ECAM State of the Art Workshop 1: Reaction Coordinates from Molecular Trajectories
Date August 29 - September 2, 2016, Lorentz Center, Leiden, The Netherlands

Organisers: Peter Bolhuis (University of Amsterdam, The Netherlands), Christoph Dellago (University of Vienna, Austria) Gerhard Hummer (Max Planck Institute of Biophysics, Germany)

1. Overview

Many processes in nature and technology are dominated by rare transitions between long-lived stable states separated by high free energy barriers. Examples include phase transitions in materials, chemical reactions and conformational changes of biological macromolecules. The state-of-the-art workshop recently held at the Lorentz Center in Leiden discussed current methods and future perspectives on statistical mechanical approaches to perform computer simulation of such rare event processes in complex many-particle systems and extract information from these to gain mechanistic insight. In particular, the workshop focused on the identification of reaction coordinates and the construction of reliable and meaningful models from atomistic simulation data. The computational methodologies discussed by the participants are of relevance in a wide range of fields ranging from physics and chemistry to materials science and molecular biology.

Advanced molecular simulation techniques running on today's high performance computers can sample transitions between long-lived states effectively, resulting in large sets of simulation trajectories, from which one has to extract useful models that capture the essential features of the process under study. These models should have a minimal number of degrees of freedom and should be expressed in terms of physically meaningful variables to render them understandable to humans. Moreover, such models serve as the basis for the construction of reaction coordinates that enable in-depth studies of the process at hand, e.g. by computing thermodynamic and kinetic properties. Many approaches have been developed to find good low-dimensional reaction coordinates, for instance, by seeking functions of physical variables that best approximate the committor, by identifying projections that separate transition paths and the transition state ensemble from equilibrium fluctuations, by extracting the dominant pathways in master equation and Markov-state models (MSM) or by using data-driven methods such as diffusion and sketch maps. However, these methods require the a priori definition of some collective variables that are capable of describing the process under study and thus rely on the physical and chemical intuition of the researcher apply these methods. Central goals of the workshop were to explore routes to improving algorithms, to minimize the required human intervention, and to develop practical analysis tools accessible to a wide scientific community.

2. Program of the workshop

Day 1, August, 29th
Creating Collective Variables and Reaction coordinates
09.00-09.30 Registration and coffee

09.30-09.45 Welcome and introduction by the LC staff and Chair of the day
09.45-10.20 Kresten Lindorff-Larsen (U. Copenhagen)
Reaction coordinates in molecular simulations – input or output?
10.30-11.05 Carlo Camilloni (TU Munich)
Collective variables, Experimental data and PLUMED
11.15-11.50 Pratyush Tiwary (Columbia U, New York)
Identifying and enhancing important fluctuations for sampling molecular systems with rare events
12.00-14.00 lunch & discussion
14.00-14.45 Discussion on mapping out of workshop goals (Discussion leader Peter Bolhuis)
14.45-15.20 Bettina Keller (FU Berlin)
Harnessing chemical intuition to find the slow dynamic subspace.
15.30-16.00 coffee
16.00-16.35 Alessandro Laio (SISSA, Trieste)
Automatic topography of complex and multidimensional probability distributions
16.45-17.30 Plenary discussion (Discussion leader: Gerhard Hummer)

Day 2, August 30, Methods to extract reaction coordinates from molecular simulations

09.00-09.35 Baron Peters (UCSB, Santa Barbara)
Rare Events Methods, Reaction Coordinates, and Useful Rate Theories
09.45-10.20 Ron Elber (U Texas, Austin)
Global and local approaches for calculations of reaction pathways.
10.30-11.15 coffee
11.15-11.40 Pietro Faccioli (Trento U, Trento)
Variational principles for reaction coordinates and biased reaction pathways
12.00-13.00 lunch & discussion
13.00-14.45 discuss & work & collaborate on goals
14.45-15.20 Titus Van Erp (NTNU, Trondheim)
Analyzing complex reaction mechanisms using path sampling
15.30-16.00 coffee
16.00-16.35 Michele Ceriotti (EPFL, Lausanne) Finding patterns and mapping landscapes.
16.45-17.30 Plenary discussion (Discussion Leader Alessandro Laio)
17.30 end of the day

Day3, 31 August, Fundamental issues in reaction coordinates

09.00-09.35 Cecilia Clementi (Rice U, Houston)
Low dimensional representation of protein conformational dynamics: tools and challenges
09.45-10.20 Robert Best (NIH, Bethesda)
Mapping protein folding dynamics onto 1D reaction coordinates — how far can we push this idea?
10.30-11.15 coffee
11.15-11.50 Sergei Krivov (U Leeds)
Nonparametric variational optimization of reaction coordinates.
12.00-13.00 Lunch & discussion

13.00-14.45 Discuss & work & collaborate on goals
14.45-15.20 Daan Crommelin (CWI, Amsterdam)
Importance functions for multilevel splitting
15.30-16.00 coffee
16.00-16.35 Peter Bolhuis (U Amsterdam)
Networks of reaction coordinates or reaction coordinates of networks
16.45-17.30 Plenary discussion (Discussion Leader Baron Peters)
17.30 end of the day
19.00 Workshop dinner

Day 4, September 1, Machine Learning and Markov Modeling
09.00-09.35 Max Welling (U Amsterdam)
Modern Machine Learning Tools Relevant for Molecular Dynamics
09.45-10.20 Frank Noe (FU Berlin)
Machine learning methods for dimension reduction, reaction coordinate identification and extracting kinetics from molecular dynamics
10.30-11.15 coffee
11.15-11.50 Edina Rosta (Kings College, London)
Identification and Analysis of Transition and Metastable Markov States
12.00-13.00 lunch & discussion
13.00-14.30 Discuss & work & collaborate on goals
14.30-15.05 Xuhui Huang (HKUST, Hong Kong)
Using the Projection Operator Approach to Identify Optimal Kinetic Lumping and Recover Slowest Conformational Dynamics of Complex Systems
15.15-15.45 coffee
15.45-15.55 David Swenson (U Amsterdam) (contributed talk)
Generation and Analysis of Arbitrary Path Ensembles using OpenPathSampling
16.00-16.35 Jocelyne Vreede (U Amsterdam)
Path sampling simulations of the mechanisms and rates of transitions between Watson-Crick and Hoogsteen base pairing in DNA
16.45-17.30 Plenary discussion (Discussion Leader Christoph Dellago)
17.30 end of the day

Day 5, Friday 2nd September, Applications and Experimental testing of reaction coordinates
9.00-9.35 Michael Woodside (U. Alberta, Edmonton)
Measuring transition paths in the folding of single molecules
9.45-10.20 Dmitrii Makarov (U Texas, Austin)
Reaction coordinates and pathways of mechanochemical transformations
10.30-11.15 coffee
11.15-11.50 Christoph Dellago (U Vienna)
Reaction coordinate for freezing: do we understand crystallization?
12.00-13.30 lunch & discussion & work

13.30-13.40 Ruben Demuynck (Genth U, Zwijnaarde) (contributed talk)
Advanced molecular dynamics simulations to construct free energy profiles of complex transformations in nanoporous materials.
13.45-13.55 Rodrigo Casasnovas (Forschungszentrum Julich) (contributed talk)
Understanding protein-ligand unbinding kinetics from metadynamics simulations
14.00-15.35 Gerhard Hummer (MPI Frankfurt)
Reaction coordinates in the analysis of single-molecule experiments
14.45-15.00 coffee
15.00-16.00 overview of the workshop + discussion plans for the future (Discussion leader Peter Bolhuis)
16.00 End of the workshop

3. List of participants

Last Name	First name	Affiliation
Best	Robert	National Institutes of Health, US
Bonella	Sara	CECAM@EPFL, CH
Camilloni	Carlo	Technische Universität München, DE
Casasnovas	Rodrigo	Forschungszentrum Jülich, DE
Ceriotti	Michele	Institute of Materials, EPFL, CH
Clementi	Cecilia	Rice University, US
Crommelin	Daan	CWI Amsterdam, NL
Dellago	Christoph	University of Vienna, AT
Demuynck	Ruben	Ghent University, BE
Elber	Ron	University of Texas at Austin, US
Everaers	Ralf	, ENS de Lyon, FR
Faccioli	Pietro	Trento University, IT
Huang	Xuhui	The Hong Kong University y, CN
Hummer	Gerhard	MPO of Biophysics, DE
Keller	Bettina	Freie Universität Berlin, DE
Krivov	Sergei	University of Leeds, UK
Lindorff-larsen	Kresten	University of Copenhagen, DK
Mackernan	Donal	University College Dublin, IR
Makarov	Dmitrii	University of Texas at Austin, US
Noe	Frank	FU Berlin, DE
Peters	Baron	UCSB, US
Rosta	Edina	King's College London, UK
Swenson	David	Universiteit van Amsterdam, NL
Tiwary	Pratyush	Columbia University, US
Van erp	Titus	NTNU, NO
Vreede	Jocelyne	University of Amsterdam, NL
Welling	Max	U. of Amsterdam, NL
Woodside	Michael	University of Alberta, CA

3. Major outcomes

The workshop discussed methods and techniques to sampling and analyzing rare events in complex systems, such as nucleation at phase transitions, drug binding, protein-protein interactions, protein folding, association, and self-assembly to name but a few. In talks and discussions the workshop participants reviewed current state of the art methods to address such process including path sampling (including Milestoning), metadynamics, Markov state modeling, diffusion maps, dimension reduction, reaction coordinate optimization, machine learning, and unsupervised cluster methods, and explored ways to improve these methods. Particular attention was devoted to the integration of popular MD packages such as Gromacs, NAMD Charmm, Amber, ACEMD, MOIL, LAMMPS with enhanced analysis and advanced sampling tools including Plumed (a package for enhanced sampling and collective variable analysis), pyEmma, and MSMBuilder (packages for Markov sate model analysis).

Notwithstanding the great capabilities of existing methods and software, several unsolved issues were identified. In particular, during the workshop the following topics were discussed:

1. *Extracting order parameters from molecular simulations to construct low dimensional models.* This point is important because there is no straightforward recipe to reduce the dimensions to meaningful variables and progress in this area is urgently needed. Speakers contributing here were Lindorff-Larsen, Camilloni, Tiwary, Keller, Laio, Ceriotti, Rosta, and Huang.
2. *Methods for sampling rare pathways*. Here the goal is to create the molecular trajectory data using advanced sampling algorithms. Speakers contributing here were Peters, Elber, van Erp, Faccioli. Vreede, and Swenson.
3. *Methods to construct reaction coordinates*. This subject was discussed by Peters, Best, Clementi, Krivov, Crommelin and Bolhuis.
4. *Machine learning algorithms*. The prospect of automatic methods to construct RCs from molecular trajectories is very appealing. Various speakers discussed this approach, including Welling, Dellago, Ceriotti, Noé.
5. *Better ways to integrate simulations and experiments.* It is important to connect the proposed computational methods to experimental probes. This was discussed by Woodside, Makarov, Hummer, Vreede, Casasnovas, and Demuynck.

More specifically, during the discussion sessions the following questions were raised and partially answered:

1. What properties should a simulation method predict?
2. What is a reaction coordinate?
3. Is the committor really the perfect reaction coordinate?
4. How to obtain the best low dimension model for the committor?
5. How to connect high dimensional reaction coordinates to physically meaningful variables?
6. How can we use machine learning to find collective variables and reaction coordinates?

7. Is it always possible to devise an optimal reaction coordinate? If so, what is the meaning of this optimal coordinate?
8. When can reaction coordinates, which often constitute the slow variables of a process, be used to coarse-grain the dynamics? When not?
9. How do we get the order parameters for testing?
10. Can we make use of diffusion or sketch maps for reaction coordinate analysis?
11. What if multiple transitions are important? Do we resort to kinetic networks or use multiple reaction coordinates? Should one identify a single (possibly complicated) reaction coordinate, or try to construct a Markov state model (MSM) using many metastable states?
12. When is it possible to reduce a complex problem to diffusion on a one dimensional free energy landscape, and when do we need a network Markov model?
13. How can experiments test reaction coordinate predictions? How do we connect to experiments?
14. What are the predictions that one can make with an optimal reaction coordinate (as opposed to reasonable RC)?
15. To what degree does the computational efficiency of free energy calculations depend on the chosen coordinates, and how can these coordinates by optimized, possibly "on the fly"?
16. Experiments are almost always interpreted in terms of 1D models. What are the limits of this approach? When does it fail? How good is 'good enough'?

To review the current state of the field, identify stumbling blocks, and point out possible future research directions, the workshop participants discussed the possibility to collectively write a review paper co-authored by all of them.

4. Community needs

Several sessions were devoted to identifying what currently slows down progress in the field and to address the specific needs of the scientific community to overcome these problems. During the discussions it became clear that here is a need for methods and software tools to carry out the following tasks:

1. Sample rare event processes, either by a) configuration or b) trajectory sampling.
2. Construct kinetic networks or Markov state models from trajectory data.
3. Construct and identify order parameters from arbitrary configurations.
4. Dimensionality reduction of large data sets and reaction coordinate analysis.

Several software packages already exist, e.g. MSMBuilder and PyEmma for item 2) and Plumed for items 1a) and 3). Therefore, it makes sense to focus on items 1b) and 4). Indeed, that is the main goal of Work Package 1 of the ECAM Center of Excellence. Developments in this direction will be carried out mainly within The Open Path Sampling (OPS) framework, a Python library to facilitate path sampling simulations (http://openpathsampling.org).

Software modules to be developed in WP1 of ECAM partly based on OPS will address the follow specific issues:

1. Specific modules that perform transition path sampling, TIS, FFS
2. Modules to optimization interface for TIS and FFS
3. Modules for executing the reactive flux algorithm
4. Module for the calculation of the transition state ensemble
5. State definition assessor modules
6. Wrappers for popular MD engines (LAMMPS, Gromacs)
7. Interface to Plumed for use of collective variables
8. Reaction coordinate analysis modules for use with TPS/TIS/FFS
9. Analysis tools to work with path sampling

The development of these modules is the objective of the extended software development workshops organized within Work Package 1 of ECAM.

5. Funding

To make progress in the development of new methods and tools for the sampling and analysis of rare events, a concerted community effort is needed. To fund such a community-wide effort in a sustainable way, a combined strategy is necessary. While individual PIs can apply for single investigator grants from their national funding agencies, the EU H2020 framework could provide opportunities to set up an international network geared towards method development. More specifically, within the Marie Skłodowska-Curie actions, an Innovative Training Network focused on rare event sampling and analysis could be used to create a collective research and training effort with strong ties to industry. Other opportunities for funding might be available in the Future and Emerging Technologies (FET) Program of H2020.

Developments for society and industry

The societal benefits are twofold. On the fundamental level, the tools will enable new discoveries and developments, for instance in material science and molecular medicine. This will be made possible by the scientific community using the sampling and analysis tools that re being developed.

On an economic level, industry would benefit from software containing efficient and easy to use simulation and analysis tools to extract observables for applications in

- Drug binding: binding affinities, kon/koff rates, effect of mutations on these quantities (this implies mechanistic insight).
- Food/dairy industry: protein aggregation, chocolate crystallization, grain size in ice cream, food preservation, all of which requires knowledge of the freezing mechanism
- Materials science: nucleation, crystallization of polymers, ageing of materials, soft matter, self-assembly of nanomaterais, colloids, defect formation in crystals.

Filially, society will benefit from methods and ultimately software that make drug development cheaper, improve food quality, and help in the development of high-tech materials.