Teemu Murtola, Mikko Kupiainen, Emma Falck, and Ilpo Vattulainen. 2007. Conformational analysis of lipid molecules by self-organizing maps. The Journal of Chemical Physics, volume 126, number 5, 054707.

Conformational analysis of lipid molecules by self-organizing maps

Teemu Murtola and Mikko Kupiainen
Laboratory of Physics, Helsinki University of Technology, P.O. Box 1100, FI-02015 HUT, Finland and Helsinki Institute of Physics, Helsinki University of Technology, P.O. Box 1100, FI-02015 HUT, Finland

Emma Falck
Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801

Ilpo Vattulainen
Laboratory of Physics, Helsinki University of Technology, P.O. Box 1100, FI-02015 HUT, Finland Helsinki Institute of Physics, Helsinki University of Technology, P.O. Box 1100, FI-02015 HUT, Finland; Institute of Physics, Tampere University of Technology, P.O. Box 692, FI-33101 Tampere, Finland; and Memphys-Center for Biomembrane Physics, Physics Department, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

(Received 25 August 2006; accepted 1 December 2006; published online 5 February 2007)

The authors have studied the use of the self-organizing map (SOM) in the analysis of lipid conformations produced by atomic-scale molecular dynamics simulations. First, focusing on the methodological aspects, they have systematically studied how the SOM can be employed in the analysis of lipid conformations in a controlled and reliable fashion. For this purpose, they have used a previously reported 50 ns atomistic molecular dynamics simulation of a 1-palmitoyl-2-linoeayl-*sn*-glycero-3-phosphatidylcholine (PLPC) lipid bilayer and analyzed separately the conformations of the headgroup and the glycerol regions, as well as the diunsaturated fatty acid chain. They have elucidated the effect of training parameters on the quality of the results, as well as the effect of the size of the SOM. It turns out that the main conformational states of each region in the molecule are easily distinguished together with a variety of other typical structural features. As a second topic, the authors applied the SOM to the PLPC data to demonstrate how it can be used in the analysis that goes beyond the standard methods commonly used to study the structure and dynamics of lipid membranes. Overall, the results suggest that the SOM method provides a relatively simple and robust tool for quickly gaining a qualitative understanding of the most important features of the conformations of the system, without *a priori* knowledge. It seems plausible that the insight given by the SOM could be applied to a variety of biomolecular systems and the design of coarse-grained models for these systems. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2429066]

I. INTRODUCTION

During the last decade, the role of lipids and lipid bilayers in many biological processes has become more recognized.[1] Lipids are a central constituent of biological membranes, and these membranes act as an environment for a wide variety of biomolecular assemblies and biochemical processes. The structural features of membranes are central in controlling these processes, and the properties of the lipid molecules determine many fundamental properties of the membrane. Studies of conformational characteristics of the lipids can therefore yield valuable information about the behavior of membranes.

With increases in available computing resources, atomistic molecular dynamics (MD) simulations have been increasingly used in studies of biological membranes.[2–6] Such simulations produce a wealth of conformational data that should be analyzed to gain more insight into the characteristic features of the system. Traditionally these data have been analyzed by calculating various structural quantities,[2] and possibly visualizing the system frame by frame. Also different projection and clustering methods have been applied to aid the interpretation of the data.[7] Based on the observations, hypotheses on the behavior of the system can subsequently be formed and tested by calculating additional quantities.

Approaches based on neural networks provide alternative methods for analysis of complex data.[8] Some of these approaches, such as the self-organizing map (SOM),[9] are based on unsupervised learning, and thus require no *a priori* knowledge of the data that are being analyzed. This makes them promising candidates for the analysis of conformational data, in particular, for the initial analysis, where one wants to quickly form a qualitative understanding of the most important features of the system. In the context of proteins, SOMs have been applied to, e.g., analysis of three-dimensional structures of amino acid sequences from the protein data bank,[10] classification of sequences within a protein family,[11] identification of overrepresented motifs in sequences,[12,13] prediction of HIV protease cleavage sites,[14] and a study of ammonium salts as ligands at the neuronal nicotinic acetylcholine receptor.[15] However, we are only aware of one preliminary study of the applicability of SOMs in the conformational analysis of lipids.[16]

Conceptually, a SOM is a mapping from high-dimensional input data vectors into a low-dimensional (usu-

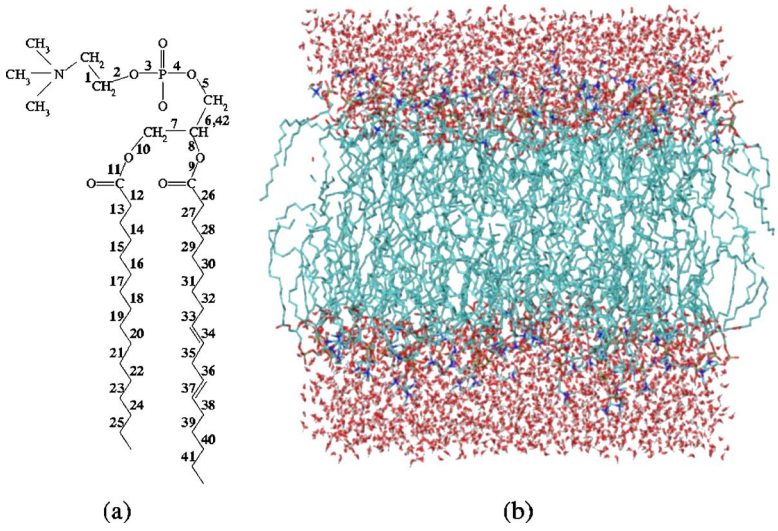

FIG. 1. (Color online) (a) Molecular structure of PLPC. The numbers show the indexing of the dihedral angles. One of the bonds in the glycerol backbone has two associated dihedral angles, and thus two indices. The saturated 16:0 *sn*-1 chain is formed by the dihedrals 12–25, while the diunsaturated 18:2 *sn*-2 chain (dihedrals 26–41) has two cis double bonds at dihedrals 34 and 37. (b) PLPC bilayer in atomic detail.

ally one- or two-dimensional) grid of so-called neurons. After successful training with input data, similar data vectors are mapped to the same or neighboring neurons. This is the origin of the self-organizing behavior of the map. The SOM can be used to convert the nonlinear statistical relationships of high-dimensional data into simple geometric relationships of the neurons.[9] The low dimensionality of the output space makes the visualization of the results simple and thus makes it possible to quickly find relevant information about the system.

The self-organizing map describes the data using only relatively few model conformations, and hence it produces an abstraction of the data. This feature can be used to extract general characteristics of the data, as well as to find the most relevant conformational states. Thus the SOM can also be seen as a coarse-graining approach. The knowledge gained by SOM analysis could be used as a basis for further studies, or the information on the most relevant conformational states could also help in the construction of coarse-grained models for the system. The latter issue is of particular interest, since a number of systematic schemes have recently been proposed to find effective interactions for coarse-grained models of biomolecular systems.[17–24] However, much less attention has been paid to develop techniques for finding appropriate coarse-grained representations for biomolecules. Our results indicate that the SOM can provide a useful tool for this purpose.

In the present work, we have focused on the different aspects of SOMs in the analysis of lipid conformations. There were three separate goals for this work: to consider the possibilities and limitations of the SOM analysis in this context, to study the conformations of a specific lipid system, and to gain more understanding of the methodology, in particular, the training parameters, for further work. A lipid system provides an ideal test case for the approach, since the structure of lipid bilayers is already well understood, and the lipid molecules are relatively small. However, the *a priori* nature of the SOMs should be advantageous for studies of more complex biological systems.

As a model system for all these studies, we have used a 50 ns MD simulation of a 1-palmitoyl-2-linoleoyl-*sn*-glycero-3-phosphatidylcholine (PLPC) bilayer.[25,26] PLPC was chosen because the double bonds in the *sn*-2 chain give rise to interesting conformational features. In addition, PLPC allows us to compare our results with the previous ones by Hyvönen *et al.*,[16] who conducted pioneering studies of SOMs applied to molecular dynamics data for PLPC. While Ref. 16 demonstrated the aptitude of SOM for conformational studies of lipid systems, it was also limited in scope due to the short sampling time (1 ns) that was feasible at that time. Here we extend this work. The molecular structure of the PLPC molecule, as well as a snapshot from the MD simulation, are shown in Fig. 1. As in Ref. 16, we have used dihedral angles to parametrize the conformations. This choice was made because of its simplicity, as well as to facilitate comparison with earlier studies. Its consequences are briefly discussed in Sec. III.

We have systematically studied how SOMs could be used in the analysis of lipid conformations, and elucidated the effect of various parameters that influence the behavior of the SOM. The effect of different training parameters on the structure and quality of the map, as well as the effect of the size of the map, have been investigated. Simple qualitative rules have been deduced for the behavior of the SOM. These rules can be used for improving the quality of the map and for tuning the level of detail of the map, and provide a sound basis for further work. We have also tested different methods for assessing the quality of the map, and propose a new method that can be used to determine whether the size of the map is sufficient.

Using the parameters deduced above, we then apply the SOM analysis to PLPC. We analyze the molecule in three separate parts: the headgroup, the glycerol region, and the *sn*-2 chain. In addition, we find that our results for the whole molecule differ significantly from those of the earlier study,[16] and discuss reasons for this. We also give examples of how SOMs can be used to gain new insights into the structure and dynamics of lipid bilayers. The SOM is used to study the dynamics of the headgroup region, and to analyze the correlations between the conformations of the different parts of the molecule. In addition, we provide an example of how the SOMs could be used in coarse graining. The results demon-

strate that the SOM can be used as a robust tool for gaining qualitative insights into the conformations of the molecules. Finally, we present a thorough discussion of different aspects of the SOM methodology for studies of biomolecular conformations. We discuss the advantages and problems of the SOM approach, and point out possible directions for future research, such as analyzing complexes of molecules to gain insight into specific interactions between the molecules.

II. METHODS

A. Self-organizing maps

A SOM is a powerful software tool for the visualization of high-dimensional data.[9] It consists of a low-dimensional grid of so-called neurons, and a model vector associated with each neuron. Each model vector represents a group of similar data vectors. SOM analysis consists of five phases: selection of data representation, selection of map structure, initialization of the model vectors, training of the map, and finally analysis of the trained map. In this section, we briefly describe the SOM method and the choices that we have made for the first four phases. Most of these choices are discussed in more detail in Sec. III. All the SOM analysis in this work was performed using the SOM Toolbox,[27] which was slightly modified to take into account the periodic nature of the used variables.

Selection of data representation. We have decided to use the values of the dihedral angles to represent the conformations, as was done in Ref. 16. In this case, an individual molecule can be visualized by using the relevant set of dihedral angles together with the average bond lengths and bond angles in the molecule. The similarity of conformations was measured using standard Euclidean metric for the dihedral vectors without weighting. We also tested different weighting schemes for the analysis of the whole molecule, but this did not significantly alter the results. The SOM implementation was modified to take into account the periodic nature of the angle variables.

Selection of map structure. We have used a nonperiodic (sheetlike) hexagonal grid of neurons for the SOMs. The map size of 48×72 was selected for the final analyses. However, for the analysis of the whole molecule, a size of 40×60 was used to speed up the training. We also studied the effect of the size of the map on the results, see Sec. III.

Initialization. The initial values for the model vectors were constructed using linear initialization[9] where the model vectors are placed on a regular lattice on a two-dimensional plane (one-dimensional if the map is linear) that is oriented such that the variance of the data, projected to the plane, is maximal. This plane is spanned by the eigenvectors of the data covariance matrix corresponding to the largest eigenvalues.[9]

Training. Once the SOM is initialized, it has to be trained with the data vectors. The training aims to modify the model vectors such that they represent the typical features of the data vectors as precisely as possible. The map should also represent the topology, i.e., the structure and internal distances of the original data. We have used sequential training[9] where the data vectors are traversed one by one, and at each step the best-matching unit (BMU) of a given data vector is moved towards the data vector. The BMU is defined to be the neuron whose model vector is closer to the data vector than that of any other neuron. In addition to the BMU, the model vectors of neighboring neurons are also updated, although by a smaller amount. This update step may be expressed as

$$\mathbf{m}_i = \mathbf{m}_i^0 + h_{\mathrm{BMU}(\mathbf{x}),i}(t)[\mathbf{x} - \mathbf{m}_i^0],$$

where $\mathbf{m}_i^0$ and $\mathbf{m}_i$ are the model vectors before and after the update, $\mathbf{x}$ is the data vector, BMU($\mathbf{x}$) is the BMU corresponding to $\mathbf{x}$, and $h_{i,j}(t)$ is a neighborhood function. The vector $\mathbf{x}-\mathbf{m}_i^0$ is the direction towards which the model vector should be moved in order to make it more similar to the data vector.

The neighborhood function determines the magnitude of the changes to the model vectors and is a decreasing function of the distance between neurons i and j. Typically the neighborhood function is written as

$$h_{i,j}(t) = \alpha(t)g(\|\mathbf{r}_i - \mathbf{r}_j\|;t),$$

where $\alpha(t)\in[0,1]$ is a learning rate, and $g(r;t)$ is a shape function, which is usually taken to be Gaussian with a time-varying variance $\sigma^2(t)$, scaled such that $g(0;t)=1$. The vectors $\mathbf{r}_i$ are the positions of the neurons on the low-dimensional grid. Both $\sigma(t)$ (called the neighborhood radius) and $\alpha(t)$ are decreasing functions of time such that in the beginning of the training the map organizes rapidly, while towards the end of the training more and more detailed features of the map are tuned.

In this study, the initial learning rate was 0.3, and it decreased exponentially during the training to a final value of 0.0015. The neighborhood radius decreased linearly from an initial value of 3 to a final value of 0.7. The length of the training was ten epochs, i.e., each data vector was presented to the map ten times during the training.

B. Molecular dynamics

For our analysis of PLPC conformations, we used the conformations produced by a 50 ns molecular dynamics simulation of a PLPC bilayer with 128 fully hydrated lipid molecules. The details of the simulation have been published elsewhere,[25] and only a brief summary will be given here. The simulation was performed in the *NpT* ensemble with the GROMACS molecular simulation package.[28] The force field and the starting configuration were taken from a previous study of a PLPC bilayer.[29] The temperature was kept at 310 K with a Nosé-Hoover thermostat,[30,31] and a Parrinello-Rahman barostat[32,33] was used to keep the pressure at 1 bar. Long-range electrostatic interactions were handled using the particle-mesh Ewald method.[34,35] After equilibration, 36 ns of the 50 ns trajectory were used for analysis, with the configurations saved every 10 ps. VMD software[36] was used for the visualization of molecular structures.

III. CHOOSING PARAMETERS FOR SOM

There are many choices that have to be made in the process of constructing and training a self-organizing map,

and some of these choices may significantly alter the properties of the resulting map. In this section, we discuss the effects of these choices. We also study the effect of different training parameters to give simple rules for selecting them. An understanding of these effects can also be used to tune the map towards a specific goal such as a desired level of detail.

It is important to notice that there are actually three goals that are being optimized simultaneously: the resolution of the map, its topological properties, and the computational effort needed to train it. In many cases improving one of these leads inevitably to worse performance for the other ones.

A. Data representation

To use SOM for conformational data, we have to choose how to represent the conformations as n-dimensional vectors. Typically, Euclidean distance between these vectors is used to measure the similarity of the conformations, which should be taken into account when selecting the representation. The results of the analysis can only be as good as the underlying representation of the data allows. Hence, selecting the data representation is perhaps the most important step in applying the SOM approach. The dihedral angles seem to work well for the analyses in this work, but in other cases careful consideration is required before fixing the representation.

Some different alternatives for the representation are discussed in Ref. 7, along with their advantages and disadvantages. For example, using the dihedral angles is rather sensitive to local conformational changes and similarities, but not very sensitive to the apparent cancellation of changes in two dihedral angles. It should also be noted that the SOM places some additional limitations on the metric (and the representation): for each pair of conformations, we should have a well-defined method for making one of the conformations more similar to the other.

It is also possible to modify the similarity measure by weighting the distance norm to place more emphasis on certain variables. This may be particularly useful in the case of complex molecules where some of the variables are clearly more important than the others. Weights can also be used to highlight some areas of the molecule such that the conformations of these parts are more likely to form clusters in the final map (by a cluster we mean a region of the map in which the model vectors are similar to each other, and different from model vectors in other clusters). However, this requires some insight into the system under study, and thus complicates the analysis.

B. Structure and initialization

We have decided to use a two-dimensional hexagonal grid of neurons. This is a typical choice for the two-dimensional structure because it gives the most isotropic structure for the map. Sheet topology was chosen for the map, i.e., the map itself is not periodic. We confirmed that although the dihedrals are periodic, this does not have any major effect on the map. We chose to use 48×72 neurons as a standard map size for most of the studies in this work. One dimension was chosen significantly larger than the other to allow the map to orient itself properly.[9] Training such a map with our conformation data (400 000 conformations, each having 12 dihedrals) takes a few days on a standard desktop computer. The effects of the size of the map on the resulting map will be discussed at the end of this section.

Linear initialization was used to set the initial values for the model vectors. The initialization routine was used as it was implemented in the SOM Toolbox, with no consideration for periodicity of the dihedral angles. For most angles this does not matter, because they are distributed around the 180° value. For the rest, the map seems to quickly converge to reasonable values, confirming that the initialization does not decrease the quality of the map. In principle, the initialization method does not affect the results, but careful initialization may lead to faster convergence in the training phase.

C. Training

The features of the map after training are mainly influenced by three factors: the behavior of the neighborhood function $h_{i,j}(t)$ as a function of time, the length of training, and the training data itself. As discussed in the previous section, the behavior of the neighborhood function is parametrized by two quantities: the learning rate $\alpha(t)$ and the neighborhood radius $\sigma(t)$. To determine how these different parameters affect the features of the resulting map, we looked at different possibilities for the selection of these training parameters. The different options were selected from the multitude of possibilities provided by the SOM Toolbox, using the default values as a starting point. For the studies described here, the data for the headgroups of the PLPC molecules were used for training. The size of the map was 48×72 neurons in all cases.

To quantitatively assess the quality of the maps with different training parameters, we use quantization and topographic errors. The quantization error is defined as the average distance between each data vector and its BMU, and the topographic error is the proportion of all data vectors for which the BMU and the second best matching unit are not adjacent on the low-dimensional grid. The quantization error measures how well the model vectors can represent the underlying data, i.e., the resolution of the map. The topographic error measures how well the map preserves the topological features of the input data.

Choice of σ. For $\sigma(t)$, we used a linearly decreasing form. The initial value was set to 3, and the final value was varied from 0 to 0.8. Generally a smaller final value results in a larger topographic error, but a smaller quantization error. This makes sense, since with a small neighborhood radius only the BMU of a data vector is updated towards the end of the training, which improves resolution, but does not conserve the topology of the map. However, one of the most important features of a SOM is that it should represent the topological properties of the data. Hence, we decided to use a value of 0.7 for the final value of the neighborhood radius. The effect of $\sigma(t)$ is also coupled to the size of the map, and is discussed further at the end of this section.

Choice of α. For $\alpha(t)$, we tested three possible forms, a

linearly decreasing one (zero at the end of training), an exponentially decreasing form (by the name "power" in SOM Toolbox), and an inverse form $[\propto(a+t)^{-1}]$. In addition to the form of the learning rate, we varied its initial value α_0 from 0.1 to 0.5. For the nonlinear forms, the built-in value of the SOM Toolbox was used for the final value of α (this is $0.005\alpha_0$ for the exponential form and $0.01\alpha_0$ for the inverse form). For this system, the inverse form gave a poor topology for the map and was discarded. The linear and exponential forms differed little from each other, with the power form usually giving a slightly better topographic error at the cost of slightly poorer quantization error. As for the initial value, we found that a smaller value leads to a better topology, again at the cost of a somewhat poorer resolution. We decided to use an initial value of 0.3 and a power form for the learning rate as a reasonable compromise for minimizing both error measures.

Length of training. The number of data samples is very large and the data contain very similar vectors. Hence, a reasonably low number of epochs (passes over the data) can be used. We trained the maps with either five or ten epochs, and noticed no significant improvement for longer training, neither in quantization nor topographic error. However, the training length of ten epochs was chosen for production runs to make sure that the individual data samples are represented well enough. It should be noted that the effect of training length is difficult to study independently since it also affects the time evolution of the other parameters.

Choice of training data. Also the choice of training data can have an effect on the resulting map since the map can be only as good as the training data. Some effects of poor training data are discussed in more detail in Sec. V. Our results show that the number of configurations that are used in training is not as crucial as a good sampling of the conformational space. In the present study, the conformational dynamics for the well and poorly sampled cases differ significantly, which could be used to distinguish these two cases (see Sec. V). For the representative sample, the trajectory of a single molecule covers most of the map, while for the short sample, such trajectories cover only a few neurons.

D. Map size

The size of the map has a significant effect on the features of the map, and on the type of errors induced by the mapping. These effects are coupled to the choice of training parameters, in particular the neighborhood radius. To study these effects in detail, we constructed several maps with different sizes and neighborhood radii and compared their properties. The selected map sizes were 48×72, 40×60, 32×48, 24×36, 16×24, and 8×12. The absolute upper limit for the size is set by the amount of training data: there should be enough data per neuron to avoid overlearning (where the map learns to represent the individual data vectors and not general features). First, we used similar training parameters for all sizes, scaling the neighborhood radius down with the size of the map. In this case, decreasing the size of the map leads to a significant increase in the topographic error, with a smaller increment in the quantization error. We also trained smaller maps (sizes 16×24 and 24×32) with training parameters identical to those of the largest map. In this case the topographic error actually decreases when the map size is decreased, accompanied with a significant increase in the quantization error.

Visualization of the different sized maps gives additional insight into the method. Two important visualization methods for the SOMs are unified distance matrices (U matrices) and component planes.[9] In the U matrix, neurons are colored based on their average distance from their neighbors. This gives a clear overview of the general structure of the map. The component planes show the dihedral angles of the model vectors in an easily readable way.

Figure 2 shows the U matrices for the different maps. In light regions the model vectors of neighboring neurons are similar, and the darker regions mark larger differences. The largest maps in Figs. 2(a)–2(c) have distinct clusters with distinct boundaries, and these clusters cover the majority of the map. When the map size is decreased further, the boundaries of the clusters become more vague [Figs. 2(d)–2(f)]. However, using the same neighborhood radius for the smaller maps as for the largest map results in clearer clustering [Figs. 2(g) and 2(h)], although for the smaller map a large portion of the map is taken up by the boundaries between clusters.

The component planes are shown in Fig. 3 for four selected maps. Each small figure shows the values of a single dihedral angle for each model vector, and the color range visualizes the angle range that is actually used by the model vectors. The component planes show that for all map sizes the map has distinct regions for different values of each dihedral angle. However, the smallest maps trained with a small neighborhood radius [Figs. 3(e) and 3(f)] have several single-neuron clusters, and neighboring neurons even within a single cluster can have substantial variability. This explains the poor topology of the map. In contrast, the component planes for a small map trained with a larger neighborhood radius [Fig. 3(h)] show nearly smooth variation of the angles, with the exception of cluster boundaries. This explains why the U matrices for the latter map show clearer structure. The component planes also show that as the map size is reduced, there is a significant reduction in the range over which the angles in the model vectors vary, indicating more "averaged" configurations (see Sec. V for more discussion). This can be most clearly seen in dihedrals 2, 5, 9, 10, and 42 [see Fig. 1(a)], in particular between Figs. 3(a) and 3(h).

In the present work, we have also developed a simple way to assess the cluster structure of the SOMs. The so-called BMU rank plot, i.e., a plot of hits (number of data vectors having the neuron as BMU) as a function of the rank of the neuron (the position of the neuron after sorting according to the number of hits), can give quick insight into the topological features of the map, beyond the topographic error.

The different map sizes above provide a good illustration of the usefulness of BMU rank plots in assessing the structure of the map. Rank plots for the different maps are shown in Fig. 4. The 40×60 case has been excluded for the sake of clarity, and because it does not significantly differ from the

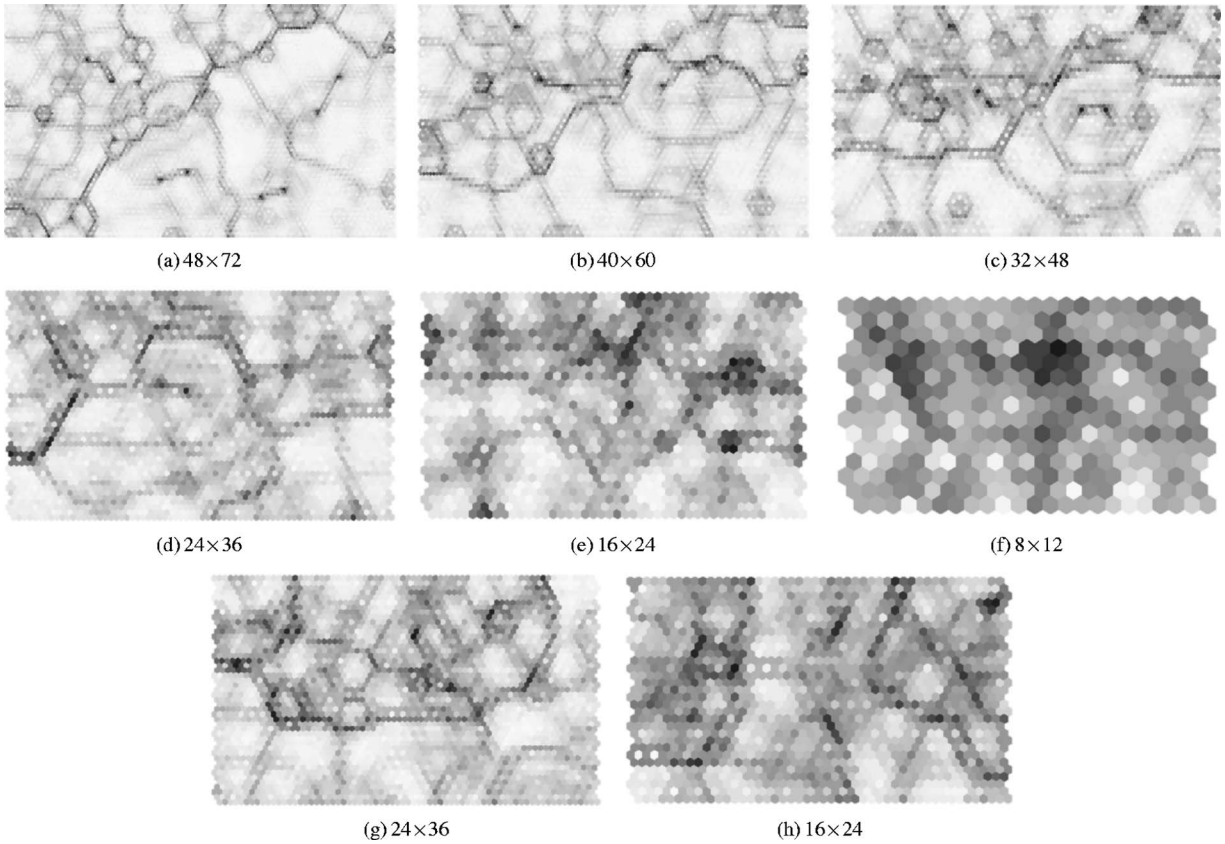

FIG. 2. U matrices for different SOM sizes and different training parameters. Headgroup data were used for the training of all maps. In light regions the neighboring neurons represent similar conformations, and dark regions mark larger differences. For the maps in (a)–(f) the neighborhood radius was scaled with the size of the map. The training parameters for maps (g) and (h) were identical to those of (a).

other nearby sizes. First, the figure shows that when the (final) neighborhood radius is increased while keeping the size of the map constant, the number of low hit count neurons increases. This is consistent with the presence of more distinct cluster boundaries. Second, it shows that the percentage of the low hit count neurons increases when the size of the map is reduced, with the exception of the smallest map size. This trend can be explained by noting that the number of clusters is roughly constant irrespective of the size of the map, and therefore the cluster boundaries (whose width does not change significantly) take up a larger portion of the smaller maps. For the smallest map there are no longer any cluster boundaries, resulting in the lack of low hit count neurons. These conclusions are in agreement with the other visualizations, highlighting the applicability of BMU rank plots in quickly making a rough assessment of the structure of the map.

To summarize, the size of the SOM and the neighborhood radius both affect the properties of the map significantly. The neighborhood radius effectively determines the size of the smallest clusters that can appear on the map, and thus increasing it makes the map smoother, decreasing resolution and improving topology. The neigborhood radius also affects the minimum width of a boundary between clusters. Too low a radius leads to a poor topology of the map, and the self-organizing property of the SOM is lost. The presence of single-neuron clusters that are very dissimilar to their neighbors indicates too small a neighborhood radius. Varying the size of the map while keeping the neighborhood radius constant changes the maximum number of clusters that the map can have, and hence it can be used to tune the resolution of the map. If the map size is too small for every significant conformation to have its own cluster, some clusters are merged, and the boundaries of the clusters may become vague, making the analysis more difficult. However, the component planes show that all maps have distinct regions for the four major conformations of the headgroup-glycerol region (see next section). In addition, the relative proportion of the map that is allocated for each conformation is quite independent of the size of the map.

IV. CONFORMATIONAL ANALYSIS OF PLPC

We first applied the SOM to analyze the conformations of the whole PLPC molecule as in Ref. 16. However, even with the largest map sizes the map shows only vague clusters and is thus very difficult to analyze. This and other measures indicate that even larger map sizes would be needed to adequately describe the whole molecule, but such sizes are not computationally feasible. The other possibility is to reduce the complexity of the conformational space, an option also suggested by Hyvönen *et al.*[16] This can be accomplished by studying only a subset of the dihedral angles. We did this for three different groups: the headgroup, consisting of dihedral angles 1–11 and 42 [see Fig. 1(a)], the glycerol region (dihedral angles 5–13, 26–27, and 42), and the diunsaturated

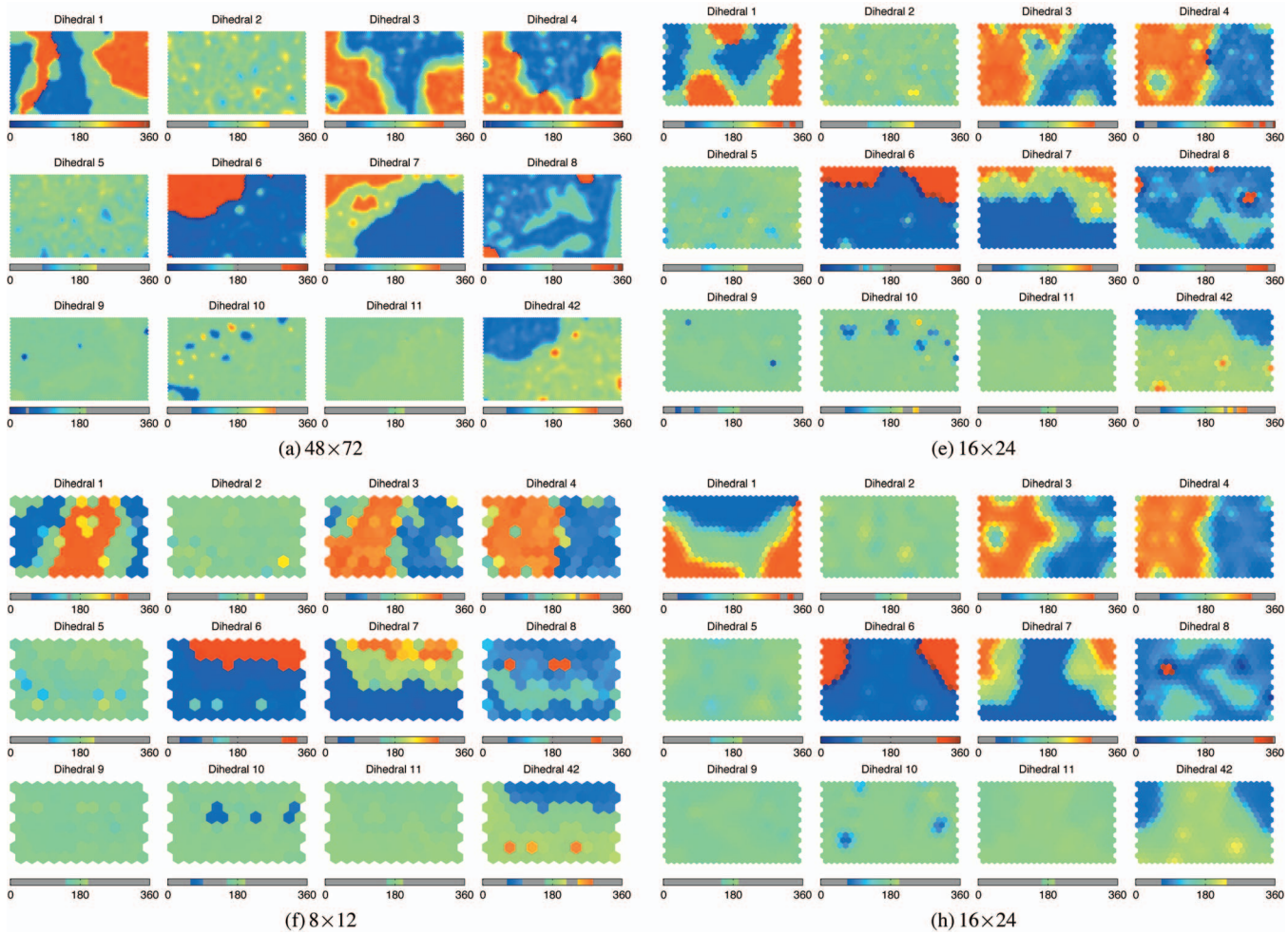

FIG. 3. (Color) Component planes for different SOM sizes and different training parameters. Headgroup data were used for the training of all maps. The maps are selected from those in Fig. 2, and the panels have been labeled accordingly. Each small figure shows the values of one dihedral angle for each model vector. The color range is the same for all figures, and the color bar below each figure shows the range of values that are present on the map.

sn-2 chain (dihedral angles 30–41). The results for these groups are presented in this section, and the more detailed discussion of the whole molecule is deferred to the next section, where we will also compare our results with those of Ref. 16.

The different regions have rather different structures, and serve different functions in the membrane. The polar headgroups prefer to be in contact with water, and they form the outermost layer of the membrane. They also shield the nonpolar parts of the membrane from contact with water. The glycerol forms the backbone of the molecule to which the other parts are attached, and thus influences the overall shape of the whole molecule. Finally, the tails form the innermost part of the bilayer. In biological membranes, this region acts, for example, as an environment for the hydrophobic parts of integral membrane proteins. The unsaturated *cis* bonds in the *sn*-2 chain have an effect on the ordering of the chains, and therefore also on other physical properties. Hence, understanding the effect of different conformations of the *sn*-2 chain is of specific interest.

A. Headgroup

The trained SOM for the headgroup data is visualized in Figs. 2(a) and 3(a). Figure 2(a) shows the U matrix and Fig. 3(a) the component planes. There is a prominent boundary in the map, particularly distinct in the U matrix, extending diagonally from the lower left corner to the upper right corner. The lower part of the map has large homogeneous clusters, while the upper part is scattered with small clusters and re-

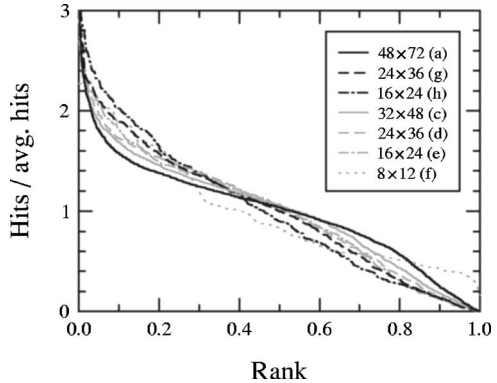

FIG. 4. Plots of hit counts as functions of neuron rank for different map sizes. Headgroup data were used for training the maps. The letters in parentheses refer to the different maps in Figs. 2 and 3. The black lines are for maps trained with identical neighborhood radii. For the gray lines, the neighborhood radius was scaled down with the size of the map. For each plot, the number of hits has been scaled by the average number of hits, and the rank has been scaled to range from zero to one.

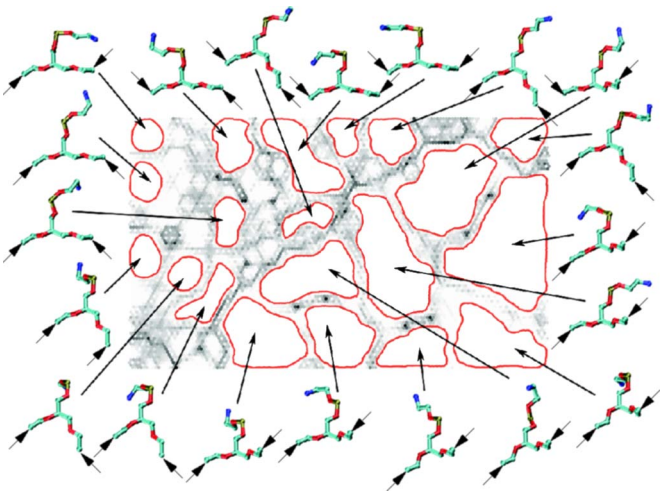

FIG. 5. (Color online) Visualization of most prominent conformations of headgroup SOM. The glycerol backbone is oriented identically for all conformations, and the carbons where the tails start are marked with arrows. The U matrix is shown in the background, and the clusters have been determined manually.

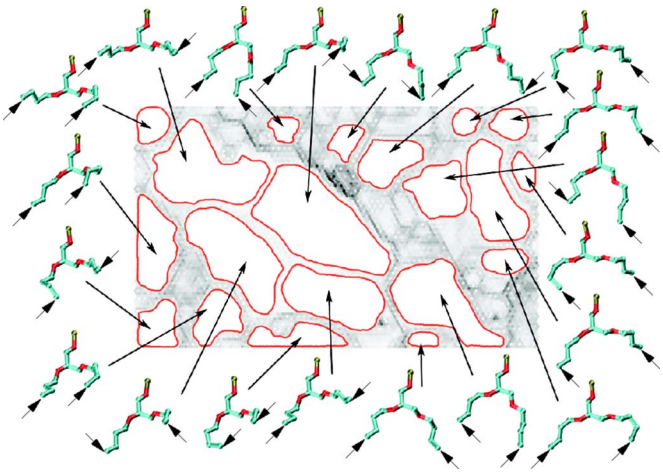

FIG. 6. (Color online) U matrix of SOM for glycerol region [dihedral angles 5–13, 26–27, and 42 in Fig. 1(a)]. The conformations for the most prominent clusters are also visualized, with the glycerol backbone in identical orientation for all conformations. The arrows mark the carbon atoms where the tails are attached. The clusters have been determined manually.

gions with no clear clusters at all. From the component planes in Fig. 3(a) it is evident that this boundary is characterized by a large change in the dihedral angle 7. The lower part of the map corresponds to the most typical value of this angle, while the upper part covers the other values. There are also strong correlations between dihedral angles 6, 7, 11, and 42. These dihedral angles are located in the region where the headgroup is attached to the glycerol backbone, and thus the boundary reflects the conformational degrees of freedom of the headgroup with respect to the glycerol region. On crossing the boundary we see a major change in more than one dihedral angle, which explains why the boundary is so clearly visible. This also suggests that there is a high potential barrier associated with the boundary, and therefore we should see fewer conformational transitions across the boundary than inside the regions.

To further visualize the conformations in different clusters we have chosen one neuron from each cluster for closer inspection. The neurons were chosen approximately in the middle of the cluster such that they would have as many hits as possible. Inside one cluster it is justifiable to look at only one neuron and the associated model vector since the neighboring neurons in the same cluster are very similar and show no qualitative difference.

Figure 5 shows the clusters (determined manually) and the conformations of the associated model vectors. The division between the upper left and lower right halves of the map is also clearly visible in the conformations: the upper left corner has six conformations where the lower part of the group (between *sn*-1 and *sn*-2 chains) is straight and bond 5 points towards the viewer (the direction is determined by dihedral 6), and for the rest of the conformations above the dark boundary the bond points away from the viewer and bond 7 is downwards, inducing a kink in the glycerol region. For the conformations below the boundary, bond 5 points away from the viewer and bond 7 towards the viewer. Thus the boundary indeed marks the differences between these three major configurations of the headgroup with respect to the glycerol region. The clusters inside these major regions are distinguished by different orientations of the *P-N* vector in the headgroup and, in some cases, different orientations of the beginning of the *sn*-2 chain.

The component planes also show that there is a two-part region where dihedral 8 is significantly different from the rest of the map (red regions in the dihedral 8 plane). There are no clear clusters within this region, indicating that the value of dihedral 8 is the main determining factor for this class of conformations. This group of conformations provides a fourth major class for the headgroup. The conformations in this class differ from other conformations in that the first carbon of the *sn*-2 chain and the *sn*-1 carbon atom are on the same side of the *sn*-2 C–O bond. Analysis of the dynamics in Sec. VI additionally shows that there is a large barrier in crossing to or from this region.

The SOM can also be used to visualize the conformational dynamics of the molecules by plotting the trajectories of individual molecules on the map.[16] A thorough analysis of the headgroup dynamics is postponed until Sec. VI, here it suffices to note that the trajectories (not shown) indicate that the molecules stay within one region of the map for tens of picoseconds before jumping to another region. The movement within a region results from small fluctuations in dihedral angles while a longer leap is caused by one or more dihedral angles changing from one local potential minimum to another. The effect of the prominent diagonal boundary on the map can also be clearly seen in the trajectories: transitions over the boundary are very rare events.

B. Glycerol region

Figure 6 shows the U matrix of the SOM for the glycerol region, together with representative configurations for the major clusters in the map. The component planes of the map are shown in Fig. 7.

The glycerol region contains many of the same dihedral angles as the headgroup region. Hence, it is not surprising that the map shows a boundary similar to the one in the

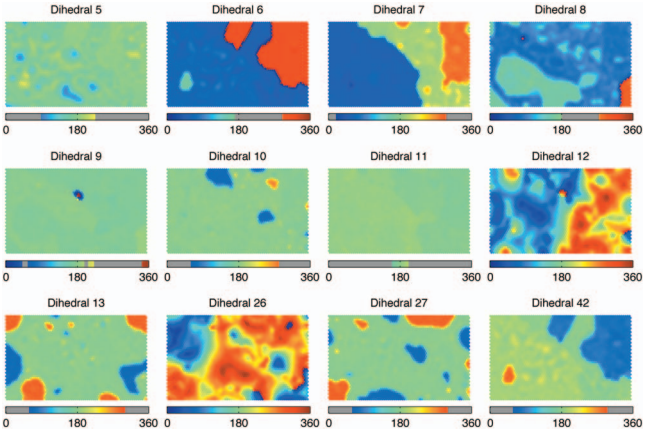

FIG. 7. (Color) Component planes of SOM for glycerol region. See Fig. 3 for details on how to interpret the figures.

headgroup map, and a separate region for high values of dihedral 8. In the case of the glycerol SOM, the boundary runs from the lower right corner to the upper left part of the map. This boundary, as the one in the headgroup SOM, is characterized by simultaneous changes in the dihedral angles 6, 7, 11, and 42. In the upper right corner, there are five conformations that have a straight configuration between *sn*-1 and *sn*-2 chains and bond 5 pointing toward the viewer, corresponding to the upper left corner of the headgroup map. The other conformations above the boundary have bond 7 pointing downwards and bond 5 away from the viewer (with one exception), and the configurations below the boundary all have bond 5 pointing away from the viewer and bond 7 towards the viewer, both in correspondence with the headgroup map. The clusters within these groups are distinguished by the different orientations of the first bonds of the tails. The correspondence between the headgroup and the glycerol map will be analyzed in more detail in Sec. VI. One should also note that the *trans* and *gauche* states of dihedral

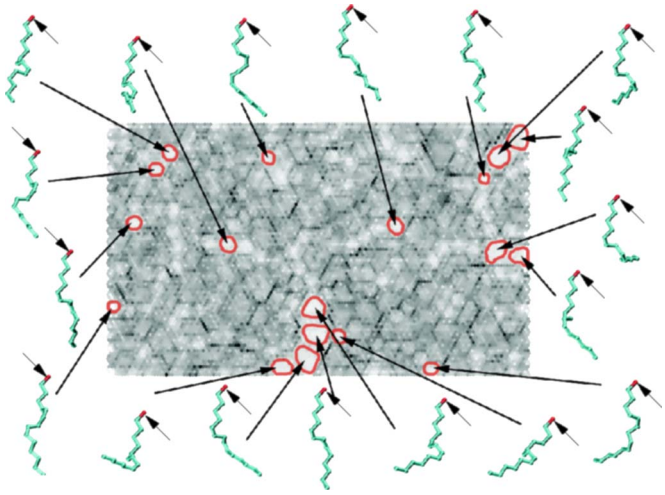

FIG. 8. (Color online) U matrix of *sn*-2 chain SOM [dihedral angles 30–41 in Fig. 1(a)]. The conformations for some of the largest clusters are also visualized. The carbon atom closest to the headgroup has been colored differently from the rest and marked with an arrow, and the first bond is oriented identically for all conformations. The clusters have been determined manually.

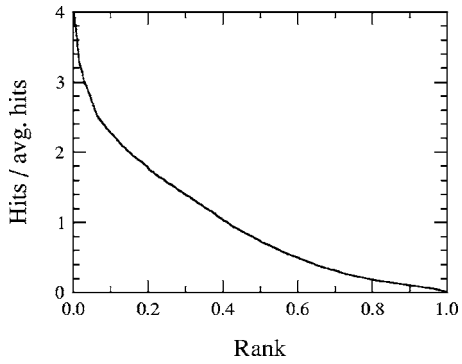

FIG. 9. BMU rank plot of *sn*-2 chain SOM. The number of hits has been scaled by the average number of hits, and the rank has been scaled to range from zero to one.

angles 13 and 27 are clearly visible and correspond to different areas of the map.

The conformational dynamics shows many features reminiscent of the headgroup case. Transitions across the boundary are rare, and short-time transitions take place predominantly between neurons in the same cluster.

C. *sn*-2 chain

Figure 8 shows the U matrix of the SOM for the diunsaturated *sn*-2 chain, together with representative configurations for some major clusters in the map. The BMU rank plot is shown in Fig. 9 and the component planes are shown in Fig. 10. In this case the structure of the U matrix is markedly different from the other two cases. The map has a large number of small clusters, but the boundaries of these clusters cannot be clearly defined. There are also areas which are difficult to classify as belonging to any cluster. However, the selected visualizations of the conformations indicate that in different parts of the map, the dihedral angles around the double bonds have different values, and that there may also be differences in the saturated regions.

The differences in the structure of the map, compared with the other two cases, can be explained by the conformational space accessible to the chain. The chain has six dihedral angles with *trans-gauche* type behavior, as well as four skew-type dihedral angles close to the double bonds. In ad-

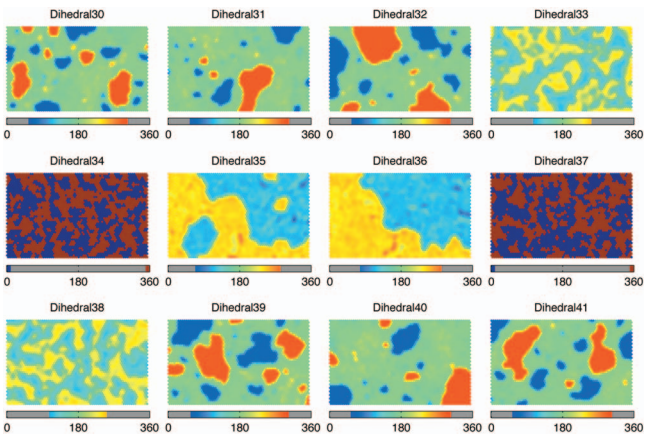

FIG. 10. (Color) Component planes of *sn*-2 chain SOM. See Fig. 3 for details on how to interpret the figures.

dition, the correlations between different angles are weak. Thus there are a large number of probable conformations of the chain, characterized by a major change in one or more of the dihedral angles. Because of rapid *trans-gauche* isomerization rate for the chain (of the order of tens of picoseconds), the chains explore this conformational space thoroughly during the simulation. As there is only a limited amount of neurons on the map, it is not possible to have a separate cluster for each of these conformations. In addition, the conformational space is intrinsically high dimensional, so it cannot be easily mapped into a two-dimensional plane while preserving the topology, as a SOM attempts to do. A quick comparison of quantization and topographic errors to the other cases shows that the topographic error is similar in all cases, but the quantization error is significantly higher for the *sn*-2 chain SOM. This is in agreement with the above discussion: if each conformation cannot have its own cluster in the SOM, the resolution suffers because several distinct conformations have to be described by a single cluster.

Also the BMU rank plot, shown in Fig. 9, is different from the other cases. For the *sn*-2 chain SOM there are a few neurons with a very large hit count, but there is no plateau at intermediate ranks. Thus a very large proportion of the neurons has rather low hit counts. This indicates the presence of many cluster boundaries, see above. A significantly larger map would be required to obtain a plateau in the rank plot and thus to have clear clusters on the map. However, there are also some similarities with the whole molecule SOM, and it is possible that overlearning-related problems could occur with a larger map (see discussion in the next section).

The component planes in Fig. 10 show that despite the problems noted above, the SOM is able to form a rather good representation of the conformational space of the chain. For each *trans-gauche*-type dihedral (dihedrals 30–32 and 39–31) the map clearly shows three distinct (possibly noncontinuous) regions, with the largest region corresponding to the *trans* state and the other regions to the two *gauche* states. Similar partitioning into two roughly equally sized domains can be seen for the skew-type dihedrals (dihedrals 33, 35, 36, and 38) in the vicinity of the double bond. Thus, each region of the map is characterized by a specific combination of the different possible values for the dihedral angles, and the boundaries for these domains are rather sharp.

Comparison of the SOM for the *sn*-2 chain with the headgroup SOMs in Figs. 2(g) and 2(h) show some interesting similarities: the U matrix has large dark areas that do not have any clear clusters. In addition, in both cases the component planes show that the map has clear regions for different conformations, despite the apparent lack of structure in the U matrices. Also the BMU rank plots show similar tails of low hit counts. This observation highlights an important issue that has to be kept in mind when analyzing the map, namely, that a single visualization of the map may give an incomplete picture of the results. It also suggests that for the smaller headgroup SOMs the size of the map is approaching the limit of being able to describe the conformational space of the headgroup adequately. Conversely, it also indicates that a larger map for the *sn*-2 chain probably would indeed have a clearer cluster structure.

For the *sn*-2 chain the trajectories of individual molecules on the map feature rapid transitions over the whole map. This is natural because the time scale for conformational transitions in the chain is of the same order of magnitude as the sampling interval of our data.

V. ANALYSIS OF WHOLE LIPID AND COMPARISON TO EARLIER RESULTS

The SOM for the whole lipid molecule shows some qualitative differences to that described in Ref. 16, which, to our knowledge, is the only SOM study of lipid systems prior to the present work. In Ref. 16, one applied a map of size 10×10 to a 1 ns MD simulation of a PLPC bilayer with 1.44×10^6 lipid conformations. It was found that the resulting model molecules had clearly distinct conformations, and it seemed that the SOM represented the data profoundly well. In the present work, however, the final neuronal model molecules were very similar to each other, in particular, with a small map such as the one used by Hyvönen *et al.* The dihedral angle values for nearly all model molecules were very close to their average values. This effect is most pronounced in the tail region. We studied several map sizes, and this holds for all of them. We argue that these differences originate mainly from the differences in the MD input data. In addition, our results show that the map size used by Hyvönen *et al.* is probably too small to take full advantage of the SOM analysis.

Leaving aside the details in different force fields used in the two studies, an important factor affecting the results of a SOM analysis is how the conformational space of the molecules is sampled. In the present work, we had 3601 samples of 128 molecules (460 918 conformations in total) at 10 ps intervals, covering 36 ns. The earlier study had 40 000 samples of 36 molecules (1 440 000 conformations in total) at 0.025 ps intervals, covering an interval of 1 ns (very reasonable at the time), and every tenth frame was used for training. The major difference is the length of the interval between sampled configurations: a short sampling interval results in major correlations between configurations and hence in many very similar conformations. This is because the time scale of conformational isomerization of the molecule is of the order of picoseconds or tens of picoseconds. Therefore many of the 1.4×10^6 conformations are, in fact, nearly identical, and the data represent only a part of the conformational space. A significant part of the variability results from the different initial configurations of different molecules. This explains the differences in the SOMs: when there are only few conformations in the data, the SOM converges to these conformations, determined mostly by the initial configuration of the simulation, and thus does not represent the general properties of the system.

To verify our arguments we performed a 1 ns extension to our PLPC simulation saving the configurations every 0.04 ps, and chose randomly 32 molecules for analysis. We then trained two SOMs, one with the 1 ns data and another with our original 36 ns data. For both maps, the training parameters were identical to those used in the analysis of molecular parts, and the map size was 40×60. We also did the analysis for the 10×10 maps used in Ref. 16. We found

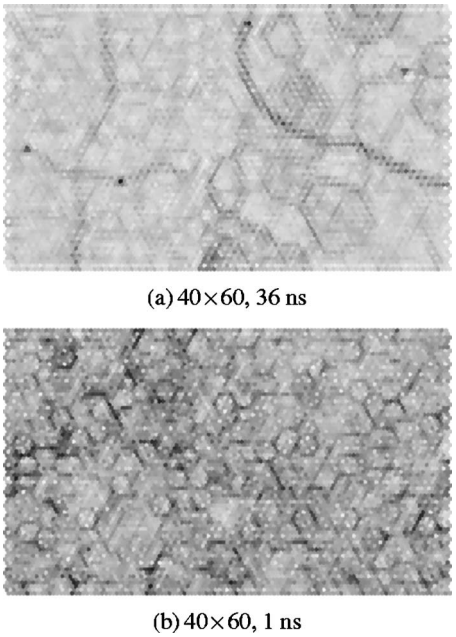

FIG. 11. U matrices for whole molecule SOMs with different training sets (see text for details).

that due to the complexity of the conformations, such a map is too small to take full advantage of the SOM analysis (results not shown).

The U matrix, selected component planes, and the BMU rank plot for the 40×60 maps trained with the 36 ns data are shown in Figs. 11(a), 12(a), and 13 solid line), respectively. For the sake of clarity, only some of the dihedral angles are shown in the component planes, but similar conclusions can be drawn from all of the angles.

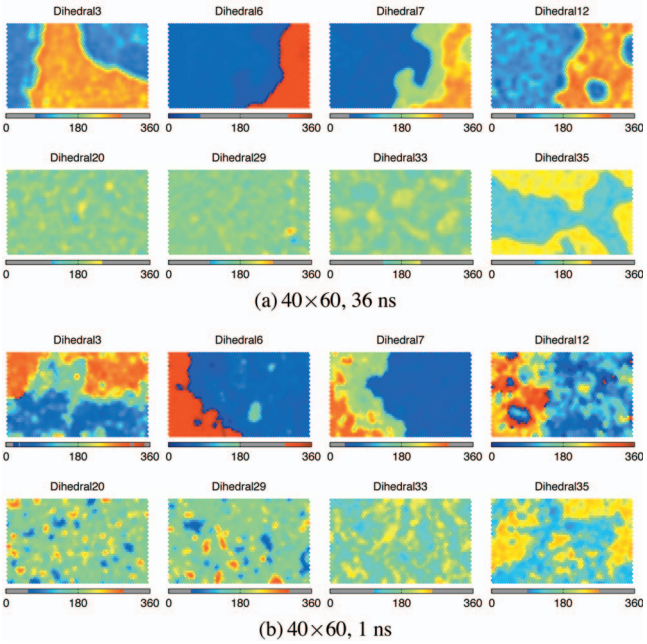

FIG. 12. (Color) Component planes of whole molecule SOMs with different training sets (see text and Fig. 11). For the sake of clarity, only a subset of the dihedral angles is shown. The dihedrals for the headgroup (3, 6, and 7), glycerol (6, 7, and 12), and *sn*-2 (20, 29, 33, and 35) regions can be compared with the respective SOMs in Figs. 3, 7, and 10.

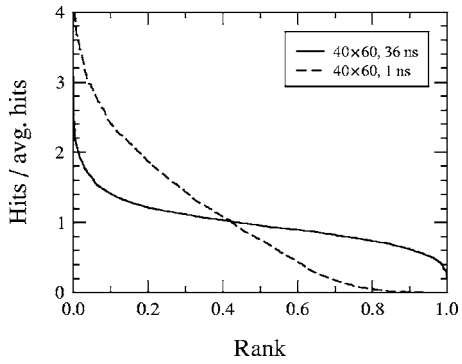

FIG. 13. BMU rank plots of whole molecule SOMs. The number of hits has been scaled by the average number of hits, and the rank has been scaled to range from zero to one.

The rank plot indicates that the size of the map may be too small, but nevertheless, the other figures show that the SOM is able to form a relatively good representation of the data. The component planes for the headgroup dihedrals show many features that are similar to the results in the previous section. However, many of the *trans-gauche* dihedrals in the tails show only conformations close to the most probable *trans* value. Also, the quantization error (divided by the number of dihedral angles to make them comparable) is higher than for the headgroup maps, indicating that there is stronger averaging than for the headgroup. In particular, this can be seen in the reduced range of the model dihedrals. It is also interesting to note that the U matrix has a prominent boundary similar to the headgroup and glycerol cases. However, in this case this boundary is characterized by simultaneous changes in dihedrals 3 and 4. Also the other boundary, characterized by changes in dihedrals 6, 7, 11, and 42, is present, but it is not as clearly visible in the U matrix.

The U matrix, selected component planes, and the BMU rank plot for the 1 ns data are shown in Figs. 11(b), 12(b), and 13 dashed line), respectively. There is a profound difference between these results and those for the 36 ns data, confirming the importance of sufficient sampling of the conformational space. The quantization error is actually 25% lower than for the 36 ns data, and the topographic error is one-third of that of the 36 ns map. Both of these agree with only partial sampling of the conformational space: when the map represents only a part of the conformational space, it can adapt to more detailed features of the data, which leads to better resolution. Further, for the 1 ns data, the ranges of the dihedral angles present in the map are much wider, indicating more diversity in the conformations represented by the map. This is in agreement with the observed averaging: the model molecules in the SOM trained with the 36 ns data are strongly averaged and the chains are mostly in straight all-*trans* states with the exception of the double bond region.

Further analysis of the 40×60 map for the 1 ns data yields additional insight into the behavior of SOMs. The number of distinct conformations per molecule can be roughly estimated by dividing the length of the sample by the average conformational isomerization time. Taking 10 ps as the isomerization time (this is of the same order of magnitude as the fastest transitions in the molecule), we can es-

timate that there are at most 100 distinct conformations for a molecule, and thus the maximum number of distinct conformations in the training data is of the order of 3000. As the 40×60 map has 2400 neurons, there is nearly one neuron for each configurations, making it possible for the map to adapt to the training data nearly perfectly. Thus there is very little averaging, and the map essentially does very little to aid in the analysis of the data. This is also demonstrated by the conformational dynamics (trajectories not shown): for most of the 1 ns trajectory each molecule stays at the same neuron, and there are only few transitions to neurons farther away. The fact that this does not happen for the 36 ns data also confirms that the conformational space is sampled better in that case. This phenomenon is related to overlearning, which occurs if the number of data vectors is comparable to the size of the map. However, as the number of data vectors is not a problem in the current case, better results could probably be obtained by a careful selection of the training parameters. The dynamics of the molecules provides a good indicator for assessing whether the sampling is sufficient: if the trajectories of individual molecules do not cover a major part of the map, the sampling is probably not long enough.

VI. APPLICATIONS

A. Headgroup dynamics

To demonstrate how the SOMs can be used in analysis that goes beyond the standard practices to study the structure and dynamics of membranes,[2] we now consider three applications. As a first case, we analyze the dynamics of the lipid headgroup. Some qualitative aspects were already discussed in Sec. IV. Here we focus on more quantitative analysis of the 48×72 headgroup map.

The simplest way to analyze the dynamics of the headgroup region would be to directly study the transition frequencies between different neurons. However, the large size of the map makes the number of possible transitions intractable. Hence, some kind of clustering is needed to group the neurons into larger units. A systematic method is to be preferred, because in subjective clustering, like the one in Fig. 5, large parts of the map cannot be easily assigned to any cluster. The transition frequencies between neurons provide an attractive measure of the similarity of the neurons: the more transitions between two neurons, the more similar they are. This measure has an additional advantage in that it has not been used in the construction of the map. Hence, the resulting clustering offers an independent method for assessing the quality of the map. The measure can also be easily extended to clusters of neurons by averaging, as discussed below.

Figure 14 shows the clustering based on transition frequencies. The clustering was constructed using hierarchical agglomerative clustering.[37] In this method, we first construct an initial clustering by putting each neuron in its own cluster. We then merge the most similar clusters, and continue until only one cluster remains. This yields a hierarchy that can then be used to select a suitable clustering. The final clustering was obtained by finding all merges where both clusters had at least 80 neurons, and undoing these and all subsequent

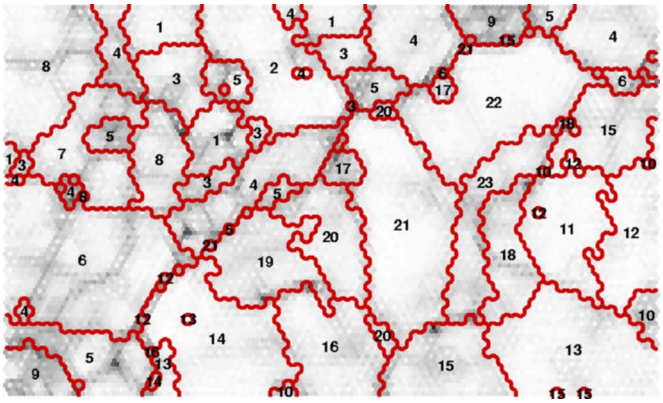

FIG. 14. (Color online) Clustering of headgroup SOM based on transition frequencies. The U matrix is shown in the background. The clusters are not necessarily continuous. Unlabeled neurons have zero hits.

merges to these clusters. After this, all neurons in clusters that had less than five neurons were considered as outliers, and they were merged to the most similar clusters. Due to the clustering method, different values for the splitting size limit result in slightly different clusters: starting from a high value and gradually lowering the limit results initially in a few big clusters, which then split into smaller and smaller pieces. The selected value of 80 was chosen manually to get a reasonably small amount of clusters that were not too large and that were either continuous or formed from a few continuous regions.

In the above process, we need to be able to calculate the distance between the new cluster (formed by merging two clusters) and all the other clusters. There are several ways of defining this distance.[37] Here, we have adopted the average distance: the distance between the new cluster N and some other cluster, say, A, is calculated as the weighted average of the distances between A and the original clusters forming N. The resulting distance has a simple physical interpretation: it gives the expectation value for the number of transitions between two randomly selected neurons belonging to different clusters.

The final clustering has 23 clusters, which makes direct studies of the transition matrix feasible. The cluster numbers in Fig. 14 have been selected manually in such a way that most of the transitions occur between numbers that are close to each other, e.g., 3 and 4.

On the average, 80% of the transitions occur within one cluster, which shows that the clustering represents the actual dynamics of the molecules. For individual clusters, the proportion of intracluster transitions ranges from 57% (cluster 10) to 98% (cluster 9), with values of 76%–92% for clusters 1–8 and 63%–83% for clusters 11–23. Clusters with a significant amount of transitions to nearby clusters typically display a smaller percentage of intracluster transitions. For example, 64%–65% of the transitions in clusters 11 and 12 are intracluster, while only 18% of the transitions are to other clusters than the two. The difference between clusters 1–9 and 10–23 is in line with the structure of the U matrix: in the region covered by clusters 10–23 the neurons are typically much more similar to their neighbors, and therefore it is understandable that there are more intercluster transitions.

The intracluster transition probabilities can be used to estimate the lifetimes of individual clusters. We define the lifetime of a cluster as the time after which the probability of a conformation belonging to the same cluster as initially is less than $1/e$ (where e is the Euler number). With this definition, the 98% probability of an intracluster transition in cluster 9 translates into 650 ps, cluster 10 has a lifetime of 18 ps, and the rest have lifetimes in the range of 22–120 ps. The long lifetime of cluster 9 warrants further study: it turns out that if a trajectory properly enters the cluster (i.e., stays there longer than tens of picoseconds), the lifetime is actually of the order of 5–10 ns. This cluster corresponds to the fourth class of conformations discussed in the headgroup analysis in Sec. IV, characterized by a large value of dihedral 8. The long lifetime of the cluster shows that transitions to and from such a conformation have a high potential barrier.

The prominent diagonal boundary in the U matrix is also seen in the transition frequencies: only 0.3% of the transitions occur over the boundary, i.e., between clusters 1–9 and 10–23. The transition matrix also shows that these transitions cannot happen arbitrarily, but that the most probable paths are from cluster 4 to 17 or 19–23, from 6 to 10 or 12–15, or from 8 to 17 or 23, and identically to the other direction.

To understand these transitions, one can look at the angle between the *P-N* vector and the vector connecting the first carbons of the tails. The latter characterizes the orientation of the glycerol backbone in the plane of the membrane, and hence this angle is related to the joint between the headgroup and the glycerol regions. Looking at the values of this angle for the conformations in different clusters, we see that for the most probable transitions, this angle does not typically change significantly. This makes the conformational change smaller, and therefore the transition is more probable.

There are also smaller blocks of clusters between which the transitions are less probable. For example, transitions between clusters 1–3 and 5–8 occur mainly either through cluster 4 or directly between clusters 1 and 8. The main difference between the blocks is in dihedrals 3 and 4, i.e., in the orientation of the *P-N* vector. Cluster 4 has a similar orientation of the *P-N* vector as clusters 1–3. In its glycerol part, cluster 4 is closer to 5–8. Hence, it functions as a transition state between 1–3 and 5–8. Direct transitions from 1 to 8 are possible because the glycerol part in these clusters is very similar. A similar pair of blocks is formed on the other side of the diagonal by clusters 10–16 and 19–23, although here the pattern of transitions is more complex.

The preliminary results shown here demonstrate how the SOM can aid in the analysis of the dynamics of a complex biological system. The SOM provides a good starting point for the clustering based on transition rates, and also offers a useful template for the visualization of the clusters. In addition, these results propose that conformations that are close to each other on the map resulting from SOM analysis are typically also close in the dynamical sense.

B. Correlations within molecules

As a second application, we discuss the relationship between the conformations of different parts of a lipid molecule. By focusing on where groups of conformations occur on SOMs that portray different parts of a lipid molecule, one can gain more insight into such relationships. The SOMs provide an easy way of visually performing such comparison. For example, we can check how a given cluster in the headgroup SOM maps onto the glycerol or *sn*-2 SOM: we first identify all conformations in the underlying data that have their BMUs in the cluster of interest in the headgroup SOM. We then calculate the BMUs of these conformations on the other maps, and plot the number of hits on top of the U matrices (or other visualizations). If some clusters map into one or a few well-defined, continuous subregions of the other SOM, we can conclude that there is some correlation between the different parts. We can also perform the mapping with the maps interchanged to check whether the found regions also map back to the original clusters.

Comparing the headgroup and glycerol SOMs in this way, we note that the three major groups of headgroup clusters (we use the numbering in Fig. 14 because it gives a convenient frame of reference), 1–8, 9, and 10–23, map to separate (more or less) continuous regions in the glycerol map. The different regions are characterized by different values of dihedrals 7 and 8, and are located similarly to the headgroup map: clusters 1–8 map to the upper right half of the glycerol map, clusters 10–23 to the lower left half, and cluster 9 is located in between. In addition, cluster 5, clusters 4 and 6, and clusters 10 and 17 map to distinct regions that are characterized by distinct values of dihedral 10 (cluster 5) or dihedral 6 (the others).

The above mappings are natural, because the dihedrals that characterize the specific groups of clusters are common to the headgroup and glycerol regions. However, the above mapping also enables us to compare how the dihedrals 12–13 and 26–27 (that are part of our definition of the glycerol region, but not part of the headgroup) behave for different conformations of the headgroup. Qualitative insight can be gained by comparing the distributions of the different dihedrals (Fig. 7) for the regions into which different headgroup clusters map. Such a comparison shows that dihedral 12 is often larger for clusters 1–4 and 6–8 than for clusters 10–23 (in Fig. 14). Armed with this knowledge, we can then make a more quantitative comparison by plotting the histograms of this dihedral for the conformations that belong to the different clusters. This confirms that the observed difference is real: the distribution of dihedral 12 (over all conformations) has two wide peaks, and conformations in clusters 1–4 and 6–8 have values mostly close to one of the peak values, while conformations in clusters 10–23 have values close to the other peak. For the remaining clusters 5 and 9, the histogram is similar to the global distribution.

The above difference originates from the fact that the orientation of the glycerol backbone with respect to the membrane normal is different on different sides of the diagonal. Dihedral 12 determines the orientation of the *sn*-1 chain with respect to the glycerol backbone, and the *sn*-1 chain is always more or less in the same orientation with respect to the membrane normal. Hence, the dihedral has to be in different orientations in different clusters to prevent the tail from pointing out of the membrane.

Similarly, we find that dihedral 26 has a specific value in cluster 9, which again originates in the requirement that the beginning of the tail is always directed into the membrane. In contrast, the following dihedral angles along the tails, i.e., dihedrals 13 and 27, do not have any clear correlations with the headgroup conformation, which indicates that the tail conformations are more or less independent of the headgroup.

The last conclusion is also supported by a comparison of the headgroup SOM to the *sn*-2 chain. All the clusters on the headgroup SOM map to the whole *sn*-2 SOM, showing that the conformations of the two regions are independent. Similar results are obtained when comparing the glycerol and *sn*-2 regions.

C. Coarse graining

Finally, we provide an example of how the information from the SOMs could be used in constructing coarse-grained models. The aim of coarse graining is to design simplified models that include only the relevant degrees of freedom for the problem at hand. The ability of the SOM to find the most relevant states with only minimal human intervention could be particularly useful for the selection of the necessary degrees of freedom. Here we discuss one possible use of this information in constructing a coarse-grained model. We separately focus on each part of the molecule, selecting a minimal coarse-grained description that is able to represent the most relevant conformations of the lipid as represented by the maps.

The headgroup map shows that the most important features in the headgroup region are related to the orientation of the headgroup with respect to the glycerol backbone. This is because there is a fairly small set of specific conformations in this part of the molecule. This finding is in line with atomic-scale molecular dynamics simulations and experiments, that have demonstrated the importance of the *P-N* vector orientation for electrostatic properties at the membrane-water interface.[38,39] To be able to describe these conformations, the coarse-grained headgroup should have at least two particles that define the direction of the *P-N* vector.

The glycerol map shows that the most important conformations in the glycerol region are related to the orientation of the *P-N* vector with respect to the glycerol backbone, and the direction of the first bonds of the tails. Hence, the glycerol region itself does not contain any significant internal degrees of freedom, and can be described by one and two particles. Two particles help us to distinguish the tails from each other, although this could also be achieved by a careful choice of bending potentials and other interactions. Two particles could also make it easier to describe the relative orientation of the glycerol and the *P-N* vector in the *x-y* plane, but again, this could also be achieved by a proper choice of intramolecular interactions.

Finally, the *sn*-2 map shows the lack of any specific important conformations. This indicates that the most pronounced effect of the double bonds is to induce generic disorder in the tail region instead of promoting a set of typical conformations. Hence, the general shape of the tails is the most important feature to consider in the coarse-grained description, and the double bonds can be included by appropriate intramolecular interactions.

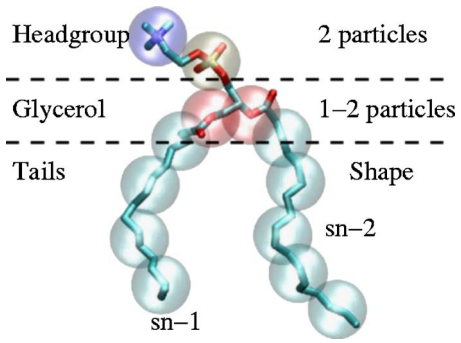

FIG. 15. (Color online) Schematic representation of SOM-derived coarse-grained model (see text for details). The three regions were considered separately based on the different SOMs.

Figure 15 summarizes the model. To complete the model, we would also need to determine the interactions for the different particles, but such analysis is beyond the scope of this article. However, it is interesting to note that the successful coarse-grained model developed by Marrink *et al.*[40] incorporates the features discussed above. That model was constructed based on experiences from the atomistic models, and the similarity demonstrates the possibilities the self-organizing maps have to offer in this context.

VII. DISCUSSION AND CONCLUSIONS

Self-organizing maps have many features that make them useful in the analysis of large amounts of conformational data. They are particularly appropriate for gaining a qualitative understanding of the most important features of a complex system. They do not require any significant *a priori* knowledge of the behavior of the system, which makes them an excellent tool for initial studies. In addition, the ease of visualizing the results helps in gaining additional understanding, which can then be used when planning for further studies. Nevertheless, some care is warranted to confirm the validity of conclusions.

Advantages of SOM. Many standard methods used to characterize lipid bilayers are global in nature. For example, electron density and lateral pressure profiles, as well as deuterium order parameters, give information on the global structure of the bilayer. Such information is important for the general behavior of the bilayer, but it is difficult to relate it to molecular details such as conformations of the molecules or specific interactions between different molecules. In contrast, the information gained by clustering methods such as SOM gives a view on the properties of individual molecules. The clustering methods are particularly effective in reducing the complexity of the molecular configurations to such levels that can be handled by simpler analysis methods or humans. One of the main advantages is that one can study a relatively small set of configurations to form hypotheses about the general behavior of the system, and then test these using other methods.

The main advantage of SOM over other clustering methods is the ease of visualization of the results. Further, this visualization can be made in such a way that it is generally easy for a human analyst to find nontrivial characteristics for the conformations. The information gained in this way can then be used in planning more quantitative analysis, and in some cases also the map itself can provide a basis for further analysis. In many ways, SOM is more qualitative than many other methods, but it also gives more freedom for a careful interpreter to get around the limitations of the method.

The present work provides several examples of these advantages. The clustering of the headgroup conformations gives an idea of the different typical conformations of the headgroup. The SOM makes it easy to group these clusters into larger entities, and thus to find the four major classes (see Sec. IV). It also turns out that different regions of the map have distinct orientations in the bilayer, characterized by the orientation of the glycerol backbone with respect to the membrane normal (data not shown). Further, the ease of visualization and assessment of the sensibility of the clustering were highly beneficial in the study of headgroup dynamics in Sec. VI, as well as in studies of correlations within the different regions.

Using SOMs. SOMs are also relatively easy to use, since the tools are freely available, and only minor modifications are needed for taking into account the periodicity of the angles. However, some care is needed in selecting the size of the map and the training parameters, and some experimentation may be required to take full advantage of the approach. Effective interpretation of the results may also require some effort if one has no previous experience of similar methods.

The effects of the various parameters were discussed in Sec. III, and the qualitative rules described there give a good idea of how the parameters affect the results. Also, the present work gives reasonable initial values for these quantities. After training a map with these values, one can then see whether the level of detail is proper for the use one needs, and possibly do fine-tuning of the values.

First, an absolute upper limit for the size of the map is given by the number of training samples, because there should be a sufficient number (preferably at least a few hundred) of training samples per neuron. Below this limit, one can then choose a size for the map such that a desirable level of detail is achieved. As for the training parameters, the choices described in Sec. II give a reasonable starting point. The most important thing is to have the initial neighborhood radius large enough to allow for the initial organization of the map, and to also have a sufficiently large final neighborhood radius for good visualization properties. Further, long enough training (in practice, slow enough variation of the training parameters) should be used to avoid trapping into a local optimum for the map. Full automation of this process is unfortunately difficult because of the qualitative nature of the SOM and the large amount of human interpretation needed to get full advantage of the map. However, the exact values of the training parameters are not very important, since a rather wide range of values leads to very similar maps.

Evaluating map quality. Evaluating the quality of the trained map is one of the central tasks in determining how successful the SOM approach has been. It is also needed to decide whether the training parameters should be tuned further to obtain better results. The quantization and topographic errors can be used to get an initial idea of the quality of the map. Both measures should be taken into account, as a map can have a very good resolution while having a very poor topology, or vice versa. Poor resolution indicates that the map cannot adequately represent the data, while poor topology makes the visualization of the map less useful. In a good quality map both of these properties are within acceptable limits. However, *a priori* estimation of these limits is not straightforward.

In addition to the above simple error measures, the BMU rank plots have been found to provide a good estimate for the quality of the map. They are particularly useful for quickly assessing information on the clusters and their boundaries on the map. Because distinct clustering increases the amount of information that can be obtained from the SOM, the rank plots can thus be used to assess the usefulness of the map. There are two features that, when present in the rank plot, indicate distinct clustering and sufficient map size: the presence of a significant amount of neurons with low hit counts, and a relatively flat plateau, indicating a peaked distribution of hit counts. In particular, there should not be a tail floating above zero caused by too small a map size. There are two reasons why such features are beneficial. First, neurons that fall between clusters generally have a low hit count, and thus their presence indicates a clear division into clusters. The second reason is related to the shape and size of the clusters: if many neurons have a nearly constant hit count, it indicates that the neurons cover the clusters more or less evenly, which leads to a more desirable cluster structure.

Robustness of SOM. The self-organizing map is a robust tool, as can be seen when comparing the results for the headgroup and glycerol regions, as well as the results for the larger SOM for the whole molecule. The headgroup and glycerol regions overlap partially, and the common region has four major conformations. For all three cases, the SOM is able to find these features without any difficulties. However, the cluster structure within these major conformations is different for the three cases, highlighting the possibility to tailor the method to study interesting features of different parts of the system. Similar robustness can also be seen in the results for the different map sizes.

Despite the robustness, the analysis of the SOM should be done carefully considering all available information, such as the U matrix and the component planes. Otherwise some features of the system might be missed. A good example of this is provided by the headgroup and glycerol data sets, where the U matrix shows only a single prominent boundary, but a more careful analysis uncovers four major conformations. The SOM for the whole lipid shows a similar boundary in the U matrix, but in that case it actually separates different conformations (see previous section). Thus the orientation of the map and the relative positions of the clusters may alter the appearance of the U matrix significantly. This is a property of the similarity measure used (standard Euclidean distance), and thus some other measure might be better for visualizing the U matrix. This could facilitate the analysis by

highlighting different types of cluster boundaries. It could also be interesting to use different distance measures for the training phase.

In the analysis, one should also note that the sequential training algorithm presents the data vectors to the SOM one by one, and therefore the results depend on the order of the data vectors. To minimize this effect, the data vectors are typically presented in a random order. This nondeterministicity contributes to the fact that there does not necessarily exist a simple mapping between clusters in different maps, even if they are trained using identical data. Linear initialization reduces this effect, but it should still be kept in mind when interpreting the results.

Prospects. The automated data analysis performed by the SOM could be taken a step further by using a clustering algorithm for locating clusters in the SOM. The clustering problem in general,[41] as well as clustering of self-organized maps,[42–44] have been studied extensively, and there are several algorithms that could be tested in the current context. This could remove one more manual step from the current procedure, and thus make the procedure less subjective. Such a two-phase process could also reduce the number of model vectors by using single model vectors for each cluster, without losing the many advantages of a larger map. Steps in this direction were already taken when considering the headgroup dynamics.

The SOM approach could also be used for conformational studies beyond single-molecule level. For example, self-organizing maps could be used to classify conformations of phospholipid-cholesterol pairs,[45–47] to study structures of lipid complexes bridged together by salt,[48,49] to characterize conformations of carbohydrate moieties in glycolipids,[50] and to explore conformational degrees of freedom associated with peptides attached to membranes.[51] The main difficulty in this approach is the proper selection of variables for describing the conformations of molecular complexes. However, with a carefully selected set of variables the SOM could give valuable insight into the interaction between the molecules, without the need of making *a priori* assumptions. This could prove advantageous in studies of specific interactions between different molecules over atomistic scales.

The SOM itself can also be used as a starting point for further studies that go beyond the standard approaches often used to characterize structural as well as dynamical properties of biomolecular systems, as we have done for the the conformational dynamics and intramolecular correlations. The data from the map could also be used, for example, for assessing the correlations between neighboring molecules, or for selecting interesting conformations for further analysis. Such information could yield further insight into the specific properties of the molecules in the system. It is also plausible to combine SOM with coarse graining to design simplified models that include only the most relevant degrees of freedom. As shown in this work, the SOM can provide exceptionally useful information for this purpose.

Concluding, the SOM is a promising, relatively simple, and robust tool for a variety of purposes in biomolecular systems. Efforts to develop the method further, as well as applications to other lipid systems, are in progress.

ACKNOWLEDGMENTS

The authors would like to thank Siewert-Jan Marrink and Marja Hyvönen for fruitful discussions and Samuli Ollila for providing the MD simulation data. This work has, in part, been supported by the Academy of Finland through its Center of Excellence Program, Grant no. 80246 [to one of the authors (I. V.)], and through Grant no. 109514 for working abroad [to another author (E. F.)], the National Graduate School in Materials Physics [to two of the authors (T. M. and E. F.)], and the National Graduate School in Nanoscience [to the first author (T. M.)]. Two of the authors (T. M. and M. K.) equally contributed to this article.

[1] O. G. Mouritsen, *The Frontiers Collection* (Springer, Berlin, 2005).
[2] D. P. Tieleman, S. J. Marrink, and H. J. C. Berendsen, Biochim. Biophys. Acta **1331**, 235 (1997).
[3] S. E. Feller, Curr. Opin. Colloid Interface Sci. **5**, 217 (2000).
[4] H. L. Scott, Curr. Opin. Struct. Biol. **12**, 495 (2002).
[5] W. L. Ash, M. R. Zlomislic, E. O. Oloo, and D. P. Tieleman, Biochim. Biophys. Acta **1666**, 158 (2004).
[6] I. Vattulainen and M. Karttunen, in *Computational Nanotechnology*, edited by M. Rieth and W. Schommers (American Scientific, Stevenson Ranch, CA, 2006).
[7] F. A. Hamprecht, C. Peter, X. Daura, W. Thiel, and W. F. van Gunsteren, J. Chem. Phys. **114**, 2079 (2001).
[8] S. Haykin, *Neural Networks—A Comprehensive Foundation* (Prentice-Hall, Englewood cliffs, NJ, 1999).
[9] T. Kohonen, *Self-organizing Maps*, Springer Series in Information Science, Vol. 30, 3rd ed. (Springer, Berlin, 2001).
[10] J. Schuchhardt, G. Schneider, J. Reichelt, D. Schomburg, and P. Wrede, Protein Eng. **9**, 833 (1996).
[11] M. A. Andrade, G. Casari, C. Sander, and A. Valencia, Biochemistry **76**, 441 (1997).
[12] S. Mahony, D. Hendrix, T. J. Smith, and A. Golden, Artif. Intell. Rev. **24**, 397 (2005).
[13] S. Mahony, D. Hendrix, A. Golden, T. J. Smith, and D. S. Rokhsar, Bioinformatics **21**, 1807 (2005).
[14] Z. R. Yang and K. C. Chou, J. Chem. Inf. Comput. Sci. **43**, 1748 (2003).
[15] J. T. Ayers, A. Clauset, J. D. Schmitt, L. P. Dwoskin, and P. A. Crooks, AAPS J. **25**, E678 (2005).
[16] M. T. Hyvönen, Y. Hiltunen, W. El-Deredy, T. Ojala, J. Vaara, P. T. Kovanen, and M. Ala-Korpela, J. Am. Chem. Soc. **123**, 810 (2001).
[17] A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **52**, 3730 (1995).
[18] A. P. Lyubartsev, M. Karttunen, I. Vattulainen, and A. Laaksonen, Soft Mater. **1**, 121 (2003).
[19] T. Murtola, E. Falck, M. Patra, M. Karttunen, and I. Vattulainen, J. Chem. Phys. **121**, 9156 (2004).
[20] S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth, J. Chem. Phys. **120**, 10896 (2004).
[21] Q. Shi and G. A. Voth, Biophys. J. **89**, 2385 (2005).
[22] A. A. Louis, P. G. Bolhuis, J. P. Hansen, and E. J. Meijer, Phys. Rev. Lett. **85**, 2522 (2000).
[23] P. G. Bolhuis, A. A. Louis, J. P. Hansen, and E. J. Meijer, J. Chem. Phys. **114**, 4296 (2001).
[24] R. Faller, Polymer **45**, 3869 (2004).
[25] M. Kupiainen, E. Falck, S. Ollila, P. Niemelä, A. A. Gurtovenko, M. T. Hyvönen, M. Patra, M. Karttunen, and I. Vattulainen, J. Comput. Theor. Nanosci. **2**, 401 (2005).
[26] S. Ollila, M. T. Hyvönen, and I. Vattulainen, J. Phys. Chem. B (in press).
[27] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, in Proceedings of the Matlab DSP Conference 1999 (Espoo, Finland, 1999), see also http://www.cis.hut.fi/projects/somtoolbox/
[28] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[29] M. Bachar, P. Brunelle, D. P. Tieleman, and A. Rauk, J. Phys. Chem. B **108**, 7170 (2004).
[30] S. Nosé, Mol. Phys. **52**, 255 (1984).
[31] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
[32] M. Parrinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981).
[33] S. Nosé and M. L. Klein, Mol. Phys. **50**, 1055 (1983).
[34] T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**, 10089 (1993).

[35] U. Essmann, L. Perera, M. L. Berkowitz, H. L. T. Darden, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).
[36] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[37] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining* (MIT, Cambridge, MA, 2001).
[38] M. Langner and K. Kubica, Chem. Phys. Lipids **101**, 3 (1999).
[39] L. Saiz and M. L. Klein, J. Chem. Phys. **116**, 3052 (2002).
[40] S. J. Marrink, A. H. de Vries, and A. E. Mark, J. Phys. Chem. B **108**, 750 (2004).
[41] A. K. Jain, M. N. Murty, and P. J. Flynn, ACM Comput. Surv. **31**, 264 (1999).
[42] J. Vesanto and E. Alhoniemi, IEEE Trans. Neural Netw. **11**, 586 (2000).
[43] M. Y. Kiang, Comput. Stat. Data Anal. **38**, 161 (2001).
[44] S. Wu and T. W. S. Chow, Pattern Recogn. **37**, 175 (2004).
[45] E. Falck, M. Patra, M. Karttunen, M. T. Hyvönen, and I. Vattulainen, Biophys. J. **87**, 1076 (2004).
[46] S. A. Pandit, D. Bostick, and M. L. Berkowitz, Biophys. J. **86**, 1345 (2004).
[47] S. A. Pandit, S. Vasudevan, S. W. Chiu, R. J. Mashl, E. Jakobsson, and H. L. Scott, Biophys. J. **87**, 1092 (2004).
[48] A. A. Gurtovenko, M. Miettinen, M. Karttunen, and I. Vattulainen, J. Phys. Chem. B **109**, 21126 (2005).
[49] R. A. Böckmann and H. Grubmüller, Angew. Chem., Int. Ed. **43**, 1021 (2004).
[50] T. Róg, I. Vattulainen, and M. Karttunen, Cell. Mol. Biol. Lett. **10**, 625 (2005).
[51] M. Ø. Jensen, O. G. Mouritsen, and G. H. Peters, Biophys. J. **86**, 3556 (2004).