Methods for analyzing continuous conformational variability of biomolecules in cryo electron subtomograms: HEMNMA-3D vs. traditional classification

Mohamad Harastani   Slavica Jonic

IMPMC - UMR 7590 CNRS, Sorbonne Université, MNHN

4 Place Jussieu, Paris, France

{mohamad.harastani, slavica.jonic}@upmc.fr

Abstract

Cryogenic electron tomography (cryo-ET) allows studying biological macromolecular complexes in cells by three-dimensional (3D) data analysis. The complexes continuously change their shapes (conformations) to achieve biological functions. The shape heterogeneity in the samples imaged in the cryo electron microscope is a bottleneck for comprehending biological mechanisms and developing drugs. Low signal-to-noise ratio and spatial anisotropy (missing wedge artefacts) make cryo-ET data particularly challenging for resolving the shape variability. Other shape variability analysis techniques simplify the problem by considering discrete rather than continuous conformational changes of complexes. Recently, HEMNMA-3D was introduced for cryo-ET continuous shape variability analysis, based on elastic and rigid-body 3D registration between simulated shapes and cryo-ET data. The simulated motions are obtained by normal mode analysis of a high- or low-resolution 3D reference model of the complex under study. The rigid-body alignment is achieved via fast rotational matching with missing wedge compensation. HEMNMA-3D provides a visual insight into molecular dynamics by grouping and averaging subtomograms of similar shapes and by animating movies of registered motions. This article reviews the method and compares it with existing literature on a simulated dataset for nucleosome shape variability.

1. Introduction

Three-dimensional (3D) volumetric images of vitrified cell sections (slices) can be obtained using cryogenic electron tomography (cryo-ET). The 3D nature of cryo-ET data permits studying macromolecular complexes despite the crowded cell environment. The most common cryo-ET data acquisition scheme is the acquisition of multiple 2D projection images of a specimen rotated around a single axis (perpendicular to the electron beam) inside the electron microscope. The obtained tilt-series 2D projections are used to reconstruct a 3D volume (referred to as tomogram) based on the Fourier slice theorem and backprojection. The tomogram of a cell section typically contains hundreds of copies of different macromolecules at random orientations. The copies of the macromolecule under study (target macromolecule) can be identified and extracted into individual volumes called subtomograms, either manually or via template matching.

Subtomograms suffer from a low signal-to-noise ratio (SNR) due to exposing the sample to a low electron dose during data acquisition to preserve the fragile biological structure from radiation damage. Additionally, subtomograms suffer from spatial anisotropies, known as missing wedge artefacts, due to the inability to include in the 3D reconstruction the images from all orientations (the maximum tilt angle in the microscope is usually limited to ±60°), which corresponds to a missing wedge-shaped region in 3D Fourier space. The missing wedge artefacts are often observed as elongations along the beam axis, blurring and distracting caustics in the subtomograms. Due to the low SNR and the missing wedge artefacts, cryo-ET data processing is mainly based on rigid-body aligning and averaging of many subtomograms to enhance the data quality and reveal the target macromolecular structure, which is known as Subtomogram Averaging (StA) [1]. However, with recent instrumentation and software development, more research moves in the direction of studying subtomograms individually (e.g., development of methods for denoising and missing wedge correction with no or a minimum of averaging [2, 3]).

The main focus of studying macromolecules in their native state is to observe their shapes and dynamics. In the last decade, cryo-electron microscopy (cryo-EM) research has shown that disentangling macromolecular shape variability and identifying the macromolecular conformational transitions is valuable for understanding biological mechanisms [4, 5]. The most popular cryo-EM technique, namely Single Particle Analysis (SPA), allows a near-atomic resolu-

tion of 3D reconstruction of macromolecules *in vitro* (from vitrified samples of a solution containing biochemically purified macromolecules), by combining alignment and classification in 2D and 3D [6]. Only some SPA methods consider continuous shape variability, meaning gradual conformational transitions of macromolecules with many intermediate conformational states. They represent images in a low-dimensional space allowing a 3D visualization of the macromolecular shape variability [7, 8]. In contrast to SPA, cryo-ET allows studying macromolecules *in situ* (in vitrified cell sections). The conformational variability in the native cellular environment is largely overlooked, due to the lack of cryo-ET image analysis methods capable of dealing with continuous shape variability in the data with low SNR and missing-wedge artefacts.

Contribution: HEMNMA-3D [9] was introduced for cryo-ET continuous macromolecular shape variability analysis, inspired by HEMNMA, a method for continuous shape variability analysis in SPA [10, 8]. HEMNMA-3D is based on elastic and rigid-body 3D registration between simulated shapes and cryo-ET data. The simulated motions are obtained by normal mode analysis of a high- or low-resolution 3D reference model of the complex under study. The rigid-body alignment is achieved via fast rotational matching with missing wedge compensation. HEMNMA-3D provides a visual insight into molecular dynamics by grouping and averaging subtomograms of similar shapes and by animating movies of registered motions. This article reviews the method and compares it with existing literature on a simulated dataset for nucleosome shape variability.

2. Related work

2.1. 3D elastic registration for medical images

Elastic registration is extensively used in medical image processing for a wide spectrum of applications, including Magnetic Resonance Imaging [11] and Computed Tomography [12]. Noteworthy, several multi-purpose machine learning and deep learning-based 3D registration methods were proposed [13, 14]. However, none of these methods was adapted or applied to cryo-ET, possibly due to severe cryo-ET data limitations such as missing wedge artefacts and poor SNR.

2.2. Cryo-ET subtomogram classification

Methods reported to deal with cryo-ET data heterogeneity are based on classification and rigid-body alignment. They can be split into two categories, namely multi-reference alignment approaches [15] and post-alignment classification approaches [16].

Multi-reference alignment approaches are based on competitive alignment. An expert user provides a number of references with different shapes of the macromolecule under study based on anticipated shape variabilities. An algorithm then aligns (rigid-body-wise) and compares each subtomogram with the multiple given references and attributes it to the reference that yields the highest similarity score. The starting references evolve by averaging the aligned subtomograms and repeating the process until stability. Such methods require prior knowledge of the biological specimen and are prone to overfitting and data misinterpretation if not carefully used.

Post-alignment classification approaches perform classification after rigid-body alignment via StA. During StA, subtomograms are rigid-body aligned against a reference to maximize a scoring function. In each iteration, the aligned subtomograms are averaged to produce a structure that becomes the reference for the next StA iteration. The iterations repeat until convergence. The result is an average structure and aligned subtomograms (centred and oriented with respect to the reference used in the last iteration). Classifying aligned subtomograms according to the shape variability is challenging because of the missing wedge artefacts. The aligned subtomograms are usually classified based on the covariance matrix calculated using a constrained correlation coefficient (CCC) between each pair of aligned subtomograms. The CCC corresponds to constraining the cross-correlation evaluation to the Fourier-space region that excludes the missing wedge of both subtomograms [17]. The covariance matrix serves as a basis for a hierarchical classification technique, or it is fed to a dimensionality reduction method first and, then, to a clustering technique [17, 18]. A drawback of post-alignment classification methods is their dependency on rigid-body alignment accuracy, which decreases as the specimen's conformational heterogeneity increases.

In practice, the classification remains efficient at resolving discrete structural variabilities such as ligand binding and macromolecular disassembly, but for classifying continuous shape variants, particles assigned to the same class will rarely, if ever, have identical conformations.

3. HEMNMA-3D

A flowchart for HEMNMA-3D is in Figure 1. The method comprises the following blocks:

1. Input: subtomograms and a **reference structure**;
2. **Normal Mode Analysis** of the reference structure;
3. **Elastic and rigid-body 3D registration**; and
4. **Visualizing macromolecular shape variability** via selective data averaging and animations.

3.1. Reference structure

The 3D registration is performed using a reference structure in the Protein Data Bank (PDB) format. If there is an
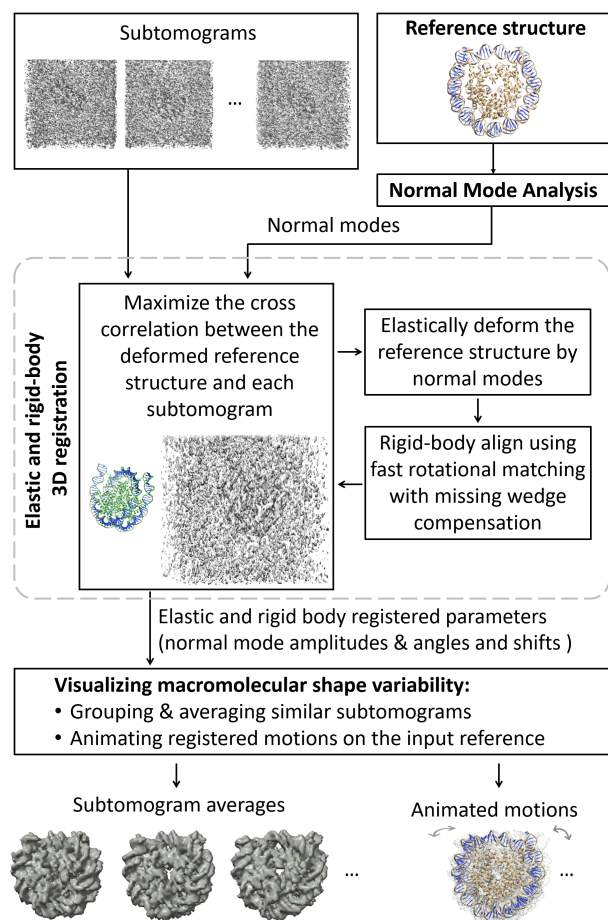
Figure 1. Flowchart of HEMNMA-3D for macromolecular shape variability analysis in cryo-ET using 3D elastic registration via normal mode analysis. Note: the displayed molecule is the nucleosome (PDB:3w98).

existing atomic model of the macromolecule under study (e.g., a model available in the PDB database and obtained by X-ray crystallography, nuclear magnetic resonance, or derived from 3D reconstructed density maps from cryo-EM images), this model can be used as the reference structure for the method. However, if no atomic model is available but a lower-resolution density map can be obtained by SPA reconstruction or subtomogram averaging, this density map can be converted into a pseudoatomic model [19, 20] and, then, this model can be used as the reference structure for the method. The PDB file of the reference structure contains 3D Cartesian coordinates of atoms or pseudoatoms.

3.2. Normal Mode Analysis

Normal Mode Analysis (NMA) is a method for molecular mechanics simulation. One of its main applications is the elastic deformation of an existing atomic structure to fit a cryo-EM density map of the same macromolecule but at a different conformation. This application is usually known

as normal mode flexible fitting and allows deriving atomic models from cryo-EM maps [21].

Normal modes are computed from the atomic or pseudoatomic reference structure based on the elastic network model [22]. This model represents the atoms or pseudoatoms as locally connected (within a cutoff distance) by elastic springs. Normal modes are vectors that describe harmonic-oscillator motions of the elastic network model (the length of each normal-mode vector is 3 times the number of atoms or pseudoatoms). Computationally, normal modes are eigenvectors of a Hessian matrix of the system (the matrix of second derivatives of the potential energy function of the elastic network model) and the corresponding eigenvalues are the squares of the normal-mode frequencies. Also, for each normal mode, its collectivity degree is computed, which measures the percentage of atoms or pseudoatoms that move with that mode. Low-frequency, high-collectivity normal modes have been shown to be relevant to macromolecular conformational variability [23].

The atomic or pseudoatomic displacement is determined by a linear combination of normal modes, where the coefficients of the linear combination are the displacement amplitudes along the modes. NMA only allows computing normal modes (vectors), meaning atomic or pseudoatomic displacement directions. The amplitudes of these displacements are computed by 3D elastic registration. In the context of 3D elastic registration, the key advantage of NMA is that selecting a subset of normal modes with the lowest frequencies and highest collectivities allows faster data analysis and a regularization against noise overfitting, which was also observed in 3D-to-2D elastic registration [8].

3.3. Elastic and rigid-body 3D registration

This module comprises simultaneous NMA-based elastic registration and Fast Rotational Matching (FRM)-based rigid-body alignment of the reference structure with each given subtomogram via numerical optimization.

Given a subtomogram, a reference structure, and a set of normal modes of the reference structure, this module searches for the displacement amplitudes along the normal modes (elastic parameters) and the angles and shifts (rigid-body parameters) of the reference structure to match the structure to the subtomogram. A numerical optimizer maximizes the similarity between the subtomogram and a density volume simulated [24] from the elastically deformed, oriented, and shifted atomic or pseudoatomic reference. The similarity measure is the CCC, which is here defined as the cross-correlation between the reference and subtomogram density maps excluding the Fourier-space region corresponding to the subtomogram's missing wedge. Such similarity measure allows to compensate for the missing wedge, which otherwise can lead to erroneous results. The numerical optimizer is a variant of Powell's UOBYQA

method known as CONDOR [25], which subjects the objective function to a trust-region radius and avoids noise overfitting by giving more credibility to smaller normal mode amplitudes during the registration.

For each subtomogram, the corresponding normal mode amplitudes are initiated with zeros, i.e. the non-deformed reference is used in the first iteration. As the iterations evolve, the reference model is deformed with the new guesses of the normal mode amplitudes, converted into a volume and rigid-body aligned with the subtomogram using FRM [26]. The rotational matching allows a fast and accurate 6D search for orientations and shits (three Euler angles and three translation parameters). At the end of each iteration, the CCC is found and fed to the numerical optimizer. The iterations repeat, refining the elastic and rigid-body registration parameters until convergence.

3.4. Macromolecular shape variability visualization

The number of elastic registration parameters (normal-mode amplitudes) is determined by the number of normal modes used for the elastic registration. If more than three normal modes were used, the ensemble of normal-mode amplitudes (for all subtomograms) can be projected onto a lower-dimensional space using a dimensionality reduction technique, e.g. Principal Component Analysis (PCA). The obtained high- or low-dimensional space of normal mode amplitudes (conformational space) allows a global data display for interpreting macromolecular shape variability. Each point in this space represents a subtomogram, and close points correspond to similar registered shapes. In this conformational space, the shape variability can be analyzed in the following two ways: 1) by averaging subtomograms of similar registered shapes; and 2) by animating the registered motion of the reference structure along different directions.

Grouping and averaging similar subtomograms: As close points in the conformational space correspond to similar registered shapes, grouping and averaging close points in dense regions can help visualize variable macromolecular shapes with better SNR and less artefacts (attenuated noise and missing wedge artefacts thanks to averaging of many similar shapes). Before computing the group averages, the rigid-body alignment parameters found via the elastic and rigid-body 3D registration module are applied to the subtomograms. The subtomogram averages from different regions of the conformational space can be overlapped for a visual comparison.

Animating registered motions of the input reference: We can further analyze the conformational space by animating registered motions for several points through data distribution manifolds (e.g., following a straight line or a curve fitting the data). If a dimensionality reduction technique is applied, the inverse mapping must be first used (e.g. inverse

PCA) to find the corresponding normal mode amplitudes for each point. These normal modes amplitudes are applied to elastically deform the reference structure (one elastically deformed structure is obtained for each selected point). By concatenating the resulting structures, a movie-like animation of registered motion can be obtained.

4. Implementation

The software of HEMNMA-3D is open-source and available on GitHub [10, 9]. It is built as part of the Continuous-Flex plugin of the open-source Scipion software package [27], commonly used for cryo-EM data processing. Our software provides a graphical user interface (GUI) and is empowered with a C++ backend with a message passing interface (MPI) parallelization scheme. The number of subtomograms that can be processed simultaneously depends on the number of available CPUs. The current implementation takes around 10 minutes to analyze a subtomogram of size 64^3 voxels using three normal modes (tested on 2.2 GHz Intel Xeon Silver 4214 CPU processor and 64 GB RAM). The more normal modes are used, the slower the processing.

5. Experiment

5.1. Simulating nucleosome shape variability

To review and compare HEMNMA-3D performance with existing literature, we synthesized a dataset comprising 1000 subtomograms with an imagined continuous shape variability of the nucleosome. We generated a linear combination of two reported motions for the nucleosome, breathing and gaping [28], with a linear dependence between the amplitudes of normal modes corresponding to the two motions.

First, we performed NMA of the nucleosome atomic structure available in the PDB database under the code 3w98 and we visualized the motions carried by the different computed normal modes. Among these modes, we identified the modes describing breathing and gaping motions as normal modes 9 and 13, respectively (note here that the mode number corresponds to the frequency of the mode and that higher numbers correspond to higher frequencies). We generated a dataset using a linear relationship between the amplitudes of normal modes 9 and 13 so that the nucleosome is simultaneously breathing and gaping. Precisely, at one end of the generated ground-truth conformational distribution, the nucleosome's two DNA ends (arms) are moving away from each other, and at the same time, the gap between the two DNA gyres increases. At the other end of the generated conformational distribution, the DNA arms approach each other and the gap between the two DNA gyres decreases. We simulated a gradual transition between the two ends, representing a continuum of nucleosome shapes, combining breathing and gaping. Equal random amplitudes
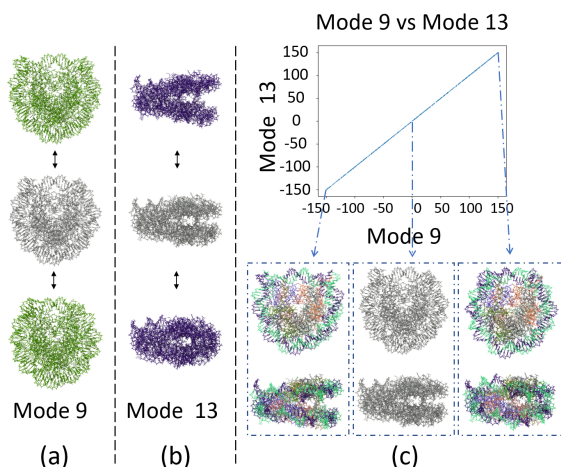
Figure 2. Synthesized combined breathing and gaping motions of the nucleosome (PDB 3w98 structure): (a) nucleosome breathing motion, (b) nucleosome gaping motion, (c) generated ground-truth conformational distribution (top) comprising 1000 synthetic nucleosome shape variants obtained by a linear combination of modes 9 and 13, with a linear dependence between the normal-mode amplitudes (blue points in the plot), and 3 representative shapes (bottom) corresponding to the two ends and the middle of the conformational distribution.

uniformly distributed in the range [-150, 150] were used for the two normal modes 9 and 13. An illustration of the simulated movements is provided in Figure 2.

To generate this dataset, for each subtomogram, we performed the following steps:

1. Elastically deform the atomic structure (PDB:3w98) using equal random amplitudes for the two normal modes 9 and 13 in the range [-150, 150].
2. Convert the elastically deformed structure to a density map of size $64 \times 64 \times 64$ voxels (voxel size: 3.45 Å $\times$ 3.45 Å $\times$ 3.45 Å), using [24].
3. Rotate and shift the volume in 3D space using random Euler angles and random x, y, z shifts (the random shift range is ±5 pixels from the center).
4. Tilt and project the randomly rotated and shifted volume, using the tilt angle from -60° to +60° with 1° step, to obtain a collection of 2D projection images (i.e. tilt series).
5. Simulate microscope conditions by adding heavy noise (SNR = 0.01) and modulating the images with the contrast transfer function (CTF) of the microscope (using the defocus of -0.5 μm).
6. Invert the CTF phase (a common CTF correction).
7. Reconstruct a volume (our synthetic subtomogram) from the obtained tilt series using a Fourier reconstruction method [29].

Figure 3 shows an example subtomogram from the synthesized dataset and the corresponding ideal volume, for comparison in real space and in Fourier space.
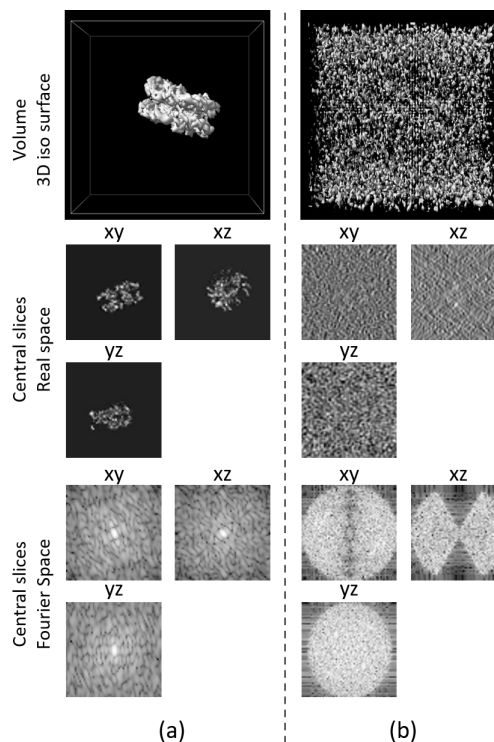

Figure 3. Example of a noisy and missing-wedge affected synthetic subtomogram compared with the corresponding ideal volume of the nucleosome: (a) ideal volume (without noise and without missing wedge artefacts), (b) noisy and missing-wedge affected synthetic subtomogram.

5.2. Traditional subtomogram averaging and post alignment classification

StA provides a global average without considering the shape variability, and it provides a basis for performing classification of subtomograms (classification of the subtomograms aligned through StA).

We applied StA on the synthesized nucleosome dataset, using the protocol based on the rigid-body alignment approach of [26] (recall that this rigid-body alignment approach is also used in the elastic and rigid-body alignment of HEMNMA-3D). This StA protocol uses an exhaustive angular search (with FRM method) and a shifts search within a region of interest, and compensates for the missing wedge by using the CCC (evaluation of the correlation between the subtomogram average of each iteration and the given subtomogram density maps, but excluding the evaluation in the missing-wedge region of the given subtomogram).

We followed the procedure in [26] and set the shifts search region to 10 voxels from the image center. We started iterations using an average of the unaligned subtomograms (this StA procedure is referred to as reference-free alignment). After six iterations, StA converged (further itera-
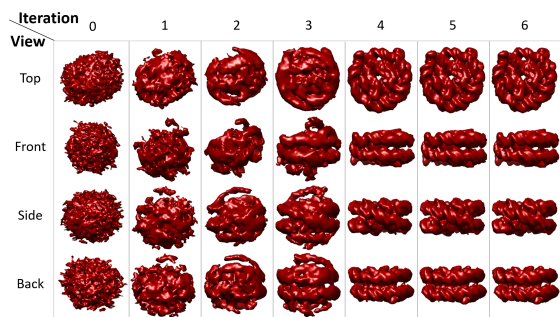
Figure 4. Subtomogram averaging applied to the synthesized nucleosome subtomograms. A reference-free alignment was performed using Fast Rotational Matching [26].
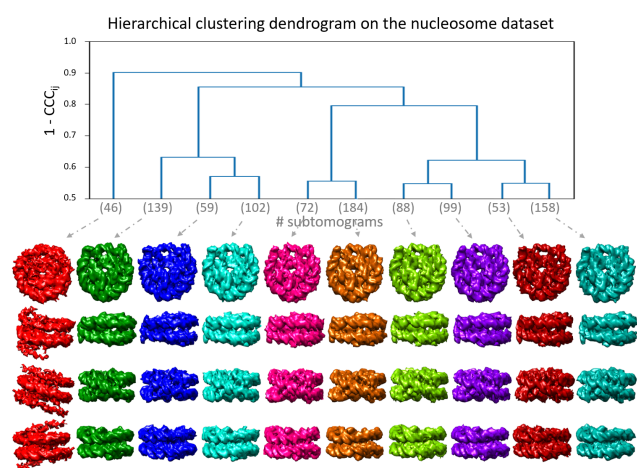

Figure 5. Hierarchical clustering applied to the synthesized nucleosome subtomograms. Top: hierarchical tree for $1\text{-}CCC_{ij}$ matrix. Bottom: views (vertically in the same color) of different subtomogram class averages (horizontally in different colors).

tions gave the same results). The StA averages are shown in Figure 4.

After StA, we applied the obtained rigid-body alignment parameters (found through StA) on the subtomograms, and we evaluated the covariance matrix CCC_{ij} of pairwise constrained cross-correlation (see section 2.2 for more details). We performed the two most common post-alignment classification techniques on the CCC_{ij} matrix, namely hierarchical clustering [18] and PCA followed by k-means [17].

The hierarchical clustering on $1\text{-}CCC_{ij}$ matrix was performed to 10 classes using the Agglomerative Clustering module of Python Scikit-Learn package (version 0.22.1 and default parameters were used) [30]. We note that applying the clustering algorithm directly on the CCC_{ij} matrix gives identical results, and we used the convention proposed in the literature [18, 17]. The clustering tree (dendrogram) and class averages are shown in Figure 5.

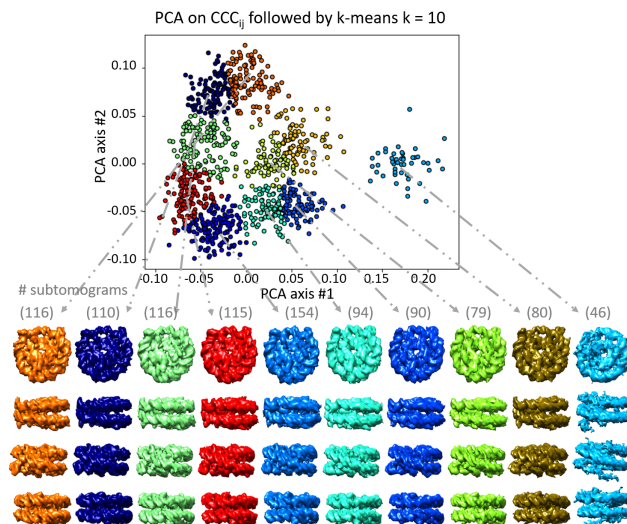The k-means clustering was performed following PCA

Figure 6. K-means clustering applied to the synthesized nucleosome subtomograms. Top: k-means clustering in the space of the first two PCA axes of CCC_{ij} matrix. Bottom: views (vertically in the same color) of different subtomogram averages (horizontally in different colors).

on the CCC_{ij} matrix. The clustering was done into 10 classes (k=10) based on the first two principal axes, using the k-means module of Scikit-Learn. In general, the choice of the number of principal axes to perform classification is arbitrary, as explained in [17]. Since the dataset was synthesized with two degrees of freedom (nucleosome breathing and gaping), we set the number of principle axes to 2, to obtain the best results. Figure 6 shows the classification of the PCA space and the resultant class averages.

We note that the two tested classification techniques give similar outputs, showing different discrete class averages of the nucleosome, at different breathing and gaping magnitudes (Figures 5 and 6). However, these outputs do not allow an unambiguous interpretation of the results in terms of the synthesized ground-truth conformational transitions of the nucleosome (from the smallest magnitudes to the largest magnitudes of breathing and gaping and vice versa).

5.3. HEMNMA-3D

Applying HEMNMA-3D to the synthesized nucleosome dataset aims at solving the inverse problem of finding the nucleosome shape variant in each subtomogram, i.e. estimating the amplitudes of normal modes 9 and 13 of the PDB structure 3w98 as close as possible to the generated ground-truth amplitudes. We set the method parameters as follows:

- **NMA settings:** To make the elastic and rigid-body 3D registration task more realistic and challenging, we used three normal modes (modes 9, 10 and 13) instead of only two modes (modes 9 and 13 used to generate the dataset).

- **FRM settings** The shift range for the rigid-body registration (FRM method) is set to 10 pixels.

The amplitudes estimated for modes 9 and 13 using HEMNMA-3D are shown in Figure 7 (a). It is graphically intelligible that the linear relationship is retrieved between the estimated amplitudes of the two modes. Figure 7 (b) shows the histogram of the amplitudes estimated for mode 10 and confirms that they are globally near zero.

Table 1 presents the mean absolute error between the estimated and ground-truth normal-mode amplitudes and the standard deviation of the error. It should be noted that 14/1000 points were excluded from the statistics as found to differ significantly (outlier points) from the remaining observations. These points were excluded for having a p-value below 10^{-4} based on the Mahalanobis distance [31].

Normal-mode amplitudes do not have a physical unit. Nonetheless, the Root Mean Square Deviation (RMSD) [32] between the reference atomic coordinates and these coordinates displaced using the calculated errors as the normal-mode amplitudes can transform these errors in physical units. The nucleosome core complex comprises eight histone proteins surrounded by 146 DNA base pairs. The synthesized movements (breathing and gaping) mainly impacted the DNA loops. Evaluating the RMSD without excluding the core histones can give a false sense of achieving higher accuracy by pulling the RMSD value towards zero. Therefore, the reported RMSD hereafter is based on the nucleosome's DNA loops only (chain I and J of the PDB structure 3w98).

We found a RMSD of 0.44 Å corresponding to the mean absolute errors in Table 1 (for a combined displacement along modes 9, 10 and 13). Also, we found a RMSD of 0.79 Å corresponding to the sum of the mean and standard deviation of the errors in Table 1. Hence, the error range is significantly inferior to the pixel size used to create the data (3.45 Å).

Figure 8 (a) shows grouping and averaging of subtomograms through the point distribution in the conformational space (ten equally distanced groups). The corresponding subtomogram averages show the expected combination of continuous motions of breathing and gaping, which can be compared with the ground-truth motion in Figure 2.

Figure 8 (b) shows the displacement of the reference structure along 10 points in the direction of the point distribution in the conformational space (the procedure explained in Section 3.4).

The obtained subtomogram averages and animation show that the ground-truth nucleosome motion (a combination of breathing and gaping) was retrieved.
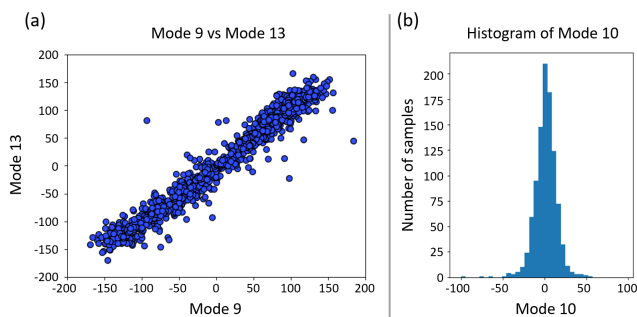

Figure 7. Output of the elastic and rigid-body 3D registration module of HEMNMA-3D using synthesized nucleosome subtomograms. The goal was the retrieval of the ground-truth amplitudes of normal modes 9, 10 and 13. Ideally, the amplitudes of mode 10 are equal to zero and there is a linear relationship between the amplitudes of modes 9 and 13 in the range [-150, 150]: (a) amplitudes of mode 9 vs amplitudes of mode 13, (b) histogram of amplitudes of mode 10.

Normal mode amplitude	Mode 9		Mode 10		Mode 13	
Actual range	[-150, 150]		0		[-150,150]	
Measure	Mean	Std	Mean	Std	Mean	Std
Absolute error	10.86	8.66	9.98	7.82	10.81	8.83

Table 1. Mean absolute error between the estimated and ground-truth normal-mode amplitudes and the standard deviation of the error, for HEMNMA-3D using synthesized nucleosome subtomograms. Points below the p-value of 10^{-4} were excluded (14/1000 points) from the error evaluation based on the Mahalanobis distance.

6. Discussion and conclusion

This article presented a review on HEMNMA-3D, a method that addresses continuous macromolecular shape variability captured in cryo-ET subtomograms. HEMNMA-3D applies elastic and rigid body 3D registration via normal mode analysis, fast rotational matching and trust-region based numerical optimization.

We tested HEMNMA-3D by synthesizing a dataset of nucleosome shape variability under challenging conditions of the simulated microscope and testing the method's capability to recover the ground-truth shapes. The test results indicate that the method recovers the ground-truth combination of two nucleosome motions (breathing and gaping) with subpixel accuracy. Also, it does not overfit the data with non-existing motions (normal mode 10 was not used to create the nucleosome dataset but used for 3D elastic registration, and it was retrieved with amplitudes near zero).

We applied two state-of-the-art methods for cryo-ET classification on the synthesized nucleosome dataset. Both methods gave similar outputs, showing different discrete class averages of the nucleosome at different breathing and
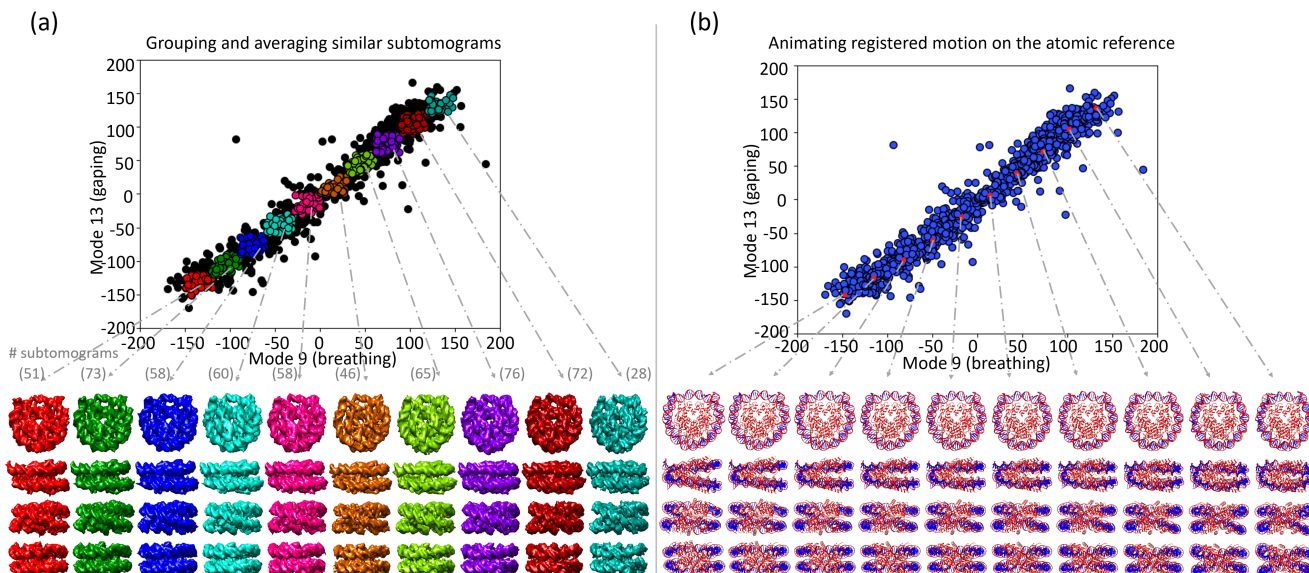
Figure 8. HEMNMA-3D applied to the synthesized nucleosome subtomograms. (a) group averages for ten equally distanced groups along the subtomogram (point) distribution in the conformational space, (b) displacement of the reference PDB structure 3w98 along the direction of the data distribution in the conformational space, 10 frames represented by red dots. Note: each column represents four different views of the same structure.

gaping magnitudes. However, the choice of the number of classes is arbitrary in these methods and the shape transitions between the obtained class averages are ambiguous, probably because of the continuous nature of the shape variability.

In contrast to these methods, HEMNMA-3D adopts a new scheme that permits revealing hidden macromolecular dynamics by i) grouping and averaging similar subtomograms at locations in the conformational space that reveal the shape transitions and ii) animating the reference structure by displacing it in different directions in the conformational space. Hence, HEMNMA-3D provides a promising new insight into what can be achieved in cryo-ET studies of macromolecular shape variability. For more details on HEMNMA-3D and an example of its use and results with experimental cryo-ET subtomograms, the reader is referred to [9].

However, unlike the classification methods, HEMNMA-3D is limited to macromolecular elastic shape variability that can be explained with NMA. It is not suitable for analyzing other structural variabilities such as macromolecular disassembly or binding and unbinding of ligands. Future work can involve combining this method with classification to first disentangle such discrete structural variabilities and then analyze continuous intraclass variability.

7. Acknowledgements

References

[1] W. Wan and J. Briggs, "Chapter thirteen - cryo-electron tomography and subtomogram averaging," in *The Resolution Revolution: Recent Advances In cryoEM* (R. Crowther, ed.), vol. 579 of *Methods in Enzymology*, pp. 329 – 367, Academic Press, 2016. 1

[2] E. Moebel and C. Kervrann, "A monte carlo framework for missing wedge restoration and noise removal in cryo-electron tomography," *Journal of Structural Biology: X*, vol. 4, p. 100013, 2020. 1

[3] X. Zhai, D. Lei, M. Zhang, J. Liu, H. Wu, Y. Yu, L. Zhang, and G. Ren, "Lottor: An algorithm for missing-wedge correction of the low-tilt tomographic 3d reconstruction of a single-molecule structure," *Scientific reports*, vol. 10, no. 1, pp. 1–17, 2020. 1

[4] P. D. Abeyrathne, C. San Koh, T. Grant, N. Grigorieff, and A. A. Korostelev, "Ensemble cryo-em uncovers inchworm-like translocation of a viral ires through the ribosome," *Elife*, vol. 5, p. e14874, 2016. 1

[5] A. Zhou, A. Rohou, D. G. Schep, J. V. Bason, M. G. Montgomery, J. E. Walker, N. Grigorieff, and J. L. Rubinstein, "Structure and conformational states of the bovine mitochondrial atp synthase by cryo-em," *Elife*, vol. 4, p. e10180, 2015. 1

[6] X. Zhang, E. Settembre, C. Xu, P. R. Dormitzer, R. Bellamy, S. C. Harrison, and N. Grigorieff, "Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction," *Proceedings of the National Academy of Sciences*, vol. 105, no. 6, pp. 1867–1872, 2008. 2

[7] A. Dashti, P. Schwander, R. Langlois, R. Fung, W. Li, A. Hosseinizadeh, H. Y. Liao, J. Pallesen, G. Sharma, V. A. Stupina, A. E. Simon, J. D. Dinman, J. Frank, and A. Ourmazd, "Trajectories of the ribosome as a brownian nanomachine," *Proceedings of the National Academy of Sciences*, vol. 111, no. 49, pp. 17492–17497, 2014. 2

[8] Q. Jin, C. O. S. Sorzano, J. M. De La Rosa-Trevín, J. R. Bilbao-Castro, R. Núñez-Ramírez, O. Llorca, F. Tama, and S. Jonić, "Iterative elastic 3d-to-2d alignment method using normal modes for studying structural dynamics of large macromolecular complexes," *Structure*, vol. 22, no. 3, pp. 496–506, 2014. 2, 3

[9] M. Harastani, M. Eltsov, A. Leforestier, and S. Jonic, "Hemnma-3d: Cryo electron tomography method based on normal mode analysis to study continuous conformational variability of macromolecular complexes," *Frontiers in molecular biosciences*, vol. 8, 2021. 2, 4, 8

[10] M. Harastani, C. O. S. Sorzano, and S. Jonić, "Hybrid electron microscopy normal mode analysis with scipion," *Protein Science*, vol. 29, no. 1, pp. 223–236, 2020. 2, 4

[11] A. Wittek, K. Miller, R. Kikinis, and S. K. Warfield, "Patient-specific model of brain deformation: Application to medical image registration," *Journal of biomechanics*, vol. 40, no. 4, pp. 919–929, 2007. 2

[12] J. Hou, M. Guerrero, W. Chen, and W. D. D'Souza, "Deformable planning ct to cone-beam ct image registration in head-and-neck cancer," *Medical physics*, vol. 38, no. 4, pp. 2088–2094, 2011. 2

[13] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical image analysis*, vol. 52, pp. 128–143, 2019. 2

[14] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9252–9260, 2018. 2

[15] M. Stölken, F. Beck, T. Haller, R. Hegerl, I. Gutsche, J.-M. Carazo, W. Baumeister, S. H. Scheres, and S. Nickell, "Maximum likelihood based classification of electron tomographic data," *Journal of structural biology*, vol. 173, no. 1, pp. 77–85, 2011. 2

[16] Y. Chen, S. Pfeffer, J. J. Fernández, C. O. S. Sorzano, and F. Förster, "Autofocused 3d classification of cryoelectron subtomograms," *Structure*, vol. 22, no. 10, pp. 1528–1537, 2014. 2

[17] F. Förster, S. Pruggnaller, A. Seybert, and A. S. Frangakis, "Classification of cryo-electron sub-tomograms using constrained correlation," *Journal of structural biology*, vol. 161, no. 3, pp. 276–286, 2008. 2, 6

[18] M. Xu, M. Beck, and F. Alber, "High-throughput subtomogram alignment and classification by fourier space constrained fast volumetric matching," *Journal of structural biology*, vol. 178, no. 2, pp. 152–164, 2012. 2, 6

[19] S. Jonić and C. Sorzano, "Coarse-graining of volumes for modeling of structure and dynamics in electron microscopy: Algorithm to automatically control accuracy of approximation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 1, pp. 161–173, 2016. 3

[20] R. Nogales-Cadenas, S. Jonic, F. Tama, A. Arteni, D. Tabas-Madrid, M. Vázquez, A. Pascual-Montano, and C. O. S. Sorzano, "3dem loupe: analysis of macromolecular dynamics using structures from electron microscopy," *Nucleic acids research*, vol. 41, no. W1, pp. W363–W367, 2013. 3

[21] 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," *Journal of Structural Biology*, vol. 147, no. 3, pp. 315 – 326, 2004. Time-Resolved Imaging of Macromolecular Processes and Interactions. 3

[22] M. M. Tirion, "Large amplitude elastic motions in proteins from a single-parameter, atomic analysis," *Physical review letters*, vol. 77, no. 9, p. 1905, 1996. 3

[23] J. Ma, "Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes," *Structure*, vol. 13, no. 3, pp. 373–380, 2005. 3

[24] L.-M. Peng, G. Ren, S. Dudarev, and M. Whelan, "Robust parameterization of elastic and absorptive electron atomic scattering factors," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 52, no. 2, pp. 257–276, 1996. 3, 5

[25] F. V. Berghen and H. Bersini, "Condor, a new parallel, constrained extension of powell's uobyqa algorithm: Experimental results and comparison with the dfo algorithm," *Journal of computational and applied mathematics*, vol. 181, no. 1, pp. 157–175, 2005. 4

[26] Y. Chen, S. Pfeffer, T. Hrabe, J. M. Schuller, and F. Förster, "Fast and accurate reference-free alignment of subtomograms," *Journal of Structural Biology*, vol. 182, no. 3, pp. 235–245, 2013. 4, 5, 6

[27] J. De la Rosa-Trevín, A. Quintana, L. Del Cano, A. Zaldivar, I. Foche, J. Gutierrez, J. Gomez-Blanco, J. Burguet-Castell, J. Cuenca-Alba, V. Abrishami, *et al.*, "Scipion: A software framework toward integration, reproducibility and validation in 3d electron microscopy," *Journal of structural biology*, vol. 195, no. 1, pp. 93–99, 2016. 4

[28] J. Zlatanova, T. C. Bishop, J.-M. Victor, V. Jackson, and K. van Holde, "The nucleosome family: dynamic and growing," *Structure*, vol. 17, no. 2, pp. 160–171, 2009. 4

[29] C. O. Sorzano, J. de la Rosa Trevín, J. Otón, J. Vega, J. Cuenca, A. Zaldívar-Peraza, J. Gómez-Blanco, J. Vargas, A. Quintana, R. Marabini, *et al.*, "Semiautomatic, high-throughput, high-resolution protocol for three-dimensional reconstruction of single particles in electron microscopy," in *Nanoimaging*, pp. 171–193, Springer, 2013. 5

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011. 6

[31] P. C. Mahalanobis, "On the generalized distance in statistics," National Institute of Science of India, 1936. 7

[32] I. Kufareva and R. Abagyan, "Methods of protein structure comparison," in *Homology Modeling*, pp. 231–257, Springer, 2011. 7