \boldsymbol{\Omega}_i)$ and $\min(\mathbf{r}_i, \boldsymbol{\Omega}_i)$ are the maximum and minimum values of $(\mathbf{r}, \boldsymbol{\Omega})$ in the cluster numbered i.

To determine Euler angles of a ligand position, we superimpose its center of mass with the center of mass of the native position and find its principal axes. The unit vectors **e** along these directions are the eigenvectors of the tensor of inertia. The orientation of the principal axes with respect to the fixed system of coordinates is determined by the Euler angles. The vectors **e** and the unit vectors **e** of the fixed axes are related with a matrix equation: $\mathbf{e} = \mathbf{A}(\theta, \varphi, \psi)\mathbf{e}$, where $\mathbf{A}(\theta, \varphi, \psi)$ is a rotation matrix [63]. We find elements of the rotation matrix and use them to find the Euler angles.

Using $\sigma_p = \sigma_{pl} = 1$, we obtain

$$k_B \Delta S_i = k_B T \ln \Gamma_i \frac{\sigma_l c_o N_a}{8\pi^2} \tag{4}$$

Eq. (4) can also be derived by a Monte-Carlo approximation of the configurational integral [49].

For consistency with common usage, through the paper, we will use the term ‘‘entropy contribution’’ for free energy contribution $-k_B T \Delta S$. To evaluate the entropy contribution we determine the set $\{\Gamma\}$, calculate

$$\Delta G_i = \Delta H_i - k_B T \ln \Gamma_i \frac{\sigma_l c_o N_a}{8\pi^2} \tag{5}$$

for all Γ_i (ΔH_i is the score of the representative position i) and select a representative position having a minimal value of ΔG_i. The representative position resides in a cluster of the binding mode. We substitute the binding volume of this cluster into Eq. (4).

Materials and methods: the test set

We used the test set of 100 PDB protein-ligand complexes [35]: 1bbz, 4xia, 8xia, 2xim, 1fkf, 1fkb, 1hvr, 1tet, 2cgr, 1abf, 1apb, 7abp, 5abp, 8abp, 9abp, 1abe, 1bap, 6abp, 1e96, 1add, 2ak3, 1adb, 9aat, 1bzm, 1cbx, 2ctc, 3cpa, 1cla, 3cla, 4cla, 2csc, 5cna, 1af2, 1dr1, 1dhf, 1drf, 1ela, 7est, 3fx2, 2gbp, 1hsl, 2qwd, 2qwe, 2qwf, 2qwg, 2qwc, 2qwb, 1mnc, 1exw, 1apw, 1apt, 1bxo, 1fmo, 2pk4, 1inc, 4sga, 5sga, 5p21, 1rbp, 1rgk, 6rnt, 1rgl, 1rnt, 1zzz, 1yyy, 1b5g, 1ba8, 1bb0, 2sns, 1sre, 7tln, 4tln, 1tmn, 2tmn, 3tmn, 5tln, 1tlp, 1etr, 1ets, 1d3d, 1d3p, 1a46, 1a5g, 1bcu, 1tha, 4tim, 6tim, 7tim, 1bra, 1tnj, 1pph, 1tnk, 1tnh, 1tni, 1ppc, 1tng, 3ptb, 1tnl, 1bhf, 2xis. All these entries have resolution better than 2.5 Å. Wang et al. [35] selected the test set from 172 protein-ligand complexes used in their previous work. They used the AutoDock program [21] to generate an ensemble of 101 docked conformations for each ligand in the test set. One of the conformations corresponds to the experimentally observed native conformation of the ligand.

The AutoDock program uses a genetic algorithm (GA) for conformational sampling. Wang et al. [35] performed 100 individual GA runs to generate 100 docked conformations. In every GA run they created a population of 100 individuals, assigning a random set of translational coordinates of ligand center of mass, a random orientational ligand position, and random torsions to each of the 100 individuals. They applied a sophisticated algorithm described in their article to achieve the diversity of the ensembles. For 72 complexes, a satisfactory conformational ensemble was not obtained, and these complexes were discarded. The final ensembles represent local minima rather than random spots on the protein-ligand energy landscape. RMSD distributions in the conformational ensembles spread from 0 Å to 15 Å. The ensembles of docked conformations are scored using above mentioned eleven scoring functions (http://sw16.im.med.umich.edu/software/xtool/).

Results and discussion: calculations of protein-ligand binding entropy

The total binding entropy can be split into translational and rotational contributions as $k_BT\Delta S = k_BT(\Delta S_{tr} + \Delta S_{rot})$. The entropy change of overall and relative translational motions has the following form

$$k_BT\Delta S_{tr} = k_BT\ln\Big(N_a c_o[\max(x_1) - \min(x_1)] [\max(y_1) - \min(y_1)][\max(z_1) - \min(z_1)]\Big), \quad (6)$$

where 1 is the number of the top-ranked representative position, scored using Eq. (5); $\max(\mathbf{r}_1)$ and $\min(\mathbf{r}_1)$ are the maximal and minimal values of $\mathbf{r}$ in the cluster of the top-ranked representative position. The entropy change of overall and relative orientational motions is

$$k_BT\Delta S_{rot} = k_BT\ln\tfrac{\sigma_l}{8\pi^2} + k_BT\ln[-\cos(\max\theta_1) + \cos(\min\theta_1)][\max(\varphi_1) - \min(\varphi_1)][\max(\psi_1) - \min(\psi_1)], \quad (7)$$

where $\max(\boldsymbol{\Omega}_1)$ and $\min(\boldsymbol{\Omega}_1)$ are the maximal and minimal values of $\boldsymbol{\Omega}$ in the cluster 1.

Summarized in Table 1 are the entropy contributions averaged over one hundred protein-ligand complexes. The averaging is done as a simple mean $< \ldots > = 1/100\sum_{s=1}^{100}\ldots$. The averaged values of $-k_BT\Delta S$ vary from 6.2 (F-Score) to 9.1 kcal/mol (ChemScore). This variation is a result of differences in scores of a docked ligand position measured by the scoring functions. As we described in Materials and Methods: Theory, the clustering procedure groups docked positions based on their scores and RMSDs. Therefore the cluster composition and the corresponding well volume Γ depend on a scoring function used to measure the scores. The standard error of the mean of the total binding entropy being lower than $k_BT = 0.6$ kcal/mol varies from 0.2 to 0.4 kcal/mol. It is interesting to note that for all scoring functions the absolute values of translational contributions are about 2-fold higher than the absolute values of rotational contributions. The results are physically reasonable and quantitatively similar to the results of other calculations [14, 18–20, 45, 64, 65, 59].

Indeed, using strained molecular dynamic simulations to study binding of benzene to a mutant T4 lysozyme, Hermans and Wang [14] estimated the total entropy contribution of 7 kcal/mol. The same complex was studied by Carlsson and Åqvist [65] using unstrained MD simulations. They obtained the total entropy of 8.3 kcal/mol using Schlitter's formula, 8.5 kcal/mol using quasiharmonic analysis, 6.0 kcal/mol using uniform distribution of the ligand center of mass and Euler angels, 5.4 kcal/mol using Gaussian distribution of the ligand center of mass and Euler angels. Luo and Gilson [64] took a different approach

Table 1 Averaged entropy contributions to binding affinity at 298 K, kcal/mol

Scoring function	Translational contribution $-T\Delta S_{tr}$	Rotational contribution $-T\Delta S_{rot}$	Total entropy $-T(\Delta S_{tr} + \Delta S_{rot})$	Standard error of the mean of the total entropy
AutoDock	4.79	2.67	7.46	0.24
ChemScore	5.41	3.70	9.11	0.37
DrugScore	4.30	2.05	6.35	0.21
D-Score	4.99	2.70	7.69	0.25
F-Score	4.22	2.02	6.24	0.21
G-Score	4.92	2.69	7.61	0.26
LigScore	4.68	2.52	7.20	0.23
LUDI	4.32	1.99	6.31	0.21
PLP	4.31	2.09	6.40	0.22
PMF	4.74	2.72	7.46	0.24
X-Score	4.80	2.66	7.46	0.27

to compute the translational and rotational entropy. They used the ''Mining Minima'' method to analyze binding of adenine to synthetic adenine receptors and obtained the value of 7 kcal/mol. Erickson [45] suggested a value of $-k_BT\Delta S$ = 11 kcal/mol for the entropy contribution in protein association. He also noted that a lower value of 7 kcal/mol gave the best fit to experimental data. Baginski et al. [18] obtained a $-k_BT\Delta S$ between 4 kcal/mol and 9 kcal/mol from free energy calculations and MD simulations of anthracycline antibiotics to DNA. Swanson et al. [20] estimated that for FK506 binding protein and ligand 4-hydroxy-2-butanone, association leads to the entropy contribution of 5.6 kcal/mol. Luo and Sharp [19] used quasi-harmonic analysis and MD simulations and estimated the entropy contribution of 7.5 kcal/mol for antibody FAB fragment-digoxigenin binding and 6.7 kcal/mol for streptavidin-biotin binding. Lazaridis et al. [59] obtained a $-k_BT\Delta S$ between 4.5 kcal/mol and 6 kcal/mol from MD simulations of ligand-avidin/streptavidin complexes. It is encouraging to see that the method applied with 11 scoring functions and the cited above force-field based methods address the entropy contribution of overall and relative motions quantitatively similar.

To evaluate the entropy effect on correlation between calculated and experimentally measured binding affinities, we compared correlation coefficients, accounting and not accounting for entropy contributions, for each of 11 scoring functions. Correlation analysis showed no statistically significant influence of the entropy on the correlation coefficients. Both approaches demonstrated from weak to medium relationships with close values of the correlation coefficients that vary from 0.3 to 0.7. These results demonstrate that the improvement of the one contribution to binding affinity can be necessary but not sufficient for accurate prediction of binding affinity. This fact is not new and has been recognized in recent publications (see, e.g., [36, 38]). We would like to attract attention to another interesting aspect that can be learned from the confrontation between ''positive'' results (the success of the method achieved in entropy calculations and predicting native ligand positions [49, 50]) and ''negative'' results (the weak influence of the entropy contribution on the score-affinity correlation). Since the entropy calculations use ensembles of bound conformations only, our results could be an indication for a lack of unbound contributions to binding affinity. These contributions have no effect on the entropy calculations and docking accuracy, but they may have an effect on estimates of binding affinity. Considering contribution of ligand unbound conformers to binding affinity, Tirado-Rives and Jorgensen noted ''that the uncertainties from this source alone are sufficient to preclude the viability of current docking methodology for rank-ordering of diverse compounds in high-throughput virtual screening'' [55]. It will be of interest in the future to combine their approach with the method accounting entropy of relative motions and to explore the performance of the combined method in the correlation test.

Characterization of native ligand wells

The concept of the energy landscape described by scoring functions is important for better understanding of protein-ligand interactions and for designing adequate docking procedures [66]. To gain insight into the structure of the ligand binding well in the protein-ligand energy landscape, we investigated the averaged range of deviation of the ligand center of mass and Euler angles in the top scored cluster having the representative position within 2 Å from the native position. Since only a part of the 100 protein-ligand complexes has such clusters, we averaged the deviations in Eqs. (8) and (9) over the number of these complexes.

For each of scoring functions, we calculated

$$\langle \Delta x \Delta y \Delta z \rangle_{av}^{1/3} = \left(\sum_{s=1}^{N} [\max(x_1^s) - \min(x_1^s)][\max(y_1^s) - \min(y_1^s)] [\max(z_1^s) - \min(z_1^s)]/N \right)^{1/3} \tag{8}$$

and

$$\langle \Delta(\cos\theta)\Delta\varphi\Delta\psi \rangle_{av} = \sum_{s=1}^{N} [-\cos(\max\theta_1^s) + \cos(\min\theta_1^s)][\max(\varphi_1^s) - \min(\varphi_1^s)][\max(\psi_1^s) - \min(\psi_1^s)]/N, \tag{9}$$

where the subscript s numerates selected complexes; N is the number of selected complexes. As summarized in Table 2, the averaged estimates allow 0.8–1.4 Å fluctuations of the ligand center of mass. The geometric mean of the standard error of $\langle \Delta x \Delta y \Delta z \rangle_{av}$ varies from 0.4 to 0.7 Å. Thus, depending on a scoring function used, the geometric mean of the fluctuations of the ligand center of mass varies from 0.4 to 2.1 Å. Since the geometric mean of the data set of unequal members (Δx, Δy, Δz) is always smaller then the maximum in the data set, the positional sizes of the native wells allow even larger fluctuations in principal directions. These estimates are similar to the values predicted by Erickson [45], Ben-Tal et al. [47], Baginski et al. [18], Swanson et al. [20], Hermans and Wang [14]. However, we should keep in mind that our estimates of the ligand movement inside the native ligand well are based on the positional sizes of the cluster of docked ligand positions on the surface of the rigid protein. Considering protein flexibility might increase the estimates of the fluctuations.

Table 2 The averaged positional size and angular volume of top-ranked clusters with the representative solutions within 2Å from the experimentally determined position

Scoring function	$\langle \Delta x \Delta y \Delta z \rangle_{av}^{1/3}$, Å	[a]SEM$^{1/3}$$\{\Delta x \Delta y \Delta z\}$, Å	$\langle \Delta(\cos\theta)\Delta\varphi\Delta\psi \rangle_{av}$, rad^2	[a]SEM$\{\Delta(\cos\theta)\Delta\varphi\Delta\psi\}$, rad^2
AutoDock	1.0	0.5	6.7	1.2
ChemScore	0.8	0.4	3.3	0.8
DrugScore	1.3	0.6	11.3	1.5
D-Score	0.9	0.5	5.6	1.0
F-Score	1.3	0.6	10.7	1.4
G-Score	1.0	0.5	6.7	1.3
LigScore	1.1	0.5	6.7	1.1
LUDI	1.3	0.6	11.8	1.5
PLP	1.4	0.7	11.8	1.5
PMF	1.0	0.5	5.1	1.0
X-Score	1.0	0.5	7.4	1.3

[a] The standard error of the mean (SEM) is calculated as $\mathrm{SEM}\{x\} = \left(\sum_{i=1}^{n}(x_i-\bar{x})^2/(n^2-n)\right)^{1/2}$, where n is the sample size, and $\bar{x}$ is the mean

Applying MD simulations to the complex the immunophilin FKBP12 and the ligand 4-hydroxy-2-butanone, Swanson et al. [20] observed 1.7 Å of relative translational motions. Erickson's qualitative estimate was 2 Å of the center of mass motions in a protein-protein complex [45]. Considering association entropy of a peptide on a lipid membrane, Ben-Tal et al. [47] obtained 2.1 Å for fluctuations of the ligand center of mass. Baginski et al. [18] reported two estimates 1 Å and 2.3 Å of $\langle \Delta x \Delta y \Delta z \rangle_{av}^{1/3}$ in complexes of anthracycline antibiotics and DNA. Using MD simulations of the motion of benzene in the cavity of mutant T4 lysozyme and the body restraint algorithms, Hermans and Wang [14] found the displacement of the molecule center of mass in the plane of the benzene of 0.6 Å and the displacement of the molecule center of mass normal to the plane of the ring of 0.26 Å.

Finkelstein and Janin [46] used the Debye-Waller temperature factor to estimate the amplitude of relative oscillations and argued that ligands move by 0.2–0.25 Å. They also allowed larger movements up to 1 Å for less tight complexes of protease-inhibitor or antigen-antibody.

The averaged angular volumes of the native energy wells are presented in Table 2. All the values fall in the range from 3.3 to 11.8rad^2. The lowest value corresponds to the ChemScore scoring function. The higher value of the angular volume corresponds to PLP and LUDI. Changing $\langle \Delta(\cos\theta)\Delta\varphi\Delta\psi \rangle_{av}$ from 3.3 to 11.8rad^2 results in 0.8 kcal/mol change of the rotational entropy contribution.

Different values of $\langle \Delta(\cos\theta)\Delta\varphi\Delta\psi \rangle_{av}$ have been reported in the literature: 9.9 rad^2 and 0.2 rad^2 [18] for DNA-ligand complexes; 6.57 rad^2 [20] and 10^{-5}–10^{-3} rad^2 [46] for protein-ligand complexes; 10^{-2} rad^2 [60] for protein-protein complexes. Comparing these results with results given in Table 2, we do not find values of $\langle \Delta(\cos\theta)\Delta\varphi\Delta\psi \rangle_{av}$ ranging from 10^{-5} to 10^{-1} rad^2. Our method estimates angular volumes higher than methods [18, 46, 60] and similar to methods [18, 20].

Recent studies [50, 67] of protein-ligand docking revealed correlations between docking accuracies of two methods of ranking using Eq. (5) and ranking by cluster occupancy. These correlations suggest that the near-native positions, in comparison with far-native ones, have a large number of neighboring conformations within a clustering-RMSD. Similar trends were observed recently in studies of protein-ligand docking [68–70] and predictions of protein-protein complexes [71, 72]. Therefore, it is of particular interest to compare dimensions of the native wells with dimensions of the most populated non-native wells. Summarized in Table 3 are the averaged positional size and angular volume of the most populated clusters not containing the native ligand position. These results show that the averaged six-dimensional volume $\langle \Delta x \Delta y \Delta z \rangle_{av}$ $\langle \Delta(\cos\theta)\Delta\varphi\Delta\psi \rangle_{av}$ of the native well is larger than the volume of the most populated non-native well in energy landscapes described by all of 11 scoring functions. The ligand preference to reside in broader wells is of entropy origin. Indeed, according to the second law of thermodynamics—the law of increasing entropy or of decreasing the entropy contribution—the entropy contribution $-k_B T \ln \Gamma_i \sigma_l c_o N_a / 8\pi^2 = -T\Delta S_i$ of the protein-ligand complex tends to approach a minimum value at equilibrium. In other words, the ligand tends to occupy wells with the larger volume Γ_i to decrease the entropy contribution. Comparing the positional and angular sizes of the wells

Table 3 The averaged positional size and angular volume of the most populated clusters not containing the native ligand position

Scoring function	$\langle\Delta x\Delta y\Delta z\rangle_{av}^{1/3}$, Å	[a]SEM$^{1/3}\{\Delta x\Delta y\Delta z\}$, Å	$\langle\Delta(\cos\theta)\Delta\varphi\Delta\psi\rangle_{av}$, rad^2	[a]SEM$\{\Delta(\cos\theta)\Delta\varphi\Delta\psi\}$, rad^2
AutoDock	0.9	0.5	4.9	0.9
ChemScore	0.7	0.4	2.7	0.7
DrugScore	1.3	0.7	8.3	1.2
D-Score	0.9	0.5	4.9	1.0
F-Score	1.2	0.6	7.6	1.1
G-Score	0.9	0.5	5.0	1.2
LigScore	1.0	0.5	8.0	1.2
LUDI	1.2	0.6	6.9	1.0
PLP	1.3	0.7	9.1	1.3
PMF	0.9	0.5	6.2	1.0
X-Score	1.0	0.5	6.0	1.1

[a] The standard error of the mean (SEM) is calculated as $\mathrm{SEM}\{x\} = \left(\sum_{i=1}^{n}(x_i-\bar{x})^2/(n^2-n)\right)^{1/2}$, where n is the sample size, and $\bar{x}$ is the mean

separately, we note that the native well allows larger fluctuations of the ligand center of mass and Euler angels for all of the scoring functions except LigScore and PMF. The results for LigScore and PMF show that the non-native well can have the larger angular volume.

Discussion

The suggested method for estimation the binding entropy is based on the coarse-grained mapping of the the protein-ligand energy landscape (see [73, 74] and references therein for the using of the method in the studies of atomic clusters and biomolecules). Within the mapping, the dominant valleys in the rugged energy landscape are replaced with flat-bottomed energy wells. This results in the simple equation (4) for the entropy contribution that depends on the 6-dimensional volume of the well. Since we use the energy-based clustering procedure to map the valleys, the well volume depends on energies of docked positions of the corresponding cluster implicitly. The well volume is determined by the coordinates of ligand center of mass and Euler angles of docked positions located on the cluster's boundary and so by the conformational ensemble. Although the ensembles developed by Wang et al. [35] consist of 101 docked positions and do not cover all possible docked positions, use of them in conjunction with Eq. (5) substantially improves predicting of native ligand positions [50]. Therefore these ensembles are good representations of the important portions of the energy landscapes. Use of inappropriate ensembles will result in a breach of a correspondence between clusters and the dominant valleys and as a result will decrease the docking accuracy.

The generation of the appropriate conformational ensemble is a difficult task due to numerous local minima on the energy landscape. Therefore, we would not recommend using of gradient-based sampling procedures with our method. It will be interesting to investigate the performance of our method on the basis of ligand docked positions sampled by Monte Carlo (MC) techniques, evolutionary and genetic algorithms, tabu searches, methods combining different docking techniques (e.g., a tabu search with an MC simulated annealing [75], Mining Minima optimization [76], or MC simulated annealing with the Dead End Elimination algorithm for side-chain optimization [77]), and sampling algorithms that use a smoothing strategy for the rugged landscape [78].

The explicit energy-based equation for the entropy contribution can be derived by a Monte-Carlo approximation of the integral (2). In this case the entropy contribution has the following form

$$-k_B T\Delta S_i = -Tk_B \ln\left[\frac{\sigma_l c_o N_a}{8\pi^2}\frac{\Gamma_i(\mathbf{r}_i,\mathbf{\Omega}_i)}{N_i}\sum_{j=1}^{N_i}\exp\left(-\frac{E_i^j - E_i)}{T}\right)\right], \tag{10}$$

where N_i is the number of conformations in the cluster numbered i; E_i^j is the energy of the conformation j in the cluster i; $E_i = E_i^1$ is the energy of the representative position in the cluster i. Subtracting Eq. (4) from Eq. (10), we obtain

$$-Tk_B \ln\frac{1}{N_i}\left[1+\sum_{j=2}^{N_i}\exp\left(-\frac{E_i^j - E_i}{T}\right)\right] \tag{11}$$

The performance of both Eqs. (4) and (10) has been evaluated in docking tests on 135 protein-ligand complexes [49]. Both equations showed absolutely the same docking results, indicating a weak influence of Eq. (11) on the success rate of docking. The explanation verified by the energy histogram approach [67] is that energies E_i^j within the cluster i are nearly isoenergetic. For $E_i \approx E_i^j$, Eq. (11) yields a weak contribution. These results are entirely in accord with those reported by Kortvelyesi et al. [69]. Considering binding of ligands to lysozyme and thermolysin, the authors showed that the energy distribution of docked positions within clusters tends to be narrowed. It is interesting to note that recent protein studies showed that near-native conformations in comparison with misfolded ones reside in broader wells of the folding landscape and are characterized by low barrier heights between them [79–82].

Conclusions

In conclusion, we have described a nonparametric statistical thermodynamic method for estimation the binding entropy of overall and relative molecular motions. The method is based on the integration of the configurational integral over clusters obtained from multiple docked positions. We have applied the method in conjunction with 11 popular scoring functions to estimate the binding entropy of 100 protein-ligand complexes. The results show that alongside with computationally expensive conformational sampling-based methods to calculate the binding entropy, protein-ligand docking generating a statistical ensemble of docked positions provide information to derive the binding entropy. Depending on a scoring function used, the averaged binding entropy varies from 6.2 to 9.1 kcal/mol, showing good agreement with other previously reported molecular dynamics-based estimates and qualitative assessments of the binding entropy. We have calculated positional sizes and the angular volume of the ligand binding well in the energy landscape of protein-ligand interactions described by each of 11 scoring functions. The averaged geometric mean of positional sizes in principal directions has been estimated from 0.8 to 1.4 Å. The calculated angular volumes vary from 3.3 to 11.8rad^2. The findings are in agreement with known estimates of the binding well dimensions. We have shown that the averaged six-dimensional volume of the native well is larger than the volume of the most populated non-native well in energy landscapes described by all of the scoring functions. Finally, the method of entropy calculations used here is very fast, as the entropy terms have simple analytical forms. We believe our results and methodology provide a promising starting point for routine incorporation of entropy calculations in screening applications.

Acknowledgments We wish to thank reviewers for their helpful comments and suggestions.

Appendix

This appendix derives the partition functions of overall and relative rotational motions. It follows the method described by Kubo [83]. The Hamiltonian of a rotator is

$$H = \tfrac{1}{2I_1 \sin^2\theta}\left((p_\varphi - p_\psi \cos\theta)\cos\psi - p_\theta \sin\theta \sin\psi\right)^2 + \tfrac{1}{2I_2 \sin^2\theta}\left((p_\varphi - p_\psi \cos\theta)\sin\psi + p_\theta \sin\theta \cos\psi\right)^2 + \tfrac{1}{2I_3} p_\psi^2 \tag{12}$$

where I_1, I_2, I_3 are the principal moments of inertia of the molecular; (θ, ϕ, ψ) are Euler angles of rotational motions; $(p_\theta, p_\phi, p_\psi)$ are the corresponding canonical conjugate moments. The classical formula for the partition function of free rotations is

$$Z_{rot} = \frac{1}{(2\pi\hbar)^3\sigma} \int_0^\pi d\theta \int_0^{2\pi} d\psi \int_0^{2\pi} d\varphi \int_{-\infty}^{\infty} dp_\theta \int_{-\infty}^{\infty} dp_\psi \int_{-\infty}^{\infty} dp_\varphi \exp\left(-\frac{H}{k_B T}\right), \tag{13}$$

where σ is the order of symmetry of the molecule; $\hbar$ is the Planck's constant. The Hamiltonian in Eq. (12) can be rewritten as

$$H = \frac{1}{2}\left(\frac{\sin^2\psi}{I_1} + \frac{\cos^2\psi}{I_2}\right) \left(p_\theta + \left(\frac{1}{I_2} - \frac{1}{I_1}\right)\frac{\sin\psi\cos\psi}{\sin\theta((\sin^2\psi)/I_1 + (\cos^2\psi)/I_2)} (p_\varphi - p_\psi\cos\theta)\right)^2 + \frac{1}{2I_1 I_2 \sin^2\theta((\sin^2\psi)/I_1 + (\cos^2\psi)/I_2)} (p_\varphi - p_\psi\cos\theta)^2 + \frac{1}{2I_3} p_\psi^2 \tag{14}$$

Substituting Eq. (14) into Eq. (13) and using the Gaussian integral formula

$$\int_{-\infty}^{\infty} \exp\left(-a(x+b)^2\right) dx = \int_{-\infty}^{\infty} \exp\left(-ax^2\right) dx = \left(\frac{\pi}{a}\right)^{1/2}, \tag{15}$$

we can perform the integrations over the canonical conjugate moments in Eq. (13). Integration over p_θ gives

$$(2\pi k_B T)^{1/2} \left(\frac{\sin^2\psi}{I_1} + \frac{\cos^2\psi}{I_2}\right)^{-1/2} \tag{16}$$

Integration over p_ϕ gives

$$(2\pi k_B T I_1 I_2)^{1/2} \sin\theta \left(\frac{\sin^2\psi}{I_1} + \frac{\cos^2\psi}{I_2} \right)^{1/2} \tag{17}$$

Integration over p_ψ gives

$$(2\pi k_B T I_3)^{1/2} \tag{18}$$

Eqs. (16, 17, 18) and (13) yield the partition function as

$$\begin{aligned} Z_{rot} &= (2\pi k_B T)^{3/2} (I_1 I_2 I_3)^{1/2} \frac{1}{(2\pi\hbar)^3 \sigma} \int_0^\pi \sin\theta d\theta \int_0^{2\pi} d\psi \int_0^{2\pi} d\varphi \\ &= \frac{(2k_B T)^{3/2} (\pi I_1 I_2 I_3)^{1/2}}{\sigma \hbar^3} \end{aligned} \tag{19}$$

The sine of the polar angle in Eq. (19) appears as a result of integrations over the canonical conjugate moments. Thus, the partition function of restricted relative rotations is proportional to

$$\int_{\theta_1}^{\theta_2} \sin\theta d\theta \int_{\psi_1}^{\psi_2} d\psi \int_{\varphi_1}^{\varphi_2} d\varphi = [\cos\theta_1 - \cos\theta_2)][\varphi_2 - \varphi_1][\psi_2 - \psi_1] \tag{20}$$

References

1. Taylor RD, Jewsbury PJ, Essex JW (2002) J Comput Aided Mol Des 16:151
2. Gohlke H, Klebe G (2002) Angew Chem Int Ed 41:2644
3. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Protein Struct Funct Bioinf 47:409
4. Brooijmans N, Kuntz I (2003) Annu Rev Biophys Biomol Struct 32:335
5. Shoichet BK (2004) Nature 432:862
6. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Nat Rev Drug Discovery 3:935
7. Stockwell BR (2004) Nature 432:846
8. Cole JC, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) Prot Str Func Bioinf 60:325
9. Sousa SF, Fernandes PA, Ramos MJ (2006) Prot Struct Funct Bioinf 65:15
10. Gilson MK, Given JA, Bush BL, McCammon JA (1997) Biophys J 72:1047
11. Gilson MK, Given JA, Head MS (1997) Chem Biol 4:87
12. Wang W, Donini O, Reyes CM, Kollman PA (2001) Annu Rev Biophys Biomol Struct 30:211
13. Villa A, Zangi R, Pieffet G, Mark AE (2003) J Comput Aided Mol Des 17:673
14. Hermans J, Wang L (1997) J Am Chem Soc 119:2707
15. Essex JW, Severance DL, Tirado-Rives J, Jorgensen WL (1997) J Phys Chem B 101:9663
16. Bostrom J, Norrby PO, Liljefors T (1998) J Comp Aided Mol Des 12:383
17. Lee MS, Salsbury FR Jr, Brooks CL III (2002) J Chem Phys 116:10606
18. Baginski M, Fogolari F, Briggs JM (1997) J Mol Biol 274: 253
19. Luo H, Sharp K (2002) Proc Natl Acad Sci USA 19:10399
20. Swanson JMJ, Henchman RH, McCammon JA (2004) Biophys J 86:67
21. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) J Comput Chem 19:1639
22. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267:727
23. Ewing TJA, Makino S, Skillman AG, Kuntz ID (2001) J Comput Aided Mol Des 15:411
24. Cerius2, version 4.6; Accelrys Inc.; http: www.accelrys.com
25. Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DB, Fogel LJ, Freer ST (1995) Chem Biol 2:317
26. Böhm HJ (1994) J Comput Aided Mol Des 8:243
27. Rarey M, Kramer B, Lengauer T, Klebe G (1996) J Mol Biol 261:470
28. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) J Comput Aided Mol Des 11:425
29. Wang R, Lai L, Wang S (2002) J Comput Aided Mol Des 16:11
30. Muegge I, Martin YC (1999) J Med Chem 42:791
31. Gohlke H, Hendlich M, Klebe G (2000) J Mol Biol 295:337
32. Ruvinsky AM, Kozintsev AV (2005) Prot Str Func Bioinf 58:845
33. Bissantz C, Folkers G, Rognan D (2000) J Med Chem 43:4759
34. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Marrone T, Rose PW (2000) J Comput Aided Mol Des 14:731
35. Wang R, Lu Y, Wang S (2003) J Med Chem 46:2287
36. Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL III (2004) J Med Chem 47:3032
37. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) J Med Chem 49:5912
38. Leach AR, Shoichet BK, Peishoff CE (2006) J Med Chem 49:5851
39. Ajay, Murcko MA, Stouten PFW (1997) In: Charifson PS (eds) Practical application of computer-aided drug design. Marcel Dekker Inc, New York, pp 355–411
40. Murray CW, Verdonk ML (2002) J Comp-Aided Mol Des 16:741
41. Murray CW, Verdonk ML (2006) In: Jahnke W, Erlanson DA, Mannhold R, Kubinyi H, Folkers G (eds), Fragment-based approaches in drug discovery. WILEY-VCH Verlag GmbH & Co KGaA, Weinheim, pp 55-66
42. Pratt LR, Chandler D (1977) J Chem Phys 67:3683
43. Shoichet BK, Leach AR, Kuntz ID (1999) Prot Str Funct Bio 34:4
44. Steinberg IZ, Scheraga HA (1963) J Biol Chem 283:172
45. Erickson HP (1989) J Mol Biol 206:465
46. Finkelstein AV, Janin J (1989) Prot Eng 3:1
47. Ben-Tal N, Honig B, Bagdassarian CK, Ben-Shaul A (2000) Biophys J 79:1180
48. Yu YB, Privalov PL, Hodges RS (2001) Biophys J 81:1632
49. Ruvinsky AM, Kozincev AV (2005) J Comput Chem 26:1089
50. Ruvinsky AM (2007) J Comput Chem 28:1364
51. Page MI, Jencks WP (1971) Proc Natl Acad Sci USA 68:1678
52. Levitt M, Sander C, Stern PS (1985) J Mol Biol 181:423
53. Tidor B, Karplus M (1994) J Mol Biol 238:405
54. Doig AJ, Sternberg MJE (1995) Prot Sci 4:2247
55. Tirado-Rives J, Jorgensen WL (2006) J Med Chem 49:5880
56. Baker BM, Murphy KP (1996) Biophys J 71:2049
57. Bradshaw JM, Waksman G (1998) Biochem 37:15400
58. Karplus M, Janin J (1999) Prot Eng 12:185
59. Lazaridis T, Masunov A, Gandolfo F (2002) Prot Str Func Bioinf 47:194

60. Minh DDL, Bui JM, Chang CE, Jain T, Swanson JM, McCammon JA (2005) Biophys J 89:L25
61. Balog E, Becker T, Oettl M, Lechner R, Daniel R, Finney J, Smith JC (2004) Phys Rev Lett 93:028103
62. Lee SA, Rupprecht A, Chen YZ (1998) Phys Rev Lett 80:2241
63. Leach AR (ed) (1996) Molecular modelling: principles and applications. Longman, Harlow, p 382
64. Luo R, Gilson MK (2000) J Am Chem Soc 122:2934
65. Carlsson J, Åqvist J (2005) J Phys Chem B 109:6448
66. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Freer ST, Rose PW (2002) Curr Opin Struct Biol 12:197
67. Ruvinsky AM, Kozincev AV (2006) Prot Str Funct Bioinf 62:202
68. Källblad P, Mancera RL, Todorov NP (2004) J Med Chem 47:3334
69. Kortvelyesi T, Silberstein M, Dennis S, Vajda S (2003) J Comp Aided Mol Des 17:173
70. Rosenfeld RJ, Goodsell DS, Musah RA, Morris GM, Goodin DB, Olson AJ (2003) J Comp-Aid Mol Des 17:525
71. Tovchigrechko A, Vakser IA Prot Sci, 10 (2001) 1572; Ruvinsky AM, Vakser IA, submitted (2007)
72. Kozakov D, Clodfelter KH, Vajda S, Camacho CJ (2005) Biophys J 89:867
73. Wales DJ (2005) Phys Biol 2:S86
74. Wales DJ, Scheraga HA (1999) Science 285:1368
75. Price MLP, Jorgensen WL (2000) J Am Chem Soc 122:9455
76. Head MS, Given JA, Gilson MK (1997) J Phys Chem A 101:1609
77. Schaffer L, Verkhivker G (1998) Protein Struct Funct Genet 33:295
78. Straub JE (1996) In: Elber R (eds) New Developments in Theoretical Studies of Proteins. World Scientific, Singapore, pp 137
79. Frauenfelder H, Wolynes PG, Austin RH (1999) Rev Mod Phys 71:S419
80. Straub JE, Thirumalai D (1993) Proc Natl Acad Sci 90:809
81. Kitao A, Hayward S, Go N (1998) Prot Str Func Genet 33:496
82. Shortle D, Simons KT, Baker D (1998) Proc Natl Acad Sci USA 95:11158
83. Kubo R (ed) (1988) Statistical mechanics: an advanced course with problems and solutions. Elsevier Science Publishers BV, Amsterdam, p 200