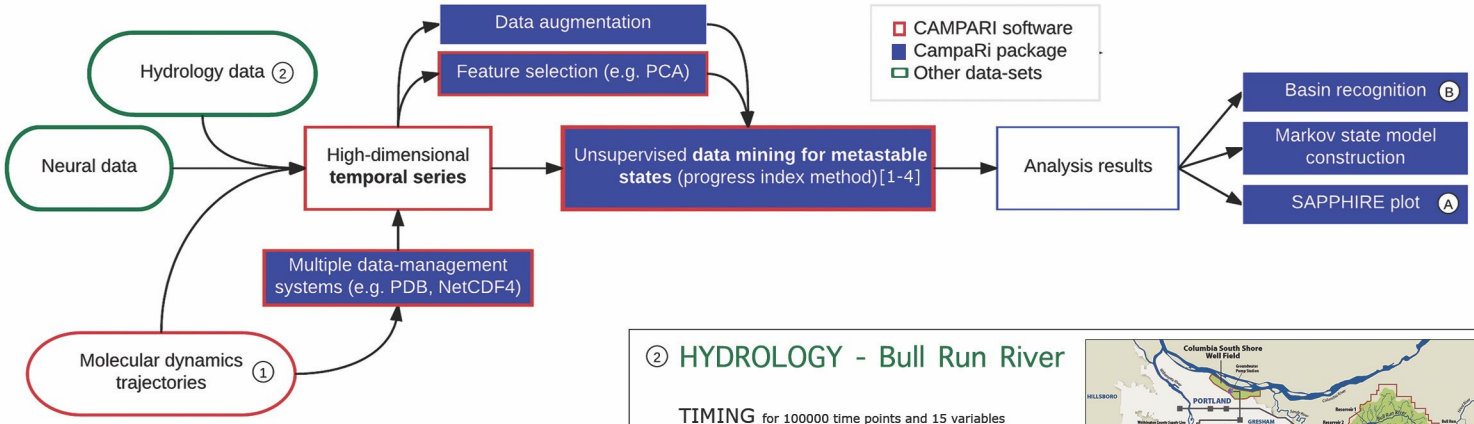
CampaRi: An R Package for Extracting Metastable States from Time Series Data

Davide Garolini, *Amedeo Caflisch, and Andreas Vitalis*
Department of Biochemistry, University of Zurich, Switzerland
davide.garolini@uzh.ch
University of Zurich[UZH]

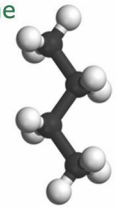
INTRODUCTION

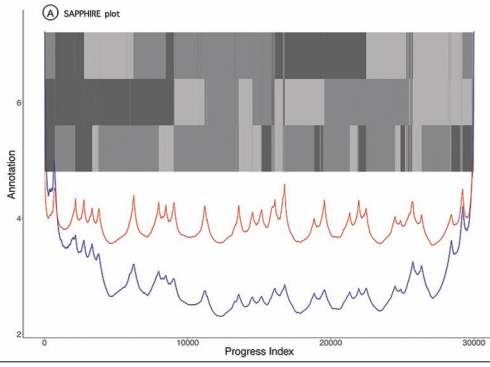
We present here an R package that is able to act as a frontend to the HPC Fortran implementation of CAMPARI (software for Molecular Dynamic simulation and time series analysis [1-4]), and expands its functionality by a set of user-friendly tools. These tools aid common analysis tasks by saving human and computing time. Moreover, the package reimplents some of CAMPARI's core analysis algorithms for testing and developing.



(1) MOLECULAR SIMULATION - *n*-butane

Stochastic dynamic simulation of the linear alkane *n*-butane described by three dihedral angles specifying rotations around all three carbon–carbon bonds.
The degeneracy of metastable states is fully observed in 27 basins, reflecting 3 distinct potential energy minima (*gauche*$^+$, *gauche*$^-$, *anti* [annotated in grey]) per angle and 3 dihedral angles.

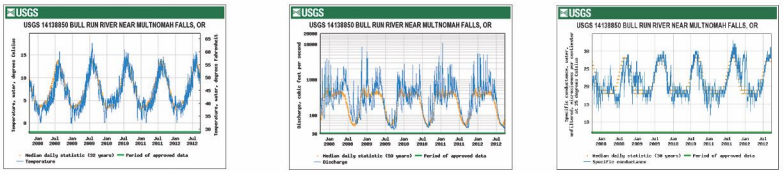


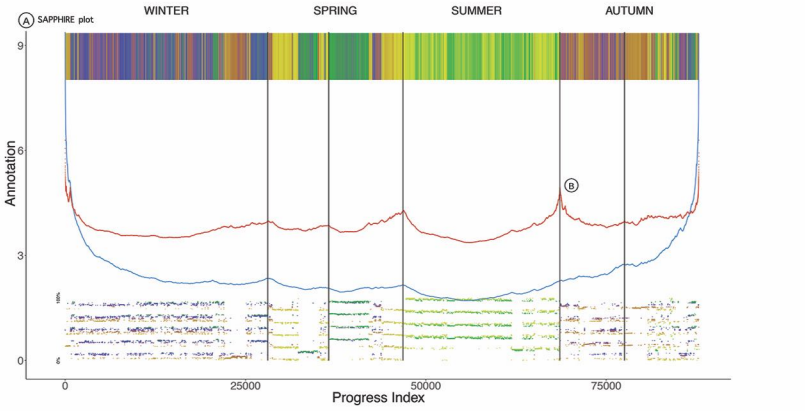


(2) HYDROLOGY - Bull Run River

TIMING for 100000 time points and 15 variables

CPU time per core: 120 s
HUMAN: 5 min (3 lines of code)





BACKEND - openMP, MPI Fortran

CAMPARIv3 is a versatile software package for **performing computer simulations of biomolecules and for analyzing large data sets**, whether they are obtained from such simulations or not. Nearly all of its features are OpenMP-parallelized to be able to exploit modern multicore CPUs. In addition, there is an outer MPI parallelization layer utilized heavily for mutli-copy simulation methods. CAMPARI complements comparable software (like GROMACS, CHARMM, etc) by a **unique combination of many simulation features** (e.g., the ABSINTH implicit solvation model, Monte Carlo and dynamics samplers in rigid-body/torsional space, advanced sampling methods such as Hamiltonian replica exchange, Wang-Landau, and progress index-guided sampling) while also supporting standard molecular dynamics simulation paradigms. CAMPARI comes with an **embedded suite of analysis tools**, including implementations of well-known biomolecule- and polymer-centric analysis (e.g., Guinier plots, NMR coupling constants, or DSSP annotation). A very powerful **unsupervised data mining utility** is included, which implements **clustering algorithms[3]**, the **progress index method[1-4]**, and a **collection of Markov state model routines** (similar to what is found ins pyEMMA or comparable tools).

FRONTEND - R package

- **run_campari(...)** is able to call directly Fortran Campari with all its installed functionalities
- **Minimum Spanning Tree** from trajectory (MST) and progress index analysis [1-4]
- **Short Spanning Tree** from trajectory (using tree-based clustering) and progress index analysis [1-4]
- **Feature selection** (e.g. PCA, tICA)
- **Data augmentation** (e.g. multiplication)
- Extensive plotting (**SAPPHIRE**[1-2], basin recognition)
- **Basin recognition** and Markov state model inference (with ramping procedure)

References

[1] Blöchliger, N., Vitalis, A. & Caflisch, A. A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems. *Comput. Phys. Commun.* 184, 2446–2453 (2013).
[2] Blöchliger, N., Vitalis, A. & Caflisch, A. High-Resolution Visualisation of the States and Pathways Sampled in Molecular Dynamics Simulations. *Sci. Rep.* 4, 6264 (2014).
[3] Vitalis, A. & Caflisch, A. Efficient Construction of Mesostate Networks from Molecular Dynamics Trajectories. *J. Chem. Theory Comput.* 8, 1108–1120 (2012).
[4] Blöchliger, N., Caflisch, A. & Vitalis, A. Weighted Distance Functions Improve Analysis of High-Dimensional Data: Application to Molecular Dynamics Simulations. *J. Chem. Theory Comput.* 11, 5481–5492 (2015).