Coarse-Graining of Volumes for Modeling of Structure and Dynamics in Electron Microscopy: Algorithm to Automatically Control Accuracy of Approximation

Slavica Jonić, *Member, IEEE*, and Carlos Óscar Sánchez Sorzano, *Senior Member, IEEE*

***Abstract*—Coarse-graining (or granularization) of structures from transmission electron microscopy (EM volumes) has been shown to be useful for a variety of structural analysis applications. Several methods perform coarse-graining of EM volumes using hard spheres or 3D Gaussian functions but they do not allow controlling automatically the volume approximation accuracy. To tackle this problem, we recently developed such a method. It is currently used by 3DEM Loupe web server and HEMNMA software to study macromolecular dynamics based on coarse-grained representations of EM volumes. In this paper, we give a detailed description of the implemented algorithm and fully analyze its performance, which was out of scope of our previous papers. The performance is analyzed in a controlled environment, in the context of studying structure and dynamics of macromolecular complexes. We show that this technique allows computing structures that are similar to atomic structures, by analyzing intermediate-resolution volumes. Additionally, we show that it allows sharpening of intermediate-resolution volumes. The full algorithm description allows its implementation in any other software package.**

***Index Terms*—Coarse-graining, dynamics, electron microscopy (EM), Gaussian functions, macromolecular complexes, modeling, radial basis functions, structure.**

I. Introduction

THE study of macromolecular structures provides valuable insight about the way proteins carry out their function in the cell and how they interact with each other. Protein functions are intimately linked to their structure and the analysis of this structure may reveal important facts about their biochemical and biophysical properties. Structural information is acquired through experimental techniques like X-ray diffraction, Nuclear Magnetic Resonance (NMR) or transmission Electron Microscopy (EM). X-ray diffraction and NMR produce high resolution structures in which the position of each atom is known (with some uncertainty due to experimental errors) but can be used only with crystalline matter (X-ray diffraction) or small-size systems (NMR). In contrast, EM produces structures of low-to-medium resolution (usually in the range between 2 nm and 0.4 nm, although this paradigm has recently changed with EM structures approaching the resolution of 0.2 nm) but allows studying large (diameter larger than 10 nm and molecular weight sometimes of several mega-Daltons) and flexible macromolecular complexes inaccessible to X-ray diffraction and NMR techniques. An EM structure contains a volumetric distribution of the electron density.

Reduced representations of EM density volumes with hard spheres or 3D Gaussian functions (coarse-graining or granularization of structures) have been shown to be useful for a variety of structural analysis applications such as studying topology of the complex [1], its conformational changes [2], [3], its hydrodynamic properties [4], or for aligning structures obtained at different resolutions (e.g., fitting of X-ray diffraction or NMR structures into EM volumes [5]). These methods represent the density using a certain type of probability density function. The majority of them is based on a neural network clustering approach for volume quantization and they use a reduced set of points whose overall probability density function approximates the density profile of the given volume (optimization of a density-weighted metric) [1], [2], [5]–[7]. The input parameters of these methods are a desired number of granules (spheres or Gaussian functions) or a desired maximum number of iterations to produce the granules. Since samples are taken randomly (with a probability distribution given by the EM density), a drawback of these methods is that more dense regions might be oversampled whereas less dense regions might be underrepresented. A solution to overcome this drawback is to use the number of samples that is sufficiently high to guarantee an adequate representation of low density regions.

The method proposed in [8] is different from previously cited methods as it uses expectation maximization algorithm to estimate parameters of a Gaussian-Mixture-Model (GMM) type of

Manuscript received February 27, 2015; revised July 13, 2015; accepted October 01, 2015. Date of publication October 09, 2015; date of current version January 21, 2016. The work was supported in part by the CNRS (France) and the CSIC (Spain) under Grant PICS 2011, in part by the French National Research Agency ANR under Grant ANR-11-BSV8-010-04, in part by the European Social Fund and the Ministerio de Educación y Ciencia under Grant "Ramón y Cajal" fellowship to COSS, in part by the Spanish Ministry of Economy and Competitiveness under Grants AIC–A–2011–0638 and BIO2010-16566, and in part by the Comunidad de Madrid under Grant CAM S2010/BMD-2305. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Charles Kervrann.

S. Jonić is with the IMPMC, Sorbonne Universités – CNRS UMR 7590, UPMC University Paris 6, MNHN, IRD UMR 206, 75005 Paris, France (e-mail: jonic@impmc.upmc.fr).

C. O. S. Sorzano is with the Biocomputing Unit, Centro Nacional de Biotecnología – CSIC, 28049 Madrid, Spain (e-mail: coss@cnb.csic.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2015.2489186

the probability density function. GMM is a linear combination of several Gaussian functions and Gaussian functions are the granules in the volume representation. However, as the other methods, the method proposed in [8] uses a fixed number of granules (the desired number of granules is the input parameter). As with other methods, allowing the user to set the final number of granules might lead to suboptimal representations (e.g., a too small number of granules might result in a significantly asymmetrical representation of a symmetrical structure). As the method in [8] has been developed for fitting multiple subunit atomic structures into low-resolution EM volumes, it is aimed at representing EM volumes with few large multivariate Gaussian functions (the experiments showed in [8] deal with less than 100 Gaussian functions for large complexes such as GroEL/ES at 0.15 nm resolution). This representation can be very useful to analyze the relative orientation of different subunits or large domains (given by their atomic structures) inside a low-resolution EM volume of the entire macromolecular complex, as shown in [8]. However, more and more structures are currently being obtained by EM at much higher resolutions (between 0.4 nm and 0.25 nm). Such high resolutions imply more precise fitting that could be realized using much larger numbers of granules but this would be a formidable optimization problem in [8]. Another solution would be to use a smaller volume approximation error (the error between the given volume and its granules-based approximation). The control of the volume approximation accuracy also helps with the underrepresentation problem, as it allows replacing the optimization of a nonuniformly weighted metric by the optimization of a uniformly weighted one. However, the cited methods do not allow controlling the volume approximation accuracy.

Normal mode analysis (NMA) is another example of application of granulated EM volumes and it is used for exploring conformational changes of macromolecular complexes [2]. NMA is based on modeling dynamics of a macromolecular complex by a linear combination of harmonic oscillations around a minimum-energy conformation. In EM, the obtained density volume is considered to contain a minimum-energy conformation of the complex. EM volumes are thus used as the reference structures for computing normal modes but NMA requires their granularization. In NMA literature, the use of a minimal number of granules that captures the overall shape is largely accepted and it is based on the principle that the low-frequency spectra will be similar when using smaller and larger numbers of granules [9], [10]. However, we have found that, in some NMA applications, the shape is not giving enough information and that more structural details are required [11]. Indeed, in NMA applications such as elastic projection matching using normal modes (elastic 3D-to-2D fitting), a good quality of volume projections (projections of a volume that is obtained from an elastically transformed granules-based representation) is required for a precise alignment [11]. This requires small volume approximation errors meaning much larger numbers of granules than those required for other NMA applications. In elastic 3D-to-2D fitting, structures with large numbers of granules are often required (usually, 10^3-10^4 granules) [11].

To tackle these problems, we have developed a granularization method for controlling the volume approximation accuracy that can result in using very large or very small numbers of granules depending on the application of interest (NMA, fitting, etc.). This method approximates a density volume using a collection of radial basis functions (RBFs), with 3D Gaussian functions as RBFs. Such functions form a set of control points that provide information about the shape and the density distribution of the macromolecular complex. Given a volume, a Gaussian-function standard deviation, and a target volume approximation error, the method adjusts the number, the position and the amplitude of the Gaussian functions to achieve the given target approximation accuracy. By making these Gaussian functions sufficiently small, we may approximate the volume to any desired level of accuracy (it has been shown that Gaussians are a Riesz basis of L2 functions [12]). A combination of very small Gaussian functions and a very small target approximation error will result in a large number of granules (Gaussian functions), and vice versa.

The granularization method that controls the volume approximation accuracy is currently used by *3DEM Loupe* web server [13] and *HEMNMA* software [14] to study macromolecular dynamics based on coarse-grained representations of EM volumes. However, it was previously only briefly mentioned as one of the software and web server tools. In this paper, we give a detailed description of the implemented algorithm and fully analyze its performance, which was out of scope of our previous papers. The full algorithm description allows its implementation in any other software package. Its performance is analyzed in a controlled environment using data of adenylate kinase (AK) and 70S ribosome (70S). The results of the performance analysis show some interesting properties of the method and we discuss their potential for structural and dynamics studies of macromolecular complexes.

II. From EM Volume to Gaussian Functions Using Volume Approximation Error Control

We formulate the problem in a function approximation framework with 3D Gaussian functions as radial basis functions (RBFs). In this way, the input volume density represented by the function $f(\mathbf{r})$ $\left(\mathbf{r} \in \mathbb{R}^3\right)$ is approximated as follows:

$$\hat{f}_N(\mathbf{r}) = \sum_{i=1}^{N} \omega_i K_\sigma\left(\|\mathbf{r} - \mathbf{r}_i\|\right), \tag{1}$$

where N is the number of RBFs, $K_\sigma(r)$ is the RBF with the width σ (i.e., the Gaussian function with the standard deviation σ, in angstroms, and the amplitude of 1), $\mathbf{r}_i$ is the position of the RBF center, $\|\mathbf{r} - \mathbf{r}_i\|$ is the Euclidean distance between the vectors $\mathbf{r}$ and $\mathbf{r}_i$, and $\omega_i > 0$ is the RBF weight (i.e., the amplitude attributed to the Gaussian function $K_\sigma(r)$).

The goal is to find a RBF representation (N, ω_i, and $\mathbf{r}_i$) so that the representation error with N RBFs, e_N, satisfies:

$$e_N = \frac{1}{V}\sum_{j=1}^{V} \frac{\left|f\left(\mathbf{r}_j\right) - \hat{f}_N\left(\mathbf{r}_j\right)\right|}{\Delta f} < \varepsilon, \tag{2}$$

where ε is the target volume approximation error, and Δf is the normalization factor describing the effective range of values in $f(\mathbf{r})$:

$$\Delta f = F^{-1}(1-\alpha) - F^{-1}(\alpha). \tag{3}$$

Here, $F^{-1}(x)$ is the inverse of the cumulative distribution function of the values of f; α determines the statistical confidence on the effective range; $\mathbf{r}_j$ is one of the locations at which the experimental volume density is compared to the approximated volume density; and V is the total number of the evaluation locations $\mathbf{r}_j$ (voxels). Note that the error e_N can be computed within a mask that defines the volume region occupied by the molecule. The advantage of using an effective range instead of the true density range is that the effective range is insensitive to data outliers that, in practice, are found in 3D EM reconstructions (the effective range is robust up to an outlier proportion of 2α). Note that α plays the role of a Type I error in the construction of a confidence interval, so that typical values for α are 0.0025, 0.025, 0.05. We usually use $\alpha = 0.025$. The choice of α does not make a practical difference for volumes without outliers. The error e_N represents the average relative error committed in the representation. For instance, $e_N < 0.01$ means that in average the function is represented with an error smaller than 1% of the effective range.

The minimization of the error e_N in a high dimensional space is prone to get trapped in local minima (e.g., for a fixed number of Gaussian functions, representations with 150,000 Gaussian functions would imply 600,000 optimization variables to determine amplitudes and positions of each Gaussian function). For this reason, we increase the number of Gaussian functions progressively, from a given initial number of Gaussian functions (referred to as the initial seeds parameter) using a given speed of adding the Gaussian functions (referred to as the grow seeds parameter) and concentrating every time the newly added Gaussian functions in regions with large errors. Given the current number of Gaussian functions N, the width σ, and the target error ε, we compute amplitudes and positions of Gaussian functions using gradient descent minimization of the error e_N. Although this approach is not guaranteed to converge to the globally optimal coarse-grain representation of the input volume, we have found that it provides rather good solutions as we show in this article.

The pseudo-code of the algorithm is given in **Fig. 1**. By analogy with atomic structures, from here on, we will refer to the obtained coarse-grain structures as "pseudoatomic structures" and to the Gaussian functions as "pseudoatoms." At each iteration, a test collection of Gaussian functions is converted into a volumetric structure that is then compared with the given volume to minimize the volume approximation error in the next iteration. This conversion is straight-forward and uses (1). The algorithm stops iterating when the desired target error is reached. Since any smooth function can be approximated with a desired accuracy using a sufficiently small RBF width, the method suggests decreasing the current RBF width when it gets stuck in a local minimum wherein the target approximation error is unattainable. The algorithm will stop iterating in this case that is tested by checking whether the number of pseudoatoms still changes between iterations.

```
$\hat{f}_0(\mathbf{r}) = 0$;
$N = 0$
while $e_N > \epsilon$ do
    /* Remove weak pseudoatoms and add new
       pseudoatoms                                  */
    if first iteration then
        // Add $N_0$ atoms
        repeat
            Place atom at $\mathbf{r}_i = \text{argmax} f(\mathbf{r}) - \hat{f}_N(\mathbf{r})$;
            $w_i = f(\mathbf{r}_i) - \hat{f}_N(\mathbf{r}_i)$;
            $N \leftarrow N + 1$
        $N_0$ times;
    else
        // Remove very light atoms
        repeat
            Remove the atom $i = \text{argmin } w_i$
            $N \leftarrow N - 1$;
        $\Delta N$ times;
        // Remove atoms where $\hat{f}_N > f$
        repeat
            $\mathbf{r}_{\max} = \text{argmax}(\hat{f}_N(\mathbf{r}) - f(\mathbf{r}))$
            if $\hat{f}_N(\mathbf{r}_{\max}) > f(\mathbf{r}_{\max})$ then
                Remove an atom in the vecinity of $\mathbf{r}_{\max}$;
                $N \leftarrow N - 1$;
            end
        $\Delta N$ times;
        // Remove very close atoms
        for $i, j$ such that $\|\mathbf{r}_i - \mathbf{r}_j\| < d_{min}$ do
            if $w_i < w_j$ then
                Remove atom $i$;
            else
                Remove atom $j$;
            end
            $N \leftarrow N - 1$;
        end
        // Add $4\Delta N$ atoms
        repeat
            Place atom at $\mathbf{r}_i = \text{argmax}(f(\mathbf{r}) - \hat{f}_N(\mathbf{r}))$;
            $w_i = f(\mathbf{r}_i) - \hat{f}_N(\mathbf{r}_i)$;
            $N \leftarrow N + 1$
        $4\Delta N$ times;
    end
    /* Optimize the weight and position of
       current pseudoatoms                          */
    Minimize till convergence $e_N$ w.r.t. $w_i$ and $\mathbf{r}_i$ using
    gradient descent;
    /* Check if the target error can be
       achieved                                     */
    if number of atoms did not change from last iteration then
        Suggest to decrease $\sigma$;
        Finish;
    end
end
```

Fig. 1. Proposed algorithm for coarse-graining of EM volumes. Given σ of Gaussian functions (pseudoatoms), an initial number of pseudoatoms N_0 (referred to as the initial seeds parameter), a percentage of pseudoatoms to add at every iteration (Δ referred to as the grow seeds parameter), the minimum distance between pseudoatoms $d_{\min}$, and a target error ε, this algorithm converts an input density volume $f(\mathbf{r})$ into a set of pseudoatoms characterized by a position $\mathbf{r}_i$ and a weight ω_i. Here, ΔN is the number of pseudoatoms calculated as the percentage Δ of the current number of pseudoatoms N.

Some pseudoatoms are added and some are removed at each iteration (this is also a feature of Growing Cell Structures [15]). More precisely, the algorithm removes weak pseudoatoms (with very small weights ω_i) and the pseudoatoms whose distance is below a given value ($d_{\min}$). We have found that the strategy of removing these pseudoatoms allows the optimization algorithm to place new pseudoatoms in the regions that are most in need of new pseudoatoms. This also allows adapting the pseudoatoms near the removed pseudoatoms to better represent the local intensity in the input volume, if necessary. For instance, the weakest pseudoatoms are the least contributing ones to the decrease of the approximation error when they are in high-density areas in which a high-weight (high-intensity) pseudoatom is already covering much of the local intensity. Because of the positivity constraint ($\omega_i > 0$), the regions in which the approximation has surpassed the function ($\hat{f}_N(\mathbf{r}) > f(\mathbf{r})$) cannot be recovered by the gradient descent. For this reason, some pseudoatoms are removed in these areas so that the gradient descent can later adapt to the function. Finally, new pseudoatoms are added at the locations where the approximation is worse. In the pseudo-code (**Fig. 1**), ΔN means the number of pseudoatoms calculated as the percentage Δ of the current number of pseudoatoms N. Given the value of the grow seeds parameter (the percentage Δ) and the current number of pseudoatoms (N), ΔN pseudoatoms are removed in each loop at most (e.g., in the loop that removes closest pseudoatoms, no pseudoatom will be removed if $d_{\min}$ is set to a very small value). This results in $3\Delta N$ removed pseudoatoms per iteration at most (three loops in **Fig. 1**). The number of added pseudoatoms per iteration is $4\Delta N$ (a single loop in **Fig. 1**). One can note that the number of pseudoatoms to add is larger than the maximum number of pseudoatoms to remove and that these numbers were set so that their ratio is 4/3. The goal is to have a ratio slightly larger than 1 so that the approximation volume can get closer to the target volume while slowly increasing the number of pseudoatoms.

Note that the method produces Gaussian functions of different amplitudes (**Fig. 1**) but can optionally be constrained to produce Gaussian functions of the same amplitude.

Summarizing, the most important parameters of the algorithm are the width of the pseudoatoms, σ, and the target approximation error, ε. The width of the pseudoatom is related to the resolution of the input volume (volumes of lower resolution can be represented with larger pseudoatoms), and it typically ranges from 0.7–0.8 for high-resolution volumes to 2–2.5 for low-resolution volumes. The target approximation error is related to the noise in the input volume. Experimental EM volumes are usually noisy. Their denoising is a delicate task as it affects structural details and can be devastating for the weakest ones. Thus, they are rarely truly denoised. Noisy, experimental EM volumes are approximated using larger target approximation errors (typically, $\varepsilon = 10 - 15\%$) to avoid approximating noise very accurately (with many pseudoatoms), whereas clean volumes such as appropriately denoised experimental or synthetic volumes are approximated using smaller target approximation errors (typically, $\varepsilon = 1 - 5\%$). The rest of parameters of the algorithm (the minimum distance between pseudoatoms $d_{\min}$, the initial seeds parameter, and the grow seeds parameter) mostly affect the speed of convergence of the algorithm. For instance, we usually choose $d_{\min}$ to be very small (10^{-3} of the voxel size), which allows pseudoatoms to be as close as they have to be in order to achieve the desired accuracy of the volume approximation. So, such small values of $d_{\min}$ may be only affecting the convergence speed but not the maximum achievable accuracy of the volume approximation. A too large number of initial seeds and a too large grow seeds parameter may affect the achievable accuracy of the volume approximation. As shown below, we have found that 300 initial seeds and the grow seeds parameter of 30% generally produce good results.

Occasionally, the algorithm cannot reach the desired accuracy. This happens because we have required a too accurate representation when using relatively large pseudoatoms. The solution is either to reduce the pseudoatom width or to sacrifice accuracy. Alternatively, we may decide to keep the current pseudoatom representation (even if it did not achieve the required accuracy). In practice, we have found that this is a useful way to remove noise from the input volume. In fact, a non-converged pseudoatomic approximation corresponds to the local minimum from which the standard deviation of the Gaussian function (σ) should be made smaller if we aim at continuing converging towards the given target approximation error (to obtain pseudoatomic representations with many more pseudoatoms that will well approximate finer details including noise). If we decide to stop the iterative process (not to reduce σ), the resulting pseudoatomic representation will contain Gaussian functions of larger σ. Approximations using Gaussian functions of larger σ will necessarily be smoother than those with Gaussian functions of smaller σ. For noisy volumes such as experimental EM volumes, smoother approximations usually mean volumes with less noise. This denoising possibility is currently being further explored as the subject of a separate research work, whereas this article shows the performance of the method using synthetic volumes without noise (computed by low-pass filtering of ground-truth atomic structures). More importantly, general guidelines acquired thanks to these synthetic-data experiments were used here to successfully analyze experimental EM volumes.

III. Results

In this section, we first show the effect of different granularization parameters on the volume approximation. Then, we show an application of the method in the context of exploring dynamics of macromolecular complexes. Also, we fully evaluate its performance for different desired coarse-grain representations and show its potential to create pseudoatomic structures approaching atomic structures. Finally, we show an approximation of an experimental EM volume.

A. Effect of Different Granularization Parameters on Volume Approximation

By tuning the parameters σ and ε, the coarse-grain representation can be varied from a representation with several tens of pseudoatoms (e.g., for a coarse 3D-to-3D fitting where a large complex described by an EM volume is fitted with atomic structures of the complex subunits or large domains of subunits [6], [8]) to a representation with several tens of thousands of pseudoatoms (e.g., for studying conformational changes using elastic 3D-to-2D fitting based on building fine pseudoatomic models from EM volumes and fine estimation of conformational motions by NMA of such models [11]). For instance, to decrease the number of pseudoatoms, larger values of σ or ε should be used. On the contrary, to increase the number of pseudoatoms, smaller values of σ or ε should be used. In Fig. 2, we show three different pseudoatomic representations of the same volume. The volume (volume size: $90 \times 90 \times 90$ voxels; voxel size: 0.2 nm

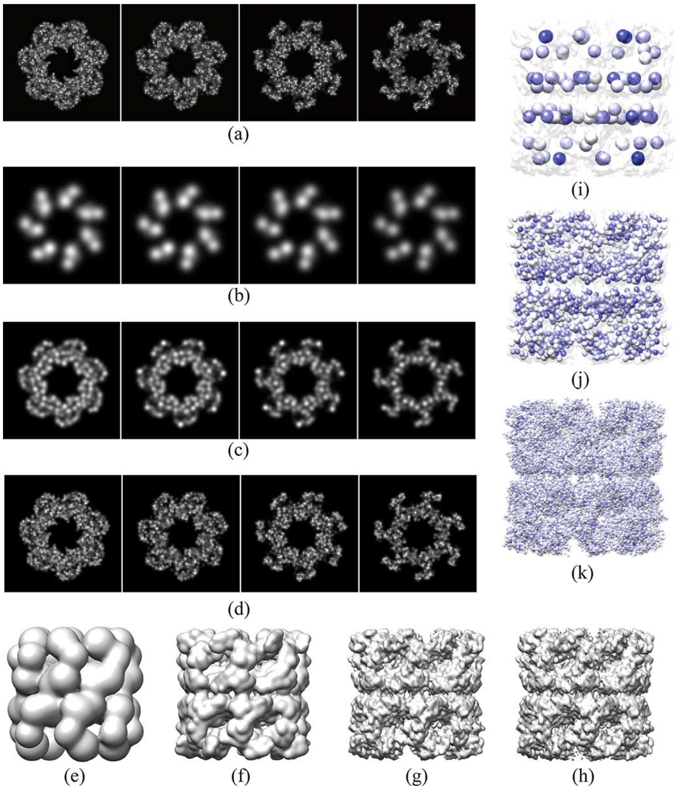


Fig. 2. Example of a volume and its approximations shown using volume slices, volume isosurface representation, and the corresponding 3D pseudoatomic representations. (a) Several central slices of a volume obtained from the GroEL atomic-resolution structure 1GRL. (b)-(d) Slices of the approximated volume for three different pseudoatomic representations, corresponding to the slices shown in (a) ((b) $\sigma = 7$ and $\varepsilon = 25\%$ resulting in 101 pseudoatoms; (c) $\sigma = 3$ and $\varepsilon = 15\%$ resulting in 1,764 pseudoatoms; (d) $\sigma = 1.5$ and $\varepsilon = 5\%$ resulting in 17,953 pseudoatoms). (e)-(g) Volumes whose slices are shown in (b)-(d), respectively. (h) Volume whose slices are shown in (a). (i)-(k) Isosurface representation of the original volume (transparent grey) and its pseudoatomic representations corresponding to (b)-(d), respectively. In (i)-(k), the pseudoatoms are shown as spheres of a radius value corresponding to the standard deviation of the Gaussian function, and the color of pseudoatoms corresponds to the weight ω_i (white: low weight; blue: high weight). In (e)-(k), the same orientation is shown for all volumes and pseudoatomic structures. The same volume isosurface is shown in (h)-(k).

× 0.2 nm × 0.2 nm) was obtained from an atomic-resolution GroEL structure (PDB entry: 1GRL [16]).

To compute a density volume from an atomic-resolution structure, throughout the paper, we used a method based on electronic-form atomic factors [17], [18], which is related to the images recorded by the microscope [19], [20]. **Fig. 2(a)** shows several central slices of the volume (its isosurface representation is shown in **Fig. 2(h))**. The volume was approximated using the following three pseudoatomic representations: 1) 101 pseudoatoms for $\varepsilon = 25\%$ and $\sigma = 7$ (**Fig. 2(b),(e),(i)**); 2) 1,764 pseudoatoms for $\varepsilon = 15\%$ and $\sigma = 3$ (**Fig. 2(c),(f),(j)**); and 3) 17,953 pseudoatoms for $\varepsilon = 5\%$ and $\sigma = 1.5$ (**Fig. 2(d),(g),(k)**). **Fig. 2** shows that the best quality of slices and isosurface representation was obtained for the volume approximated using $\varepsilon = 5\%$ and $\sigma = 1.5$, whereas the worse quality was obtained for $\varepsilon = 25\%$ and $\sigma = 7$. These results clearly show that the improvement of pseudoatomic representation quality can be achieved by reducing σ and ε.

B. Application: Analysis of Macromolecular Conformational Changes

We now show how such pseudoatomic structures can be used to explore macromolecular conformational changes. In this context, we use NMA. Given a structure, we compute normal modes using a standard elastic network model as it considers that the structure represents the minimum-energy conformation and, thus, it does not require energy minimization [21]. Normal modes are computed by diagonalizing a 3 N × 3 N Hessian matrix of second derivatives of the potential energy, where N is the number of nodes in the elastic network model. Nodes are 3D point particles that are connected with springs simulating harmonic restraints on displacements around the minimum-energy conformation, where each node is connected via springs only with those nodes that are within the interaction cutoff distance, R_c. The R_c parameter, thus determines the distance between nodes beyond which they do not interact. The nodes are atoms in the case of an atomic-resolution structure. In the case of an EM volume, nodes are a set of control points that have to be extracted from the volume. The coordinates of the nodes are modified to simulate the structural flexibility. For EM volumes, here, we use the coordinates of centers of the Gaussian functions as nodes of the elastic network model. The diagonalization of the Hessian matrix is done by solving an eigenvalue problem. The eigenvectors of the matrix are normal modes and the eigenvalues are the squares of the normal-mode frequencies. In the case of atomic structures, the R_c value of 0.8 nm usually gives good results. In the case of pseudoatomic structures, the interaction cut-off distance R_c should be adjusted to each protein complex (its size), the structure of the complex (its resolution), and the representation of that structure by nodes in the elastic network model (the number of nodes). In practice, typical pseudoatomic R_c values are in the range 1–3 nm. They can be adjusted by applying an ad-hoc rule by which the value of R_c is set so that 95% of pseudoatoms interact among each other (we have found it to work in most cases [13]). Pseudoatomic R_c values are usually larger than atomic R_c values because pseudoatomic distances are usually larger than atomic distances (the number of pseudoatoms is usually smaller than the number of all atoms in the complex).

Using adenylate kinase (AK) data, we now give an example showing that quality of the pseudoatomic structures is essential to correctly describe the conformational changes. We compute normal modes of an atomic AK structure (atomic modes) and normal modes of two different pseudoatomic representations of a synthetic AK density volume (pseudoatomic modes). Then, we compute the overlap of an AK experimentally observed conformational change with pseudoatomic modes and with atomic modes. The overlap between a normal mode and the experimental conformational change is computed as a normalized inner product between two vectors, where the pseudoatomic modes are extended to the atomic resolution by thin-plate spline interpolation [22], [23]. The overlap will be higher for the modes that contribute more to the conformational change than for those that contribute less. Ideally, the overlap should be the same for atomic and pseudoatomic modes. The same overlap for the two types of modes would mean that pseudoatomic modes describe the experimentally observed conformational change in the same way as atomic modes. This would also mean that pseudoatomic modes reproduce perfectly atomic modes. In practice, pseudoatomic and atomic modes are different, and the quality of pseudoatomic modes and their overlap with experimentally observed conformational changes will depend on the quality of the pseudoatomic representation (determined by σ and ε and a resulting number of pseudoatoms). Here, we compare pseudoatomic representations of two different qualities by analyzing the overlap of their modes with the experimental conformational change, and by comparing these two overlaps to the one obtained for the atomic representation.

A volume (volume size: $64 \times 64 \times 64$ voxels; voxel size: 0.2 nm × 0.2 nm × 0.2 nm) was obtained from an atomic-resolution structure of open-form AK (PDB entry 4AKE, containing 1,656 atoms [24]) and, then, low-pass filtered at 3 nm so that we get a very low resolution volume (**Fig. 3(a)–(c)**). This low-pass filtered volume was approximated using $\sigma = 2.5$ and $\varepsilon = 0.5\%$, which resulted in a pseudoatomic representation with 367 pseudoatoms (**Fig. 3(d)**). A higher-quality pseudoatomic representation of the same volume was obtained using $\sigma = 2$ and $\varepsilon = 0.4\%$, which resulted in a pseudoatomic representation with 922 pseudoatoms (**Fig. 3(e)**). The difference between the open-form and closed-form atomic AK structures was here used as the experimentally observed conformational change. The closed-form structure was obtained from the PDB entry 1AKE [25] and is also shown in **Fig. 3(a)**.

Fig. 3(f) shows the overlap of the experimental conformational change with atomic modes 7–20 (ground-truth overlap) where modes 7–20 correspond to frequencies 7–20, respectively. If the conformational difference could be completely explained by a single normal mode, the overlap of the difference with this mode would be 1 and its overlap with all other modes would be 0. This is clearly not the case here, which means that several modes contribute to the conformational change. Indeed, **Fig. 3(f)** shows that the conformational change is mostly contributed by mode 7 (the maximum overlap is at mode 7). However, other prominent peaks of the overlap are at modes 11 and 15 (**Fig. 3(f)**), which means that these modes also contribute to the conformational change, though less than mode 7.

Fig. 3(g) shows that the normal modes from the lower-quality pseudoatomic representation overlap with the observed conformational change similarly to the normal modes from the ground-truth atomic structure. This is an important result and we show that it is valid for different values of R_c between 1 nm and 2 nm. As shown in **Fig. 3(g)**, this pseudoatomic representation reproduces atomic normal modes well enough for several R_c values (all tested values below 1.8 nm), which means that such pseudoatomic representation is robust to the choice of the value of R_c. This is interesting when the R_c value is set automatically (e.g., using the mentioned rule of 95% interacting pseudoatoms) as other R_c values around the value set automatically could also work.

More importantly, **Fig. 3** shows that the ground-truth overlap is more similar to the overlap obtained for the modes from the higher-quality pseudoatomic representation than for those from the lower-quality pseudoatomic representation. More precisely, for the same value of R_c ($R_c = 1$ nm), the peaks of the overlap for the atomic modes are found at the same frequencies (modes

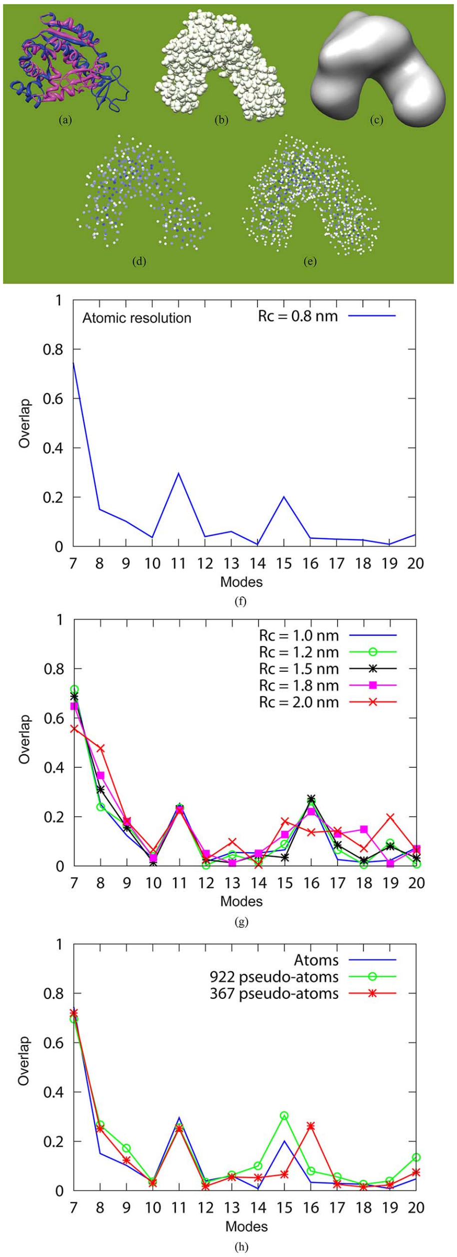


Fig. 3. Pseudoatomic-structure quality determining quality of approximation of atomic normal modes. (a) Overlapped atomic-resolution structures of the open and closed AK conformations (PDB codes: 4AKE (blue) and 1AKE (magenta), respectively). (b) All-atom representation of the open AK shown in (a) (atoms are shown as spheres of a radius corresponding to the atom radius). (c) Synthetic volume representing the open AK conformation at the resolution of 3 nm. (d) Pseudoatomic structure from the volume shown in (c), obtained for $\varepsilon = 0.5\%$ and $\sigma = 2.5$ (367 pseudoatoms). (e) Pseudoatomic structure from the volume shown in (c), obtained for $\varepsilon = 0.4\%$ and $\sigma = 2$ (922 pseudoatoms). (f)-(g) Overlap of the AK conformational change observed experimentally (difference between the open and closed conformations shown in (a)) with atomic modes (f) and with pseudoatomic modes of the 367-pseudoatom structure for different values of R_c (g). (h) Comparison of the 922-pseudoatom and 367-pseudoatom structures by showing the overlap of their normal modes with the experimental conformational change (the ground-truth overlap shown in (f) is also shown here, and $R_c = 1$ nm was used for both sets of pseudoatomic modes). In (d)-(e), pseudoatoms are represented by spheres using the same size and color coding as in Fig. 2.

7, 11, and 15) as for the modes from the higher-quality pseudoatomic representation (**Fig. 3(h)**). These peaks were obtained at slightly different frequencies for the modes from the lower-quality pseudoatomic representation (atomic mode 15 *vs* pseudoatomic mode 16, **Fig. 3(h)**).

If the pseudoatomic representation quality were lowered further, the overlap at lower frequencies (e.g., 7 and 11) would probably be also affected, beside the overlap at higher frequencies (e.g., 15). For instance, the peaks corresponding to modes 7 and 11 could move to different frequencies (examples of such shifts for mode 7 are given in [10]) or some additional but false peaks could appear (as in the example shown for the structure with 367 pseudoatoms and). For other applications of normal modes (elastic 3D-to-3D or 3D-to-2D fitting), it may be useful to pick automatically or semi-automatically the modes that are the most relevant to the ground-truth conformational change, without its knowledge [11], [26]–[29]. In a separate work, we have developed a protocol for picking such modes [11], [14]. Though such protocols could successfully deal with local changes in the order of normal modes, they cannot detect false-relevant modes that may result from a lower-quality pseudoatomic representation. Thus, high quality of pseudoatomic representations is generally recommended.

C. Towards Atomic Approximations: Detailed Performance Analysis

We now show a detailed performance analysis using a large system such as *E. coli* 70S ribosome. An atomic-resolution structure of the ribosome, with 10,204 Cα and phosphate atoms, referred to as 3I1OP structure (a composite of structures with PDB codes 3I1O and 3I1P [30]) (**Fig. 4(a)**) was converted into a volume (volume size: $128 \times 128 \times 128$ voxels; voxel size: 0.3 nm $\times$ 0.3 nm $\times$ 0.3 nm) (**Fig. 4(b)**) that was then low-pass filtered at an intermediate EM resolution of 1.5 nm (**Fig. 4(c)**).

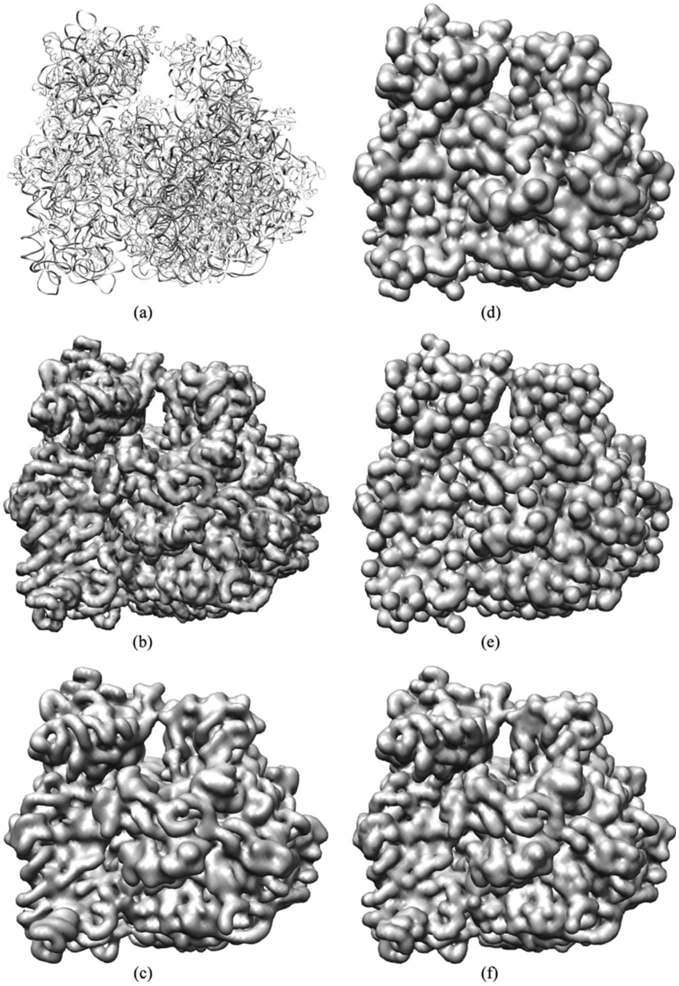
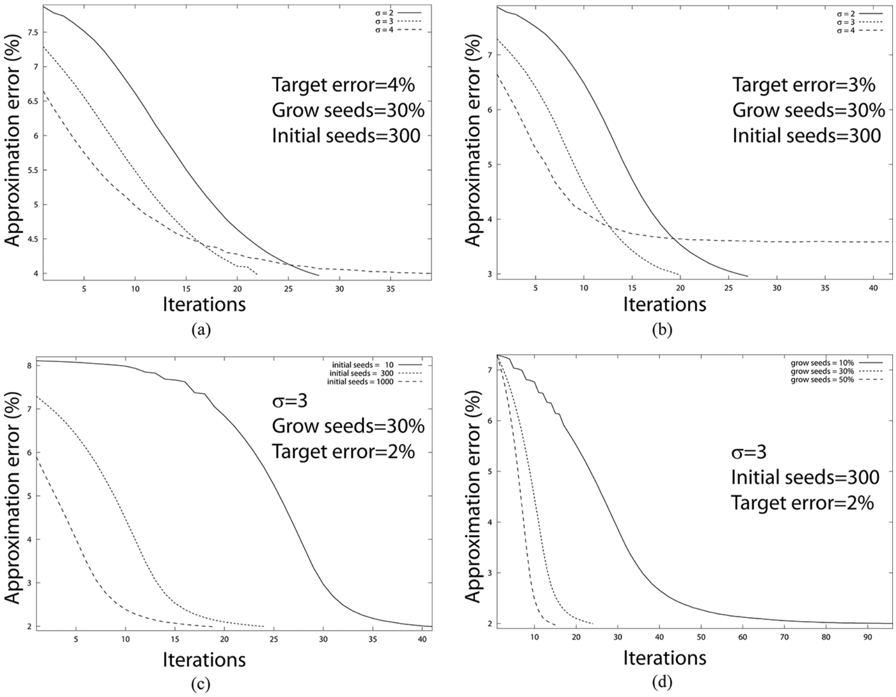


Fig. 4. Atomic structure of 70S ribosome, unfiltered synthetic volume, and low-pass filtered volume that was used for different pseudoatoms representations, along with three approximated volumes. (a) Atomic structure 3I1OP. (b) Synthetic volume from the structure shown in (a). (c) Volume shown in (b) after a low-pass filtering at 1.5 nm. (d)-(f) Volumes produced using three pseudoatomic representations of the volume in (c) ($\sigma = 4$ and $\varepsilon = 3\%$ resulting in 1,778 pseudoatoms (d); $\sigma = 3$ and $\varepsilon = 4\%$ resulting in 2,235 pseudoatoms (e); $\sigma = 3$ and $\varepsilon = 2\%$ resulting in 5,776 pseudoatoms (f)). The same values of the grow seeds parameter (30%) and the initial seeds parameter (300) were used to obtain the volumes in (d)-(f).

Pseudoatomic structures were computed using the low-pass filtered volume (**Fig. 4(c)**) and different combinations of the values of σ (2, 3, and 4), ε (2%, 3%, and 4%), the initial seeds parameter (10, 300, and 1,000), and the grow seeds parameter (10%, 30%, and 50%). The algorithm performance was compared for these different combinations of parameters regarding the convergence of the volume approximation error to its target value. **Fig. 5** shows that σ and ε are the most important parameters because they affect the final approximation result (**Fig. 5(a)–(b)**), whereas the initial seeds and grow seeds parameters affect the speed of convergence (the number of iterations) (**Fig. 5(c)–(d)**). **Fig. 5** also shows that σ and ε are tightly linked. For instance, when using $\sigma = 4$, it is possible to achieve $\varepsilon = 4\%$ (**Fig. 5(a)**) but not smaller ε such as $\varepsilon = 3\%$ (**Fig. 5(b)**). To achieve $\varepsilon = 3\%$, we need to reduce σ to a smaller value such as $\sigma = 3$ or $\sigma = 2$ (**Fig. 5(b)**). **Fig. 5** seems suggesting the use of larger initial seeds and grow seeds parameters to obtain a faster convergence (**Fig. 5(c)–(d)**). However, with some combinations of input data and parameter settings, a too large number of initial seeds or a too large grow seeds parameter could stop the iterations too early, which would prevent the method from achieving ε (the final approximation result would be affected in such cases). We thus recommend the use of 300 initial seeds and the grow seeds parameter of 30% that were found to produce good results in most cases. For instance, the use of $\varepsilon = 4\%$, $\sigma = 3, 300$ initial seeds, and the grow seeds parameter of 30% produced a 2235-pseudoatom representation of the ribosome volume (**Fig. 4(c)**) with the volume approximation error of 3.98% (the error of 0.5% with respect to target error). For the same ε and σ setting ($\varepsilon = 4\%$, $\sigma = 3$) but a very large number of initial seeds (3000) and a large grow seeds parameter (50%), we obtained a 3000-pseudoatom representation with the volume approximation error of 3.58% (the error of 10.5% with respect to target error). In fact, a single iteration was only performed in this second ("pathological") case, whereas 23 iterations were performed in the first case. In the second case, we used a larger number of initial seeds than the final number of pseudoatoms obtained in the first case. The reason for this is that, in reality, we would not know how to choose this (large) number and it could be "blindly" set to any value.



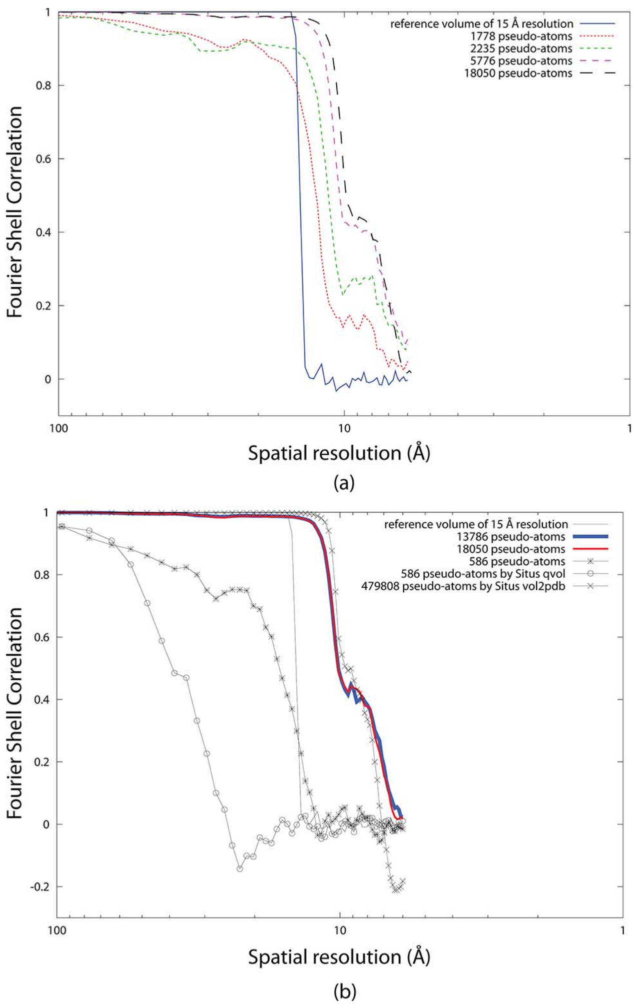
(a) (b) (c) (d)

Fig. 5. Approximation error at each iteration until the convergence for different combinations of parameters, for the synthetic volume of 70S ribosome shown in Fig. 4(c). (a)-(b) Use of three different values of σ (2, 3, and 4), the grow seeds parameter of 30%, 300 initial seeds, and $\varepsilon = 4\%$ (a) or $\varepsilon = 3\%$ (b). (c) Use of three different values of the initial seeds parameter (10, 300, and 1000), $\sigma = 3$, $\varepsilon = 2\%$, and the grow seeds parameter of 30%. (d) Use of three different values of the grow seeds parameter (10%, 30%, and 50%), $\sigma = 3$, $\varepsilon = 2\%$, and 300 initial seeds.

The approximation error shown in **Fig. 5** was also plotted against the number of pseudoatoms obtained at different iterations. This showed a virtually identical convergence behavior of the algorithm to the one already shown in **Fig. 5**, and the only new result was that different settings of parameters produce different final numbers of pseudoatoms. For instance, the number of pseudoatoms required for reaching the target approximation error of 2% increases from 5,776 pseudoatoms to 18,050, when reducing σ from 3 to 2.

As the number of pseudoatoms in the volume representation increases (as we reduce σ or ε), the curve of the Fourier shell correlation (FSC) between the resulting volume and the ideal, unfiltered volume (**Fig. 4(b)**) extends to higher (spatial) frequencies (**Fig. 6(a)**). The higher FSC at higher frequencies means that the resolution of the resulting volume increases, as shown in **Fig. 4(d)–(f)**. In the example shown in **Fig. 6(a)**, the highest FSC at all spatial frequencies was obtained for the 18,050-pseudoatom volume ($\sigma = 2, \varepsilon = 4\%$). However, this 18,050-pseudoatom FSC curve is very close to the 5,776-pseudoatom FSC curve (**Fig. 6(a)**). This suggests that the representation with 5,776 pseudoatoms ($\sigma = 3$, $\varepsilon = 2\%$) is sufficient to describe this complex, though it can still be improved (e.g., using smaller pseudoatoms ($\sigma = 2$)). Indeed, the 5,776-pseudoatom volume (**Fig. 4(f)**) looks as a good approximation of the ideal, unfiltered volume (**Fig. 4(b)**). This result also shows that the use of pseudoatom sizes that are similar to the voxel size can produce a sufficiently good approximation of the given volume ($\sigma = 3$ was used in the 5,776-pseudoatom representation while the voxel size was 3 Å × 3 Å × 3 Å).

Additionally, we show that σ can be adjusted to obtain an optimal pseudoatomic representation in terms of the similarity between the pseudoatomic and atomic distance histograms and that such "reproduction" of atomic distances does not require knowing the 3D atomic structure but the number of coarse-grain atoms only (Cα and phosphate atoms in the case of ribosome). In the example shown here, the ground-truth atomic structure contains around 10,000 Cα and phosphate atoms. The number of pseudoatoms closest to 10,000 was obtained for the volume approximation using $\sigma = 2.3$, $\varepsilon = 2\%$ (13,786 pseudoatoms). **Fig. 7** shows that the histogram of pseudoatomic distances is the most similar to the histogram of atomic distances for this representation ($\sigma = 2.3$, $\varepsilon = 2\%$, 13,786 pseudoatoms), when comparing it to the representations whose number of pseudoatoms was further away from 10,000 such as the 18,050-pseudoatom representation and the 5,776-pseudoatom representation. Interestingly, the 13,786-pseudoatom FSC curve is almost the same as the 18,050-pseudoatom FSC curve (**Fig. 6(b)**), which means that the two corresponding volumes are almost identical. Though reducing the size of pseudoatoms to increase their number from 5,776 to 13,786 still improves the representation of the complex (the corresponding FSC curves are slightly different), this last result suggests that increasing this number above 13,786 (by reducing the pseudoatom size further) does not significantly change the representation. Thus, these two complementary measures (the FSC and the distance histogram) show that the σ and ε combination producing the 13,786-pseudoatom representation is optimal regarding both the volume representation quality (the FSC remains almost unchanged or it degrades for the σ and ε combinations producing other pseudoatomic representations) and the "reproduction" of atomic distances (the pseudoatomic distance histogram is more dissimilar to the atomic distance histogram



(a)

(b)

Fig. 6. Fourier shell correlation (FSC) of the volume from the 70S ground-truth atomic structure (Fig. 4(b)) with the volumes obtained by the proposed method, the volumes obtained by the Situs software, and the reference volume used to compute the pseudoatoms (Fig. 4(c)). (a) FSC for the volumes from the pseudoatomic structures computed using the proposed method and FSC for the reference volume. (b) FSC for the volumes from the pseudoatomic structures obtained by two methods of Situs (*qvol* and *vol2pdb*), compared to the FSC for the volumes obtained by the proposed method (the FSC for the pseudoatomic structure that is optimal in terms of the distance histogram (Fig. 7) is also shown) and to the FSC for the reference volume. The proposed method was used with different σ and ε settings to obtain these different pseudoatom representations (586 pseudoatoms: $\sigma = 5, \varepsilon = 5.2\%$; 1,778 pseudoatoms: $\sigma = 4, \varepsilon = 3\%$; 2,235 pseudoatoms: $\sigma = 3, \varepsilon = 4\%$; 5,776 pseudoatoms: $\sigma = 3, \varepsilon = 2\%$; 13,786 pseudoatoms: $\sigma = 2.3, \varepsilon = 2\%$; 18,050 pseudoatoms: $\sigma = 2, \varepsilon = 2\%$).

for the σ and ε combinations producing other pseudoatomic representations).

Fig. 6 also shows that the volumes computed from pseudoatoms are sharpened versions of the intermediate-resolution volume that was used to compute the pseudoatoms. More precisely, **Fig. 6** shows that, at higher spatial frequencies (here, higher than 15 Å i.e., 1.5 nm), the ideal, unfiltered volume (**Fig. 4(b)**) correlates better with the volumes computed from pseudoatoms than with the 1.5 nm resolution volume (**Fig. 4(c)**) that was used to compute the pseudoatoms (the FSC is closer to 1 at the higher frequencies for the volumes computed from pseudoatoms). This can be explained by our regularization of the approximation problem using a prior information that is the approximate form of the density (radially symmetric functions around a discrete set of points), which allows us to recover information that is "not in the data." Furthermore, our choice of Gaussian functions as radially symmetric functions allows regulating volume sharpness by varying the standard deviation of Gaussian functions σ. Indeed, a smaller value of σ means a smaller support of each Gaussian function in the pseudoatomic representation, which in turn results in a sharper density volume from pseudoatoms.

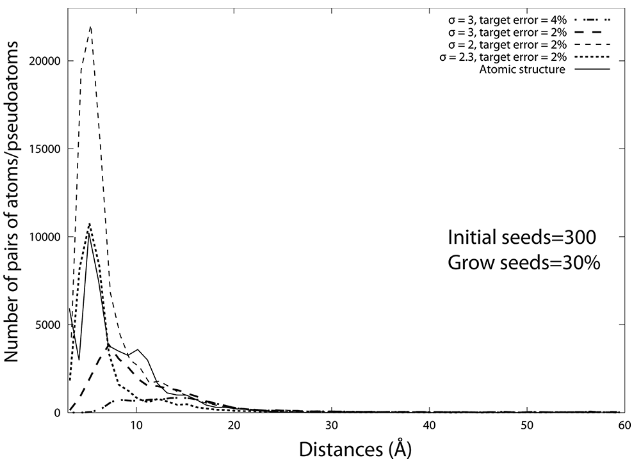


Fig. 7. Histogram of pseudoatomic distances overlapped with the histogram of 70S atomic distances. Different pseudoatomic-distance histograms were obtained for different pseudoatomic representations of the reference volume (Fig. 4(c)). The atomic-distance histogram was computed using the ground-truth atomic structure (Fig. 4(a)).

We compared our method with the methods *qvol* and *vol2pdb* of the current version of Situs (ver. 2.7.2) [31]. The *vol2pdb* method first rescales all density values to [0,99.99] and then places pseudoatoms at all densities that are above a cutoff density. The default value of 0.005 can be used for this cutoff or the cutoff can be set by the user. In both cases, this method produces a huge number of pseudoatoms (we tested several cutoff values for the 70S ribosome, which produced representations with 100,000–400,000 pseudoatoms). As it is difficult to decide which cutoff to select to describe the volume sufficiently well, this method was used with the default cutoff density setting for the comparison with other methods in this article. The *qvol* method is based on vector quantization. It requires that the user specifies a desired number of pseudoatoms. Though the method was previously shown to work with up to 2,000 pseudoatoms [2], we were unable to use it with more than 600 pseudoatoms. In fact, the default maximum number of pseudoatoms in *qvol* is 50 and Situs allows changing this number, by changing the corresponding parameter in situs.h and recompiling the software. In our work, we were able to make *qvol* working by changing this parameter only up to the value of 600. For larger values of this parameter, we were receiving a segmentation error message on a machine with enough memory for all other, currently standard EM methods (Dual Intel Xeon X5472 processor [3.00 GHz, 1600 FSB, 2 × 6 MB, Quad Core], 2GB RAM per core [800 MHz ECC Memory, 8 × 2 GB]). Note however that the number of 600 pseudoatoms may be large enough to produce good results in typical applications of *qvol*. Actually, *qvol* is usually employed for a fast rigid-body fitting of atomic structures of subunits or large subunit domains of a complex into its EM density volume [6], where smaller numbers of pseudoatoms can be used.

The volume shown in **Fig. 4(c)** was converted into pseudoatoms using *qvol* and *vol2pdb*, which produced 600 and 479,808 pseudoatoms, respectively. The same volume was converted using the proposed method and the σ and ε values producing 586 pseudoatoms (this was the closest number of pseudoatoms to the one obtained by *qvol* and it was obtained using $\sigma = 5$ and $\varepsilon = 5.2\%$). The obtained pseudoatomic representations were then converted into density volumes (the pseudoatomic representations obtained with Situs were converted into volumes using the *pdb2vol* method of Situs). The FSC of these three volumes with respect to the unfiltered ground-truth volume (**Fig. 4(b)**) are shown in **Fig. 6(b)**. **Fig. 6(b)** shows that *vol2pdb* produces a high-resolution pseudoatomic representation that is comparable to the one obtained with the proposed method (the 13,786-pseudoatom FSC curve obtained by the proposed method is very similar to the 479,808-pseudoatom FSC curve obtained by *vol2pdb*). However, the proposed method uses 35 times fewer pseudoatoms than *vol2pdb* to achieve such high-resolution representation. A larger number of pseudoatoms usually results in a slower analysis of the structure and a larger use of memory. For instance, with the huge number of pseudoatoms resulting from *vol2pdb*, it would be impossible to compute normal modes (the Hessian matrix would be huge and its diagonalization difficult). Finally, **Fig. 6(b)** shows that the 586-pseudoatom volume obtained by the proposed method (**Fig. 8(b)**) correlates with the ground-truth volume (**Fig. 8(a)**) better than the 600-pseudoatom volume obtained by *qvol* (**Fig. 8(c)**).

D. Experiments With Real EM Volumes

The main focus of this article is on evaluating the performance of the proposed method by comparing its results to the ground truth. This is the reason for the use of simulated data. However, we have also performed the tests using experimental EM density maps, which showed consistent results to those obtained using synthetic density maps. The results of some of those experiments are available on the *3DEM Loupe* web server [13]. Actually, the *3DEM Loupe* web site contains the precomputed results for EM density maps of 70S ribosome, GroEL, ribosome-bound termination factor RF2, and connector of bacteriophage T7, and allows a web user to visualize these results and download them.

In **Fig. 8(d)–(g)**, we show the results of the experiment with the real EM volume of the 70S ribosome (EMDB entry code EMD-5262; volume size: 125 × 125 × 125 voxels; voxel size: 0.3 nm × 0.3 nm × 0.3 nm). This EM volume (**Fig. 8(d)–(e)**) was approximated using $\sigma = 3.5$ and $\varepsilon = 10\%$, which resulted in a 5913-pseudoatom volume (**Fig. 8(d),(f)**). The most important result is given in **Fig. 8(g)** and it shows that the FSC between these two volumes falls below 0.5 at 13 Å (1.3 nm). This value is very close to the one declared as the resolution of the EM volume (13.2 Å [32]). As the ground-truth atomic structure is often unavailable, the resolution of real EM volumes is usually obtained as the value at which the FSC between two "half-volumes" (reconstructed from two halves of the total set of images) falls below 0.5. This was also the case with this EM volume [32]. Note however that the FSC computed in such way does not really measure the resolution of the volume but its consistency with other volumes that could be computed from the same set of images. The approximated volume (**Fig. 8(f)**) is thus consistent with the given EM volume (**Fig. 8(e)**) and with other volumes that could be computed from the same set of images up to the spatial resolution of 13 Å (**Fig. 8(g)**).

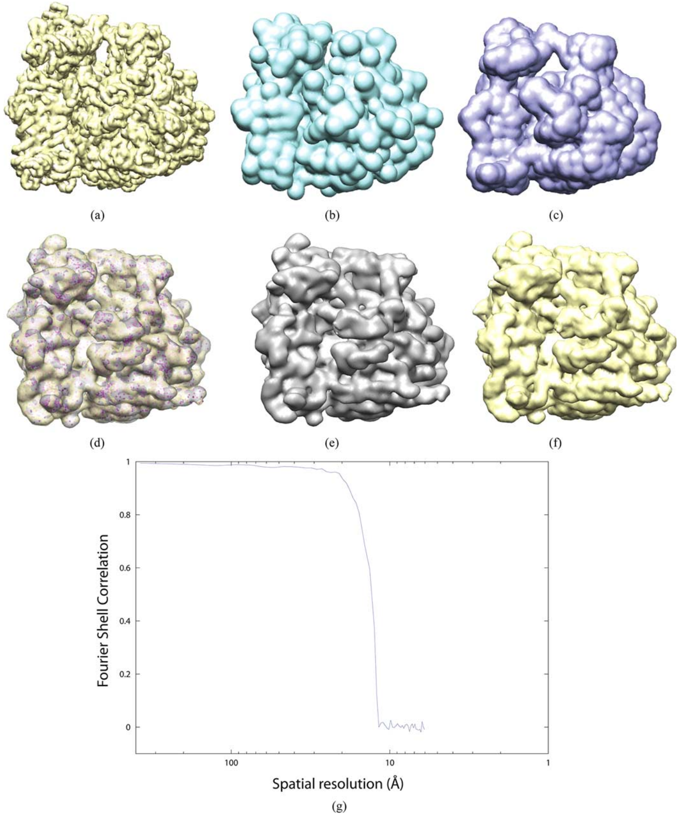


Fig. 8. Comparison of the proposed method with the *qvol* Situs method using synthetic data (a)-(c) and processing of experimental data (d)-(g) of the 70S ribosome using the proposed method. (a) Volume from the ground-truth atomic structure (also shown in Fig. 4(b)). (b) Volume from 586 pseudoatoms obtained by approximating the volume shown in Fig. 4(c) using the proposed method. (c) Volume from 600 pseudoatoms obtained by approximating the volume shown in Fig. 4(c) using *qvol*. (d) Overlap between the experimental EM volume (EMDB:EMD-5262), its pseudoatomic representation obtained by the proposed method, and the corresponding volume (EM volume: gray; approximated volume: yellow; pseudoatoms: magenta). (e) EMD-5262 volume. (f) EMD-5262 volume approximated by the proposed method. (g) Fourier shell correlation between the volumes shown in (e) and (f).

IV. Discussion and Conclusion

In this paper, we presented an algorithm to convert a three-dimensional (3D) transmission electron microscopy (EM) density volume into a granulated model by controlling the volume approximation error. The advantage of this approach is that the target approximation accuracy can be chosen to suit a particular application. For instance, normal mode analysis (NMA) of experimental noisy volumes or 3D-to-3D fitting of complexes with their subunits can be done using larger target approximation errors (e.g., $\varepsilon = 10 - 15\%$), whereas NMA of clean volumes or elastic 3D-to-2D fitting can be done using smaller target approximation errors (e.g., $\varepsilon = 1 - 5\%$). By analogy with atomic-resolution structures, the granulated model obtained by the proposed method is referred to as pseudoatomic structure. The granules are Gaussian functions and they are referred to as pseudoatoms. Given a volume, a pseudoatom size (Gaussian-function standard deviation), and a target volume approximation error, the algorithm adjusts the number of pseudoatoms, their position, and their mass (amplitudes of Gaussian functions) to obtain the desired quality of the volume approximation.

The quality of pseudoatomic structures for NMA was evaluated by analyzing the overlap of their normal modes with an experimental conformational change. Also, this overlap was compared with the one obtained between the atomic modes and the same experimental conformational change. We showed that the method allows constructing pseudoatomic structures whose normal modes are highly similar to those of atomic-resolution structures.

The method was compared with *qvol* and *vol2pdb* methods of the Situs package [31]. Our experiments have shown that *qvol* cannot handle many pseudoatoms. We showed that the quality of the volume resulting from *qvol* is lower than the one of the volume obtained by the proposed method for similar small numbers of pseudoatoms (around 600) in the two volume representations. Additionally, we showed that *vol2pdb* does not allow defining a desired quality of the volume approximation neither in terms of the desired number of pseudoatoms (contrary to Situs *qvol*) nor in terms of the desired target approximation error (contrary to the method proposed here). As it places pseudoatoms wherever the volume density is positive, the quality of the approximated volume will be virtually the same for any given positive-density volume. More precisely, the quality of the approximated volume will be around the highest one because the method produces a huge number of pseudoatoms (10^5-10^6). However, such huge numbers of pseudoatoms cannot be used in some applications (e.g., in normal modes computation where the diagonalization of the Hessian matrix would be difficult with such numbers of pseudoatoms). The proposed method can result in a volume that is similar to the one resulting from *vol2pdb* (using a low desired approximation error) but it requires at least 30 times fewer pseudoatoms to achieve such high volume quality.

The source codes of the implemented algorithms for volume-to-pseudoatoms and pseudoatoms-to-volume conversions are available in the Xmipp software [33], [34] (the Xmipp methods volume_to_pseudoatoms and volume_from_pdb, respectively). On our architecture, volume_to_pseudoatoms takes from several seconds to a few tens of minutes depending on the volume size, the target volume approximation error, and the standard deviation of the Gaussian functions. For example, for the 70S volumes used in this paper, it took between 10 seconds (around 300 pseudoatoms) and 15 minutes (around 20,000 pseudoatoms). The method volume_from_pdb is run within the method volume_to_pseudoatoms but it can also be run separately as it is additionally available as a separate Xmipp program (e.g., pseudoatomic structures obtained by another software can be converted into volumes using this method). As a separate program, volume_from_pdb takes from a few hundreds of milliseconds to a few tens of seconds on our architecture.

The volume-to-pseudoatoms conversion can additionally be performed with the user-friendly web application *3DEM Loupe* and the user-friendly graphical Xmipp interface of *HEMNMA* [13], [14]. They also allow a computation of normal modes and an interactive animation of pseudoatomic structures displaced along normal modes. Our previous papers are focused on the presentation of the web server and the HEMNMA methodology with the graphical interface. As such, they mention the volume-to-pseudoatoms conversion method as one of the tools that were used, without showing the details of the algorithm or the parameter sensitivity analysis that are exclusively described in this paper. The description of the algorithm details allows implementing the method in other software packages.

The proposed method results in pseudoatomic representations that are optimal in terms of the volume approximation error defined in (2). Additionally, by varying the Gaussian-function standard deviation, the method can reproduce atomic distances. This means that the method can give a solution that is additionally optimal in terms of the similarity with the atomic structure. However, the user does not always have access to the 3D atomic structure and, thus, cannot always check this similarity. The presented results show that the "reproduction" of atomic distances does not require knowing the 3D atomic structure but only the number of coarse-grain atoms such as Cα and phosphate atoms. Indeed, 1D amino acid sequences are available for most protein complexes and they can give the information about the number of atoms in the complex. By varying the Gaussian-function standard deviation, we can get the number of pseudoatoms that is similar to the given number of atoms. We showed that the pseudoatomic structure constructed in such way has the distance histogram that is similar to the one of the atomic structure.

Additionally, we showed that this technique allows sharpening of intermediate-resolution volumes (the standard deviation of Gaussian basis functions regulates the volume sharpness for a given target approximation error). To produce sharpening, the target approximation error and the standard deviation of Gaussian basis functions did not have to be specifically adjusted. More precisely, the FSC curves extend to higher spatial frequencies than the reference FSC curve for all settings of these two parameters that were tested here, except for the one resulting in a very small number of pseudoatoms with respect to the large size of the complex that was used in these tests (70S ribosome). Such small number of pseudoatoms (around 600) was used only for a comparison with a similar representation obtained by Situs but, otherwise, it would rarely be used with such large complexes (at least 1000–1500 pseudoatoms would be recommended in such cases). In practice, we rarely can really check the sharpening results because the ground-truth structure is often unavailable. As discussed in the previous paragraph, the 3D ground-truth structure may be unavailable but the total number of coarse-grain atoms is often available, which allows selecting the pseudoatomic representation whose number of pseudoatoms is similar to this number. We showed that the quality of the pseudatomic representation selected in this way is around the best one for the given volume and it is realistic to expect that such pseudatomic representations should produce sharpening in general.

Too small target errors and Gaussian-function standard deviations would result in pseudoatomic representations with too many details. Such settings should be avoided with noisy volumes such as those obtained by EM to avoid representing noise too accurately. The possibility to lower the volume approximation accuracy by increasing the values of these two parameters is currently being explored for EM volume denoising as a part of a separate research work.

Acknowledgment

We thank GENCI-CINES/IDRIS (France) for HPC resources [2013-x2010072174, 2014-x2010072174, 2015-x2010072174], and Dr I. Callebaut (IMPMC, Paris, France) and Dr J. M. Carazo (CNB, Madrid, Spain) for stimulating discussions.

References

[1] N. Jimenez-Lozano, M. Chagoyen, J. Cuenca-Alba, and J. M. Carazo, “FEMME database: Topologic and geometric information of macromolecules,” *J. Struct. Biol.*, vol. 144, pp. 104–113, Oct.-Nov. 2003.

[2] P. Chacon, F. Tama, and W. Wriggers, “Mega-Dalton biomolecular motion captured from electron microscopy reconstructions,” *J. Mol. Biol.*, vol. 326, pp. 485–492, Feb. 2003.

[3] D. Ming, Y. Kong, M. A. Lambert, Z. Huang, and J. Ma, “How to describe protein motion without amino acid sequence and atomic coordinates,” *Proc. Nat. Acad. Sci.*, vol. 99, pp. 8620–8625, Jun. 2002.

[4] J. Garcia de la Torre, O. Llorca, J. L. Carrascosa, and J. M. Valpuesta, “HYDROMIC: Prediction of hydrodynamic properties of rigid macromolecular structures obtained from electron microscopy images,” *Eur. Biophys. J.*, vol. 30, pp. 457–462, Oct. 2001.

[5] S. Birmanns and W. Wriggers, “Multi-resolution anchor-point registration of biomolecular assemblies and their components,” *J. Struct. Biol.*, vol. 157, pp. 271–280, Jan. 2007.

[6] W. Wriggers, R. A. Milligan, K. Schulten, and J. A. McCammon, “Self-organizing neural networks bridge the biomolecular resolution gap,” *J. Mol. Biol.*, vol. 284, pp. 1247–1254, Dec. 1998.

[7] P. A. De-Alarcon, A. Pascual-Montano, A. Gupta, and J. M. Carazo, “Modeling shape and topology of low-resolution density maps of biological macromolecules,” *Biophys. J.*, vol. 83, pp. 619–632, Aug. 2002.

[8] T. Kawabata, “Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a Gaussian mixture model,” *Biophys. J.*, vol. 95, pp. 4643–4658, Nov. 2008.

[9] O. Kurkcuoglu, R. L. Jernigan, and P. Doruker, “Mixed levels of coarse-graining of large proteins using elastic network model succeeds in extracting the slowest motions,” *Polymer*, vol. 45, pp. 649–657, 2004.

[10] F. Tama, W. Wriggers, and C. L. Brooks 3rd, “Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory,” *J. Mol. Biol.*, vol. 321, pp. 297–305, Aug. 2002.

[11] Q. Jin, C. O. Sorzano, J. M. de la Rosa-Trevin, J. R. Bilbao-Castro, R. Nunez-Ramirez, O. Llorca, F. Tama, and S. Jonic, “Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes,” *Structure*, vol. 22, pp. 496–506, Mar. 2014.

[12] B. A. Bailey, T. Schlumprecht, and N. Sivakumar, “Nonuniform sampling and recovery of multidimensional bandlimited functions by Gaussian radial-basis functions,” *J. Fourier Anal. Applicat.*, vol. 17, pp. 519–533, 2011.

[13] R. Nogales-Cadenas, S. Jonic, F. Tama, A. A. Arteni, D. Tabas-Madrid, M. Vazquez, A. Pascual-Montano, and C. O. Sorzano, “3DEM Loupe: Analysis of macromolecular dynamics using structures from electron microscopy,” *Nucleic Acids Res.*, vol. 41, pp. W363–W367, Jul. 2013.

[14] C. O. Sorzano, J. M. de la Rosa-Trevin, F. Tama, and S. Jonic, “Hybrid electron microscopy normal mode analysis graphical interface and protocol,” *J. Struct. Biol.*, vol. 188, pp. 134–141, Nov. 2014.

[15] B. Fritzke, “Growing cell structures – A self-organizing network for unsupervised and supervised learning,” *Neural Netw.*, vol. 7, pp. 1441–1460, 1994.

[16] K. Braig, Z. Otwinowski, R. Hegde, D. C. Boisvert, A. Joachimiak, A. L. Horwich, and P. B. Sigler, “The crystal structure of the bacterial chaperonin GroEL at 2.8 A,” *Nature*, vol. 371, pp. 578–586, 1994.

[17] L. M. Peng, G. Ren, S. L. Dudarev, and M. J. Whelan, “Robust parameterization of elastic and absorptive electron atomic scattering factors,” *Acta Crystallographica Sec. A*, vol. 52, pp. 257–276, 1996.

[18] C. O. S. Sorzano, J. Vargas, J. Otón, V. Abrishami, J. M. de la Rosa-Trevín, S. del Riego, A. Fernández-Alderete, C. Martínez-Rey, R. Marabini, and J. M. Carazo, “Fast and accurate conversion of atomic models into electron density maps,” *AIMS Biophysics*, vol. 2, pp. 8–20, 2015.

[19] M. Vulovic, R. B. Ravelli, L. J. van Vliet, A. J. Koster, I. Lazic, U. Lucken, H. Rullgard, O. Oktem, and B. Rieger, “Image formation modeling in cryo-electron microscopy,” *J. Struct. Biol.*, vol. 183, pp. 19–32, Jul. 2013.

[20] H. Rullgard, L. G. Ofverstedt, S. Masich, B. Daneholt, and O. Oktem, “Simulation of transmission electron microscope images of biological specimens,” *J. Microsc.*, vol. 243, pp. 234–256, Sep. 2011.

[21] M. M. Tirion, “Large amplitude elastic motions in proteins from a single-parameter, atomic analysis,” *Phys. Rev. Lett.*, vol. 77, pp. 1905–1908, Aug. 1996.

[22] F. L. Bookstein, *Morphometric Tools for Landmark Data*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[23] J. N. Stember and W. Wriggers, “Bend-twist-stretch model for coarse elastic network simulation of biomolecular motion,” *J. Chem. Phys.*, vol. 131, p. 074112, Aug. 2009.

[24] C. W. Muller, G. J. Schlauderer, J. Reinstein, and G. E. Schulz, “Adenylate kinase motions during catalysis: An energetic counterweight balancing substrate binding,” *Structure*, vol. 4, pp. 147–156, Feb. 1996.

[25] C. W. Muller and G. E. Schulz, “Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 A resolution. A model for a catalytic transition state,” *J. Mol. Biol.*, vol. 224, pp. 159–177, Mar. 1992.

[26] M. Delarue and P. Dumas, “On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models,” *Proc. Nat. Acad. Sci. USA*, vol. 101, pp. 6957–6962, May 2004.

[27] K. Suhre, J. Navaza, and Y. H. Sanejouand, “NORMA: A tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps,” *Acta Crystallogr D Biol. Crystallogr.*, vol. 62, Sep. 2006, 1098–100.

[28] F. Tama and Y. H. Sanejouand, “Conformational change of proteins arising from normal mode calculations,” *Protein Eng.*, vol. 14, pp. 1–6, Jan. 2001.

[29] Y. Wang, A. J. Rader, I. Bahar, and R. L. Jernigan, “Global ribosome motions revealed with elastic network model,” *J. Struct. Biol.*, vol. 147, pp. 302–314, Sep. 2004.

[30] W. Zhang, J. A. Dunkle, and J. H. Cate, “Structures of the ribosome in intermediate states of ratcheting,” *Science*, vol. 325, pp. 1014–1017, Aug. 2009.

[31] W. Wriggers, R. A. Milligan, and J. A. McCammon, “Situs: A package for docking crystal structures into low-resolution maps from electron microscopy,” *J. Struct. Biol.*, vol. 125, pp. 185–195, Apr.-May 1999.

[32] J. Fu, J. B. Munro, S. C. Blanchard, and J. Frank, “Cryoelectron microscopy structures of the ribosome complex in intermediate states during tRNA translocation,” *Proc. Nat. Acad. Sci. USA*, vol. 108, pp. 4817–4821, Mar. 2011.

[33] C. O. Sorzano, R. Marabini, J. Velazquez-Muriel, J. R. Bilbao-Castro, S. H. Scheres, J. M. Carazo, and A. Pascual-Montano, “XMIPP: A new generation of an open-source image processing package for electron microscopy,” *J. Struct. Biol.*, vol. 148, pp. 194–204, Nov. 2004.

[34] J. M. de la Rosa-Trevin, J. Oton, R. Marabini, A. Zaldivar, J. Vargas, J. M. Carazo, and C. O. Sorzano, “Xmipp 3.0: An improved software suite for image processing in electron microscopy,” *J. Struct. Biol.*, vol. 184, pp. 321–328, Nov. 2013.

Slavica Jonić received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Belgrade, Serbia, in 1996 and 1999, respectively, and the Ph.D. degree in image processing from the Swiss Federal Institute of Technology in Lausanne – EPFL, Switzerland, in 2003.

From 1996 to 2003, she held Research and Teaching Assistant positions at the University of Belgrade (1996–1999) and at the EPFL (2000–2003). From 2004 to 2008, she was a Research Scientist at the University Pierre and Marie Curie (UPMC) – Paris 6, France. Since 2008, she has been an Associate Scientist (“Chargé de Recherche”) at the French National Centre for Scientific Research (CNRS), and currently with the IMPMC-CNRS UMR 7590, Paris, France. Her first research activities (1996–1999) were focused on development of methods for modeling and control of walking in paraplegic subjects. They evolved towards electron-microscopy image processing during her Ph.D. at the Biomedical Imaging Group, EPFL. Her postdoctoral stay at an electron-microscopy laboratory (IMPMC-CNRS UMR 7590) motivated her current research activities that are focused on developing new image analysis methods for capturing conformational changes of imaged macromolecular complexes, which shall open new doors towards high throughput four-dimensional electron microscopy.

Carlos Óscar Sánchez Sorzano received the B.Sc. and M.Sc. in electrical engineering with two specialities (electronics and networking, Univ. Málaga), B.Sc. in computer science (Univ. Málaga), B.Sc. and M.Sc. in mathematics, (speciality in statistics, UNED), Ph.D. in biomedical engineering (Univ. Politécnica de Madrid) and Ph.D. in pharmacy (Univ. San Pablo-CEU).

He served as Secretary of the Dept. of Engineering of Electronic and Telecommunication Systems of the Univ. CEU-San Pablo (Madrid) between 2005 and 2008, Coordinator of the Section on Signal and Communications theory between 2004 and 2009, head of the Bioengineering Laboratory of that University between 2007 and 2008, Director of the Summerschool on Advanced Data Analysis and Modelling between 2006 and 2009, and Codirector of the Master on Computational Biotechnology between 2007 and 2009. He did his Ph.D. at the Biocomputing Unit of the National Center of Biotechnology (CSIC), and a post-doc at the Biomedical Imaging Group of the Swiss Federal Institute of Technology Lausanne (EPFL). In 2006, he received the Ángel Herrera research prize. He has been senior member of the IEEE since 2008 and that same year he was accredited as “profesor titular de universidad” by ANECA. In 2009, he was appointed as “Profesor Agregado” at Univ. San Pablo CEU, awarded a Ramón y Cajal research contract and appointed as technical director of the INSTRUCT Image Processing Center for Microscopy. In 2011 and 2012, he was president of the National Association of Ramón y Cajal researchers. He coordinates the service of image processing and statistical analysis of the CNB since 2011. In 2013, he was accredited as Full Professor.