Determining Equilibrium Constants for Dimerization Reactions from Molecular Dynamics Simulations

DJURRE H. DE JONG,[1,2*] LARS V. SCHÄFER,[1,2*] ALEX H. DE VRIES,[1,2] SIEWERT J. MARRINK,[1,2]
HERMAN J. C. BERENDSEN,[1,2] HELMUT GRUBMÜLLER[3]

[1]*Groningen Biomolecular Science and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747 AG, Groningen, The Netherlands*

[2]*Zernike Institute for Advanced Materials, University of Groningen, Nijenhorgh 4, NL-9747 AG Groningen, The Netherlands*

[3]*Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany*

Received 19 November 2010; Accepted 26 January 2011
DOI 10.1002/jcc.21776
Published online 5 April 2011 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: With today's available computer power, free energy calculations from equilibrium molecular dynamics simulations "via counting" become feasible for an increasing number of reactions. An example is the dimerization reaction of transmembrane alpha-helices. If an extended simulation of the two helices covers sufficiently many dimerization and dissociation events, their binding free energy is readily derived from the fraction of time during which the two helices are observed in dimeric form. Exactly how the correct value for the free energy is to be calculated, however, is unclear, and indeed several different and contradictory approaches have been used. In particular, results obtained via Boltzmann statistics differ from those determined via the law of mass action. Here, we develop a theory that resolves this discrepancy. We show that for simulation systems containing two molecules, the dimerization free energy is given by a formula of the form $\Delta G \propto \ln(P_1/P_0)$. Our theory is also applicable to high concentrations that typically have to be used in molecular dynamics simulations to keep the simulation system small, where the textbook dilute approximations fail. It also covers simulations with an arbitrary number of monomers and dimers and provides rigorous error estimates. Comparison with test simulations of a simple Lennard Jones system with various particle numbers as well as with reference free energy values obtained from radial distribution functions show full agreement for both binding free energies and dimerization statistics.

 J Comput Chem 32: 1919–1928, 2011

Key words: free energy; law of mass action; bayesian statistics; statistical mechanics; thermodynamic ensemble

Introduction

The computation of free energy differences is the aim of many molecular simulation studies. As a thermodynamic state function, the free energy provides insights into the molecular driving forces for the studied process, and often enables a direct and quantitative comparison to experiments. However, in many cases, it is not trivial to obtain free energy differences from simulations of large condensed-phase systems, because it requires proper and extensive sampling of the underlying thermodynamic ensemble, for example through Monte Carlo (MC) or molecular dynamics (MD) techniques.

A number of MD-based simulation protocols for calculating free energy differences has been devised. Thermodynamic integration and free energy perturbation approaches, based on an alchemical transformation of one group of atoms into another, are frequently used.[1–4] Also nonequilibrium methods have been successfully applied to calculate free energies of molecular systems.[5–10] To study the energetics of self-assembly processes, such as the binding of two (or more) molecules, the umbrella sampling technique[11] is often applied, in which harmonic (umbrella) potentials drive the system along a pre-defined reaction coordinate, for example the distance between the molecules.

Today, the increasing available computer power and ongoing development of efficient algorithms and coarse-grain force fields

*These authors contributed equally to this work.

Correspondence to: H. Grubmüller; email: hgrubmu@gwdg.de

Grant sponsor: The Netherlands Organisation for Scientific Research; grant numbers: Veni grant 700.57.404; Top grant 700.57.303

enables longer simulations of large systems, thus opening the way to a straightforward alternative: To carry out an extended equilibrium MD simulation and obtain the free energy difference from directly counting the fractions of simulation time spent in the respective states. Such an approach has been applied to, for example, the dimerization of chirally related organic molecules,[12] folding / unfolding of small peptides,[13,14] dimerization of methane molecules[15] as well as of charged[16] and hydrophobic[17] amino acid pairs in water, and a dimer of transmembrane helices in a lipid bilayer.[18]

However, contradicting approaches and formulae have been employed to calculate free energy differences. In particular, as further explained in the Theory section, directly using the ratio of the observed Boltzmann probabilities[13,14,16,17,19] yields different results compared to approaches adopting the law of mass action to simulations of two dimerizing molecules.[12,15,18] For example, if the system is found in a dimerized state during a fraction P_1 of the total simulation time, and in a monomeric state during a fraction $P_0 = 1 - P_1$, the free energy difference would in the first case be given by an equation of the form $\Delta G \propto \ln(P_1/P_0)$, whereas in the latter case, it is $\propto \ln(P_1/P_0^2)$. Which of the two approaches is correct? Furthermore, it is not trivial to provide a generalized formalism as well as reliable error estimates for simulations with more than two molecules, which may provide better sampling.

Here, we develop a rigorous theory for dimerization reactions involving an arbitrary number of molecules, including only two, and derive how dimerization free energies can be calculated from simulations by direct counting. First, we will use thermodynamic arguments to show that an equation of the form $\Delta G \propto \ln(P_1/P_0)$ is the correct formula for simulations of two dimerizing molecules. Second, we present a general statistical mechanical treatment of dimer association/dissociation reactions of any number of molecules, and demonstrate how the law of mass action is recovered. Third, we discuss how the counter-intuitive disagreement between the Boltzmann treatment and the naive application of the law of mass action is resolved by careful consideration of the respective ensembles. We finally test our theoretical results against MD simulations, and compare free energies obtained from direct counting with those from radial distribution functions.

Theory

As shown in Figure 1, we consider the dimerization of two molecules in solution, A and B,

$$A + B \rightleftharpoons AB, \tag{1}$$

for which the law of mass action reads

$$K_a = \frac{[AB]c^{\varnothing}}{[A][B]}, \tag{2}$$

with association constant K_a, concentrations $[X]$, and $c^{\varnothing}$ an agreed standard concentration, usually 1 mol/L. Equation (2) assumes that the system is sufficiently diluted, such that concentrations can be used instead of activities.

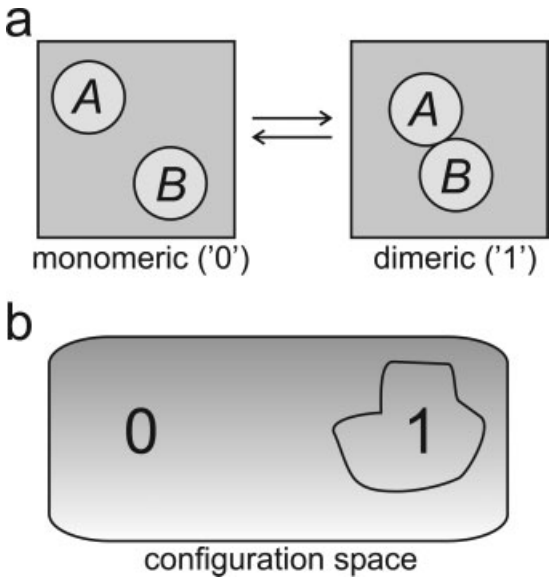

Figure 1. Dimerization of two molecules A and B within a given volume (a). The configuration space (b) is divided into two parts, monomeric (no dimer, "0") and dimeric (one dimer, "1").

Two Particles

For a mixture of n_i mol of species i, the Gibbs free energy is related to the thermodynamic (chemical) potentials through

$$G = \sum_i n_i \mu_i. \tag{3}$$

Thus, for two molecules,

$$G_1 = \frac{1}{N_{Av}}\mu_{AB} + \frac{N_s}{N_{Av}}\mu_s \tag{4}$$

and

$$G_0 = \frac{1}{N_{Av}}\mu_A + \frac{1}{N_{Av}}\mu_B + \frac{N_s}{N_{Av}}\mu_s \tag{5}$$

for the dimer and monomer states, 1 and 0, respectively. Here, N_s is the number of solvent molecules, μ_s the thermodynamic potential of the solvent, and N_{Av} is Avogadro's number. Thus,

$$N_{Av}G_1 = \mu_{AB}^{\varnothing} + RT \ln \frac{1}{c^{\varnothing} N_{Av} v} + N_s \cdot \left[\mu_s^{\varnothing} + RT \ln x_{s,1}\right] \tag{6}$$

and

$$N_{Av}G_0 = \mu_A^{\varnothing} + \mu_B^{\varnothing} + 2RT \ln \frac{1}{c^{\varnothing} N_{Av} v} + N_s \cdot \left[\mu_s^{\varnothing} + RT \ln x_{s,0}\right], \tag{7}$$

where $R = k_B N_{Av}$ is the gas constant, v is the total volume of the system, and $x_{s,1}$ and $x_{s,0}$ are the mole fractions of the solvent (assuming ideal solution) for the dimer and monomer, respectively. Taking the difference of eqs. (6) and (7) and neglecting the small difference between $x_{s,1}$ and $x_{s,0}$ yields

$$N_{Av}(G_1 - G_0) = \mu_{AB}^{\varnothing} - \mu_A^{\varnothing} - \mu_B^{\varnothing} + RT \ln(c^{\varnothing} N_{Av} v). \quad (8)$$

We now assume an MD simulation of two molecules A and B, surrounded by solvent molecules in a simulation box. The monomers can form a dimer according to eq. (1), defined by an unambiguous definition (for example a distance criterion). The system is observed to be in its dimeric state during a fraction P_1 ("one dimer") of the total simulation time and in its monomeric state ("zero dimers") during a fraction $P_0 = 1 - P_1$. We further assume that the simulation time is long enough for sufficiently many transitions to occur between the two states, such that the system can be considered to be in thermodynamic equilibrium. Our aim is to calculate the equilibrium constant K_a for the association reaction eq. (1), or, equivalently, the (standard) association free energy

$$\Delta G^{\varnothing} = -k_B T \ln K_a \quad (9)$$

from the simulation, that is, from P_0 and P_1.

To derive an expression for the free energy difference, the configuration space is divided into two parts 0 and 1, representing the monomeric and dimeric states, respectively (Fig. 1b). For simulations at constant v, T, the probability to be in a defined state is proportional to $\exp(-A/k_B T)$, where A is the Helmholtz free energy of that state. Thus,

$$\frac{P_1}{P_0} = \exp\left[\frac{-(A_1 - A_0)}{k_B T}\right]. \quad (10)$$

To obtain ΔG, the difference between the Gibbs functions,

$$G_1 - G_0 = -k_B T \ln \frac{P_1}{P_0} + (p_1 - p_0)v \quad (11)$$

is desired. The pv term is small for most reactions in solution and is therefore disregarded here; however, it can be determined from the pressures during the simulation if needed. For simpler notation, all partition functions further below refer to simulation ensembles at constant v, T.

Combining eqs. (8) and (11) yields the expected Boltzmann relation between the standard free energy change and the probability ratio observed in the simulation,

$$\Delta G^{\varnothing} = -RT \ln \frac{P_1}{P_0} - RT \ln(c^{\varnothing} N_{Av} v) = \Delta G - RT \ln(c^{\varnothing} N_{Av} v). \quad (12)$$

Alternatively, the equilibrium constant can be expressed using eq. (9),

$$K_a = \frac{P_1}{P_0} c^{\varnothing} N_{Av} v. \quad (13)$$

Subsequently, we will omit explicit reference to the standard state; it can easily be reintroduced via the above eqs. (12) and (13).

Generalization for Many Particles

We now assume an equilibrium between N molecules, N_A of which are of type A, and $N_B = N - N_A$ of which are of type B. We further assume that particles of the same type do not dimerize, and that polymerization does not occur. For example, for $N = 4$ and $N_A = N_B = 2$, the relevant states are

$$A + A + B + B \rightleftharpoons AB + A + B \rightleftharpoons AB + AB. \quad (14)$$

Generalizing eqs. (12) and (13), we aim at relating the statistics of observed dimers AB in the MD simulation to the (macroscopic) association constant K_a and free energy ΔG for the dimerization reaction eq. (1). As in Figure 1, the configuration space of the N molecules within the (3-dimensional) volume v is divided into ($3N$-dimensional) subvolumes V_0 (only monomers), V_1 (1 dimer, $N - 2$ monomers), ..., V_m (m dimers, $N - 2m$ monomers).

From the respective Helmholtz free energies, eq. (10), the probability of finding the system in a fully monomeric state reads

$$P_0 = \frac{Z_0}{Z} = \frac{\int_{V_0} e^{-U/k_B T} dV}{\int_V e^{-U/k_B T} dV}, \quad (15)$$

and that of finding exactly $m \leq \min(N_A, N_B)$ dimers is

$$P_m = \frac{Z_m}{Z} = \frac{\int_{V_m} e^{-U/k_B T} dV}{\int_V e^{-U/k_B T} dV}, \quad (16)$$

with configurational partition functions Z_m, integrated over those regions of configuration space with m dimers and $N-2m$ monomers, interaction potential U, and partition function $Z = \sum_{m=0}^{\min(N_A,N_B)} Z_m$. We note that the kinetic part of the partition functions can always be factored out and therefore cancels in the above equations. From these probabilities, the average number $\langle m \rangle$ of dimers is readily obtained,

$$\langle m \rangle = \frac{1}{Z} \sum_{m=1}^{\min(N_A,N_B)} m Z_m. \quad (17)$$

Neglecting the interaction energy of distant (unbound) particles, the above partition functions can be expressed in terms of an excess free energy per dimer (with respect to the ideal gas term),

$$G^* = k_B T \ln \langle e^{U/k_B T} \rangle_{V_1}, \quad (18)$$

which is independent of the box volume. With this abbreviation, one obtains

$$Z_0 = \int_{V_0} e^{-U/k_BT} dV = V_0 \tag{19}$$

and

$$Z_1 = \int_{V_1} e^{-U/k_BT} dV = V_1 e^{-G^*/k_BT}. \tag{20}$$

Assuming additive interaction potentials, the partition function for m dimers then reads

$$Z_m = \int_{V_m} e^{-U/k_BT} dV = V_m e^{-mG^*/k_BT}, \tag{21}$$

which reduces the problem to estimating the configuration space volumes V_m.

To determine the configuration space volume V_0 for which the system consists of only monomers, note that the first molecule can be placed anywhere within the simulation box volume v. For the second molecule, only those positions are allowed for which it does not form a dimer with the first molecule, which yields the reduced volume $v - v_D$. Here, v_D denotes the dimerization volume. Further, molecules A and B are assumed to occupy the same volume, that is, the dimerization volume v_D equals the volume excluded by the repulsive interaction between *AA* or *BB*, respectively. For the third molecule, similarly, only a volume $v - 2v_D$ remains, etc.

With $x = v_D/v$, one thus obtains to second order in the particle concentrations $[A] + [B] = N/(N_{Av}v)$,

$$V_0 = v(v - v_D)(v - 2v_D) \cdot \cdots \cdot (v - (N-1)v_D) = v^N \prod_{j=1}^{N-1} (1 - jx). \tag{22}$$

For the configuration space volume V_m of all states of m dimers and $N - 2m$ monomers, from similar but somewhat more involved reasoning

$$V_m = v^N m! \binom{N_A}{m} \binom{N_B}{m} x^m \prod_{j=1}^{m-1} (1 - jy) \prod_{j=0}^{N-2m-1} (1 - my - jx) \tag{23}$$

follows, where v_{AB} is the (average) volume excluded by each dimer *AB* (Fig. 2b), and $y = v_{AB}/v$. To see why eq. (23) holds true, first place m molecules to form a monomeric state, which yields the first product term as in eq. (22). Next, place further m molecules to form m dimers, such that each of these molecules is restricted to a volume v_D, thus yielding the x^m-term. Finally, place $N - 2m$ further monomers within the remaining volume fraction $v^N(1 - my)$, with each monomer further reducing the available volume fraction by x. To verify the combinatorics note that this procedure yields only one out of all possible ways to select m molecules from the N_A molecules of type A and from the N_B molecules B, hence the two binomials. Finally, having selected m molecules of each type for dimerization, there are $m!$ ways of joining those into m dimers.

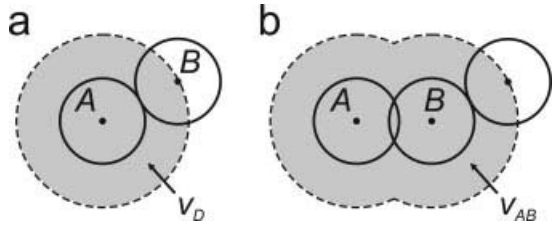

Figure 2. Definition of (a) dimerization volume v_D and (b) average dimer exclusion volume v_{AB} for the special case of spherical particles. If the center of particle B falls within the dimerization volume (gray) centered at particle A, the two particles are considered a dimer. The average volume of two combined overlapping dimerization volumes defines the dimer exclusion volume, within which a third particle is not considered monomeric.

Note that with the convention that products for which the final value of the running index is smaller than the starting value equal unity (i.e., $\prod_{j=1}^{0} = 1$), eq. (23) reduces to eq. (22) for $m = 0$.

Equations (22) and (23) are independent of the shapes of the volumes. However, v_D and v_{AB} (and thus x and y) may in general be hard to determine. For spherical particles, a useful estimate is obtained by assuming a constant (average) interaction energy between the particles (Fig. 2). In this case, the distance distribution between overlapping spheres is $p(r) \propto r^2$, yielding an average overlap of $v_D/8$ and, hence, $y = x \cdot 15/8$. For the general case, it is important to realize that by choosing a criterion to define the dimer state, for example a distance cut-off, one implicitly determines the dimer volume, v_D. It is thus also possible to determine V_m numerically, without prior knowledge of v_D or v_{AB} by placing N non-interacting particles in a volume through a Monte Carlo run, as is demonstrated in section "Molecular Dynamics Simulations".

Limiting Cases

It is noteworthy to consider the case of moderately large N and low concentrations (i.e., $Nv_D \ll v$ and, therefore, $x < y \ll 1$). In this case, using $\binom{N}{m} \approx N^m/m!$ and expanding the logarithm of the result to first order, eq. (23) simplifies to

$$V_m \approx v^N \frac{(N_A N_B v_D/v)^m}{m!} e^{-\frac{1}{2}N^2 v_D/v}. \tag{24}$$

Combining eq. (24) with eqs. (16) and (21), after proper normalization, a Poisson distribution follows for the probabilities of finding m dimers:

$$P_m = \frac{\lambda^m e^{-\lambda}}{m!}, \tag{25}$$

with

$$\lambda = N_A N_B \frac{v_D}{v} e^{-G^*/k_BT}. \tag{26}$$

For very large N, eq. (23) can be approximated by a Gaussian function in m of width $m^{1/2}$ and with a maximum at $m = m_{\max}$ given by

$$\frac{m_{\max}}{(N_A - m_{\max})(N_B - m_{\max})} = \frac{v_D}{v} e^{-G^*/k_B T}. \quad (27)$$

Because the relative width of this function tends to zero for large m, and with $[A] = (N_A - m)/(N_{Av}v)$, $[B] = (N_B - m)/(N_{Av}v)$, and $[AB] = m/(N_{Av}v)$, the law of mass action is readily recovered,

$$K_a = \frac{[AB]}{[A][B]} = v_D N_{Av} e^{-G^*/k_B T}. \quad (28)$$

K_a from Counting and Error Estimate

The above framework enables to determine K_a from the number of dimers and monomers observed during a simulation.

For $N = 2$, eqs. (17) and (21), and using eqs. (22) and (23), yield

$$\frac{n_1}{n_0} \approx \frac{P_1}{P_0} = \frac{V_1}{V_0} e^{-G^*/k_B T} = \frac{1}{v/v_D - 1} e^{-G^*/k_B T}, \quad (29)$$

or

$$G^* = -k_B T \ln \left[\frac{n_1}{n_0} \left(\frac{v}{v_D} - 1 \right) \right], \quad (30)$$

where n_1 and n_0 are the number of snapshots from the trajectory containing one dimer or two monomers, respectively. As can be seen, for simulation volumes v that are small compared to the molecular volumes v_D, an estimate for the latter is required. The association constant is then obtained by combining eqs. (28) and (30),

$$K_a = N_{Av} \frac{n_1}{n_0} (v - v_D). \quad (31)$$

For the general case of N particles, proceeding along similar lines, eq. (17) serves to relate the average number of dimers $\langle m \rangle$ to G^*. Hence, G^* can be obtained from the dimer frequencies observed in a simulation of few particles — either via the second order approximations eqs. (22) and (23), or via numerical integration, for example, through Monte Carlo approaches, as demonstrated further below.

Moreover, simulations with $N > 3$ provide an independent approach to calculate G^*: For the number n_m of snapshots from the trajectory containing m dimers, eqs. (17) and (21) yield

$$\frac{n_m}{n_0} \approx \frac{P_m}{P_0} = \frac{V_m}{V_0} e^{-mG^*/k_B T}. \quad (32)$$

Therefore, the quantity

$$g_m = k_B T \ln \frac{n_0 V_m}{n_m V_0} \quad (33)$$

should satisfy $g_m = mG^*$, that is, be proportional to m. Plotting g_m as a function of m thus provides the excess free energy per dimer G^* as the slope, and the plot yields a straight line only for properly chosen v_D.

Finally, error estimates are readily obtained from a Bayesian approach by considering the conditional probability

$$P(G^{*\prime}|n_0, n_1, n_2, \ldots) \propto P(n_0, n_1, n_2, \ldots |G^{*\prime}) \cdot P(G^{*\prime}) \quad (34)$$

$$\approx P(n_0, n_1, n_2, \ldots |G^{*\prime}) \cdot 1 = \prod_{m=0}^{\min(N_A, N_B)} \left[\frac{V_m e^{-mG^{*\prime}/k_B T}}{\sum_k V_k e^{-kG^{*\prime}/k_B T}} \right]^{\tilde{n}_m}. \quad (35)$$

Here, a uniform *a priori* probability $P(G^{*\prime})$ for the dimer interaction free energy is assumed, and $\tilde{n}_m = n_m \cdot \Delta t / t_c$ is the effective (i.e., statistically independent) number of snapshots containing $m = 0, 1, 2, \ldots$ dimers. Several methods to determine this number are available, which critically determines the obtained error estimate, for example, correlation analysis,[20–22] block averaging,[23,24] and bootstrap analysis.[25,26] For the test simulations presented below, the statistically independent number of snapshots is estimated from the average time t_c between collisions relative to the time spacing Δt of snapshots in the trajectory.

This probability distribution serves to calculate G^* and an estimate of its statistical uncertainty σ_{G^*} via

$$G^* = \int_{-\infty}^{\infty} G^{*\prime} P(G^{*\prime}|n_0, n_1, n_2, \ldots) dG^{*\prime} \quad (36)$$

and

$$\sigma_{G^*}^2 = \int_{-\infty}^{\infty} (G^{*\prime} - G^*)^2 P(G^{*\prime}|n_0, n_1, n_2, \ldots) dG^{*\prime}. \quad (37)$$

An example python program is provided in the Appendix.

Resolving the Seeming Contradiction

We have shown above that for dilute systems of two dimerizing molecules, the Boltzmann approach, $\Delta G \propto \ln P_1/P_0$ [eqs. (12) and (13)], yields correct free energies, whereas direct application of the law of mass action provides wrong results. Nevertheless, as the above results show, the law of mass action can be derived from the Boltzmann approach and is in this sense compatible — as must be. What, then, is wrong with the expression $\Delta G \propto \ln P_1/P_0^2$ suggested by the law of mass action? As we will show in the following, both approaches are in fact correct; however, to apply the law of mass action to simulations of, for example, only two dimerizing molecules requires a careful consideration of the relevant thermodynamic ensembles.

To demonstrate this, we again consider a two-particle MD system, and define the concentrations [A], [B], and [AB] from the respective probabilities,

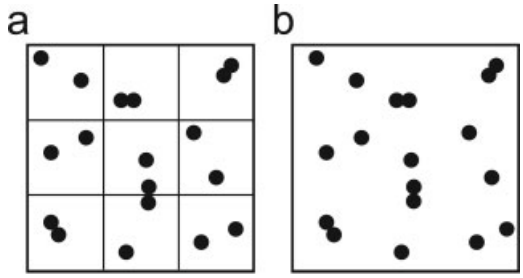

Figure 3. Two different ensembles are considered. Using dimerization frequencies obtained from an MD simulation of two molecules in a periodic box implicitly assumes an ensemble (a) consisting of replicas of the (microscopic) simulation system. Accordingly, the molecules cannot interact across the boxes [the configurations in (a) represent snapshots at different time points, not periodic boxes]. Straightforward application of the law of mass action, in contrast, refers to a macroscopic system of molecules (b), which can all form dimers with each other. As a consequence, the two ensembles generate different monomer/dimer ratios.

$$[A] = \frac{N_A P_0}{v^{\text{eff}} N_{Av}}, \tag{38}$$

$$[B] = \frac{N_B P_0}{v^{\text{eff}} N_{Av}}, \tag{39}$$

and

$$[AB] = \frac{N_{AB} P_1}{v^{\text{eff}} N_{Av}}, \tag{40}$$

where, for our two-particle system, $N = 2$, $N_A = N_B = N_{AB} = 1$, and v^{eff} is an appropriate effective volume. Inserting these concentrations into the law of mass action yields the puzzling result

$$K_a = \frac{[\text{AB}]c^{\varnothing}}{[\text{A}][\text{B}]} = \frac{P_1}{P_0^2} c^{\varnothing} N_{Av} v^{\text{eff}}. \tag{41}$$

Note, however, that it has not yet been defined how v^{eff} relates to the volume v of the simulation box. In contrast to what is saliently assumed by the above application of the law of mass action in eq. (41), the MD ensemble does *not* represent a (macroscopic) volume $v^{\text{eff}} = v \cdot N/2$ containing N interacting molecules (Fig. 3b). Rather, it is an ensemble of $N/2$ simulation systems (Fig. 3a), that is, $N/2$ separate (periodic) boxes, each of volume v and containing two molecules. The crucial difference is that molecules from different boxes can never dimerize and, thus, the association rate, k^+, differs for the two ensembles. Since the dissociation rate, k^-, is unaffected, and $K_a \propto k^+/k^-$, the equilibrium concentrations also differ for the two ensembles. In particular, the time-averaged fraction P_0 of molecules in the monomeric state obtained from the MD simulation is, generally, not equal to the ensemble fraction expected for the macroscopic volume.

The two different ensembles can be reconciled by compensating for the fact that for each molecule within the $N/2$ MD boxes, a dimerization partner is available only during a fraction P_0 of the time. This is achieved by decreasing the effective volume by this factor, i.e., $v^{\text{eff}} := P_0 v N/2$ (note that properly correcting P_0 instead of v^{eff} yields the same results). Inserting this expression into eq. (41) recovers eq. (13), thus resolving the apparent contradiction.

Molecular Dynamics Simulations

We carried out equilibrium MD simulations to demonstrate how the above framework can be applied to obtain association constants from simulations. Our simulation systems comprised of N van der Waals particles in a box, with N ranging from 2 to 64. $N/2$ of the particles were considered to be of type A and $N/2$ of type B, respectively (for uneven N, there was one excess A particle). The systems were simulated within periodic boundary conditions at constant volume using the Gromacs (v. 4.0.5) simulation package.[27] The temperature was kept constant using stochastic temperature coupling with an inverse friction coefficient of 5 ps. The neighbor list was updated at every integration time step, which was set to 50 fs. Test simulations with 10 fs and 20 fs integration time steps yielded identical results within the statistical errors. The particles had a mass of 72 amu and were interacting through a Lennard-Jones 6-12 potential

$$V_{LJ}(r) = 4\epsilon \left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right), \tag{42}$$

with $\sigma = 0.47\,\text{nm}$ and $\epsilon = 4\,\text{kJ mol}^{-1}$. The potential was smoothly shifted to zero between 0.9 and 1.2 nm. For each N, we studied three different concentrations, corresponding to a volume per particle of $v/N = 27$, 54, and $108\,\text{nm}^3$, respectively. Using these volumes, and at the simulation temperature of 298 K, the systems are in the gaseous state.

To obtain comparable statistics for a given computational cost for the different systems, each simulation was carried out for $64/N\ \mu$s of simulation time. For example, the simulation system with $N = 2$ particles was simulated for $32\ \mu$s, whereas the simulation time for the system with $N = 64$ particles was $1\ \mu$s. These simulation times are long enough to have sufficiently many (i.e., thousands) of dimer association / dissociation events.

To obtain K_a from the simulations, the number of particles A within a distance r_c of any particle B was counted along the trajectory using the g_mindist tool of Gromacs. Subsequently, the obtained set of contacts was filtered for higher-order oligomers, which were discarded in order to not erroneously count them as dimers (the average number of particles in higher order oligomers was less than 1%, even at the highest concentration used). The dimer cut-off distance $r_c = 0.7\,\text{nm}$ was chosen such that the dimer peak observed in the radial distribution function is included. We did not investigate the dependence of K_a on the chosen dimer cut-off r_c,[28,29] as our aim here was to compare different approaches for the calculation of K_a using the same cut-off.

Using the final (filtered) set of dimer contacts, G^* and its statistical uncertainty σ_{G^*} were calculated using the above Bayesian approach [eq. (36) and eq. (37); see Appendix for an example python program]. To estimate the configuration space volumes, we used both the analytic approximation [eq. (23)] and a Monte Carlo approach (see below), which is numerically exact. Finally,

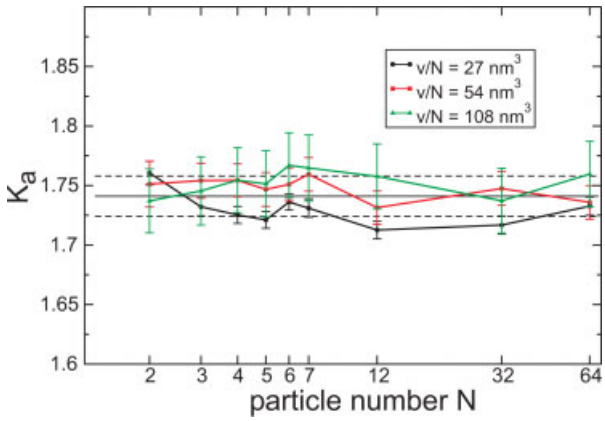

Figure 4. Association constant K_a obtained from counting the snapshots containing m dimers and $N - 2m$ monomers in equilibrium MD simulations of a van der Waals gas. Systems with $N = 2$, 3, 4, 5, 6, 7, 12, 32, and 64 particles were simulated at volumes $v/N = 27$, 54, and 108 nm^3, respectively (black, red, and green curves, respectively); statistical errors were estimated using Bayesian statistics. The solid horizontal line gives K_a as obtained from integrating over the bound and unbound parts of the radial distribution function [eq. (44)]; here, the statistical error (dashed lines) is the difference between the K_a's obtained from separately analyzing the two halves of the trajectory.

the equilibrium constant was calculated from G^* using the standard state form of eq. (28),

$$K_a = \frac{v_D}{v^\varnothing} e^{-G^*/k_B T}, \tag{43}$$

where $v^\varnothing = (c^\varnothing N_{Av})^{-1}$ is the (molecular) standard volume (1.66 nm^3) and $v_D = 4/3\ \pi\ r_c^3$ (Fig. 2). A uniform distribution $P(G^{*\prime})$ was used as a prior. To estimate the number of statistically independent snapshots, we calculated the average time between particle collisions according to $t_c = \sigma_c^{-1}\ \bar{c}^{-1/2}\ v/N$, with collision cross section $\sigma_c = \pi\ (2\sigma_{LJ})^2$ and mean velocity $\bar{c}$. We obtained $t_c \approx 25$, 50, and 100 ps for the systems with $v/N = 27$, 54, and 108 nm^3, respectively.

Results and Discussion

Figure 4 summarizes the results obtained for the three different particle concentrations studied. The association constant K_a should be independent of the number of particles and of the volume of the simulation box. Figure 4 shows that this is indeed the case: Averaged over all particle numbers N, the obtained association constants ($\pm$ std. dev.) are 1.730 ± 0.014, 1.748 ± 0.009, and 1.752 ± 0.011 for $v/N = 27$, 54, and 108 nm^3, respectively. Furthermore, the statistical errors (bars in Fig. 4) turn out to be independent of N, due to the comparable simulation times of 64/N μs per system.

As a check for the calculated K_a, Figure 5a shows a plot of g_m versus m. The linear dependence suggested by eq. (33) holds, and the fit yields $G^* = -0.6951$ k$_B$T ($K_a = 1.734$), in excellent agreement with $K_a = 1.733 \pm 0.008$ obtained from counting the number of dimers in the simulation with $N = 64$, $v/N = 27$ nm^3 (Fig. 4). Figure 5b confirms that for the simulated systems with moderately large N, the probabilities of finding m dimers follow a Poisson distribution, as derived above, eq. (25).

As another, independent check for K_a, we integrated over the bound and unbound parts of the radial distribution function $g(r)$, as obtained from the simulation with $N = 2$ particles, according to

$$K_a = \frac{4\pi R^3 \int_0^{r_c} r^2 g(r) dr}{3v^\varnothing \int_{r_c}^{R} r^2 g(r) dr}. \tag{44}$$

The thus obtained equilibrium constant of $K_a = 1.74 \pm 0.02$ (solid line in Fig. 4) also agrees with the result from counting.

To assess the accuracy of the analytical second-order approximation for the configuration space volumes V_m, eq. (23), we placed $N = N_A + N_B$ non-interacting particles in a periodic 3-dimensional volume through Monte Carlo (MC) sampling, and counted the number of snapshots containing m dimers and $N - 2m$ monomers. As above, a dimer was defined by a distance criterion between any two

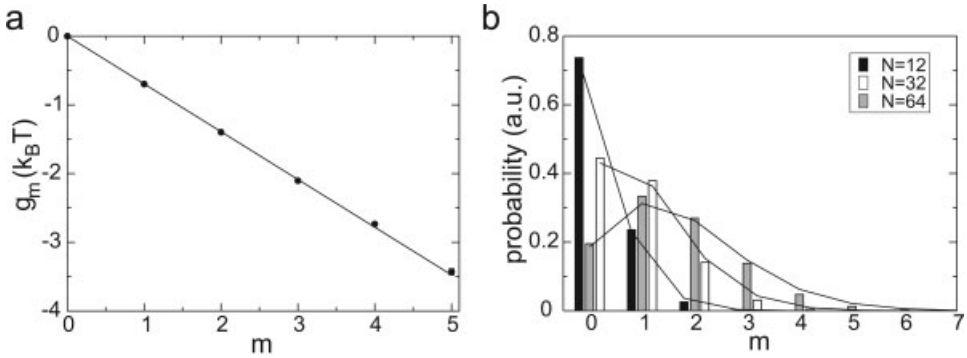

Figure 5. (a) Plot of g_m over m, obtained from the simulation with $N = 64$ particles in a box with volume $v/N = 27$ nm^3. The slope of the linear fit (solid line) yields the same G^* as obtained from counting. (b) The bars show the distributions of snapshots containing m dimers for the simulation systems with 12, 32, and 64 particles, respectively ($v/N = 27$ nm^3). They follow the corresponding Poisson distributions, eq. (25), plotted as lines. The individual distributions are slightly shifted along the m-axis for clarity.

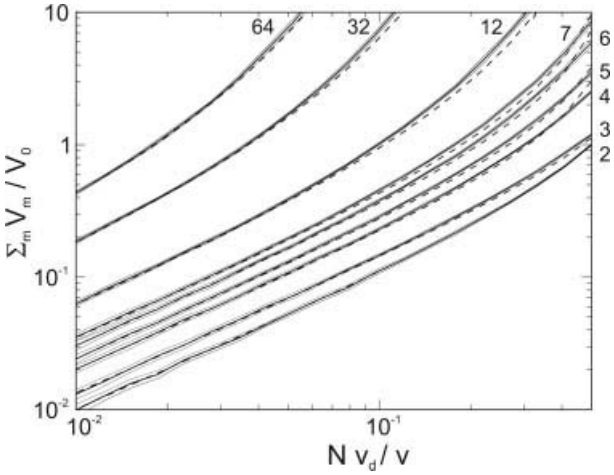

Figure 6. Comparison of dimer/monomer frequencies as a function of particle concentration, obtained from the second-order analytical approximation eq. (23) (dashed lines) and Monte Carlo sampling (solid lines). Results are shown for $N = 2, 3, 4, 5, 6, 7, 12, 32$, and 64 (bottom to top). Statistical errors are shown as side lines.

particles A and B. Trimers and higher-order oligomers were discarded. The volume v and particle number N were systematically varied.

Figure 6 shows that at low concentrations ($Nv_D/v \leq 0.1$), eq. (23) is a very good approximation for the ratio of dimer/monomer configuration space volumes, particularly for small N. For higher concentrations $0.1 < Nv_D/v < 0.4$, eq. (23) predicts a slightly too low ratio, whereas for even higher concentrations, the analytical approximation might slightly overestimate V_m/V_0, at least for $N \leq 7$. For $N = 2$, the results obtained from eq. (23) and from MC sampling agree along the entire concentration range, because the analytical formula is exact in this case: The dimer configuration space volume V_m only contains terms beyond the second order for $N > 2$ (the monomer configuration space volume V_0 lacks higher order terms). In summary, from Figure 6 we conclude that for the diluted systems studied here ($0.01 < Nv_D/v < 0.05$), the second order analytical approximation Eq. (23) yields sufficiently accurate estimates of the configuration space volumes, and no further concentration-dependent correction is required.

Summary and Conclusions

We presented a derivation of the thermodynamics of dimerization reactions, and laid out how to calculate equilibrium constants and corresponding free energies from simulations of a limited number of dimerizing molecules. Using thermodynamic arguments, we have shown that naive application of the law of mass action, eq. (41), yields wrong results in particular for simulations of only few dimerizing molecules, and that correct results are obtained via the Boltzmann factor of the ratio of the observed frequencies, eq. (13). The difference between the two approaches can be significant, in particular if $P_0 \ll 1$ of course, as is the case for the systems studied in Ref. 12, 18.

We further derived through statistical mechanics the thermodynamics of dimerization reactions of any number of particles, and a Bayesian statistics approach to estimate equilibrium constants and free energies with their statistical errors from simulations. Finally, we showed that the two approaches can be reconciled by carefully considering the different underlying thermodynamic ensembles. We applied our approach to extract equilibrium constants from molecular dynamics simulations of systems containing different numbers of dimerizing particles.

One may ask whether there is an optimal system size to obtain statistically accurate free energies of dimerization from MD simulations through direct counting, given a certain available amount of computer time. From our results, we would argue in favor of simulating systems with $N = 2$ dimerizing molecules, for the following reasons. First, eq. (23) is exact for the two-particle case, thus no concentration-dependent correction needs to be applied. Second, no trimers (or higher order oligomers) can occur, thus simplifying the analysis. Third, the computer time grows at least linearly with N, while our results show that simply having a larger number of molecules in the simulation box does not *per se* improve the statistical accuracy as compared to a simulation with two dimerizing molecules and correspondingly longer sampling time. For example, it would be better to carry out, for example, four independent simulations with $N = 2$ molecules instead of one single simulation with $N = 8$, as the latter would suffer from non-optimal parallel scaling.

Another question of practical importance is the choice of the simulation volume. One might argue that a large volume is desirable, because in that case concentrations can be used instead of activities, and second order effects (and thus the dimerization volume v_D) can be neglected. However, such an approach would suffer from a low statistical accuracy due to the small number of associations/dissociations. In addition, for simulations with explicit solvent, one seeks to reduce the number of solvent molecules as much as possible, since their treatment is computationally usually the most expensive part of the simulation. This discussion also underscores the importance of including v_D within our theory for obtaining free energies from counting in MD simulations.

Finally, we would like to discuss the advantages of running extended equilibrium simulations over biased simulations, such as umbrella sampling. From the latter, K_a can be calculated from the obtained potential of mean force $V_{\mathrm{mf}}(r)$ according to eq. (44), with $g(r) = \exp(-V_{\mathrm{mf}}(r)/k_BT)$. This approach may seem more straightforward. However, it has the disadvantage that the system needs to be driven along a pre-defined reaction coordinate. This may be, for example, distances, angles, dihedrals, or (linear) combinations of these — even curvilinear coordinates may be required in certain cases. In general, the definition of a proper reaction coordinate may not always be straightforward and, furthermore, involve the derivation of Jacobian corrections that can become cumbersome for more complicated reaction coordinates. In such cases, it appears more convenient to run an unbiased simulation and choose the parameters to define the different states in an *a posteriori* manner during the analysis.

Acknowledgment

We thank Mark Sansom and Ruud M. Scheek for helpful discussions.

Appendix

Example python program using Bayes statistics for calculating G^* from an MD simulation. The configuration space volumes are estimated using eq. (23).

```
       from numpy import *
from math import factorial

def VM(na,nb,m,v,vd,y=None):
    '''Returns VM as a function of #particles (na,nb),\
    #dimers(m), volume(v), dimer volume(vd), excl.volume(y)'''
    x = vd/(1.0*v)
    if y == None: y = 15.*x/8.
    n = na + nb
    prod=1.0*factorial(m)*v**(1.0*n)
    prod=prod*factorial(na)/(factorial(m)*factorial(na-m))
    prod=prod*factorial(nb)/(factorial(m)*factorial(nb-m))
    prod=prod*x**m
    for j in range(1,m):
        prod=prod*(1.0-y*j)
    for j in range(1,n-2*m):
        prod=prod*(1.0-m*y-x*j)
    return prod

def bayes(vmarr,nf,na,nb,v,vd):
    '''A probability distribution for Gd is calculated, \
    given vmarr (array containing the # monomers,\
    single dimers (m=1), double dimers (m=2), etc \
    for every sample set, e.g. obtained with g_mindist) \
    and nf, the effective nr of samples (dt/tcoll), and na/nb.'''
    # Define the range and resolution of Gd
    gdmax = 2.; res=2000
    gdarr=2*gdmax*.arange(res)/(1.0*res)-gdmax
    #Choose a flat prior
    parr=ones(res)/(1.0*res)
    for n in range(len(vmarr)):
        for i in range(0,res):
            norm =0.0
            max_M=len(vmarr[n,:])
            for k in range(0,max_M):
                norm = norm + VM(na,nb,k,v,vd) * exp(-k*gdarr[i])
            prod=1.0
            for m in range(0,max_M):
                VMc=VM(na,nb,m,v,vd)
                prod=prod*(VMc*exp(-m*gdarr[i])/norm)**(1.0*nf*vmarr[n,m])
            parr[i]=prod*parr[i]
        parr = parr/(sum(parr*gdarr))
        gd = sum(gdarr * (parr/sum(parr)));
        print gd
    gdsigma = sum((gdarr -gd)**2 * (parr/sum(parr)));
    print 'gd = %.8f +/- %.8f'%(gd,sqrt(gdsigma))
```

References

1. Shirts, M. R.; Mobley, D. L.; Chodera, J. D. Annu Rep Comput Chem 2007, 3, 41.
2. Knight, J. L.; Brooks, C. L., III. J Comput Chem 2009, 30, 1692.
3. Deng, Y. Q.; Roux, B. J Phys Chem B 2009, 113, 2234.
4. Christ, C. D.; Mark, A. E.; van Gunsteren, W. F. J Comput Chem 2010, 31, 1569.
5. Jarzynski, C. Phys Rev Lett 1997, 78, 2690.
6. Jarzynski, C. Phys Rev E 1997, 56, 5018.
7. Crooks, G. E. J Stat Phys 1998, 90, 1481.
8. Ytreberg, F. M.; Swendsen, R. H.; Zuckerman, D. M. J Chem Phys 2006, 125, 184114.
9. Goette, M.; Grubmüller, H. J Comput Chem 2009, 30, 447.
10. Hummer, G. Free energy calculations: theory and applications in chemistry and biology; Springer, Berlin and Heidelberg, 2007.
11. Torrie, G. M.; Valleau, J. P. J Comput Phys 1977, 23, 187.
12. Hünenberger, P.; Granwehr, J. K.; Aebischer, J. N.; Ghoneim, N.; Haselbach, E.; van Gunsteren, W. F. J Am Chem Soc 1997, 119, 7533.
13. Daura, X.; van Gunsteren, W. F.; Mark, A. E. Proteins Stuct Funct Genet 1999, 34, 269.
14. de Groot, B. L.; Daura, X.; Mark, A. E.; Grubmüller, H. J Mol Biol 2001, 309, 299.
15. Zhang, Y.; McCammon, J. A. J Chem Phys 2003, 118, 1821.
16. Thomas, A. S.; Elcock, A. H. J Am Chem Soc 2004, 126, 2008.
17. Yang, H.; Elcock, A. H. J Am Chem Soc 2003, 125, 13968.
18. Psachoulia, E.; Fowler, P. W.; Bond, P. J.; Sansom, M. S. P. Biochemistry 2008, 47, 10503.
19. Moro, G. J.; Severin, M. G. J Chem Phys 2001, 114, 4565.
20. Jenkins, G. M.; Watts, D. G., Spectral analysis and its applications; Holden-Day: San Francisco, 1968.
21. Schiferl, S. K.; Wallace, D. C. J Chem Phys 1985, 83, 5203.
22. Straatsma, T. P.; Berendsen, H. J. C.; Stam, A. J. Mol Phys 1986, 57, 89.
23. Flyvbjerg, H.; Petersen, H. G. J Chem Phys 1989, 91, 461.
24. Hess, B. J Chem Phys 2002, 116, 209.
25. Efron, B. Ann Stat 1979, 7, 1.
26. Knecht, V.; Grubmüller, H. Biophys J 2003, 84, 1527.
27. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. J Chem Theory Comput 2008, 4, 435.
28. Shoup, D.; Szabo, A. Biophys J 1982, 40, 33.
29. Jorgensen, W. L.; Severance, D. L. J Am Chem Soc 1990, 112, 4768.