OPEN ACCESS Freely available online

PLoS one

Estimating Absolute Configurational Entropies of Macromolecules: The Minimally Coupled Subspace Approach

Ulf Hensen[1], Oliver F. Lange[2]*, Helmut Grubmüller[1]

1 Department of Theoretical and Computational Biophysics, Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany, **2** Department of Biochemistry, University of Washington, Seattle, Washington, United States of America

Abstract

We develop a general minimally coupled subspace approach (MCSA) to compute absolute entropies of macromolecules, such as proteins, from computer generated canonical ensembles. Our approach overcomes limitations of current estimates such as the quasi-harmonic approximation which neglects non-linear and higher-order correlations as well as multi-minima characteristics of protein energy landscapes. Here, Full Correlation Analysis, adaptive kernel density estimation, and mutual information expansions are combined and high accuracy is demonstrated for a number of test systems ranging from alkanes to a 14 residue peptide. We further computed the configurational entropy for the full 67-residue cofactor of the TATA box binding protein illustrating that MCSA yields improved results also for large macromolecular systems.

Citation: Hensen U, Lange OF, Grubmüller H (2010) Estimating Absolute Configurational Entropies of Macromolecules: The Minimally Coupled Subspace Approach. PLoS ONE 5(2): e9179. doi:10.1371/journal.pone.0009179

Editor: Jörg Langowski, German Cancer Research Center, Germany

Received May 10, 2009; **Accepted** January 25, 2010; **Published** February 23, 2010

Funding: U.H. was supported by the Deutsche Forschungsgemeinschaft (research training group 782). O.F.L. was supported by the Human Frontiers of Science Program and by the Volkswagen Foundation, Grant I/80436. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: olange@u.washington.edu

Introduction

Entropies are key quantities in physics, chemistry, and biology. While free energy changes govern the direction of all chemical processes including reaction equilibria, entropy changes are the underlying driving forces of ligand binding, protein folding and other phenomena driven by hydrophobic effect. Traditionally calculating entropies from atomistic ensembles $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ of n configurations $\mathbf{x}_i\in\mathbb{R}^{3N}$ of a macromolecule of N atoms remains notoriously difficult.

We here propose and apply a method for calculating configurational entropies

$$S_c \sim -\int \rho(\mathbf{x}) \ln \rho(\mathbf{x}) d\mathbf{x}, \qquad (1)$$

where $\rho(\mathbf{x})$ denotes the configurational probability density $\rho(\mathbf{x}) = \exp(-\beta V(\mathbf{x}))/Z_c$ in the $3N$ dimensional configurational space governed by the potential energy $V(\mathbf{x})$ of the system. The fact that N is usually on the order of several hundreds or thousands renders the evaluation of this integral quite challenging despite a number of successful attempts. [1–4] These broadly fall into three classes, (i) special-purpose perturbation type approaches, also known as thermodynamic integration [5], (ii) step-by-step reconstruction methods, in particular the scanning procedures introduced by Meirovitch [6,7], (iii) direct approaches which analyse information readily available in standard equilibrium simulation trajectories [8–10].

While perturbation approaches provide relatively accurate free energy differences also for larger systems, accurate entropies are obtained only for smaller molecules. The main obstacle, which aggravates with system size, is the sampling problem, which severely limits the accuracy, in particular for explicit solvent models [2,5].

The most widely used direct method is the quasi-harmonic approximation [8] (QH), which provides an upper limit to the configurational entropy in terms of $3N$ independent classical or quantum mechanical harmonic oscillators [9,10], which is equivalent to approximating the configurational density $\rho(\mathbf{x})$ by a multi-variate Gaussian function,

$$\rho(\mathbf{x}) = (2\pi)^{-3N/2} \det \mathbf{A} \exp\left[-\frac{1}{2}(\mathbf{x}-\langle\mathbf{x}\rangle)^{\mathrm{T}}\mathbf{A}(\mathbf{x}-\langle\mathbf{x}\rangle)\right],$$

with $\mathbf{A}^{-1}=\mathbf{C}$ derived from the covariance matrix [9,10] $\mathbf{C}=\langle(\mathbf{x}-\langle\mathbf{x}\rangle)(\mathbf{x}-\langle\mathbf{x}\rangle)^{\mathrm{T}}\rangle$. However, for macromolecules undergoing large conformational motions the entropy is likely to be considerably smaller than this QH upper limit due to coupling and anharmonicities and, in particular, due to the existence of multiple conformational states [11–14]. Indeed, for smaller systems such as di-saccharides [15] or lipids [16], or small subsets of larger proteins [17] significantly lower entropies than with QH were obtained by inclusion of anharmonicities [11–13,18,19] and pairwise correlation of QH modes [20].

Results

The MCSA Scheme

Here we develop a direct method consisting of three building blocks. Results for small test systems will be presented during this

introduction of the methodology to illustrate the effect of each building block. Figure 1 shows that indeed for various small test systems (alkanes, dialanine and a complete 14-residue β-turn) the quasi-harmonic approximation severely overestimates the reference entropy. The reference values were obtained by thermodynamic integration (TI) gradually perturbing the systems towards an analytically tractable reference state consisting of non-interacting particles in harmonic wells, as described in methods and Refs. [21,22]. Entropy estimates obtained for all test systems are also summarized in Table 1.

Non-Parametric Density Estimation

As the first of the three building blocks of the methodology we recently introduced a non-parametric density estimation resting on adaptive anisotropic ellipsoidal kernels [21] that captures the configurational density in sufficient detail. Briefly, the configurational part of the entropy in a d-dimensional space is estimated from n configurations according to

$$S_c = \frac{k_B}{n}\sum_{i=1}^{n} \ln \frac{n Z_d(\mathbf{x}_i) r_{i,k}^d}{k(\mathbf{x}_i, r_{i,k})}, \qquad (2)$$

where $k(\mathbf{x}_i, r_{i,k}) = \langle K(\mathbf{x}_i, (\mathbf{x}_i - \mathbf{x})/r_{i,k})\rangle_{\mathbf{x}}$ denotes the ensemble average of an adaptive anisotropic kernel function K, whose anisotropy and scaling $r_{i,k}$ depends on the local density at point $\mathbf{x}_i$, and whose L_1-measure is denoted by $Z_d(\mathbf{x}_i)$. This formula simplifies to the well-known k-nearest neighbour entropy (k-NN) by fixing the kernel function to an (isotropic) sphere whose radius $r_{i,k}$ is chosen such that exactly k configurations are within the sphere centered at configuration $\mathbf{x}_i$. In this limiting case, Z_d is the volume of the d-dimensional unit sphere. NN estimators in general are entirely non-parametric and, at a finite sample size n, have minimal bias [23] in any given number of dimensions d. A major drawback, however, is the fact that due to the so-called 'curse of dimensionality' [24] simple k-NN estimators are applicable for up to ten dimensional configurational spaces only [25]. In contrast, as can be seen in Fig. 1 (left, "dir"-bar), adaptive anisotropic kernels yield accurate results even for the 45-dimensional configurational space of dialanine. For the more than 500-dimensional configurational space of the 14-residue β-turn, however, the 'curse of dimensionality' [24] renders it impossible to improve on the quasi-harmonic approximation with direct density estimation alone (Fig. 1 right). Convergence properties and full technical details of this first MCSA module are discussed in Ref. [21].

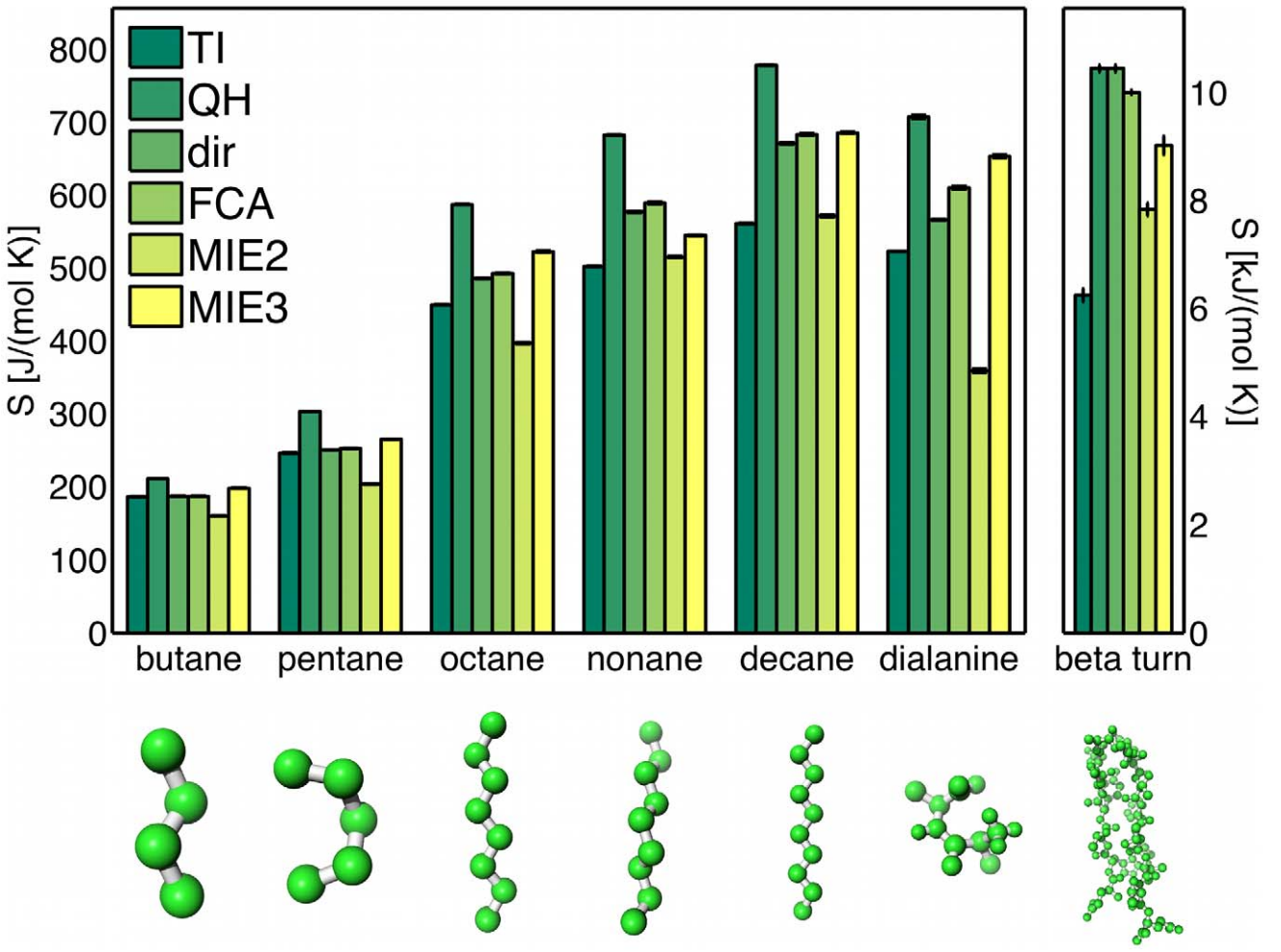

Figure 1. Entropy estimates for a set of small test systems. Five selected alkane systems, dialanine (left), and the C-terminal $\beta-$turn of Protein G (right, please note that here the units are kJ/(mol K)). Thermodynamic integration (TI), density estimates over the whole configurational space (dir), full correlation analyis with subsequent clustering and kernel density estimation (FCA), quasi-harmonic (QH) and mutual information expansion estimates of 2nd (MIE2) and 3rd (MIE3) order were obtained as described in the text.
doi:10.1371/journal.pone.0009179.g001

Table 1. Entropy estimates obtained for all systems.

System	N	S_{TI}	S_{dir}	S_{FCA}	S_{MIE2}	S_{MIE3}	*clust*	S_{QH}
Butane	4	185 ± 0.29	187 ± 0.11	187 ± 0.36	160 ± 0.24	197 ± 0.34	5	211 ± 0.18
Pentane	5	245 ± 0.30	251 ± 0.17	252 ± 0.69	203 ± 0.44	265 ± 0.25	8	303 ± 0.08
Hexane	6	307 ± 0.68	319 ± 0.21	323 ± 0.40	244 ± 0.55	383 ± 1.15	11	395 ± 0.17
Heptane	7	388 ± 0.92	399 ± 0.34	407 ± 0.33	317 ± 1.26	484 ± 1.58	13	492 ± 0.17
Octane	8	450 ± 0.48	485 ± 0.67	492 ± 0.59	397 ± 1.13	522 ± 1.15	15	587 ± 0.07
Nonane	9	502 ± 0.46	577 ± 0.88	589 ± 1.8	515 ± 0.95	544 ± 0.88	19	682 ± 0.14
Decane	10	564 ± 0.75	670 ± 1.10	683 ± 1.3	571 ± 1.57	685 ± 0.88	21	778 ± 0.13
Dialanine	15	524 ± 1.1	566 ± 0.4	610 ± 2.2	359 ± 2.67	653 ± 2.23	32	707 ± 2.1
β-turn	169	6246 ± 119	10446 ± 66	10002 ± 42	7834 ± 123	9018 ± 174	84–108	10446 ± 66
TBP cofactor	696	–	–	22250 ± 58	21543 ± 152	21853 ± 93	32–88	23226 ± 88
TBP complex	696	–	–	24918 ± 229	24371 ± 392	24514 ± 500	56–80	25880 ± 197

Alkane test systems butane to decane, dialanine, the 14-residue β-turn, as well as free and complexed TATA box binding protein (TBP) cofactor. S_{TI}: absolute configurational entropy obtained by TI (in J/(mol K)); S_{dir}: direct density estimate without clustering; S_{FCA}: sum of density estimates after subspace clustering; S_{MIE2} and S_{MIE3} : Mutual information expansion estimates of 2nd (MIE2) and 3rd order (MIE3); *clust*: size of largest cluster; S_{QH}: QH entropy estimate.
doi:10.1371/journal.pone.0009179.t001

Generation of Minimally Coupled Subspaces

As the second building block of our method, we apply an entropy invariant transformation $\mathbf{T}$ such that the usually highly coupled degrees of freedom separate into optimally uncoupled subspaces, each of which being sufficiently low-dimensional to render non-parametric density estimation applicable. As the most straightforward class of entropy invariant transformations, we consider here linear orthonormal transformations of the form $\mathbf{y}=\mathbf{T}(\mathbf{x}-\langle\mathbf{x}\rangle)$, with $\det\,\mathbf{T}=1$. More general transformations are currently explored [26]. We apply Full Correlation Analysis (FCA) [27] which minimizes mutual information by considering

$$H[\mathbf{T}]=-\frac{k_B}{\ell}\sum_{i=1}^{3N}\int\rho_i^{(1)}(y_i)\ln\rho_i^{(1)}(y_i),$$

where y_i denote the components of $\mathbf{y}$ and $\rho_i^{(1)}(y_i)=\ell^{3N-1}\int\rho(\mathbf{y})dy_{j\neq i}$ the 1-dimensional marginal density along y_i. This procedure minimizes non-linear correlations of second and higher order [27] and therefore generalizes the principal component analysis (PCA) which only considers linear correlations of second order. For complex macromolecules, however, even for the optimal linear FCA transformation $\mathbf{T}$, considerable non-linear correlations between several degrees of freedom will remain and cannot be neglected. To address this issue, the FCA modes are subsequently clustered according to the generalized correlation coefficient [25,28]

$$r_{\mathrm{MI},ij}=\left(1-\exp\left[-2I_{i,j}^{(2)}\right]\right)^{1/2},$$

with the mutual information

$$\begin{aligned}I_{i,j}^{(2)}&=H_i^{(1)}[\mathbf{T}]+H_j^{(1)}[\mathbf{T}]-H_{i,j}^{(2)}[\mathbf{T}]\\&=-\frac{k_B}{\ell}\int\rho_{i,j}^{(2)}(y_i,y_j)\ln\frac{\rho_{i,j}^{(2)}(y_i,y_j)}{\rho_i^{(1)}(y_i)\rho_j^{(1)}(y_j)}\end{aligned}$$

between components y_i and y_j. This is achieved by assigning mode indices j to m clusters C_s such that all modes with correlation coefficients larger than a certain threshold θ are assigned to the same cluster. This disjoint clustering defines an approximate factorization $\rho(\mathbf{y})\approx\Pi_{s=1}^{m}\rho_s^{(d_s)}\left(\otimes_{j\in C_s}y_j\right)$, where $\rho_s^{(d_s)}$ denotes the generalized d_s-dimensional marginal density along $\otimes_{j\in C_s}y_j$. This factorization is approximate in the sense that for the entropy

$$S[\rho(\mathbf{y})]=\sum_{s=1}^{m}S\left[\rho_s^{(d_s)}\left(\bigotimes_{j\in C_s}y_j\right)\right]+S_{\mathrm{res}}\left[\{C_s\}_{s=1,\ldots,m}\right]\qquad(3)$$

the residual entropy $S_{\mathrm{res}}\left[\{C_s\}_{s=1,\ldots,m}\right]$ is small.

Such approximate factorization, of course, neglects all inter-cluster correlations. These can be pairwise correlations, and thus are small ($<\theta$) by construction, or higher-order correlations. For the latter we have to assume that they are also effectively eliminated by our threshold criterion. This assumption is supported by the observation that for the alkanes and for dialanine, with $\theta=0.025$, $S_{\mathrm{dir}}\approx S_{\mathrm{FCA}}$ (cf. Fig. 1). Thus, our factorization yields accurate entropies and S_{res} is indeed small.

Mutual Information Expansions for Oversized Clusters

However, for the larger molecules considered here, the necessarily small threshold typically results in at least one cluster being too large for a sufficiently accurate density estimate (e.g., for the β-turn $d_1=108$). Accordingly, while our factorization still improves the entropy estimate (cf. Fig. 1), S_{res} cannot be neglected anymore. The third building block of our method addresses this issue by subdividing each oversized cluster into h_s disjoint subclusters $D_i^{(s)}$ of sizes $d_1^s,\ldots,d_{h_s}^s<15$, $C_s=\bigcup_{i=1}^{h_s}D_i^{(s)}$, irrespective of the necessarily remaining strong correlations between these. The residual entropy contributions to the configurational entropy

$$\begin{aligned}S[\rho(\mathbf{y})]=&\sum_{s=1}^{m}\sum_{a=1}^{h_S}S\left[\rho_s^{(d_s)}\left(\bigotimes_{j\in D_a^{(s)}}y_j\right)\right]\\&+\sum_{s=1}^{m}S_{\mathrm{res}}\left[\left\{D_a^{(s)}\right\}_{a=1,\ldots,h_s}\right]+S_{\mathrm{res}}\left[\{C_s\}_{s=1,\ldots,m}\right]\end{aligned}$$

will be drastically increased due to non-neglegible intra-cluster

contributions $S_{\rm res}\left[\left\{D_a^{(s)}\right\}_{a=1,\ldots,h_s}\right]$ from all subdivided clusters C_s, where we have omitted the argument ρ in the rightmost two terms for brevity. We here propose to compute each $S_{\rm res}\left[\left\{D_a^{(s)}\right\}_{a=1,\ldots,h_s}\right]$ via the mutual information expansion (MIE) as

$$S_{\rm res}\left[\left\{D_a^{(s)}\right\}_{a=1,\ldots,h_s}\right] = -\sum_{a<b} I_2^{(d_a^s+d_b^s)}[\rho_a,\rho_b] + \sum_{a<b<c} I_3^{(d_a^s+d_b^s+d_c^s)}[\rho_a,\rho_b,\rho_c] - \ldots (-1)^{h_s+1} I_{h_s}[\rho_a,\ldots,\rho_{h_s}], \quad (4)$$

where $\rho_a \equiv \rho^{(da)}\left(\otimes_{j\in D_a^{(s)}} y_j\right)$. Expanding the mutual information terms

$$I_k^{(\sum_1^k da)}[\rho_1,\ldots,\rho_{h_s}] = \sum_{a=1}^{k}(-1)^{a+1}\sum_{i_1<\ldots<i_a} S[\rho_{i_1},\ldots,\rho_{i_a}], \quad (5)$$

up to second or third order, respectively, with the right-hand sum running over all possible permutations $\{i_1,\ldots,i_a\}\in\{1,\ldots,k\}$, has proven sufficiently accurate in liquid state theory [29] and information theory [30,31]. Indeed, for the β-turn, inclusion of the remaining correlations via this expansion improved the entropy estimate (Fig. 1). For the other test systems $S_{\rm dir}\approx S_{\rm FCA}\approx S_{\rm MIE3}$. In contrast, for some of the test systems $S_{\rm MIE2}<S_{\rm TI}$, such that from our observations, 3rd order MIE provides a better estimate and an upper bound to the true entropy.

Applications of MIE to macro-molecular systems can be hampered by the curse of dimensionality and combinatorial explosion of the number of terms [32,33]. In this work, the problem is circumvented by clustering into sufficiently high-dimensional ($\sim$15) subspaces which minimizes residual inter-D_a correlations and delays the onset of the combinatorial explosion. At the same time the subspaces are sufficiently small that even for the 3rd-order MIE no direct density estimates beyond the critical dimensionality of $d_s=45$ are required.

TATA Box Binding Protein: Protein Test Case and Error Estimate

Together, these three building blocks enable one to calculate configurational entropies even for larger biomolecules. We considered the 67-residue TATA box binding protein (TBP, pdb code 1TBA) inhibitor in two different configurations; complexed (Fig. 2 top left) and free (Fig. 2 top right). To estimate the statistical error of MCSA and QH configurational entropy estimates, for both states five independent molecular dynamics (MD) simulations were carried out using the OPLS force-field [34] and the TIP4P explicit solvent model [35] (see methods section for full simulation details). Fig. 2 shows the results obtained by the five entropy estimation methods for both complexed (left) and free (right) inhibitor. All methods estimate the free cofactor's entropy to be significantly higher than that of the bound cofactor. As can be seen, for both complexed and free cofactor, QH yields the largest estimate. The first two MCSA modules combined (kernel density estimation on little correlated configurational subspaces obtained from FCA) already yield remarkably smaller estimates, irrespective of whether a high or a low clustering threshold θ was chosen (hi thresh and low thresh in Fig. 2), i.e., chosing small but higher correlated subspaces or larger but lowly correlated subspaces provides similar estimates. Finally, employing all the three MCSA modules including MIE of 2nd (MIE2) and 3rd (MIE3) lowered the estimate again with, as before, the 2nd-order estimate being lower than the 3rd-order estimate.

The fact that the QH estimate is the largest in all cases corroborates the observations for the small test cases, and generally shows that MCSA yields improved estimates also for large macromolecules. Already the first two MCSA modules provide lower entropy estimates, even though relatively large configurational subspaces ($d_s=35\ldots88$, see Table 1) were obtained from FCA, which illustrates that indeed our kernel density estimator works accurately also for the complex high-dimensional configurational spaces spanned by proteins. Further, the fact that the clustering threshold did not affect the final estimate very much naturally reflects the fact that clustering with a high threshold yields small subspaces C_s which are correlated, such that $S_{\rm res}\left[\{C_s\}_{s=1,\ldots,m}\right]$ in Eq. 3 is large, increasing our estimate $S[\rho(\mathbf{y})]$. On the other hand, clustering with a small threshold gives rise to a small $S_{\rm res}\left[\{C_s\}_{s=1,\ldots,m}\right]$ but sparse sampling due to large d_s then entails higher $S\left[\rho_s^{(d_s)}\left(\otimes_{j\in C_s} y_j\right)\right]$, such that $S[\rho(\mathbf{y})]$ is also increased in this case. As expected, the third MCSA module, MIE, circumvents this problem and lowers the MCSA estimate further by 404 or 397 J/(molK) for the free and the complexed cofactor, respectively. The 2nd-order estimate is lower than the 3rd-order estimate in all cases, which shows that also for proteins the pair correlations are generally overestimated, and inclusion of 3rd-order correlations is indeed crucial.

The statistical errors are relatively small in all cases, but generally twice as large for the free than for the complexed cofactor. We attribute this observation to the larger inherent flexibility of the free state, and hence to insufficient molecular dynamics sampling. Consequently, the MIE error for the free cofactor is over three times larger than that of the the complex. Interestingly, the MIE estimate is slightly more affected with the error for the free cofactor being three- to fourfold as high as for the complex. Due to the high number of terms to be evaluated for the MIEs (Eq. 5), already small errors of each $S[\rho_{i_1},\ldots,\rho_{i_a}]$ result in relatively large errors in $S_{\rm res}\left[\left\{D_a^{(s)}\right\}_{a=1,\ldots,h_s}\right]$.

Discussion

We have developed a minimally coupled subspace approach (MCSA) to estimate absolute macromolecular configurational entropies from structure ensembles which takes anharmonicities and higher-order correlations into account. The approach combines three building blocks which together allow one to calculate absolute entropies even for the highly complex configurational densities generated by the dynamics of biological macromolecules such as proteins. MCSA shares the versatility of the quasi-harmonic approach as it can be applied to unperturbed equilibrium trajectories while achieving the accuracy of special-purpose perturbation type methods. The effective dimension reduction provided by the Full Correlation Analysis allows for the application of mutual information expansions to large macromolecules. Further, the adaptive kernel non-parametric density estimation method developed for MCSA requires much weaker a-priori assumptions about the properties of the configurational densities than (quasi-)harmonic approaches. The method is applicable also to large macromolecules such as proteins. In this study, we showed that MCSA applied to the TATA box binding protein yielded significantly smaller and thus improved entropy estimates.

We note that here we focus at configurational entropies of the solute only, thus missing both the solvent as well as the solvent/

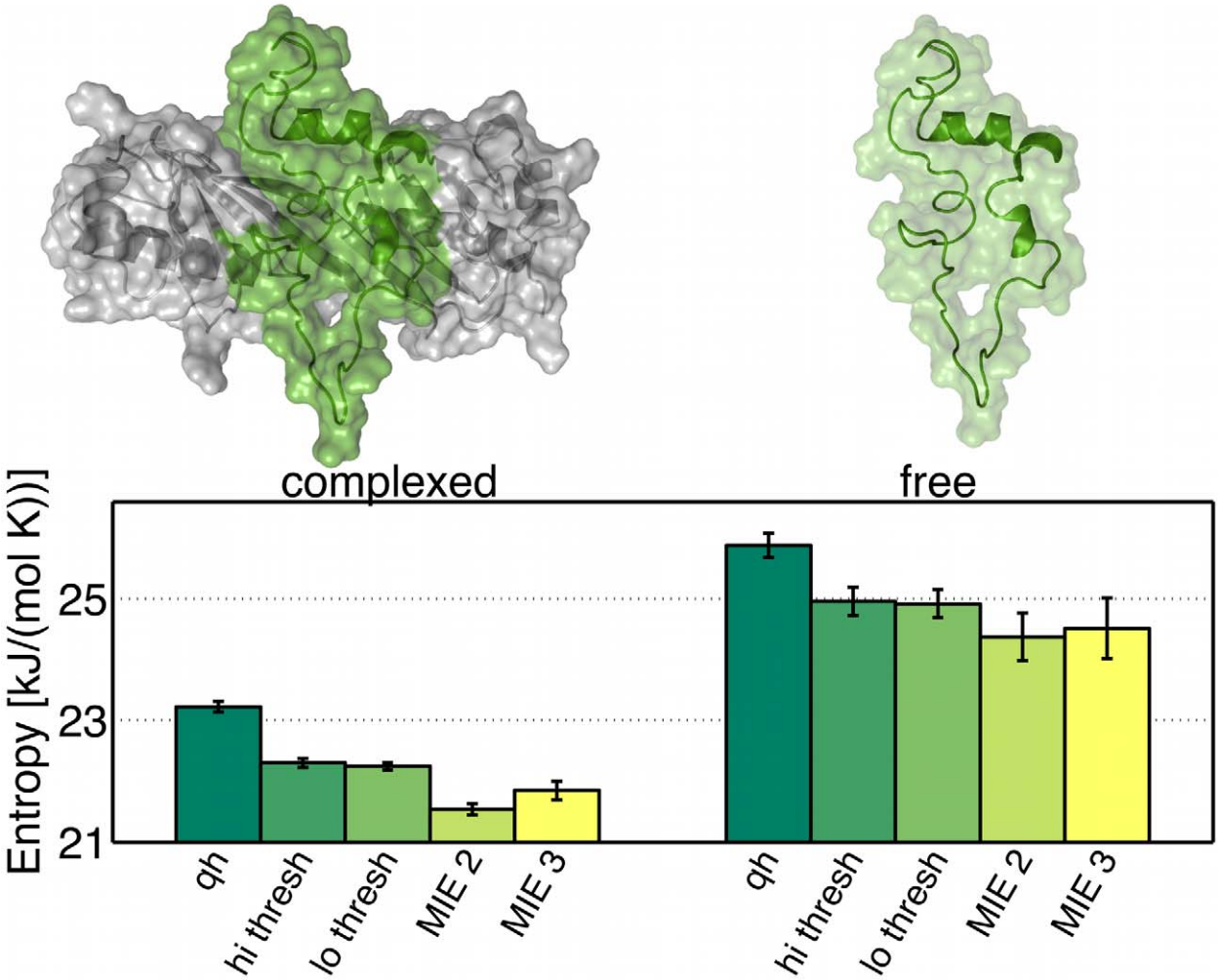

Figure 2. Entropy estimates for the TATA box binding protein (TBP) inhibitor in complex (left) and free (right). The following techniques are used: quasi-harmonic approximation (QH); FCA with subsequent density estimation using a high clustering threshold θ (hi thresh) or, respectively, a low threshold (lo thresh); mutual information expansion of order 2 (MIE2) or, respectively, of order 3 (MIE3). The displayed entropy estimates are averages over five independent simulations of 100 ns each, the error bars indicate standard deviations of the mean.
doi:10.1371/journal.pone.0009179.g002

solute parts. Using permutation reduction techniques [36], our method should be capable of capturing also these important contributions, which however lies outside the scope of the present work.

Methods

Thermodynamic Integration Reference Entropy

Absolute free energies for the test systems butane to decane, dialanine, and the ProteinG β-turn were calculated by thermodynamic integration (TI). Simulation parameters cf. below. The TI scheme we have chosen to obtain the Helmholtz free energy A of the fully interacting particles consists of two phases. Harmonic position restraints with a force constant $k = 25000\ \mathrm{kJ/(mol\ nm^2)}$ were slowly switched on for each atom in the first phase, and in the second phase all force-field components were gradually switched off. Within the second phase, the charges were switched off prior to the rest of the force field. After the second phase, the system consisted of non-interacting dummy particles with mass m oscillating in their respective harmonic position restraint potentials, i.e.,

$$V = \frac{1}{2} k \sum_{j=1}^{N} (\mathbf{x} - \mathbf{x}_j)^2 .$$

The free energy of this harmonic system can be obtained analytically,

$$A_0 = -\beta^{-1} \frac{3}{2} \sum_{j=1}^{N} \left[\ln \left(\frac{1}{\hbar^2 \beta^2 k_j} \right) \right]$$

where $k_j = \tilde{k}_j / m_j$ denotes the mass-weighted force constant. Hence, the thermodynamic integration yields the absolute free energy

$$A = A_0 - \Delta A_2 - \Delta A_1$$

and the entropy by $S = (A - \langle V \rangle)/T$, where $\langle V \rangle$ denotes the ensemble average of the potential energy.

For the TI between the systems given by V_s (start) and V_f (end), 21 intermediate steps $V_i(\lambda) = \lambda V_s + (1-\lambda) V_f$, $i = 1, \ldots, 21$ were

used, and the intermediate values of $\lambda_i = 0$, 1e-6, 5e-6, 1e-5, 5e-4, 1e-4, 1e-3, 1e-2, 2e-2, 3e-2, 5e-2, 7e-2, 9e-2, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1 were distributed unevenly to obtain approximately balanced ΔA_i values. For each value of λ a trajectory of 12.5 ns (alkanes and dialanine) or 125 ns (β-turn), respectively, was generated.

The error estimates of the TI reference entropies detailed in Table 1 were obtained via two ways for the alkane test systems and dialanine. First, by averaging over five independent simulations and, second, by performing blockwise averaging as derived in Ref. [37] over each of the 23 $V_i(\lambda)$ of each of these five trajectories. We found that the error estimates obtained by these two methods agree very well. Accordingly, for the β-turn only the block averaging method was applied and the resulting error estimates are also given in Table 1.

Molecular/Stochastic Dynamics Simulations

The test systems that were compared with a thermodynamic integration reference (butane to decane, dialanine, and the ProteinG β-turn) were set up as follows. Force-field parameterizations were obtained from the Dundee Prodrug server [38] based on the GROMOS united-atom force field [39]. Stochastic Dynamics simulations were performed using the molecular simulations package GROMACS [40] in vacuo at 400 K with friction constant γ set to 10, dielectric constant $\varepsilon = 1$, integration step size of 0.0005 ps and no bond constraints. Positional restraints were applied to three adjacent terminal heavy atoms. To obtain MCSA error estimates, each of the simulations was carried out five times using different starting velocities. MCSA and QH entropy estimates were obtained from trajectories of lengths 12.5 ns (alkanes and dialanine) or 125 ns (β-turn), respectively, i.e. the TI entropy references required 23 times as much computing time as MCSA and QH estimates.

The TATA box binding protein (TBP) complex (protein database entry 1TBA) was simulated using the OPLS all atom force field [34] in explicit TIP4P solvent [35] and periodic boundary conditions. NpT ensembles were simulated, with the protein and solvent coupled separately to a 300-K heat bath ($\tau = 0.1$ ps). [41] The systems were isotropically coupled to a pressure bath at 1 bar ($\tau = 1.0$ ps) [41]. Application of the Lincs [42] and Settle [43] algorithms allowed for an integration time step of 2 fs. Short-range electrostatics and Lennard–Jones interactions were calculated within a cut-off of 1.0 nm, and the neighbour list was updated every 10 steps. The particle mesh Ewald (PME) method was used for the long-range electrostatic interactions [44], with a grid spacing of 0.12 nm. The free cofactor was simulated using the same parameters as above. The starting structure was obtained by removing the TBP from the X-ray structure of the complex and equilibrating for 2 ns. Entropy estimates and corresponding errors for both complexed and free cofactor were obtained from five trajectories of 200 ns length each.

Mutual Information Expansions Implementation Details

Fill modes. Due to the moderate regularization assumptions, our adaptive kernel density estimator is sensitive to the sparse sampling problem whose effect is highly dependent on the dimensionality. To guarantee the same accuracy of all density estimates required for the computation of the correlation terms I_n of Eq. 5 despite different dimensionality it is, thus, necessary to ensure the same local densities around points $\mathbf{y}_i$ in different terms. This is normally not provided. The mutual information between two modes y_i and y_j,

$$I_2 = \int_{i,j} \rho(y_i,y_j) \ln \frac{\rho(y_i,y_j)}{\rho(y_i)\rho(y_j)}, \tag{6}$$

contains differently well sampled terms in denominator and numerator, because the number of sampling points available to estimate $\rho(y_i,y_j)$ is only half the number of sampling points available for estimating the marginal densities $\rho(y_i)$ and $\rho(y_j)$ (see Fig. 3). The accuracy for the estimation of the marginal densities is, consequently, possibly higher than the joint estimate yielding an inaccurate correlation estimate. To overcome this problem, we devised the concept of fill modes. Accordingly, artificially decorrelated modes $y_i' : \{y'_{i,1}, \ldots, y'_{i,3N}\} = \mathrm{perm}\{y_{i,1}, \ldots, y_{i,3N}\}$ are created by permuting its components $y_{i,j}$, with $1 \le j \le 3N$. The marginal densities $\rho(y_i') = \rho(y_i)$ and $\rho(y_i',y_j) = \rho(y_i)\rho(y_j)$, yielding a new expression for Eq. 6,

$$I_2 = \int_{i,j} \rho(y_i,y_j) \ln \frac{\rho(y_i,y_j)}{\rho(y_i',y_j)}, \tag{7}$$

where the product of the marginal densities $\rho(y_i)$ and $\rho(y_j)$ is now computed from the synthetically decorrelated joint distribution $\rho(y_i',y_j)$, such that the same accuracy for the joint estimate is guaranteed as for the marginal estimates. Conducting this scheme on the 3rd order correlation function of three modes y_i, y_j and y_k,

$$I_3 = \int_{i<j<k} \rho(y_i,y_j,y_k) \ln \frac{\rho(y_i,y_j,y_k)}{\frac{\rho(y_i,y_j)\rho(y_i,y_k)\rho(y_j,y_k)}{\rho(y_i)\rho(y_j)\rho(y_k)}},$$

yields

$$I_3 = \int_{i<j<k} \rho(y_i,y_j,y_k) \ln \frac{\rho(y_i,y_j,y_k)}{\frac{\rho(y_i,y_j,y_k')\rho(y_i,y_k,y_j')\rho(y_j,y_k,y_i')}{\rho(y_i',y_j',y_k')^2}}, \tag{8}$$

where the pairwise joint distributions have been 'filled up' with permuted 'fill modes', as described above, e.g. $\rho(y_i,y_j) = \rho(y_i,y_j,y_k')/\rho(y_k')$.

Consistent dimensions. The sensitivity of the nearest-neigbour estimates, Eq. 2, towards the sparse sampling problem also affects the different terms of Eq. 5, which inevitably suffer from different sparse sampling problems if computed separately. Furthermore, a huge number of probability density distributions $\rho(y_i),\rho(y_i,y_j),\ldots,\rho(y_i,y_j,\ldots,y_k)$ is computed more than once for the many instances of identical correlation terms appearing in that equation. Expanding over entropy terms rather than correlation terms, in contrast, yields

$$S[\rho(y_1,\ldots,y_n)] = \sum_{k=1}^{t} \sum_{m_1<\ldots<m_k} \vartheta_{k,t} S[\rho(y_1,\ldots,y_k)], \tag{9}$$

where the first summation runs over different orders $k = 1,\ldots,t$ until truncation order $t \le n$. $\vartheta_{k,t} = \sum_{i=k}^{t} (-1)^{i+k} \binom{n-k}{i-k}$ designates how many times a certain order appears and whether it needs to be added or subtracted, and the second sum over all $\binom{n}{k}$ possible combinations $\{m_1,\ldots,m_k\} \in \{1,\ldots,n\}$. To guarantee the same estimation accuracy for all $\rho(y_1,\ldots,y_k)$ of Eq. 9, each term is filled up to truncation order t yielding

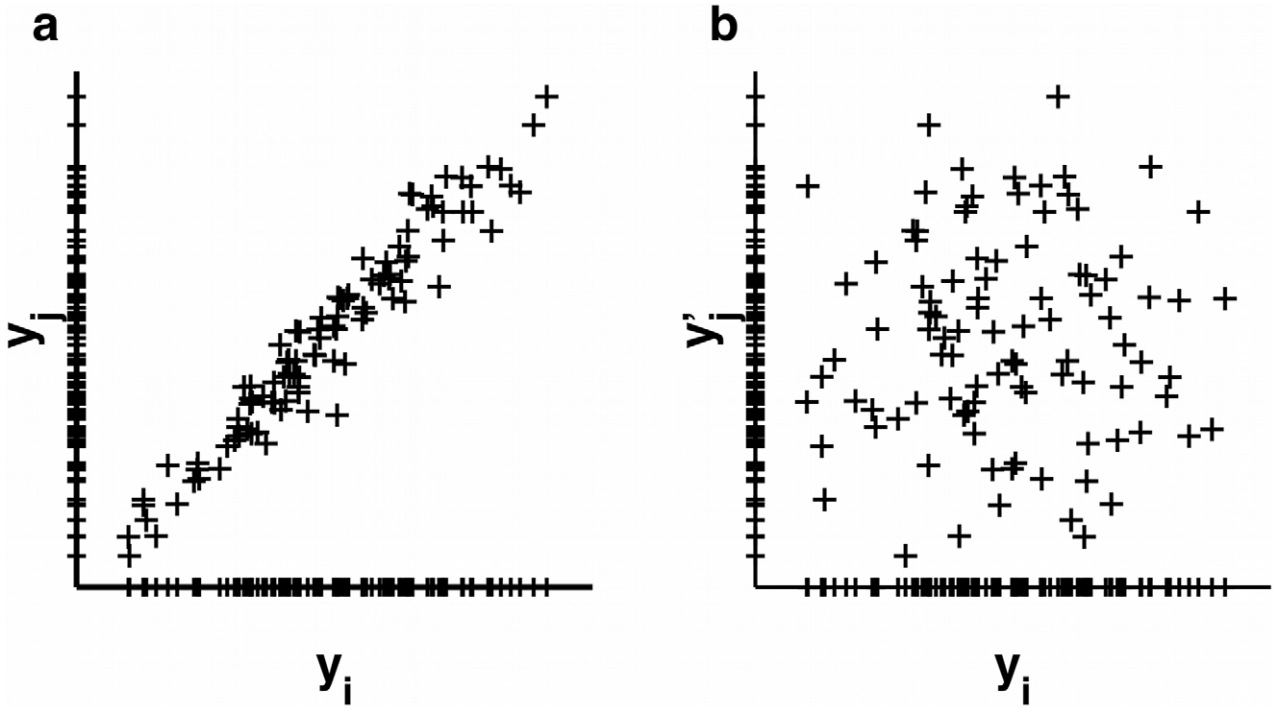

Figure 3. Principle of fill modes. a) Two arbitrarily correlated modes y_i and y_j marginally distributed on the axes. Correlation is clearly visible from the y_j-distributed y_i. The joint distribution $\rho(y_i,y_j)$ is more sparsely sampled than both marginal distributions. b) The y_j'-distributed y_i is decorrelated and has exactly as many sample points as the joint distribution in a), allowing precise computation of $I_2(y_i,y_j)$.
doi:10.1371/journal.pone.0009179.g003

$\rho(y_1,\ldots,y_k,y_{k+1}',\ldots,y_t')$. Under this modification, Eq. 9 reads

$$S[\rho(y_1,\ldots,y_n)] = \underbrace{\vartheta'_{1,t}\sum_{m_1,\ldots,m_t} S[\rho(y_1',\ldots,y_t')]}_{\text{marginal entropies/fill modes}} + \sum_{k=2}^{t}\sum_{m_1<\ldots<m_k}\vartheta_{k,t}S[\rho(y_1,\ldots,y_k,y_{k+1}',\ldots,y_t')], \tag{10}$$

with the number of marginal entropies,

$$\vartheta'_{1,t} = \underbrace{\sum_{i=1}^{t}(-1)^{i+1}\binom{n-1}{i-1}}_{\text{normal first-order indexing}} - \underbrace{\sum_{i=2}^{t-1}\vartheta'_t\frac{\binom{n}{i}\binom{n-i}{t-i}(t-i)}{n}}_{\text{fill modes}},$$

which depends on the fill mode weighting index

$$\vartheta'_t = \sum_{k=2}^{t}\sum_{i=k}^{t}(-1)^{i+k}\binom{n-k}{i-k},$$

where, like above, primes indicate permuted entries.

Author Contributions

Conceived and designed the experiments: UH OFL HG. Performed the experiments: UH OFL. Analyzed the data: UH OFL. Contributed reagents/materials/analysis tools: UH OFL. Wrote the paper: UH OFL HG.

References

1. Beveridge DL, DiCapua FM (1989) Free energy via molecular simulation: Applications to chemical and biomolecular systems. Annual Review of Biophysics and Biophysical Chemistry 18: 431–492.
2. Straatsma TP, McCammon JA (1992) Computational alchemy. Annual Review of Physical Chemistry 43: 407–435.
3. Kollman P (1993) Free energy calculations: Applications to chemical and biochemical phenomena. Chem Rev 93: 2395–2417.
4. Meirovitch H (2007) Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. Curr Opin Struct Biol 17: 181–186.
5. Peter C, Oostenbrink C, van Dorp A, van Gunsteren WF (2004) Estimating entropies from molecular dynamics simulations. J Chem Phys 120: 2652–2661.
6. Cheluvaraja S, Meirovitch H (2004) Simulation method for calculating the entropy and free energy of peptides and proteins. PNAS 101: 9241–9246.
7. Cheluvaraja S, Meirovitch H (2006) Calculation of the entropy and free energy of peptides by molecular dynamics simulations using the hypothetical scanning molecular dynamics method. J Chem Phys 125: 024905.
8. Karplus M, Kushick JN (1981) Method for estimating the configurational entropy of macromolecules. Macromolecules 14: 325–332.
9. Schlitter J (1993) Estimation of absolute and relative entropies of macromolecules using the covariance matrix. Chemical Physics Letters 215: 617–621.
10. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Mol Biol 9: 646–652.
11. Chang C, Chen W, Gilson M (2005) Evaluating the accuracy of the quasiharmonic approximation. J Chem Theory Comput 1: 1017–1028.
12. Chang C, Chen W, Gilson MK (2007) Ligand configurational entropy and protein binding. PNAS 104: 1534–1539.
13. Gilson MK, Zhou HX (2007) Calculation of protein-ligand binding affinities. Ann Rev Biophys Biomol Struct 36: 21–42.

14. Minh DDL, Bui JM, Chang C, Jain T, Swanson JMJ, et al. (2005) The entropic cost of protein-protein association: A case study on acetylcholinesterase binding to fasciculin-2. Biophys J 89: 25–27.
15. Pereira CS, Kony D, Baron R, Müller M, van Gunsteren WF, et al. (2006) Conformational and dynamical properties of disaccharides in water: a molecular dynamics study. Biophysical Journal 90: 4337–4344.
16. Baron R, deVries A, Hünenberger P, van Gunsteren W (2006) Comparison of atomic-level and coarse-grained models for liquid hydrocarbons from molecular dynamics configurational entropy estimates. J Phys Chem B 110: 8464–8473.
17. Baron R, McCammon JA (2008) (thermo)dynamic role of receptor flexibility, entropy, and motional correlation in protein-ligand binding. ChemPhysChem 9: 983–988.
18. Kolossvary I (1997) Evaluation of the molecular configuration integral in all degrees of freedom for the direct calculation of conformational free energies: Prediction of the anomeric free energy of monosaccharides. J Phys Chem A 101: 9900–9905.
19. Chang C, Potter M, Gilson M (2003) Calculation of molecular configuration integrals. J Phys Chem B 107: 1048–1055.
20. Baron R, van Gunsteren W, Hünenberger P (2006) Estimating the configurational entropy from molecular dynamics simulations: anharmonicity and correlation corrections to the quasi-harmonic approximation. Trends Phys Chem 11: 87–122.
21. Hensen U, Grubmüller H, Lange OF (2009) Adaptive anisotropic kernels for nonparametric estimation of absolute configurational entropies in high-dimensional configuration spaces. Phys Rev E 80: 011913.
22. Tyka M, Clarke A, Sessions R (2006) An efficient, path-independent method for free-energy calculations. J Phys Chem B 110: 17212–17220.
23. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E 69: 066138.
24. Bellman RE (1961) Adaptive Control Processes Princeton University Press.
25. Hnizdo V, Darian E, Fedorowicz A, Demchuk E, Li S, et al. (2007) Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. J Comp Chem 28: 655–668.
26. Hennig M (2007) Entropy invariant transformations. Master's thesis, Universität Jena.
27. Lange OF, Grubmüller H (2008) Full correlation analysis of conformational protein dynamics. Proteins 70: 1294–1312.
28. Lange OF, Grubmüller H (2006) Generalized correlation for biomolecular dynamics. Proteins 62: 1053–1061.
29. Baranyai A, Evans DJ (1989) Direct entropy calculation from computer simulation of liquids. Phys Rev A 40: 3817–3822.
30. Attard P, Jepps OG, Marčelja S (1997) Information content of signals using correlation function expansions of the entropy. Phys Rev E 56: 4052–4067.
31. Attard P (1999) Statistical Physics on the Eve of the Twenty-First Century, World Scientific, chapter Markov Superposition Expansion for the Entropy and Correlation Functions in Two and Three Dimensions.
32. Killian BJ, Kravitz JY, Gilson MK (2007) Extraction of configurational entropy from molecular simulations via an expansion approximation. J Chem Phys 127: 024107.
33. Hnizdo V, Tan J, Killian BJ, Gilson MK (2008) Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. J Comp Chem 29: 1605–1614.
34. Kaminski G, Friesner R, Tirado-Rives J, Jorgensen W (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105: 6474–6487.
35. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79: 926–935.
36. Reinhard F, Grubmüller H (2007) Estimation of absolute solvent and solvation shell entropies via permutation reduction. J Chem Phys 126: 014102.
37. Hess B (2002) Determining the shear viscosity of model liquids from molecular dynamics simulations. J Chem Phys 116: 209–217.
38. Schüttelkopf AW, van Aalten DMF (2004) PRODRG - a tool for high-throughput crystallography of protein-ligand complexes. Acta Crystallographica D 60: 1355–1363.
39. van Gunsteren WF, Daura X, Mark AE (1998) GROMOS force field. Encyclopaedia of computational chemistry edition. pp 1211–1216.
40. van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, et al. (2005) Gromacs: Fast, flexible, and free. J Comp Chem 26: 1701–1718.
41. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81: 3684–3690.
42. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) Lincs: A linear constraint solver for molecular simulations. Journal of Computational Chemistry 18: 1463–1472.
43. Miyamoto S, Kollman PA (1992) Settle: An analytical version of the shake and rattle algorithm for rigid water models. J Comp Chem 13: 952–962.
44. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. J Chem Phys 98: 10089–10092.