JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

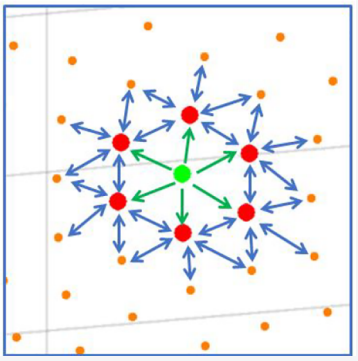
Propagation of Conformational Coordinates Across Angular Space in Mapping the Continuum of States from Cryo-EM Data by Manifold Embedding

Suvrajit Maji, Hstau Liao, Ali Dashti, Ghoncheh Mashayekhi, Abbas Ourmazd, and Joachim Frank*

Cite This: *J. Chem. Inf. Model.* 2020, 60, 2484−2491 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Recent approaches to the study of biological molecules employ manifold learning to single-particle cryo-EM data sets to map the continuum of states of a molecule into a low-dimensional space spanned by eigenvectors or "conformational coordinates". This is done separately for each projection direction (PD) on an angular grid. One important step in deriving a consolidated map of occupancies, from which the free energy landscape of the molecule can be derived, is to propagate the conformational coordinates from a given choice of "anchor PD" across the entire angular space. Even when one eigenvector dominates, its sign might invert from one PD to the next. The propagation of the second eigenvector is particularly challenging when eigenvalues of the second and third eigenvector are closely matched, leading to occasional inversions in their ranking as we move across the angular grid. In the absence of a computational approach, this propagation across the angular space has been done thus far "by hand" using visual clues, thus greatly limiting the general use of the technique. In this work we have developed a method that is able to solve the propagation problem computationally, by using optical flow and a probabilistic graphical model. We demonstrate its utility by selected examples.



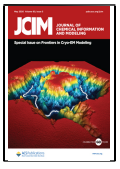
1. INTRODUCTION

Recent approaches exploit the information contained in large single-particle cryo-EM data sets of a molecular machine to map its continuum of states in thermal equilibrium.[1,2] The free energy landscape derived from such a mapping provides a rich source of information to study the molecule's functional dynamics. A manifold embedding-based technique recently introduced[3−7] enables the quantitative study of continuous conformational motions over the energy landscape. Given a projection direction (PD) of interest, cryo-EM molecule images falling into a small angular range forming a "cone" around that PD are considered. This cone is narrow enough, so that the image variability due to changes in orientation is much smaller than that due to conformational and compositional heterogeneity. These images form a cloud of points in the high-dimensional space of pixels and are arranged according to their mutual similarities, reflecting the conformational states occupied by molecules in this view range. This point cloud may be considered as a hypersurface or "manifold"—a topological space in which a local Euclidean geometry can be defined at any point on the manifold.[8] "Manifold embedding"[9,10] maps the unordered data points lying in the high-dimensional space into a low-dimensional manifold representation of the data set, such that a certain form of geometric relationships between the data points is preserved. In our case,[3−7,11,12] the embedding technique describes a manifold of images containing the information on conformational changes, by means of Euclidean coordinates found via a diagonalization procedure. This low-dimensional description of the manifold is given by a set of eigenfunctions which is generated by operating on a nonlinear manifold embedded from the original high-dimensional image space using diffusion maps.[13]

The conformational changes reflected in this manifold can be "decomposed" along each of these coordinates, via the Nonlinear Laplacian Spectral Analysis (NLSA).[6,11] The first few (usually less than ten) highest-ranking eigenfunctions, as determined by the spectrum of eigenvalues, are retained, and a subset of these is chosen as *conformational coordinates*. The number of conformational coordinates, which is based on the eigenspectrum, is related to the manifold type and its intrinsic dimensionality, which in turn reflects the degrees of freedom exercised by the system. The NLSA technique produces a sequence of images, or a *movie*, showing the conformational changes along any path on the manifold. Next, the low-dimensional representations of the manifolds in different PDs must be related to one another, such that a consolidated map of occupancies describing the continuously varying conformations can be generated. From this consolidated map the free energy landscape of the molecule is subsequently derived

Special Issue: Frontiers in Cryo-EM Modeling

Received: December 2, 2019
Published: March 24, 2020



https://dx.doi.org/10.1021/acs.jcim.9b01115
J. Chem. Inf. Model. 2020, 60, 2484−2491

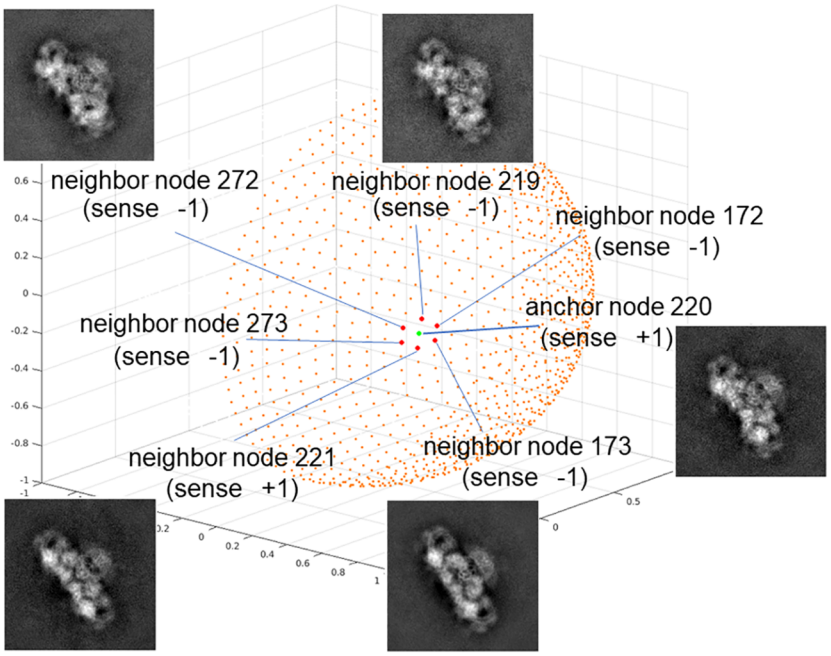


Figure 1. Propagation of conformational coordinates for the RyR data on the orientation sphere. The orange points represent the projection directions (hemisphere shown here). The fat green point is one of the selected anchor nodes, and the fat red points are its neighbors. As a representative example, PD 220 is selected as an anchor node, and PDs 219, 221, 172, 173, 272, 273 are its neighbors. One snapshot of the movie corresponding to the first eigenfunction of the anchor node 220 is shown here, and the sense of the movie is assigned as +1. The corresponding movies in the six neighboring nodes are selected such that they have the same type of motion as the movie for green node. The movies for the anchor node and some of the neighboring nodes are shown in Movies S1−S5. The sense for the movie in the neighbor node is denoted as +1 if it has the same direction as the anchor node, otherwise it is −1. A more illustrative graphical representation of the propagation is shown in the Supplementary Figure S1.

through the Boltzmann relationship. One of the main assumptions is that the same conformational spectrum is viewed in all the PDs—in other words, the conformational variability of a molecule is independent of its orientation on the EM grid. Identifying the same conformational motion across the entire angular grid is a difficult optimization problem, but it becomes more tractable when we propagate across the angular grid using the information from a limited number of neighboring PDs. Hence, we aim to determine the correspondence between the conformational coordinates (and their respective conformational movies) of neighboring PDs.

The difficulty in finding the correspondence is that the order (ranking) and the sense of the conformational coordinates in one PD may be different from those in another PD. The former difference is due to the fact the conformational change of a given coordinate is easier to observe at certain viewing angles than at others. The latter difference is due to the sign ambiguity in standard eigen-decomposition solvers:[14] a movie reverses its directionality or sense when the corresponding coordinate flips its sign. The task of propagation of coordinates is to match coordinates in adjacent (and also, by inference, farther located) PDs, such that they express the same or closely similar motions with matching directionality. A successful outcome of this matching procedure will ensure that we are summing over corresponding conformational occupancies from all PDs, from which the consolidated occupancy map and, ultimately, the energy landscape of the molecule may be obtained.

In this paper we present an approach to automate the propagation of conformational coordinates across all PDs, in contrast to the current manual-selection approach, which entails user inspection of the movies for each PD and visual comparison of movies in adjacent PDs. Our automated method requires the labeling of the movies in only a few PDs (*anchors*), and then the information is propagated across the remaining PDs. The propagation process is formulated as an optimization problem, where the "closeness" of the conformational changes in the movies is maximized. We use a feature-extraction technique[15,16] to encode the conformational change and then employ *graphical models*[17] to propagate this information across angular space. The details of the individual techniques are provided in the Methods section.

2. METHODS

2.1. Overall Approach. The problem we seek to solve is the propagation of the conformational information from a few anchors to all other PDs on the orientation sphere S^2, as shown in Figure 1. Our approach is based on exploiting the distinguishing features in a movie of a macromolecular structure undergoing a certain motion. We use optical flow[15,18] to compute the motion vector for a movie. It is a widely used method for computing the approximate motion of image structures using local gradients (Supplementary Text S1). Some of the popular optical flow algorithms are Lucas-Kanade (LK),[19] Horn-Shunck (HS),[15] and Gunner Farneback (GF).[20] The LK method produces local and sparse flow vectors, whereas the HS and GF methods generate global and dense flow vectors, which we use here. Next we use the technique of Histogram of Oriented Gradients[16] (HOG) to obtain the characteristic features of the movies. HOG is a popular method in computer vision to detect objects in images. HOG has also been used as a morphological feature shape

https://dx.doi.org/10.1021/acs.jcim.9b01115
J. Chem. Inf. Model. 2020, 60, 2484−2491

descriptor for videos with moving objects in conjunction with optical flow.[21] The HOG feature vector encodes the type of motion by registering in each movie the position, magnitude, and direction of the optical flow vectors. Some variants of this approach have been applied to human motion analysis by computing histogram of optical flow (HOF)[22] and oriented histograms of differential optical flow.[23] To discriminate two movies s and t, we estimate the closeness measure HD_{st} between the two HOG feature vectors (Supplementary Text S2) H_s and H_t, given by the l_p-norm of their difference as

$$HD_{st} = \|H_s - H_t\|_p, \quad p = 1, 2 \tag{1}$$

If the number of states of a node is S_K (see Section 2.2), then the $S_K \times S_K$ matrix of pairwise compatibility measurements between the movies in nodes i and j is given by

$$(HD)_{ij} = \{HD_{s_i t_j}\}_{s_i=1\ldots S_K, t_j=1\ldots S_K} \tag{2}$$

We then set up the optimization problem of selecting the correct movie in each PD in the form of a *graphical model*[17] (Supplementary Text S3), where the PDs are the nodes and the relationship between two nodes is given by the closeness measure (eq 1). To solve this optimization problem, we have used a *message-passing* method called *belief propagation*,[24−26] which is a type of dynamic programming algorithm typically used to solve *inference* problems in graphical models. For our work, we will focus on belief propagation (BP) applied to graphs with loops, also known as loopy belief propagation.[27−29] The belief propagation works by evaluating the information on a node and the relationship with the neighboring nodes, which are provided in the form of node and edge potential functions (Supplementary Text S3), and then iteratively propagating the information throughout the graph (Supplementary Text S3.1.1). We compute the node potentials ϕ and edge potentials ψ as follows

$$\phi_i(x_i) = \begin{cases} e^a, & \text{if node } i \text{ is an anchor} \\ e^b,\ a > b, & \text{otherwise} \end{cases} \tag{3}$$

$$\psi_{ij}(x_i, x_j) = e^{-(HD)_{ij}} \tag{4}$$

where x_i and x_j are the states of nodes i and j, respectively (Text S3), and $(HD)_{ij}$ is calculated using eqs 1 and 2.

After the belief propagation has converged to within a tolerance level, or stopped after *maxIter* iterations (Supplementary Text S3.1.1), we can obtain the beliefs for each node using step (v) in Supplementary Text S3.1.1. The beliefs obtained using the sum-product are the marginal probabilities. We then select the state of the nodes with the highest marginal probability estimates. For the max-product algorithm, the beliefs are not marginal probabilities, but the maxima of the beliefs will give us the most likely state configuration and thus approximate the set of *maximum a posteriori* (MAP) probabilities for all nodes. The BP calculations thus provide us with the optimal states of the nodes for the entire graphical model, given the appropriate node and edge potential values.

2.2. Formulating the Selection of Conformational Coordinate as an Optimization Problem. We pose the optimization problem of selecting the conformational coordinates as finding the solution to inference problems of a graphical model. The graph nodes are the projection directions (PDs), and the neighbors of a node are the PDs that are within a certain distance ϵ on the orientation sphere S^2. Let *minDist* be the minimum Euclidean distance between any two nodes on S^2, nG is the total number of tessellated bins on S^2 (orange dots in Figure 1), and *numPDs* is the total number of PDs for a data set, then the distance threshold ϵ is calculated as

$$\epsilon = \min\{\max(minDist, \epsilon_{ball}), \text{min}Dist \times 2\sqrt{2}\}$$
$$\text{where } \epsilon_{ball} = minDist \times (nG/numPDs) \tag{5}$$

The ϵ_{ball} and ϵ values can be adjusted as necessary. We chose the values such that the number of neighbors for a node is around 10 or less (generally around 6), so that the viewing angles of the immediate neighbors are close to each other and the computational cost is reasonable. The nodes of the graph can assume any of the states *state* $s = 1, \ldots, S_K$ depending on the number of eigenfunctions, K. Now, to convert this graph of PDs into a probabilistic graphical model (Section 3), we denote the state of the nodes with random variables $x_1, x_2, \ldots, x_N$, where N is the number of nodes (Text S3). In our case $S_K = 2K$, as we select the proper movie out of K candidates, each of which with "sense" equal to either +1 or −1. The problem can be better understood from Figure 1, where the green point (node 220) is an anchor node. The red nodes are the immediate neighbors of the anchor node. The first K states represent the K movies in the "forward" direction (sense +1) and the last K states for the same K movies in the "backward" direction (sense −1). The task is to determine which one among the $2K$ movies represents the proper eigenfunction and sense in each PD. The optical flow vectors for the K movies for each PD are computed according to the procedure outlined in Supplementary Text S1, and the pairwise edge measurement values are estimated using the HOG feature distance described in Supplementary Text S2. The corresponding edge potential functions ψ_{ij} are calculated using eq 4. The node potential functions ϕ_i, given by eq 3, are set to be uniform priors except for the nodes chosen as "anchor PDs", in which case the node potential values are set with relatively high values. Then we perform the graphical model inference with belief propagation as described in Supplementary Text S3.1 and Supplementary Text S3.1.1 of Supplementary Text S3. The iterative propagation of the information starting from the anchor PDs to all other PDs in the graphical model occurs through the immediate neighbors. This ensures that the change in the features between immediate neighbors is small and, subsequently, that the message passaging between the neighbors is as accurate as possible.

The essential steps of the ManifoldEM[6,12] approach followed by our propagation method are summarized as follows:

1. The first step of ManifoldEM is to create a tessellation of the orientation sphere such that the 2D cryo-EM images are classified into bins known as projection directions (PDs).
2. The images in each PD are aligned, and all the pairwise Euclidean distances are calculated for the images.
3. Using the distances, the images in each PD are embedded into a low K-dimensional manifold described by K eigenvectors using diffusion maps. The manifold of images contains the information on conformational changes along each of the eigenvectors.
4. The conformational information contained in this manifold is now decomposed and sorted along each of the K eigenvectors in the form of conformational

https://dx.doi.org/10.1021/acs.jcim.9b01115
J. Chem. Inf. Model. 2020, 60, 2484−2491

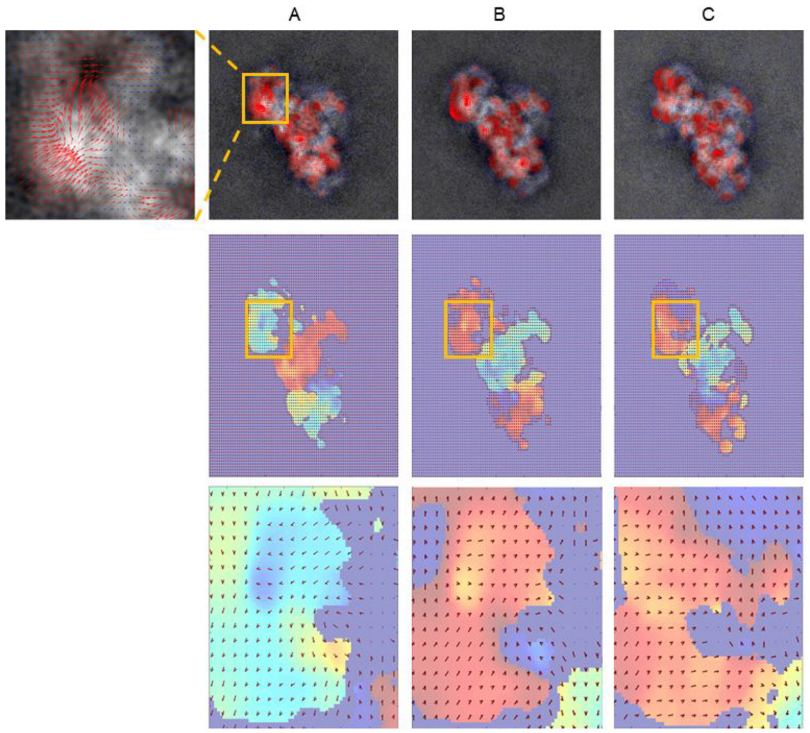


Figure 2. Optical flow and HOG feature calculations for the RyR data set. Column A represents the first eigenfunction of PD 220, column B represents the first eigenfunction of PD 220 with reverse sign, and column C represents the first eigenfunction of PD 219. First row: optical flow vectors for the movies corresponding to the respective eigenfunctions for the PDs. The zoomed-out box shows the optical flow vectors for the region inside the yellow box on the first image in column A. The blue vectors are the optical flow vectors computed using the Horn-Shunck method, and the red ones are the selected optical flow vectors with magnitude above the 85th percentile, for the movie. Second row: *rose plot* visualization of the HOG feature vectors, superimposed on an orientation heat map of the optical flow vectors. Third row: zoomed-in detailed view of the region within the yellow boxes in the second row. We can see that the HOG feature vector plot for column C is more similar to column B than to column A.

movies, using the Nonlinear Laplacian Spectral Analysis. Due to the sign ambiguity between the same eigenvectors in different PDs, we have forward and reverse directions, so there are $2K$ movies to compare in each PD.

5. At this point our propagation method steps in. We now compute the dense optical flow field using the Horn-Schunck method for each movie and then compute the histogram of oriented gradients (HOG) from the optical flow vectors. The HOG feature vector encodes the motion present in the movie.

6. To compare two movies we compute the Euclidean distance between their HOG feature vectors. Thus, to compare all pairwise combinations of movies in projection directions i and j, we obtain a $2K \times 2K$ HOG feature distance matrix $(HD)_{ij}$.

7. We then formulate the conformational coordinate selection process as an optimization problem using a graphical model, defined over the angular grid. The PDs are the nodes of the graph, and two nodes are connected by an edge if they are within a certain distance ϵ.

8. Next, the edge potentials between node i and node j are calculated using the HOG distance matrices $(HD)_{ij}$. Nodes of the graph can assume any one of the $2K$ states corresponding to the K forward and K reverse direction movies.

9. The goal is to find a consensus labeling across the entire angular grid using our propagation algorithm by comparing movies in neighboring nodes. We achieve this goal by using belief propagation, which tries to determine the probability (belief) of a node being in a particular state in an iterative manner, given the reference movies in a few PDs called "anchors". To obtain the best state label we compute the maxima of the beliefs.

3. RESULTS AND DISCUSSION

We tested our method with three data sets, two experimental cryo-EM data sets, ryanodine receptor (RyR),[12,30] and *E. coli* ribosome (Acosta-Reyes, F. J.; Holm, M.; Sanyal, S.; Frank, J., to be published elsewhere), which were analyzed with the manifold embedding method,[6,12] and one synthetic data set. The labels of the conformational movies were obtained by visual inspection, which were then compared against the labels obtained from our automated selection. The RyR data predominantly showed a combination of wing motions and channel opening vs closing type of motion of the molecule in most PDs (Movies S1−S5). The total number of PDs for this data set is 1,117, with 5 movies for each PD, and 9 of those PDs were randomly designated as anchors. For the particular settings of ϵ_{ball} and ϵ in eq 5, the number of edges in the graph was 2,769. It is evident from the eigenvalue spectra that there is mainly one dominant eigenfunction in most PDs. We denote the eigenfunctions by γ_K. Figure S4 shows the eigenvalue spectrum for several PDs, including the anchor PD 220. Movie S1 is the conformational movie for γ_1 of PD 220, and it shows a clear wing motion approximately along the direction of the

https://dx.doi.org/10.1021/acs.jcim.9b01115
J. Chem. Inf. Model. 2020, 60, 2484−2491

vertical axis (Figure S5). We will denote this as the *vertical wing motion*. Movie S2 is the conformational movie for γ_1 of PD 220, and it shows a subtle motion of the outer part of the wings roughly along the direction of the lateral axes (Figure S5). We will refer to this as the *lateral wing motion*. We chose to propagate the most dominant vertical wing motion (conformational coordinate 1; Movie S1) and the less dominant lateral wing motion (conformational coordinate 2; Movies S2 and S5), across all PDs. Both types of motion involve channel opening−closing as seen from top view (e.g., Movie S8) and are not easily distinguishable. At first we computed the optical flow vectors for each movie in every PD using the Matlab function *opticalflowHS*, where the smoothness parameter was set to 1.5 and the number of iterations for solving the numerical scheme was set to 200. In our application, the movies are first averaged over a block of five frames, and then optical flow vectors are computed between successive pairs of averaged frames. The final optical flow vectors for the entire movie are obtained by adding up those intermediate sets of vectors. The results of the optical flow calculations for the RyR data set are shown in the first row of Figure 2. The optical flow vectors are shown in blue, and the red ones are the selected vectors with magnitude above the 85th percentile. For HOG feature calculations we used the Matlab function *extractHOGFeatures*. The parameter values used for the HOG feature calculations are as follows:

$$CellSize = [4, 4],\ BlockSize = [2, 2],\ BlockOverlap = ceil\left(\frac{BlockSize}{2}\right),$$

$$NumBins = 9,\ UseSignedOrientation = 1$$

The top row of Figure 2 represents the movies for eigenfunction candidate γ_1 of PD 220 in the forward direction, PD 219 in the forward direction, and PD 219 in the backward direction, respectively. The second and third rows of Figure 2 show the corresponding HOG feature vectors, which are calculated according to Supplementary Text S2. The HOG feature vectors are visualized using a *rose plot*, and they are overlaid on the orientation heat map of the flow vectors. It is evident from the HOG rose plots and the orientation heat map that γ_1 of PD 220 in the forward direction matches better with γ_1 of PD 219 in the backward direction than in the forward direction. Hence, if γ_1 for the anchor PD 220 has *sense* +1, then γ_1 of PD 219 has *sense* −1, which is shown in Figure 1. The corresponding movies can be seen in Movies S1 and S3.

We calculated the l_2-norm of the HOG feature vector difference according to eqs 1 and 2, which were then used as input to the edge potential function in eq 4 (step (i) of Supplementary Text S3.1.1). We implemented the standard loopy belief propagation[27,28] with a damping factor for message updates (Supplementary Text S3.1.1). For the first type of motion, we compared our result with the manual selection, which excluded the assignment of labels to movies for 108 PDs that were corrupted, or where motions of prominent features were occluded. Those unassigned 108 PDs were found to be spread out on the orientation sphere and were included in the propagation but not in the accuracy calculations. For the particular choice of optical flow parameters and HOG feature measurement as mentioned above, the sum-product BP produced an accuracy of 95.7% in less than 200 iterations and with a tolerance level of 1×10^{-4} (Text S3.1.1). The max-product BP for the same settings produced an accuracy of 95.6%. We performed a series of 100 trials by randomly sampling 10 anchors and performed the belief propagation for each trial. The accuracy measurements are provided in Table 1 (CC1). We report the mean and standard deviation (SD) and also the median, mean absolute deviation (MAD-mean), and median absolute deviation (MAD-median) as they provide a more robust measure of variability. The accuracy for the sum-product and max-product BP remained virtually the same when we ran the BP trials for the first type of motion, with fewer than 10 PDs, even up to 3 or 2 PDs as anchors.

Table 1. RyR Conformational Coordinate (CC) Propagation Accuracy Measurements with 100 Random Trials[a]

propagation method	no. of anchors	mean (%)	median (%)	SD (%)	MAD (mean) (%)	MAD (median) (%)
CC1						
sum-product BP	10	95	96	2	1	<0.5
max-product BP	10	95	96	1	<0.5	<0.5
CC2	set of 632 PDs from which anchors were drawn					
sum-product BP	15	61	62	5	4	3
	30	65	65	3	2	2
CC2	set of 403 remaining PDs from which anchors were drawn					
sum-product BP	15	52	53	6	5	5
	30	57	57	5	4	3

[a]CC1 values are for the first type of motion, and CC2 values are for the second type of motion.

For the second motion, we manually created labels for the possible choices of eigenvectors associated with this motion. There were 82 PDs that were unassigned (for the same reason as stated for the first motion), and for a particular choice of 30 anchor PDs, we obtained an accuracy of 71.5% with the sum-product BP and 70.3% with the max-product BP. Among the 100 trials that we performed for the second motion (Table 1, CC2), there were a few trials which produced accuracy ≥70%, and the above measures are from one such trial. The typical accuracy measurements for about 30 anchors were ∼65% (Table 1, CC2). The max-product BP did not converge within 200 iterations under the same settings. From the same set of anchor PDs as above, when we select a subset of 20 and 15 anchors, the accuracy was 67.8% and 64.7%, respectively, for the sum-product. On the other hand, the max-product BP did not converge within 200 iterations and produced an accuracy of only 68.8% and 63.2%, respectively. Most of the errors in the propagation method were in the selection of movies which had orientational artifacts (e.g., Movie S6) and exhibited only partial similarity to the actual motion (e.g., Movie S5). For different combinations of same-size subsets of anchor PDs, we obtained accuracy measurements which were significantly different in some cases (Table 1, CC2). The number and choice of anchor PDs seem to be much more important for propagating the second motion, as it is subtler in most PDs compared to the first type of motion. Therefore, to obtain a sense for the degree of variability in the accuracy measurements for the propagation of the second eigenvector, we performed sets of 100 trials by randomly sampling 15 and 30 PDs as anchors for the second motion. We performed the above experiments separately on two pools of anchors, one with 632 PDs, where the second motion is comparatively more prominent, and another set with the remaining 403 PDs, which excluded the 82 unassigned PDs. As expected, the first

https://dx.doi.org/10.1021/acs.jcim.9b01115
J. Chem. Inf. Model. 2020, 60, 2484−2491

experiment with a relatively "good" pool of 632 anchors produced better accuracy than the second. We are providing more details regarding the accuracy measurement for the RyR data set, as an example case for analyzing the performance of the method and also because it has two types conformational motions for propagation. We only report the sum-product BP values in Table 1, CC2. The values for the max-product accuracy were generally ∼2% less than the sum-product values. We should note here again that there were five eigenfunctions, each with sign +1 or −1, so there are a total of ten choices in each PD to be considered as candidates for selection of the conformational coordinate. However, for the selection of CC2 in each PD, we set the node potential value in the BP step to a very low value for the eigenfunction selected as CC1. This step is usually not needed if the second motion is very different from the first type, but it could be useful when the two motions share some similarity in many PDs. This is in fact the case for the two types of RyR motion in many side views (e.g., Movies S5 and S7) and all top views (Movie S8). All the measurements for conformational coordinate selections are provided in Table 1.

We compared the 2D occupancy maps (Figure 3) for the RyR data set by combining the first and second eigenfunctions

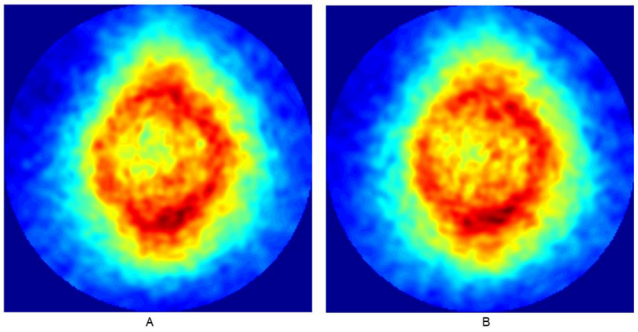


Figure 3. Occupancy map for the RyR data set. Column A is for the manually selected eigenfunctions. Column B is for automatically selected eigenfunctions using our propagation method. Top row shows the occupancy map (OM). We can see that the OM for A and B match closely overall. The characteristic differences in the contours of A and B are primarily due to the differences in the occupancies (computed by ManifoldEM) of the conformational states as present in the second eigenfunctions that were selected from the multiple choices in each PD.

from all PDs. The occupancy maps (OMs) generated from labels selected manually and by our propagation method are seen to match closely. To estimate how similar the two resulting occupancy maps are, we computed the relative error with the Frobenius matrix norm[31] also called the normalized root mean squared error (NRMSE)

$$\mathrm{NRMSE} = \frac{\|\hat{A} - A\|_F}{\|A\|_F} \cdot 100\% \tag{6}$$

where $\hat{A}$ is the approximation of the matrix A.

In our case $\hat{A}$ is the occupancy map using labels obtained by our propagation method, A is the occupancy map using the manual labels, and the NRMSE value is 9.5%. Besides the difference in the selection of the eigenfunctions (in this case the second one) from multiple choices in each PD, the finer characteristic variation in the OM contours in Figure 3A and 3B is due to the differences in the computed occupancies (by *ManifoldEM*) of similar conformational states represented by the different eigenfunctions and therefore not a direct problem of our propagation method.

The time taken for the optical flow calculation and pairwise HOG feature vector calculation for the RyR data set with movie frame size of 336 × 336 pixels was about 4 h with 32 parallel jobs. The computer specifications are as follows: 64 processors with speed 1200 MHz with a maximum of 2600 MHz, memory of 250 GB. Down-sampling the images can obviously make the computations faster. The BP step, when it converges, takes between ten and several tens of seconds.

The second example tested was the ribosome data set, which has 505 PDs with 5 movies for each PD, and 7 of those PDs were designated as anchors. The most dominant motion (Figure S6) for this data set is the "intersubunit rotation", also known as small subunit (SSU) "ratchet-like motion" (Movies S9 and S10), and very few PDs contained evidence for the small subunit's "rolling motion"[34] and "head rotation"[35]. We chose to propagate the intersubunit rotation motion in this case. For the optical flow calculations (Figure S7) the smoothness parameter was 1.5, and the number of iterations was set to 200. For the HOG feature calculations (Figures S9 and S10) we used the following parameters: *CellSize* = [8,8], *BlockSize* = [4,4], *NumBins* = 9. There are 11 PDs which had corrupted movies, and those were excluded from accuracy calculations. The accuracy of the propagation for this data set was 98.6% for the sum-product and 98.4% for the max-product BP for these settings and feature calculations.

In addition, we also tested the method with a simple synthetic data set of 2D movies (fictitious eigenfunctions) with known labels within a region of 200 PDs of the angular sphere (Figure S8). (We chose *not* take the route of creating a stack of 2D cryo-EM projection images and applying manifoldEM to obtain eigenfunctions in each PD, because in that case we would have to manually obtain the labels). The data set was derived from the structure of splicing factor TIA-1 (pdb 2mjn), a protein with two symmetrically placed domains joined by a linker. We morphed one domain into different positions to represent its stepwise movement (Movies S11, S12, S14, and S15), creating a total of 21 volumes ("states"). From these volumes, a stack of 21 projections was created for each PD, to simulate the movie associated with fictitious eigenfunction 1 (Movie S11). To simulate the movie along the second eigenfunction, we rotated the first movie by 180°, as if the opposite domain of the protein were moving (Movie S12). To simulate the movie along the third eigenfunction (Movie S13), we added a substantial amount of noise to the first movie, such that we barely observe any change of the structure. Next we reversed the directionality and also shuffled the order of the first two movies in some PDs, specifically every fourth PD starting with the first PD. This mimics the variable order (ranking) and sign ambiguity of eigenfunctions[14] encountered in the real situations. In this way we created a test data set with known ground truth labels. We used two anchor nodes for the propagation. For the optical flow calculations (Figures S9 and S10), the smoothness parameter was chosen as 1.5, and the number of iterations was set to 200. For the HOG feature calculations (Figures S9 and S10) we used the following parameters: *CellSize* = [4,4], *BlockSize*= [2,2], *NumBins* = 9. The accuracy of the propagation of the first fictitious eigenfunction for the synthetic data set was 100.0% for the sum-product and max-product BP for these settings and feature calculations.

https://dx.doi.org/10.1021/acs.jcim.9b01115
J. Chem. Inf. Model. 2020, 60, 2484−2491

We also experimented with the method using an oriented histogram of differential optical flow vectors[23] which is typically used for analyzing movies, but the results were not better compared to the procedure we implemented using optical flow followed by HOG. It should be noted here that movies corresponding to some lower-ranked eigenfunction (i.e., with lower-ranked eigenvalue in the spectrum) can show a very similar conformational movie as the higher-ranked eigenfunction.[1] In such cases, the movies corresponding to the lower-ranked eigenfunction can get selected instead of the higher-ranked one depending on the optical flow and HOG feature values for such movies. In our work, the movies for the PDs that were not optimally selected in the experimental data sets were the ones which either had several corrupted frames, conformational changes that were small, or the PD represented a view where feature extraction was difficult.

While demonstrating robust performance for strongly articulated motions, the performance analysis provided above for different data sets also gives us an idea about the limitations of the proposed propagation method. In general, the propagation method has difficulty in matching neighboring views when they are relatively far apart on the angular grid, or if there are occlusions of moving components in many PDs. In addition, it can perform poorly for motions that are subtle as the noise affects the computation of smooth flow vectors depicting the actual motion. The performance can also be affected by the sensitivity of the discrimination metric used to compare the movies between different PDs.

4. CONCLUSIONS

In the applications of the ManifoldEM technique until now, the selection of the conformational coordinates across the angular sphere has been performed manually, by inspecting each movie individually, to generate the consolidated maps of occupancy on which the energy landscapes are based.[6,12] Therefore, it was imperative to automate the propagation of conformational coordinates, for the manifold embedding workflow to be general-purpose. The method presented thus provides a solution to this long-standing problem, in a relatively fast and efficient manner and with good accuracy. However, the task of propagating the conformational coordinate is challenging for data sets where the distribution of images on S^2 has poor orientational coverage (changes between neighboring PDs are not small anymore) and a significant number of PDs have low occupancies, or the conformational changes are subtle, and hence for all such cases the optical flow vectors might not be able to discriminate those changes appropriately against the background of nonrelevant pixel motions. For future work, it is worth mentioning that with improved feature extraction techniques, or better feature estimates with different optical flow and HOG features, the loopy belief propagation approach could potentially provide a better solution. Alternatively, one has to implement a fast and memory-efficient implementation of generalized belief propagation[32] or junction tree algorithm[33] for large graphs with loops as in our case. These generalized variants of the BP algorithm have better theoretical convergence performance with more accurate marginal and MAP probabilities compared to conventional loopy belief propagation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b01115.

Background description of optical flow, HOG feature extraction, and probabilistic graphical model with inference using belief propagation (PDF)

Movies S1−S16 (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Joachim Frank − *Department of Biochemistry and Molecular Biophysics and Department of Biological Sciences, Columbia University, New York, New York 10032, United States;* orcid.org/0000-0001-5449-6943; Email: jf2192@cumc.columbia.edu

Authors

Suvrajit Maji − *Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032, United States;* orcid.org/0000-0001-5547-1975

Hstau Liao − *Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032, United States*

Ali Dashti − *Department of Physics, University of Wisconsin Milwaukee, Milwaukee, Wisconsin 53211, United States*

Ghoncheh Mashayekhi − *Department of Physics, University of Wisconsin Milwaukee, Milwaukee, Wisconsin 53211, United States;* orcid.org/0000-0001-9891-190X

Abbas Ourmazd − *Department of Physics, University of Wisconsin Milwaukee, Milwaukee, Wisconsin 53211, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.9b01115

Author Contributions

S.M., H.L., and J.F. conceived the conformational coordinate propagation algorithm. S.M. implemented and applied the present conformational coordinate propagation algorithm and analyzed all the data sets for this work. J.F. supervised this work. A.D., G.M. under A.O.'s guidance, applied the ManifoldEM algorithm to the RyR data and made the files and relevant ManifoldEM Matlab software codes available. S.M., H.L., and J.F. interpreted the propagation data and wrote the manuscript. A.O. provided helpful comments.

Notes

The authors declare no competing financial interest.

Software Distribution: A python software package called *ManifoldEM* has been developed following the original Matlab code,[6,12] which will be released for beta testing. It will be released, in the near future, to the broader academic community to study the conformational dynamics of macromolecular machines using cryo-EM data. The proposed conformational-coordinate propagation method not only is included as a part of *ManifoldEM* but also can be downloaded separately from this link https://github.com/suvrajitm/ConformationalCoordPropagation.git.

■ ACKNOWLEDGMENTS

We would like to thank Francisco Acosta Reyes for providing us with the ribosome data set. We also thank the reviewers for their helpful comments and suggestions. The work has been

supported by NIH grant R01 GM 55440 and R01 GM 29169 (to J.F.)

REFERENCES

(1) Moscovich, A.; Amit, H.; Joakim, A.; Amit, S. Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes. *Inverse Problems* **2020**, *36*, 024003.

(2) Zhong, E. D.; Bepler, T.; Davis, J. H.; Berger, B., Reconstructing continuously heterogeneous structures from single particle cryo-EM with deep generative models. *CoRR*. arXiv:1909.05215. 2019. https://arxiv.org/abs/1909.05215 (accessed 2020-03.27).

(3) Schwander, P.; Fung, R.; Phillips, G. N.; Ourmazd, A. Mapping the conformations of biological assemblies. *New J. Phys.* **2010**, *12* (12), 035007.

(4) Schwander, P.; Fung, R.; Ourmazd, A. Conformations of macromolecules and their complexes from heterogeneous datasets. *Philos. Trans. R. Soc., B* **2014**, *369* (1647), 20130567.

(5) Ourmazd, A. Machine-learning Routes to Dynamics, Thermodynamics and Work Cycles of Biological Nanomachines. In *X-Ray Free Electron Lasers: Applications in Materials, Chemistry and Biology*; The Royal Society of Chemistry: 2017; Chapter 22, pp 418−433, DOI: 10.1039/9781782624097-00418.

(6) Dashti, A.; Schwander, P.; Langlois, R.; Fung, R.; Li, W.; Hosseinizadeh, A.; Liao, H. Y.; Pallesen, J.; Sharma, G.; Stupina, V. A.; Simon, A. E.; Dinman, J. D.; Frank, J.; Ourmazd, A. Trajectories of the ribosome as a Brownian nanomachine. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (49), 17492−7.

(7) Frank, J.; Ourmazd, A. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods* **2016**, *100*, 61−67.

(8) Riemann, B. 'Über die Hypothesen, welche der Geometrie zugrunde liegen'(1854, published posthumously by Dedekind). *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen* **1867**, *13*, 133−152.

(9) Tenenbaum, J. B.; de Silva, V.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290* (5500), 2319.

(10) Roweis, S. T.; Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290* (5500), 2323.

(11) Giannakis, D.; Majda, A. J. Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (7), 2222−2227.

(12) Dashti, A.; Hail, D. B.; Mashayekhi, G.; Schwander, P.; Georges, A. d.; Frank, J.; Ourmazd, A. Functional Pathways of Biomolecules Retrieved from Single-particle Snapshots. *bioRxiv* **2018**, 291922.

(13) Coifman, R. R.; Lafon, S. Diffusion maps. *Appl. Comput. Harmon A* **2006**, *21* (1), 5−30.

(14) Bro, R.; Acar, E.; Kolda, T. G. Resolving the sign ambiguity in the singular value decomposition. *J. Chemom.* **2008**, *22* (2), 135−140.

(15) Horn, B. K. P.; Schunck, B. G. Determining optical flow. *Artif Intell* **1981**, *17* (1), 185−203.

(16) Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol 1, Proceedings* **2005**, 886−893.

(17) Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*; The MIT Press: 2009; p 1208.

(18) Beauchemin, S. S.; Barron, J. L. The computation of optical flow. *Acm Comput. Surv* **1995**, *27* (3), 433−467.

(19) Lucas, B. D.; Kanade, T., An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2*; Morgan Kaufmann Publishers Inc.: Vancouver, BC, Canada, 1981; pp 674−679.

(20) Farnebäck, G. In *Two-Frame Motion Estimation Based on Polynomial Expansion*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2003; pp 363−370.

(21) Hariyono, J.; Hoang, V. D.; Jo, K. H. Moving Object Localization Using Optical Flow for Pedestrian Detection from a Moving Vehicle. *Sci. World J.* **2014**, *2014*, 196415.

(22) Pers, J.; Sulic, V.; Kristan, M.; Perse, M.; Polanec, K.; Kovacic, S. Histograms of optical flow for efficient representation of body motion. *Pattern Recogn Lett.* **2010**, *31* (11), 1369−1376.

(23) Dalal, N.; Triggs, B.; Schmid, C. Human detection using oriented histograms of flow and appearance. *Computer Vision - Eccv 2006, Pt 2, Proceedings* **2006**, *3952*, 428−441.

(24) Weiss, Y.; Pearl, J. Belief Propagation. *Commun. ACM* **2010**, *53* (10), 94−94.

(25) Pearl, J. Fusion, Propagation, and Structuring in Belief Networks. *Artif Intell* **1986**, *29* (3), 241−288.

(26) Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc.: 1988; 552 pp, DOI: 10.1016/C2009-0-27609-4.

(27) Murphy, K. P.; Weiss, Y.; Jordan, M. I. Loopy belief propagation for approximate inference: An empirical study. *Uncertainty in Artificial Intelligence, Proceedings* **1999**, 467−475.

(28) Ihler, A. T.; Fisher, J. W.; Willsky, A. S. Loopy belief propagation: Convergence and effects of message errors. *J. Mach Learn Res.* **2005**, *6*, 905−936.

(29) Frey, B. J.; MacKay, D. J. C., A revolution: belief propagation in graphs with cycles. In *Proceedings of the 1997 conference on Advances in neural information processing systems 10*; MIT Press: Denver, Colorado, USA, 1998; pp 479−485.

(30) des Georges, A.; Clarke, O. B.; Zalk, R.; Yuan, Q.; Condon, K. J.; Grassucci, R. A.; Hendrickson, W. A.; Marks, A. R.; Frank, J. Structural Basis for Gating and Activation of RyR1. *Cell* **2016**, *167* (1), 145−157.e17.

(31) Golub, G. H.; Van Loan, C. F. *Matrix computations*, 3rd ed.; Johns Hopkins University Press: Baltimore, MD, 1996; p xxvii, 694 pp.

(32) Yedidia, J. S.; Freeman, W. T.; Weiss, Y. Generalized belief propagation. *Adv. Neur In* **2001**, *13*, 689−695.

(33) Shafer, G. R.; Shenoy, P. P. Probability propagation. *Annals of Mathematics and Artificial Intelligence* **1990**, *2* (1), 327−351.

(34) Budkevich, T. V.; Giesebrecht, J.; Behrmann, E.; Loerke, J.; Ramrath, D. J.; Mielke, T.; Ismer, J.; Hildebrand, P. W.; Tung, C. S.; Nierhaus, K. H.; Sanbonmatsu, K. Y.; Spahn, C. M. Regulation of the mammalian elongation cycle by subunit rolling: a eukaryotic-specific ribosome rearrangement. *Cell* **2014**, *158* (1), 121−131.

(35) Spahn, C. M.; Gomez-Lorenzo, M. G.; Grassucci, R. A.; Jorgensen, R.; Andersen, G. R.; Beckmann, R.; Penczek, P. A.; Ballesta, J. P.; Frank, J. Domain movements of elongation factor eEF2 and the eukaryotic 80S ribosome facilitate tRNA translocation. *EMBO J.* **2004**, *23* (5), 1008−1019.

https://dx.doi.org/10.1021/acs.jcim.9b01115
J. Chem. Inf. Model. 2020, 60, 2484−2491