Perspectives to Coarse-Graining through Self-Organizing Maps

Teemu Murtola[1], Mikko Kupiainen[1], Emma Falck[2], and Ilpo Vattulainen[1,3,4]

[1] Helsinki University of Technology, Laboratory of Physics,
P.O. Box 1100, FI–02015 Espoo, Finland
[2] Beckman Institute for Advanced Science and Technology,
University of Illinois at Urbana-Champaign,
405 N. Mathews Ave., Urbana, IL 61801, United States
[3] Tampere University of Technology, Institute of Physics,
P.O. Box 692, FI–33101 Tampere, Finland
[4] MEMPHYS – Center for Biomembrane Physics, University of Southern Denmark
tkm@fyslab.hut.fi; http://www.fyslab.hut.fi/bio/

Abstract. We discuss the use of self-organized maps, an unsupervised learning algorithm based on neural networks, as a tool in a coarse-graining process of biomolecular systems, particularly membranes. First, some results on the conformational analysis of unsaturated lipids in a bilayer are presented. These show that the method is able to extract relevant information about the most important conformations of the molecule. Based on these results, we then discuss applications of this method in construction of coarse-grained models. The properties of the method make it an attractive candidate to be used as a tool to more systematically construct structural descriptions for coarse-grained models.

1 Introduction

In many particle-based coarse-graining processes [1], the construction of the coarse-grained model can be essentially divided into two major phases. First, the system is reduced to a simpler one by reducing the degrees of freedom. This is commonly done by grouping several atoms together to obtain effective interaction sites. In the second phase, effective interactions for the coarse-grained particles are constructed. It would be desirable to do the coarse-graining as systematically as possible by utilizing information from e.g. simulations of the more detailed model and from experiments. To date, the research on coarse-graining has mostly focused on the second phase, and several different methods have been proposed for systematical construction of the effective interactions.

The choices made during the first phase can severely limit the descriptive power of the model if they are not made properly. On the other hand, the simpler the model, the faster it is to simulate and the larger systems can be studied. It is also possible to include internal states for the coarse-grained particles. This can be particularly useful for simpler models, where the coarse-grained structural description is not able to capture all the important features of the underlying

system, such as the ordering of lipid tails. The results from a phenomenological model by Nielsen et al.[2] show that even a very simple, two-state description of the ordering can give qualitatively correct phase diagrams for a binary phospholipid/cholesterol mixture. The ability to systematically derive such models could thus provide an attractive approach to studies of large-scale structure of e.g. membrane systems. Also, systematic methods for selecting interaction centers would help in constructing efficient models.

Molecular conformations are closely linked to the choice of the interaction centers, as well as the possible internal states of the coarse-grained particles. Hence, analysis of the conformations can also give valuable insight to be used in the coarse-graining process. In a recent study [3], we have demonstrated that self-organizing maps (SOMs) [4] can provide a useful tool for analysis and classification of the molecular conformations. The SOM method is an unsupervised learning strategy based on neural networks. It has the advantage that it does not require any significant a priori knowledge of the behavior of the system, and thus provides a sound basis for systematic analysis of the conformations.

In this contribution, we present the SOM method, and describe our results for the conformational analysis of lipid molecules. Based on these results, several applications to coarse-graining are then discussed. The discussion is focused on membrane systems, but same principles could also be applied to other biomolecular systems.

2 Self-Organizing Maps

A self-organizing map (SOM) is a powerful software tool for the visualization of high-dimensional data [4]. Conceptually, it is a mapping from high-dimensional input data vectors into a low-dimensional (typically two-dimensional) grid of so-called neurons. A model vector is associated with each neuron, describing a typical data vector mapped to that neuron. Similar data vectors are mapped to the same or neighboring neurons, which is the origin of the self-organizing behavior of the map. The self-organizing map uses only relatively few model conformations to describe the data, and hence it produces an abstraction of the data. This feature can be used to extract general characteristics and to find the most relevant features of the data.

SOM analysis consists of four phases: selection of variables, initialization of the model vectors, training of the map, and finally analysis of the trained map. A freeware implementation of the SOM algorithm is available [5], making the analysis easy to implement.

Selection of Variables. The first phase is specific to the data that are being analyzed. For conformational analysis it means that we have to describe the shape of the molecule as an n-dimensional vector. One convenient choice is to use all or some of the dihedral angles of the molecule to describe the conformations. Other variables may be beneficial in analysis of more complex structures or mutual conformations of several molecules.

Initialization and Training. Before training, some initial values need to be given to the model vectors. Commonly used method for this is to use linear initialization where the model vectors are positioned on a regular grid on the hyperplane formed by the principal values of the data distribution.

The training has two goals: to find such model vectors that represent the typical features of the data as precisely as possible, and simultaneously to preserve the topological structure of the original data. During the training, the data vectors are traversed one by one, and at each step the most similar model vector is moved towards the data vector. In addition to the most similar neuron, the model vectors of neighboring neurons are also updated, although by a smaller amount. This gives rise to the self-organization of the map. The size of the updated neighborhood, as well as the magnitude of the changes, is determined by a so-called neighborhood function. Typically a smooth Gaussian shape is used for the neighborhood function, and both its magnitude and variance are decreased during the training. Thus 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.

Analysis. Finally, the trained map is analyzed. This is done mostly manually, using different visualizations of the map and analyzing the model vectors in different parts of the map. However, the ease of visualization and the self-organizing behavior of the map makes visual identification of relevant areas of the map relatively easy.

3 Application to Conformational Analysis

In a recent study [3], we applied the SOM analysis to a set of molecular conformations produced by a 50 ns atomistic molecular dynamics simulation of a 1-palmitoyl-2-linoleoyl-*sn*-glycero-3-phosphatidylcholine (PLPC) bilayer. The goals of the study were threefold: to consider the possibilities and limitations of the method in the context of lipid conformations, to analyze a specific lipid system, and to gain better understanding of the methodology for further work.

Results of the study were positive: the SOM analysis was able to find distinct conformations from the data, and also to classify these conformations into more general classes. For example, the conformations of the headgroup and glycerol regions were classified into three classes distinguished by different orientation of the headgroup with respect to the glycerol backbone. These classes were further divided into different conformations based on the orientation of the P–N-vector and the orientation of the beginning of the tails with respect to the glycerol. For the diunsaturated linoleoyl chain, the different conformations of the double bond region were clearly classified into different regions on the map. The obtained results demonstrate that SOMs provide a useful and robust tool for analysis of lipid conformations.

4 Discussion: Prospects in Coarse-Graining

Information gained by self-organizing maps has several applications in coarse-graining. These applications are based on the ability of the SOM to robustly find and classify the most relevant conformational states of the molecules.

First of all, the model conformations could be used to select the interaction centers for the coarse-grained model. An efficient choice would be such that also the coarse-grained description would be able to distinguish between the most important conformations. The model conformations could also give clues about the form of possible intramolecular constraints such as bond angles. The information on the most relevant states could also be used to assign internal states to the coarse-grained particles as described in the introduction.

As an example, consider coarse-graining of a bilayer containing unsaturated lipids. The SOM analysis of the chain conformations could give valuable information about how the interaction centers could be chosen to represent the conformational variability of the chain, simultaneously taking into account the constraints induced by the double bonds. SOM could also be applied to the glycerol and the headgroup to find the most relevant conformations in these regions. As described above, these show that the largest differences in the conformations relate to the junction between the glycerol and the headgroup, and to the orientation of the P–N-vector. Based on the analysis, the interaction centers could be chosen such that they can efficiently describe these different conformations.

The SOM analysis could also be extended beyond single-molecule level to analyze pairs of molecules, for example phospholipid–cholesterol pairs in a binary mixture. Such a study could yield additional information about the most relevant conformations with respect to the intermolecular interactions, and thus further aid in the selection of the coarse-grained description.

To summarize, self-organizing maps provide a promising tool for more or less systematic selection of coarse-grained interaction centers, as well as for choosing the form of the intramolecular interactions. Hence, the method can complement the various methods developed for systematically obtaining the effective interactions after the coarse-grained description has been selected.

References

1. Karttunen, M., Vattulainen, I., Lukkarinen, A. (Eds.): Novel Methods in Soft Matter Simulations. Springer, Berlin (2004).
2. Nielsen, M., Miao, L., Ipsen, J. H., Zuckermann, M. J., Mouritsen, O. G.: Off-lattice model for the phase behavior of lipid-cholesterol bilayers. Phys. Rev. E **59** (1999) 5790–5802
3. Kupiainen, M., Murtola, T., Falck, E., Vattulainen, I.: Conformational analysis of lipid molecules using self-organizing maps. (submitted to J. Phys. Chem. B)
4. Kohonen, T.: Self-Organizing Maps. 3rd ed., Springer, Berlin (2001)
5. Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J.: Self-organizing maps in Matlab: the SOM Toolbox, in *Proceedings of the MatLab DSP Conference 1999*, Finland, 35–40