Comparison between HEMNMA-3D and Traditional Classification Techniques for Analyzing Biomolecular Continuous Shape Variability in Cryo Electron Subtomograms

Mohamad Harastani, Slavica Jonic

This is the author's version of an article accepted for publication in IEEE Proc. BioSMART 2021, https://doi.org/10.1109/BioSMART54244.2021.9677643

Comparison between HEMNMA-3D and Traditional Classification Techniques for Analyzing Biomolecular Continuous Shape Variability in Cryo Electron Subtomograms

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 cryo-ET is a bottleneck for comprehending biological mechanisms and developing drugs. Cryo-ET data suffer from a low signal-to-noise ratio and spatial anisotropies (missing wedge artefacts), making it 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 using normal mode analysis and 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 HEMNMA-3D and compares it with existing literature on a simulated dataset for nucleosome shape variability.**

***Index Terms*—Conformational Variability, HEMNMA-3D, Classification, Cryo-ET, Subtomograms**

I. Introduction

Three-dimensional (3D) volumetric images of vitrified cell sections can be obtained using cryogenic electron tomography (cryo-ET). The 3D nature of cryo-ET data allows studying macromolecular complexes despite the crowded cell environment. Cryo-ET data is obtained by acquiring multiple 2D projection images of a specimen around a single tilting axis inside the electron microscope (so-called tilt-series). The tilt-series are used to reconstruct a 3D volume (tomogram) based on the Fourier slice theorem. The tomogram of a cell section typically contains hundreds of copies of different macromolecules at random orientations. The copies of a macromolecule of interest can be identified and extracted into individual volumes called subtomograms.

Subtomograms suffer from a low signal-to-noise ratio (SNR) and spatial anisotropies, often observed as elongations along the beam axis, blurring and distracting caustics, known as missing wedge artifacts, due to the limitation in the tilting range (usually limited to ±60°), which corresponds to a missing wedge-shaped region in 3D Fourier space.

Due to the difficulties mentioned above, cryo-ET data processing is mainly based on rigid-body aligning and averaging many subtomograms to enhance the data quality and reveal the target macromolecular structure, which is known as Subtomogram Averaging (StA) [1]. Observing continuous shape variability using cryo-ET for *in situ* biomolecules is largely overlooked due to the lack of data analysis methods.

HEMNMA-3D [2] was recently introduced for cryo-ET continuous macromolecular shape variability analysis, inspired by HEMNMA, a method from 2D image analysis [3], [4]. HEMNMA-3D is based on elastic and rigid-body 3D registration between simulated shapes and cryo-ET data based on normal mode analysis (NMA) and fast rotational matching (FRM) with missing wedge compensation. HEMNMA-3D provides visual insights 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.

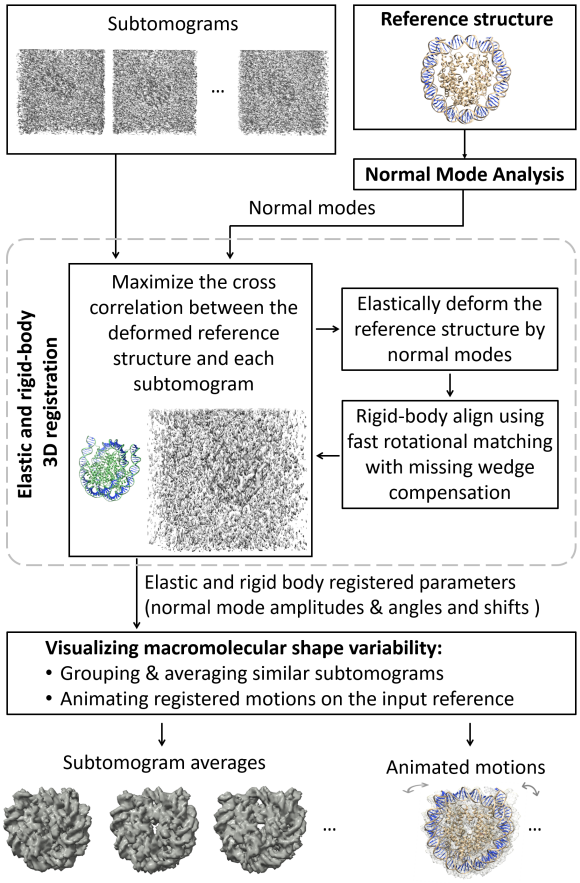

Fig. 1: Flowchart of HEMNMA-3D.

II. Methods

A. Cryo-ET StA and classification

Methods reported to deal with cryo-ET data heterogeneity are based on classification and rigid-body alignment. The most common classification techniques are known as post-alignment classification approaches [5].

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 aligned subtomograms are then classified based on the covariance matrix calculated using pair-wise constrained correlation coefficient (CCC - constraining the cross-correlation evaluation to the Fourier-space region that excludes the missing wedge). 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.

B. HEMNMA-3D

HEMNMA-3D (flowchart shown in Figure 1) takes as input a **reference structure** in Protein Data Bank (PDB) format. The PDB file contains 3D Cartesian coordinates (point cloud) of atoms or pseudoatoms. HEMNMA-3D applies Normal Mode Analysis (NMA) to the reference structure (NMA is a method for molecular mechanics simulation). Different conformations are represented by a linear combination of normal modes amplitudes. [2].

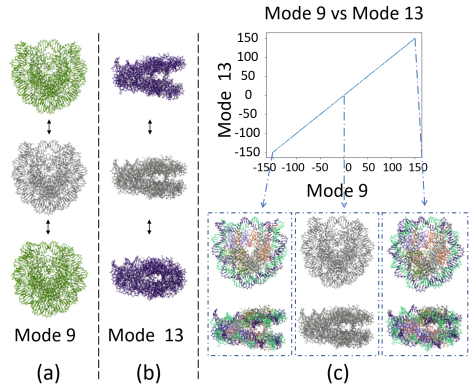

Fig. 2: Synthesized combined breathing and gaping motions of the nucleosome (PDB 3w98 structure): (a) nucleosome breathing motion, (b) nucleosome gaping motion, (c) ground-truth conformational distribution of 1000 points.

HEMNMA-3D performs **elastic and rigid-body 3D registration** using 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. A numerical optimizer maximizes the similarity between the subtomogram and a density volume simulated from the elastically deformed, oriented, and shifted atomic or pseudoatomic reference. The similarity measure is the CCC, which compensates for the subtomogram's missing wedge.

Once the registration is done for all the subtomograms, HEMNMA-3D allows **macromolecular shape variability visualization** by processing the conformational space. The elastic registration parameters (normal-mode amplitudes) are projected onto a lower-dimensional space using a dimensionality reduction technique, e.g., Principal Component Analysis (PCA). In this conformational space, each point represents a subtomogram, and close points correspond to similar registered shapes, which allows: 1) grouping close points in dense regions (similar shapes) and averaging the corresponding aligned subtomograms; and 2) animating the registered motions on the reference structure by fitting curves through data distribution manifolds.

III. Experiment

A. 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 [6], [7], with a linear relationship between normal modes amplitudes 9 and 13 so that the nucleosome is simultaneously breathing and gaping. An illustration of the simulated movements is provided in Figure 2.

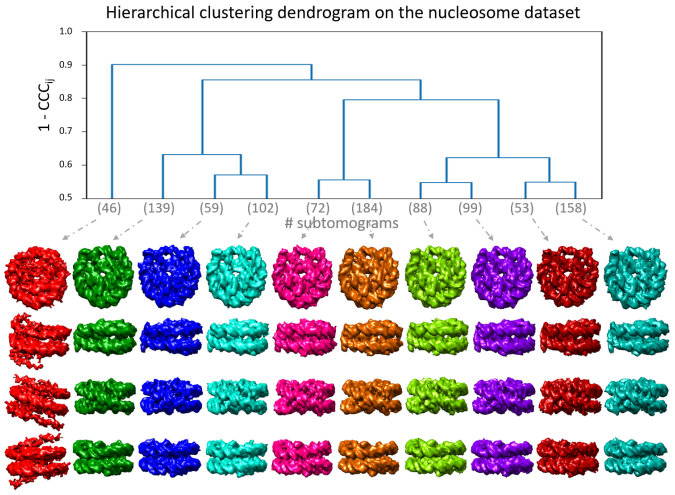

Fig. 3: Hierarchical clustering applied to the nucleosome dataset.

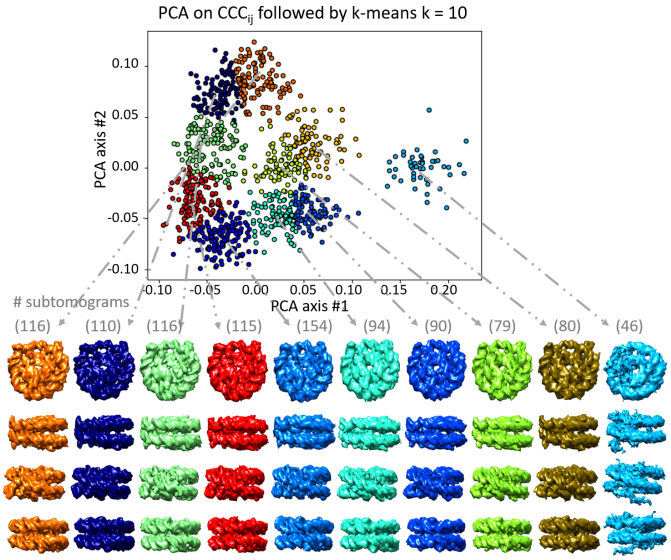

Fig. 4: K-means clustering applied to the nucleosome dataset.

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

1) 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 deformed structure to a density map of size 64^3 voxels (voxel size: (3.45 Å)^3).
3) Rotate and shift the volume in 3D space using random Euler angles and random x, y, z shifts (±5 voxels).
4) Tilt and project the volume, using the tilt angle from -60° to +60°.
5) Simulate microscope conditions by adding heavy noise (SNR = 0.01) and modulating the images with the microscope's contrast transfer function (CTF) using the defocus of -0.5 µm.
6) Invert the CTF phase.
7) Reconstruct a subtomogram from the obtained tilt series using a Fourier Central Slice method.

B. Traditional StA and post alignment classification

StA provides a global average without considering the shape variability, and it provides a basis for performing classification on subtomograms.

We applied StA on the synthesized nucleosome dataset, using the protocol based on the rigid-body alignment approach of [8]. This StA protocol uses an exhaustive angular search (with FRM method), a shifts search within a region of interest, and compensates for the missing wedge using the CCC.

After StA, we applied the obtained rigid-body alignment parameters on the subtomograms, and we evaluated the covariance matrix CCC$_{ij}$ of pairwise CCC. We performed the two most common post-alignment classification techniques on the CCC$_{ij}$ matrix, namely hierarchical clustering [9] and PCA followed by k-means [5].

The hierarchical clustering on 1-CCC$_{ij}$ matrix was performed to 10 classes using the Agglomerative Clustering module of Python Scikit-Learn package [10]. The clustering tree (dendrogram) and class averages are shown in Figure 3.

The k-means clustering was performed following PCA 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. Figure 4 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 3 and 4).

C. 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.

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). We used HEMNMA-3D open-source software, a plugin called ContinuousFlex [2], [3] for Scipion [11].

Figure 5 (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.

Figure 5 (b) shows the displacement of the reference structure along 10 points in the direction of the point distribution in the conformational space.

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

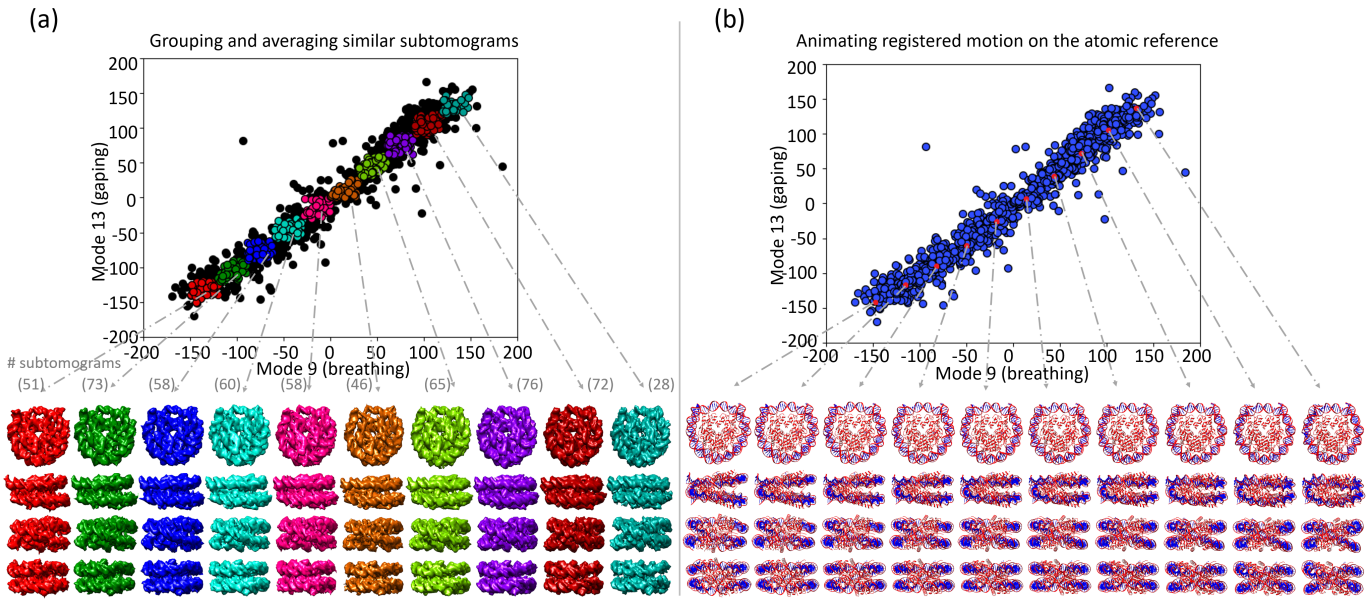

Fig. 5: HEMNMA-3D applied to nucleosome dataset. (a) group averages for ten equally distanced groups in the conformational space, (b) displacement of the reference along the direction of the data distribution (10 frames represented by red dots).

IV. Discussion and conclusion

This article presented a comparative review on HEMNMA-3D with traditional classification methods for analyzing macromolecular shape variability captured in cryo-ET. The comparison was performed on a dataset of nucleosome shape variability under simulated microscope conditions and testing the capabilities of the methods to recover the ground-truth shapes.

The test results indicate that HEMNMA-3D recovers the ground-truth combination of two nucleosome motions (breathing and gaping). Two state-of-the-art methods for cryo-ET classification were applied and showed different discrete class averages of the nucleosome at different breathing and gaping magnitudes. However, the shape transitions between the obtained class averages are ambiguous due to the continuous nature of the shape variability.

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 [2].

V. Acknowledgements

We acknowledge the support of the French National Research Agency - ANR (ANR-20-CE11-0020-03 and ANR-19-CE11-0008-01 to S.J.); the Sorbonne University (2019 ”Interface pour le Vivant” PhD scholarship grant to M.H.); and the access to the HPC resources of CINES and IDRIS granted by GENCI (A0100710998, A0070710998, AP010712190, AD011012188 to S.J.).

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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] M. Eltsov, D. Grewe, N. Lemercier, A. Frangakis, F. Livolant, and A. Leforestier, “Nucleosome conformational variability in solution and in interphase nuclei evidenced by cryo-electron microscopy of vitreous sections,” *Nucleic acids research*, vol. 46, no. 17, pp. 9189–9200, 2018.

[8] 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.

[9] 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.

[10] 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.

[11] 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.