Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods

Frank Noé[1] and Cecilia Clementi[2]

Collective variables are an important concept to study high-dimensional dynamical systems, such as molecular dynamics of macromolecules, liquids, or polymers, in particular to define relevant metastable states and state-transition or phase-transition. Over the past decade, a rigorous mathematical theory has been formulated to define optimal collective variables to characterize slow dynamical processes. Here we review recent developments, including a variational principle to find optimal approximations to slow collective variables from simulation data, and algorithms such as the time-lagged independent component analysis. Using these concepts, a distance metric can be defined that quantifies how slowly molecular conformations interconvert. Extensions and open questions are discussed.

Addresses
[1] Department of Mathematics and Computer Science, FU Berlin, Arnimallee 6, 14195 Berlin, Germany
[2] Center for Theoretical Biological Physics, and Department of Chemistry, Rice University, 6100 Main Street, Houston, TX 77005, United States

Corresponding authors: Noé, Frank (frank.noe@fu-berlin.de) and Clementi, Cecilia (cecilia@rice.edu)

Current Opinion in Structural Biology 2017, **43**:141–147

This review comes from a themed issue on **Theory and simulation**

Edited by **Carol Post** and **Ronald Levy**

For a complete overview see the Issue and the Editorial

Available online 17th March 2017

http://dx.doi.org/10.1016/j.sbi.2017.02.006

0959-440X/© 2017 Elsevier Ltd. All rights reserved.

Introduction

It is increasingly easy to generate extensive simulation data of complex molecular systems, such as molecular dynamics (MD) of biomolecules and fluids [1–5], or *ab initio* MD of solids and chemical compounds [6,7].

The number of microscopic degrees of freedom in a molecular simulation is usually large, for example, thousands or more. In order to generate understanding and to be able to control and enhance the simulation, it is very useful to find a few collective variables (CVs) of the system. CVs are descriptors that can identify interesting collective phenomena over long timescales, and separate macroscopically different structures or aggregation states.

A property of fundamental importance is whether a set of CVs is able to differentiate between metastable states of a system and can thus resolve the rare event transitions, or the slow dynamics/kinetics between them. Examples of rare events include freezing and melting in liquids close to the critical point, and conformational changes, folding or association of macromolecules. In the following we will use the term *slow CVs* to refer to CVs that can resolve slow processes. Slow CVs are not only important to characterize rare event kinetics but also to extract the essential information from the simulation for its interpretation (in particular in comparison with experimental results): reducing the description of a high dimensional system to a few coordinates implies that it is possible to somehow renormalize over the remaining coordinates, that is, these remaining coordinates either average on timescales faster than those of the resolved CVs, or have no effect on the process of interest.

Slow CVs are also essential to enhance the sampling of rare events in simulation, either by biasing the system's energy function to increase the probability of visiting scarcely populated states (as in Metadynamics [8], Umbrella Sampling [9], Adiabatic Free Energy Calculations [10,11]), by iteratively redistributing large ensembles of unbiased simulations to increase the probability of observing rare transitions [12,13•], or by path sampling along progress variables [14,15].

The concept of slow CVs naturally leads to the concept of eigenfunctions of the dynamical operator underlying MD simulation [16•,17–19], and we will focus on this notion here. We will mostly avoid the term 'reaction coordinate' which is frequently used in the context of CVs [20], but it is subject to different interpretations, as it may either refer to slow CVs or CVs that resolve the progress along a transition between two pre-defined states, A and B. The latter definition leads to the concept of 'committor probability', that is, the probability of proceeding towards the target state B rather than relapsing to the source state, A [21–24]. Although the two interpretations are related, methods specific for the A/B-reaction coordinates and committor probabilities will not be reviewed here.

Historically, physical intuition has played a key role in identifying good CVs, for example, based on physical interpretability [25–28]. While intuition will always be an essential tool for a scientist, the size and dimensionality of the data that are now generated calls for automatable ways of extracting good CVs from simulation data. In particular

we are interested in finding CVs in a principled and systematic way that avoids too much ambiguity in the interpretation of simulation results and allows a direct connection to experiment [29]. For example, Anton supercomputer simulations (e.g. [1]) have been widely used as benchmark systems for analysis of molecular trajectories, and led to hundreds of publications analyzing the conformations and kinetics contained in them, with often different and sometimes seemingly contradictory results [30–34].

A rigorous theory has been formulated that shows that the optimal slow CVs are given by the eigenfunctions of the molecular dynamics operator. While usually these cannot be exactly computed a priori (simply from the knowledge of the Hamiltonian of the system), in practice they can be estimated from MD simulation data. We review a recently described variational principle which allows to search for optimal approximations of these eigenfunctions, and the relation to other analysis methods. Ongoing research focuses on the design of strategies to solve the variational problem algorithmically and how to gain physical intuition from the interpretation of the optimal CVs in terms of physical observables. For instance, in the case of a protein complex, one would like to understand the dynamics in terms of groups of atoms, angles, interatomic distances, or formation of specific contacts.

Molecular kinetics

Consider an ensemble of molecules, each characterized by the microscopic state $\mathbf{r}$ (e.g. positions and momenta). In equilibrium the probability of finding a molecule in state $\mathbf{r}$ is given by the Boltzmann distribution $\pi(\mathbf{r})$, and we will measure any off-equilibrium probability distribution *relative* to π, that is, if a perturbed ensemble is distributed as $p(\mathbf{r})$, we say its relative distribution is $u(\mathbf{r}) = p(\mathbf{r})/\pi(\mathbf{r})$. Next consider the dynamics: each molecule changes its state over time, with a certain probability $p_\tau(\mathbf{r}_1, \mathbf{r}_2)$ to transition from its current state $\mathbf{r}_t$ at time t to another state $\mathbf{r}_{t+\tau}$ after a lag time τ. If each molecule evolves according to this transition probability, we can define a transfer operator $\mathcal{T}_\tau$ to express the evolution of the relative distributions u with time: $u_{t+\tau} = \mathcal{T}_\tau u_t$ (see [35] for a more complete discussion). Since the relative distribution associated to $\pi(\mathbf{r})$ itself is the constant 1-function $\pi(\mathbf{r})/\pi(\mathbf{r}) = 1(\mathbf{r})$, it is unchanged by $\mathcal{T}_\tau$:

$$\mathcal{T}_\tau 1 = 1. \tag{1}$$

This is an eigenvalue equation with eigenvalue 1. However, the eigenvalue equation for the transfer operator $\mathcal{T}_\tau$ has also other solutions, analogously to the time-independent Schrödinger equation in quantum mechanics:

$$\mathcal{T}_\tau \psi_i = \psi_i \lambda_i(\tau) \tag{2}$$

where the eigenfunctions $\psi_i(\mathbf{r})$ describe the relaxation back to equilibrium of perturbations of the equilibrium distribution. The corresponding eigenvalues decay exponentially with increasing lag time τ: $\lambda_i(\tau) = e^{-\tau\kappa_i}$ with rates κ_i or timescales $t_i = 1/\kappa_i$ (see Figure 1d, [35,36]).

While the first eigenfunction $\psi_1 = 1$ is trivial (Eqn [1]), eigenfunctions ψ_i with $i > 1$ are ideal slow CVs in the sense that they identify the ith-slowest transitions in the system (occurring with corresponding rate κ_i) and order the system configurations along the direction of these transitions (Figure 1a–c), thus characterizing the metastable states and transition states along these processes [17,35,37].

Estimating Slow CVs from simulation

How do we compute the eigenfunctions in practice? Eqn 2 can generally not be solved analytically, however ψ_i can be estimated from MD simulation data if at least some rare events have been sampled. Methods to approximate slow CVs from MD simulation data have been developed over the last several years, for example Markov State Models (MSMs) [17,35,38,39•], or Diffusion Maps [18,19]. These methods explicitly or implicitly approximate the first few eigenfunctions of (Eqn 2), but differ in the approximations and assumptions made. For example, Markov State Models rely on a discretization of the configurational space, while Diffusion Maps use a diffusion model of the data. Interestingly, all these methods can be subsumed by a recently developed variational principle and a variational algorithm derived thereof. Furthermore, these methods rely on the definition of a distance metric that can separate slowly mixing configurations from rapidly interconverting ones, and an optimal metric can also be derived from variational considerations.

Variational principle

Recently, a general principle has been proposed that can be used to systematically approximate ψ_i from data for molecular systems in thermal equilibrium. In particular, if the dynamics are statistically reversible, that is, if the equilibrium probability to go from any state $\mathbf{r}_1$ to any other state $\mathbf{r}_2$ is equal to its reverse, the following variational principle can be proved [40•,41••]: The m eigenfunctions $\psi_1, \ldots, \psi_m$ of (2) with smallest eigenvalues $\kappa_i, \ldots, \kappa_m$ (i.e. longest relaxation timescales, $t_i = 1/\kappa_i$) are the functions $f_1(\mathbf{r}), \ldots, f_m(\mathbf{r})$ that maximize the so called Rayleigh trace (or generalized Rayleigh quotient) R_m:

$$R_m = \sum_{i=1}^{m} \mathbb{E}[f_i(\mathbf{r}_t) f_i(\mathbf{r}_{t+\tau})], \tag{3}$$

such that $\mathbb{E}[f_i(\mathbf{r}_t) f_j(\mathbf{r}_{t+\tau})] = \delta_{ij}$,

where $\mathbb{E}[\cdot]$ is the equilibrium expectation value and δ_{ij} is the Kronecker delta (note that different, equivalent formulations of this principle exist [42]). In words: the first m eigenfunctions are statistically uncorrelated and have maximum autocorrelation at a lag time τ. It is

Figure 1

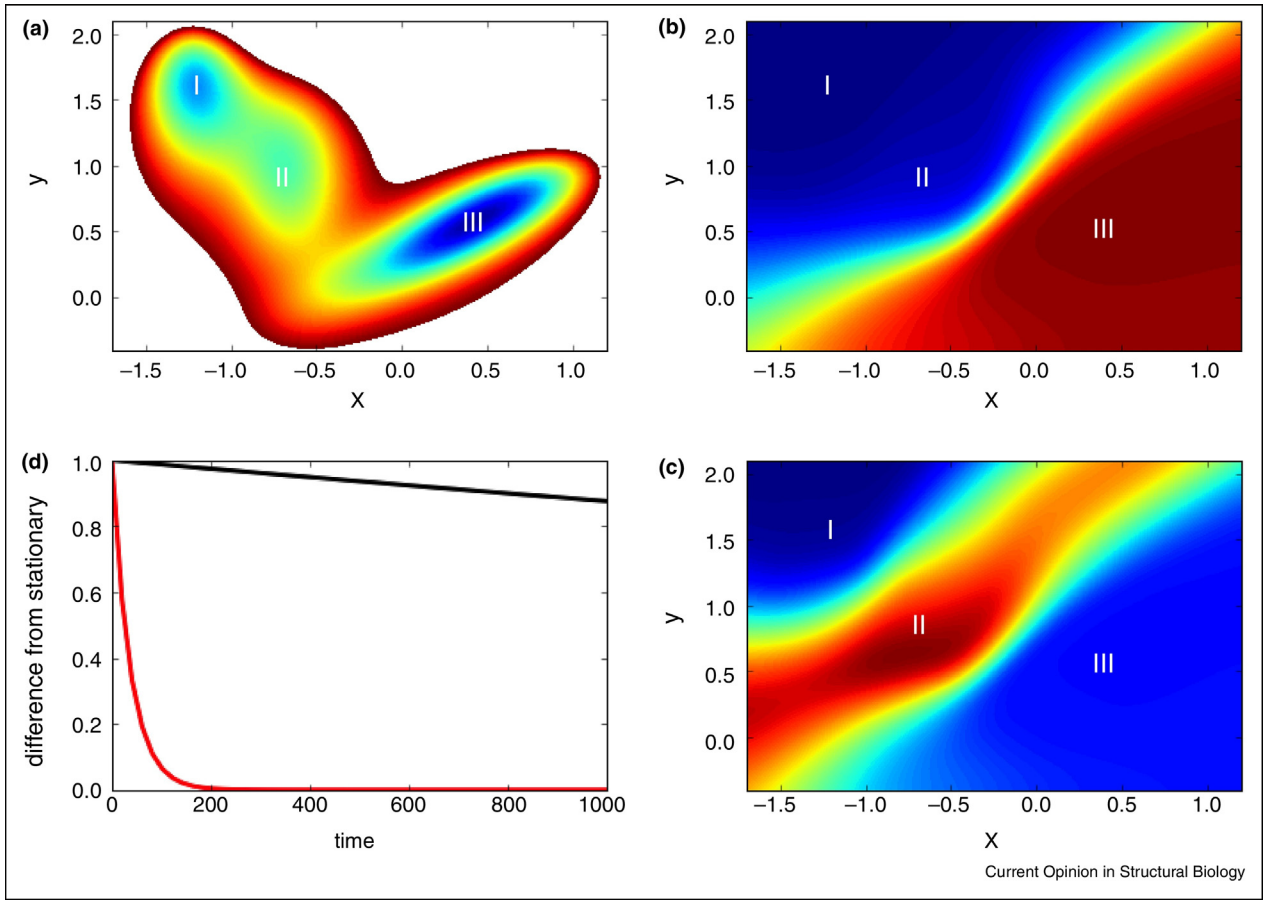


Illustration of eigenfunctions of a two-dimensional diffusion process. **(a)** Müller–Brown potential. **(b,c)** Slowest and second-slowest eigenvectors. **(d)** Relaxation from a perturbation along the slowest (black) and second-slowest (red) eigenvector.

straightforward to show that their autocorrelation function is the corresponding eigenvalue [40•]: $max_{f_i \ldots f_m} R_m = \sum_{i=1}^{m} \lambda_i$. Any other set of functions will have a smaller value of R_m, and this result can be used in practice as a guiding principle to design functions to optimally approximate eigenvalues and eigenfunctions. This principle is closely reminiscent of the popular variational method in quantum mechanics, and their derivations are analogous [41••].

Variational approach (linear version)

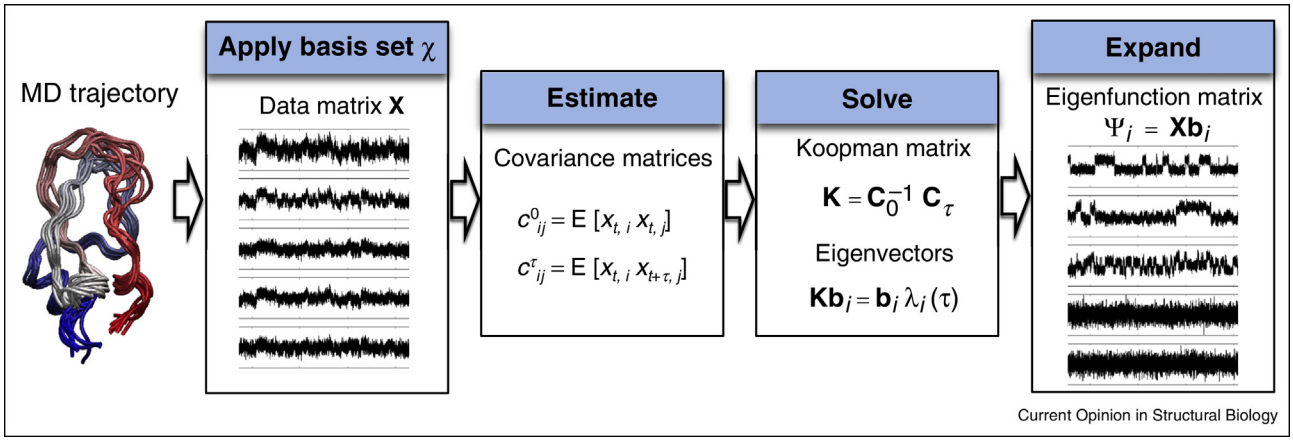
How do we practically search for CVs using the principle above? A simple approach is to define a basis set of functions $\chi_j(\mathbf{r})$ and express the eigenfunctions as a linear expansion:

$$\psi_i = \sum_j b_{ij} \chi_j(\mathbf{r}). \tag{4}$$

The basis functions may represent molecular order parameters such as distances between atoms or dihedral angles. More sophisticated choices include Gaussians in Ramachandran space to indicate the rotamer state of a residue in a protein [41••], MSM approximations of the eigenfunctions from individual amino acid simulations [43•], diffusion coordinates [30], or products between one-dimensional basis functions [44]. As detailed in [40•,41••], inserting the linear expansion (4) into (3), and maximizing the resulting expression yields the algorithm to find optimal expansion coefficients b_{ij} described in Figure 2.

The optimal coefficients b_{ij} resulting from the variational approach provide insight into the interpretation of the eigenfunctions. If molecular order parameters are used in the definition of the basis functions, the values of b_{ij} indicate the relative contributions of the corresponding different parameters to the slow processes. For instance, if one uses a set of inter-residue distances as basis functions to study a macromolecular configurational change, the distances that best describe the slowest dynamical process in the system will have the largest absolute values of the coefficients $\mathbf{b}_2$ in Ψ_2, etc. For each eigenfunction, the

Figure 2



Variational approach: in the method of linear variation the eigenfunctions are represented by linear expansion of basis functions [(4). The optimal approximation to the eigenfunctions evaluated on the MD trajectories is obtained as follows: **(1)** evaluate each basis function on all configurations sampled in a trajectory, $\mathbf{r}_t$. That is, define the data matrix, $\mathbf{X}$, as $x_{t,i} = \chi_i(\mathbf{r}_t)$, **(2)** estimate the covariance, $\mathbf{C}_0$, and time-lagged covariance matrix, $\mathbf{C}_\tau$, **(3)** compute the Koopman matrix $\mathbf{K} = \mathbf{C}_0^{-1}\mathbf{C}_\tau$ and its eigenvalues/eigenvectors, **(4)** project the data matrix onto the eigenfunctions. Note that TICA and MSMs are special cases of this algorithm.

order parameters that have a maximum correlation are the best physical descriptors for the corresponding dynamical process [16•]. Physical interpretation can be further facilitated by enforcing sparsity in the solution b_{ij} [45].

A number of algorithms that have been proposed to characterize slow molecular processes are special cases of the variational approach above. For example, if a state space partitioning (e.g. using clustering and Voronoi partition) is performed and cluster indicator functions (e.g. the ith function equals 1 on the ith state and 0 elsewhere) are used as basis set, then the method is equivalent to an MSM analysis [36,40•,41••]. In a similar way, milestoning or core-set MSMs [38,46] approximate the eigenfunctions ψ_i via a basis set expansion of committor functions between cores [46]. In the fuzzy MSMs introduced in [47], scaled Gaussians are used as basis set functions $\chi_i(\mathbf{r})$.

Although the algorithm presented above finds linear solutions (4), the basis functions themselves can be nonlinear functions of the original coordinates $\mathbf{r}$, thus the variational approach can in principle represent arbitrarily complex and nonlinear CVs. However, if the solution is nonlinear in the parameters (e.g. widths of Gaussian distributions), nonlinear optimization algorithms need to be employed [40•]. Alternatively, Tensor-based algorithms can approximate eigenfunctions by sums of higher-order products of simple (e.g. 1-dimensional) basis functions [44,48].

An important aspect in the implementation of the variational approach is that the covariance matrices $\mathbf{C}^0$ and $\mathbf{C}^\tau$ estimated from finite data (see Figure 2) should approximate the true covariance matrices of an equilibrium trajectory ensemble. Previous implementations have used direct (empirical) estimates, which converge to the true covariances in the limit of very long trajectories, but not in the limit of many short trajectories, unless their starting points were sampled from the equilibrium distribution. In [49] a reweighting method was introduced that achieves asymptotically unbiased estimates even for short trajectories started out of equilibrium, and we recommend to use this implementation.

Finally, the number of terms used in the basis set, and hyper-parameters (e.g. such as, for instance, the width if Gaussian ansatz functions are used), need to be regularized in order to avoid overfitting. Ref [50•] has proposed to use cross-validation with the Rayleigh trace R_m as a score. Cross-validation avoids overfitting by splitting the dataset into a training set, from which the matrices of the variational approach are estimated, and a test set, on which the score is computed [51].

Time-lagged independent component analysis (TICA)

The signal separation method proposed in [52], later named TICA [53], is a special case of the linear variational approach (Figure 2) using the mean-free input coordinates themselves as a basis set and empirical estimates of the covariance matrices [16•]. Although TICA has been designed as a signal decomposition algorithm, it provides an initial approximation to the optimal slow CVs.

For this reason, TICA has proven successful and become widely used to reduce the dimensionality of MD simulation data to a few slow CVs, in which MSMs and other

kinetic models can be built efficiently [16•,30,54–57]. A Kernel version of TICA [58] and multi-lagtime variations [59] have been proposed.

A general concern with TICA and the variational approach is computational efficiency for large systems and long trajectories. The computational effort scales with N^2 for N input coordinates, which becomes intractable if huge basis sets are used, such as the set of distances between all residues or atoms of a protein. The recently proposed hierarchical TICA method can obtain a coarse-grained yet accurate solution of the full TICA problem more efficiently [60], but there is still room for improvements in this area.

Kinetic maps and commute maps

Essentially all MD analysis methods require a distance metric in order to define which configurations are similar or dissimilar. Consider clustering algorithms as an example: completely different results and interpretations may be derived from the same MD data when structures are distinguished by measuring their root mean square deviation (RMSD) or instead the overlap of their contact maps. These ambiguities can be removed when the slow process eigenfunctions themselves are used to define a metric. Indeed, by measuring distances after transforming the molecular configurations into variationally approximated eigenfunctions, for example, using TICA, significant improvements over results obtained with a purely geometric metric have been reported for the construction of MSMs [2,16•,54,56,61], the computation of rates [29,62,63] and diffusion coordinates [30]. However, two questions need to be addressed when measuring distances in the space spanned by the first few eigenfunctions: (i) How many coordinates ψ_i should be used, and (ii) as the scaling of eigenfunctions is arbitrary, how should they be scaled in the projection step? Previous studies have found that the quality of the analysis critically depends on these choices [57,64•].

Recently, the concept of *kinetic distance* has been introduced in order to address these issues [64•,65••]. It builds upon the idea of 'diffusion distance' that was proposed for diffusion processes [66] as the measure of the overlap, after a lag time τ, between probability distributions initiated in two distinct states. Using this idea in the more general context of Markovian dynamics, leads to the concept of kinetic distance [64•,65]. A kinetic map can be simply obtained by scaling the normalized eigenfunctions Ψ_i by the corresponding eigenvalues:

$$\tilde{\Psi}_i^{(\tau)}(\mathbf{r}) = \lambda_i(\tau)\Psi_i(\mathbf{r}), \tag{5}$$

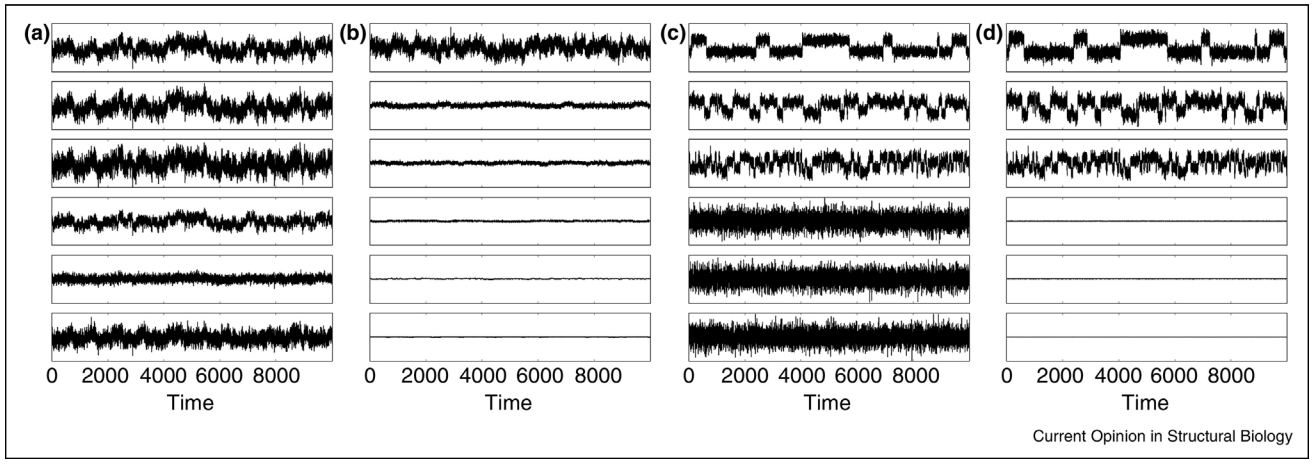
and the kinetic distance is obtained as Euclidean distance in this scaled space $\tilde{\Psi}_i^{(\tau)}$.

The dependence of the results on the choice of lag time τ was eliminated in a later development [65••] by integrating the definition of kinetic distance over τ. The resulting 'commute distance' is the Euclidean distance in the space obtained by scaling the eigenfunctions by the square root of the corresponding relaxation timescales $t_i = -\tau/\ln|\lambda_i(\tau)|$:

$$\Psi'_i(\mathbf{r}) = \sqrt{t_i/2}\Psi_i(\mathbf{r}). \tag{6}$$

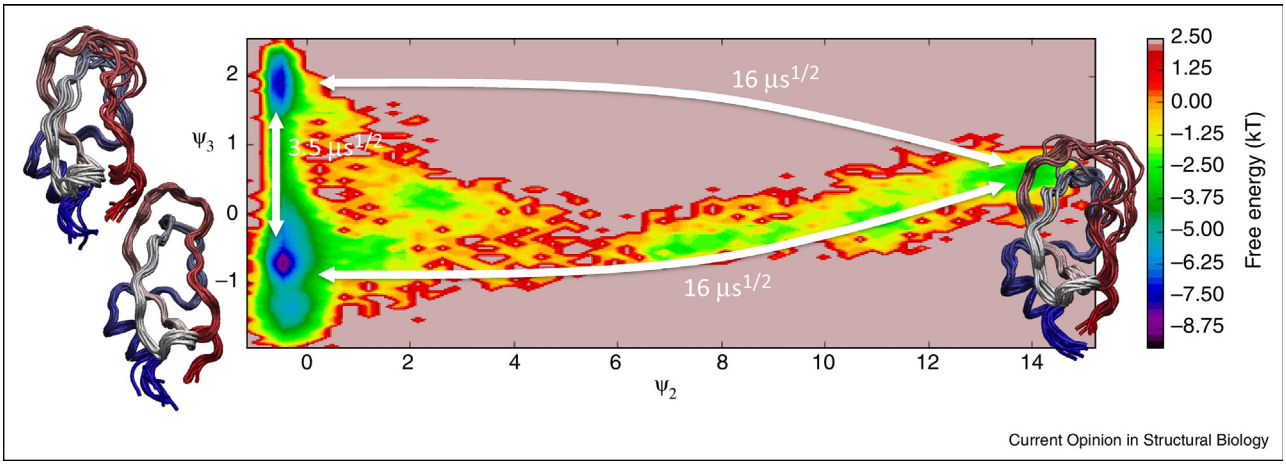
This distance has also a particularly interesting interpretation as it can be shown to approximate the commute time, that is, the mean time needed to go from one configuration to the other, and back [65••]. As a result, Ψ'_i can be interpreted as a *commute map*, in which

Figure 3



TICA illustration: **(a)** 6-dimensional signal which contains three slow dimensions with metastable jump processes (rare events) and three random noise processes. A 6 × 6 confusion matrix with random elements has produced a seemingly random time series with no obvious metastability. **(b)** Result of applying principal component analysis on (a). **(c)** Result of applying TICA on (a) — the three slow coordinates are recovered and appear first as they have the largest eigenvalues. **(d)** Kinetic map scaling — same as (c), but scaled by TICA eigenvalue.

Figure 4



Commute map illustration using the 1 ms BPTI trajectory from [67]. Free energy plotted on the commute map approximated by applying the variational approach to the basis set of C_α-distances. Distances in this plot correspond approximately to the square root of the half commute time (white numbers: square root of actual half commute times, computed from a Markov state model with 1000 *k*-means states).

Euclidean distances measure the commute times (Figure 4), thus directly quantify how slowly molecular conformations interconvert. In practice, Eqns 5 and 6 can be estimated by using a TICA approximation of Ψ_i.

The definition of commute distance answers both issues mentioned above: in addition to providing the scaling of the eigenfunctions, it also eliminates the choice on how many dimensions should be kept in the dimensionality reduction. In principle all dimensions can be kept, but faster processes only contribute with smaller amplitude. Figure 3 shows how the kinetic map suppresses fast processes, and [64•,65••] demonstrate the usefulness of kinetic maps in practice.

Conclusions

The systematic definition and identification of slow CVs is key to an insightful analysis of high-dimensional molecular systems. Thanks to the theoretical results of the last decade, it is now clear that the eigenfunctions of the dynamical operator underlying MD simulation define optimal slow CV. The variational principle and its algorithmic incarnations, such as the linear variational approach and TICA provide an optimal approach to find slow CVs from molecular simulation data. A key property of the optimal slow CVs is that they can be scaled to define a kinetic distance that separate fast and slowly mixing states. In such a metric space, transition times can be measured in terms of Euclidean distances, and geometric analyses, such as clustering, can be employed to reliably unravel rare event transitions. Future research directions include modeling problems, such as the design or data-based learning of nonlinear basis sets in the variational approach, as well as algorithmic improvements such as the efficient solution of the variational problem for large-scale simulation data. An important application of the CV identification methods described here is their use in enhanced and adaptive sampling methods, in order to efficiently sample rare events in complex molecular systems.

Acknowledgements

We gratefully acknowledge funding from European Research Council (ERC StG pcCell to F.N.), National Science Foundation (grants CHE-1152344 and CHE-1265929 to CC) and the Welch Foundation (grant C-1570).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE: **Science,** 2011, **334**:517.
2. Plattner N, Noé F: **Nat Commun,** 2015, **6**:7653.
3. Kohlhoff KJ, Shukla D, Lawrenz M, Bowman GR, Konerding DE, Belov D, Altman RB, Pande VS: **Nat Chem,** 2014, **6**:15.
4. Doerr S, Harvey MJ, Noé F, Fabritiis GD: **J Chem Theory Comput,** 2016, **12**:1845.
5. Shirts M, Pande VS: **Science,** 2000, **290**:1903.
6. Ufimtsev IS, Martinez TJ: **Comput Sci Eng,** 2008, **10**:26.
7. Marx D, Hutter J: Edited by Grotendorst J. **Modern Methods and Algorithms of Quantum Chemistry, NIC Series,** . John von Neumann Institute for Computing; 2000.
8. Laio A, Gervasio FL: **Rep Prog Phys,** 2008, **71**:126601.
9. Torrie GM, Valleau JP: **J Comput Phys,** 1977, **23**:187.
10. Rosso L, Minary P, Zhu Z, Tuckerman M: **J Chem Phys,** 2002, **116**:4389.
11. Maragliano L, Vanden-Eijnden E: **J Phys Lett,** 2006, **426**:168.
12. Zheng W, Rohrdanz MA, Clementi C: **J Phys Chem B,** 2013, **117**:12769.

13. Preto J, Clementi C: **Phys Chem Chem Phys,** 2014, **16**:19181.
• Collective variables obtained from eigenfunctions estimated during a molecular dynamics simulation are used for iteratively enhance the sampling.

14. Du W-N, Marino KA, Bolhuis PG: **J Chem Phys,** 2011, **135**:145102.

15. Faradjian AK, Elber R: **J Chem Phys,** 2004, **120**:10880.

16. Perez-Hernandez G, Paul F, Giorgino TG, Fabritiis D, Noé F: **J**
• **Chem Phys,** 2013, **139**:015102.
TICA is derived as a special case of the variational approach. This article shows that it is a linearly optimal method to approximate Markov operator eigenfunctions.

17. Schütte C, Fischer A, Huisinga W, Deuflhard P: **J Comput Phys,** 1999, **151**:146.

18. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker SW: **Proc Natl Acad Sci U S A,** 2005, **102**:7426.

19. Rohrdanz MA, Zheng W, Maggioni M, Clementi C: **J Chem Phys,** 2011, **134**:124116.

20. Rohrdanz MA, Zheng W, Clementi C: **Annu Rev Phys Chem,** 2013, **64**:295.

21. Banushkina PV, Krivov SV: **Wiley Interdiscip Rev Comput Mol Sci,** 2016, **6**:748.

22. Hummer G: **J Chem Phys,** 2004, **120**:516.

23. Vanden-Eijnden E: **Annu Rev Phys Chem,** 2010, **61**:391.

24. Peters B: **J Chem Phys,** 2006, **125**:241101.

25. Clementi C, Nymeyer H, Onuchic JN: **J Mol Biol,** 2000, **298**:937.

26. Bolhuis PG, Dellago C, Chandler D: **Proc Natl Acad Sci U S A,** 2000, **97**:5877.

27. Cho S, Levy Y, Wolynes PG: **Proc Natl Acad Sci U S A,** 2006, **103**:586.

28. Peters B, Trout BL: **J Chem Phys,** 2006, **125**:054108.

29. Zheng W, Rohrdanz MA, Maggioni M, Clementi C: **J Chem Phys,** 2011, **134**:144109.

30. Boninsegna L, Gobbo G, Noé F, Clementi C: **J Chem Theory Comput,** 2015, **11**:5947.

31. Noé F, Wu H, Prinz J-H, Plattner N: **J Chem Phys,** 2013, **139**:184114.

32. Scherer MK, Trendelkamp-Schroer B, Paul F, Perez-Hernandez G, Hoffmann M, Plattner N, Prinz J-H, Noé F: **J Chem Theory Comput,** 2015, **11**:5525.

33. Beauchamp KA, McGibbon R, Lin YS, Pande VS: **Proc Natl Acad Sci U S A,** 2012, **109**:17807.

34. Guarnera E, Vanden-Eijnden E: **J Chem Phys,** 2016, **145**:024102.

35. Prinz J-H, Wu H, Sarich M, Keller BG, Senne M, Held M, Chodera JD, Schütte C, Noé F: **J Chem Phys,** 2011, **134**:174105.

36. Sarich M, Noé F, Schütte C: **Multiscale Model Simul,** 2010, **8**:1154.

37. Martini L, Kells A, Hummer G, Buchete N-V, Rosta E. arXiv:1605.04328 (2016).

38. Buchete NV, Hummer G: **J Phys Chem B,** 2008, **112**:6057.

39. Bowman GR, Pande VS, Noé F (Eds): **An Introduction to Markov**
• **State Models and Their Application to Long Timescale Molecular Simulation, Advances in Experimental Medicine and Biology,** . Springer Heidelberg; 2014.
An extensive overview of methods and applications of classical Markov state models.

40. Noé F, Nüske F: **Multiscale Model Simul,** 2013, **11**:635.
• The original mathematical study that introduces the variational principle for molecular kinetics and the linear implementation in the variational approach. •• A presentation of the variational approach of molecular kinetics for a physical chemist community, with a new algorithmic approach.

41. Nüske F, Keller BG, Pérez-Hernández G, Mey ASJS, Noé F: **J**
•• **Chem Theory Comput,** 2014, **10**:1739.
A presentation of the variational approach of molecular kinetics for a physical chemist community, with a new algorithmic approach.

42. Fan K: **Proc Natl Acad Sci U S A,** 1949, **35**:652.

43. Vitalini F, Noé F, Keller BG: **J Chem Theory Comput,** 2015,
• **11**:3992.
A transferable basis functions in the variational approach is suggested.

44. Nüske F, Schneider R, Vitalini F, Noé F: **J Chem Phys,** 2016, **144**:054105.

45. McGibbon RT, Pande VS. arXiv:1602.08776 (2016).

46. Schütte C, Noé F, Lu J, Sarich M, Vanden-Eijnden E: **J Chem Phys,** 2011, **134**:204105.

47. Weber M: *ZIB report 09-27*. 2011.

48. Klus S, Schütte C. arXiv:1512.06527 (2015).

49. Wu H, Nüske F, Paul F, Klus S, Koltai P, Noé F. arXiv:1610.06773 (2016a).

50. McGibbon RT, Pande VS: **J Chem Phys,** 2015, **142**:124105.
• The variational principle is used to propose the Rayleigh trace as a cross-validation score, in order to optimize kinetic models without overfitting.

51. Arlot S, Celisse A: **Stat Surv,** 2010, **4**:40.

52. Molgedey L, Schuster HG: **Phys Rev Lett,** 1994, **72**:3634.

53. Aapo Hyvärinen EO, Juha Karhunen: *Independent Component Analysis*. John Wiley & Sons; 2001.

54. Schwantes CR, Pande VS: **J Chem Theory Comput,** 2013, **9**:2000.

55. Naritomi Y, Fuchigami S: **J Chem Phys,** 2011, **134**:065101.

56. Stanley N, Esteban-Martin S, Fabritiis GD: **Nat Commun,** 2014, **5**:5272.

57. Suárez E, Adelman JL, Zuckerman DM: **J Chem Theory Comput,** 2016, **12**:3473.

58. Schwantes CR, Pande VS: **J Chem Theory Comput,** 2015, **11**:600.

59. Ziehe A, Müller K-R: *ICANN 98,* . Springer Nature; 1998:675-680.

60. Pérez-Hernández G, Noé F: **J Chem Theory Comput,** 2016, **12**:6118-6129.

61. Husic BE, McGibbon RT, Sultan MM, Pande VS: **J Chem Phys,** 2016, **145**:194103.

62. Prinz J-H, Chodera JD, Noé F: **Phys Rev X,** 2014, **4**:011020.

63. Wu H, Paul F, Wehmeyer C, Noé F: **Proc Natl Acad Sci U S A,** 2016, **113**:E3221.

64. Noé F, Clementi C: **J Chem Theory Comput,** 2015, **22**:5002.
• Kinetic distances and kinetic maps are introduced for the study of molecular kinetics.

65. Noé F, Banisch R, Clementi C: **J Chem Theory Comput,** 2016,
•• **12**:5620.
The commute distances and commute maps are developed as an extension of kinetic maps. This article shows that a metric space can be derived in which Euclidean distance is related to commute time between molecular configurations.

66. Nadler B, Lafon S, Coifman RR, Kevrekidis IG: **Adv. Neural Inf. Process. Syst. (NIPS),** . 2005:955-962.

67. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror R, Eastwood M, Bank J, Jumper J, Salmon J, Shan Y, Wriggers W: **Science,** 2010, **330**:341.