Manuscript submitted to:

AIMS Biophysics

Volume 2, Issue 1, 8-20

DOI:10.3934/biophy.2015.1.8

Received February 2015, Accepted 11 March 2015, Published 18 March 2015

Research article

Fast and accurate conversion of atomic models into electron density maps

Carlos O.S. Sorzano[1,2*], Javier Vargas[1,2], Joaquín Otón[1], Vahid Abrishami[1], José M. de la Rosa-Trevín[1], Sandra del Riego[2], Alejandro Fernández-Alderete,[2], Carlos Martínez-Rey[2], Roberto Marabini[3], José M. Carazo[1]

[1] National Center of Biotechnology (CSIC), c/Darwin, 3, Campus Univ. Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain

[2] Bioengineering Lab., Escuela Politécnica Superior, Univ. San Pablo CEU, Campus Urb. Montepríncipe s/n, 28668, Boadilla del Monte, Madrid, Spain

[3] Escuela Politécnica Superior, Univ. Autónoma de Madrid, Campus Univ. Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain

[*] **Correspondence:** coss@cnb.csic.es; Tel: +34-91-585-4510; Fax: +34-91-585-4506

Abstract: New image processing methodologies and algorithms have greatly contributed to the significant progress in three-dimensional electron microscopy (3DEM) of biological complexes we have seen over the last decades. Naturally, the availability of accurate procedures for the objective testing of new algorithms is a crucial requirement for the further advancement of the field. A good and accepted testing workflow involves the generation of realistic 3DEM-like maps of biological macromolecules from which some measure of "ground truth" can be derived, ideally because their 3D atomic structure is already known. In this work we propose a very accurate generation of maps using atomic form factors for electron scattering. We thoroughly review current approaches in the field, quantitatively demonstrating the benefits of the new methodology. Additionally, we study a concrete example of the use of this approach for hypothesis testing in 3D Electron Microscopy.

Keywords: Atom models; Electron scattering; Filter design; Image formation; 3D signals; Electron microscopy

1. Introduction

Knowledge of the 3D structure of macromolecules and their complexes is crucial for understanding their biological function. Three-dimensional Electron Microscopy (3DEM) addresses the problem of determining the structure of macromolecular complexes from projection images recorded by transmission electron microscopy [9, 29, 15].

The resolution levels routinely achieved with this approach have improved significantly over time, and have now reached the level of 3Å [47, 11] (the corresponding spatial frequency is $1/3$ Å^{-1}). At the level of image processing, these continuous improvements in resolution are due, among other reasons, to: increases in the number of particles processed; improvements at the level of automation of the data-collection instrument [34]; the introduction of new electron detectors; and, in general, the design of new algorithms extracting more information from each available micrograph [9, 29]. One of the best ways of testing the strengths and limitations of such new algorithms is to first simulate the computational experiments using artificially-generated test data (also known as "phantom" data). Indeed, 3D model structures can be projected into 2D images and can then be used, for example, to generate artificial micrographs. The accuracy of these simulations, which should closely resemble experimental situations, is becoming critical as resolution is improved.

3D model structures can consist of simple geometrical forms like spheres, cylinders and other geometrical features [2, 31], but this may have the disadvantage of generating data which have an unrealistic amplitude spectrum distribution. It can thus be advantageous to use randomly generated noise with a precisely defined spectral behavior [12]. In general, the most realistic model structures for single particle 3DEM are simulated

biological complexes based on the atomic coordinates of known structures as deposited in the Protein Data Bank (PDB or wwPDB).

The Protein Data Bank (PDB, part of the world-wide: wwPDB) is the standard repository for the atomic structure of macromolecules [1]. A PDB entry is formed of a list of atoms and their coordinates in space (see Fig. 1). Thus, a volume can be created simply by assigning different densities to the different kinds of atoms. Many standard 3DEM software packages (including Spider [10], EMAN [21], Situs [46], IMAGIC [42], Xmipp [30], and Bsoft [14]) can be used to perform this conversion from atomic PDB coordinates to a 3D density model. However, as we will show in this work, the atom shapes traditionally used may be too simple, leading to problems with the generated models. For example, using a simple 3D hard sphere to model the atoms of a 3D structure (see Fig. 1, the common choice of visualization programs. *e.g.*, Chimera or Jmol) leads to a Fourier transform of the 3D model structure with artificial zero-power rings. The most standard choice, a single Gaussian per atom [46, 45], does not reproduce well atom shapes in Fourier space (as shown below). The more complex model based on a sum of Gaussians defined in Fourier space [44] does correlate well (as shown below) to EM structures, but its implementation in Fourier space is extremely slow.

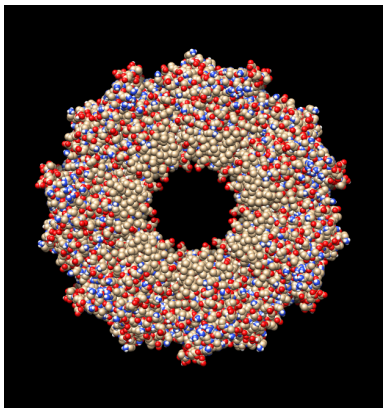

Figure 1. Example of the atomic structure of the GroEL chaperonin (PDB entry: 1GRL). For display purposes, most atomic visualizers substitute each atom by a sphere whose color depends on the kind of atom.

Considering the need to increase the resemblance between high-resolution cryoEM images and computer generated models, in this work we promote the use of PDB along with more elaborated functions describing individual atoms such as Electron Atomic Scattering Factors (EASF) [28], namely, the actual shape of the atoms seen by the electrons traveling down the microscope column. EASFs have been studied extensively in physics [28, 23] and have been measured experimentally [32]. A review of available experimental data, along with other physical properties related to X-ray and neutron scattering of the corresponding atoms, is presented in [4]. This experimental information can be approximated analytically as a sum of a collection of Gaussian functions [24] and we here propose to use these sums to replace the simpler atomic models currently in use in 3DEM. Additionally, there are a number of applications, like fitting [37, 39, 38] that may also need to convert atomic structures into density volumes. As we show in this article, EASFs are much better suited for this task than any other function currently available. EASFs have already been used in biological EM applications. In particular, Bsoft [14], TEM simulator [26, 43] and RSref [5] have made use of EASFs. But, as we will show in the Results section, these implementations are either too slow, their accuracy can be improved, or they do not return the converted volume.

In material science, where samples are usually less sensitive to damage by the electron beam than biological samples, high resolution electron microscopy (HREM) studies can achieve the visualization of individual atoms. Indeed, over the last decade EASFs have been used extensively in this field to help interpret the (relatively) noise free images so obtained [16, 17] and special software is now available for the specific needs of this field [33]. However, the specificities of macromolecular samples (large number of atoms and not in a regular arrangement) makes these programs inapplicable to structural biology.

In this article we present a method for generating three-dimensional volumes from atomic coordinates obtained by X-ray crystallography, Nuclear Magnetic Resonance, or computational modeling, as well as their two-dimensional projections resembling those obtained in an electron microscope. Contrary to the standard approach in the field (that works in Fourier space), the proposed method works in real space with very low memory

requirements since the need of upsampling to have accurate results has been eliminated. Additionally, we present an example in which the design of new image processing algorithms would have benefited from this development by providing an accurate hypothesis testing tool.

2. Materials and methods

2.1. Generation of volumes from atomic models

Most programs in 3DEM represent volumes as a collection of samples of some underlying continuous function. Samples are taken with a fixed spacing that is the sampling rate and is measured in Å/pixel. Sampling an atomic structure means sampling a continuous function with a support that has a diameter in the order of twice the atomic radius (the empirical atomic radius ranges from 0.25Å for the hydrogen atom, to 1.40Å for the iron atom). The sampling frequency must be carefully chosen so that so small functions are appropriately sampled, avoiding aliasing. The resulting volume could, later on, be downsampled to the desired final sampling rate.

2.1.1. Atomic model

The EASF of an atom is related to the scattering of electrons by matter [23]. Electrons may interact elastically (deviated from their trajectory without any loss of energy) or inelastically (part of the energy is transferred to the interacting atom). Both kinds of interactions can be modeled to a good-extent using the complex optical potential [23]. In our model, only the elastic component of the interaction is considered, since electrons that interact inelastically are supposed to be filtered out by electron optical components [9]. It must be pointed out that the interaction of an electron with an atom is related to the atomic number more than to the atomic weight of that atom, since neutrons do not contribute to this interaction.

It has been shown that the elastic scattering of electrons can be accurately approximated by a sum of Gaussians up to a frequency of 6 $Å^{-1}$ [24]:

$$f_e(R) = \sum_{i=1}^{n} a_i \exp(-b_i R^2), \tag{1}$$

where $f_e(R)$ is the electron scattering factor of a given atom at the spatial frequency R (in $Å^{-1}$). Parameters a_i and b_i are specific for each atom. Five Gaussians are employed by Peng and coworkers to faithfully represent the experimentally collected data, and the specific values of a_i and b_i for each atom are tabulated in [24]. Having the expression of the EASF in Fourier space automatically allows us to derive the corresponding expression of the EASF in real space, by Fourier inversion in the continuous variable R:

$$f_e(r) = \sum_{i=1}^{n} 2\pi a_i \sqrt{\frac{\pi}{b_i}} \exp(-\frac{\pi^2 r^2}{b_i}), \tag{2}$$

Fig. 2 shows the EASF of some of the most common atoms present in biological samples in Fourier space as well as in real space. This figure presents three interesting properties of the EASFs: first, the shape of the EASF in real space is not so similar to the shape of a Gaussian (note the peak at the origin $r = 0$, although, by construction, it is a differentiable peak); second, the differences among the shape of the different atoms cannot be modeled by a simple multiplicative factor (EASFs are not parallel to each other in Fourier space), as is usually done when atoms are modeled by a single Gaussian [45]; third, an atom can interact elastically with an electron far beyond the atomic radius (for instance, the atomic radius of iron is 1.40Å while its EASF is significantly different from zero up to a radius of 5Å).

EASFs can be compared in Fourier space to other models that have been used for the simulation of atoms in 3DEM: the step function (a solid sphere), the Gaussian, the modified Kaiser-Bessel window, and the blob. The formula for the step function is

$$f_{step}(r) = u(r + r_0) - u(r - r_0), \tag{3}$$

where $u(r)$ is the Heaviside step function and r_0 is the atomic radius. The formula for the Gaussian is

$$f_{Gaussian}(r) = \exp\left(-\frac{r^2}{\left(\frac{r_0}{3}\right)^2}\right). \tag{4}$$

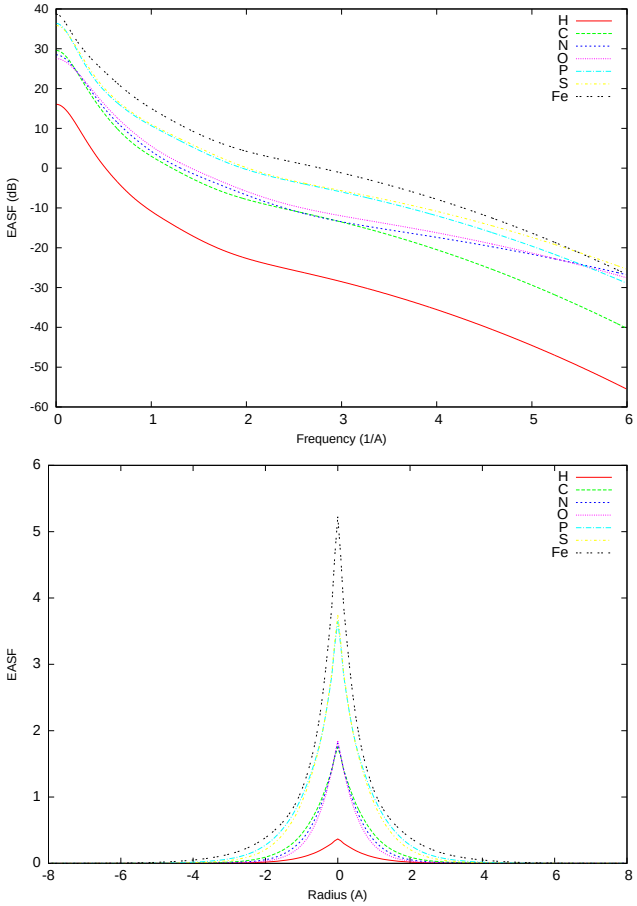

Figure 2. Top: EASFs in Fourier space for several atoms. Bottom: Corresponding EASFs in real space.

The expression for the modified Kaiser-Bessel window is

$$f_{Kaiser}(r) = \frac{I_0\left(\beta\sqrt{1-\left(\frac{r}{r_0}\right)}\right)}{I_0(\beta)}. \tag{5}$$

Finally, the expression for the blob [19] is

$$f_{blob}(r) = \frac{\left(\sqrt{1-\left(\frac{r}{r_0}\right)}\right)^m I_m\left(\beta\sqrt{1-\left(\frac{r}{r_0}\right)}\right)}{I_m(\beta)}. \tag{6}$$

In the previous equations, I_m is the modified Bessel function of the first kind and order m (I_0 is the particular case in which $m = 0$) and β is a parameter controlling the atom shape. Fig. 3 shows the spectra of all these functions (they have been normalized in frequency so that all of them have the same amplitude at frequency $R = 0$). We presume the EASF to be a more reliable estimate of what actually happens within the microscope since it has been experimentally measured, while the other functions are some smooth functions, just convenient from a signal processing or implementation point of view. The modified Kaiser-Bessel and the blob models have been computed for $\beta = 3.6$ and $m = 2$ which are the standard values implemented in Xmipp. Remarkably, the most widely used function, the Gaussian, is a specially bad model both at low and high frequency. The zoomed plot of the different spectra corresponds to a final sampling rate of 1.5Å/pixel (a rather common sampling rate in 3DEM). It can be observed, for the carbon atom, that all the models introduce unnecessary extra energy at high frequencies with respect to the EASF.

2.1.2. Sampling+Downsampling

Since EASFs are represented by a sum of Gaussians, their power spectra never vanish and, hence, EASFs are not band-limited signals. However, it can be easily checked that, at a frequency of 6Å^{-1}, most of the spectra of the EASFs of biologically relevant atoms are more than 60dB below their maximum value (achieved at frequency $R = 0$, see Fig. 2). Therefore, we can assume that at a sampling frequency of 12Å^{-1}, the aliasing is negligible. In any case, this is the maximum sampling frequency usable with our atomic model, since the EASF approximation provided in [24] is only valid up to 6Å^{-1}.

In conclusion, to sample the atomic model of a macromolecular structure it suffices to substitute every atom in the list of atomic coordinates provided by PDB by the EASF corresponding to every type of atom and, then, sample this new function:

$$f_{model}(\mathbf{r}) = \sum_i f_{ei}(\|\mathbf{r} - \mathbf{r}_i\|), \tag{7}$$

where f_{model} is the continuous function representing the atomic model of the macromolecule, $\mathbf{r} \in \mathbb{R}^3$ is the coordinate at which the volume is being evaluated, $\mathbf{r}_i$ is the center of the i-th atom as provided by PDB, and f_{ei} is the EASF of the i-th atom.

Once the model has been finely sampled, it can be safely resampled to any sampling rate that is a rational multiple of the fine sampling rate applying standard low-pass filters, together with the stretching/decimation operators [22].

A typical macromolecular complex has a diameter of around 200Å. Thus, the sampled volume needs to be stored in a volume of at least 2400x2400x2400 (2400=200x12) voxels. If double precision numbers are used (8 bytes/sample), the memory needed to store this volume is about 103 GB. As this amount of memory cannot be currently efficiently handled, the standard approach to downsampling cannot be implemented. Alternatively, in the following section we propose to combine the sampling, low-pass filtration and decimation in one single step, so that the aliasing at the final sampling rate is minimized and there is not such a formidable memory requirement.

2.1.3. Sampling directly at a low sampling rate

We propose a low-pass filtered version of the EASF, $\tilde{f}_e$ (LEASF), to represent each atom. This results in a new continuous function:

$$\tilde{f}_{model}(\mathbf{r}) = \sum_i \tilde{f}_{ei}(\|\mathbf{r} - \mathbf{r}_i\|). \tag{8}$$

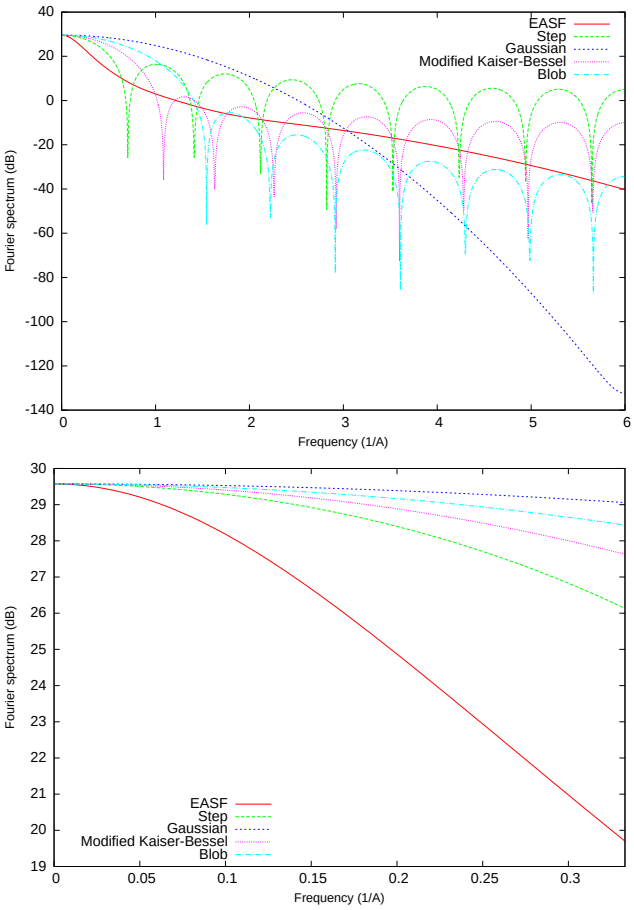

Figure 3. Fourier spectra of different atom models proposed in the literature to simulate atoms in electron microscopy. The EASF corresponds to that of the carbon atom. The lower plot is a zoomed version of the top plot up to a frequency of 0.33 $Å^{-1}$, corresponding to a standard sampling rate in 3DEM of 1.5Å/pixel. The region corresponding to the zoomed plot at the bottom is marked by a rectangle at the top figure.

Since the new function $\tilde{f}_e$ is effectively band-limited, it can be safely sampled at the final sampling rate without any significant aliasing. We discuss in this section how to generate such a low-pass filtered EASF.

The EASF can be convolved at the finely sampled space by the impulse response of a low-pass filter designed by windowing a sinc function with a Kaiser-Bessel window [22]. Although the expression of the low-pass filter is widely known, we reproduce it here for completeness. If $h_{LPF}[n]$ is the impulse response of the sought-after low-pass filter, then

$$h_{LPF}[n] = \frac{\omega_c}{\pi}\text{sinc}\left(\frac{\omega_c}{\pi}n\right)\frac{I_0\left(\beta\sqrt{1-\left(\frac{n}{M}\right)}\right)}{I_0(\beta)}, \tag{9}$$

where ω_c is the discrete cut-off frequency of the filter (normalized so that the maximum frequency is π), the total length of the filter is $2M+1$ (M is calculated below), and β is a parameter that is calculated from the width of the transition band $\Delta\omega$ as follows: first, an intermediate parameter A is calculated

$$A = -20\log_{10}(\delta),$$

where δ is the ripple allowed in the resulting low-pass filter, and then

$$M = \left\lceil \frac{A-8}{2.285\Delta\omega} \right\rceil \tag{10}$$

and

$$\beta = \begin{cases} 0.1102(A-8.7) & \text{if } 50 \leq A \\ 0.5842(A-21)^{0.4} + 0.07886(A-21) & \text{if } 21 \leq A < 50 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Thus, it can be observed that the design parameters for the low-pass filter are ω_c, $\Delta\omega$ and δ.

If the EASF is sampled at the fine sampling rate, it results in the discrete sequence $f_e[n]$. This sequence is convolved with the low-pass filter $h_{LPF}[n]$ to produce $\tilde{f}_e[n]$, i.e., the samples of the proposed continuous LEASF. The continuous function $\tilde{f}_e(r)$ is then interpolated using cubic B-splines [40, 41].

For every atom we search for the low-pass filter (i.e., for the parameters ω_c, $\Delta\omega$ and δ) that minimizes the quadratic error between the LEASF and the EASF within the frequency interval $\left[0, \frac{1}{2T}\right]$, being T the final sampling rate of the volume. Fig. 4 shows the EASF and LEASF for the carbon atom. Note the close agreement between the carbon EASF and the corresponding LEASF in the range [0,0.33] (this example was computed assuming a final sampling rate of 1.5Å/pixel). Fig. 5 shows the EASF and LEASF of the carbon atom in real space.

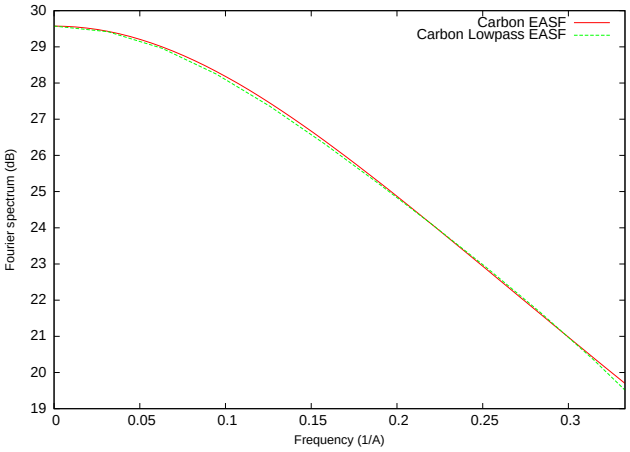

Figure 4. Fourier spectrum of the carbon EASF up to a frequency of 0.33Å^{-1} along with its corresponding LEASF. Note the perfect agreement up to a sampling frequency of 1.5 Å/pixel.

Eq. 8, along with the different LEASFs for every kind of atom, allows us to sample an atomic model directly at the final sampling rate without having to use an intermediate step with a finely sampled volume. In this way,

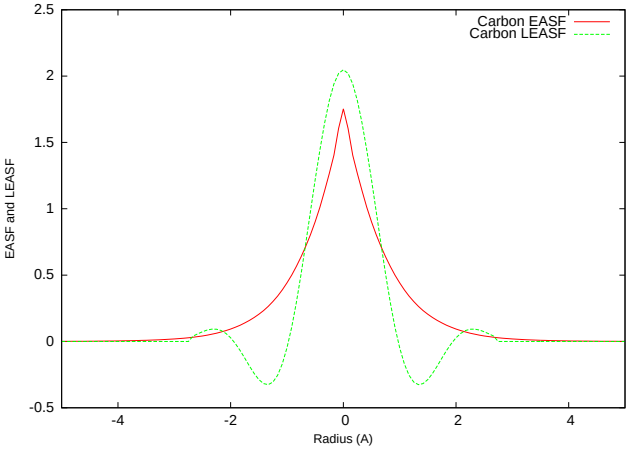

Figure 5. EASF and LEASF (T=1.5 Å/pixel) of the carbon atom.

we have tackled the problem of the huge amount of memory (and time) needed to process that finely sampled volume.

Note that the parameter of a single Gaussian model could also have been optimized to maximally resemble the EASF profile. However, this is not the standard approach in the field and we have not performed this exercise since the proposed LEASF approximation is a better approximation of the truly underlying electron atomic scattering factors.

3. Results and Discussion

3.1. Generation of volumes

To show the applicability of our atomic models to the generation of phantoms, we have chosen the chaperonin GroEL [3] (PDB accession code: 1GRL; see Fig. 1). However, similar results and conclusions have been derived from other structures (not shown in this paper for space constraints). We generated the corresponding voxel volume at a sampling rate of 1Å/pixel from the PDB by using the most standard software tools available:

- Situs [46, 45] (`pdb2vol`): this program projects the atomic structure on a cubic lattice by trilinear interpolation. Then, it convolves the lattice points with a Gaussian kernel whose width is internally calculated to achieve a given resolution (in our case 2Å).
- Spider [10] (`cp from pdb`): this program uses a similar algorithm to the previously described employing trilinear interpolation.
- EMAN [21] (`pdb2mrc`): this program uses a Gaussian in real-space that is sampled at the appropriate lattice locations.
- Bsoft [14] (`bsf`): this program follows the standard method in X-ray crystallography as programmed in the CCP4 [6] program `sfall` and reimplemented for EM in Bsoft [14]. The program follows a different set of Gaussians [44] (different from the ones proposed in this paper) and it performs all calculations in Fourier space instead of real space (as proposed in this paper).
- Xmipp [30] (`xmipp_volume_from_pdb`): this program implements LEASFs as proposed in this paper.

Fig. 6 shows the radial average of the 3D Power Spectrum Density of each one of the models. We see that Situs and Bsoft produced almost identical results (although Situs took a 12 seconds to produce this result, while Bsoft took more than 16 hours; the experiment was run on a single Intel Xeon 2.66 GHz with 2 Gb of memory). Xmipp LEASFs was executed in 5 seconds (twice faster than Situs and more than 11,000 times faster than Bsoft; another speed comparison is reported below with a wider set of structures). Situs, Bsoft and LEASFs coincide up to a resolution about 20Å. Beyond this point LEASFs power density is much smaller compared to the other two programs. Spider produces a structure whose power spectrum is above the previous ones, and EMAN has the largest amount of energy of all. This may be pointing out that simple models may be overemphasizing the

importance of high frequency components in atomic structures. Note that in EM, many times volumes are scaled at the end of the process to match a certain intensity profile in Fourier space. Although this practice would homogenize all packages regarding amplitudes, Fig. 6 highlights the fact that the underlying functions used by the different approaches are introducing different distortions in Fourier space that show up in amplitude (as shown in the figure) as well as in phase (not shown).

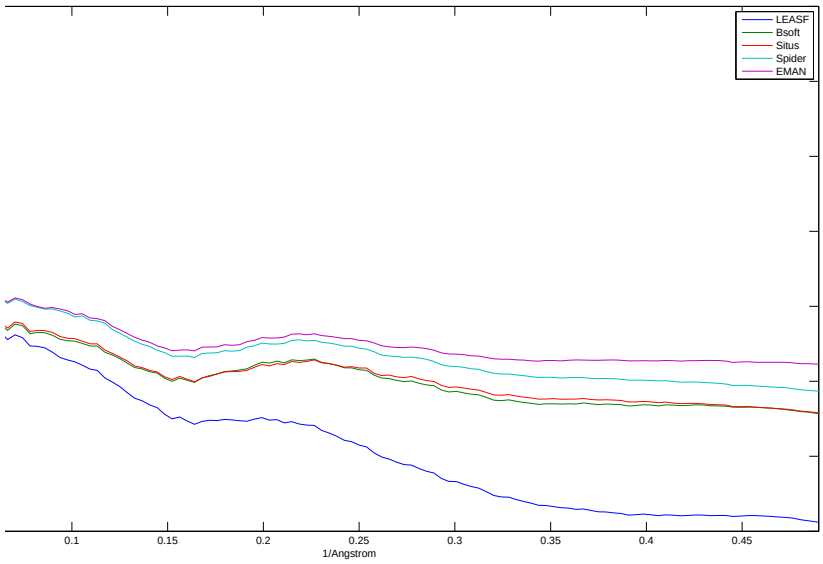

Figure 6. Radial average of the 3D Power Spectrum Density (in decibels) of different volume generation strategies.

To test whether the proposed simulation is more faithful to the EM volume than the standard Gaussian approach, we downloaded all volumes from the EMDB (Electron Microscopy Data Bank, http://www.emdatabank.org) [18] that were fully modeled with a single PDB file (Protein Data Bank, http://www.rcsb.org) [25], a total of 27 (on Feb. 15th, 2012). We measured the correlation of the EM reconstructed volume and the simulated atomic volume. We compared the correlation of the volumes simulated with the proposed method and with Gaussians+trilinear downsampling. The hypothesis that both methods were equally performing was rejected with a p-value of 0.019, that is, the correlation of the volumes generated with the new method was significantly larger (in a statistical sense) than the correlation of the volumes generated by a widely used method (the correlation with the new method was 2.6% larger). The execution time of the new method was twice faster, and the memory requirements of the new method are much smaller since the volume needs not to be sampled first at a high-resolution that is later downsampled.

We performed the same experiment with TEM Simulator 1.3. The correlation of the new method was on average 6.2% larger than the one of TEM Simulator. The difference was significant with a p-value of 0.048. Despite the larger difference in correlation, the larger p-value is due to a larger variance of the TEM Simulator results. TEM Simulator produced an empty volume for 4 of the 27 structures without any error message.

Finally, we repeated the previous experiment comparing the newly proposed method to the results from the standard method in X-ray crystallography. `bsf` failed to produce an appropriate volume for 4 out of the 27 volumes (either it produced an empty volume or it produced a volume with strange strip patterns and no recognizable molecule). The correlation of the remaining volumes with the EMDB volume was 56% larger using the new method compared to the correlation using the standard crystallographic method (with a p-value of 0, i.e., the `bsf` correlation was systematically lower than the one of the newly proposed method). `bsf` took 8 days to produce the 27 volumes while the proposed method took only 3 minutes (about 4,000 times faster). `bsf` allows to define any set of weights for describing the atoms. We exploited this option and we substituted the original weights of `bsf` with the weights proposed in this article. `bsf` still failed to produce some molecules (taking the

same amount of time as the one reported in the previous experiment). However, for the remaining structures, the correlation of the proposed method and `bsf` with the EMDB structure was not, as expected, significantly different (confidence level of 95%).

Note that the main difference between the proposed method and the standard crystallographic method, as implemented in `bsf` and `sfall`, is that the newly proposed method is implemented in real space without any downsampling, while the standard crystallographic method is implemented in Fourier space. There is a performance difference of more than 3 orders of magnitude in speed.

3.2. Application to Bayesian models in single particle analysis

In this section we show how the model developed above can be used to check statistical distributions such as the one put forward in [27] in which Fourier coefficients of the reconstructed volumes were supposed to be independent, identically distributed random variables with a Gaussian distribution where the real and imaginary parts are independent and have zero mean. For doing so, we analyzed the Fourier transform of 500 atomic structures randomly selected from the Protein Data Bank [1]. Our simulation framework provides us with a very accurate way to empirically estimate the statistical distribution of 3DEM map properties, like their Fourier coefficients.

We computed the Fourier transform of the 500 structures as the sum of the Fourier transforms of each of their atoms modeled with LEASFs. The Fourier transform of the LEASF is a continuous function that we sampled every 0.01Å^{-1} in the frequency range between 0 and 0.5Å^{-1} (maximum resolution 2Å). For each possible pair of Fourier coefficients, V_{l_1} and V_{l_2}, we tried to fit a linear regression model of the real part of V_{l_1} (analogously, the imaginary part) on the real and imaginary parts of V_{l_2}, i.e.

$$\text{Re}\{V_{l_1}\} = b_0 + b_1\text{Re}\{V_{l_2}\} + b_2\text{Im}\{V_{l_2}\}.$$

If the independence assumption is true, then the b_1 and b_2 coefficients should not be significantly different from 0. We tested these hypothesis using the standard confidence intervals for linear regression coefficients [8] with a 95% confidence. At least one of the hypothesis ($b_1 = 0$ or $b_2 = 0$) was rejected in 83.13% of the cases, i.e., the vast majority of the macromolecule spectrum has a statistically significant linear dependence with other Fourier coefficients.

We also studied the statistical distribution of the real and imaginary parts of the Fourier coefficients corresponding to the aforementioned 500 atomic structures. If each complex coefficient follows a bivariate Gaussian distribution, then its marginal distributions (the ones of the real and imaginary parts) should be univariate Gaussians (we used Lilliefors' normality test, [20]). The hypothesis that either the real part of the Fourier coefficient or its imaginary part are univariate Gaussians was rejected in 86% of the cases with 95% of confidence. Consequently, Fourier coefficients cannot be Gaussianly distributed.

We used a Wilcoxon signed rank test (non-parametric test, [7]) to test the hypothesis that the real and imaginary parts of the Fourier coefficients were actually centered at 0. The hypothesis was rejected in 54% of the cases with a 95% of confidence (interestingly, the hypothesis was rejected in 100% of the cases up to a resolution of 3.3Å).

Next, we tested the assumption that the real and imaginary parts of the Fourier coefficients have the same variance. We did this by computing a bootstrap estimate (1000 resamplings) of the ratio between the variance of the real part over the variance of the imaginary part (normally this ratio is distributed as a Snedecor-F, but the lack of normality prevents the use of this distribution). The variance of the real part was 1.9 times larger than the variance of the imaginary part, and the confidence interval with a level of confidence of 95% was between 1.3 and 3.3.

Finally, we tested the assumption that the real part and the imaginary parts of the Fourier coefficients are uncorrelated (this is implied by the zeros of the covariance matrix Σ_l). As we did before, we tested whether the real part had a significant linear dependence on the imaginary part or vice-versa. The hypothesis that both parts are independent was rejected in 56% of the cases with a 95% confidence (the hypothesis was rejected in 100% of the cases up to 3.7Å).

3.3. Discussion

In this paper we have introduced a new basis function (LEASF) for the conversion of atomic models into density volumes based on Electron Atomic Scattering Factors (EASF, an accurate representation of the electron behavior

inside the electron microscope in the presence of atoms). The EASF in Fourier space was already known and partly used by the EM community. However, our real-space counterpart allows a much faster implementation. We have compared the other models to the newly introduced basis. We have shown that previous methods tend to overemphasize high-frequency components with respect to low-frequency ones, this is specially true for EMAN and Spider. An exhaustive statistical analysis using both the PDB and EMDB databases has shown that our method outperforms currently used approaches in the field. Additionally, based on our new approach, we have derived more accurate ways to compute very realistic projection images.

The fact that the volumes simulated with the newly proposed basis functions correlate better with experimental electron microscopy volumes is an important issue that opens the door to all tasks in which converting from atoms to voxels is needed, like fitting [37, 39, 38] or normal mode analysis [36, 35]. In fact, recently [13] it has been proposed that one of the ways to validate EM volumes is to correlate the EM reconstructed volume with the simulated atomic structure of those domains that are known at atomic resolution. Additionally, we have used the very accurate EASF-based modeling of EM-like density maps for PDB files to check the consistency of some statistical assumptions underlying a recently introduced method in the field, specifically, a Maximum *a posteriori* method used for 3D classification. The goal was to see if by using the information already existing in structural databases along with the the new LEASF approach we could check the validity of a number of statistical assumptions. The possibility of accurately converting atomic models into density volumes allows to estimate the empirical distribution of Fourier coefficients and their joint probability density functions without the need to use closed-form *a priori* distributions such as the Gaussian.

4. Conclusions

We have here presented a strategy for generating reliable models of macromolecules for 3DEM that can be used for generating very realistic two-dimensional projections as well as 3D models. The availability of good models is important for developing new algorithms and for testing image analysis procedures prior to applying them to real data as well as for molecular fitting purposes. With the methodology described in this paper, excellent three-dimensional model structures can be generated based on the sampling of the low-pass electron form factors (LEASF) for electron scattering. We have introduced an efficient way of sampling these functions at any desired sampling rate such that aliasing effects are exploited in order to faithfully reproduce the original EASF up to half the sampling frequency ("Nyquist frequency"). All operations are very fast, ranging from 2 to 11,000 times faster than currently used approaches. Additionally, memory requirements are much smaller since there is no need to explicit downsampling. The developed methodologies are now available in the Xmipp software system (http://xmipp.cnb.csic.es) [30].

Acknowledgments

The authors would like to acknowledge economical support from the Spanish Ministry of Economy and Competitiveness through Grants AIC-A-2011-0638, BIO2010-16566, CAM(S2010/BMD-2305), as well as the Instituto Nacional de BioinformÃątica (a project of Instituto de Salud Carlos III), and NSF Grant 1114901. Postdoctoral Juan de la Cierva Grants with references JCI-2011-10185 and JCI-2010- 07594 are also acknowledged. C.O.S. Sorzano is recipient of a RamÃşn y Cajal fellowship. This work was partially funded by Instruct, part of the European Strategy Forum on Research Infrastructures (ESFRI) and supported by national member subscriptions.

Conflict of interests

The authors declare no conflict of interests.

REFERENCES

1. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov and P. Bourne, The protein data bank, *Nucleic Acids Research*, **28** (2000), 235–242.
2. J. R. Bilbao-Castro, C. O. S. Sorzano, I. García and J. J. Fernández, Phan3D: design of biological phantoms in 3D electron microscopy, *Bioinformatics*, **20** (2004), 3286–3288.

3. 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 2.8Å, *Nature*, **371** (1994), 578–86.

4. C. T. Chantler, Detailed tabulation of atomic form factors, photoelectric absorption and scattering cross section, and mass attenuation coefficients in the vicinity of absorption edges in the soft x-rays (z=30-36, z=60-89, e=0.1kev-10kev), addressing convergence issues of earlier work, *J. Phys. Chem. Ref. Data*, **29** (2000), 597–1048.

5. M. S. Chapman, A. Trzynka and B. K. Chapman, Atomic modeling of cryo-electron microscopy reconstructions - joint refinement of model and imaging parameters., *J. Structural Biology*, **182** (2013), 10–21, URL http://dx.doi.org/10.1016/j.jsb.2013.01.003.

6. Collaborative computational project no. 4, The CCP4 Suite: Programs for Protein Crystallography, *Acta Crystallographica*, **D50** (1994), 760–763.

7. H. A. David and H. N. Nagaraja, *Order statistics*, John Wiley and Sons, 2003.

8. W. R. Dillon and M. Goldstein, *Multivariate analysis: Methods and applications*, John Wiley, New York, USA, 1984.

9. J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*, Oxford Univ. Press, New York, USA, 2006.

10. J. Frank, M. Radermacher, P. Penczek, J. Zhu, Y. Li, M. Ladjadj and A. Leith, SPIDER and WEB: Processing and visualization of images in 3D electron microscopy and related fields., *J. Structural Biology*, **116** (1996), 190–9.

11. P. Ge and Z. H. Zhou, Hydrogen-bonding networks and rna bases revealed by cryo electron microscopy suggest a triggering mechanism for calcium switches., *Proc. Natl. Acad. Sci. USA*, **108** (2011), 9637–9642.

12. G. Harauz and M. van Heel, Exact filters for general geometry three dimensional reconstruction, *Optik*, **73** (1986), 146–156.

13. R. Henderson, A. Sali, M. L. Baker, B. Carragher, B. Devkota, K. H. Downing, E. H. Egelman, Z. Feng, J. Frank, N. Grigorieff, W. Jiang, S. J. Ludtke, O. Medalia, P. A. Penczek, P. B. Rosenthal, M. G. Rossmann, M. F. Schmid, G. F. SchrÃűder, A. C. Steven, D. L. Stokes, J. D. Westbrook, W. Wriggers, H. Yang, J. Young, H. M. Berman, W. Chiu, G. J. Kleywegt and C. L. Lawson, Outcome of the first electron microscopy validation task force meeting., *Structure*, **20** (2012), 205–214.

14. B. Heymann and D. Belnap, Bsoft: Image processing and molecular modeling for electron microscopy, *J. Structural Biology*, **157** (2007), 3–18.

15. S. Jonic, C. O. S. Sorzano and N. Boisset, Comparison of single-particle analysis and electron tomography approaches: an overview, *J. Microscopy*, **232** (2008), 562–579.

16. D. C. Joy, *Monte Carlo Modeling for Electron Microscopy and Microanalysis*, Oxford Univ. Press, London, England, 1995.

17. E. Kirkland, *Advanced computing in electron microscopy*, Plenum press, New York, USA, 1998.

18. C. L. Lawson, M. L. Baker, C. Best, C. Bi, M. Dougherty, P. Feng, G. van Ginkel, B. Devkota, I. Lagerstedt, S. J. Ludtke, R. H. Newman, T. J. Oldfield, I. Rees, G. Sahni, R. Sala, S. Velankar, J. Warren, J. D. Westbrook, K. Henrick, G. J. Kleywegt, H. M. Berman and W. Chiu, Emdatabank.org: unified data resource for cryoem., *Nucleic Acids Res*, **39** (2011), D456–D464.

19. R. M. Lewitt, Alternatives to voxels for image representation in iterative reconstruction algorithms, *Physics in Medicine & Biology*, **37** (1992), 705–716.

20. H. Lilliefors, On the kolmogorov-smirnov test for normality with mean and variance unknown, *J. American Statistical Association*, **62** (1967), 399–402.

21. S. J. Ludtke, P. R. Baldwin and W. Chiu, EMAN: Semiautomated software for high-resolution single-particle reconstructions, *J. Structural Biology*, **128** (1999), 82–97.

22. A. Oppenheim, R. Schafer and J. Buck, *Discrete-time signal processing*, 2nd edition, Prentice-Hall, 1999.

23. L. M. Peng, Electron atomic scattering factors, debye-waller factors and the optical potential for high-energy electron diffraction, *J. Electron Microscopy*, **54** (2005), 199–207.

24. L. M. Peng, G. Ren, S. L. Dudarev and M. J. Whelan, Robust parameterization of elastic and absorptive electron atomic scattering factors, *Acta Crystallographica*, **A52** (1996), 257–276.

25. P. W. Rose, B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, J. D. Westbrook, J. Young, B. Yukich, C. Zardecki, H. M. Berman and P. E. Bourne, The rcsb protein data bank: redesigned web site and web services., *Nucleic Acids Res*, **39** (2011), D392–D401.

26. H. Rullgård, L.-G. Ofverstedt, S. Masich, B. Daneholt and O. Oktem, Simulation of transmission electron microscope images of biological specimens., *J Microsc*, **243** (2011), 234–256, URL http://dx.doi.org/10.1111/j.1365-2818.2011.03497.x.

27. S. H. W. Scheres, A Bayesian view on cryo-EM structure determination., *J. Molecular Biology*, **415** (2012), 406–418.

28. G. H. Smith and R. E. Burge, The analytical representation of atomic scattering amplitudes for electrons, *Acta Crystallographica*, **15** (1962), 182–186.

29. C. O. S. Sorzano, S. Jonic, M. Cottevieille, E. Larquet, N. Boisset and S. Marco, 3D electron microscopy of biological nanomachines: principles and applications, *European Biophysics Journal*, **36** (2007), 995–1013.

30. C. O. S. Sorzano, R. Marabini, , J. Velázquez-Muriel, J. R. Bilbao-Castro, S. H. W. Scheres, J. M. Carazo and A. Pascual-Montano, XMIPP: A new generation of an open-source image processing package for electron microscopy, *J. Structural Biology*, **148** (2004), 194–204.

31. C. O. S. Sorzano, R. Marabini, N. Boisset, E. Rietzel, R. Schröder, G. T. Herman and J. M. Carazo, The effect of overabundant projection directions on 3D reconstruction algorithms, *J. Structural Biology*, **133** (2001), 108–118.

32. J. C. H. Spence, On the accurate measurement of structure-factor amplitudes and phases by electron diffraction, *Acta Crystallographica*, **A49** (1993), 231–260.

33. P. A. Stadelmann, EMS - a software package for electron diffraction analysis and HREM image simulation in materials science, *Ultramicroscopy*, **21** (1987), 131–146.

34. S. M. Stagg, J. Pulokas, D. Fellmann, A. Cheng, J. D. Quispe, S. P. Mallick, R. M. Avila, B. Carragher and C. S. Potter, Automated cryoem data acquisition and analysis of 284,742 particles of groel, *Nature*, **439** (2006), 234–238.

35. F. Tama, O. Miyashita and C. L. Brooks, Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis., *J Mol Biol*, **337** (2004), 985–999.

36. F. Tama, O. Miyashita and C. L. Brooks III, Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM, *J. Structural Biology*, **147** (2004), 315–326.

37. E. Tjioe, K. Lasker, B. Webb, H. J. Wolfson and A. Sali, Multifit: a web server for fitting multiple protein structures into their electron microscopy density map., *Nucleic Acids Res*, **39** (2011), W167–W170.

38. M. Topf, K. Lasker, B. Webb, H. Wolfson, W. Chiu and A. Sali, Protein structure fitting and refinement guided by cryo-em density., *Structure*, **16** (2008), 295–307.

39. L. G. Trabuco, E. Villa, K. Mitra, J. Frank and K. Schulten, Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics., *Structure*, **16** (2008), 673–683.

40. M. Unser, A. Aldroubi and M. Eden, B-Spline signal processing: Part I - theory, *IEEE Trans. Signal Processing*, **41** (1993), 821–832.

41. M. Unser, A. Aldroubi and M. Eden, B-Spline signal processing: Part II–Efficient design and applications, *IEEE Trans. Signal Processing*, **41** (1993), 834–848.

42. M. van Heel, G. Harauz, E. V. Orlova, R. Schmidt and M. Schatz, A new generation of the IMAGIC image processing system, *J. Structural Biology*, **116** (1996), 17–24.

43. M. Vulović, R. B. G. Ravelli, L. J. van Vliet, A. J. Koster, I. Lazić, U. Lücken, H. Rullgård, O. Öktem and B. Rieger, Image formation modeling in cryo-electron microscopy., *J Struct Biol*, **183** (2013), 19–32, URL http://dx.doi.org/10.1016/j.jsb.2013.05.008.

44. A. J. C. Wilson (ed.), *International tables for crystallography*, 500, Kluwer Academics Publisher, 1995.

45. W. Wriggers, Using situs for the integration of multi-resolution structures., *Biophys Rev*, **2** (2010), 21–27.

46. W. Wriggers, R. A. Milligan and J. A. McCammon, Situs: A package for docking crystal structures into low-resolution maps from electron microscopy, *J. Structural Biology*, **125** (1999), 185–195.

47. X. Zhang, L. Jin, Q. Fang, W. H. Hui and Z. H. Zhou, 3.3 a cryo-em structure of a nonenveloped virus reveals a priming mechanism for cell entry., *Cell*, **141** (2010), 472–482.