CoCo-MD: A Simple and Effective Method for the Enhanced Sampling of Conformational Space

Journal:	*Journal of Chemical Theory and Computation*
Manuscript ID	ct-2017-00327m
Manuscript Type:	Article
Date Submitted by the Author:	27-Mar-2017
Complete List of Authors:	Shkurti, Ardita; Sci-Tech Daresbury Styliari, Ioanna; University of Hertfordshire School of Life and Medical Sciences Balasubramanian, Vivek; Rutgers University Department of Electrical and Computer Engineering Bethune, Iain; The University of Edinburgh, Edinburgh Parallel Computing Centre Jha, Shantenu; Rutgers, Electrical and Computer Engineering Laughton, Charles; University of Nottingham, School of Pharmacy

CoCo-MD: A Simple and Effective Method for the Enhanced Sampling of Conformational Space

Ardita Shkurti[†§], Ioanna Danai Styliari[†○], Vivek Balasubramanian[‡], Iain Bethune[#], Shantenu Jha[‡], Charles A. Laughton[†]*

[†]School of Pharmacy and Centre for Biomolecular Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK.

[‡]Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854, USA.

[#]EPCC, The University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh, , UK.

CoCo ("Complementary Coordinates") is a method for ensemble enrichment based on Principal Component Analysis (PCA) that was developed originally for the investigation of NMR data. Here we investigate the potential of the CoCo method, in combination with molecular dynamics simulations (CoCo-MD), to be used more generally for the enhanced sampling of conformational space. Using the alanine penta-peptide as a model system, we find that an iterative workflow, interleaving short multiple-walker MD simulations with long-range jumps through conformational space informed by CoCo analysis, can nearly double the rate of sampling of conformational space for the same computational effort (total number of MD timesteps). Rare states that in conventional simulations are slow to be reached can be encountered more than twenty times sooner using CoCo-MD. The PCA-based approach means that optimal collective variables to enhance sampling need not be defined in advance by the user, but are identified automatically and are adaptive, responding to the characteristics of the developing ensemble. In addition the approach does not require any adaptations to the associated MD code, and is compatible with any conventional MD package.

Introduction

Adequate sampling remains a challenge in molecular simulation. This, as much as the accessing of increasingly long timescales, has been the driver for many decades of development in both hardware and software. Individual simulations have been accelerated through hardware developments such as MD-GRAPE[1] and ANTON[2]. New integrators open up the possibility of accelerating simulations by extending the fundamental timestep[3]. An alternative approach has been to replace single, long simulations with multiple, shorter ones[4][5], and possibly integrating the data through approaches such as Markov Chain Models[6] or Replica Exchange strategies[7]. A complementary approach is to accelerate sampling by modifying – generally flattening - the effective potential energy surface, e.g. as in metadynamics[8], accelerated dynamics[9] and simulated tempering[10].

Underlying many of these methods is a process of reducing the dimensionality of the search space, and collective variables (CVs) are commonly used to do this. However, choosing an effective set of CVs for use in such methods is a major challenge, particularly if one does not wish to impose a set of these a priori, or if one wishes to adapt them during the simulation. A variety of novel, and established, algorithms for the unsupervised and adaptive construction of CVs exist, but software to close the loop by providing on-the-fly integration of MD simulation and analysis is limited. The work of Preto and Clementi [11], interleaving cycles of MD simulation with data analysis through Locally Scaled Diffusion Maps [12], is an example of this approach and has been shown to be able to increase sampling rates by an order of magnitude.

Related methods include the non-targeted PaCS-MD method of Harada and Kitao[13], variants thereof[14], and the PCA-based method of Peng and Zhang[15].

CoCo ("Complementary Coordinates") is a PCA-based method developed originally for assessing and potentially enhancing conformational variety within NMR-derived ensembles of molecular structures[16]. The basis of the CoCo method is as follows. Firstly, the original ensemble of size *N* (which in the case of deposited NMR data, is typically is between ten and fifty structures) is analysed by PCA in Cartesian coordinate space. Next the coordinates of the structures in a low-dimensional PC subspace (typically three or four dimensions) are used to construct a multidimensional histogram; the histogram boundaries in each dimension being chosen to just enclose all *N* points. Following this, a set of *N* diverse unoccupied bins in the histogram is identified by an iterative procedure: the unoccupied bin most distant (in PC space) from any occupied bin is chosen, then the bin most distant from all occupied bins plus this new one, and so on until *N* bins are chosen. Finally the PC coordinates of the midpoints of the chosen bins are back-transformed to generate *N* new conformations of the molecule of interest, distinct from those in the initial set. In [16] we have shown that though these new structures are not guaranteed to be physically realistic, on assessment against the original NMR restraint data they can prove to be valid and valuable extensions to the ensemble.

We hypothesise that this procedure, within an iterative simulation/analysis workflow, could be applied more generally as an enhanced conformational sampling method. The approach is similar to that discussed by Harada, but the CoCo approach has a particular feature – that the new "seed"

structures for each round of MD are not taken from within the existing ensemble, but from outside it (see Figure 1). We hypothesise that this feature could have a major impact on the ability of the method to rapidly explore new regions of conformational space.

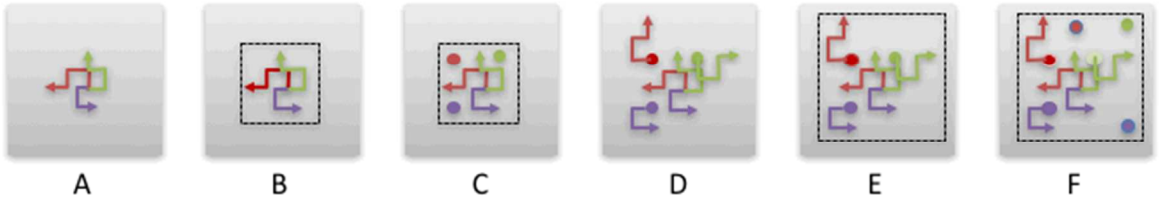

Figure 1. Key steps in the CoCo-MD method. **A**: *N* short independent MD simulations are initiated from a starting conformation. **B:** The bounding box in a PC-space is identified. **C:** *N* unsampled regions within the bounding box are identified. **D:** New short simulations are run, starting from the points identified in C. The process then iterates: **E:** The new bounding box is identified. **F:** *N* new unsampled regions are selected and MD runs performed, etc. Note that there is no particular connection between individual simulations and the subsequently-identified new start points; i.e. the colour coding of the circles is arbitrary, but illustrates that one can regard the method as generating a set of discontinuous trajectories, where conventional MD-directed sections are separated by long-distance "jumps" through conformational space.

Here we describe the implementation of the CoCo-MD workflow, and benchmark its performance against conventional MD (CMD) simulations on the alanine penta-peptide, a popular test-bed for the investigation of simulation methods and performance[17][18]. We first describe the construction of the reference CMD ensemble, and the metrics used to assess its convergence. Then we analyse the sampling power of the CoCo-MD method, investigating the

sensitivity of its performance to various user-definable parameters and against a variety of metrics.

Computational Methods

All simulations were performed using AMBER12[19] and the parm99SB force field. Models of Ace-Ala_5-NMe were built in both the extended and helical state and immersed in a truncated octahedral box of TIP3P water with a minimum buffer of 15 angstroms between any solute atom and the box boundary. After relaxation, production simulations were run at 300K in an NPT ensemble. SHAKE was applied to all bonds and a 2 fs time step was used. Snapshots were saved every ps.

Principal component analysis was performed using the pyPcazip toolkit[20] and CoCo analysis using pyCoCo (www.bitbucket.org/extasy-project/coco). CoCo-MD workflows were written in Python, using the ExTASY wrappers (www.bitbucket.org/extasy-project/wrappers) to interface the MD and analysis codes. Large-scale CoCo-MD simulations were performed using the ExTASY domain specific workflow system (which in turn uses Ensemble-toolkit) [21,22].

Results and Discussion

Construction and Analysis of the Reference Dataset. Using large-scale CMD simulations, we constructed a reference ensemble for the alanine penta-peptide against which to benchmark the performance of the CoCo-MD method. One hundred independent 10 ns simulations were run

starting from each of the extended and helical states of the peptide, differing in the initial assignment of randomised velocities at 300 K. Coordinates were saved every picosecond, giving a total of two million configurations representative of two microseconds of aggregated simulation time.

There is no definitive way to define convergence, but a variety of tests can be useful[23]. Because CoCo is a PCA-based method, we used PCA-related metrics for this. After iterative, "Procrustes", least-squares fitting[24], we used the pyPcazip package to perform PCA in Cartesian space on the heavy atoms of the complete dataset (2 million configurations) and calculated the coordinates for each snapshot in the three-dimensional subspace defined by the three top PCs (which capture 32%, 13%, and 10% of the total variance respectively). The scores data were then discretised into 30 uniform bins in each dimension. Subsets of the data thus formed three-dimensional histograms that could be compared with each other using a variety of metrics. Though one can argue that since this PC1-3 subspace only captures c. 55% of the total variance, it will not measure complete convergence in conformational sampling, a) this 3D subspace is enough to delineate key conformational states (see below), and the first few PCs are typically (though not always) capture conformational transitions that are slowest and most non-Gaussian in their distributions.

Comparison of block averages is one test of convergence, so firstly we examined the similarity in the histograms generated from the data in the first and second 5 ns portions of the aggregated simulations. The metric was the mean relative unsigned error in the bin occupancies:

$$err(a,b) = \frac{1}{Nocc}\sum_{Nocc} 2 \cdot \frac{|n_a - n_b|}{n_a + n_b} \qquad \text{(Eq 1)}$$

where n_a and n_b are the number of counts in equivalent bins in histograms *a* and *b*, and *Nocc* is the total number of significantly occupied bins (taken as those with a mean occupancy of > 0.01%, to avoid noise from very rarely sampled bins).

To test for relaxation in the distributions, we repeated the error calculation after trimming off increasing numbers of snapshots from the start of each simulation (always dividing the remaining dataset into two equal portions). The result (Figure 2) shows that the error measure between the sampling in the two halves of the dataset is minimal if the initial 3.1 ns of each trajectory is discarded. The degree of similarity in this situation can be appreciated visually by the comparison of the 2D free energy plots (Figure 2). For the first half of the dataset the global minimum is in bin (5, 16, 14) and has free energy 6.24 kT, while for the second half it is in bin (5, 17, 14) and has free energy 6.12 kT.

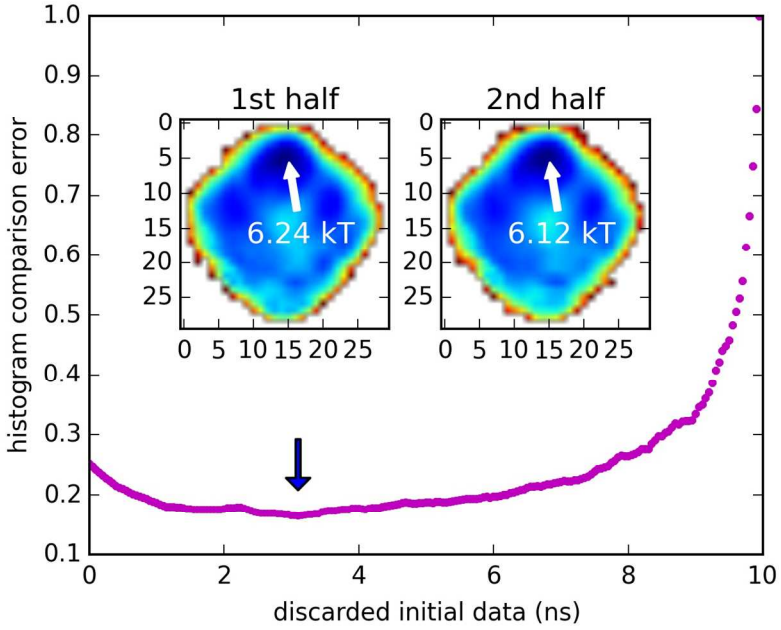

Figure 2. Analysis of relaxation and convergence. The main plot shows *err(a,b)* (see eqn 1) as a function of the amount of the initial simulations excluded from analysis. The arrow indicates the position of the minimum, for which the inset plots show the free energy surfaces for the two halves of the remaining data set, in the PC1/PC3 plane. The locations and values of the global free energy minima are shown.

Clustering is another approach to assessing convergence[25]. Using the 3D free energy surface generated from the histogram of the complete dataset (minus the initial 3.1 ns equilibration phase), watershed partitioning (implemented in pyPcazip [20]) identified seven basins of attraction and corresponding local minima (Figure 3). The major basin corresponds to the extended state; basins 2 and 3 are loose turns while basin 4 is a hydrogen-bonded beta turn. Basin 5 is the alpha-helical state while 6 is helix-like but with a distorted H-bonding pattern. The rarest state, basin 7, features bifurcated hydrogen bonds between the N-capping acetyl carbonyl group and the backbone amides of alanines 4 and 6. The basin occupancies for the data from the two halves of the simulation show high similarity.

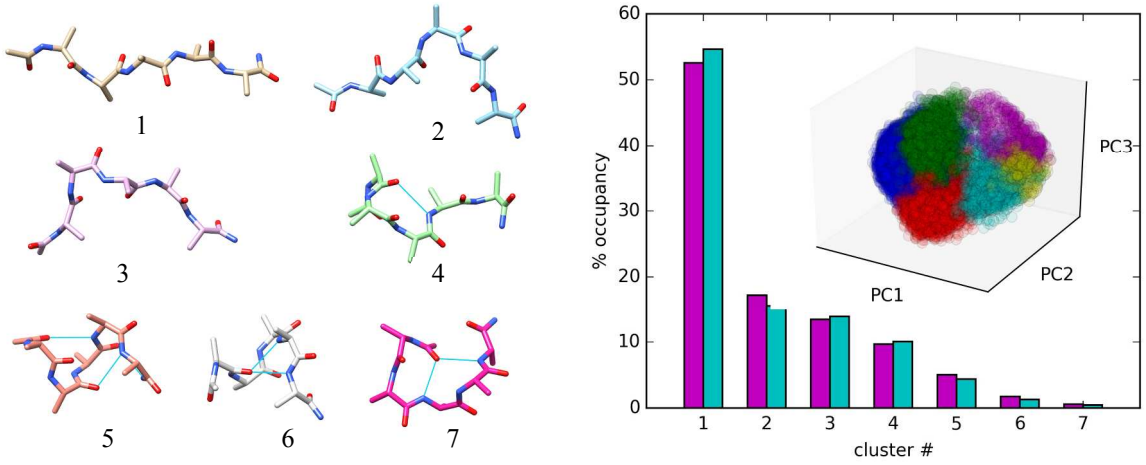

Figure 3. Left panel: examples of the seven conformational states identified by watershed partitioning. Right panel: the percentage occupancy of each basin/cluster in the two halves of the equilibrated portion of the dataset (first half: magenta, second half: cyan). The inset shows the locations of the basins in the PC subspace. The colour code is: blue, green, red, cyan, magenta, yellow, and black for basins 1-7 respectively.

Next we benchmarked the rate at which CMD simulations explored conformational space. We calculated, as a function of simulation time, how many of the bins in the 3D histogram were visited at least once. We observe that the relationship between simulation time, *t*, and volume of conformational space sampled, *Vt*, (Figure 4) can be fitted to a function of the same form as has been used previously to describe the convergence in configurational entropy with sampling time[26]:

$$Vt = Vinf - A.t^n \qquad \text{(Eq 2)}$$

where *Vinf* is the volume sampled for infinite simulation time, and *A* and *n* are fitting parameters. This model predicts that although the volume of space sampled may appear visually to be close to the limit, in fact we have about 80-90% coverage (based on parameter errors estimated from the non-linear least-squares fitting process).

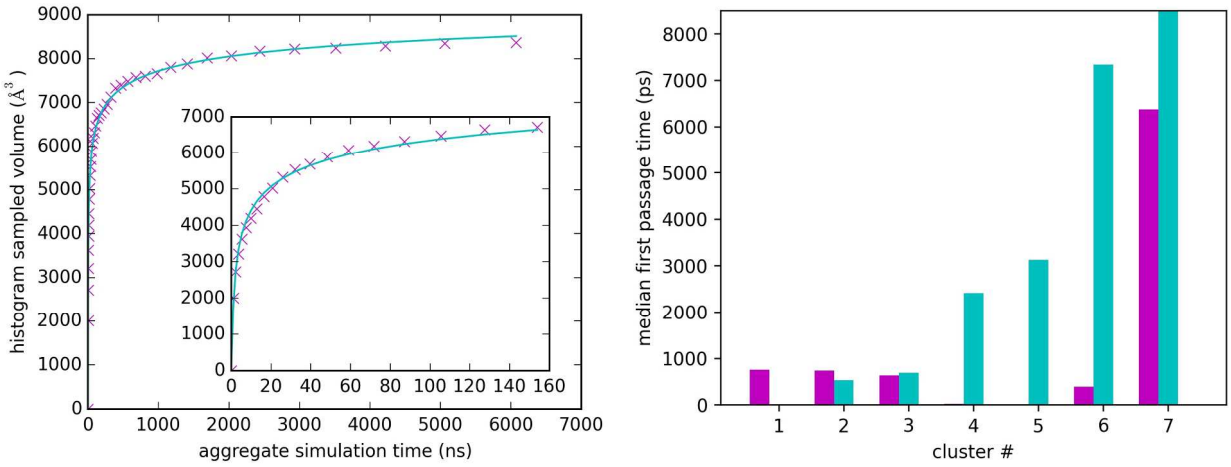

Figure 4. Left panel: conformational space sampled as a function of aggregate simulation time. The cyan curve is the fit of the data to the model in eq. 2. The inset shows just the data for the first 160 ns. Right panel: median first passage time from the helical state (cluster 5, magenta bars), and from the extended state (cluster 1, cyan bars), to all other states. Note that the median first passage time from cluster 5 to cluster 4 is very short so the magenta bar for cluster 4 is almost indiscernable.

Finally we have calculated the median first passage time for each of the two hundred independent simulations from their start conformation to each of the basins/states identified by the partitioning procedure above. As half the simulations begun from the extended state, and half from the helical state, the two sets were treated independently. We use median, rather than mean, first passage time as the metric because we observe that it is not guaranteed that every walker will sample every state over the 10ns time period (checks confirm that in general using medians rather than means has no significant impact on the analysis). Figure 4b shows that the median first passage time for simulations begun from the helical state (which corresponds to cluster 5) to all other states is of the order of 0.75 nanoseconds, with the exception of state 7 for which the

median first passage time is over 6 ns. The situation is worse for simulations begun from the extended state (cluster 1): states 5-7 have median first passage times from 3 to over 8 nanoseconds.

Performance of the CoCo-MD Method. Our first CoCo-MD dataset was generated as follows. Ten independent simulations of 0.02 ns were run (protocol as above), starting from the extended conformation, as this is the starting state from which it appears most difficult to reach some of the other states by CMD. The 200 conformations thus generated were analysed using the CoCo approach, and ten new start points generated. As discussed above, CoCo-generated structures are not guaranteed to by physically realistic, so they were not used directly as start point for the next round of simulations, rather they were used as the target structures for short restrained MD simulations (10 ps, force constant 1 $kCal.mol^{-1}$ on each heavy atom) that started from the well-equilibrated initial structure. The final structures from these restrained simulations were then used as start points for the next 0.02 ns production phase MD. The trajectory data from these simulations was then added to that from the first round of simulations, and the CoCo analysis performed again. This loop was repeated 100 times in total, to give a total of 20000 conformations representing an aggregate of 20ns of simulation time.

The performance of the CoCo-MD dataset is shown in Figure 5; for comparison, analysis of an equivalent volume of data (ten replicates of 2 ns each) from the reference CMD dataset is also shown. We see that the CoCo-MD data explores conformational space over twice as fast as the CMD dataset (Figure 5a), has sampled all seven conformational basins while CMD has sampled only five, and has a median first passage time to each cluster of < 400 ps (Figure 5b). On average, CoCo-MD decreases the median first passage time from the extended state to all other

states by a factor of more than 20. The difference between the increase in rate of general sampling (two-fold), and the increase in the rate of discovery of new conformational states (20-fold) indicates the effectiveness of this PCA-directed method.

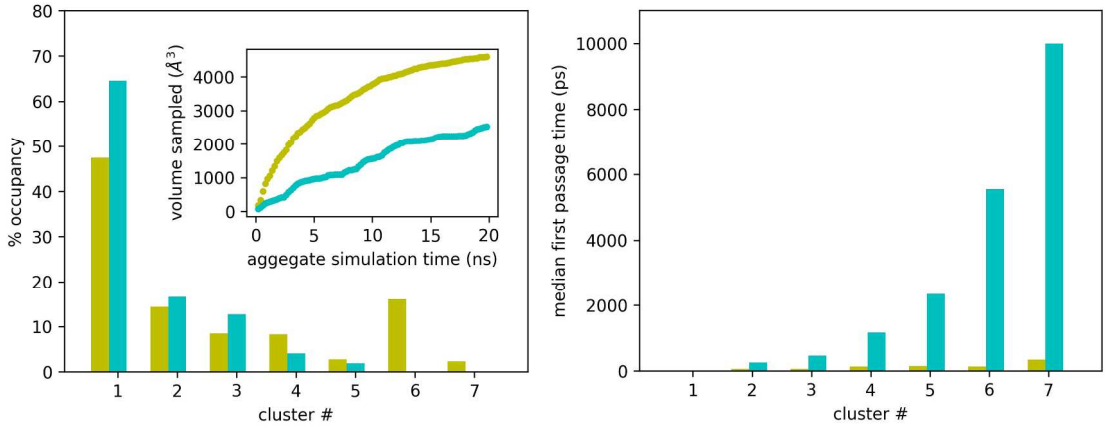

Figure 5. Performance of the CoCo-MD method (mustard) compared to that for an equivalent collection of CMD simulations (cyan), all begun from the extended state. Left panel: cluster occupancies and (inset) rates of exploration of conformational space. Right panel: median first passage times from the start cluster (1) to all others.

Impact of CoCo-MD Parameters on Performance. The initial CoCo-MD dataset was generated using a workflow that used ten independent "walkers" and interleaved ten cycles of 0.02ns MD with ten CoCo analyses of the aggregating data. We surmised that optimal performance, for a set total amount of simulation time, would come from the right balance between the number of walkers, the lengths of individual MD simulations, and the number of CoCo-based "jumps".

The first parameter we have explored is the frequency with which MD simulations are interrupted for CoCo-directed "jumps" to new, so-far unsampled, regions of space. A high jump frequency might mean many new areas are sampled, but the intervening unsupervised MD stages

would be short, and so restricted in their ability to a) explore the local region of space and b) sample outside the PC subspace that was defined by the CoCo procedure, thus making the process less adaptive.

For this evaluation we fixed the number of independent walkers to ten, as used above, and fixed the total length of each MD simulation (between CoCo-jumps) at 2 ns – also as above. We evaluated performance against two metrics. The first was the total volume of the PC1-3 subspace that was sampled by the ensemble, the second was the average of the median first passage times from the initial state (extended) into each of the other six basins. Each run was repeated three times, varying the initial randomized velocity assignments.

The results (Figure 6a) show that, by chance, our initial choice of 100 CoCo-driven jumps over 2 ns of simulation was a good one. The total volume sampled is actually rather insensitive to this parameter, but mean median first passage time bottoms out at about this value. Fifty jumps might do as well, but performance begins to degrade significantly if fewer than fifty jumps are used.

The second parameter we explored was the total number of walkers to use. A large number of walkers will permit the simultaneous exploration of many different regions of conformational space, but, if we insist the total aggregate simulation time is kept constant, the length of each MD is shorter. Again, this raises the possibility that the process becomes less adaptive, as PCs identified in early rounds of the process are less likely to be significantly challenged by the short interleaving unsupervised MD stages.

The results show however (Figure 6b) that this does not seem to be an issue. More walkers means faster median first passage times, though beyond fifty the improvement is marginal. Interestingly, the total volume of space sampled seems to decrease with the number of walkers, though the trend is not very statistically significant. We hypothesise this may result from a greater likelihood that walkers will cross each other's paths, and so resample already-visited regions of conformational space.

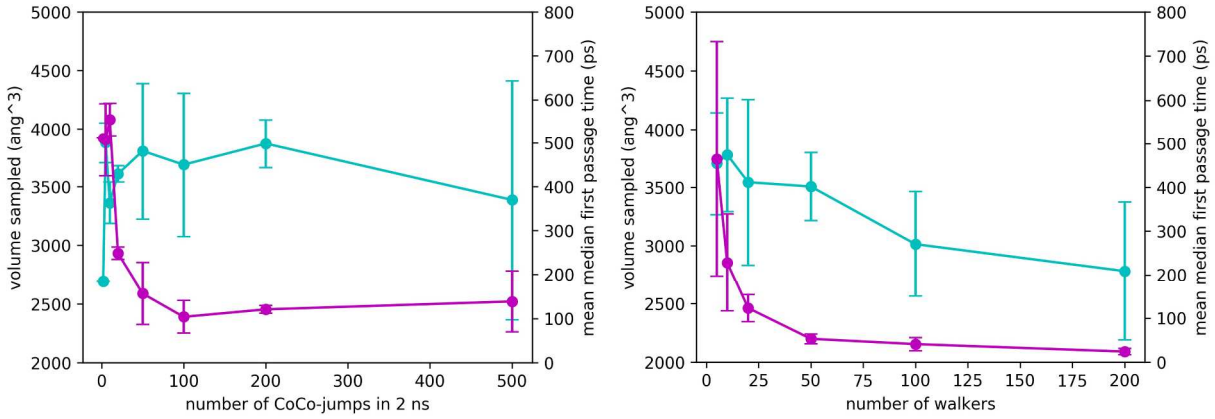

Figure 6. Performance of the CoCo-MD method as a function of selected parameters. Metrics are total volume sampled by the ensemble (cyan), and mean median first passage time (magenta) (see text for details). Error bars are standard deviations calculated from three independent experiments. Left: the frequency of CoCo-driven 'jump' steps during the MD simulations, for a constant number (ten) of independent walkers. Right: the number of walkers, for a constant number of CoCo-jumps (100) and a constant aggregate simulation time (2 ns).

Conclusions

We have found that within the context of an iterative simulation/data analysis workflow, the CoCo ensemble expansion method is a very effective approach to enhanced sampling. Applied to the classic alanine penta-peptide test-case, CoCo-MD can generate structures corresponding to all of the significantly occupied conformational states with a small fraction – 5-10% - of the computational effort required to do this using conventional MD. The method does not require any adaptations to the MD code, so this can run at its fully-optimised speed. The method is unsupervised – the user does not need to specify in advance the "interesting" reaction coordinates, these emerge and adapt automatically as the sampling progresses. The limitation of the approach is that it is not easy to see how to reconstruct a free energy surface from the ensemble. Typical trajectory reweighting methods require each simulation fragment to begin from an already-sampled region of space, but the unique feature of the CoCo method is that it does not do this. However we do not see this as a significant issue; a) there are many situations in which a rapid exploration of conformational space that reaches physically-reasonable conformations is more important than a complete knowledge of the free energy surface, b) CoCo-MD can be used as a 'pre-processor' for other approaches that can generate a free energy surface (such as the DMDMD method[12] or indeed a variety of MSM approaches [6]), and c) there are a number of emerging methods such as DHAM[27] that may be able to work with non-equilibrium trajectory ensembles of the type CoCo-MD produces anyway.

AUTHOR INFORMATION

Corresponding Author

*Charles Laughton, Centre for Biomolecular Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK. Email: Charles.laughton@nottingham.ac.uk

Present Addresses

[§]Hartree Centre, STFC Daresbury Laboratory, Scientific Computing Department, Cheshire WA4 4AD, UK.

°School of Life and Medical Sciences, Department of Pharmacy, Pharmacology and Postgraduate Medicine, University of Hertfordshire, College Lane, AL10 9AB, UK

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

ACKNOWLEDGMENT

This work is a product of the ExTASY consortium (http://www.extasy-project.org) supported by EPSRC Grant EP/K039490/1 and NSF SSI Awards (CHE-1265788 and CHE- 1265929). This work used the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk), with additional access through HECBioSim (EPSRC Grant EP/L000253/1). We acknowledge access to XSEDE computational facilities via TG-MCB090174.

REFERENCES

[1] T. Narumi, K. Yasuoka, M. Taiji, F. Zerbetto, S. Höfinger, Fast calculation of electrostatic potentials on the GPU or the ASIC MD-GRAPE-3, Comput. J. 54 (2011) 1181–1187. doi:10.1093/comjnl/bxq079.

[2] D.E. Shaw, J.C. Chao, M.P. Eastwood, J. Gagliardo, J.P. Grossman, C.R. Ho, D.J. Lerardi, I. Kolossváry, J.L. Klepeis, T. Layman, C. McLeavey, M.M. Deneroff, M. a. Moraes, R. Mueller, E.C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, S.C. Wang, R.O. Dror, J.S. Kuskin, R.H. Larson, J.K. Salmon, C. Young, B. Batson, K.J. Bowers, Anton, a special-purpose machine for molecular dynamics simulation, Commun. ACM. 51 (2008) 91. doi:10.1145/1364782.1364802.

[3] P. Minary, M.E. Tuckerman, G.J. Martyna, Long time molecular dynamics for enhanced conformational sampling in biomolecular systems, Phys. Rev. Lett. 93 (2004). doi:10.1103/PhysRevLett.93.150201.

[4] J.J. Perez, M.S. Tomas, J. Rubio-Martinez, Assessment of the Sampling Performance of Multiple-Copy Dynamics versus a Unique Trajectory, J. Chem. Inf. Model. 56 (2016) 1950–1962. doi:10.1021/acs.jcim.6b00347.

[5] A. Atzori, N.J. Bruce, K.K. Burusco, B. Wroblowski, P. Bonnet, R.A. Bryce, Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics, J. Chem. Inf. Model. 54 (2014) 2764–2775. doi:10.1021/ci5003334.

[6] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, T.R. Weikl, Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations., Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 19011–6. doi:10.1073/pnas.0905466106.

[7] Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding, Chem. Phys. Lett. 314 (1999) 141–151. doi:10.1016/S0009-2614(99)01123-9.

[8] A. Barducci, M. Bonomi, M. Parrinello, Metadynamics, Wiley Interdiscip. Rev. Comput. Mol. Sci. 1 (2011) 826–843. doi:10.1002/wcms.31.

[9] L.C.T. Pierce, R. Salomon-Ferrer, C. Augusto F. De Oliveira, J.A. McCammon, R.C. Walker, Routine access to millisecond time scale events with accelerated molecular dynamics, J. Chem. Theory Comput. 8 (2012) 2997–3002. doi:10.1021/ct300284c.

[10] A.C. Pan, T.M. Weinreich, S. Piana, D.E. Shaw, Demonstrating an Order-of-Magnitude Sampling Enhancement in Molecular Dynamics Simulations of Complex Protein Systems, J. Chem. Theory Comput. 12 (2016) 1360–1367. doi:10.1021/acs.jctc.5b00913.

[11] J. Preto, C. Clementi, Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics., Phys. Chem. Chem. Phys. 16 (2014) 19181–19191. doi:10.1039/c3cp54520b.

[12] W. Zheng, M.A. Rohrdanz, M. Maggioni, C. Clementi, Polymer reversal rate calculated via locally scaled diffusion map, J. Chem. Phys. 134 (2011). doi:10.1063/1.3575245.

[13] R. Harada, A. Kitao, Parallel cascade selection molecular dynamics (PaCS-MD) to generate conformational transition pathway, J. Chem. Phys. 139 (2013). doi:10.1063/1.4813023.

[14] R. Harada, Y. Takano, T. Baba, Y. Shigeta, Simple, yet powerful methodologies for conformational sampling of proteins, Phys. Chem. Chem. Phys. 17 (2015) 6155–6173. doi:10.1039/C4CP05262E.

[15] J. Peng, Z. Zhang, Simulating large-scale conformational changes of proteins by accelerating collective motions obtained from principal component analysis, J. Chem. Theory Comput. 10 (2014) 3449–3458. doi:10.1021/ct5000988.

[16] C.A. Laughton, M. Orozco, W. Vranken, COCO: A simple tool to enrich the representation of conformational variability in NMR structures, Proteins Struct. Funct. Bioinforma. 75 (2009) 206–216. doi:10.1002/prot.22235.

[17] C.J. Margulis, C.J. Margulis, H. a. Stern, H. a. Stern, B.J. Berne, B.J. Berne, Helix Unfolding and Intramolecular Hydrogen Bond Dynamics in Small α-Helices in Explicit Solvent, J. Phys. Chem. B. 106 (2002) 10748–10752. doi:10.1021/jp0205158.

[18] W.A. Hegefeld, S.E. Chen, K.Y. Deleon, K. Kuczera, G.S. Jas, Helix formation in a pentapeptide: Experiment and force-field dependent dynamics, J. Phys. Chem. A. 114 (2010) 12391–12402. doi:10.1021/jp102612d.

[19] D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W.. Götz, F. Kolossváry, I. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J.. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, P.. Kollman, AMBER 12, (2012).

[20] A. Shkurti, R. Goni, P. Andrio, E. Breitmoser, I. Bethune, M. Orozco, C.A. Laughton, pyPcazip: A PCA-based toolkit for compression and analysis of molecular simulation data, SoftwareX. 5 (2016) 44–50. doi:10.1016/j.softx.2016.04.002.

[21] V. Balasubramanian, I. Bethune, A. Shkurti, E. Breitmoser, E. Hruska, C. Clementi, C.A. Laughton, S. Jha, ExTASY: Scalable and Flexible Coupling of MD Simulations and Advanced Sampling Techniques, 2016. arXiv:1606.00093.

[22] V. Balasubramanian, A. Treikalis, O. Weidner, S. Jha, Ensemble Toolkit: Scalable and Flexible Execution of Ensembles of Tasks, in: Proc. Int. Conf. Parallel Process., 2016. doi:10.1109/ICPP.2016.59.

[23] L. Sawle, K. Ghosh, Convergence of Molecular Dynamics Simulation of Protein Native States: Feasibility vs Self-Consistency Dilemma, J. Chem. Theory Comput. 12 (2016) 861–869. doi:10.1021/acs.jctc.5b00999.

[24] I.L. Dryden, K.V. Mardia, Statistical Shape Analysis, Stat. Med. 19 (2000) 2716–2717. doi:10.1002/1097-0258(20001015)19:19<2716::AID-SIM590>3.0.CO;2-O.

[25] E. Lyman, D.M. Zuckerman, Ensemble-based convergence analysis of biomolecular trajectories., Biophys. J. 91 (2006) 164–72. doi:10.1529/biophysj.106.082941.

[26] S.A. Harris, C.A. Laughton, A simple physical description of DNA dynamics: quasi-harmonic analysis as a route to the configurational entropy, J. Phys. Condens. Matter. 19 (2007) 76103. doi:10.1088/0953-8984/19/7/076103.

[27] E. Rosta, G. Hummer, Free energies from dynamic weighted histogram analysis using unbiased Markov state model, J. Chem. Theory Comput. 11 (2015) 276–285. doi:10.1021/ct500719p.

for Table of Contents use only

CoCo-MD: A Simple and Effective Method for the Enhanced Sampling of Conformational Space

Ardita Shkurti[†§], Ioanna Danai Styliari[°†], Vivek Balasubramanian[‡], Iain Bethune[#], Shantenu Jha[‡], Charles A. Laughton[†*]