Identification of slow molecular order parameters for Markov model construction

Guillermo Pérez-Hernández,[1] Fabian Paul,[2,a)] Toni Giorgino,[3,a)] Gianni de Fabritiis,[4,b)] and Frank Noé[1,c)]

[1)] *Freie Universität Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin, Germany.*

[2)] *Max-Planck-Institut for Colloids and Interfaces, Division Theory and Bio-Systems, Science Park Potsdam-Golm, 14424 Potsdam, Germany*[d)]

[3)] *Institute of Biomedical Engineering (ISIB), National Research Council of Italy (CNR), Corso Stati Uniti 4, I-35127 Padua, Italy.*

[4)] *GRIB, Barcelona Biomedical Research Park (PRBB), C/ Dr. Aiguader 88, 08003, Barcelona, Spain.*

[a)] equal contribution

[b)] Electronic Address: gianni.defabritiis@upf.edu

[c)] Electronic Address: frank.noe@fu-berlin.de

[d)] Freie Universität Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin, Germany.

SUPPLEMENTARY INFORMATION

Derivation of TICA

The generalized eigenvalue problem of Eq. (9) of the article, and more specifically the TICA problem can be derived in different ways. It goes back to the classical Ritz method[1] and can be found in many mathematical texts. In the following we sketch a standard derivation using variational calculus, see also[2,3] for a thorough discussion of this approach.

Let $\mathbf{r} \in \mathbb{R}^d$ be a vector of coordinates used, for example distances or Cartesian positions. Without restriction of generality we assume that $\mathbf{r}$ is mean-free, i.e. the mean of the data has already been subtracted. Note that $\mathbf{r}$ is contains generally only a subset of the full phase space coordinates, thus $\mathbb{R}^d$ is a subset of Ω.

We now seek new coordinates $\mathbf{z} \in \mathbb{R}^m$ as a linear transformation of $\mathbf{r}$ such that

1. $\mathbf{z}$ are uncorrelated

2. the autocovariances of $\mathbf{z}$ at a fixed lag time τ are maximal.

We will show that if coordinates $\mathbf{z}$ are given by a weighted sum of $\mathbf{r}$

$$z_i(\mathbf{r}) = \sum_{k=1}^{d} u_{ik} r_k \tag{1}$$

the weight coefficients have to fulfill the generalized eigenvalue problem (see theory section)

$$\mathbf{C}^r(\tau)\mathbf{u}_i = \lambda_i^{\ddagger}\mathbf{C}^r(0)\mathbf{u}_i \tag{2}$$

where $\mathbf{C}^r_\tau(\tau)$ is the time-lagged covariance matrix that is defined by:

$$c^r_{ij}(\tau) = \langle r_i(t) r_j(t+\tau)\rangle_t \tag{3}$$

To prove this we rewrite the covariance matrix of $\mathbf{z}$ and the time-lagged covariance matrix of $\mathbf{z}$ using the defining equations (1), and (3).

$$c^z_{ij}(0) = \langle z_i(t) z_j(t)\rangle_t = \sum_{k,l} u_{ik} u_{jl} \langle r_k(t) r_l(t)\rangle_t = \sum_{k,l} u_{ik} u_{jl} c^r_{kl}(0)$$

$$c^z_{ij}(\tau) = \langle z_i(t) z_j(t+\tau)\rangle_t = \sum_{k,l} u_{ik} u_{jl} \langle r_k(t) r_l(t+\tau)\rangle_t = \sum_{k,l} u_{ik} u_{jl} c^r_{kl}(\tau)$$

We wish to maximize $c^z_{ij}(\tau)$ (property 2) under the constraint, that $c^z_{ij}(0) = \delta_{ij}$ (property 1). We start by computing one coordinate z_1 with maximal autocovariance. It is given by the weighted sum $z_1 = \sum_{i,j} u_i u_j c^r_{ij}(0)$, where we used the shorthand notation $u_i = u_{1i}$. The constraint (1) for z_1 is now $c^z_{11}(0) = \sum_{i,j} u_i u_j c^r_{ij}(0) = 1$.

Since the matrix-elements $c^r_{ij}(0)$ are fixed, the autocovariance $c^z_{11}(\tau)$ can be treated as a differentiable function of the coefficients u_i. Therefore, we need to maximize the function

$$F(u_1, \ldots, u_d) = \left(\sum_{k,l} u_k u_l c^r_{kl}(\tau) \right) - \lambda^\ddagger_1 \left(\sum_{k,l} u_k u_l c^r_{kl}(0) - 1 \right)$$

where $\lambda^\ddagger$ is the Lagrange multiplier. We perform the maximization by setting the partial derivatives of F with respect to the weight coefficients to zero.

$$0 = \frac{\partial F}{\partial u_k} = \left(\sum_{l} u_l c^r_{kl}(\tau) \right) - \lambda^\ddagger_1 \left(\sum_{l} u_l c^r_{kl}(0) \right)$$

Rearranging and rewriting this equation in matrix-vector form leads to (2) for $i = 1$. The same argument is used for the subsequent eigenvalues. We now prove that the solutions of (2) fulfill the properties requested above:

1. **The IC's obtained by solving** (2) **are uncorrelated**: Let $\mathbf{u}_i$ be a generalized eigenvector with eigenvalue $\lambda^\ddagger_i$ and let $\mathbf{u}_j$ be a generalized eigenvector with eigenvalue $\lambda^\ddagger_j \neq \lambda^\ddagger_i$. Then the orthogonality condition

$$\mathbf{u}_i^T \mathbf{C}^r(0) \mathbf{u}_j = \delta_{ij} \tag{4}$$

will hold if $\mathbf{C}^r(0)$ and $\mathbf{C}^r(\tau)$ are symmetric matrices. If $\mathbf{u}_i$ and $\mathbf{u}_j$ are used as the weights in (1) this is equivalent to $c^z_{ij}(0) = \delta_{ij}$.

Proof:

$$\lambda^\ddagger_i \mathbf{C}^r(0)\mathbf{u}_i \cdot \mathbf{u}_j = \mathbf{C}^r(\tau)\mathbf{u}_i \cdot \mathbf{u}_j = \mathbf{u}_i \cdot \mathbf{C}^r(\tau)\mathbf{u}_j = \mathbf{u}_i \cdot \lambda^\ddagger_j \mathbf{C}^r(0)\mathbf{u}_j = \lambda^\ddagger_j \mathbf{C}^r(0)\mathbf{u}_i \cdot \mathbf{u}_j$$

Therefore $0 = (\lambda^\ddagger_i - \lambda^\ddagger_j)(\mathbf{u}_i^T \mathbf{C}^r(0)\mathbf{u}_j)$. Because $\lambda^\ddagger_i \neq \lambda^\ddagger_j$ the orthogonality condition must hold. This does not hold, if eigenvectors are degenerate i.e. $\lambda^\ddagger_i = \lambda^\ddagger_j$ for some i, j. However degeneracy can be avoided by changing the lag time τ such that no eigenvalues with large magnitude coincide. Solutions with smaller eigenvalues might still be degenerate, but this is unproblematic since these solutions are discarded for clustering. In addition, “fast” modes will necessarily be uncorrelated with “slow” modes, because their eigenvalues are far apart.

2. **The autocovariances at a fixed lag time τ are maximal**:

We show that the autocovariances are identical to the Lagrange multipliers, and thus to the eigenvalues in Eq. 2:

$$c^z_{ij}(\tau) = \lambda^\ddagger_i \delta_{ij} \tag{5}$$

To see this, multiply (2) with $\mathbf{u}_i^T$ from the left, to obtain

$$c^z_{ji}(\tau) = \mathbf{u}_j^T \mathbf{C}^r(\tau)\mathbf{u}_i = \lambda^\ddagger_i \mathbf{u}_j^T \mathbf{C}^r(0)\mathbf{u}_i$$

now use the orthogonality condition (4)

$$c^z_{ji}(\tau) = c^z_{ij}(\tau) = \mathbf{u}_j^T \mathbf{C}^r(\tau)\mathbf{u}_i = \lambda^\ddagger_i \delta_{ij}$$

To show that the optimum found is indeed a maximum, we calculate the Hessian of the constrained autocovariance $c^z_{11}(\tau)$. Its elements are:

$$H_{kl} = \frac{\partial^2 F}{\partial u_k \partial u_l} = c^r_{kl}(\tau) - \lambda^\ddagger_1 c^r_{kl}(0)$$

and show that it is a positive definite matrix

$$\mathbf{x}^T\mathbf{H}\mathbf{x} = \mathbf{x}^T\mathbf{C}^r(\tau)\mathbf{x} - \lambda^\ddagger_1 \mathbf{x}^T\mathbf{C}^r(0)\mathbf{x} < 0 \ \forall \mathbf{x}$$

We first expand $\mathbf{x}$ in the basis of the generalized eigenvectors $\mathbf{x} = \sum_i^m \mathbf{u}_i(\mathbf{u}_i \cdot \mathbf{x}) = \sum_i^m \mathbf{u}_i c_i$ and use equations (4) and (5)

$$\sum_{i,j} c_i c_j \mathbf{u}_i^T \mathbf{C}^r(\tau)\mathbf{u}_j - \lambda^\ddagger_1 \sum_{i,j} c_i c_j \mathbf{u}_i^T \mathbf{C}^r(0)\mathbf{u}_j = \sum_i c_i^2 \lambda^\ddagger_i - \lambda^\ddagger_1 \sum_i c_i^2$$

Without loss of generality, we assume that the solution vectors of Eq. (2) were sorted by descending eigenvalues $\lambda^\ddagger_i$ to obtain an ordering from "slow" modes to "fast" modes, $\lambda^\ddagger_1 > \lambda^\ddagger_2 > \ldots > \lambda^\ddagger_m$. From this follows $\sum_i c_i^2 \lambda^\ddagger_i - \lambda^\ddagger_1 \sum_i c_i^2 \leq 0$ for the first solution $\mathbf{u}_1$. The second solution is restricted to a subspace that is orthogonal to $\mathbf{u}_1$ according to (4) and (5)

$$\mathbf{x}^T\mathbf{C}^r(0)\mathbf{u}_1 = \mathbf{x}^T\mathbf{C}^r(\tau)\mathbf{u}_1 = 0$$

Therefore we can ignore the coefficient c_1 in the development of $\mathbf{x}$ and obtain the quadratic form

$$\sum_{i=2} c_i^2 \lambda^\ddagger_i - \lambda^\ddagger_2 \sum_{i=2} c_i^2$$

Again, this is negative, because $\lambda^\ddagger_2$ is the largest eigenvalue in the sum. This procedure can be extended to the third, fourth,... eigenvalue, showing that the optima are minima for all solutions.

As a result, we can sort the solution vectors of Eq. (2) by descending eigenvalues $\lambda_i^\ddagger$ to obtain an ordering from "slow" modes to "fast" modes.

Symmetry and Symmetrization of the time-lagged covariance matrix

Consider the correlation matrix of mean-free coordinates $\mathbf{r}$ for lag time τ:

$$c^r_{ij}(\tau) = \langle r_i(t)\, r_j(t+\tau)\rangle$$

and the correlation matrix for lag time τ:

$$\begin{aligned}\mathrm{cor}^r_{ij}(\tau) &= \frac{c^r_{ij}(\tau)}{\sigma_i\sigma_j} = \frac{\langle r_i(t) r_j(t+\tau)\rangle}{\sqrt{\langle r_i^2(t)\rangle\langle r_j^2(t)\rangle}} \\ &= \int_x\int_y dxdy\; xy\; p(r_i(t)=x, r_j(t+\tau)=y)\end{aligned}$$

where $p(x(t)=x, y(t+\tau)=y)$ is the unconditional transition probability between the set $S_1=\{r_i=x\}$ and the set $S_2=\{r_j=y\}$ within time lag τ. In statistically reversible dynamics, such an unconditional set-transition probability is symmetric (this follows directly from integrating the detailed balance condition $\mu(\mathbf{x})p_\tau(\mathbf{y}\mid\mathbf{x}) = \mu(\mathbf{y})p_\tau(\mathbf{x}\mid\mathbf{y})$ over the sets). Thus, we can exchange time indexes and show:

$$\begin{aligned}\mathrm{cor}^r_{ij}(\tau) &= \int_x\int_y dxdy\; xy\; p(r_i(t+\tau)=x, r_j(t)=y) \\ &= \int_y\int_x dydx\; yx\; p(r_j(t)=y, r_i(t+\tau)=x) \\ &= \mathrm{cor}^r_{ji}(\tau).\end{aligned}$$

And then trivially

$$c^r_{ij}(\tau) = c^r_{ij}(\tau) \quad \forall\tau$$

When estimating correlation or covariance matrices from simulations, one cannot expect $c_{ij}=c_{ji}$ to hold. A trivial method is to use

$$c_{ij}(\tau) = \frac{1}{2}(c_{ij}(\tau) + c_{ji}(\tau))$$

where $c_{ij}(\tau)$ is the simulation estimate.

Simulation setup, KID

The coordinates of the phosphorylated KID domain (28 residues, CREB residues 119-146) were extracted from chain B of the entry 1KDX deposited in the Protein Data Bank. The entry represent the folded configuration of the pKID-CBP bound structure, determined through NMR[4]. Neutral acetylated and N-methyl caps were added to avoid artifactual charges at the peptide's termini; the protein was solvated with 6572 water molecules and a 85 mM KCl concentration (matching the experimental ionic strength). The system was then parametrized with the AMBER ff99SB-ILDN forcefield[5]; water and ions were modeled respectively with the TIP3P and Joung-Cheatham parameter sets[6,7]. The system thus prepared was first equilibrated for 24 ns in the constant-pressure ensemble, during which it stabilized at a volume of approximately 60 Å^3. The peptide was then denatured by heating it at 500 K for 17.6 ns in constant-volume conditions; 176 frames were extracted from this trajectory and used as starting configurations for the production runs. All of the simulations were performed with a time step of 4 fs, enabled by the hydrogen mass repartitioning scheme[8]; long-range electrostatic forces were computed with the particle-mesh Ewald summation method. A nonbonded cutoff distance of 9 Å was used with a switching distance of 7.5 Å for Van der Waals interactions, while the lengths of bonds involving hydrogen atoms were constrained with the SHAKE algorithm.

A set of 7706 production runs was executed on the GPUGRID distributed computing network[9]. Each simulation was performed in the constant-volume ensemble at 315K for 24 ns with the same parametrization used for equilibration, storing structural snapshots every 100 ps. Each production simulation begun either from one of the configurations visited during the denaturation run, or frames visited during preceding production trajectories. Starting frames were selected iteratively with an adaptive strategy in order to minimize the statistical uncertainty on the largest eigenvalue, computed on the already available simulation data, based on Singhal and Pande's algorithm[10].

Error estimation for the implied timescales

In order to provide an estimate of the errors in the implied timescales (both in KID and MR121-GSGSW) we produced 100 samples of our microstate-trajectory data using a bootsrapping procedure. We then estimated timescales for the 100 samples separately and computed mean and standard deviation. The results are shown in figure 1 and Tables I and II.

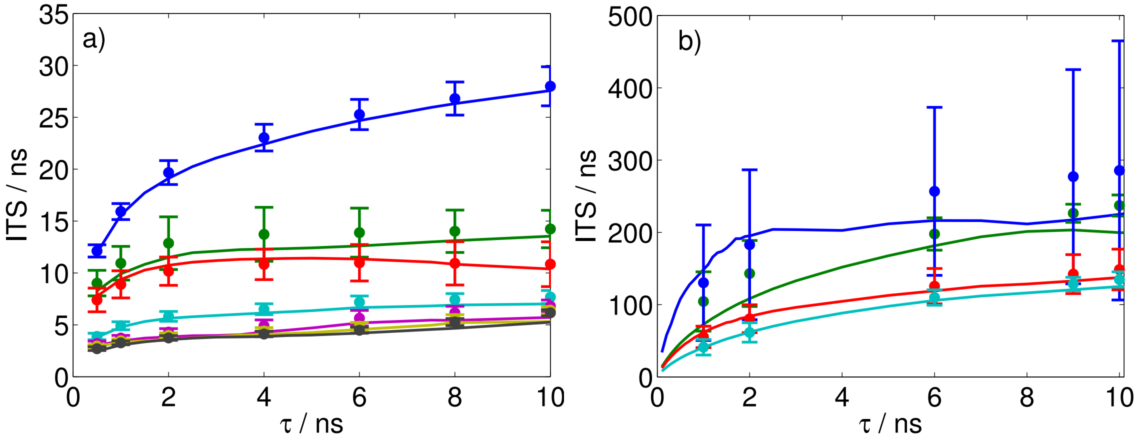

Figure 1. Estimated implied timescales $\tilde{t}_i(\tau)$ of the slowest processes in MR121-GSGSW (panel a) and KID (panel b). Solid curves indicate the values estimated for the whole datasets of MR121-GSGSW and KID, respectively (cf. Figs. 2 and 4 in the article). The solid dots show the averages over the 100 bootstrapped samples and the errorbars the respective standard deviations

Whereas the MR121-GSGSW trajectories show very little uncertainty, the second slowest timescale of the KID dataset has still a quite large uncertainty (cf. Table II, second row). This indicates that the process associated with the second eigenvector of our MSM has not been sampled very well. Still, the fact that the expectation values of the bootstrapped timescales reach a plateau with respect to τ supports the validity of the underlying MSM.

	$\mu(\sigma)$				
i	$\tilde{t}_i(\tau = 2)$	$\tilde{t}_i(\tau = 4)$	$\tilde{t}_i(\tau = 6)$	$\tilde{t}_i(\tau = 8)$	$\tilde{t}_i(\tau = 10)$
2	16(0.8)	20(1.2)	23(1.3)	25(1.5)	27(1.6)
3	11(1.6)	13(2.5)	14(2.6)	14(2.4)	14(2.0)
4	9(1.3)	10(1.4)	11(1.5)	11(1.8)	11(2.1)
5	5(0.4)	6(0.5)	6(0.5)	7(0.6)	7(0.6)
6	4(0.2)	4(0.3)	5(0.6)	6(0.7)	6(0.6)
7	3(0.2)	4(0.2)	4(0.3)	5(0.4)	6(0.4)
8	3(0.2)	4(0.2)	4(0.3)	4(0.3)	5(0.3)

Table I. MR121-GSGSW: sample expectation values(μ) and sample standard deviations(σ) of the first 7 implied timescales $\tilde{t}_i(\tau)$ averaged over 100 bootstrapped samples. All quantities in ns.

	$\mu(\sigma)$				
i	$\tilde{t}_i(\tau = 1)$	$\tilde{t}_i(\tau = 2)$	$\tilde{t}_i(\tau = 6)$	$\tilde{t}_i(\tau = 9)$	$\tilde{t}_i(\tau = 10)$
2	130(80)	183(104)	257(116)	277(148)	286(179)
3	104(41)	143(46)	198(22)	226(13)	237(15)
4	55(15)	81(19)	126(24)	142(27)	149(28)
5	41(11)	62(13)	110(11)	129(9)	135(11)

Table II. KID: sample expectation values(μ) and sample standard deviations(σ) of the first 4 implied timescales $\tilde{t}_i(\tau)$ averaged over 100 bootstrapped samples. All quantities in ns.

REFERENCES

[1] W. Ritz, J. Reine Angew. Math. **135**, 1 (1909).

[2] I. Jolliffe, *Principal Component Analysis*, 2nd ed. (Springer, New York, 2002).

[3] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, “Independent Component Analysis,” (John Wiley & Sons, 2001) Chap. 18, p. 344.

[4] K. Sugase, H. J. Dyson, and P. E. Wright, Nature **447**, 1021 (2007).

[5] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, Proteins **78**, 1950 (2010).

[6] I. S. Joung and T. E. Cheatham, J. Phys. Chem. B **112**, 9020 (2008).

[7] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).

[8] K. A. Feenstra, B. Hess, and H. J. C. Berendsen, J. Comput. Chem. **20**, 786 (1999).

[9] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis, J. Chem. Inf. Model. **50**, 397 (2010).

[10] N. S. Hinrichs and V. S. Pande, J. Chem. Phys. **126**, 244101 (2007).