THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

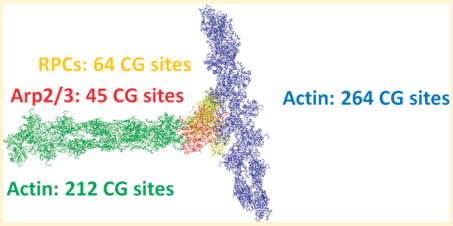
Optimal Number of Coarse-Grained Sites in Different Components of Large Biomolecular Complexes

Anton V. Sinitskiy, Marissa G. Saunders, and Gregory A. Voth*

Department of Chemistry, Institute for Biophysical Dynamics, James Franck Institute, and Computation Institute, University of Chicago, Chicago, Illinois 60637, United States

Supporting Information

ABSTRACT: The computational study of large biomolecular complexes (molecular machines, cytoskeletal filaments, etc.) is a formidable challenge facing computational biophysics and biology. To achieve biologically relevant length and time scales, coarse-grained (CG) models of such complexes usually must be built and employed. One of the important early stages in this approach is to determine an optimal number of CG sites in different constituents of a complex. This work presents a systematic approach to this problem. First, a universal scaling law is derived and numerically corroborated for the intensity of the intrasite (intradomain) thermal fluctuations as a function of the number of CG sites. Second, this result is used for derivation of the criterion for the optimal number of CG sites in different parts of a large multibiomolecule complex. In the zeroth-order approximation, this approach validates the empirical rule of taking one CG site per fixed number of atoms or residues in each biomolecule, previously widely used for smaller systems (e.g., individual biomolecules). The first-order corrections to this rule are derived and numerically checked by the case studies of the *Escherichia coli* ribosome and Arp2/3 actin filament junction. In different ribosomal proteins, the optimal number of amino acids per CG site is shown to differ by a factor of 3.5, and an even wider spread may exist in other large biomolecular complexes. Therefore, the method proposed in this paper is valuable for the optimal construction of CG models of such complexes.



1. INTRODUCTION

Modeling large heterogeneous biomolecular complexes, such as molecular machines or cytoskeletal filaments, is a significant challenge.[1,2] Experimental methods (X-ray crystallography, cryoelectron microscopy, etc.) provide critical but sometimes somewhat limited information because, for example, these complexes are dynamic in nature, have finite lifetimes, vary in their composition over time, and may be difficult to crystallize. In many cases, experimental data are supplemented by computational studies to yield detailed all-atom models.[3−8] On the other hand, molecular dynamics (MD) simulations of these systems on biologically relevant temporal scales (milliseconds to minutes) are often beyond the capability of present day computers.[9−12] For this reason, it is desirable to build and study coarse-grained (CG) models of large biomolecular complexes. Since it is possible to study individual components of such complexes (i.e., isolated proteins or nucleic acids), it is valuable to also assemble larger-scale CG models of the complexes out of CG models of their constituents.

One of the first steps in this process is to choose the number of CG sites in each part of the complex, a problem that does not seem to have been systematically addressed to date. There are several general techniques for choosing the number of CG sites that can be applied to the particular case of constituents of large biomolecular complexes. Among those techniques, the most widespread approach is to introduce one CG site per fixed number of atoms or monomers; for example, one CG site per amino acid residue in proteins, one CG site per four "heavy" (non-hydrogen) atoms in lipids, or three CG sites per nucleotide in nucleic acids.[13−20] A second method is to choose the minimal number of CG sites that retains the overall shape of a biomolecule; this is often used for membrane-forming lipids.[15,21,22] In addition, the number of CG sites may be related to the number of domains, subdomains, or motifs defined in a protein based on traditional biochemical criteria (e.g., functional or dynamical autonomy, cross-species conservation, occurrence in different proteins, etc.).[23] Using principal component analysis (PCA)[24] of all-atom MD trajectories, one can choose the number of essential degrees of freedom on the basis of consideration of the fraction of variance of atomic coordinates ("percentage of information") retained by different numbers of principal components.[25−33] The number of degrees of freedom desired could then be used to determine how many CG sites are necessary. A method based on maximizing the difference of the total number of atoms in all domains and the number of atoms in the largest domain has also been proposed.[34,35]

All the above methods share a number of shortcomings. First, they are not fully theoretically grounded by the methods of statistical mechanics, and their implications have not been

Special Issue: Macromolecular Systems Understood through Multiscale and Enhanced Sampling Techniques

Received: November 12, 2011
Revised: January 22, 2012
Published: January 25, 2012

   dx.doi.org/10.1021/jp2108895 | *J. Phys. Chem. B* 2012, 116, 8363−8374

systematically studied from that viewpoint. Second, it is unclear how transferable these rules are between different complexes of the same protein or between different time snapshots of a single MD simulation (except for simple approaches, such as one CG site per amino acid residue). Third, they do not offer a unified treatment equally applicable to different types of biomolecules: proteins, RNAs, lipids, etc. Fourth, many of them require substantial human intervention and thus cannot be automated. Finally, it is not clear whether all these methods lead to consistent results and, if not, then how to evaluate which one is the best for a given problem.

The present paper addresses the problem of choosing the optimal number of CG sites. The proposed approach is based on universal characteristics revealed in the scaling of a variational residual χ^2, which have been previously defined within the context of the essential dynamics coarse-graining (ED-CG) method,[36] as a function of the number of CG sites and the number of atoms in the molecule of interest. Using these scaling laws allows us to introduce, by means of the Lagrange multiplier formalism, a global parameter (L^2) that regulates the number of CG sites in each biomolecule in the complex in a way somewhat analogous to regulation of the number of particles in an open system by the chemical potential. This approach offers both fundamental and practical advantages. It is strictly derived from a general principle with transparent connections to statistical mechanics, and it provides a unified approach to various types of biomolecules. The computational cost is much lower than that of explicit brute-force coarse-graining of the whole system, and it scales linearly with the size of the system. Finally, this approach justifies the empirical rule of taking one CG site per fixed number of atoms or residues in each molecule as the zeroth-order estimation, at the same time providing the higher-order corrections to this rough estimate.

The remainder of this article is organized as follows: First, a theoretical derivation of the power law scaling of the residual χ^2 defined in the ED-CG method is presented. Then this result is used to derive the criterion for the optimal number of CG sites. Third, numerical estimations of the validity of the scaling law and the criteria for the choice of the number of CG sites, including the approximation of taking one CG site per constant number of monomers, are provided, and comparative analysis with PCA-based MD approaches (such as 95% rule) is performed. The paper ends with final concluding remarks.

2. THEORY AND METHODS

Summary of ED-CG. In the ED-CG method, CG sites (or, equivalently, domains) are defined as the parts of a molecular system such that the atoms belonging to the same site move in a correlated manner under the approximation of Gaussian thermal fluctuations. Technically, coarse-graining of a molecular system for a given total number of CG sites n is performed by means of minimization of the variational residual χ^2 defined as:[36]

$$\chi^2 = \frac{1}{3n}\sum_{I=1}^{n}\left\langle \sum_{i\in I}\sum_{j\geq i\in I} |\Delta\vec{r}_i(t) - \Delta\vec{r}_j(t)|^2 \right\rangle_t \qquad (1)$$

where I numerates CG sites, $i \in I$ means that atom i belongs to the CG site number I, t denotes the frames in the MD trajectory, Δr_i is the deviation of atom i from its average position, and the triangular brackets denote averaging over t. The value of χ^2, therefore, in some ways characterizes the amount of intrasite (intradomain) thermal disorder (noise) averaged over all CG sites, and defining the CG mapping by the minimization of χ^2 provides the most deterministic (the least noisy) representation of the biomolecular system possible for the given finite temperature and for the given total number of CG sites, n. This procedure can effectively project out high-frequency noise and leave only the functionally relevant low-frequency, large-scale, thermally accessible motions. Performing unconstrained minimization of χ^2 scales exponentially with respect to the number of atoms in the system; hence, a number of additional restrictions have been proposed to guide the search for the minimum in the sequence-based[36] or space-based[37] ED-CG methods. In the latter, the computational cost scales proportionally to the square of the number of atoms and is not particularly sensitive to the number of CG sites. However, even this scaling makes brute-force coarse-graining of large biomolecular complexes computationally challenging (for an example, the case of the ribosome is given in ref 11.).

In the ED-CG approach, the CG model is constructed on the basis of the thermal fluctuations of a given biomolecule. It may not guarantee that the same CG mapping provides the best CG approximation for the *inter*molecular potentials. However, in the most general case, if no specific information on the intermolecular potentials is available, this approach is justified by the following argument in the spirit of linear response theory: Because of a strong correlation of the spatial movements of atoms from the same CG site, the covariance matrix for the Cartesian coordinates of the atoms should be close to a block-diagonal form, with the blocks formed by the atoms assigned to the same CG sites. In the harmonic approximation, the covariance matrix equals, up to a factor of k_BT, the inverse of the Hessian matrix.[37] The response of the molecule to a perturbation (namely, to intermolecular interactions or to a constant external potential) is determined in the lowest order by the product of the inverse of the unperturbed Hessian matrix and the vector of external forces. Because of the nearly block-diagonal structure of the inverse Hessian, the dynamics of each atom in the biomolecule is therefore determined mainly by the external intermolecular forces exerted on the atoms from the *same* CG site. Thus, the relatively autonomous CG domains in the biomolecule defined based on the thermal fluctuations in the unbound state remains applicable, in the limit of weak interactions, for the same biomolecule when in a bound state or placed into an external potential.

Power Law Scaling of χ^2. To analyze the behavior of χ^2 with changes in the number of CG sites and of atoms in the system, the ED-CG method developed for low-resolution structural data[37] can be employed. This method naturally includes the concept of changing the number of pseudoatoms, N_{atoms}, in a biomolecule, while the nature of the biomolecule is kept unaltered (for a discussion of pseudoatoms, see ref 37.). Neglecting the details of the distribution of atoms in space, for a fixed number of CG sites, n, an increase in N_{atoms} by a factor of k leads roughly to a proportional increase in the number of atoms per CG site and, therefore, to rescaling of χ^2 by a factor of k^2, since the definition (1) includes summation over indices i and j running over all atoms in each CG site:

$$n \to n,\ N_{atoms} \to kN_{atoms} \Rightarrow \chi^2 \to k^2\chi^2 \qquad (2)$$

dx.doi.org/10.1021/jp2108895 | J. Phys. Chem. B 2012, 116, 8363−8374

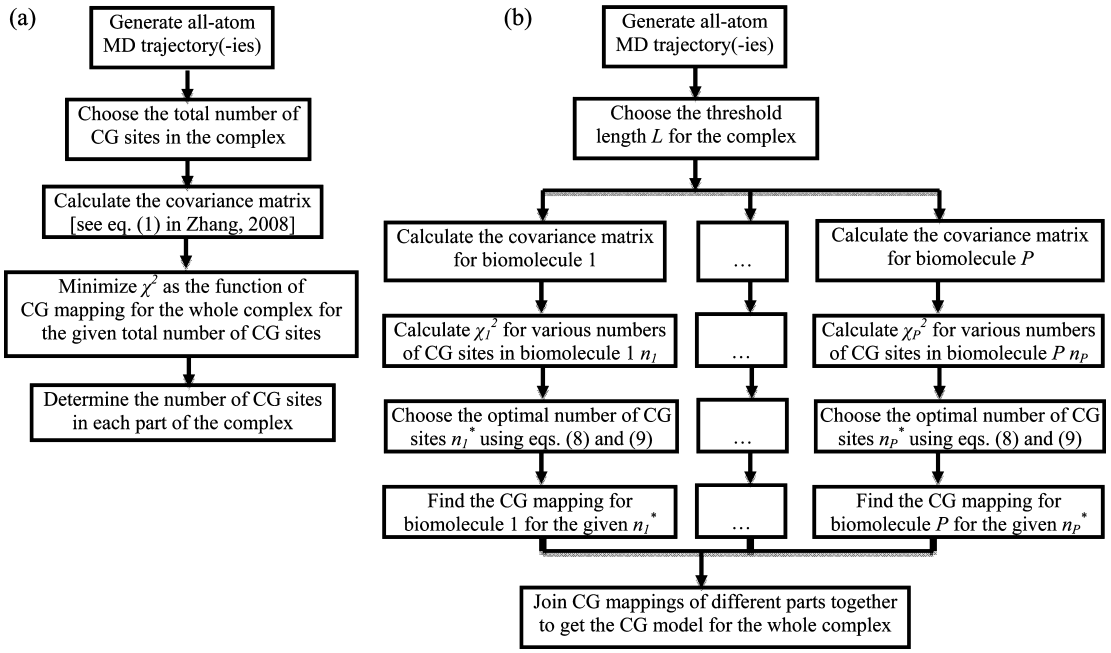


Figure 1. Two brute-force methods of building CG models of large biomolecular complexes were considered: (a) a sequential algorithm based on definition 1, and (b) a parallelizable algorithm based on definition 2.

The simultaneous increase in the number of CG sites and the number of atoms by a factor of k would keep χ^2 roughly unchanged. Indeed, the average number of atoms per site does not change; therefore, the double sum in the triangular brackets in eq 1 is not rescaled; the sum over I now includes k times more terms, but this effect is canceled out due to division of the sum by the number of CG sites:

$$n \to kn, \qquad N_{\text{atoms}} \to kN_{\text{atoms}} \Rightarrow \chi^2 \to \chi^2 \tag{3}$$

Combining scaling laws 2 and 3 leads to the following approximation:

$$\chi^2 \approx \frac{C(T){N_{\text{atoms}}}^2}{n^2} \tag{4}$$

where $C(T)$ is a constant, independent of the number of atoms and the number of CG sites (it should still be temperature-dependent, since χ^2 is a measure of the intrasite thermal fluctuations).

Equations 2 and 3 provide only a general estimation of the scaling behavior, and thus, neither they nor eq 4 are expected to be exact. Anticipating the numerical results discussed in detail in the next section, one may suggest that in real systems, the power law scaling of χ^2 as a function of N_{atoms} and n still holds, although the scaling exponents may be different. For these reasons, in the subsequent derivation, the following generalized version of eq 4 will be used:

$$\chi^2 = \frac{C(T){N_{\text{atoms}}}^{2+\gamma+\delta}}{n^{2+\gamma}} \tag{5}$$

where γ and δ are (presumably small) constants. We call γ "the anomalous dimension", analogous to the scaling of correlators in quantum field theory.[38] The anomalous dimensions γ can be system-dependent, since fitting the dependence of χ^2 on n with a power law function leads to different values of the exponents for different biomolecules (for details, see the Supporting Information (SI), section S1). On the other hand, the power law dependence of the prefactor in these biomolecule-specific approximations for $\chi^2(n)$ on N_{atoms} is by construction universal for all biomolecules, and therefore, the parameter δ is system-independent. At present, the origin of nonzero values of γ and δ has no clear theoretical justification, and they are introduced as empirical parameters. The choice of the functional form in eq 5 among various possible generalizations of eq 4 is discussed in the SI, section S1.

Definitions of the Optimal Number of CG Sites. There are several different approaches to determining how many CG sites to place in each part of a biomolecular complex and how to optimally place these CGs. Below, two methods are outlined:

Definition 1. Consider the whole complex of biomolecules as a single system and perform brute-force coarse-graining by the ED-CG method, minimizing the overall χ^2_{total} value. The resulting CG domains may contain atoms from different parts of the complex at the same time. To determine the number of CG sites in a particular biomolecule of that complex, the CG sites that span two or more parts of the complex must be appropriately apportioned. We can split the CG site on the basis of the fraction of atoms in the biomolecule of interest relative to the total number of atoms in the CG site.

Definition 2. Coarse-graining is again performed by means of minimization of the overall ${\chi_{\text{total}}}^2$, but the set of CG mappings is restricted such that every CG domain lies entirely within one biomolecule of the complex. In this case, the number of CG domains in each constituent of the complex is an integer and can be determined more simply.

Definition 1 does not provide CG domains localized within a single biomolecule, and definition 2 may not lead to the global minimum of the χ^2_{total} value. However, these definitions should lead to similar results, insofar as the complex under investigation is *a complex* indeed (an assembly of relatively autonomous units rather than a single highly correlated unit without modular structure). In this limit, thermal movements of any two atoms belonging to two different constituents will be uncorrelated or weakly correlated, and minimization of ${\chi_{\text{total}}}^2$

dx.doi.org/10.1021/jp2108895 | J. Phys. Chem. B 2012, 116, 8363−8374

will always place these atoms into two different CG sites. Note that the strict absence of correlation between molecules is not required. If two biomolecules in a complex are strongly coupled in terms of the correlation of atomic movements, they should be considered as a single constituent of the complex for the purpose of the current consideration. Practically, definition 2 has the added advantage of being compatible with complex disassembly. If two parts of a complex are mapped onto a single CG site, these two parts will not be able to dissociate unless the resolution of the CG model is increased. Comparison of the numerical results of these two approaches for the cases of *Escherichia coli* ribosome and actin filaments junction is presented in Section 3.

Derivation of the Criterion for the Optimal Number of CG Sites. Definition 2 allows one to calculate the number of CG sites more efficiently than by using the algorithm given in definition 2 itself. Since this definition requires that every CG site be localized within one biomolecule of a complex, the overall χ_{total}^2, according to eq 1, splits into the sum of the residuals χ^2 for each of the parts of the complex:

$$\chi_{\text{total}}^2 = \frac{1}{N}\sum_{i=1}^{P} n_i \chi_i^2 \tag{6}$$

where P is the number of parts (constituents) in the system, n_i is the number of CG sites in the ith part, and N is the total number of CG sites in the system:

$$N = \sum_{i=1}^{P} n_i \tag{7}$$

Then the optimal number, n_i^* of CG sites in each of the parts for a given total number of CG sites, N, can be determined by minimization [constrained by eq 7] of χ_{total}^2 as a function of (n_1, ..., n_P). Note that by doing so, the initial problem of minimization of χ_{total}^2 as a function of the CG mapping is split into P+1 separate problems: minimization of χ_{total}^2 as a function of only the number of CG sites in each biomolecule in the system, and P separate problems of minimization of χ_i^2 for each of those biomolecules, to obtain the CG mapping. Comparison of the algorithm for brute-force coarse-graining (according to definition 1) and the fast algorithm described in this subsection (leading to the same results as definition 2) is shown in Figure 1.

Applying the discrete-variable version[39] of the Lagrange multipliers method leads to the following result for each of the parts of the system:

$$n_i^* = \arg\min_n \left\{ n\chi_i^2(n) + \lambda n \right\} \tag{8}$$

where λ is the Lagrange multiplier—the continuous parameter, simultaneously regulating the number of CG sites in all parts of the system. $\lambda > 0$; otherwise, eq 8 may not have a reasonable solution. It is convenient to represent λ as L^2, with L being interpreted as a "threshold length", since it has the units of distance and functions as a threshold for determining the optimal number of sites n^*. Then condition 8 resolves into the eq 9 system of equations

$$\begin{cases} \displaystyle\sum_{I=1}^{n_i^*-1} \left\langle \sum_{i\in I}\sum_{j\geq i\in I} |\Delta\bar{r}_i(t) - \Delta\bar{r}_j(t)|^2 \right\rangle_t \\ \quad \displaystyle\sum_{I=1}^{\bar{n}_i^*} \left\langle \sum_{i\in I}\sum_{j\geq i\in I} |\Delta\bar{r}_i(t) - \Delta\bar{r}_j(t)|^2 \right\rangle_t \geq 3L^2 \\ \displaystyle\sum_{I=1}^{n_i^*} \left\langle \sum_{i\in I}\sum_{j\geq i\in I} |\Delta\bar{r}_i(t) - \Delta\bar{r}_j(t)|^2 \right\rangle_t \\ \quad \displaystyle\sum_{I=1}^{\bar{n}_i^*+1} \left\langle \sum_{i\in I}\sum_{j\geq i\in I} |\Delta\bar{r}_i(t) - \Delta\bar{r}_j(t)|^2 \right\rangle_t \leq 3L^2 \end{cases} \tag{9}$$

for each part in the system. L^2 is then a cutoff for the minimum improvement in $n\chi^2(n)$, and n^* is the largest value for which increasing the number of CG sites from $n^* - 1$ to n* improves $n\chi^2(n)$ by at least $3L^2$. Note that eq 9 provides only the necessary conditions for the optimum, whereas the original eq 8 provides both necessary and sufficient conditions. In this framework, the Lagrange multiplier, L^2, plays a role analogous to that of the chemical potential in the grand canonical ensemble: L^2 regulates the number of CG sites in a biomolecule or assembly of biomolecules.

To this point, the formulas in this subsection have been exact within the approach of definition 2 for the optimal number of CG sites in a complex. Further simplification is possible if the value of χ^2 is approximated by eq 5 and if n is considered as a continuous variable. First, setting the derivative of the expression to be minimized in eq 8 to zero and taking into account the functional form (5), yields the following relation:

$$(1 + \gamma_i)\,\chi_i^2(n_i^*) = L^2 \tag{10}$$

Naïvely, one might expect that the average intrasite disorder χ^2 from the optimal CG mapping would be the same in different parts of a biomolecular complex. However, eq 10 shows that the molecules with higher anomalous dimension γ should have a lower residual χ^2 and, therefore, should contain more CG sites. The explicit estimate for the optimal number of CG sites incorporating the effect of the anomalous dimension is

$$n_i^* = \left(\frac{(1 + \gamma_i) C N_{i,\text{atoms}}^{\,2+\delta+\gamma_i}}{L^2} \right)^{1/(2+\gamma_i)} \tag{11}$$

Further, the average number of atoms per CG site for a given biomolecule is given by

$$\frac{N_{i,\text{atoms}}}{n_i^*} = \left(\frac{L}{\sqrt{(1 + \gamma_i)C}} \right)^{2/(2+\gamma_i)} (N_{i,\text{atoms}})^{-\delta/(2+\gamma_i)} \tag{12}$$

dx.doi.org/10.1021/jp2108895 | J. Phys. Chem. B 2012, 116, 8363−8374

The number of atoms per CG site is the same for different parts of the biomolecular complex if and only if the following two conditions are satisfied:

(1) the anomalous dimensions, γ_i, of all constituents are equal to each other and
(2) either the number of atoms, $N_{i,\mathrm{atoms}}$, in all constituents is the same or the parameter δ equals zero.

These two conditions form the assumptions underlying the methodology of taking one CG site per fixed number of atoms or monomers in different parts of the complex. In the case that the anomalous dimensions, γ_i, or numbers of atoms, $N_{i,\mathrm{atoms}}$, of each of the parts of the biomolecules are not equal, the number of atoms per CG site will vary in different parts of the complex. To find the corrections to the above-mentioned empirical rule, one can perform a Taylor series expansion of eq 12, expanding about $\delta(\ln N_{i,\mathrm{atoms}} - \overline{\ln N_{\mathrm{atoms}}})$ and $(\gamma - \overline{\gamma})$, where $\overline{\gamma}$ and $\overline{\ln N_{\mathrm{atoms}}}$ are typical values of γ_i and $\ln N_{i,\mathrm{atoms}}$ respectively. This leads to

$$\frac{N_{i,\mathrm{atoms}}}{n_i} = C''(\overline{\gamma}, \overline{\ln N_{\mathrm{atoms}}}, L, C) \times \left\{ 1 - \frac{1}{2+\overline{\gamma}}\delta(\ln N_{i,\mathrm{atoms}} - \overline{\ln N_{\mathrm{atoms}}}) - \frac{\ln\left(\frac{N_{i,\mathrm{atoms}}}{n_i}\right) + \frac{1}{1+\overline{\gamma}}}{2+\overline{\gamma}}(\gamma - \overline{\gamma}) + \text{higher order terms} \right\} \quad (13)$$

Therefore, in the first-order approximation, the correction to the approximation of taking one CG site per fixed number of atoms splits into the sum of two independent contributions. One of these stems from the deviation of the anomalous dimension of the specific part from the average anomalous dimension for all parts of the complex, and the other correction comes from the heterogeneity in the sizes of the parts of the complex.

Molecular Dynamics Simulations. To verify the power law scaling, eq 5, of the residual intrasite disorder χ^2 and to estimate the corresponding parameters C, γ, and δ, all-atom MD trajectories of a number of unbound proteins and biomolecular complexes were used. The unbound proteins were chosen to sample various classes of biological functions and different numbers of domains defined in the existing literature.

Ubiquitin (76 Amino Acid (aa) Residues). The initial geometry was taken from the X-ray study of crystallized human erythrocytic ubiquitin[40] (1UBQ). Hydrogen atoms, pre-equilibrated TIP3P water molecules and K^+ and Cl^- ions in the amount corresponding to typical cytosolic ionic strength (0.15 M) were added using the standard tools of VMD software.[41] The resulting system contained ~23 000 atoms. MD simulations were performed with the NAMD suite of programs,[42] in the NPT (constant *n*umber, *p*ressure, and *t*emperature) ensemble using the Langevin thermostat and barostat,[43,44] at 310 K and 1 atm. The simulation was equilibrated for 3 ns (on the basis of rmsd analysis, the equilibration took ~1 ns), followed by 100 ns of production simulation. Three independent parallel MD trajectories were generated. No unfolding was observed: the C_α rmsd averaged ~2 Å relative to the initial geometry and ~1 Å relative to the averaged geometry.

Lysozyme (129 aa). The initial geometry was taken from the X-ray study of triclinic crystals of hen egg-white lysozyme[45] (3LZT). After addition of hydrogen atoms, water molecules, and ions, the system contained ~28 000 atoms. Each of the three independent MD trajectories was equilibrated for 6 ns, after which production simulations continued for 100 ns. The structures remained stable: the C_α rmsd averaged ~1−2 Å relative to the initial geometry and ~1 Å relative to the averaged geometry. Other conditions were the same as above.

Pyruvate Kinase (498 aa). The initial geometry was taken from the X-ray study of *L. mexicana*[46] (1PKL, chain G). After addition of hydrogen atoms, water molecules, and ions, the system contained ~85 000 atoms. In each of the three independent MD trajectories, the system was equilibrated for 30 ns (equilibration time from rmsd analysis ~ 7 ns), after which the simulations continued for 100 ns. The structures remained stable: the C_α rmsd averaged ~2−4 Å relative to the initial geometry and ~1−2 Å relative to the averaged geometry. Other conditions were the same as above.

G-Actin (375 aa). Six parallel trajectories, 50 ns long, were generated starting from the crystal structure for ADP-bound actin with a folded D-loop[47] (1J6Z). Simulation setup was performed as has been described previously,[48] changing the cation to Mg^{2+} and including the crystal structure waters within 10 Å of the cation. The system was protonated, solvated, and ionized as described above, followed by energy minimization, heating, and equilibration for 100 ps in the NVT (constant *n*umber, *v*olume, and *t*emperature) ensemble and 200 ps in the NPT ensemble, constraining the protein backbone throughout. Simulation was performed as described above. Structures remained stable: the C_α rmsd averaged ~2 Å relative to the initial geometry and ~1 Å relative to the averaged geometry. The first 8 ns of the trajectory was excluded from further analysis.

One 50-ns MD trajectory of actin filament branch junction Arp2/3 (31 proteins, 10980 a/a in total) was taken from ref 49. ("branch10" model), and one 100-ns MD trajectory of *E. coli* ribosome (61 proteins and RNAs, 10 986 aa and nucleotides in total) was taken from ref 11. Please see the original papers for details.

3. RESULTS AND DISCUSSION

Scaling Behavior of χ^2. For each of the MD trajectories of the unbound proteins, CG models were built by the space-based ED-CG method[37] using only the coordinates of C_α atoms in the interest of computational efficiency. The corresponding χ^2 values were plotted against the number of CG sites in log−log coordinates (Figure 2a). The slope of the graphs for different molecules is statistically different. The coefficient of determination, R^2, for linear fitting of the log−log transformed data from each of the trajectories was determined to be at least 0.9958, with a median value of 0.9989, suggesting

dx.doi.org/10.1021/jp2108895 | J. Phys. Chem. B 2012, 116, 8363−8374

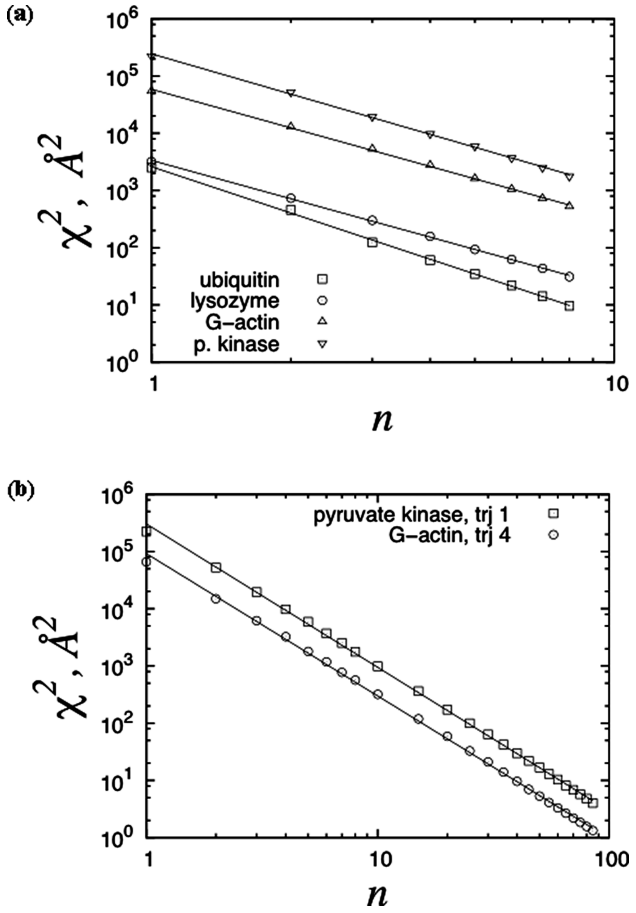


Figure 2. (a) The dependence of the residual intradomain fluctuation on the number of CG sites is linear in log–log coordinates for various proteins, which confirms the scaling law, eq 14. (b) The linearity in log–log coordinates holds over a wide range of n values.

that the scaling law below describes the behavior of χ^2 with a high level of precision.

$$\chi^2 = \frac{C'(N_{\text{atoms}})}{n^{2+\gamma}} \tag{14}$$

The component parts of the actin filament junction and the ribosome follow this equation as accurately as the whole complex. To verify this, each biomolecule within the Arp2/3 junction and the ribosome was treated in the same way as the unbound proteins. In these cases, the values of R^2 ranged between 0.985 and 0.99998, with a median value of 0.9994.

To demonstrate that the power law scaling holds in a wide range of the number of CG sites, additional calculations were performed for 10–85 CG sites per protein, using one MD trajectory for pyruvate kinase and one MD trajectory for G-actin. For each protein, a typical trajectory (that with median values of χ^2) was used. The results confirm the power law scaling of $\chi^2(n)$ (Figure 2b).

The anomalous dimensions were calculated from the slopes of the linearized log–log transformations. In all the systems studied, its value was in the range $0 < \gamma < 1$. No statistically significant correlation between the anomalous dimension and the size of the corresponding biomolecule was found.

The power law scaling of the χ^2 values has an important consequence for the physical interpretation of the CG domains. It is often suggested that biomolecules (especially multidomain proteins) can be represented as objects consisting of nearly rigid domains that move with respect to each other.[31,50–52] In this case, one might expect that the plots for χ^2 vs the number of CG sites, n, would have the following behavior. For $n < n^*$, where n^* is the true number of such rigid domains, the residual values of χ^2 would be relatively high, since a system of n^* rigid bodies cannot be adequately represented by a model with a smaller number n of rigid bodies. At $n = n^*$, the value of χ^2 would decrease to a small [in comparison to $\chi^2(n^* - 1)$] positive value and would remain small for $n > n^*$ (see Figure 3).

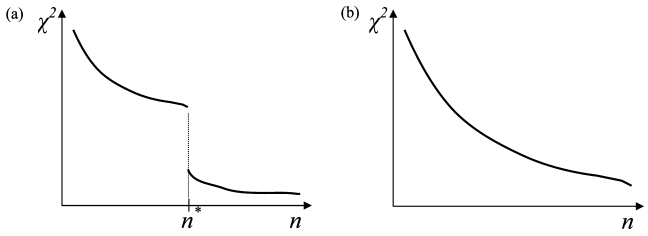


Figure 3. Comparison of the $\chi^2(n)$ curves (a) expected for n^* nearly rigid-body domains in a biomolecule and (b) calculated from all-atom MD simulations demonstrates that the dynamics of biomolecules cannot be properly described by a single model with a fixed number of CG domains. (For proper comparison with Figure 2, note that the latter is plotted in log–log coordinates.)

The plots with the described behavior are not scale-free with respect to the variable n, in the sense that simultaneous multiplication of the variables χ^2 and n by any constants k_1 and k_2 other than unity will change the shape of the curve. The sharp decrease in χ^2 would shift from n^* to $k_2 n^*$, a change that cannot be compensated for by the choice of k_1. In Figure 2, the power law scaling behavior of $\chi^2(n)$ is shown to be scale-free (similar to examples of scale-free behavior as seen in natural, social, and technical systems[53]), suggesting that biomolecular complexes may possess a complex, fractal-like dynamic structure. (In comparing Figures 2 and 3, note that the former is given in log–log coordinates.)

Instead of a set of some fixed number of nearly rigid domains, a biomolecule from the dynamical viewpoint could be represented by a hierarchy of models with different levels of resolution, such that a CG site in a coarser model can be resolved into several CG sites in a finer model, and the system does not have any internal scale for the "best" level of resolution. To the best of our knowledge, the CG models of biomolecules have not been considered from this viewpoint so far, although a number of examples of fractal-like biological systems have been published[54,55] and multiscale organization of proteins has been studied from the structural viewpont.[56]

According to eq 5, the dependence of the value C' from eq 14 on the number of atoms, N_{atoms}, should follow the power law, as well. This approximation appears to be less accurate than eq 14. To estimate the parameters C, γ_i and δ in eq 5, the least-squares multidimensional treatment of all available data (in total, 611 data points) on the χ^2 values was performed (see details in the SI, S1). The model has a residual standard error (the error in predicting ln χ^2) of 0.12, and $R^2 = 0.96$. Some of the values of the coefficients, together with their standard deviations, are given in Table 1; the full information is provided in the SI, Table S1.

The values of γ_i estimated in this way are in fair agreement with those estimated from separate treatment of each molecule with the use of eq 14. The only exceptions are the anomalous

dx.doi.org/10.1021/jp2108895 | J. Phys. Chem. B 2012, 116, 8363–8374

Table 1. The Values of the Parameter δ and the Anomalous Dimensions, γ_i, for Different Unbound Proteins and for the Constituents of the *E. coli* Ribosome and the Actin Filament Arp2/3 Junction, Calculated by the Method of Least Squares from the Linearized Version of Eq 5[a]

parameter δ	0.35 (0.05)
anomalous dimensions γ_i:	
ubiquitin	0.50 (0.01)
lysozyme	0.37 (0.01)
G-actin	0.33 (0.02)
pyruvate kinase	0.42 (0.02)
parts of *E. coli* ribosome	0−0.91
parts of actin filaments junction	0.25−0.61

[a]Standard deviations are shown in parentheses. Additional data is provided in the SI, Table S1.

dimensions for 23S rRNA and 16S rRNA in the ribosome, which are both less than zero: −0.14 (0.09) and −0.09 (0.08), where standard deviations are shown in parentheses. Note, however, that the hypotheses that both of them are ≥0 cannot be rejected at the credibility level of 10% and 27%, respectively. Therefore, it is not incorrect to claim that the anomalous dimensions in all the cases studied are restricted by the relation $0 \leq \gamma < 1$. All other coefficients in the model are nonzero, with a credibility level of at least 0.1% (for one of the tRNAs in the ribosome, 0.6%). It is of note that in the case of actin, the value of γ is fairly transferable between the unbound state and the actin filament junction: 0.33 in the first case and an average of 0.32 (between 0.25 and 0.45), in the second case.

The Exact Number of CG Sites. To calculate the optimal number of CG sites in different parts of the multiprotein complex systems under investigation (the *E. coli* ribosome and the actin filaments junction), we carried out brute-force direct coarse-graining of the MD trajectories by the space-based ED-CG method. This took more than 45 000 CPU hours for each complex, and the lowest values of χ^2 were achieved in only one or two of at least 100 000 independent runs of simulated annealing (see details on the methodology in ref 37.). For each complex, the three runs of simulated annealing with the lowest values of χ^2 were chosen and used to calculate the corresponding numbers of CG sites (based on definition 1) for each molecule in the complex. In most cases, the number of CG sites from each run were close to each other, suggesting that the obtained estimates for the number of CG sites in each part of the complexes should be close to those for the global minima of χ^2.

The variation in the numbers of CG sites obtained from the three lowest χ^2 simulated annealing sessions was used to estimate the error in this calculation. In the case of the ribosome, brute force coarse-graining was performed using 480 CG sites (chosen to be the same as in ref 11.). The results are shown in Table 2 (abbreviated version) and in the SI, Table S2 (full version).

In addition, a separate coarse-graining calculation was performed for the last 40 ns of the trajectory. The numbers of CG sites found in different parts of the ribosome using the 40 ns fragment compared with the 100 ns trajectory agree reasonably well, typically differing by no more than 2 CG sites for the L and S proteins. The two largest constituents differ more: 72 vs 76 CG sites in 16S RNA and 129 vs 132 CG sites in 23S RNA. This difference can serve as an additional error estimate for the brute-force coarse-graining calculation. In the case of the actin junction, brute-force coarse-graining was carried using a total of 585 CG sites in the complex. This value was chosen to allow about 20 CG sites in each actin monomer, a level of detail corresponding to that in the ribosome. The results are shown in Table 2 (abbreviated version) and in the SI, Table S3 (full version).

The numbers of CG sites were also calculated using definition 2 for the ribosome and the actin filament junction by means of the accelerated algorithm provided by eqs 8 and 9. First, coarse-graining of every biomolecule in the complexes was performed for a range of numbers of CG sites, and the

Table 2. The Number of CG Sites in Different Parts of the Ribosome and the Actin Filament Arp2/3 Junction Calculated by Means of Different Methods[a]

	exact results				standard methods of PCA		
fragment	definition 1	definition 2	fixed number of monomers per CG site	with first-order correction, eq. 13	*X*% rule*	95% rule	Kaiser rule
			Ribosome (*X% = 88.5%)				
23S rRNA	134 ± 7	129	127	138	19	94	37
16S rRNA	78 ± 6	72	67	73	19	80	29
other RNAs	15 ± 4	18	13	20	29	56	23
L proteins	162 ± 4	167	169	163	251	472	225
S-proteins	90 ± 3	94	104	86	162	309	147
total	480	480	480	480	480	1011	461
			Actin Filament Junction Arp2/3 (*X% = 93.7%)				
24 actins	471 ± 17	476	480	474	463	583	331
Arp2	19 ± 1	21	21	23	25	31	16
Arp3	23 ± 6	24	22	26	13	17	11
RPC1	24 ± 6	17	20	18	24	30	15
RPC2	19 ± 6	18	16	17	18	22	12
RPC3	10 ± 2	10	9	9	14	17	10
RPC4	10 ± 2	8	9	7	18	21	10
RPC5	11 ± 4	11	8	11	10	12	8
total	585	585	585	585	585	733	413

[a]The percentage, *X*, was chosen such that the total number of CG sites predicted by PCA was equal to the number of sites selected in this paper for the given system. Additional data provided in the SI, Tables S2 and S3.

dx.doi.org/10.1021/jp2108895 | J. Phys. Chem. B 2012, 116, 8363−8374

values of $\chi^2(n)$ were determined. The threshold length, L, was iteratively adjusted in such a way that the sum of the numbers of CG sites in different parts of the ribosome and actin filament junction would equal 480 (L = 11.17 Å) and 585 (L = 9.85 Å), respectively. The number of CG sites in each part of the complex is given in Table 2 (abbreviated version) and in the SI, Tables S2 and S3 (full version). Note that the use of eqs 8 and 9 instead of the algorithm used in definition 2 made calculations an order of magnitude faster. The computational cost of space-based coarse-graining grows proportionally to the square of the size of the system. Therefore, coarse-graining each of k parts of equal size of a complex would take k^2 less time than brute-force coarse-graining of the complex as a whole; since k coarse-graining calculations must be solved in parallel, coarse-graining the complex by parts using eqs 8 and 9 takes k times less time than coarse-graining of the whole complex.

The numbers of CG sites per every part of the complexes, calculated according to definitions 1 and 2, are in excellent agreement relative to the approximate error in the brute-force coarse-graining calculation (see Tables S2 and S3 in the SI).

Zeroth-Order Approximation: Constant Number of Atoms (Residues) per CG Site. The simplest technique to estimate the number of CG sites in different parts of a complex is to take one CG site per constant number of atoms or monomer units. In contrast to the ED-CG-based methods described above or PCA-based methods discussed below, it does not require generating all-atom MD trajectories, minimization of χ^2 or diagonalization of the covariance matrix and can be done very rapidly. In the ribosome, to divide 480 total sites into the component parts, it was necessary to use 22.82 protein or RNA subunits per CG site. For the actin filament Arp2/3 junction, 19.0 residues per CG site was used to yield a total of 585 sites in the complex. The results of this technique are shown in the fourth column in Table 2 (abbreviated version), and in the SI, Tables S2 and S3 (full version).

Analysis of the results demonstrates that this simple rule provides a good estimate for the number of CG sites at a simpler level of analysis (except for the S-proteins in the ribosome). In the two largest and most important parts of the ribosome (23S and 16S rRNAs), the errors in the number of CG sites (compared with definition 2) are as small as 2% and 7%, respectively. However, as is discussed below, this low level of error in these cases takes place by a cancellation of errors. The actual accuracy of the rule is better demonstrated by the other components of the ribosome.

On closer inspection, the number of monomers per CG site (according to definition 2) varies from 9 (mRNA) to 46 (L34 protein); other extreme cases are L1 and L10 proteins (13.2 and 13.7 residues per CG site), and L17 and L20 proteins (40 and 39 residues per CG site). In the actin filament Arp2/3 junction, the extreme cases are RPC5 protein (13.7 residues per CG site) and one of the actin monomers located at the end of a filament (22.1 residue per CG site). The number of residues per CG site differs by up to a factor of 5 in the ribosome (by a factor of 3.5 within the subset of L proteins) and by a factor of 1.6 in the actin filament junction. The optimal numbers of CG sites according to definition 2 differ from those estimated by the rule of apportioning CG sites on the basis of the fractional number of atoms in a particular part by less than 10% only for 19 (out of 61) constituents of the ribosome and for 22 (out of 31) constituents of the actin filaments junction. Ensuring this level of precision requires further study of the reasons for deviations from the zeroth-order estimate and new ways to introduce the required corrections.

To complete the present discussion of the "fixed number of atoms/residues per CG site" rule, one should note that it is usually used in the context of "one CG site per amino acid" or "3 CG sites per nucleotide", as mentioned in the Introduction, rather than at the much coarser level of coarse-graining of interest in this work. Exceptions are found in ref 15, in which CG models of virus capsids were built at a resolution level of ~200 atoms per CG site, and in ref 32, in which ENM models of a protein with 40 residues per CG site were studied.

First-Order Corrections: The Significance of the Anomalous Dimension. The first-order corrections to the empirical rule discussed above are provided by eq 13. First we demonstrate that eq 13 (with omitted second- and higher-order terms) leads to improved results in the two cases under investigation.

Calculations for the case of the ribosome were performed with $\overline{\gamma}$ = 0.366 (arithmetic average of all 61 γ_i) and $\overline{\ln N_{\mathrm{atoms}}}$ = 5.67 (arithmetic average of the minimal and maximal $\ln N_{i,\mathrm{atoms}}$ in the complex); other coefficients were taken from Table 1 (see also the SI, Table S1). The optimal number of CG sites n_i^* for the ith constituent of the ribosome was calculated from eq 13 and rounded to the nearest integer. The coefficient C'' from eq 13 was fitted in such a way that the sum of all n_i^* was constrained to equal 480. The results are provided in Table 2 and the SI, Table S2.

The overall quality can be estimated from Figure 4a, which shows a histogram of the ratio of the number of CG sites obtained using the first-order correction method to the number of CG sites obtained using definition 2. After introducing the first-order correction, the results for the number of CG sites more closely agree with those using definition 2. The approximate results lie in the range between $0.9(n_i^*)_{\mathrm{exact}}$ and $1.1(n_i^*)_{\mathrm{exact}}$ in 70% of the cases (43 out of 61 constituents of the ribosome), whereas in the zeroth-order approximation, this was true in only 31% of the cases (19 out of 61 constituents). Before introducing the correction, the average ratio of the number of CG sites predicted from the zeroth-order rule to the number of CG sites obtained using definition 2 is 1.11, and its standard deviation is 0.33; after introducing the first-order correction, the average ratio shifts to 0.98, and the standard deviation decreases to 0.17.

In the case of the actin filament Arp2/3 junction, the calculations were performed with $\overline{\gamma}$ = 0.337 (arithmetic average of all 31 γ_i), and $\overline{\ln N_{\mathrm{atoms}}}$ = 5.53 (arithmetic average of the minimal and maximal $\ln N_{i,\mathrm{atoms}}$ in the complex) with all other details kept the same as above. The results are given in Table 2 and in the SI, Tables S2 and S3. The overall quality can be estimated from Figure 4b. Introduction of the first-order correction increased the percentage of results that fall between $0.9(n_i^*)_{\mathrm{exact}}$ and $1.1(n_i^*)_{\mathrm{exact}}$ from 71% to 90% (from 22 to 28 out of 31 constituents). The average ratio of the approximate number of CG sites to the exact number of CG sites equals 1.00 in both the zeroth-order and the first-order approximations, whereas the standard deviation decreased from 0.10 to 0.06. In both complexes, introduction of the first-order correction significantly improves the agreements between the approximate and the exact numbers of CG sites in various parts of the complex.

The main sources of the residual errors are the following: (a) deviation of the exact values of χ^2 from those approximated by

dx.doi.org/10.1021/jp2108895 | J. Phys. Chem. B 2012, 116, 8363–8374

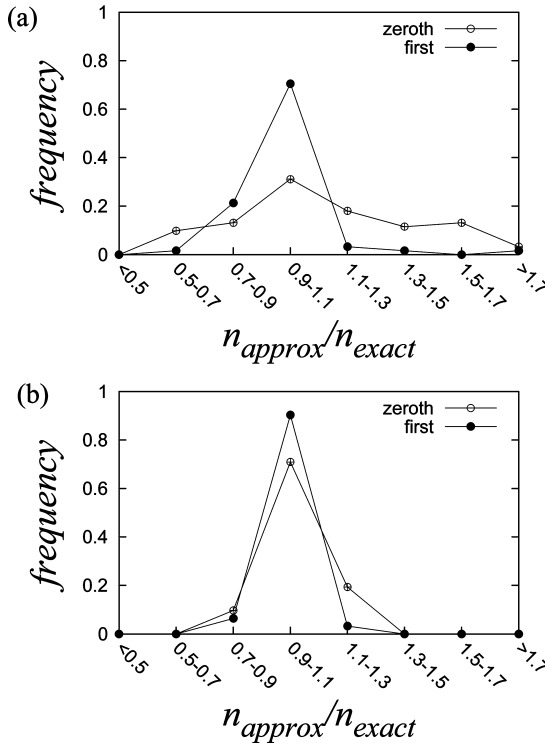


Figure 4. The histograms for the ratio of the approximate number n_{approx} to the exact number n_{exact} of CG sites in different parts of biomolecular complexes confirm the necessity of introducing the first-order corrections [(a), the *E. coli* ribosome, (b) the actin filaments junction]. The zeroth-order approximation corresponds to a constant number of residues (amino acids or nucleotides) per CG site. The first-order approximation is provided by eq 13. The values of n_{exact} are calculated according to definition 2. The perfect prediction of the number of CG sites would correspond to a δ function centered at 1.00.

eq 5 [recall that eq 5 was used in the derivation of eq 13]—the available data on the two complexes suggest (details not shown) that this is the main source of the errors; (b) uncertainty of the parameters γ_i and δ estimated from a finite-size sample (see the corresponding standard deviations in the results of the least-squares treatment in the SI, Table S1); (c) omitting higher-order terms in eq 13; (d) uncertainty in the exact values of the numbers of CG sites themselves (see the comparison of definitions 1 and 2, above).

According to eq 13, the first-order correction splits into two independent contributions: one arising from heterogeneity in the sizes of different parts of a complex and the other one coming from the heterogeneity in their anomalous dimensions. The fundamental reasons for the existence of these corrections remain unclear because of the lack of understanding of the origin of the anomalous dimensions. In the case of the simple scaling defined by eqs 2 and 3, no corrections appear at all, and the zeroth-order approximation would be exact.

Decomposition of the corrections by eq 13 allows one to understand the reasons why the rule "one CG site per fixed number of residues" leads to reasonable results in the two cases under consideration. The reasons differ between the actin filament Arp2/3 junction and the ribosome. In the first case, all proteins in the system have comparable size (with the extreme cases being 151 and 417 residues; note that eq 13 includes only logarithms of these numbers) and close values of the γ_i parameters (distributed in the range 0.25 to 0.61). Such homogeneity leads to a small variation in the number of amino acid residues per CG site (up to a factor of 1.61). By contrast, the ribosome is a very heterogeneous system, with the number of residues varying by 2 orders of magnitude in different constituents, and the γ_i parameters vary between 0 to 0.91, a range nearly 3 times wider than that of the actin filament junction. It is not surprising that the optimal number of residues per CG site varies more in the ribosome (up to a factor of 5) and, therefore, that the first-order corrections are more important in this case. However, for both the two largest and the four smallest parts of the ribosome, 23S rRNA and 16S rRNA and the L7/L12 proteins, respectively, the two corrections mostly cancel each other out. As a result, the zeroth-order approximation leads to reasonable results for these and some other, but not all, constituents. This cancellation of the corrections is possible, according to eq 13, only if the anomalous dimension is inversely proportional to biomolecule size. Coincidentally, this is true for 23S and 16S rRNAs and for the L7/L12 proteins. However, it is doubtful that this relation would hold in general. The available data demonstrate no statistically significant correlation between γ_i and $\ln N_{i,\mathrm{atoms}}$. Therefore, one could expect that at least in some other large heterogeneous biomolecular complexes (such as nucleic acids polymerases, ion channels, nuclear pore complex, etc.), this cancellation of corrections will not take place, and the optimal number of residues per CG site will vary by more than an order of magnitude in different parts of a complex. (In this case, the first-order approximation, eq 13, will not be valid, and either higher-order corrections or the initial nonlinear equation will have to be used.)

An unquestionable advantage of the empirical rule "one CG site per fixed number of atoms/residues" is its simplicity. Its application does not require performing expensive MD simulations and coarse-graining; however, the placement of these sites would require some additional information. The first-order approximation offers a comparably simple, but more precise estimation of the optimal number of CG sites. Indeed, most of the parameters in eq 13 are either known or defined by the researcher. The number of atoms/residues in each part of the complex is well-known. The number of CG sites can be solved for self-consistently. The value of C'' can also be determined iteratively to get the desired number of CG sites in the complex in total. The average value of γ for the whole complex is more stable that the individual values of γ_i; one could approximate it as 0.35 (the values for the ribosome and the actin filaments junction given above are 0.366 and 0.337). Note also that the results should not be very sensitive to the specific values of $\overline{\gamma}$ and $\overline{\ln N_{\mathrm{atoms}}}$, since they play the role of the point around which the Taylor series expansion is performed and affect the result only because the second- and higher-order terms are neglected. As a result, the only nontrivial parameters required by eq 13 are the anomalous dimensions γ_i of all constituents of the complex. It remains unclear how to calculate γ_i without use of the MD trajectories. However, a better understanding of the physical origin of the anomalous dimension might enable simple and reliable estimation of the number of CG sites in different parts of large biomolecular complexes in the future.

PCA-Based Methods. Finally, it is instructive to compare the results of the proposed methods with the techniques based on the principal components analysis (PCA). The "95% rule" coarse-grains the system based on the number of principal components (PCs) that retains 95% of the original data variance ("of original information"); similarly, the 90% rule, the

dx.doi.org/10.1021/jp2108895 | J. Phys. Chem. B 2012, 116, 8363−8374

99% rule, etc. are defined.[24−33] The Kaiser rule coarse-grains the system on the basis of the number of PCs that have eigenvalues larger than the average eigenvalue of all PCs.[57]

Fundamental problems with these approaches are illustrated by the case of ubiquitin, in which 72 out of 76 amino acid residues in this protein form a compact core, while the 4 C-terminal residues form a highly flexible tail. As a result, most of the total variance of the atomic coordinates comes from this tail. The percentage of the retained information thus rapidly increases and reaches a plateau slightly below 100% for a relatively small number of PCs. Application of the 95% rule leads to the conclusion that there are only 35−39 essential degrees of freedom; most of the corresponding PCs describe fluctuations in the tail. This level of coarse-graining would lead to a CG model with 7−8 CG sites. If the tail were removed from the protein, one would expect to find fewer than 7−8 CG sites in the remaining core. Instead, the 95% rule suggests that there should be 14 CG sites. The reason is that the disorder left in the molecule after removing the tail is delocalized throughout the core more or less evenly. In this situation, the accumulated percentage of the retained information does not increase as quickly as the increase of the number of PCs, and therefore, more essential degrees of freedom are required to describe 95% of the fluctuations. The fundamental drawback of all the above-mentioned standard approaches of PCA is that they neglect the absolute scale of thermal fluctuations in the molecules of interest. This leads to instability in the number of CG sites when parts of a molecule are removed or when the molecule associates with or dissociates from a complex. In the case of the ribosome, these methods give the same weight to fluctuations in the huge 23S RNA as those in the small L proteins. These two extremes differ in size and, therefore, in the absolute amount of thermal fluctuation, by 2 orders of magnitude. As Table 2 demonstrates, the standard PCA-based methods can poorly estimate the optimal number of CG sites in large biomolecular complexes.

4. CONCLUSIONS

A great deal of research in molecular biology is focused on obtaining detailed structural information about large biomolecular complexes and a detailed understanding of their function. Atomic resolution experimental structures are generally not available because of the large size of these systems. Computational studies can contribute to the understanding of these systems, but to study biologically relevant length and time scales, CG models become essential. In addition, any reasonable analysis of these systems cannot include all possible degrees of freedom in the system. Information must be compressed to include only the relevant details to facilitate understanding. Both of the aforementioned issues can be addressed by utilizing CG models. As a result, it is reasonable to expect rapid development in the field of CG modeling in the near future, especially for biomolecular complexes of relatively large size.

To ensure the efficiency of CG models, several fundamental methodological issues must be resolved. One of the most important of these issues is the proper choice of the number of CG sites in different parts of the biomolecular complex under investigation. High-resolution description of some parts of the system may prove to be a waste of computational resources if other parts of the system are modeled too coarsely and introduce unphysical artifacts.

The criterion for the optimal number of CG sites, proposed in this article (eqs 8 and 9), addresses these issues. It is local by its nature, relying on the corresponding molecular χ^2 function only. A consistent level of coarse-graining is maintained throughout the molecular complex by the Lagrange multiplier $\lambda = L^2$, a parameter analogous to the chemical potential. Adjusting L^2 allows for adaptive changes to the number of CG sites and redistributing the sites through the system.

The proposed formalism, however, leaves open the question of the choice of the threshold length L (or, equivalently, the *total number* of CG sites in the entire complex). It is possible that "the best in all respects" level of coarse-graining does not exist; many complex systems, including those of a biological nature, have been shown to possess hierarchical fractal-like structure.[54] In such systems, increasing the resolution of the model leads to splitting of single elements into blocks of smaller elements, which may be further resolved at even finer resolutions.[56] To the best of our knowledge, this phenomenon has not been systematically studied in relation to the dynamical CG models of biomolecules, although much of the general theory of networks is known to be applicable to biomolecules.[55]

In this paper, we have demonstrated that the residual thermal fluctuation χ^2 in CG models of molecular dynamics data depends on the number of CG sites in a scale-free manner, similar to the scale-free behavior of known hierarchical systems. If CG models do, in fact, have fractal-like organization, then the optimal total level of coarse-graining will be defined by the specific question being addressed by the CG model and will not have a universal objective character. This is analogous to saying that there is no optimal resolution level of, for example, the Mandelbrot set, and different levels of detail can be used, depending on the purpose of research. However, this required flexibility in the level of resolution does necessitate a criterion for the *distribution* of CG sites within a biomolecular complex. Reducing the problem to the choice of the single parameter L, as proposed in this work, ensures a consistent level of representation for all the parts of a complex.

The method has been successfully applied to two cases: the actin filament Arp2/3 junction and the *E. coli* ribosome. The results obtained using two different approaches to calculation of the number of CG sites (definitions 1 and 2) are in excellent agreement with each other. Separating the CG problem into coarse-graining each of the component parts demonstrates the relatively autonomous dynamics of the constituents of these complexes (and, presumably, the transferability of the characteristics of these constituents between unbound and bound states). The empirical rule of taking one CG site per fixed number of atoms or residues is derived as an approximation to the general criterion for the number of CG sites, and the level of precision for this method for the systems under consideration is on the order of ∼10−30%. The first-order correction to this rule is derived, reducing the typical error in the number of CG sites in the two studied complexes by a factor of 2. In contrast, PCA-based methods used previously in the literature are shown to yield inconsistent solutions.

The universal scaling of the intrasite disorder characteristic function χ^2 as a function of the number of CG sites in the molecule, used for derivation of the criterion, is itself of notable value. The degree of exactness to which the power law scaling is maintained (the median value of $R^2 = 0.999$) is unusual among the relations discovered for quantitative characteristics of biomolecules. This high level of agreement may indicate a

new realm of applicability of the renormalization group theory: that of biological components. The approach developed in this article may be important both for fundamental research and for applications to specific biomolecular complexes. Simpler methods for direct estimation of the values of the anomalous dimension γ would greatly facilitate calculation of the optimal number of CG sites in biomolecular complexes in conjunction with the approach developed in the present paper.

■ ASSOCIATED CONTENT

Supporting Information

Statistical characteristics of the least-squares treatment of eq 5 and detailed data on the number of CG sites in different parts of the ribosome and the actin filament Arp2/3 junction calculated by means of different methods. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 773-702-9092. Fax: 773-795-9106. E-mail: gavoth@uchicago.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation through the Center for Multiscale Theory and Simulation, grant CHE-1136709. The authors are grateful to Drs. Jim Pfaendtner and Karissa Sanbonmatsu for providing the MD trajectories of the actin filament junction and the ribosome, respectively, and to Dr. Anuj Chaudhri for his comments on the manuscript. Calculations of the all-atom MD trajectories were performed using Teragrid computational resources [Kraken (NICS), Athena (NICS) and Ranger (TACC) clusters].

■ REFERENCES

(1) Robinson, C. V.; Sali, A.; Baumeister, W. *Nature (London, U. K.)* **2007**, *450*, 973−982.
(2) Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*; CRC Press: Boca Raton, 2009.
(3) Miao, J. W.; Ishikawa, T.; Shen, Q.; Earnest, T. *Annu. Rev. Phys. Chem.* **2008**, *59*, 387−410.
(4) Lindert, S.; Stewart, P. L.; Meiler, J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 218−225.
(5) Mertens, H. D. T.; Svergun, D. I. *J. Struct. Biol.* **2010**, *172*, 128−141.
(6) Lasker, K.; Sali, A.; Wolfson, H. J. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 3205−3211.
(7) Lasker, K.; Phillips, J. L.; Russel, D.; Velazquez-Muriel, J.; Schneidman-Duhovny, D.; Tjioe, E.; Webb, B.; Schlessinger, A.; Sali, A. *Mol. Cell. Proteomics* **2010**, *9*, 1689−1702.
(8) Gumbart, J.; Schreiner, E.; Trabuco, L. G.; Chan, K. Y.; Schulten, K. In *Molecular Machines in Biology*; Frank, J., Ed.; Cambridge University Press: New York, Cambridge, 2011, pp 142−157.
(9) Sanbonmatsu, K. Y.; Tung, C. S. *J. Struct. Biol.* **2007**, *157*, 470−480.
(10) Vendruscolo, M.; Dobson, C. M. *Curr. Biol.* **2011**, *21*, 68−70.
(11) Zhang, Z.; Sanbonmatsu, K. Y.; Voth, G. A. *J. Am. Chem. Soc.* **2011**, *133*, 16828−16838.
(12) Zuckerman, D. M. *Annu. Rev. Biophys.* **2011**, *40*, 41−62.
(13) Doruker, P.; Jernigan, R. L.; Bahar, I. *J. Comput. Chem.* **2002**, *23*, 119−127.
(14) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750−760.
(15) Arkhipov, A.; Freddolino, P. L.; Schulten, K. *Structure (Cambridge, MA, U.S.)* **2006**, *14*, 1767−1777.
(16) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812−7824.
(17) Knotts, T. A.; Rathore, N.; Schwartz, D. C.; de Pablo, J. J. *J. Chem. Phys.* **2007**, *126*, 084901.
(18) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.
(19) Xia, Z.; Gardner, D. P.; Gutell, R. R.; Ren, P. Y. *J. Phys. Chem. B* **2010**, *114*, 13497−13506.
(20) Hills, R. D.; Lu, L. Y.; Voth, G. A. *PLoS Comput. Biol.* **2010**, *6*, e1000827.
(21) Lyubartsev, A. P. *Eur. Biophys. J.* **2005**, *35*, 53−61.
(22) Ayton, G. S.; Voth, G. A. *J. Phys. Chem. B* **2009**, *113*, 4413−4424.
(23) Chu, J. W.; Voth, G. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13111−13116.
(24) Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer: New York, 2002.
(25) Amadei, A.; Linssen, A. B.; Berendsen, H. J. *Proteins: Struct., Funct., Bioinf.* **1993**, *17*, 412−425.
(26) VanAalten, D. M. F.; DeGroot, B. L.; Findlay, J. B. C.; Berendsen, H. J. C.; Amadei, A. *J. Comput. Chem.* **1997**, *18*, 169−181.
(27) Tozzini, V.; McCammon, J. A. *Chem. Phys. Lett.* **2005**, *413*, 123−128.
(28) Rueda, M.; Chacon, P.; Orozco, M. *Structure (Cambridge, MA, U.S.)* **2007**, *15*, 565−575.
(29) Yang, L.; Song, G.; Carriquiry, A.; Jernigan, R. L. *Structure (Cambridge, MA, U.S.)* **2008**, *16*, 321−330.
(30) Spiwok, V.; Kralova, B.; Tvaroska, I. *J. Mol. Model.* **2008**, *14*, 995−1002.
(31) Potestio, R.; Pontiggia, F.; Micheletti, C. *Biophys. J.* **2009**, *96*, 4993−5002.
(32) Bahar, I. *J. Gen. Physiol.* **2010**, *135*, 563−573.
(33) Martinez, R.; Schwaneberg, U.; Roccatano, D. *Protein Eng., Des. Sel.* **2011**, *24*, 533−544.
(34) Stepanova, M. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2007**, *76*, 051918.
(35) Blinov, N.; Berjanskii, M.; Wishart, D. S.; Stepanova, M. *Biochemistry* **2009**, *48*, 1488−1497.
(36) Zhang, Z. Y.; Lu, L. Y.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 5073−5083.
(37) Zhang, Z. Y.; Voth, G. A. *J. Chem. Theory Comput.* **2010**, *6*, 2990−3002.
(38) Wilson, K. G. *Phys. Rev. D: Part., Fields, Gravitation, Cosmol.* **1970**, *2*, 1478−1493.
(39) Wah, B. W.; Wu, Z. In *Principles and Practice of Constraint Programming*; Jaffar, J., Ed.; Springer: Berlin/Heidelberg, 1999; Vol. *1713*, pp 28−42.
(40) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531−544.
(41) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33−38.
(42) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(43) Martyna, G. J.; Tobias, D. J.; Klein, M. L. *J. Chem. Phys.* **1994**, *101*, 4177−4189.
(44) Feller, S. E.; Zhang, Y. H.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613−4621.
(45) Walsh, M. A.; Schneider, T. R.; Sieker, L. C.; Dauter, Z.; Lamzin, V. S.; Wilson, K. S. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1998**, *54*, 522−546.
(46) Rigden, D. J.; Phillips, S. E.; Michels, P. A.; Fothergill-Gilmore, L. A. *J. Mol. Biol.* **1999**, *291*, 615−635.
(47) Otterbein, L. R.; Graceffa, P.; Dominguez, R. *Science (Washington, DC, U.S.)* **2001**, *293*, 708−711.
(48) Saunders, M. G.; Voth, G. A. *J. Mol. Biol.* **2011**, *413*, 279−291.

(49) Pfaendtner, J.; Volkmann, N.; Hanein, D.; Dalhaimer, P.; Pollard, T. D.; Voth, G. A. *J. Mol. Biol.* **2012**, *416*, 148−161.
(50) Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y. H. *Proteins: Struct., Funct., Bioinf.* **2000**, *41*, 1−7.
(51) Li, G.; Cui, Q. *Biophys. J.* **2002**, *83*, 2457−2474.
(52) Painter, J.; Merritt, E. A. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62*, 439−450.
(53) Albert, R.; Barabasi, A. L. *Rev. Mod. Phys.* **2002**, *74*, 47−97.
(54) Ravasz, E.; Somera, A. L.; Mongru, D. A.; Oltvai, Z. N.; Barabasi, A. L. *Science (Washington, DC, U.S.)* **2002**, *297*, 1551−1555.
(55) Barabasi, A. L.; Oltvai, Z. N. *Nat. Rev. Genet.* **2004**, *5*, 101−113.
(56) Delmotte, A.; Tate, E. W.; Yaliraki, S. N.; Barahona, M. *Phys. Biol.* **2011**, *8*, 055010.
(57) Kaiser, H. *Educ. Psychol. Meas.* **1960**, *20*, 141−151.