JCTC Journal of Chemical Theory and Computation

Coarse-Grained Representations of Large Biomolecular Complexes from Low-Resolution Structural Data

Zhiyong Zhang and Gregory A. Voth*

Department of Chemistry, James Franck and Computation Institutes, University of Chicago, 5735 S. Ellis Avenue, Chicago, Illinois 60637

Received July 3, 2010

Abstract: High-resolution atomistic structures of many large biomolecular complexes have not yet been solved by experiments, such as X-ray crystallography or NMR. Often however low-resolution information is obtained by alternative techniques, such as cryo-electron microscopy or small-angle X-ray scattering. Coarse-grained (CG) models are an appropriate choice to computationally study these complexes given the limited resolution experimental data. One of the important questions therefore is how to define CG representations from these low-resolution density maps. This work provides a space-based essential dynamics coarse-graining (ED-CG) method to define a CG representation from a density map without detailed knowledge of its underlying atomistic structure and primary sequence information. This method is demonstrated on G-actin (both the atomic structure and its density map). It is then applied to the density maps of the *Escherichia coli* 70S ribosome and the microtubule. The results indicate that the method can define highly CG models that still preserve functionally important dynamics of large biomolecular complexes.

Introduction

Large biomolecular complexes are involved in many important biological processes. For example, the ribosome is a very large RNA−protein assembly that plays a central role in protein biosynthesis,[1−3] while microfilaments[4,5] and microtubules[6−11] serve as structural components of the cytoskeleton.[12] It is an important but challenging task for computational biologists to simulate the large-scale functional dynamics of these biomolecular systems over the necessarily long time scales. Atomistic molecular dynamics (MD) simulation remains an important tool for studying functional dynamics of nanometer scale biomolecules (at present) up to microsecond time scales.[13,14] However, the functional dynamics of large biomolecular complexes often occur on much longer time scales than those accessible to atomistic MD. Therefore it is necessary to perform coarse-grained (CG) modeling of these biological systems in order to simulate their long-time behaviors.[15−19]

Generally speaking, a CG model is a reduction of the large number of degrees of freedom in an atomic structure into a significantly smaller set. In order to establish a reasonable mapping between the atomistic and CG models, one needs to define the desired number of CG sites and then determine where to place them. We have addressed this issue previously by developing a systematic and quantitative methodology, which is particularly useful to define an aggressive CG model with a resolution lower than one site per residue for a large biomolecule.[20,21] The method is called essential dynamics coarse graining (ED-CG), in which a CG representation is determined variationally to preserve the functional essential motion of the biomolecule.[22] In the previous ED-CG implementations, the essential dynamics were characterized by principal component analysis (PCA) of an atomistic MD trajectory[20] or an elastic network model (ENM) of a single atomic structure.[21] In both cases, an underlying detailed (high resolution) atomic structure of the biomolecule was needed. Furthermore, the CG sites were assumed to be contiguous along the primary sequence of the biomolecule, therefore the sequence information was also necessary.

However, it is very difficult sometimes to solve the atomic-resolution structure of a large biomolecular complex by current high-resolution experimental techniques, such as

* Corresponding author. E-mail: gavoth@uchicago.edu. Telephone: 773-702-7250.

10.1021/ct100374a
Published on Web 08/23/2010

X-ray crystallography and NMR, due to complications arising from the sheer size of such a complex and other factors, such as difficulty in crystallization (membrane complexes), etc. Instead cryo-electron microscopy (cryo-EM)[23,24] and small-angle X-ray scattering (SAXS)[25−27] are two experimental techniques that can obtain low-resolution models (molecular shapes) of these biomolecular complexes. There are neither atomistic details nor sequence information in these low-resolution structures, which means that the previous ED-CG methodology cannot be applied to these systems directly.

In this work, we introduce a new ED-CG scheme, which is used to define ED-CG models from a cryo-EM or SAXS structure. A technique called vector quantization (VQ) has been widely used to discretize a density map into pseudoatoms and preserve the shape of the low-resolution structure.[28−31] It has been found that an ENM built on the pseudoatom model can describe low-frequency functional dynamics of the biomolecule quite well.[32−35] Therefore ED-CG models are defined in the present work to capture the essential dynamics of the pseudoatom model. In the previous sequence-based ED-CG method, a CG site is a representation of a group of atoms that move together in a highly correlated fashion and are contiguous along the primary amino acid sequence at the same time. As the sequence is not directly related to the pseudoatom model, a new way of defining sites, which does not need sequence information, is developed. The new algorithm defines a space-based ED-CG model, in which a CG site, as before, represents a group of atoms that move together but are close in space instead of contiguous along the sequence. This new space-based ED-CG method can be used to define CG models of biomolecular complexes directly from their cryo-EM or SAXS structures without atomic details and sequence information. It should be noted that this method can of course be applied to atomistically detailed structures as well.

In the subsequent sections, a parameter-free ENM and the ENM-ED-CG method[21] will first be reviewed. The new development of the space-based ED-CG method is then introduced. As a test, the resulting method is applied to the G-actin system using both its atomistic structure and a 10 Å density map, respectively. The space-based ED-CG models from the atomic structure of G-actin are also compared to the sequence-based ED-CG models. The aforementioned VQ method is a technique to define shape-based CG models from atomistic structures[36−39] or density maps,[34] so the space-based ED-CG models are compared to the VQ-CG models. Two other applications of the space-based ED-CG method are to the cryo-EM density maps of the *Escherichia coli* 70S ribosome (11.5 Å) and the microtubule (8 Å), respectively, which will be compared with the VQ-CG method as well. Finally, concluding remarks are provided.

Theory and Methods

Elastic Network Model. In a typical residue-based ENM of a biomolecule,[40−42] the positions of C_α atoms for amino acids and P atoms for nucleotides are used.[43−46] However there is no inherent restriction on the number of atoms or residues they can represent. These CG "atoms" are connected by effective harmonic bonds, therefore, the harmonic potential of the ENM can be written as

$$V = \sum_{i,j>i} \frac{1}{2} k_{ij} \Delta r_{ij}^2 \tag{1}$$

Here, $\Delta r_{ij} = r_{ij} - r_{ij}^0$ is the fluctuation of the bond connecting atoms i and j, where r_{ij}^0 is the equilibrium bond distance. The spring constants, k_{ij}, define the interactions between atoms i and j. There are different rules to determine force constants in ENM.[47−57] Most popularly, a given cutoff distance is used to define the connections, and only the atoms within the cutoff distance are connected by springs, then a uniform force constant is placed for all connected atoms.[47,49] Recently, Hinsen argued that distance-weighted interactions in ENM are physically better motivated, which are superior to the cutoff-based interactions in reproducing crystallographic B-factors.[54] Therefore a "parameter-free" ENM (pfENM) is used here, in which the force constants between atoms are weighted by the inverse square of their distances.[57]

$$k_{ij} = c(r_{ij}^0)^{-2} \tag{2}$$

where c is a constant, which simply scales the overall range of B-factors.

For a system with n atoms, the second derivatives of the overall potential (eq 1) can be organized in a Hessian matrix $\mathbf{H} \in \mathbb{R}^{3n} \times \mathbb{R}^{3n}$. The elements of this matrix are

$$H(i_x, j_y) = \partial^2 V / \partial r_{i_x} \partial r_{j_y} \tag{3}$$

where r_{i_x} and r_{j_y} are the x (= 1, 2, 3) component of the position of the atom i, and the y (= 1, 2, 3) component of the position of the atom j, respectively. $\mathbf{H}$ can be diagonalized to yield a matrix of eigenvectors and corresponding eigenvalues,

$$H(i_x, j_y) = \sum_{q=1}^{3n} \Psi_q^{i_x} \lambda_q \Psi_q^{j_y} \tag{4}$$

Here $\Psi_q^{i_x}$ and $\Psi_q^{j_y}$ are the two components corresponding to the x coordinate of the atom i and the y coordinate of the atom j, respectively, in the eigenvector $\mathbf{\Psi}_q \in \mathbb{R}^{3n}$ (normal mode), which is the q^{th}column of the matrix $\mathbf{\Psi} \in \mathbb{R}^{3n} \times \mathbb{R}^{3n}$. There are a total of $3n$ eigenvalues λ_q, the first six of which are zero because rigid-body translations and rotations leave the Hamiltonian invariant. Each nonzero eigenvalue and corresponding eigenvector represents the frequency and the Cartesian components of this normal mode, respectively. It should be noted that different choices of the constant c in eq 2 will only change the eigenvalues but not the eigenvectors. Many studies have indicated that the first few low-frequency normal modes describe functionally important motions in biomolecules.[58,59]

ENM-ED-CG Method. The details of the ENM-ED-CG methodology are described in ref 21. From pfENM, an essential subspace is obtained, which consists of the first n_{ED} of the low-frequency normal modes with nonzero eigenvalues. For an N-site CG model of a biomolecule, $n_{\mathrm{ED}} = 3N - 6$ since it has $3N - 6$ internal degrees of freedom. In the biomolecule, those atoms that move together in a highly correlated fashion (called a dynamic domain) are mapped into one CG site by minimizing the following residual

$$\chi^2 = \frac{1}{3N}\sum_{I=1}^{N}\sum_{i\in I}\sum_{j\geq i\in I}\langle(\Delta\mathbf{r}_i^{\mathrm{ED}})^2 - 2\Delta\mathbf{r}_i^{\mathrm{ED}}\cdot\Delta\mathbf{r}_j^{\mathrm{ED}} + (\Delta\mathbf{r}_j^{\mathrm{ED}})^2\rangle \tag{5}$$

where $\Delta\mathbf{r}_i^{\mathrm{ED}}$ is the fluctuation of atom i in the essential subspace. If another atom j moves together with the atom i, their fluctuation difference, $|\Delta r_i^{\mathrm{ED}} - \Delta r_j^{\mathrm{ED}}|^2$, would be very small. Thus the residual (eq 5) can be minimized by grouping them into the same CG site I. A CG model defined by this algorithm can therefore preserve dynamic domains in the atomistic model and approximate the functional essential dynamics of the biomolecule.

According to the classical theory of networks,[60] the mean-square fluctuation of atom i in the essential subspace is

$$\langle(\Delta\mathbf{r}_i^{\mathrm{ED}})^2\rangle = k_{\mathrm{B}}T\,\mathrm{tr}[(\mathbf{h}^{\mathrm{ED}})_{ii}^{-1}] \tag{6}$$

where k_{B} is the Boltzmann constant, and T is the absolute temperature. The term $(\mathbf{h}^{\mathrm{ED}})_{ii}^{-1} \in \mathbb{R}^3 \times \mathbb{R}^3$ is the i^{th} diagonal superelement (a 3 × 3 matrix) in the inverse matrix of $\mathbf{H}^{\mathrm{ED}} \in \mathbb{R}^{3n} \times \mathbb{R}^{3n}$ ($\mathbf{H}$ in the essential subspace), and tr[ ] represents the trace. That is to say,

$$\mathrm{tr}[(\mathbf{h}^{\mathrm{ED}})_{ii}^{-1}] = \sum_{x=1}^{3}\sum_{q=7}^{n_{\mathrm{ED}}+6}\Psi_q^{i_x}\lambda_q^{-1}\Psi_q^{i_x}$$

It follows that eq 5 may be recast in the following form:

$$\chi^2 = \frac{k_{\mathrm{B}}T}{3N}\sum_{I=1}^{N}\sum_{i\in I}\sum_{j\geq i\in I}(\mathrm{tr}[(\mathbf{h}^{\mathrm{ED}})_{ii}^{-1}] - 2\mathrm{tr}[(\mathbf{h}^{\mathrm{ED}})_{ij}^{-1}] + \mathrm{tr}[(\mathbf{h}^{\mathrm{ED}})_{jj}^{-1}]) \tag{7}$$

The original ENM-ED-CG method systematically defined CG sites in a protein along its primary amino acid sequence. However, the atomistic details and sequence information underlying a low-resolution structure may not be known, which precludes the coarse graining of many proteins and the protein complexes using the ENM-ED-CG scheme. A different approach is therefore needed and described in the next two sections.

Discretization of a Density Map. Low-resolution structural information, such as density maps measured by cryo-EM or SAXS, can be discretized using a technique called vector quantization (VQ).[28−31] The VQ approach allows one to represent a density map by a finite number of n pseudoatoms, which approximate the density (mass distribution) according to a statistical optimization criterion. There are several algorithms and utilities to solve the VQ problem. In the program packages Situs[61,62] and Sculptor,[63] a VQ tool is provided for generating a pseudoatom model given an input volumetric structure, by using a so-called "neural gas" network algorithm.[30,31] This approach has also been implemented in the VMD program[64] recently. In the present work, Sculptor is used to generate a pseudoatom model from a density map.

A Space-Based ED-CG Scheme. After a density map is discretized into n pseudoatoms, a pfENM is built. Then one can define a CG model with N sites from the pseudoatom model. The spirit of ED-CG is to group atoms that move in a highly correlated fashion into a CG site. When the atoms are close in three-dimensional (3D) space, they may have a high tendency to move together and define the best CG unit in the ED-CG scheme. Therefore a space-based ED-CG algorithm is proposed for the pseudoatom model. The details of the algorithm are as follows:

(1) N cluster seeds are generated randomly. In practice, the position of one seed S (R_{S_x}, $x = 1-3$) is the position of one randomly selected pseudoatom, plus a small random offset (from −1.0 to 1.0).

(2) The pseudoatoms are clustered into N groups (domains) according to the N seeds, such that every atom in one domain is closer to the corresponding seed than any other $N - 1$ seeds. Once the N domains are determined, the position of central-of-mass (COM) of each domain, denoted as I, is computed (R_{I_x}, $x = 1-3$).

(3) The average difference between R_{I_x} and R_{S_x} is calculated as

$$R_{\mathrm{diff}} = \frac{1}{3N}\sum_{I=S=1}^{N}\sum_{x=1}^{3}|R_{I_x} - R_{S_x}| \tag{8}$$

Sometimes R_{diff} is rather large, which may indicate that in one or more domains, the atoms are not close in space. In the worst case, one domain may contain separate pieces that are far apart. To avoid this, R_{diff} needs to be decreased in the following way. The positions of the cluster seeds are updated/replaced by the COM of the domains, that is $R_{S_x} = R_{I_x}$. Repeat step 2 with the new cluster seeds to get the updated positions of the COM of the domains and a new R_{diff} (eq 8). Repeat this until R_{diff} is below a certain value $R_{\mathrm{diff}}^{\max}$.

(4) The CG sites are taken as the COM of the domains, and the residual of this N-site model is calculated by eq 7.

(5) Randomly pick a cluster seed and update its position. Repeat the steps 2−4 and obtain an updated CG model with a new residual. This new CG model is accepted or rejected based on the Metropolis criterion,[65] as introduced in refs 20, 21. Step 5 is iterated for a number of steps (n_{SA}) to minimize the residual (eq 7) using a simulated annealing algorithm.[66]

(6) The above steps 1−5 are performed, beginning with different initial sets of cluster seeds. Finally, the CG model with the lowest residual is taken.

It should be noted that step 3 is used to avoid domains with distant groups of atoms, which cannot be achieved by minimizing the residual (eq 7) only. Even if two groups of atoms are far apart in space, their fluctuation difference could still be small, and thus the residual may be minimized if they are grouped into the same CG site. Step 3 serves as a constraint to find domains without distant groups of atoms while minimizing the residual (eq 7).

One might imagine that the value of $R_{\mathrm{diff}}^{\max}$, the maximum allowed difference between the COM of the domains and the cluster seeds, is a critical parameter which determines the size of the space for ED-CG searching. When $R_{\text{diff}}^{\max}$ is too small, the searchable space for ED-CG is highly limited, and the final CG model will mostly depend on eq 8 rather than eq 7. Actually, eq 8 is a so-called Linde−Buzo−Gray (LBG) algorithm to solve the VQ problem when $R_{\text{diff}}^{\max}$ is very small.[28] A larger value of $R_{\text{diff}}^{\max}$ can certainly broaden the searchable space for ED-CG but may render this constraint meaningless. We have tested different values and found a $R_{\text{diff}}^{\max}$ around 0.5 Å to be a good compromise. The values within this range could avoid distant groups of atoms in the same domain, and in the mean time, a reasonable space for ED-CG searching was allowed. In this work, $R_{\text{diff}}^{\max}$ is therefore chosen as 0.5 Å.

CG Models from Different Methods. For all the systems studied in the next section, the corresponding space-based ED-CG models will be mainly presented by using the new algorithm described above, which can be applied to both an atomistic structure and a pseudoatom model from a density map. CG models from other methods are also introduced for comparison. The VQ method can define space-based CG models, from the atomistic structure, as well as the pseudoatom model or the density map directly. The previous published ENM-ED-CG scheme[21] can be used to define sequence-based ED-CG models only if the atomistic structure of the system is available. Therefore, three kinds of CG models (space- and sequence-based ED-CG and VQ-CG models, respectively) are all discussed when defining CG models from an atomistic structure of the system. For a density map, a pseudoatom model is constructed by the VQ method. In that case, both the space-based ED-CG and VQ-CG models are defined based on the pseudoatom model.

Results and Discussion

CG Models of G-Actin from Atomistic Structure. The protein G-actin is a globular protein with 375 residues,[67] which constitutes the subunit of the actin filament.[68−70] Low-resolution CG models of the G-actin have been widely used to explore the elastic properties of the actin filament.[71−73] The atomistic structure of the G-actin is available and was used in the previous two ED-CG papers[20,21] to define sequence-based CG models. As a comparison, the space-based ED-CG method was also applied to the atomic model of the G-actin to define space-based ED-CG models. Only the 375 C_α atoms in the atomic structure were used to build a pfENM.

Four-Site CG Models. By usual inspection, G-actin (Figure 1) can be divided into four spatial domains.[68] The residue numbers of these domains are D1 (1−32, 70−144, and 338−375; Figure 1a, blue), D2 (33−69; Figure 1a, red), D3 (145−180 and 270−337; Figure 1a, orange), and D4 (181−269; Figure 1a, green). By taking the COM of each domain as a CG site, Chu and Voth[71,72] have developed an intuitive four-site CG model of G-actin, which is in fact a space-based CG model (Figure 1a). The first mode from pfENM indicates a propeller-like motion of the domains, which is in agreement with the atomistic normal-mode analysis.[74] In this mode (Figure 1a), the domain D1 has a tendency of coming into the plane of the page as the domain D2 moves out of the plane and vice versa. The domains D3 and D4 perform a similar but antiparallel motion to the domains D1 and D2. That is to say, the domain D4 comes into the plane, while the domain D2 moves out of the plane and vice versa. The intuitive four-site model of G-actin thus naturally allows one to study this propeller motion, which may be related to the opening/closing of the ATP binding cleft.[67]

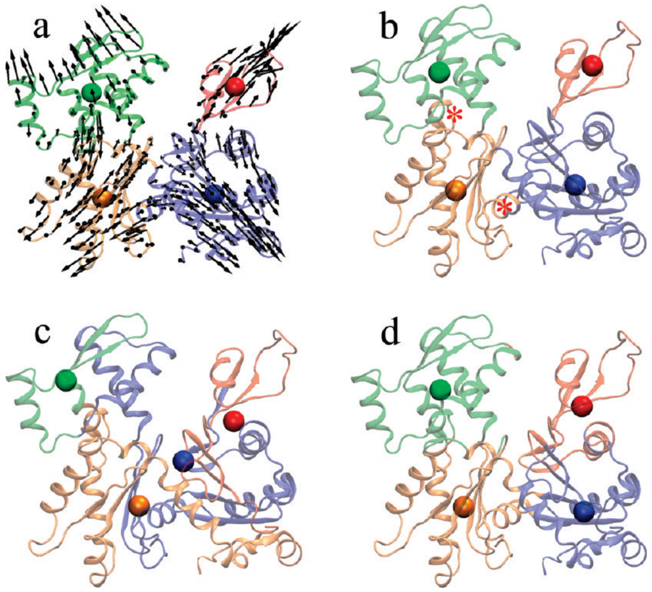

Figure 1. Four-site models of the G-actin. (a) The intuitive four-site model: D1 (1−32, 70−144, 338−375) blue; D2 (33−69) red; D3 (145−180, 270−337) orange; and D4 (181−269) green. First mode from pfENM is shown by arrows. (b) Space-based ENM-ED-CG four-site model: D1 (1−33, 70−140, 337−375) blue; D2 (34−69) red; D3 (141−181, 261−336) orange; and D4 (182−260) green. Its differences to the intuitive four-site model are marked by asterisks. (c) Sequence-based ENM-ED-CG four-site model: (1−66) red; (67−219) blue; (220−256) green; and (257−375) orange. (d) VQ four-site model. Figures 1 and 3−9 were created using VMD.[64]

Here, a four-site CG model was defined with the space-based ED-CG method, using the first six normal modes (n_{ED} = 6). Cluster seeds from 4000 random initial sets were used, and n_{SA} = 5000 steps were calculated for each set, to minimize the residual (eq 7). The lowest residual after these SA iterations was 5391 but only two out of the total 4000 cluster-seed sets reached the lowest residual value. The result indicates that the convergence of the space-based ED-CG method is not as good as the previous sequence-based method.[20,21] The space-based search is much more complicated than the sequence-based one, and it would be difficult to sufficiently sample all the possibilities in limited steps. Nevertheless, the space-based ED-CG four-site model with the lowest residual after the SA iterations looks similar to the intuitive four-site model, but the latter has a higher residual of 5458.

There are seven sequence-contiguous subdomains in the intuitive four-site model (Figure 2a), and the space-based ED-CG model has seven similar subdomains as well,

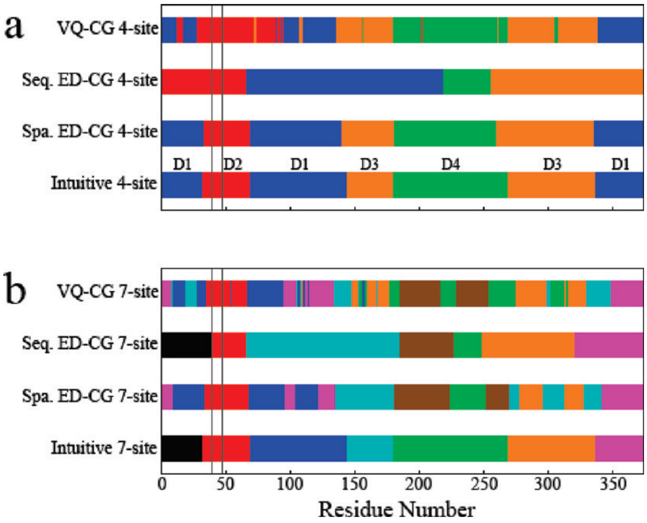

Figure 2. Allocation of domains in different CG models of the G-actin. (a) Four- and (b) seven-site models. The *x* axis is the residue number of the G-actin, and *y* axis represents different CG models. Four subdomains (D1−D4) are indicated on top of the intuitive four-site model. Domains are colored according to Figures 1 and 3 (four- and seven-site models, respectively). DB-loop region (residue 40−48) is marked with vertical gray lines.

although they are not completely contiguous in sequence. For example, there is one subdomain that is contiguous from residues 70−144 in the intuitive model, and the corresponding subdomain in the space-based ED-CG model spans the residues 70−140. Most of the residues from 70−140 belong to the same subdomain but a few interspersed residues belong to other subdomains. Interestingly, if we smoothed the subdomain by changing these a few residues to the same subdomain that most residues (from 70−140) belong to, the residual was further minimized.

It was relatively straightforward to smooth the space-based ED-CG model obtained by the SA iterations along the primary sequence, in order to make the contiguous subdomains while minimizing the residual further. Thus a new space-based ED-CG four-site model was obtained with a little lower residual of 5335 (Figure 1b) than the one without smoothing along the sequence. The residue numbers of the domains in this model are D1 (1−33, 70−140, and 337−375; Figure 1b, blue), D2 (34−69; Figure 1b, red), D3 (141−181, and 261−336; Figure 1b, orange), and D4 (182−260; Figure 1b, green). This space-based ED-CG four-site model is very close to the intuitive four-site model (Figure 1a) with just a few differences (marked by asterisks in Figure 1b). For example, the domain D4 is from residues 181−269 in the intuitive four-site model, but it is from residues 182−260 in the space-based ED-CG four-site model. According to the motions of residues 261−269 in the first ENM mode, it is better to place them in the domain D3 since they move correlated with the other residues in this domain (Figure 1b). That is to say, the space-based ED-CG four-site model is somewhat better at dividing the dynamics domains than the intuitive one. These results indicate that the space-based ED-CG method can define a robust CG model that is consistent with intuition but in a more systematic and quantitative way.

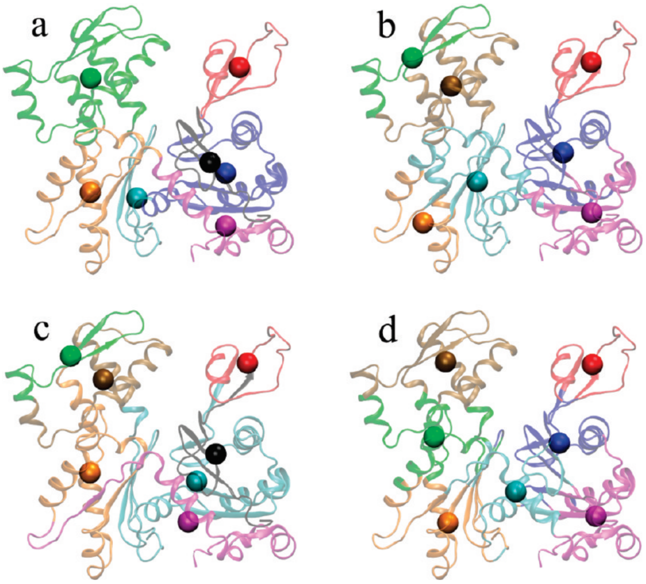

Figure 3. Seven-site models of the G-actin. (a) The intuitive seven-site model that is a sequence-based model: (1−32) black; (33−69) red; (70−144) blue; (145−180) cyan; (181−269) green; (270−337) orange; and (338−375) magenta. (b) The space-based ENM-ED-CG seven-site model. (c) The sequence-based ENM-ED-CG seven-site model: (1−39) black; (40−66) red; (67−185) cyan; (186−227) ocher; (228−249) green; (250−321) orange; and (322−375) magenta. (d) VQ seven-site model.

A sequence-based four-site model (Figure 1c) was also defined by the previous ENM-ED-CG method.[21] A total of 86 out of the total 200 initial boundary atom sets reached the minimal residual of 5672, which indicated a better convergence in the sequence-based ED-CG search than the space-based one. The model contains four sequence-contiguous domains: (1−66; Figure 1c, red), (67−219; Figure 1c, blue), (220−256; Figure 1c, green), and (257−375; Figure 1c, orange). These domains, which obviously deviate from intuition, are very different from those in the space-based ED-CG four-site model (Figure 1b) because of the additional constraint of having primary sequence-contiguous domains. Furthermore, the domain motions in the low-frequency normal mode (Figure 1a) cannot be described properly by the sequence-based ED-CG four-site model, which explains why it has a significantly higher residual than the space-based ED-CG four-site model. That is, the sequence-based model is not as good at preserving the essential low-frequency dynamics of G-actin as the space-based ED-CG model for this highly coarse-grained (four sites) model.

A four-site model defined by the VQ-CG method is shown in Figure 1d. This model looks like the space-based ED-CG and intuitive models (Figure 1b and 1a, respectively) but with a significantly higher residual of 5865. Instead of the seven sequence-contiguous subdomains in those two models, the VQ-CG four-site model becomes rather mixed along the primary sequence (Figure 2). This result is actually quite important because it shows that having any motion of the protein included in the CG model development (even an ENM motion based on VQ pseudoatoms) causes the resulting CG model to have a much greater (and physical) "molecule-like" character.

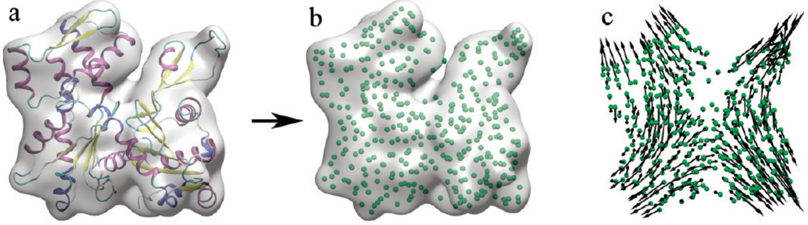

Figure 4. (a) From the atomistic structure of the G-actin, a density map of 10 Å resolution was generated by Situs.[61,62] Then a (b) pseudoatom model with 375 pseudoatoms was built by the VQ algorithm in Sculptor.[63] (c) First mode from pfENM of the pseudoatom model.

Seven-Site CG Models. The intuitive four-site model of G-actin consists of seven sequence-contiguous subdomains (Figures 1a and 2a) from which an intuitive seven-site model was defined (Figure 2b and Figure 3a). It is a sequence-based model with a high residual of 4334, where the first 15 normal modes were used ($n_{\mathrm{ED}} = 15$).

A space-based ED-CG seven-site model was defined by using 4000 random initial sets of cluster seeds and 5000 SA steps for each set. The model was then smoothed along the primary sequence to further minimize the residual, and the final seven-site model is shown in Figure 3b. This model has a much lower residual (2410) than the intuitive seven-site model (Figure 3a), which means it does a much better job of preserving the essential protein dynamics. The space-based ED-CG seven-site model obviously includes more detail than the space-based ED-CG four-site model (Figure 1b). The domain D2, which includes the most flexible DB-loop,[70] almost remains the same between the two models (Figures 1b and 3b, red). The other three sites (blue, green, and orange) in the space-based ED-CG four-site model (Figure 1b) are all divided into two CG sites, respectively, in the space-based ED-CG seven-site model (Figure 3b). Besides the DB-loop, residues 220−252 in the D4 domain are also highly mobile according to the first ENM mode (Figure 1a). Therefore it is defined as a separate CG site in the space-based ED-CG seven-site model (Figure 3b), in order to minimize the residual. However in the intuitive seven-site model (Figure 3a), the domain D4 remains intact as that in the intuitive four-site model (Figure 1a). Instead the domain D1, which is less flexible, is divided into three domains (black, blue, and magenta in Figure 3a). That explains why the residual of the intuitive seven-site model is so high because it does not capture these highly mobile domains properly.

The sequence-based ED-CG seven-site model (Figure 3c) has a residual of 2701, which is higher than the space-based ED-CG seven-site model (Figure 3b) but still much lower than the intuitive seven-site model (Figure 3a). Three CG sites (red, green, and ocher sites in Figure 3c) are similar to those in the space-based ED-CG seven-site model (Figure 3b). Therefore the residual of the sequence-based ED-CG seven-site model is reasonably low, since it can well describe these mobile regions in the domains D2 and D4.

Interestingly the residual of the VQ-CG seven-site model (Figure 3d) is 3185, still significantly lower than the intuitive seven-site model (Figure 3a). In fact, the sites in the VQ-CG seven-site model (Figure 3d) look like those in the space-based ED-CG seven-site model (Figure 3b), except for the relative locations between the green and ocher sites. However the VQ-CG seven-site model significantly scrambles the primary protein sequence (Figure 2b), making it wonder how the motions of such a CG model would correspond to underlying atomistic motions.

CG Models of G-Actin from Density Map. By lowering the resolution of the atomistic structure of G-actin, one can create a density map at a specified resolution (Figure 4a). This was done with the Situs package[61,62] by using its program "pdb2vol". A volumetric density map of G-actin at 10 Å resolution was thus generated (Figure 4a). Then 375 pseudoatoms (Figure 4b) were defined to represent the density map by using the VQ method in Sculptor,[63] which are not the same as the 375 C_α atoms from the high-resolution atomic structure. However, the first mode from pfENM of the pseudoatom model exhibits a similar propeller-like motion (Figure 4c) as in the first mode from the atomistic structure (Figure 1a). The space-based ED-CG method was then applied to group the pseudoatoms into CG sites. As in the case of the atomic structure, the same number (4000) of random initial sets of cluster seeds and SA steps (5000) were used. For a single initial cluster-seed set, the calculation was finished in a few seconds on a 2.4 GHz desktop personal computer.

Four-Site CG Models. The space-based ED-CG four-site model from the pseudoatom model has the lowest residual of 5127 (it should be noted here that residuals between the pseudoatom and atomistic models are not comparable). This CG model (Figure 5a) is very similar to that from the atomic model (Figure 1b). A VQ-CG four-site model (Figure 5b) from the same pseudoatom model is also rather close to the space-based ED-CG four-site model (Figure 5a), which is consistent with the results from the atomistic structure of G-actin (Figure 1b and d). However the VQ-CG four-site model has a higher residual of 5352 than the space-based ED-CG four-site model.

Seven-Site CG Models. The space-based ED-CG seven-site model, with the lowest residual of 2213, is shown in Figure 5c. The model looks similar to the space-based ED-CG seven-site model from the atomistic structure (Figure 3b), except that the relative positions of the green and ocher sites and the cyan and orange sites between them are somewhat different. The red CG site that includes the critically important DB-loop[70] appears the same in both the ED-CG four- and seven-site models from the pseudoatom model. The other three CG sites (blue, green, and orange) in the space-based ED-CG four-site model (Figure 5b) are

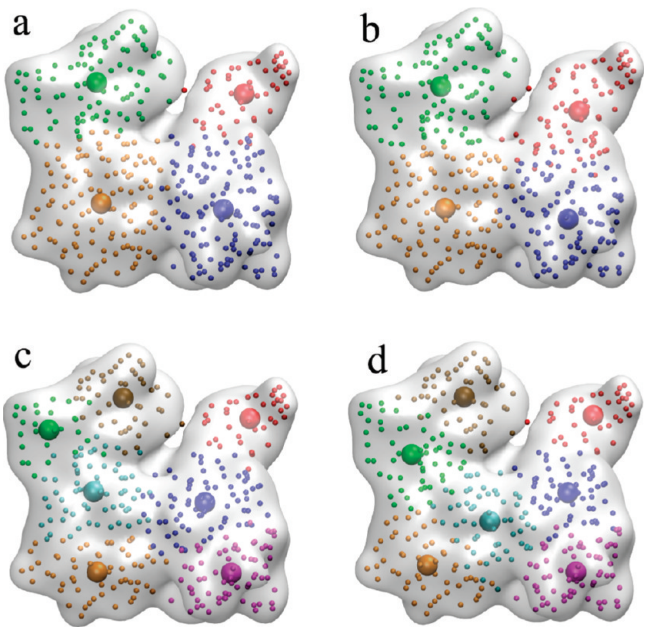

Figure 5. Space-based CG models of the G-actin from its pseudoatom model (Figure 4c). (a) Space-based ED-CG four-site model. (b) VQ-CG four-site model. (c) Space-based ED-CG seven-site model. (d) VQ-CG seven-site model.

divided into two sites each in the space-based ED-CG seven-site model (Figure 5c). These results are consistent with those derived from the underlying atomic structure of G-actin (Figures 1b and 3b), which suggests that the space-based ED-CG method also does a good job of preserving dynamic domains derived from the pseudoatom model. The VQ-CG seven-site model from the pseudoatom model (Figure 5d) may not be adequate in this regard because some of its CG sites, such as the green and cyan sites, cross different domains in the four-site model (Figure 5a). That may explain why the VQ-CG seven-site model has a higher residual (2623) than the space-based ED-CG seven-site model.

The above results are quite encouraging, demonstrating that it is a viable strategy to discretize a low-resolution density map into pseudoatoms by VQ and then to apply the space-based ED-CG approach directly to the pseudoatom model to define CG sites that preserve the essential dynamics of the system. The space-based ED-CG models built from these pseudoatoms are quite similar to those built from an underlying high-resolution atomistic structure directly. In the following sections, the space-based ED-CG method will therefore be applied directly to density maps determined by cryo-EM.

CG Models of *E. coli* 70s Ribosome from Density Map. The *E. coli* 70S ribosome is a large RNA–protein complex, which plays a central role in protein biosynthesis.[1−3] The complete ribosome consists of a small and large subunit. The 30S small subunit contains a 16S rRNA and about 20 S proteins. The 50S large subunit contains a 23S and 5S rRNA and over 30 L proteins. To date, bacterial ribosome structures are available from both low-resolution cry-EM density maps[75,76] and high-resolution atomic structures.[77−82] It is computationally very expensive to study the functional dynamics of the ribosome by atomistic MD simulations since it is a very large macromolecular assembly.[83−86] Coarse-grained ENMs have predicted certain global motions in the ribosome, such as the ratchet-like reorganization between the small and large subunits.[43−46] In this work, a cryo-EM density map of the ribosome (Figure 6a) at a 11.5 Å resolution[76] is used to define space-based ED-CG models. A total of 2000 pseudoatoms were generated by the VQ method (Figure 6b) from the ribosome density map, and a pfENM was built from the pseudoatom model. The first normal mode (Figure 6c) does describe a ratchet-like motion between the small and large subunits, which is in agreement with the results from experiments[75,87] and other atom-based ENMs.[43−46] We then define space-based ED-CG models directly from this pseudoatom model.

40-Site CG Ribosome Models. To define a space-based ED-CG 40-site model, 2000 random initial sets of cluster seeds were used, and n_{SA} = 80 000 steps were performed for each set to minimize the residual (eq 7). For one initial set of cluster seeds, the SA minimization was done in about 4−5 min on a 2.4 GHz CPU. To speed the calculations of the 2000 initial sets, they were distributed onto multiple computer nodes at the same time, since they were completely independent. The model with the lowest residual (535) is shown in Figure 7a. For comparison, a VQ-CG 40-site model was also generated (Figure 7b) that has a much higher residual of 933. Although the resolution of the ribosome density map is only 11.5 Å, the small and large subunits and some other structural details (like the head, spur, L7/L12 stalk base, and protein L1) are visible (Figure 6a). In the space-based ED-CG 40-site model, there are 17 (Figure 7a, orange) and 21 sites (Figure 7a, blue) in the small and large subunits, respectively, and 2 sites are located in the bridges between the two subunits (Figure 7a, magenta). In the VQ-CG 40-site model, there are 13 (Figure 7b, orange) and 25 sites (Figure 7b, blue) in the small and large subunits, respectively, and 2 sites in the bridge area (Figure 7b, magenta).

Since the molecular weight of the small subunit is approximately half of the weight of the large subunit,[1] the CG-site distribution in the VQ-CG 40-site model reflects the mass distribution in the ribosome. However, the space-based ED-CG 40-site model contains more CG sites in the small subunit than in the VQ-CG 40-site model. It is well-known that the 30S small subunit fluctuates more than the 50S subunit in the ribosome dynamics.[75,77,78] Therefore, more CG sites are located in the small subunit relative to its size, in order to better represent its dynamics. For example, the spur region in the small subunit has large fluctuations, according to the first normal mode (Figure 6c). The space-based ED-CG 40-site model has 2 sites in this region to preserve its dynamics (Figure 7a), but the VQ-CG 40-site model only defines 1 site in that region (Figure 7b). The site that represents the spur region in the VQ-CG 40-site model has a too large fluctuation (a so-called "tip effect"),[50] which means a single site may be not enough to well describe the dynamics of the spur.

The same behavior happens in the 50S large subunit. Although the space-based ED-CG 40-site model has less CG sites in the large subunit than the VQ-CG 40-site model, more CG sites are located in the functionally important regions, such as the L1 and L7/L12 stalks (Figure 7a). The

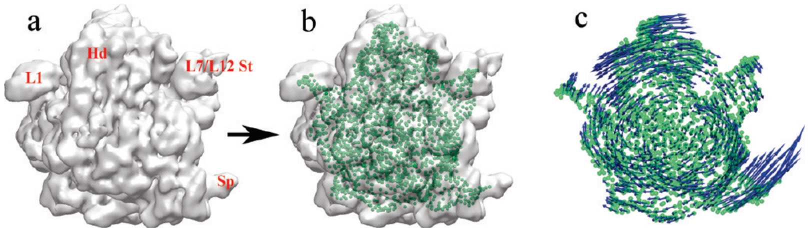

Figure 6. From (a) the 11.5 Å density map of *E. coli* 70S ribosome,[76] a (b) pseudoatom model with 2000 pseudoatoms was built by the VQ algorithm in Sculptor.[63] (c) First normal mode from pfENM of the pseudoatom model, which describes the ratchet-like motion between the small and large subunits.

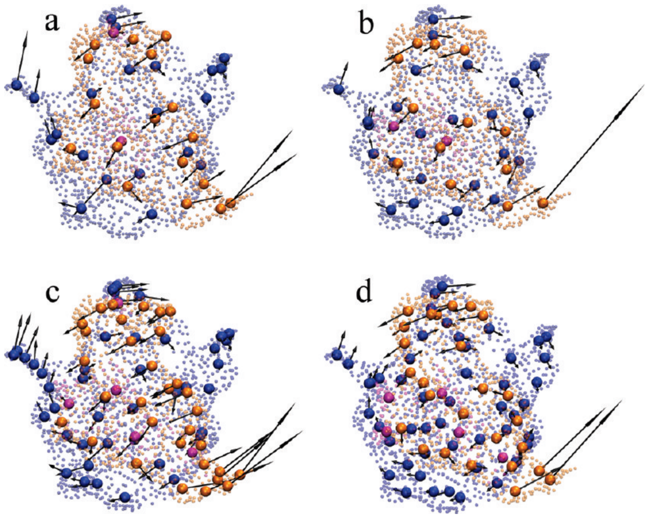

Figure 7. Space-based CG models of the ribosome from its pseudoatom model (Figure 6b). (a) Space-based ED-CG 40-site model. (b) VQ-CG 40-site model. (c) Space-based ED-CG 80-site model. (d) VQ-CG 80-site model. CG sites that belong to the small subunit are colored by orange, CG sites that belong to the large subunit are colored by blue, and those CG sites that make the bridges between the subunits are colored by magenta. In each CG model, the first normal mode from pfENM is shown by arrows.

body of the 50S subunit is massive, so a large number of CG sites are needed in the VQ-CG 40-site model in order to reflect its mass distribution (Figure 7b). However the 50S body is fairly rigid, and fewer CG sites are sufficient to preserve its dynamics (Figure 7a). According to experimental data,[76] there are some intersubunit bridges that are located in both the head and body regions of the ribosome. For the two bridge CG sites in the space-based ED-CG 40-site model (magenta in Figure 7a), one site is located in the head, and the other one is in the body of the ribosome. However in the VQ-CG 40-site model, both bridge CG sites (magenta in Figure 7b) are in the body because it is much more massive than the head. The locations of the bridge sites in the space-based ED-CG 40-site model better describe the association between the two subunits. In summary, the space-based ED-CG model appears to be superior in identifying and preserving the functional dynamics between the ribosome subunits.

80-Site CG Ribosome Models. To define a space-based ED-CG 80-site model from the pseudoatom model of the ribosome, again 2000 initial sets of cluster seeds were used and $n_{\mathrm{SA}} = 160\,000$ steps were performed for each set to minimize the residual (eq 7). It took about 7−8 min to minimize a single cluster-seed set. The lowest residual of the space-based ED-CG 80-site model (Figure 7c) is 221, while the VQ-CG 80-site model (Figure 7d) has a higher residual of 376. Although 80 CG sites are still very coarse compared to the size of the ribosome, the space-based ED-CG 80-site model adds more detail than the space-based ED-CG 40-site model. As with the 40-site models, the regions with functional importance in the ribosome are better represented in the space-based ED-CG 80-site model than in the VQ-CG 80-site model. The small subunit contains 32 CG sites in the space-based ED-CG 80-site model (Figure 7c, orange), but this number in the VQ-CG 80-site model is only 26 (Figure 7d, orange). Five CG sites are defined to represent the dynamics of the spur region in the space-based ED-CG 80-site model (Figure 7c, orange), whereas the VQ-CG 80-site model has only 2 sites in this region (Figure 7d, orange). The space-based ED-CG 80-site model has a smaller number of 43 CG sites in the large subunit (Figure 7c, blue) than that of the VQ-CG 80-site model (49 sites, Figure 7d, blue). Nevertheless the former has more CG sites located in the L1 region and the L7/L12 stalk than the latter. There are five CG sites defined for the bridges in the space-based ED-CG 80-site model, one is in the head and the other four are in the body (magenta in Figure 7c). Importantly, the locations of these intersubunit bridges are in agreement with the experimental data.[76] In the VQ-CG 80-site model, all the five bridge sites are in the body (magenta in Figure 7d).

CG Models of the Microtubule from Density Map. Microtubules (MTs) are long and stiff hollow cylindrical tubes in eukaryotic cells, which play fundamental roles in many cellular processes, such as mitosis, cytokinesis, and vesicular transport.[6−11] The structural subunit of a MT is the $\alpha\beta$-tubulin heterodimer.[88−93] The MT assembly involves two steps: the tubulin dimers bind head to tail to form protofilaments (pfs), and then pfs assemble side by side to complete the microtubule.[94] The atomic structure of the tubulin dimer has been determined by electron crystallography[89] and refined to 3.5 Å resolution.[92] However, MTs have not yet been found to be suitable for X-ray crystallography, since they are highly polymorphic. In this case, cryo-EM was well adapted for obtaining a 3D reconstruction of the MT.[95−98]

In this work, space-based ED-CG models are defined from an EM density map of the MT at an 8 Å resolution.[98] In this 3D map (Figure 8a), there are 13 parallel pfs, and lateral interactions between them complete the MT wall. In each pf, there are approximately four tubulin monomers (three

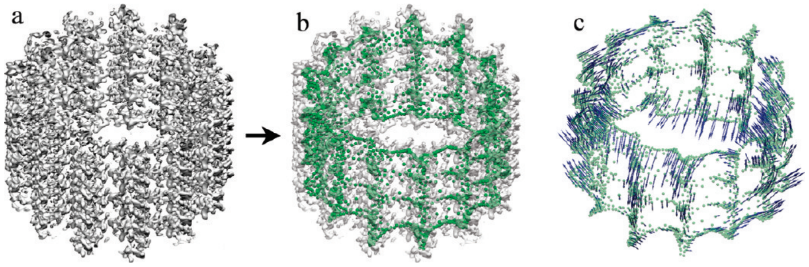

Figure 8. From (a) the 8.0 Å density map of the microtubule,[98] with its plus end towards the top, (b) pseudoatom model with 3200 pseudoatoms was built by the VQ algorithm in Sculptor.[63] (c) First mode from pfENM of the pseudoatom model.

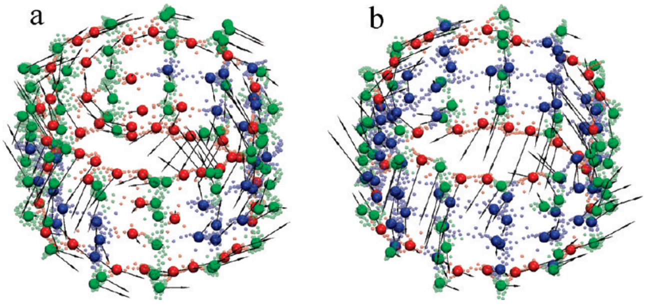

Figure 9. Space-based CG models of the microtubule from its pseudoatom model (Figure 8b). (a) Space-based ED-CG 156-site model. (b) VQ-CG 156-site model. Longitudinal CG sites that represent the pseudoatoms along the pfs are colored by green, lateral CG sites that represent the pseudoatoms between the pfs are colored by red, and mixed CG sites that contain pseudoatoms both along and between the pfs are colored by blue. First pfENM mode of each model is shown by arrows, respectively.

complete monomers in the middle, an incomplete monomer on the top, and another incomplete one at the bottom), which are connected though longitudinal contacts. At this resolution, the α and β tubulins are essentially not distinguishable because their structures are very similar.[99] However, many secondary structures are still visible in the monomers.

156-Site CG Microtubule Models. A total of 3200 pseudoatoms were generated by the VQ method (Figure 8b) from the MT density map, and the first pfENM mode of this pseudoatom model is shown in Figure 8c. As far as we know, no work about the functional modes of the MT has been reported yet, except for the tubulin dimer.[100,101] The top mode of the MT (Figure 8c) indicates a bending motion of the MT and a twist between pfs, which may change the shape of the MT. The fluctuations of the plus and minus ends are antiparallel. A space-based ED-CG 156-site model was directly defined from the pseudoatom model. A total of 2240 initial sets of cluster seeds were used, and $n_{\mathrm{SA}} =$ 500 000 steps were performed for each set to minimize the residual (eq 7). This system is larger than the ribosome, and more CG sites were defined, therefore, almost 1 h was needed to complete the calculation of a single set of cluster seeds. The space-based ED-CG 156-site model (Figure 9a) has the lowest residual of 264, whereas the VQ-CG 156-site model (Figure 9b) has a higher residual of 366. In the space-based ED-CG 156-site model, the top and bottom of the MT (incomplete monomers) have more CG sites than the complete monomers in the middle (Figure 9a). The pseudoatoms near the top and bottom are more flexible in pfENM due to fewer interactions than those in the middle (Figure 8b); therefore, more CG sites reside in the top and bottom regions of the MT in order to represent their larger fluctuations (Figure 8c). In the VQ-CG 156-site model, the distribution of CG sites between the top, bottom, and middle is more even (Figure 9b).

Another notable feature of the space-based ED-CG 156-site model is that more CG sites are found to represent lateral interactions between pfs (Figure 9a). In the MT, longitudinal contacts between tubulin monomers are stronger than lateral contacts between pfs.[94,102] By looking at the pseudoatom model (Figure 8b), many pseudoatoms are located along the pfs with a high density, but the pseudoatoms between the pfs are much less. According to the representing pseudoatoms, the CG sites are divided into three groups: CG sites that represent the pseudoatoms along the pfs (longitudinal sites, Figure 9, green), CG sites that represent the pseudoatoms between the pfs (lateral sites, Figure 9, red), and CG sites that contain pseudoatoms both along and between the pfs (mixed sites, Figure 9, blue). In the space-based ED-CG 156-site model, there are 74 longitudinal, 55 lateral, and 27 mixed sites. The corresponding numbers in the VQ-CG 156-site model are 59, 34, and 63, respectively. There are only 34 lateral sites in the VQ-CG 156-site model, which is consistent with its mass distribution. The majority of the CG sites (63) are mixed sites. It is believed that the lateral interactions between pfs are critical to regulate assembly/disassembly (dynamic instability)[103] of the MT,[94,98] so fewer lateral CG sites indicates that the VQ-CG 156-site model may not be able to describe the MT dynamics as well. In contrast, there are a larger number (55) of the lateral CG sites in the space-based EG-CG 156-site model and only 27 mixed sites (Figure 9a). This distribution of CG sites suggests that the longitudinal and lateral interactions are more clearly separated in the space-based ED-CG model. The lateral interactions are better represented than those in the VQ-CG model by comparing the first mode between the two CG models (Figure 9a and b). The space-based ED-CG 156-site model seems better to preserve the essential MT dynamics than the VQ-CG 156-site model, so the former may be used to more faithfully study elastic properties of the MT at the CG level.[104]

Conclusions

The sheer size of many large biomolecular complexes greatly complicates attempts to solve their high-resolution atomistic structures by X-ray crystallography or NMR. Alternative techniques, such as cryo-electron microscopy (cryo-EM)[23,24] and small-angle X-ray scattering (SAXS),[25−27] provide low-resolution structural information in many cases. These large biological assemblies are therefore also ideal candidates for coarse-grained computational modeling. It is thus an important priority to directly define coarse-grained (CG) models using the available low-resolution structural data for these systems. The main focus of this article is a new and important extension of the essential dynamics coarse-graining (ED-CG) methodology,[20,21] in order to build CG models directly from three-dimensional (3D) density maps obtained from cryo-EM or SAXS. First, a density map is discretized into a pseudoatom model by the vector quantization (VQ) method to retain the molecular shape,[32−34] and a pfENM is then constructed. A number of studies suggests that such a pseudoatom model is indeed sufficient to describe the low-frequency dynamics of the biomolecular system because they are mainly shape dependent.[105,106] Second, the essential dynamics defined by the low-frequency modes are used to determine space-based CG sites (eq 7). By definition in the present method, a CG site is the central-of-mass (COM) of a group of pseudoatoms that are close in space and move in a correlated way. Therefore, the search algorithm here is different from that in the previous sequence-based ED-CG method.[20,21]

The space-based ED-CG method can certainly also be applied to detailed atomistic structures. The resulting space-based ED-CG four-site model from the G-actin atomistic structure (Figure 1b) is almost the same as the intuitive four-site model (Figure 1a) but has a slightly lower variational residual. Upon comparison with the sequence-based ED-CG four-site model (Figure 1c), the space-based ED-CG model looks much more reasonable, and its residual is significantly lower. For the seven-site CG models from the atomistic structure of G-actin, the space-based ED-CG model (Figure 3b) also has a significantly lower residual than the sequence-based model (Figure 3c). These results indicate that the space-based ED-CG algorithm may find a CG model which can be better suited to preserve the essential dynamics than the CG model found by using the sequence-based ED-CG method as long as no large-scale conformational changes (e.g., unfolding) occur between groups of atoms that are space-closed in the structure used for coarse-graining. However, the space-based ED-CG search is more complicated than the sequence-based search since the former needs to explore many more possibilities. Therefore, a global minimum could not be reached even for the four-site CG model of G-actin, and the residual must be further minimized by smoothing the domains along the protein sequence. The space-based ED-CG method is also computationally more expensive than the sequence-based ED-CG method because there are additional distance calculations when clustering the atoms (steps 2 and 3 in the space-based ED-CG algorithm).

For a biomolecule with an atomic-resolution structure, one can use either space- or sequence-based ED-CG methods to define CG models. If a CG site needs to represent a large number of atoms (a relative low-resolution CG model), such as the four-site model of G-actin, then the sequence-based model may be unreasonable (Figure 1c), and one should instead choose the space-based ED-CG method (Figure 1b). For a relatively high-resolution model, i.e., each CG site represents only a small number of atoms, the sequence-based ED-CG method should perform well. In particular, when the system is large and the resolution of the CG model is high (i.e., many CG sites), the space-based ED-CG calculation becomes time consuming, and the sequence-based method is recommended.

For a density map with no atomic detail, the space-based ED-CG method is the only option to define ED-CG models. When a 10 Å density map of G-actin was created from its atomistic structure (Figure 4), the space-based ED-CG four-site model from the density map (Figure 5a) is found to be very close to the one obtained from the atomistic structure (Figure 1b). Furthermore, the space-based ED-CG seven-site model from the density map (Figure 5c) looks similar to the one from the atomistic structure (Figure 3b). These results indicate that the space-based ED-CG models from the low-resolution density map still do a good job of preserving the functional essential dynamics, which is in turn better than the corresponding VQ-CG models (Figure 5 b and d).

The ED-CG calculations form the density maps of the *E. coli* 70S ribosome and the microtubule (MT) (11.5 and 8 Å resolution, respectively) are very promising. At a very aggressive CG level, such as a 40-site representation of the ribosome, a space-based ED-CG model (Figure 7a) can still describe the ratchet-like motion between the small and large subunit and does so better than the VQ-CG 40-site model. Furthermore, regions that are important in functional dynamics, such as the head and spur in the small subunit and the L1 and L7/L12 stalks in the large subunit, contain more CG sites in the space-based ED-CG model than in the corresponding VQ-CG model (Figure 7b). In the space-based ED-CG 156-site model of the MT (Figure 9a), the lateral interactions between pfs are better represented than those in the VQ-CG 156-site model (Figure 9b). Therefore the space-based ED-CG model may be superior in preserving the MT dynamics (a precursor to dynamic instability) at the CG level. In the MT density map (Figure 8a), there are incomplete tubulin monomers on the top and at the bottom. After obtaining pfENM modes from the pseudoatom model, we can just use the subset of pseudoatoms that are in the two layers of complete monomers to define CG sites. A CG model of a longer MT may be built by duplicating the CG sites in this very short MT segment.

Computational cost of the space-based ED-CG method depends on both the number of pseudoatoms (n) from the density map and the number of CG sites (N) to be defined. The number of SA steps starting from one cluster-seed set is at least set as $n \times N$, and also the CPU time needed in one step is increasing with the system size. For a small system like G-actin, the calculation of a single set of cluster seeds is really fast (in a few seconds), and all the 4000 sets can be done within a couple of hours on a regular desktop computer. When systems become larger, such as the ribosome and the microtubule, they are computationally more expensive. However, the CPU time for a single cluster-seed set of the ribosome and the microtubule is still within 10 min and 1 h, respectively. Furthermore, all the cluster-seed sets can be distributed onto multiple CPUs (as many as one can have) at the same time, since they are independent of each other.

It should be noted that a space-based ED-CG *N*-site model is defined to capture the functional essential dynamics of a pseudoatom model constructed from a continuous density map by the VQ method. The number of pseudoatoms *n* should be much larger than the number of the CG sites *N*, since the ED-CG method is best suited to define a relatively small number of CG sites for a large biomolecule. One can also define a shape-based VQ-CG *N*-site model from the pseudoatom model, but it is not as good at describing the CG functional dynamics of the system as the space-based ED-CG model.

Before applying the space-based ED-CG method to general low-resolution structures, one should pay attention to the quality of density maps because they may contain noisy data. A real density map does not look like that shown in Figure 6a. There are actually many small satellite densities floating around the molecular surface. Fortunately these satellite densities can be ignored by using a proper threshold in the VQ calculation, so no pseudoatoms will be placed for these small noisy regions. Another subject of future research is to improve the convergence of the space-based ED-CG search. However, this approach as it stands now can lead the way to the systematic development of highly coarse-grained models of many large biomolecular complexes and thus ultimately to the CG computational modeling of such systems without the existence of high-resolution experimental structures.

Acknowledgment. This work is supported by a Collaborative Research in Chemistry grant from the National Science Foundation (CHE-0628257). The authors wish to thank Dr. Kenneth Downing for providing them with the 8 Å density map of the microtubule.[98] Z.Z. thanks Dr. Andrea Grafmüller, Dr. Edward Lyman, and Marissa Saunders for comments on the manuscript. Computer software is available upon request.

References

(1) Steitz, T. A. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 242–253.

(2) Schmeing, T. M.; Ramakrishnan, V. *Nature* **2009**, *461*, 1234–1242.

(3) Yonath, A. *J. R. Soc., Interface* **2009**, *6*, S575–S585.

(4) Reisler, E.; Egelman, E. H. *J. Biol. Chem.* **2007**, *282*, 36133–36137.

(5) Pollard, T. D.; Cooper, J. A. *Science* **2009**, *326*, 1208–1212.

(6) Nogales, E. *Cell. Mol. Life Sci.* **1999**, *56*, 133–142.

(7) Nogales, E. *Annu. Rev. Biochem.* **2000**, *69*, 277–302.

(8) Nogales, E.; Wang, H. W.; Niederstrasser, H. *Curr. Opin. Struct. Biol.* **2003**, *13*, 256–261.

(9) Nogales, E.; Wang, H. W. *Curr. Opin. Struct. Biol.* **2006**, *16*, 221–229.

(10) Brun, L.; Rupp, B.; Ward, J. J.; Nedelec, F. *Proc. Natl Acad. Sci. U.S.A.* **2009**, *106*, 21173–21178.

(11) van der Vaart, B.; Akhmanova, A.; Straube, A. *Biochem. Soc. Trans.* **2009**, *37*, 1007–1013.

(12) Li, J.; Lykotrafitis, G.; Dao, M.; Suresh, S. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4937–4942.

(13) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(14) Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589–1615.

(15) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.

(16) Ayton, G. S.; Noid, W. G.; Voth, G. A. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192–198.

(17) Sherwood, P.; Brooks, B. R.; Sansom, M. S. P. *Curr. Opin. Struct. Biol.* **2008**, *18*, 630–640.

(18) *Coarse-graining of condensed phase and biomolecular systems*; Voth, G. A., Ed.; CRC Press: New York, 2009.

(19) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1869–1892.

(20) Zhang, Z.; Lu, L.; Noid, W. G.; krishna, V.; Pfaendtner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 5073–5083.

(21) Zhang, Z. Y.; Pfaendtner, J.; Grafmüller, A.; Voth, G. A. *Biophys. J.* **2009**, *97*, 2327–2337.

(22) Amadei, A.; Linnsen, A. B. M.; Berendsen, H. J. C. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412–425.

(23) Saibil, H. R. *Nat. Struct. Biol.* **2000**, *7*, 711–714.

(24) Joachim, F. *Three-dimensional electron microscopy of macromolecular assemblies*; Oxford University Press: New York, 2006.

(25) Koch, M. H. J.; Vachette, P.; Svergun, D. I. *Q. Rev. Biophys.* **2003**, *36*, 147–227.

(26) Lipfert, J.; Doniach, S. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 307–327.

(27) Putnam, C. D.; Hammel, M.; Hura, G. L.; Tainer, J. A. *Q. Rev. Biophys.* **2007**, *40*, 191–285.

(28) Linde, Y.; Buzo, A.; Gray, R. M. *IEEE Trans. Commun.* **1980**, *28*, 84–95.

(29) Martinetz, T.; Schulten, K. *Neural Networks* **1994**, *7*, 507–522.

(30) Wriggers, W.; Milligan, R. A.; Schulten, K.; McCammon, J. A. *J. Mol. Biol.* **1998**, *284*, 1247–1254.

(31) Wriggers, W.; Chacón, P.; Kovacs, J.; Tama, F.; Birmanns, S. *Neurocomputing* **2004**, *56*, 165–179.

(32) Ming, D.; Kong, Y.; Lambert, M. A.; Huang, Z.; Ma, J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8620–8625.

(33) Ming, D.; Kong, Y.; Wakil, S. J.; Brink, J.; Ma, J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7895–7899.

(34) Tama, F.; Wriggers, W.; Brooks, C. L. *J. Mol. Biol.* **2002**, *321*, 297–305.

(35) Chacón, P.; Tama, F.; Wriggers, W. *J. Mol. Biol.* **2003**, *326*, 485–492.

(36) Arkhipov, A.; Freddolino, P. L.; Imada, K.; Namba, K.; Schulten, K. *Biophys. J.* **2006**, *91*, 4589–4597.

(37) Arkhipov, A.; Freddolino, P. L.; Schulten, K. *Structure* **2006**, *14*, 1767–1777.

(38) Arkhipov, A.; Yin, Y.; Schulten, K. *Biophys. J.* **2008**, *95*, 2806–2821.

(39) Yin, Y.; Arkhipov, A.; Schulten, K. *Structure* **2009**, *17*, 882–892.

(40) Bahar, I.; Rader, A. J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586–592.

(41) Ma, J. P. *Structure* **2005**, *13*, 373–380.

(42) Yang, L.; Song, G.; Jernigan, R. L. *Biophys. J.* **2007**, *93*, 920–929.

(43) Tama, F.; Valle, M.; Frank, J.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9319–9323.

(44) Wang, Y. M.; Rader, A. J.; Bahar, I.; Jernigan, R. L. *J. Struct. Biol.* **2004**, *147*, 302–314.

(45) Trylska, J.; Tozzini, V.; McCammon, J. A. *Biophys. J.* **2005**, *89*, 1455–1463.

(46) Kurkcuoglu, O.; Doruker, P.; Sen, T. Z.; Kloczkowski, A.; Jernigan, R. L. *Phys. Biol.* **2008**, *5*, 046005(14).

(47) Tirion, M. M. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.

(48) Hinsen, K.; Petrescu, A. J.; Dellerue, S.; Bellissent-Funel, M. C.; Kneller, G. R. *Chem. Phys.* **2000**, *261*, 25–37.

(49) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. *Biophys. J.* **2001**, *80*, 505–515.

(50) Lu, M. Y.; Poon, B.; Ma, J. P. *J. Chem. Theory Comput.* **2006**, *2*, 464–471.

(51) Moritsugu, K.; Smith, J. C. *Biophys. J.* **2007**, *93*, 3460–3469.

(52) Hinsen, K. *Bioinformatics* **2008**, *24*, 521–528.

(53) Lyman, E.; Pfaendtner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 4183–4192.

(54) Hinsen, K. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, E128−E128.

(55) Riccardi, D.; Cui, Q.; Phillips, G. N. *Biophys. J.* **2009**, *96*, 464–475.

(56) Stember, J. N.; Wriggers, W. *J. Chem. Phys.* **2009**, *131*.

(57) Yang, L.; Song, G.; Jernigan, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 12347–12352.

(58) Kitao, A.; Go, N. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164–169.

(59) Berendsen, H. J. C.; Hayward, S. *Curr. Opin. Struct. Biol.* **2000**, *10*, 165–169.

(60) Flory, P. J.; Gordon, M.; McCrum, N. G. *Proc. R. Soc. London, Ser. A* **1976**, *351*, 351–380.

(61) Wriggers, W.; Milligan, R. A.; McCammon, J. A. *J. Struct. Biol.* **1999**, *125*, 185–195.

(62) Wriggers, W. *Biophys. Rev.* **2010**, *2*, 21–27.

(63) Heyd, J.; Birmanns, S. *Microsc. Today* **2008**, *16*, 6–8.

(64) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

(65) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(66) Kirkpatrick, S. C. D.; Gelatt, J.; Vecchi, M. P. *Science* **1983**, *220*, 671–680.

(67) Graceffa, P.; Dominguez, R. *J. Biol. Chem.* **2003**, *278*, 34172–34180.

(68) Kabsch, W.; Mannherz, H. G.; Suck, D.; Pai, E. F.; Holmes, K. C. *Nature* **1990**, *347*, 37–44.

(69) Khaitlina, S. Y.; Moraczewska, J.; Strzeleckagolaszewska, H. *Eur. J. Biochem.* **1993**, *218*, 911–920.

(70) Pfaendtner, J.; Branduardi, D.; Parrinello, M.; Pollard, T. D.; Voth, G. A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 12723–12728.

(71) Chu, J. W.; Voth, G. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13111–13116.

(72) Chu, J. W.; Voth, G. A. *Biophys. J.* **2006**, *90*, 1572–1582.

(73) Pfaendtner, J.; Lyman, E.; Pollard, T. D.; Voth, G. A. *J. Mol. Biol.* **2010**, *396*, 252–263.

(74) Tirion, M. M.; Benavraham, D. *J. Mol. Biol.* **1993**, *230*, 186–195.

(75) Frank, J.; Agrawal, R. K. *Nature* **2000**, *406*, 318–322.

(76) Gabashvili, I. S.; Agrawal, R. K.; Spahn, C. M. T.; Grassucci, R. A.; Svergun, D. I.; Frank, J.; Penczek, P. *Cell* **2000**, *100*, 537–549.

(77) Wimberly, B. T.; Brodersen, D. E.; Clemons, W. M.; Morgan-Warren, R. J.; Carter, A. P.; Vonrhein, C.; Hartsch, T.; Ramakrishnan, V. *Nature* **2000**, *407*, 327–339.

(78) Yusupov, M. M.; Yusupova, G. Z.; Baucom, A.; Lieberman, K.; Earnest, T. N.; Cate, J. H. D.; Noller, H. F. *Science* **2001**, *292*, 883–896.

(79) Schuwirth, B. S.; Borovinskaya, M. A.; Hau, C. W.; Zhang, W.; Vila-Sanjurjo, A.; Holton, J. M.; Cate, J. H. D. *Science* **2005**, *310*, 827–834.

(80) Korostelev, A.; Trakhanov, S.; Laurberg, M.; Noller, H. F. *Cell* **2006**, *126*, 1065–1077.

(81) Selmer, M.; Dunham, C. M.; Murphy, F. V.; Weixlbaumer, A.; Petry, S.; Kelley, A. C.; Weir, J. R.; Ramakrishnan, V. *Science* **2006**, *313*, 1935–1942.

(82) Korostelev, A.; Noller, H. F. *Trends Biochem. Sci.* **2007**, *32*, 434–441.

(83) Sanbonmatsu, K. Y.; Joseph, S.; Tung, C. S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15854–15859.

(84) Sanbonmatsu, K. Y.; Tung, C. S. *J. Struct. Biol.* **2007**, *157*, 470–480.

(85) Gumbart, J.; Trabuco, L. G.; Schreiner, E.; Villa, E.; Schulten, K. *Structure* **2009**, *17*, 1453–1464.

(86) Villa, E.; Sengupta, J.; Trabuco, L. G.; LeBarron, J.; Baxter, W. T.; Shaikh, T. R.; Grassucci, R. A.; Nissen, P.; Ehrenberg, M.; Schulten, K.; Frank, J. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 1063–1068.

(87) Ermolenko, D. N.; Majumdar, Z. K.; Hickerson, R. P.; Spiegel, P. C.; Clegg, R. M.; Noller, H. F. *J. Mol. Biol.* **2007**, *370*, 530–540.

(88) Downing, K. H.; Nogales, E. *Eur. Biophys. J. Biophy.* **1998**, *27*, 431–436.

(89) Nogales, E.; Wolf, S. G.; Downing, K. H. *Nature* **1998**, *391*, 199–203.

(90) Downing, K. H.; Nogales, E. *Cell Struct. Funct.* **1999**, *24*, 269–275.

(91) Downing, K. H. *Annu. Rev. Cell. Dev. Biol.* **2000**, *16*, 89–111.

(92) Lowe, J.; Li, H.; Downing, K. H.; Nogales, E. *J. Mol. Biol.* **2001**, *313*, 1045–1057.

(93) Downing, K. H. *Scanning* **2003**, *25*, 74–75.

(94) Nogales, E.; Whittaker, M.; Milligan, R. A.; Downing, K. H. *Cell* **1999**, *96*, 79–88.

(95) Sosa, H.; Milligan, R. A. *J. Mol. Biol.* **1996**, *260*, 743–755.

(96) Sosa, H.; Dias, D. P.; Hoenger, A.; Whittaker, M.; WilsonKubalek, E.; Sablin, E.; Fletterick, R. J.; Vale, R. D.; Milligan, R. A. *Cell* **1997**, *90*, 217–224.

(97) Meurer-Grob, P.; Kasparian, J.; Wade, R. H. *Biochemistry* **2001**, *40*, 8000–8008.

(98) Li, H. L.; DeRosier, D. J.; Nicholson, W. V.; Nogales, E.; Downing, K. H. *Structure* **2002**, *10*, 1317–1328.

(99) Nogales, E.; Wolf, S. G.; Khan, I. A.; Luduena, R. F.; Downing, K. H. *Nature* **1995**, *375*, 424–427.

(100) Keskin, O.; Durell, S. R.; Bahar, I.; Jernigan, R. L.; Covell, D. G. *Biophys. J.* **2002**, *83*, 663–680.

(101) Gebremichael, Y.; Chu, J. W.; Voth, G. A. *Biophys. J.* **2008**, *95*, 2487–2499.

(102) Sept, D.; Baker, N. A.; McCammon, J. A. *Protein Sci.* **2003**, *12*, 2257–2261.

(103) Mitchison, T.; Kirschner, M. *Nature* **1984**, *312*, 237–242.

(104) Janosi, I. M.; Chretien, D.; Flyvbjerg, H. *Eur. Biophys. J. Biophy.* **1998**, *27*, 501–513.

(105) Lu, M. Y.; Ma, J. P. *Biophys. J.* **2005**, *89*, 2395–2401.

(106) Tama, F.; Brooks, C. L. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 115–133.

CT100374A