High-Resolution Visualisation of the States and Pathways Sampled in Molecular Dynamics Simulations

Nicolas Blöchliger, Andreas Vitalis & Amedeo Caflisch

University of Zurich, Department of Biochemistry, Winterthurerstrasse 190, CH-8057 Zurich.

We have recently developed a scalable algorithm for ordering the instantaneous observations of a dynamical system evolving continuously in time. Here, we apply the method to long molecular dynamics trajectories. The procedure requires only a pairwise, geometrical distance as input. Suitable annotations of both structural and kinetic nature reveal the free energy basins visited by biomolecules. The profile is supplemented by a trace of the temporal evolution of the system highlighting the sequence of events. We demonstrate that the resultant SAPPHIRE (States And Pathways Projected with HIgh REsolution) plots provide a comprehensive picture of the thermodynamics and kinetics of complex, molecular systems exhibiting dynamics covering a range of time and length scales. Information on pathways connecting states and the level of recurrence are quickly inferred from the visualisation. The considerable advantages of our approach are speed and resolution: the SAPPHIRE plot is scalable to very large data sets and represents every single snapshot. This minimizes the risk of missing states because of overlap or prior coarse-graining of the data.

Present day science or more broadly society record observations as a function of time in diverse contexts. Data on meteorological phenomena, communication tracking, or financial markets, to name a few, are all mined for the generation of predictive models. The raw data are usually unfit for human consumption due to their high dimensionality and sheer size. Both aspects limit the types of analyses performed on these “big data” to those algorithms with satisfactory scaling properties. In biophysics, long computer simulations of the trajectories of complex macromolecules with high-dimensional representations have become commonplace. Molecular dynamics (MD) simulations of proteins and other biomolecules record stochastic trajectories, in which the macromolecule visits a number of different, metastable states (free energy basins) connected by an ensemble of pathways of interconversion. The latter report on the barriers of the underlying free energy landscape. Because millions of snapshots are now routinely recorded for thousands of coupled degrees of freedom, MD trajectories call for scalable algorithms that are able to provide information-preserving projections for this specific class of complex systems. We have recently introduced such an algorithm and provide a brief description next.

Given a definition of distance between trajectory snapshots, the entire data set is considered as a complete graph with vertices corresponding to snapshots and edge weights given by the pairwise distances between snapshots. Either the exact or an approximation to the minimum spanning tree are computed. The available edges at each point are those connecting any snapshot not considered yet with any snapshot already included. The resulting sequence has the crucial property of stepping through high density regions one by one. It can therefore be expected that all free energy basins will appear as groups of nearby points along the progress index. Importantly, the progress index does not reflect the temporal nature of the input data in any way, i.e., it is generally independent of input order. Because every snapshot is considered, the limiting resolution is optimal given the time resolution of the input trajectory.

The progress index can be annotated both kinetically and structurally to provide an informative and compact representation of all major states visited by the input trajectory. The procedure has several advantages over projections using geometric or kinetic distances from a reference state to order snapshots. First, it maximizes resolution as mentioned above. Second, it avoids overlap precisely because the ordering is not with respect to a
We present results on two different proteins. The data on Fip35 exhibits reversible folding at a simulation temperature of 395 K in explicit solvent molecular dynamics runs of a total length of 200 μs. Specifically, the trajectories show that FiP35 converts 10–15 times between an unfolded state that is very low in secondary structure and a three-stranded β-sheet, and a coil-like unfolded state. The small WW domain, come from two long MD trajectories and allow us to stipulate a two-state Markov state model, and we can derive the mean first passage times in either direction. A cut function as used elsewhere allows us to stipulate a two-state Markov state model, and we can derive the mean first passage times in either direction.

Results

We demonstrate that the information summarized in the resultant SAPPHIRE plot threefold, viz., structurally, kinetically, and with times of occurrence in the original trajectories (called dynamical trace hereafter). We demonstrate that the information summarized in the resultant SAPPHIRE (States And Pathways Projected with High Resolution) plot provides an efficient means of identifying the statistically reliable states visited by a complex, dynamical system while enabling a rapid assessment of state interconversion and recurrence, which provide information on kinetic pathways and simulation convergence, respectively.

Reversible folding of a 35-residue protein domain. FiP35 exhibits reversible folding at a simulation temperature of 395 K in explicit solvent molecular dynamics runs of a total length of 200 μs. Specifically, the trajectories show that FiP35 converts 10–15 times between an unfolded state that is very low in secondary structure and the native topology, viz., a twisted, three-stranded β-sheet. All following results refer to a specific computational model and sampling protocol underlying the trajectories being analysed. Due to the protein’s small size, it is possible to provide a comprehensive, structural representation at the backbone level using a DSSP annotation resolved by residue. Fig. 1(a) shows the

Figure 1 | SAPPHIRE plot for FiP35. (a) The progress index, of 10^6 snapshots from 200 μs of MD data, is annotated with kinetic information (tMFP, black curve), dynamical trace (red dots), DSSP assignment by residue (legend on top) and the state partitioning of Berezovska et al. These annotations are only shown for every 1000th, 100th, 1000th and 500th snapshots, respectively, in order to maintain readability at fixed figure resolution. The limits of possible definitions of the folded and unfolded states for the computation of transition path times are indicated by the blue, horizontal lines. Cartoons of a snapshot in the native state and an unfolded conformation are shown. (b) Zoom-in on the transition region of the SAPPHIRE plot shown in (a). The various annotations are shown for every 100th, 100th, 500th, and 250th snapshots, respectively. Representative conformations of I1 and I2 are shown as cartoons. The box highlights a particular state (see text).
and smaller of the two unlabelled states as L and S, respectively. By
SCIENTIFIC
two intermediates 18–20, and we have additionally annotated the
Previous analyses of the same data suggested the existence of at least
i.e
no consistent secondary structure and is kinetically homogeneous,
Berezovska et al.
SAPPHIRE plot with the state partitioning proposed by
states identified by NMR experiments25,26 are populated significantly
in the trajectories albeit with inaccurate weights. These metastable
states can often be correlated with isomerizations of the disulphide
bridges, in particular Cys14-Cys3827-28. Shaw et al.15 used a stochastic
algorithm to obtain a coarse, kinetic clustering of their 1.03 ms MD
trajectory of BPTI relying on the autocorrelation of interatomic
distances. Empirically, they found that five states with significant
populations could be identified reliably. These states were
annotated structurally. The most important states (the smaller of
the two is the one most resembling the crystal structure) are clearly
seen in the SAPPHIRE plot as the first two basins from the left
(Fig. 2). The structural annotation we select here confirms that the
barrier identified by the kinetic analysis (black line) is related to the
isomerization of the disulphide bond Cys14-Cys38. The dynamical
trace uses the colour scheme of Shaw et al. (distinguishing the red,
blue, green, purple, and black states). It unmasks that both major
states are long-lived and that there is a clear separation of time scales
with respect to the mixing time within each basin.

Fig. 2 indicates that there is a mismatch in assignment between
that of Shaw et al. and the positions on the x-axis for several, short
excursions into a given state. As an example, we consider the
highlighted trajectory segment sampled at ~0.5 ms that is annotated by
Shaw et al. to be in the red state but that is placed by the SAPPHIRE
plot in the basin corresponding to the blue state. To understand
why this may be the case, we first note that the structural annotations
generally reveal a small amount of mixing that may be considered
erroneous. Indeed, for the segment in question, inspection of instant-
aneous values yields that the Cys14 side chain angles adopt the values
for the blue state, but the χ₁ angle and the χ₃ angle of Cys38 do not
(not shown). The combination of values for the dihedral angles
places this segment outside of the list of states characterized previ-
ously29. It appears kinetically homogeneous and may correspond to
an incomplete or blocked transition. Its sampling weight is so low
that neither the SAPPHIRE plot nor the kinetic clustering are sensi-
tive enough to resolve it as an independent state. Due to its inter-
mediate nature, it is lumped into either one of the adjacent states. A
very similar effect is observed for a second, highlighted segment (at
~0.75 ms), for which just the two Cys38 side chain angles deviate from
the blue state.

The SAPPHIRE plot for BPTI also reveals that over the course of
the 1.03 ms trajectory the purple and black states are sampled exten-
sively just once and twice, respectively. This allows us to infer a lack
of recurrence, i.e., sampling weights are unlikely to be converged.
Poor sampling may also limit the number of states obtainable from
Markov state models29 and decrease the accuracy of any extracted
passage times. The bottom panel of Fig. 2 zooms into a very thin time
slice to illustrate the pathway taken to reach the black state. This
is annotated by cartoons and a specific, interatomic distance involving
a residue identified by the original authors as being discriminative for
this state31. The final result we want to mention in this short note is
that the SAPPHIRE plot suggests the green state to be partitioned
further. The kinetic annotation is consistent with the dynamical trace
in that the two major substates of the green state are homogeneous
with respect to the times they were sampled at (no mixing). This is
despite the fact that they appear to be directly adjacent to one another in terms of transition pathways.

We conclude the description of the performance of the SAPPHIRE plot with a note of caution. In Fig. 2, toward the right side of the largest basin, there is a region of both temporal and geometric ambiguity most clearly seen by the overlap of blue and red dots in the dynamical trace. Here, the progress index is placing “fringe” regions of both basins. This weakness results from an insufficient sampling density for these lower likelihood regions that immediately surround well-defined states. It is rectified by having better time resolution or, at the risk of a decrease in resolution, by lowering the dimensionality of representation. We show the data on the sparsely sampled trajectory here to illustrate both the general robustness and possible errors encountered with smaller data sets.

**Discussion**

With growing computing resources and growing data sets, it has become paramount to use tools that quickly and efficiently improve our understanding of a system as complex as a biomolecule. The data required for the SAPPHIRE plot with all three annotations can usually be computed in near-linear time in a single run by the CAMPARI simulation and analysis package (http://campari.sourceforge.net). The plots are ideally generated as fully scalable vector graphics. At fixed resolution, readability may be improved by displaying annotations more sparsely, and we have done this for both figures. The required user input is the definition of a suitable measure of pairwise distance, and this choice may also help determine which structural annotations to use.

In Figs. 1 and 2, we have shown that SAPPHIRE plots offer an efficient procedure for the analysis and comprehensive pictorial description of complex systems undergoing stochastic evolution, such as proteins. Thermodynamics are resolved quantitatively, and the construction of the ordering of snapshots minimizes the risk of state overlap. Major basins are delineated easily by all three annotation functions. Qualitative information about pathways is available at the temporal resolution offered by the trajectory itself. The rapid availability of this information is not only valuable per se but can also be used to guide further simulations and analyses.

**Methods**

The algorithm underlying the SAPPHIRE plot has been described qualitatively above (see Introduction and Results). For a complete description we refer the reader to the original publication9. In terms of efficiency, the overall annotation procedure requires
linear time with respect to the number of snapshots. The calculation of the required spanning tree is the most expensive step of the algorithm and is aided by heuristics in either variant (exact or approximate). The approximate version can be scaled to very large data sets. When using this version, it will generally be useful to rerun the analysis a few times due to the stochastic nature of the spanning tree. In particular, the kinetic annotation function is sensitive to where a basin appears in the progress index and how well basins to the left have been captured.

The FiP35 trajectory encompasses 10^6 snapshots saved every 200 ps, while the 41250 snapshots of data on BPTI have a coarser time resolution of 25 ns. Pairwise distances were defined as the coordinate root mean square deviation (RMSD) computed over the backbone oxygen and nitrogen atoms of residues 7–29 for FiP35 and over 695 nonsymmetric atoms for BPTI. These choices reflect the different levels of variance in the two data sets. The approximate algorithm was used for both systems. It requires additional parameters as follows. The number of guesses to find putative nearest neighbours from within a limited space defined by preorganization of the data via clustering was set to a value of 1000 throughout. The lower threshold radii for clusters were 3.0 and 2.5Å for FiP35 and BPTI, respectively, and the upper threshold radii were 10.0 and 3.0Å. The required input data took ~11 and ~5 hours to compute on a single Intel Xeon core (either E5435 or E5410) for Figs. 1 and 2, respectively.

1. Fu, T.-C. A review on time series data mining. Eng. Appl. Artif. Intel. 24, 164–181 (2011).

2. Keffer, J. & Hauser, H. Visualization and visual analysis of multifaceted scientific data: A survey. IEEE Trans. Vis. Comput. Graph. 19, 495–513 (2013).

3. Rysavy, S. J., Bromley, D. & Daggett, V. DIVE: A graph-based visual-analytics framework for big data. IEEE Comput. Graph. Appl. 34, 26–37 (2014).

4. Bohlodi, M. et al. in Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives (ed Fatih, M.), Ch. Towards an integrated platform for big data analysis, 47–56 (Springer, 2013).

5. Dör, R. O., DIRKS, R. M., Grossman, J. P., Xu, H. & Shaw, D. E. Biomolecular simulation: A computational microscope for molecular biology. Annu. Rev. Biophys. 41, 429–452 (2012).

6. Spiliopoulos, D. & Caflisch, A. Molecular dynamics simulations of bromodomain reveal binding-site flexibility and multiple binding modes of the natural ligand acetyl-lysine. Isr. J. Chem. in press. DOI: 10.1002/jic.201400009 (2014).

7. Kaplansky, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. Nat. Struct. Biol. 9, 646–652 (2002).

8. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of protein folding: The energy landscape perspective. Annu. Rev. Phys. Chem. 48, 545–600 (1997).

9. Blochliger, N., Vitalis, A. & Caflisch, A. A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems. Comp. Phys. Comm. 184, 2446–2453 (2013).

10. Xu, R. & Wunsch II, D. C. Clustering algorithms in biomedical research: A review. IEEE Rev. Biomed. Eng. 3, 120–154 (2010).

11. Shaw, D. E. et al. Atom-level characterization of the structural dynamics of proteins. Science 330, 341–346 (2010).

12. Liu, F. et al. An experimental survey of the transition between two-state and downhill protein folding scenarios. Proc. Natl. Acad. Sci. USA 105, 2369–2374 (2008).

13. Berndl, K. D., Güntert, P., Orbanos, L. P. M. & Wüthrich, K. Determination of a high-quality nuclear magnetic resonance solution structure of the bovine pancreatic trypsin inhibitor and comparison with three crystal structures. J. Mol. Biol. 227, 757–775 (1992).

14. Krivov, S. V. & Kaplansky, M. One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. J. Phys. Chem. B 110, 12689–12698 (2006).

15. Lindorff-Larsen, K., Piana, S., Dör, R. O. & Shaw, D. E. How fast-folding proteins fold. Science 334, 517–520 (2011).

16. Kellogg, E. H., Lange, O. F. & Baker, D. Evaluation and optimization of discrete models of protein folding. J. Phys. Chem. B 116, 11405–11413 (2012).

17. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637 (1983).

18. Krivov, S. V. The free energy landscape analysis of protein (FiP35) folding dynamics. J. Phys. Chem. B 115, 12315–12324 (2011).

19. Lane, T. J., Bowman, G. R., Beauchamp, K., Voez, V. A. & Pande, V. S. Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. J. Am. Chem. Soc. 133, 18413–18419 (2011).

20. Bereczvics, G., Prada-Garcia, D. & Rao, F. Consensus for the FiP35 folding mechanism? J. Phys. Chem. B 139, 035102 (2013).

21. Wintjens, R. et al. 1H NMR study on the binding of Pin1 Trp-Trp domain with phosphothreonine peptides. J. Biol. Chem. 276, 25150–25156 (2001).

22. Becerra, S., Skrbic, T., Covino, R. & Faccioï, P. Dominant folding pathways of a WW domain. Proc. Natl. Acad. Sci. USA 109, 2330–2335 (2012).

23. McGibbon, R. T. & Pande, V. S. Learning kinetic distance metrics for Markov state models of protein conformational dynamics. J. Chem. Theor. Comput. 9, 2900–2906 (2013).

24. Liu, F., Nakamae, M. & Gruebele, M. The transition state transition time of WW domain folding is controlled by energy landscape roughness. J. Phys. Chem. 131, 195101 (2009).

25. Otting, G., Liepinsh, E. & Wüthrich, K. Disulfide bond isomerization in BPTI and BPTI(G36S): An NMR study of correlated mobility in proteins. Biochemistry 32, 3571–3582 (1993).

26. Greu, M. J., Wang, C. & Palmer III, J. A. Disulfide bond isomerization in basic pancreatic trypsin inhibitor: Multisite chemical exchange quantified by CPMG relaxation dispersion and chemical shift modeling. J. Am. Chem. Soc. 125, 14324–14335 (2003).

27. Long, D. & Brüschweiler, R. Atomicistic kinetic model for population shift and allostery in biomolecules. J. Am. Chem. Soc. 133, 18999–19005 (2011).

28. Xue, Y., Ward, J. M., Yuwen, T., Podkorytov, I. S. & Skrynnikov, N. R. Microsecond time-scale conformational exchange in proteins: Using long molecular dynamics trajectory to simulate NMR relaxation dispersion data. J. Am. Chem. Soc. 134, 2555–2562 (2012).

29. Noe, F., Wu, H., Prinz, J.-H. & Plattner, N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. J. Phys. Chem. B 118, 18414–18418 (2013).

30. Vitalis, A. & Caflisch, A. Efficient construction of mesoscale networks from molecular dynamics trajectories. J. Chem. Theor. Comput. 8, 1108–1120 (2012).

31. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. J. Mol. Graph. 14, 33–38 (1996).

Acknowledgments

The authors thank D.E. Shaw Research for sharing the trajectory data and their state annotation for BPTI (colour code in Fig. 2). We thank Dr. Francesco Rao for providing the state partitioning for FiP35 we use in Fig. 1. AV acknowledges financial support from the Holcim foundation. This work was supported in part by a grant from the Swiss National Science Foundation to AC.

Author contributions

N.B., A.V. and A.C. contributed to study design. N.B. analysed the data and created the figures and captions. A.V. wrote the manuscript, which was reviewed by all authors.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Blochliger, N., Vitalis, A. & Caflisch, A. High-Resolution Visualisation of the States and Pathways Sampled in Molecular Dynamics Simulations. Sci. Rep. 4, 6264; DOI:10.1038/srep06264 (2014).

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third-party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/.

www.nature.com/scientificreports