Self Organizing Maps to efficiently cluster and functionally interpret protein conformational ensembles

Domenico Fraccalvieri* [1] Laura Bonati [1] Fabio Stella [2]
[1]Department of Earth and Environmental Sciences, University of Milano Bicocca, Milano, IT
[2]Department of Informatics, Systems and Communication, University of Milano Bicocca, Milano, IT
*Corresponding Author = domenico.fraccalvieri@unimib.it

1 Introduction

The wide range of protein biological functions, such as enzymatic activity, ligand- and protein-protein interactions and allosteric regulation, is strictly related to their flexibility and dynamics[1]. To model the influence of protein motions across this broad spectrum of events Molecular Dynamics (MD) simulations are now routinely used. The identification of the most functionally relevant conformations is generally done by grouping the conformations according to a criterion of geometrical similarity and popular choices include hierarchical clustering, single, complete and average linkage and k-means [10]. These geometrical approaches rely on the assumption that the identified conformational states also correspond to the energetic states [10]. Good candidates to improve this match are Artificial Neural Networks (ANNs) which are capable to discover the relationships between the measured variables only from the available dataset [9]. Among the class of ANNs, successfully applied to artificial life systems [2], Self Organizing Maps (SOMs) [6] represent a particularly powerful data driven model that has been widely applied for exploration and clustering of high-dimensional datasets [9, 6]. We recently developed an approach which combines SOMs and hierarchical clustering to efficiently compare conformational ensembles obtained from multiple MD simulations of proteins [3]. To reliably apply the SOM analysis to these specific input data we identified and optimized a small number of SOM parameters. In particular, we confirmed that the map size is a crucial parameter and that a well-selected number of neurons is crucial both to reduce the computational cost of the analysis and to provide an intermediate topological representation of the input conformational space [3]. As a result the original MD trajectories can be represented by the few conformations that best represent the clusters obtained. Here we make a step further; the proposed approach consists of processing all the atom positions of each conformation won by a given SOM neuron using a similarity network. Different similarity measures between atoms behavior are used to compile a similarity matrix which is inputted to a network model [8]. Network algorithms are then used to automatically discover and interpret the behavior of the original protein conformational ensembles exploiting the atomic coordinates enclosed in the SOM neuron.

2 Methods

Self Organizing Map clustering. The details of the proposed SOM approach can be summarized as follow. First, an ensemble of conformations is extracted from one (or more) MD trajectory with an optimized sampling rate of 1/100 ps. Each conformation of the trajectories is represented using only the Cα Cartesian coordinates. The ensemble is encoded in a matrix and is used for the learning stage,

A. Graudenzi, G. Caravagna, G. Mauri, M. Antoniotti (Eds.):
Wivace 2013 - Italian Workshop on
Artificial Life and Evolutionary Computation
EPTCS 130, 2013, pp. 83–86, doi:10.4204/EPTCS.130.13

performed with the set of optimal SOM parameters previously identified by experimental design [3]. The most relevant parameters are: *Map size=100, Radius=3, Training length=5000 and Neighbor function=gaussian*. The neurons of this trained SOM, i.e. the proto-clusters of the original conformational ensemble, are grouped using the complete linkage algorithm and the number of clusters is automatically selected by means of the Mojena's stopping rule [7]. To provide a graphical representation of the neurons and their connections, each neuron is described by a hexagon and the neurons belonging to each cluster are homogeneously colored (see Figure 1, central panel). Finally, to provide the most synthetic structural representation of the original ensemble, only the input conformation closest to the centroid of each cluster is represented (see Figure 1, left panel). The centroid is defined as the mean vector of all the prototype vectors of the neurons in that cluster.

Network Analysis. The set of all the Cα's atomic coordinates of the conformations won by each SOM neuron, are used to compile a network model. The following similarity measures were analyzed and empirically compared; x-y-z average absolute correlation, cosine, Pearson, Spearman correlation and biweight mid-correlation. Networks are constructed from similarity matrices by means of threshold as described in [5]. The SOM, which summarizes the atomic positions of each conformation associated with a neuron, allows building a network model which is then exploited to automatically discover and interpret the behavior of the inputted protein conformational ensembles. Graph partitioning algorithms [5] to cluster atoms with similar behavior and community detection algorithms [8, 5] were applied to the network model (see Figure 1, right panel).

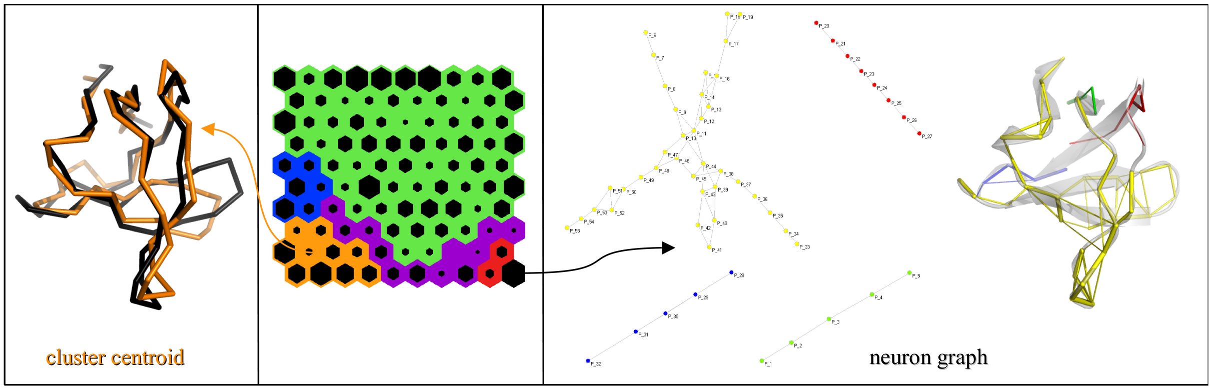

Figure 1: Central panel: the clustered SOM. The background colors of the neurons (hexagons) correspond to the cluster attribution. The inner hexagon is proportional to the number of conformations won by the neuron. Left panel: ribbon representation of one centroid cluster (in orange) superimposed to a reference structure (in black). Right panel: graphical analysis of one neuron. On the left, each node represents an atom of the neuron conformations and the edges link pairs of correlated atoms. The groups of nodes (atoms) with similar behavior are homogeneously colored. On the right, the nodes and the edges are represented on a reference 3D structure (in gray) using the same color code of the graph. The width of the edge is proportional to the correlation among the atoms (nodes).

3 Results and Discussion

The SOM approach developed by our group has been successfully applied showing its potential in the comparison of MD simulations of different proteins with dimensions ranging from 60 to 500 amino acids[3, 4]. These applications highlighted key features, related to the topological nature of the SOM. First, a single map was able to detect both low fluctuations, i.e. little conformational changes with respect to a reference structure, and large concerted motions, i.e. simultaneous conformational changes of different region of the protein.[3, 4]. Second, the intrinsic time dependence of the data, not given as input, was correctly recovered; in fact consecutive conformational transitions were correctly assigned to adjacent neurons, thus to adjacent clusters, on the map [3]. Moreover, the values of geometrical descriptors based on a subset of Cα atoms included in the SOM training [3, 4], as well as excluded features [4], were consistently clustered, indicating the potential of the map to classify unknown features [4]. Finally the ability of a SOM trained with a set of protein to classify conformations of similar protein was verified and exploited [4]. The extension of our protocol here presented is aimed at a deeper inspection of the dynamic information contained in each neuron in the output map. In fact, not only the subset of conformations won by each neuron, but also the specific motions caught by that neuron are analyzed. To this end, the matrices composed by the Cα Cartesian coordinates of the original data in each neuron are analyzed using similarity matrices and represented by graphs (see Figure 1, right panel). This analysis has a double outcome. First it provides an interesting and informative picture of the joint behavior of the atoms belonging to the conformations in each neuron, highlighting both local motions, i.e. motions involving structurally neighbor atoms, and long-range communications mediated by atoms able to transmit conformational signals to remote sites of the protein. Second this information can be used to cluster the data according to this graph. Cluster analysis based on similarity of the Cα dynamic behavior enclosed in each SOM neuron should increase the ability of our approach to compare the dynamics of different proteins. In conclusion, we already showed that the use of this SOM approach to cluster protein conformational ensembles increases the cluster quality compared to other methods routinely used [3]. Moreover both the approaches proposed for the map interpretation allow an efficient comparison of the dynamics of different proteins. In the former approach, the characteristics of the conformations belonging to each cluster are summarized and described by the cluster centroid, in the latter, the characteristic pathways of motion described by each neuron are highlighted by the graphs (Figure 1). Applications of this approach should range from protein engineering (to design specific mutation able to reduce, or increase, the function of a protein) to computer-based drug design (to select specific conformations of a protein suitable to interact with specific ligands).

References

[1] D.D. Boehr, R. Nussinov & P.E. Wright (2009): *The role of dynamic conformational ensembles in biomolecular recognition*. *Nat. Chem. Biol.* 5(11), pp. 789–796, doi:10.1038/nchembio.232.

[2] A. Cangelosi, Nolfi S. & Parisi D. (2003): *Artificial Life Models of Neural Development*. In: *On Growth, form and Computers*, Elsevier Academic Press, pp. 339–52. doi:10.1016/B978-012428765-5/50051-7

[3] D. Fraccalvieri, A. Pandini, F. Stella & L. Bonati (2011): *Conformational and functional analysis of molecular dynamics trajectories by Self-Organising Maps*. *BMC Bioinformatics* 12, p. 158, doi:10.1186/1471-2105-12-158.

[4] D. Fraccalvieri, M. Tiberti, A. Pandini, L. Bonati & E. Papaleo (2012): *Functional annotation of the mesophilic-like character of mutants in a cold-adapted enzyme by self-organising map analysis of their molecular dynamics*. *Mol. Biosyst.* 8, pp. 2680–91, doi:10.1039/C2MB25192B.

[5] S. Horvath (2011): *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer. doi:10.1007/978-1-4419-8819-5

[6] T.K. Kohonen (2013): *Essentials of the Self-Organizing Map*. Neural Networks 37, pp. 52–65, doi:10.1016/j.neunet.2012.09.018.

[7] R. Mojena (1977): *Hierarchical Grouping Methods and Stopping Rules: An Evaluation.* Comput. J. 20(4), pp. 359–63, doi:10.1093/comjnl/20.4.359.

[8] M.E.J. Newman (2010): *Networks: An Introduction*. Oxford University Press, New York.

[9] A. Pandini, D. Fraccalvieri & L. Bonati (2013): *Artificial Neural Networks for Efficient Clustering of Conformational Ensembles and their Potential for Medicinal Chemistry*. Curr. Top. Med. Chem. 13(5), pp. 642–51, doi:10.2174/1568026611313050007.

[10] J. Shao, S.W. Tanner, N. Thompson & T.E. Cheatham (2007): *Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms*. J. Chem. Theory Comput. 3(6), pp. 2312–34, doi:10.1021/ct700119m.