Subject Section

SOMMD: An R Package for the Analysis of Molecular Dynamics Simulations using Self-Organising Map

Stefano Motta[1,*], Lara Callea[1], Shaziya Ismail Mulla[2], Hamid Davoudkhani[2] Laura Bonati[1], and Alessandro Pandini[2,3,*]

[1]Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milan 20126, [2]Department of Computer Science, Brunel University of London, Uxbridge UB8 3PH, U.K., [3]TheThomas Young Centre for Theory and Simulation of Materials, London SW7 2AZ, U.K.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Molecular Dynamics (MD) simulations provide critical insights into biomolecular processes but they generate complex high-dimensional data that are often difficult to interpret directly. Dimensionality reduction methods like Principal Component Analysis (PCA), Time-Lagged Independent Component Analysis (TICA) and Self-Organising Maps (SOMs) have helped in extracting essential information on functional dynamics. However, there is a growing need for a user-friendly and flexible framework for SOM-based analyses of MD simulations. Such a framework should offer adaptable workflows, customizable options, and direct integration with a widely adopted analysis software.

Results: We designed and developed SOMMD, an R package to streamline MD analysis workflows. SOMMD facilitates the interpretation of atomistic trajectories through SOMs, providing tools for each stage of the workflow, from importing a wide range of MD trajectories data types to generating enhanced visualizations. The package also includes three example projects that demonstrate how SOM can be applied in real-world scenarios, including cluster analysis, pathways mapping and transition networks reconstruction.

Availability: SOMMD is available on CRAN (https://CRAN.R-project.org/package=SOMMD) and on GitHub (https://github.com/alepandini/SOMMD).

Contact: stefano.motta@unimib.it; alessandro.pandini@brunel.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Molecular Dynamics (MD) simulations are invaluable tools to study the dynamic behaviour of biomolecules, offering a detailed view of molecular processes at the atomistic level and the ability to gain insights on the interplay between conformations and functions (Dror et al., 2012). However, to understand these molecular processes, it is crucial to describe the conformational states sampled by the system during the simulation and their relationships (Hollingsworth & Dror, 2018).

Exploring the conformational space through MD simulations can generate high dimensional data due to the number of degrees of freedom in complex molecular systems. A significant challenge lies in interpreting this large volume of data and transforming it into meaningful representations that can reveal the relationships between different states (Hollingsworth & Dror, 2018). Such representations should be not only informative but also

K.Takahashi et al.

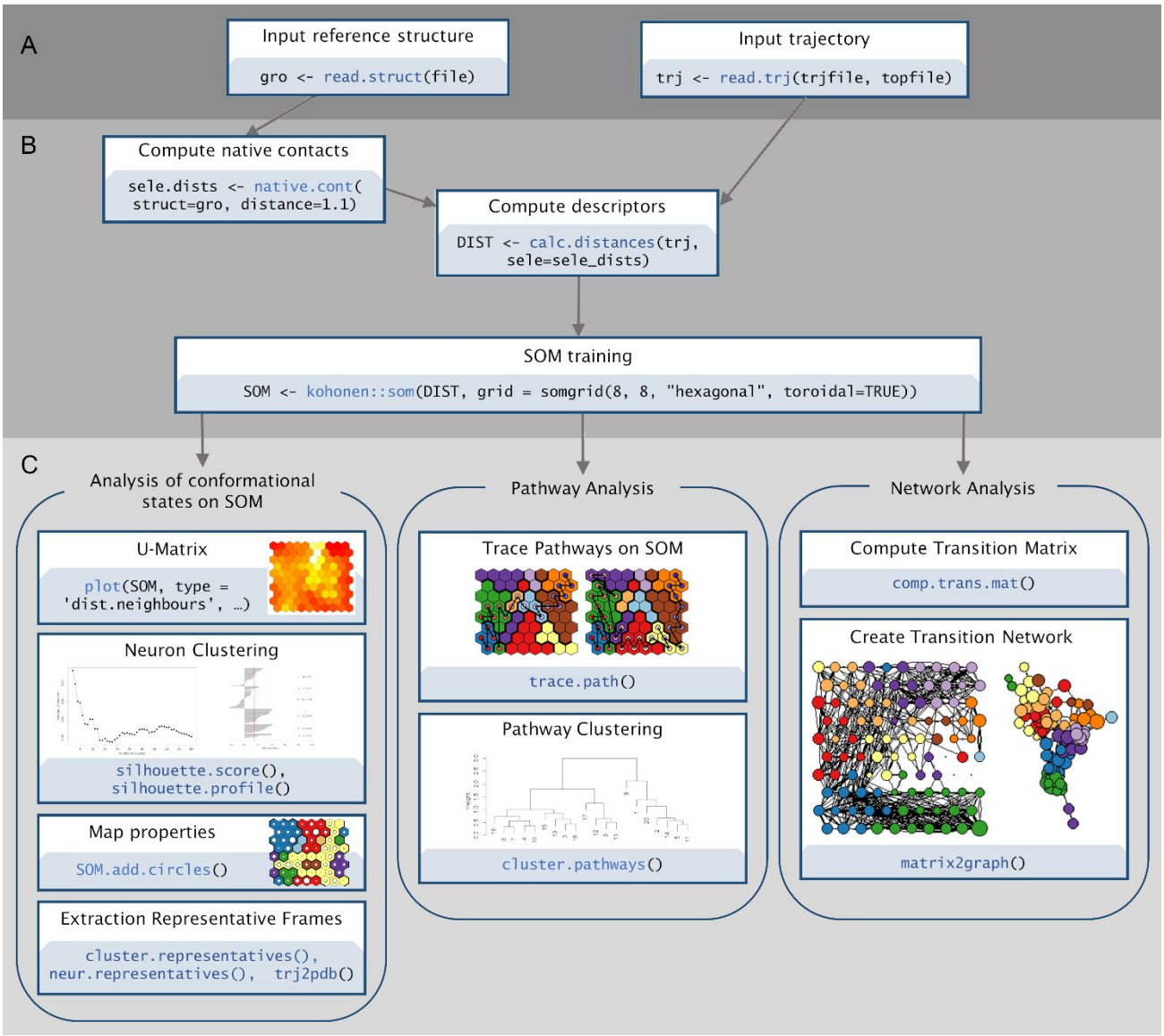


Figure 1: Summary diagram of the SOMMD package architecture and analysis workflow. The package provides functions to read the structure topology and trajectory files (A. *input*), to compute the desired conformational descriptors and to train the SOM (B. *preprocessing and training*). From a trained SOM analysis and visualization can be performed on three levels: conformational states, pathways and transition networks (C. *analysis and visualization*).

interpretable, enabling researchers to gain insights into the functional dynamics of the systems being studied. Dimensionality reduction methods, which transform high-dimensional data into lower-dimensional representations, have become indispensable in this context. Among the available approaches, methods preferentially used for simulation data include Principal Component Analysis (PCA), Time-Lagged Independent Component Analysis (TICA) (Molgedey & Schuster, 1994; Noé & Nüske, 2013), and Self-Organising Maps (SOMs) (Bouvier et al., 2015; Fraccalvieri et al., 2011; Kohonen, 2013; Shao et al., 2007; T, 1990).

SOMs provide a distinctive approach to organise complex, high-dimensional data into a two-dimensional grid of neurons, where each neuron represents a specific region in the data space. Topological relationships between these regions in the original data space are preserved in the SOM. In recent years, SOMs have been successfully used to cluster conformational ensembles (Bouvier et al., 2015; Fraccalvieri et al., 2011; Motta et al., 2023), and to reconstruct conformational pathways of protein (un)folding (Hendrix et al., 2022; Motta et al., 2021), protein-protein binding (Yuan et al., 2024a, 2024b) and ligand binding (Motta et al., 2022; Tripathi & Nair, 2023), with the ability to provide estimates of binding kinetics constants (Callea et al., 2024; Rubina & Moin, 2023). Despite their potential, there is currently no user-friendly and extendable framework that allows researchers to work seamlessly in a single environment and to develop ad-hoc workflows for specialized SOM-based analyses. To address this gap, we have chosen the R environment for its versatility and extensive ecosystem of packages. Currently available packages on CRAN (e.g. kohonen (Wehrens & Kruisselbrink, 2018), aweSOM (Boelaert et al., 2021), popsom7 (Hamel et al., 2025)) are general-purpose implementations that have not been designed to process MD data and do not include functions to reconstruct conformational pathways and model transitions between states. Therefore, we have developed SOMMD, a dedicated package for the analysis of MD simulations using SOMs.

SOMMD is an R package specifically designed to meet the needs of the MD community by providing a format-agnostic tool, support for processing large datasets, flexible user-defined geometrical descriptors of dynamics, effective visualization tools, and the ability to convert trajectory data into interpretable models of conformational transitions. To this end, SOMMD offers R functions to read and process MD trajectories of various formats, calculate descriptors for SOM training, extract representative structures of the sampled states, trace pathways followed during the simulations on the trained SOM, construct transition graphs between SOM neurons, and generate customizable visualizations for effective analysis. SOMMD offers a user-friendly and integrated solution, allowing researchers to easily model and interpret the dynamics of biomolecular systems using SOMs.

The SOMMD package includes both pre-defined workflows in the form of R notebooks which serves as "scenario recipes", as well as a set of R functions for developing ad-hoc workflows for the analysis of molecular simulations. Simulation data required to execute the workflows are hosted on Figshare (see Data availability) and can be automatically downloaded and loaded from the notebooks.

2 Description

The SOMMD package was designed to streamline the generation of interpretable Self-Organising Maps from MD data in R. To this end it takes advantage of powerful data classes from bio3d (Grant et al., 2006) and machine learning functions from the kohonen (Wehrens, 2007; Wehrens & Kruisselbrink, 2018) and cluster (Maechler et al., 2022) packages, but it extends them. Specifically, SOMMD introduces a more general-purpose trajectory class and dedicated visualization functions for MD data. The architecture of SOMMD is structured into three distinct components, each fulfilling a specific role in the analysis workflow (Figure 1):

A) input: this initial step involves the handling and preprocessing of input data, introducing a new core class for trajectory data ('trj'). An additional class is provided to represent structural files in GROMOS format ('gro'). These two new classes are designed to complement the existing molecular structure classes in the bio3d package. This implementation ensures compatibility with widely used input formats and extends the capabilities of the bio3d package (Grant et al., 2006) for structural and trajectory processing.

B) preprocessing and training: the input data is pre-processed to compute descriptors for the variables used in training the SOM. These descriptors typically consist of a subset (or the entire set) of relevant interatomic distances for a group of atoms, but the package also supports user-defined geometrical measures. The map is trained using a wrapper around functions provided by the kohonen package (Wehrens and Kruisselbrink, 2018; Wehrens, 2007).

C) analysis and visualization: this step provides functions to analyse and identify microstates and macrostates. Time-dependent relationships between states can be reconstructed and visualised as pathways on the map and as transitions in a graph model. SOMMD also offers functions for ad-hoc mapping of time-dependent properties on both the SOM and the transition graph. Various workflows can easily be constructed to extract system-specific information tailored to the biomolecular process under investigation.

The next section presents three example scenarios that are included as R Markdown notebooks in the package. These notebooks serve as prototypical examples of the most common use cases for SOMMD.

3 Usage scenarios

The R package includes example notebooks that illustrate how SOM analysis can be performed on previously published and validated cases.

Clustering of MD trajectories: The first scenario provides a brief introduction to SOM training with SOMMD. In this case, the user is interested in describing the conformations sampled during MD simulations. The study case is a set of multiple unbiased simulations of the FOXP1 protein DNA-binding domain. By applying SOMMD, it is possible to extract representative structures of clusters and create informative plots for selected property. Additionally, sampling can be assessed by remapping multiple replicas.

Analysis of Pathways: This scenario demonstrates how to use SOMMD to compare different replicas for a process of interest recovering alternative pathways. In the present case, the process was the unfolding of a protein domain generated through steered MD (Motta et al., 2021). SOMMD can be used to obtain a clustering of pathways based on the sequence of neurons sampled during the different replicas.

Transition network analysis: This scenario demonstrates how to build a transition matrix starting from the mapping of each frame of the simulation on the SOM. Starting from a metadynamics reconstruction of the ligand-binding process (Callea et al., 2021), an approximate transition matrix is built according to the observed number of transitions between pairs of neurons. The visualization of the SOM neurons as a graph provides a unified picture of the sampled pathways.

Summary descriptions of the workflows and detailed results for each scenario are reported in the Supplementary Data (sections 1-3). Moreover, for users working with large datasets, Supplementary Section 4 outlines practical tips for managing memory and computational efficiency, ensuring that SOMMD remains scalable and effective even on standard workstations.

4 Conclusions

The SOMMD package is a versatile tool for the analysis of molecular conformations using unsupervised learning. It integrates dimensionality reduction and clustering of conformational states sampled during MD simulations. Additionally, through dedicated functions, it facilitates the visualization of alternative pathways in molecular processes and the construction of state graphs based on the transition matrix between pairs of neurons.

SOMMD addresses the need for a comprehensive framework to generate interpretable and informative representations of conformational states and their interrelationships. While other methods for dimensionality reduction, such as PCA and TICA, are commonly used for this scope, SOMs have the distinct advantage of preserving the topological relationships between microstates, providing a visual model for clustering data into distinct regions of the map. Supplementary Section 5 provides a detailed comparison between SOMs and PCA, demonstrating that SOMs offer a more intuitive and comprehensive understanding of molecular unfolding pathways with only a minimal increase in computational time.

SOMMD also supports customizable visualizations, simplifying the process to map specific properties or features onto the SOM grid. This flexibility can be valuable for researchers interested in visualizing and understanding system-specific properties from MD simulation data.

The package is designed for efficient analysis of large datasets within the limits of available RAM. Details on computational performance and strategies to mitigate memory limitations are provided in Supplementary Section 4. Like other methods for analysing conformational dynamics, SOMMD is limited by the sampling in the original trajectory, as detection

of transition pathways and state modeling requires sufficient sampling of critical functional events. Additionally, the package does not include an automated method for selecting the best input features, so identifying the important degrees of freedom is a prerequisite.

Acknowledgements

The authors would like to thank Anja Estermann for useful suggestions on the design and implementation of the code.

Funding

This project made use of time on HPC granted via the UK High-End Computing Consortium for Biomolecular Simulation, HECBioSim (https://www.hecbiosim.ac.uk), supported by EPSRC [EP/X035603/1].

Conflict of Interest: none declared.

Data availability

The MD trajectory data required to run the examples in SOMMD are openly available on Figshare under CC-BY licence, and the accompanying R notebooks in the package include the code to automatically download and process these datasets.

References

Bouvier, G., Desdouits, N., Ferber, M., Blondel, A., & Nilges, M. (2015). An automatic tool to analyze and cluster macromolecular conformations based on self-organizing maps. *Bioinformatics*, 31(9), 1490–1492. https://doi.org/10.1093/bioinformatics/btu849

Callea, L., Bonati, L., & Motta, S. (2021). Metadynamics-Based Approaches for Modeling the Hypoxia-Inducible Factor 2α Ligand Binding Process. *Journal of Chemical Theory and Computation*, 17(7), 3841–3851. https://doi.org/10.1021/acs.jctc.1c00114

Callea, L., Caprai, C., Bonati, L., Giorgino, T., & Motta, S. (2024). Self-organizing maps of unbiased ligand–target binding pathways and kinetics. *The Journal of Chemical Physics*, 161(13). https://doi.org/10.1063/5.0225183

Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., & Shaw, D. E. (2012). Biomolecular Simulation: A Computa-tional Microscope for Molecular Biology. *Annual Review of Biophysics*, 41(1), 429–452. https://doi.org/10.1146/annurev-biophys-042910-155245

Fraccalvieri, D., Pandini, A., Stella, F., & Bonati, L. (2011). Conformational and functional analysis of molecu-lar dynamics trajectories by self-organising maps. BMC *Bioinformatics*, 12(1), 158. https://doi.org/10.1186/1471-2105-12-158

Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCam-mon, J. A., & Caves, L. S. D. (2006). Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics*, 22(21), 2695–2696. https://doi.org/10.1093/bioinformatics/btl461

Hendrix, E., Motta, S., Gahl, R. F., & He, Y. (2022). Insight into the Initial Stages of the Folding Process in On-conase Revealed by UNRES. *The Journal of Physical Chemistry B*, 126(40), 7934–7942. https://doi.org/10.1021/acs.jpcb.2c04770

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron*, 99(6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

Kohonen, T. (1990). The Self-organizing Map. *Proceedings of the IEEE*, 78, 1464–1480.

Kohonen, T. (2013). Essentials of the self-organizing map. Neural Networks, 37, 52–65. https://doi.org/10.1016/j.neunet.2012.09.018

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2022). cluster: Cluster Analysis Basics and Extensions.

Molgedey, L., & Schuster, H. G. (1994). Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72(23), 3634–3637. https://doi.org/10.1103/PhysRevLett.72.3634

Motta, S., Callea, L., Bonati, L., & Pandini, A. (2022). PathDetect-SOM: A Neural Network Approach for the Identification of Pathways in Ligand Binding Simulations. *Journal of Chemical Theory and Computation*, 18(3), 1957–1968. https://doi.org/10.1021/ACS.JCTC.1C01163

Motta, S., Pandini, A., Fornili, A., & Bonati, L. (2021). Re-construction of ARNT PAS-B Unfolding Pathways by Steered Molecular Dynamics and Artificial Neural Networks. *Journal of Chemical Theory and Computation*, 17(4), 2080–2089. https://doi.org/10.1021/acs.jctc.0c01308

Motta, S., Siani, P., Donadoni, E., Frigerio, G., Bonati, L., & Di Valentin, C. (2023). Metadynamics simulations for the investigation of drug loading on functional-ized inorganic nanoparticles. *Nanoscale*, 15(17), 7909–7919. https://doi.org/10.1039/D3NR00397C

Noé, F., & Nüske, F. (2013). A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. *Multiscale Modeling & Simulation*, 11(2), 635–655. https://doi.org/10.1137/110858616

Rubina, & Moin, S. T. (2023). Attempting Well-Tempered Funnel Metadynamics Simulations for the Evalua-tion of the Binding Kinetics of Methionine Ami-nopeptidase-II Inhibitors. *Journal of Chemical In-formation and Modeling*, 63(24), 7729–7743. https://doi.org/10.1021/acs.jcim.3c01130

Shao, J., Tanner, S. W., Thompson, N., & Cheatham, T. E. (2007). Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clus-tering Algorithms. *Journal of Chemical Theory and Computation*, 3(6), 2312–2334. https://doi.org/10.1021/ct700119m

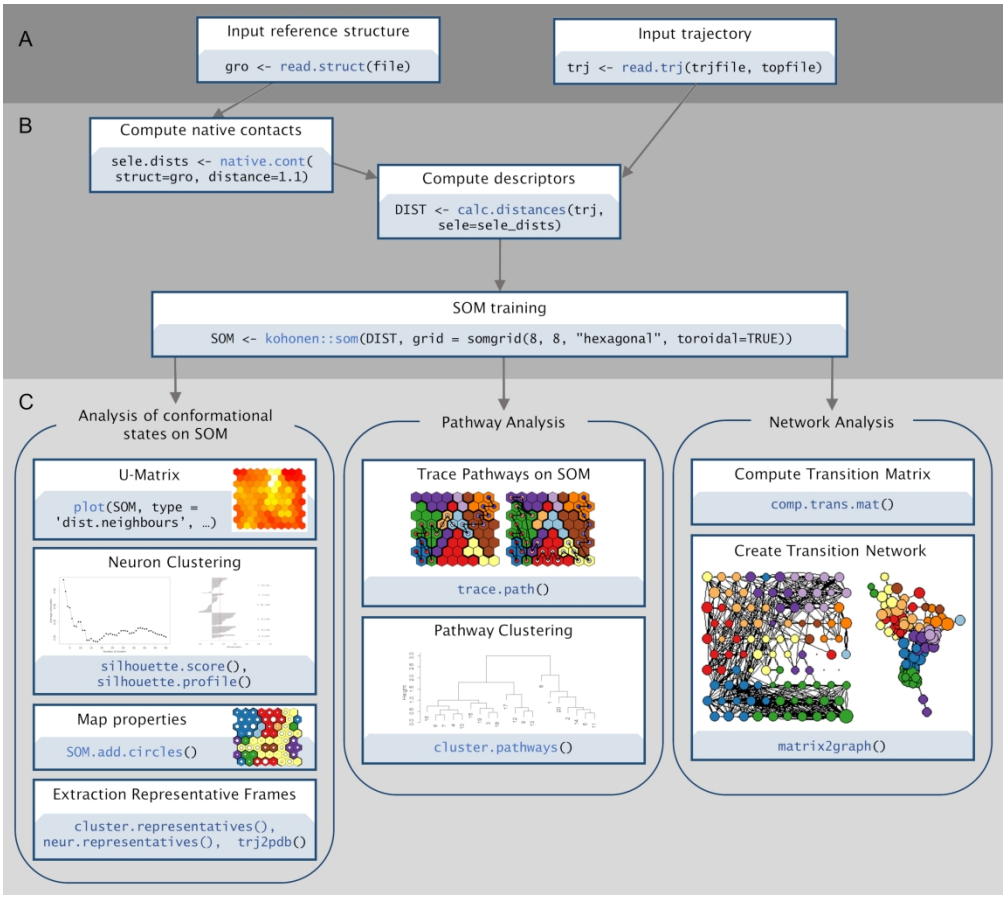
Tripathi, S., & Nair, N. N. (2023). Temperature Accelerated Sliced Sampling to Probe Ligand Dissociation from Protein. *Journal of Chemical Information and Mod-eling*, 63(16), 5182–5191. https://doi.org/10.1021/acs.jcim.3c00376

Wehrens, R., & Buydens L.M.C. (2007). Self- and Super-organizing Maps in R: The kohonen Package. *Journal of Statistical Software*, 21(5), 1-19. https://doi.org/10.18637/jss.v021.i05

Wehrens, R., & Kruisselbrink, J. (2018). Flexible Self-Organizing Maps in kohonen 3.0. *Journal of Statistical Software*, 87, 1–18. https://doi.org/10.18637/jss.v087.i07

Yuan, L., Liang, X., & He, L. (2024a). Insights into the Dis-sociation Process and Binding Pattern of the BRCT7/8-PHF8 Complex. *ACS Omega*, 9(19), 20819–20831. https://doi.org/10.1021/acsomega.3c09433

Yuan, L., Liang, X., & He, L. (2024b). Unveiling dissocia-tion mechanisms and binding patterns in the UHRF1-DPPA3 complex via multi-replica molecular dynamics simulations. *Journal of Molecular Modeling*, 30(6), 173. https://doi.org/10.1007/s00894-024-05946-9



Summary diagram of the SOMMD package architecture and analysis workflow. The package provides functions to read the structure topology and trajectory files (A. in-put), to compute the desired conformational descriptors and to train the SOM (B. preprocessing and training). From a trained SOM analysis and visualization can be performed on three levels: conformational states, pathways and transition networks (C. analysis and visualization).

162x144mm (600 x 600 DPI)