Generalized Yvon-Born-Green Theory for Molecular Systems

J. W. Mullinax and W. G. Noid*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

(Received 6 June 2009; published 6 November 2009)

We employ a basis set representation for classical force fields to derive an original system of exact integral equations relating each mode in the force field to an associated set of structural correlation functions. This generalized Yvon-Born-Green theory provides a framework for interpreting complex many-body correlations and also for variationally determining optimal interaction potentials for proteins and other complex molecules directly from structural correlation functions.

DOI: 10.1103/PhysRevLett.103.198104 PACS numbers: 87.15.A−, 05.20.Jj, 61.20.Gy, 87.15.−v

The direct quantitative deduction of molecular interactions from structural information has significant ramifications for nanotechnology, molecular biology, and a variety of physical systems. The many-body correlations inherent to condensed phase systems make this a challenging inverse statistical physics problem. In the case of monatomic liquids interacting via central pair potentials, the Yvon-Born-Green (YBG) equation provides an exact linear relation determining these pair potentials from knowledge of two- and three- particle correlation functions [1]. Subsequently, the YBG equation has been generalized for molecular systems with pair additive potentials [2,3]. However, accurate models for many molecular systems of interest often involve many-body potentials to describe, e.g., angle and torsional interactions. The YBG equation has not been generalized to treat intramolecular many-body potentials and has not been applied to the inverse problem for complex molecular systems. Instead, current methods for deducing interaction potentials from molecular structures implicitly treat many-body correlations by relying upon either approximate closure relations that relate the unknown potentials to simple distribution functions [4–7] or iterative nonlinear regression techniques [8–10], e.g., reverse Monte Carlo methods [11,12]. Consequently, exact linear equations for determining many-body potentials from structural correlation functions would represent a significant advance in addressing this inverse problem and would, moreover, provide considerable insight into the important role played by many-body correlations in complex molecular systems.

This Letter introduces a generalized YBG theory for complex molecular systems. We employ a basis set representation for classical force fields to derive exact linear relations between force functions and related structural correlation functions. We prove that, given appropriate canonical distribution functions for an unknown potential function, the set of potential functions that satisfy the generalized YBG equation determine an optimal approximation to the unknown potential. Consequently, force-matching variational principles for determining approximate potentials [13–16] can be implemented directly from structural correlation functions. We numerically demonstrate the quantitative accuracy of this theory for a model protein system.

We consider the canonical ensemble for an n particle system with a potential energy function that may be expressed as a function of the Cartesian coordinates $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)$:

$$u(\mathbf{r}) = \sum_{\zeta} \sum_{\lambda} u_{\zeta}(\psi_{\zeta}(\{\mathbf{r}\}_{\lambda})), \tag{1}$$

where ζ identifies a particular type of interaction (e.g., a bond angle interaction) that is a function of a particular mode ψ_{ζ} (e.g., a bond angle) that is itself a function of the Cartesian coordinates for a particular set of particles λ in the system. The force on particle i may be expressed

$$\mathbf{f}_{i}(\mathbf{r}) = \sum_{\zeta} \int dz \phi_{\zeta}(z) \mathcal{G}_{i;\zeta}(\mathbf{r}; z), \tag{2}$$

where $\phi_{\zeta}(z) = -du_{\zeta}(z)/dz$ is the force on mode ζ and

$$\mathcal{G}_{i;\zeta}(\mathbf{r}; z) = \sum_{\lambda} \frac{\partial \psi_{\zeta\lambda}(\mathbf{r})}{\partial \mathbf{r}_i} \delta(\psi_{\zeta\lambda}(\mathbf{r}) - z), \tag{3}$$

with $\psi_{\zeta\lambda}(\mathbf{r}) \equiv \psi_{\zeta}(\{\mathbf{r}\}_{\lambda})$. The force field $\mathbf{f}$ is identified as a set $\{\mathbf{f}_1(\mathbf{r}), \ldots, \mathbf{f}_n(\mathbf{r})\}$ of n vector valued functions that determine the force on each particle in a given configuration $\mathbf{r}$. This set may be considered to be an element in a vector space of force fields [17]. Equation (2) represents $\mathbf{f}$ as a linear combination of force field basis vectors, $\mathcal{G}_{\zeta}(z)$, with elements given by Eq. (3). The set of basis vectors included in Eq. (2) spans a subspace in the vector space of force fields and a particular continuous set of coefficients, $\phi_{\zeta}(z)$, identifies a particular force field in this subspace [18]. The inner product between basis vectors is defined as a canonical ensemble average according to the equilibrium configuration distribution function, $p(\mathbf{r}) \propto \exp[-u(\mathbf{r})/k_B T]$:

$$\mathcal{G}_{\zeta}(z) \odot \mathcal{G}_{\zeta'}(z') = \left\langle \sum_{i} \mathcal{G}_{i;\zeta}(\mathbf{r}; z) \cdot \mathcal{G}_{i;\zeta'}(\mathbf{r}; z') \right\rangle$$
$$= \bar{G}_{\zeta\zeta'}(z, z') + \delta_{\zeta,\zeta'} g_{\zeta}(z) \delta(z - z'),$$

where

$$g_\zeta(z) = \left\langle \sum_\lambda |\nabla \psi_{\zeta\lambda}(\mathbf{r})|^2 \delta(\psi_{\zeta\lambda}(\mathbf{r}) - z) \right\rangle$$

$$\bar{G}_{\zeta\zeta'}(z, z') = \left\langle \sum_{\lambda \neq \lambda'} (\nabla \psi_{\zeta\lambda}(\mathbf{r}) \cdot \nabla \psi_{\zeta'\lambda'}(\mathbf{r})) \delta(\psi_{\zeta\lambda}(\mathbf{r}) - z) \times \delta(\psi_{\zeta'\lambda'}(\mathbf{r}) - z') \right\rangle,$$

are structural correlation functions and $\nabla = (\partial/\partial \mathbf{r}_1, \ldots, \partial/\partial \mathbf{r}_n)$. In general, the basis vectors are not orthogonal because they correspond to correlated molecular interactions.

We derive our distribution function theory from the force balance relation:

$$k_B T \frac{\partial p(\mathbf{r})}{\partial \mathbf{r}_i} = \mathbf{f}_i(\mathbf{r}) p(\mathbf{r}). \tag{4}$$

Each side of this relation is manipulated by taking the scalar product with $\mathcal{G}_{i;\zeta}$, summing over i, and integrating over the configuration space. The resulting right hand expression is the inner product of the force field with the basis vector $\mathcal{G}_\zeta(z)$: $b_\zeta(z) = \mathcal{G}_\zeta(z) \odot \mathbf{f}$ and may be calculated directly from correlation functions involving the force field. Alternatively, $\mathbf{f}$ may be expanded according to Eq. (2) and b_ζ may be decomposed into direct and indirect contributions

$$b_\zeta(z) = \phi_\zeta(z) g_\zeta(z) + \sum_{\zeta'} \int dz' \phi_{\zeta'}(z') \bar{G}_{\zeta\zeta'}(z, z'). \tag{5}$$

Upon applying the same operations and performing integration by parts, the left hand expression of Eq. (4) becomes

$$k_B T \int d\mathbf{r} \sum_i \mathcal{G}_{i;\zeta}(\mathbf{r}; z) \cdot \frac{\partial p(\mathbf{r})}{\partial \mathbf{r}_i} = k_B T \left[\frac{d g_\zeta(z)}{dz} - L_\zeta(z) \right] \tag{6}$$

where $L_\zeta(z) = \langle \sum_\lambda \Delta \psi_{\zeta\lambda}(\mathbf{r}) \delta(\psi_{\zeta\lambda}(\mathbf{r}) - z) \rangle$ and $\Delta = \sum_i (\partial/\partial \mathbf{r}_i)^2$. Equations (5) and (6) demonstrate that the inner product of the force field with each basis vector may be expressed in terms of structural correlation functions. Equating these expressions, we obtain

$$k_B T \left[\frac{d g_\zeta(z)}{dz} - L_\zeta(z) \right] = g_\zeta(z) \phi_\zeta(z) + \sum_{\zeta'} \int dz' \bar{G}_{\zeta\zeta'}(z, z') \phi_{\zeta'}(z') \tag{7}$$

for each ζ included in Eq. (1).

Equation (7) generalizes the YBG equation by defining a system of linear integral equations relating each force function ϕ_ζ (and therefore each interaction potential u_ζ) to an associated correlation function for that mode and a set of higher order distribution functions describing the correlation between the modes in the potential. In the case of a simple monatomic fluid, the system potential is comprised of central pair potentials, λ identifies a particular pair of particles $\{i, j\}$, $\psi_{\zeta\lambda}(\mathbf{r})$ is the pair distance r_{ij}, $g_\zeta(z) = 2\langle \sum_\lambda \delta(\psi_{\zeta\lambda}(\mathbf{r}) - z) \rangle$, $L_\zeta(z) = 2 g_\zeta(z)/z$, the sums over λ and λ' incorporate the correct combinatorial factors, and Eq. (7) becomes equivalent to the YBG equation for simple liquids. However, Eq. (7) remains valid for any molecular potential that can be expressed in the form of Eq. (1). We note that the generalized YBG equation is not valid at singularities that may appear in $\nabla \psi_{\zeta\lambda}$ and $\Delta \psi_{\zeta\lambda}$.

This derivation may be generalized in the case that each force function, ϕ_ζ, is represented with a linear combination of discrete basis functions $f_{\zeta d}(z)$:

$$\phi_\zeta(z) = \sum_d \phi_{\zeta d} f_{\zeta d}(z), \tag{8}$$

where $\phi_{\zeta d}$ are the constant coefficients of the basis functions and, equivalently, discrete coefficients for the force field. The appropriate generalization of Eq. (7) is a system of coupled linear algebraic equations for each ζd combination:

$$- k_B T \left[\left\langle \sum_\lambda |\nabla \psi_{\zeta\lambda}(\mathbf{r})|^2 f'_{\zeta d}(\psi_{\zeta\lambda}(\mathbf{r})) \right\rangle + L_{\zeta d} \right] = \sum_{\zeta' d'} G_{\zeta d; \zeta' d'} \phi_{\zeta' d'}, \tag{9}$$

where $f'_{\zeta d}(z) = d f_{\zeta d}(z)/dz$, $G_{\zeta d;\zeta' d'} = \mathcal{G}_{\zeta d} \odot \mathcal{G}_{\zeta' d'}$, and $G_{\zeta d;\zeta' d'}$ and $L_{\zeta d}$ are discrete analogs of the corresponding continuous correlation functions that are defined by replacing each instance of $\delta(\psi_{\zeta\lambda}(\mathbf{r}) - z)$ with $f_{\zeta d}(\psi_{\zeta\lambda}(\mathbf{r}))$. Equations (7) and (9) readily generalize for the case that the force field includes both continuous, $\phi_\zeta(z)$, and discrete, $\phi_{\zeta d}$, parameters.

Given an appropriate set of equilibrium structural correlation functions for an unknown potential, the generalized YBG equation, Eq. (7), determines the potential functions of a given form that provide a variationally optimal approximation to the unknown potential. In the case that the true potential, $U(\mathbf{r})$, for the system is of some unknown and arbitrarily complex form and the approximate potential is assumed to be of the form given by Eq. (1), then Eq. (7) determines the interaction potentials, $u_\zeta(z)$, that minimize the positive semidefinite quadratic functional of force fields:

$$\chi^2[\mathbf{f}'] \equiv ||\mathbf{F} - \mathbf{f}'||^2 \equiv \left\langle \sum_i |\mathbf{F}_i(\mathbf{r}) - \mathbf{f}'_i(\mathbf{r})|^2 \right\rangle_U, \tag{10}$$

where $||\mathbf{A}|| = \sqrt{\mathbf{A} \odot \mathbf{A}}$ is a norm in the vector space of force fields, $\mathbf{F}$ is the force field defined by gradients of the unknown potential U with elements $\mathbf{F}_i(\mathbf{r}) = -\partial U(\mathbf{r})/\partial \mathbf{r}_i$, $\mathbf{f}'$ is a trial force field with elements $\mathbf{f}'_i(\mathbf{r})$, and the subscripted angular brackets denote a canonical ensemble average according to the Boltzmann distribution for U. If $U = u$, then $\mathbf{f}$ with elements $\mathbf{f}_i$ given by Eq. (2) provides the only force field for which $\chi^2 = 0$ and, in addition, $\mathbf{f}$ also satisfies Eq. (7). If $U \neq u$, then the force field $\mathbf{f}$ with elements given by Eq. (2) and that satisfies Eq. (7) for each ϕ_ζ provides the unique minimum of this functional for all force fields defined by potentials of the form given by Eq. (1). This follows because the force field $\mathbf{f}$ that minimizes χ^2 is the orthogonal projection of $\mathbf{F}$ onto the basis set specified by the assumed form of the approximate potential [17–19]. Therefore, $\mathcal{G}_\zeta(z) \odot \mathbf{F} = \mathcal{G}_\zeta(z) \odot \mathbf{f} = b_\zeta(z)$, which can be related to structural correlation functions as in Eq. (5). However, in this case, the correlation functions are canonical ensemble averages for the unknown potential, U.

TABLE I. Comparison of parameters for the HT model and parameters calculated directly from structural correlation functions. Type 1 dihedral angles include less than two N sites, while type 2 dihedral angles include two or more N sites.

Nonbonded Parameters				
Pair Potential	$u_\zeta(r) = C_{12}(r/a)^{-12} - C_6(r/a)^{-6}$			
Pair	Parameter	Exact	Calculated	% Error
$B - B$	C_{12}	4.00	3.98	−0.57
	C_6	4.00	3.98	−0.50
$B - L$, $L - L$	C_{12}	2.67	2.61	−1.94
	C_6	−2.67	−2.66	−0.31
$N - B$, $N - L$, $N - N$	C_{12}	4.00	3.99	−0.15
	C_6	0.00	−0.006	N/A
Bonded Parameters				
Bond Stretches	$u_\zeta(r) = \frac{1}{2}k_r(r - r_0)^2/a^2$			
all bonds	k_r	100.00	99.88	−0.12
	r_0	1.00	0.9999	−0.01
Bond Angles	$u_\zeta(\theta) = \frac{1}{2}k_\theta(\theta - \theta_0)^2$			
all angles	k_θ	20.00	19.97	−0.13
	θ_0	105.00	104.98	−0.02
Dihedral Angles	$u_\zeta(\psi) = \sum_{n=1}^{5} A_n \cos^n(\psi - 180^0)$			
dihedral type 1	A_1	2.400	2.388	0.51
	A_2	0.000	−0.028	N/A
	A_3	−4.800	−4.765	−0.73
	A_4	0.000	0.020	N/A
	A_5	0.000	−0.016	N/A
dihedral type 2	A_1	0.600	0.606	0.93
	A_2	0.000	−0.005	N/A
	A_3	−0.800	−0.801	0.08
	A_4	0.000	0.008	N/A
	A_5	0.000	0.006	N/A

The multiscale coarse-graining method [15,16] employs a force-matching variational principle [13,14] analogous to Eq. (10) and explicitly uses force correlation functions to determine the unique coarse-grained (CG) force field for a specified basis set that provides an optimal approximation to the atomistic many-body potential of mean force [17,18,20]. The preceding argument implies that this variationally optimal potential can be determined from the generalized YBG equation.

We have employed this theory to determine the interaction potentials for the Honeycutt-Thirumalai (HT) protein model [21,22] directly from canonical structural correlation functions calculated from equilibrium molecular dynamics simulations. The HT protein model is an implicit solvent CG model that represents each amino acid with a single interaction site that is one of three distinct types, either hydrophobic (B), hydrophilic (L), or neutral (N). The interaction potential for the HT model includes bond stretch, bond angle, and two distinct types of dihedral angle potentials, as well as nonbonded pair potentials between sites. Table I defines the precise form and parameters for the HT model. In Table I and subsequently, all energies are reported in terms of ε, the well depth of the attractive $B - B$ interaction, and all distances are reported in terms of the equilibrium bond length, a.

The 46 amino acid HT protein sequence $B_9N_3(LB)_4N_3B_9N_3(LB)_5L$ was simulated with GROMACS 3.3.3 [23,24], using a stochastic dynamics algorithm to generate a canonical ensemble of 10^5 configurations at a temperature approximately equal to the protein's folding

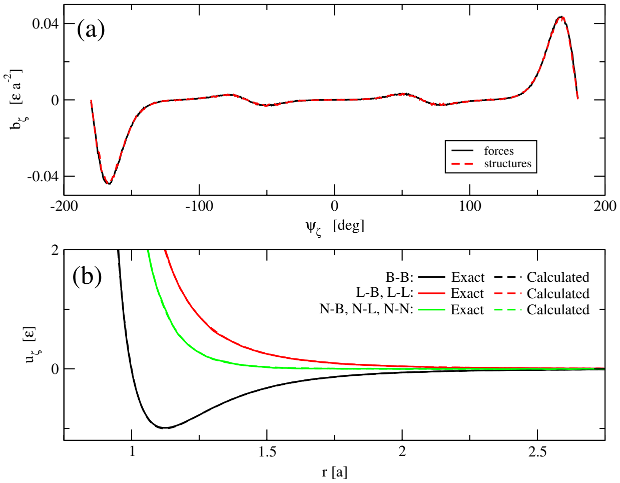

FIG. 1 (color online). Quantitative validation of the generalized YBG theory for the HT model peptide. (a) Comparison of $b_\zeta(z)$ for type 1 dihedral angles calculated using force (solid line) and structural (dashed line) correlation functions. (b) Comparison of the exact (solid line) nonbonded pair potentials with those calculated from structural correlation functions (dashed line).

temperature. At this temperature, the protein sampled both folded and also more extended conformations. We obtained similar results at both higher and lower temperatures.

As a first numerical test, we calculated each ϕ_ζ in the force field directly from Eq. (7) using structural correlation functions and without any assumptions regarding the form of the force functions. Figure 1(a) compares calculations of $b_\zeta(z)$ using forces sampled from the simulations according to $\mathcal{G}_\zeta(z) \odot \mathbf{f}$ with calculations of $b_\zeta(z)$ using Eq. (6) for ζ corresponding to the type 1 dihedral angle potential. Singularities in calculations for the dihedral angle at $\psi = 0^0, \pm 180^0$ were treated by neglecting their contribution to $L_\zeta(z)$ and by setting $dg_\zeta(z)/dz = 0$ at $\psi = \pm 180^0$. A centered running average over 11 grid points was also employed to smooth the numerical derivative, except for the first and last five grid points, which were linearly interpolated. Despite these approximations, Fig. 1(a) demonstrates that the necessary inner products of the force field are accurately determined from structural correlation functions. Figure 1(b) compares the nonbonded pair potentials employed in the original simulations with those calculated directly from structural correlation functions and demonstrates the quantitative accuracy of the calculated potentials.

As a second test, the parameters for the HT protein force field were calculated using the discrete YBG equation, Eq. (9). Table I presents the numerical errors in the calculated force field parameters and demonstrates similar quantitative accuracy. The C_6 parameter for the $B - L$ and $L - L$ pair interactions is miscalculated by -1.9%, most likely as a result of the limited sampling for these interactions. All of the other force field parameters are accurately recovered to within less than 1.0%.

The generalized YBG theory variationally determines optimized potentials directly from structural correlation functions. This theory provides a powerful method for calculating CG force fields from structure ensembles determined from, e.g., atomistic simulations or NMR measurements [25]. It provides the basis for developing approximate perturbation theories for calculating force fields from partial knowledge of the structure ensemble obtained from, e.g., small angle scattering measurements. Finally, because the generalized YBG theory exactly treats many-body correlations in protein structures, when combined with a recent theory for transferable CG models [26], the present Letter provides the foundation for variationally calculating optimal knowledge-based [27] potentials from the protein databank.

The authors gratefully acknowledge helpful conversations with Vinod Krishna, Irina Navrotskaya, Jorge Sofo, and Jayanth Banavar. W. G. N. gratefully acknowledges start-up funds from the Pennsylvania State University and also support from Research Computing and Cyberinfrastructure, a unit of Information Technology Services at Penn State.

*wnoid@chem.psu.edu

[1] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids* (Academic Press, New York, 1990), 2nd ed.
[2] S. G. Whittington and L. G. Dunfield, J. Phys. A **6**, 484 (1973).
[3] K. E. Gubbins, Chem. Phys. Lett. **76**, 329 (1980).
[4] H. C. Andersen and D. Chandler, J. Chem. Phys. **57**, 1918 (1972).
[5] H. C. Andersen and D. Chandler, J. Chem. Phys. **57**, 1930 (1972).
[6] K. S. Schweizer and J. G. Curro, Phys. Rev. Lett. **58**, 246 (1987).
[7] A. A. Louis, P. G. Bolhuis, J. P. Hansen, and E. J. Meijer, Phys. Rev. Lett. **85**, 2522 (2000).
[8] M. C. Rechtsman, F. H. Stillinger, and S. Torquato, Phys. Rev. Lett. **95**, 228301 (2005).
[9] A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **52**, 3730 (1995).
[10] F. Müller-Plathe, Chem. Phys. Chem. **3**, 754 (2002).
[11] R. L. McGreevy and L. Pusztai, Mol. Simul. **1**, 359 (1988).
[12] D. A. Keen and R. L. McGreevy, Nature (London) **344**, 423 (1990).
[13] F. Ercolessi and J. B. Adams, Europhys. Lett. **26**, 583 (1994).
[14] S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth, J. Chem. Phys. **120**, 10896 (2004).
[15] S. Izvekov and G. A. Voth, J. Phys. Chem. B **109**, 2469 (2005).
[16] S. Izvekov and G. A. Voth, J. Chem. Phys. **123**, 134105 (2005).
[17] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, J. Chem. Phys. **128**, 244114 (2008).
[18] W. G. Noid, P. Liu, Y. T. Wang, J.-W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth, J. Chem. Phys. **128**, 244115 (2008).
[19] A. J. Chorin and O. H. Hald, *Stochastic Tools in Mathematics and Science* (Springer, New York, 2006).
[20] W. G. Noid, J.-W. Chu, G. S. Ayton, and G. A. Voth, J. Phys. Chem. B **111**, 4116 (2007).
[21] J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **87**, 3526 (1990).
[22] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).
[23] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comput. Chem. **26**, 1701 (2005).
[24] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[25] K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson, and M. Vendruscolo, Nature (London) **433**, 128 (2005).
[26] J. W. Mullinax and W. G. Noid, J. Chem. Phys. **131**, 104110 (2009).
[27] J. Skolnick, Curr. Opin. Struct. Biol. **16**, 166 (2006).