JCTC Journal of Chemical Theory and Computation
Letter
pubs.acs.org/JCTC

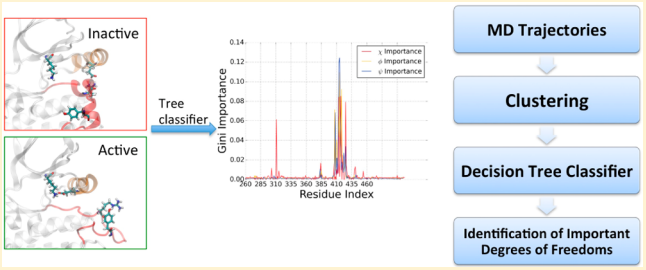
Automatic Selection of Order Parameters in the Analysis of Large Scale Molecular Dynamics Simulations

Mohammad M. Sultan,[†] Gert Kiss,[†,‡,§] Diwakar Shukla,[†,‡] and Vijay S. Pande*[,†,‡,§]

[†]Department of Chemistry, Stanford University, 318 Campus Drive, Stanford, California 94305, United States
[‡]SIMBIOS NIH Center for Biomedical Computation and [§]Center for Molecular Analysis and Design, Stanford University, Stanford, California 94305, United States

Supporting Information

ABSTRACT: Given the large number of crystal structures and NMR ensembles that have been solved to date, classical molecular dynamics (MD) simulations have become powerful tools in the atomistic study of the kinetics and thermodynamics of biomolecular systems on ever increasing time scales. By virtue of the high-dimensional conformational state space that is explored, the interpretation of large-scale simulations faces difficulties not unlike those in the big data community. We address this challenge by introducing a method called clustering based feature selection (CB-FS) that employs a posterior analysis approach. It combines supervised machine learning (SML) and feature selection with Markov state models to automatically identify the relevant degrees of freedom that separate conformational states. We highlight the utility of the method in the evaluation of large-scale simulations and show that it can be used for the rapid and automated identification of relevant order parameters involved in the functional transitions of two exemplary cell-signaling proteins central to human disease states.



INTRODUCTION

Modern molecular dynamics (MD) simulations have matured to a point at which simulations of many complex biological systems can be carried out routinely. Recent performance boosts from software and hardware developments have further added to the adaptability of biophysical simulations and have pushed the computational study of protein dynamics into the micro- and millisecond regime.[1,2] As the resulting trajectories approach the terabyte scale, conventional analysis techniques tend to encounter a sustainability limit and are faced with challenges typical for big data: what is the information content and how can we organize it.

Markov state models (MSMs)[3−6] are kinetic representations of complex dynamical systems and have been used to study MD trajectories. They partition the accessible protein conformational landscape by first finely discretizing the data into microstates, which can then be lumped together to form macrostates.[3,7] MSMs can be used to gain holistic insight into conformational preferences of protein dynamics via the identification of metastable intermediates. Transition path theory[8−10](TPT) can subsequently be applied to identify pathways that connect various regions of the phase space.[11] MSMs have been shown to be useful for exploring and understanding of underlying dynamics in protein folding and conformational change.[3,4,12]

While, MSMs can used to locate metastable conformational states and TPT can be employed to find and assess the probabilistic paths that connect them, these methods do not provide an atomistic level of detail into the specifics that set the states apart. A basic challenge in this process is choosing a low-dimensional projection that best captures experimentally determined properties of the biomolecular system under investigation. How do we identify the important degrees of freedom in the clustered simulation data? Which measurements are relevant and can be used to elucidate the functional dynamics? Typically the decision is based on chemical intuition or on prior knowledge about the system at hand, which provides a good starting point for the data analysis. While useful for proteins that are well studied, the approach grows increasingly biased as the prior information content decreases. When experimental observables cannot be translated directly into computational measurements or when there is no obvious way to do so, the challenge of discerning the signal from the noise can become rate limiting or even prohibitive in the discovery process.

Here we show that feature selection algorithms coupled with reaction coordinate identification methods[13,14] and techniques from supervised machine learning can be used to interpret clusters of MD trajectories by finding the relevant degrees of freedom that separate these clusters. Consider a protein that possesses j metastable states, which are defined by m degrees of freedom. Supervised machine learning (SML) algorithms can pick out k critical features (where $k < m$) that can distinguish between the j states. The features are quantifiable geometric properties—such as dihedral angles and distances—within

Received: April 23, 2014
Published: October 22, 2014

   dx.doi.org/10.1021/ct500353m | J. Chem. Theory Comput. 2014, 10, 5217−5223

individual MD frames. We hypothesize that an SML algorithm capable of drawing a decision boundary to distinguish human faces[15] can also be used to differentiate between active, intermediate, and inactive protein conformations at an atomistic level. We show that the Gini importance criterion[16] used in the construction of decision tree (DT) and random forest (RF)[17] classifiers can also be employed in the search for degrees of freedom that correlate most strongly with the assignment of a conformation to a particular MSM macrostate.

Here, we utilize the CB-FS approach for the analysis of MSMs. It can, however, be trivially expanded to the interpretation of results from any high-dimensional clustering algorithms, such as K-means or K-medoids. We chose to analyze data that was clustered into MSMs because of its biophysical relevance and the link it provides between theory and experiment. To that end, we applied CB-FS to identify the key degrees of freedom that are involved in the conformational transitions of the signal hub-protein ubiquitin and those of the regulatory protein Src kinase.

METHODS

Supervised machine learning techniques have been effectively employed for tasks that include spam filtering, optical character recognition, search engines, and computer vision—all of which require the assignment of unseen data to an output class. Typically, an SML algorithm is tasked to find a high-dimensional decision boundary between training examples in order to predict an output. Similarly, we use decision trees (DTs) to predict the Markov state for protein conformations represented in a high-dimensional vector space. We choose DTs over other classification methods primarily because of their natural translation to biology. Since the early days of the Monod-Wyman-Changeux (MWF)[18] and Koshland-Némethy-Filmer (KNF)[19] model of allostery, proteins have been viewed as adopting distinct active and inactive conformations, many of which vary by no more than a few degrees of freedom. MSMs allow us to locate these different conformations, while DTs can use the information content of each state to identify the characteristic degrees of freedom that set them apart.

Throughout the Methods section, we adhere the following convention: D = data set of vectorized conformations, X = conformation, c_m = Markov state assigned to conformation, θ = feature/degree of freedom under investigation, x_j = value of j^{th} feature of the x^{th} conformation, $\{t\}$= set of threshold values for each feature, τ = nodes of trees, Δ = normalized gain, and $I(:)$ = Gini impurity of a node.

Markov State Models. Markov state models are kinetic models of complex, dynamic systems. They partition the conformational landscape of proteins into discrete states where the transitions between states are considered as a memory less jump process (Markovian). This is equivalent to representing the system as an $N*N$ transition matrix where N is the number of states in the model. The eigenvectors of the transition matrix correspond to dynamic events that—when added together—can describe complex events such as folding pathways or functional activation sequences. For more information about the analysis of protein dynamics we direct the interested reader to recent reviews of this topic.[3,20]

Decision Tree (DT) Classifiers. Tree classifiers are hierarchical structures that model the output class as indicator functions over a restricted set of values of the input variables. An indicator function, defined over a set R, designates membership to that set. DTs can be represented as binary trees where the nodes correspond to a decision criterion and the leaves represent an output variable. Applied to biomolecular simulation data, we can determine if a protein is in different Markov states by querying the values of its features at every node. DTs are capable of modeling the following class of functions

$$C(x) = \sum_{m=1}^{M} c_m I(x \in R_{\text{m}}) \tag{1}$$

where $\{R_m\}_1^M$ is the set of *disjointed* subregions of the input features. $\{c_m\}_1^M$ is the set of output classes (Markov states), and $I(x \in R_m)$ is the indicator function. The indicator function denotes membership of conformation x to the m^{th} region of phase space. The function assigns a particular Markov state (c_m) to a conformation (x) if x is in a certain region of phase space (R_m). The high-dimensional partitioning is learned by the DTs based on a labeled training set that it has previously seen. We can expand this indicator function into a product of "m" individual features as shown in eq 2

$$C(x) = \sum_{m=1}^{M} c_m \prod_{j=1}^{m} I(x_j \in s_{jm}) \tag{2}$$

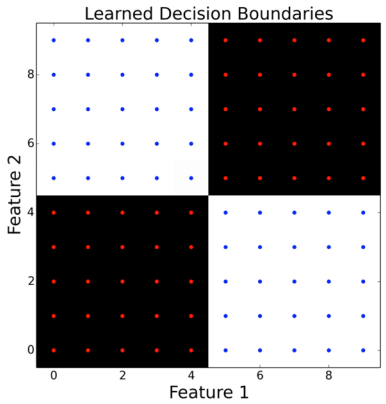
where $I(x_j \in s_{jm})$ is the indicator function. The value of feature j of the conformation x belongs to the set of values that the j^{th} feature has in the m^{th} region (s_{jm}). This expansion highlights the natural link that DTs have to the description of biological systems. Structural and computational studies[4,12,21] have shown that residue side chains adopt distinct orientations characterized by specific sets of contacts, dihedrals, etc., that can be depicted as metastable conformational wells. We propose that such multistate landscapes can be captured using the piecewise indicator functions shown in eq 2. Compiling our data in the form of vector representations allows us to include virtually any degree of freedom such as backbone, side chain dihedrals, contact maps, and hydrogen bond (H-bond) networks.

The parameters from eq 1 can be optimized by employing a recursive greedy algorithm.[17,22] This works by dividing the data along a feature that can best separate the output states. For instance, if the data at any arbitrary node "τ" is represented by "D", then we can divide it into two subsets at each possible threshold "t" for every possible feature "θ".

$$D_{\text{left}}(\theta, t) = D|\theta \le t \tag{3}$$

$$D_{\text{right}}(\theta, t) = D \backslash D_{\text{left}}$$

The threshold "t" depends on the type of feature θ and can be obtained through an exhaustive search over all possible midpoint values in the training data, if the feature is confined to linear positive values. Periodic variables like dihedrals can be represented as features in positive real space and divided through multiple single boundary divisions to account for their periodicity. The resulting sequential divisions are equivalent to using two boundaries at once. The curious reader is encouraged to review the Supporting Information for an algorithm that converts an arbitrary tree into a binary tree. To select the best possible split, we calculate the impurity at node "τ" and at all of its children. The normalized gain Δ[16,22] can be used as criterion to choose the best split for the data at every node



$$MI(F,S) = H(F) + H(S) - H(F,S)$$

$$r = \frac{\sum_{i=1}^{N}(F_i - \bar{F})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{N}(F_i - \bar{F})^2}\sqrt{\sum_{i=1}^{N}(S_i - \bar{S})^2}}$$

Feature	Mutual Information (MI[F,S])	Pearson Correlation (r[F,S])
Feature 1	0.0	0.0
Feature 2	0.0	0.0

Figure 1. Toy example that highlights the advantage of using a decision tree over correlation metrics (mutual information (MI) and Pearson correlation (r)). The image on the left depicts a two state model (red and blue dots) with the learned decision boundaries in black and white. To the right are the equations that define the two comparison metrics. $H(F)$ is the entropy of a given feature, $H(S)$ is the state entropy, and $H(F,S)$ is the joint entropy. N, $\bar{F}$, and $\bar{S}$ represent the total number of examples, the mean value of the feature, and the mean value of the state, respectively. The table lists the values obtained from both the Pearson correlation and the mutual information on each of those features with the corresponding state. The details of these calculations are given in the Supporting Information.

$$\Delta(\theta, t) = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)*I(v_j)}{N} \tag{4}$$

where $I(:)$ is the impurity (statistical heterogeneity) criterion, k is the number of child nodes, $N(v_j)$ is the number of data points with child node v_j, and N is the total number of points. We then select the parameters (θ,t) that maximizes the normalized gain.

$$(\theta, t)* = \text{argmax}_{(\theta,t)}\Delta(\theta, t) \tag{5}$$

This procedure is repeated recursively until either all the data has been divided into completely pure samples or only a single sample is left at the leaf node. Alternatively, the process can be terminated upon satisfaction of a threshold criterion. The information gain in eq 4 is equivalent to an entropy reduction in the target state given the potential split and is comparable to the mutual information (MI) between the target variable and the feature under investigation. In contrast to MI calculations and due to a more complex objective function (eq 1), DTs are capable of modeling additive effects (Figure 1 and the Supporting Information).

Equation 6 defines the Gini impurity criterion

$$I(\tau) = 1 - \sum_{i=1}^{c} [p(i|\tau)]^2 \tag{6}$$

where τ is the node under consideration, and c is the number of states. The Gini impurity is a computationally efficient approximation to the Shannon information entropy. Similar to the latter, the Gini impurity of a pure sample is 0 and is maximized when the sample is uniformly heterogeneous. Since DTs built with either criterion give consistent results,[22] we used the Gini impurity. Intuitively speaking, DTs greedily and recursively divide the data in an attempt to make it more pure. The gain in information—when summed and normalized over the entire tree for individual features θ over all the nodes τ where it was employed—allows us to calculate the significance of each feature, better known as the Gini importance

$$\text{Gini importance}(\theta) = \sum_{\tau} \Delta(\theta)_\tau \tag{7}$$

where θ is the feature under investigation.

We propose using DTs over other classification methods, such as regularized kernel support vector machines or logistic regression, due to the interpretability of the final model, its ability to ignore correlated variables, and because it can handle both continuous and binary data.

Choosing or pruning the right sized DT can be an involved process. The complexity of the tree is proportional to its depth where smaller trees tend to have a large bias and large trees are typically characterized by a large variance. The simplest way to choose the right sized tree is via test set estimates on the maximum tree depth. In short, we divide our data set into training and testing sets (ranging from a 50−50 split to 80−20 splits). We build iteratively more complex trees by increasing the maximum allowable depth. At each iteration, we use the learned tree to predict the output states for the testing set and pick the model that gives the lowest error. An alternative approach is one in which complexity parameters[17] with multifold cross-validation can be used. In the context of analyzing the output of clustering algorithms, however, this is not necessary.

One of the most common extensions to decision trees is the concept of random forests (RF).[17] Here, multiple trees are generated from a random subset of the data, after which each tree votes over the outcome of new unseen data. Due to the greedy search strategy employed in parametrizing eq 1, DTs can potentially produce very different sequences of cuts upon even small perturbations in the data. RFs reduce this variance by letting individual trees survey a bootstrapped sample of the original data set. RFs are trivially parallelizable, can handle large amounts of data quite easily, and can provide upper and lower bounds on the importance of individual features.

Recent work by Schwantes et al.[23] and Perez-Hernandaz et al.[24] have used time-structure independent components analysis (tICA) to find the slowest decorrelating degrees of freedom in the automatic construction of MSMs. In principle, we can perform similar order parameter selection by looking at the degrees of freedom that are the most correlated with the eigenvectors of the transition matrix or the independent components. Analyses of this nature are limited to the dynamic eigenvectors of the Markov models, and it is not immediately clear how this approach can be extended to arbitrary combinations of states. Clustering based feature selection has

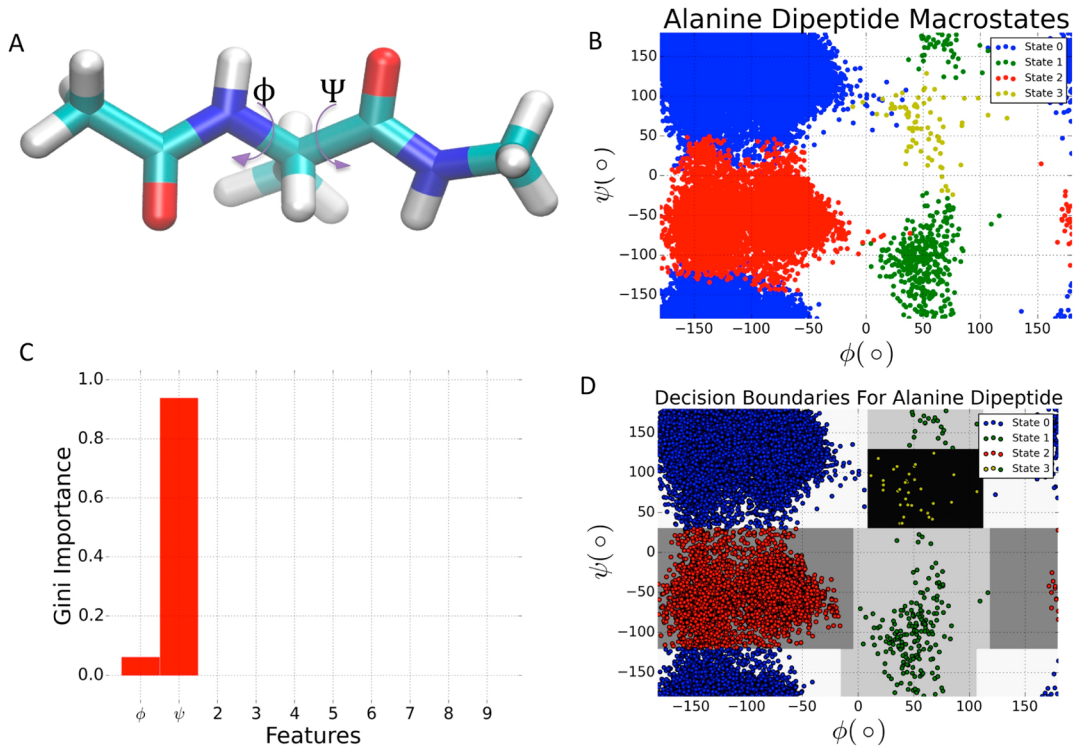


Figure 2. A) Terminally capped alanine dipeptide with the two central dihedrals marked. B) ϕ vs ψ scatter plot of the Markov macrostate assignment, based on the heavy atom RMSD as the clustering criterion. C) The Gini importance of the features with the two dihedrals marked. Features 2 to 9 are Gaussian and uniform in noise. D). Recreation of plot B) by training a single classifier on the system dihedrals and employing the classifier to predict the Markov states of an unseen test data set. The contour lines going from white to black represent the decision boundaries learned on the data set. The plot contains half the data points of B since the first half was used to train the model. Protein images were generated using Visual Molecular Dynamics (VMD).[27] The code used to generate these plots can be found online at https://github.com/msultan/alanine_tree.

the ability to work outside the Markov framework with arbitrary clustering methodologies. Insights about the partitioning can be drawn from the tree structure itself (Figure 1 and Figure 2d). Moreover, calculating the correlation of a single feature against dominant eigenvectors or independent components would suffer from its inability to model higher order additive effects—similar to calculating a single mutual information data point. The toy example in Figure 1 highlights the difficulties that arise from a symmetric interaction effect. A DT can readily model the additive affects to find the relevant features in the data. Projecting the resulting decision boundaries back onto the original space gives the state boundaries one would intuitively expect. Correlation based metrics such as mutual information (MI) and Pearson correlation (r) are incapable of modeling this effect and incorrectly show no association between either of the features and the output state. The details of how to calculate the mutual information and Pearson correlation of a target feature against the output state are given in the SI.

Feature selection methods are not without limitations. Most ignore correlated variables in the data, which is often considered a strength. This can, however, be a concern in the context of biological systems. The features that best explain the data here are not necessarily biologically relevant. In practice, such a situation might never arise given enough sampling, through use of an ensemble classifier, and from sparsity in the underlying model. However, we recommend the use of this method to generate starting points in the interpretation of clustered MD trajectories rather than as a solution.

Feature Vector Generation. DTs and other classification methods require a vector representation of the individual protein conformations. These features can be a subset or an exhaustive combination of several metrics like backbone dihedrals, α-carbon distances, hydrogen bond networks, or the root mean squared deviation (RMSD) of secondary structure elements. The choice of the features depends on our prior knowledge about the system.

All the DTs and RFs used here were built with the Scikit-learn library[25] in Python. The Markov models and vector representations of MD data were generated using the MDTraj library in the MSM builder software.[26]

RESULTS AND DISCUSSION

Alanine Dipeptide Model. As a proof of concept, we applied DT classifier to identify the important degrees of freedom of the alanine dipeptide model system. A four state macrostate model was built from the backbone RMSDs of the heavy atoms (Figure 2b). The trajectories were then vectorized using the Φ and Ψ backbone dihedrals (relevant features), along with eight normally distributed noisy features. This was done to quantify the degree to which the noise contributes to the analysis. Using the macrostate labels as a target variable, a single decision tree classifier was trained on 25,048 conformations that represent 50% of each state within our data set. The maximum depth of the decision tree classifier was increased until the test error on the previously unseen 50% data set was less than 1% (max depth = 4). The learned boundaries and the Gini importance of the input features was then

dx.doi.org/10.1021/ct500353m | J. Chem. Theory Comput. 2014, 10, 5217−5223

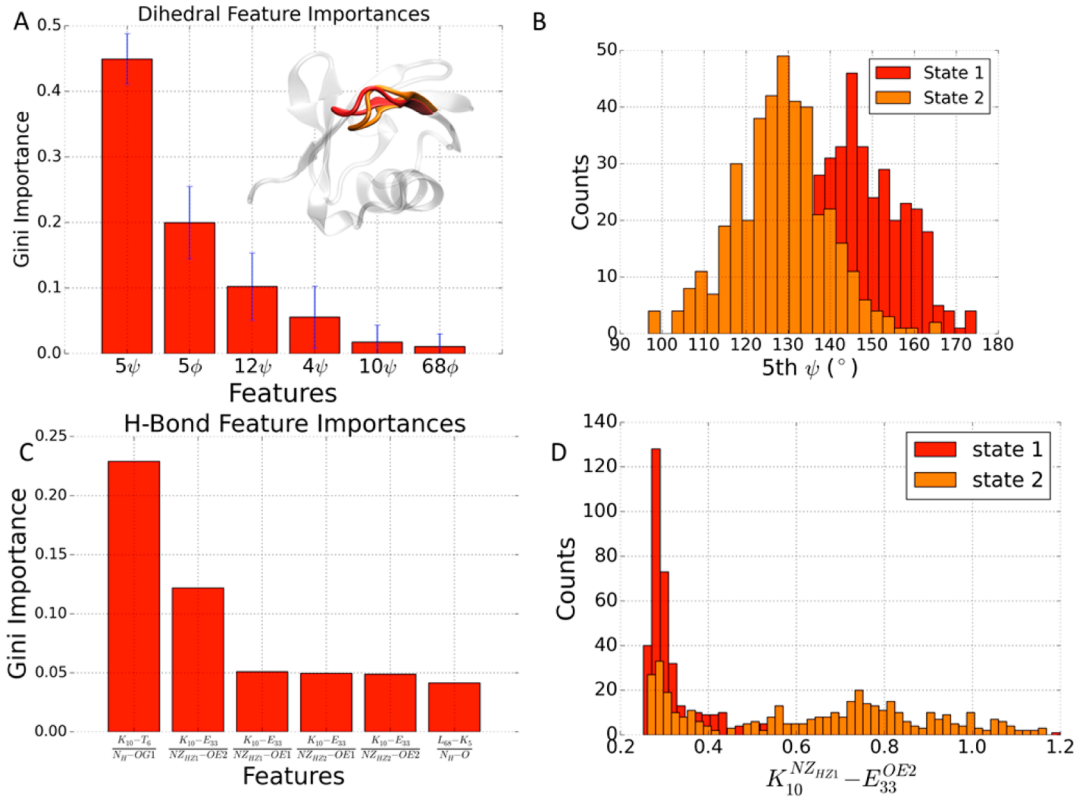


Figure 3. A) Results from building a random forest classifier on a two state hidden Markov model of human ubiquitin. The fifth and 12th dihedrals correspond to up and down conformations of the loop and the error bars are from the different DTs in the ensemble. B) Two state behavior of fifth ψ dihedral in the two states. C) The two H-bonds that stabilize the loop in the up (red) state. D) Histogram showing the length of the H-Bond between Glu33 and Lys10.

computed using this model (Figure 2c and Figure 2d). The plots highlight the ability of the Gini importance to find the key backbone dihedrals without assigning significance to the noise variables. The Supporting Information contains Web links for the data set and Python scripts that were used to generate the images shown in Figure 2.

Ubiquitin. We tested the performance of CB-FS in the context of a biological system and analyzed the dynamics of human ubiquitin—a signaling hub protein that connects multiple cellular pathways.[28,29] Its misregulation has been implicated in numerous pathologies, including neurodegeneration and tumor progression. A two state model was generated from an aggregate 100 μs of simulation data using a hidden Markov model formalism.[30] It discerns two distinct conformations of a functionally selective loop (Figure 3a inset) and provides us with insights into the degrees of freedom that correspond to this conformational change. 800 structures were randomly pulled from the two states and further analyzed with CB-FS. Two different vectorized representations, dihedral angles and hydrogen bond networks, were used to break down the states. Two random forest classifiers with 40 trees each and a maximum depth of 4 for the dihedrals features and a maximum depth of 7 for hydrogen bond networks were trained. The results are shown in Figure 3. The H-bond random forest revealed two important interactions. A backbone hydrogen bond between K10 and T6 breaks as the system switches to state 2 (orange). The H-bond network (Figure 3c) also revealed the functionally important interaction between the side chains of K10 and E33. The finding is in line with previous work that experimentally validated the significance of the K10-E33 contact.[29] The mutation of K10 into a neutral residue gives a markedly increased pK_a of E33. Further work by Wickliffe et al.[31] and Bremm et al.[32] showed that this noncovalent interaction is important for orienting the K10 in a position suitable for selection by the Ube2s enzyme via substrate-assisted catalysis.

Src Kinase. Kinases are a family of proteins responsible for catalyzing the transfer of the gamma phosphate group of ATP to a target substrate. They are key regulators of cell signaling and are therapeutic targets for a wide spectrum of diseases. The Src tyrosine kinase is involved in the cellular signaling pathways associated with cell proliferation. Its signaling has been implicated in uncontrolled cancerous growth.[33] Understanding the atomic level interactions involved in the activation pathways of Src and other kinases can potentially help in the design of better and more selective drugs, which has given rise to significant research activity over the past decade.[12,34,35]

Recently we generated a four state model for Src that was built using a combined 500 μs of sampling on the Folding@home distributed computing platform.[36] The simulation data was extensively analyzed in recent work,[12] and serves as a large test case for the CB-FS analysis of the Src kinase activation pathway.[37] An activation trajectory consisting of 20,000 frames was generated from the Markov model. Using the dihedral degrees of freedom as the feature space and the four MSM macrostates as the target variable, feature selection was performed using a random forest classifier comprising of 20 DTs.

The key results and their biological interpretation are summarized in Figure 4. The Src kinase activation is a

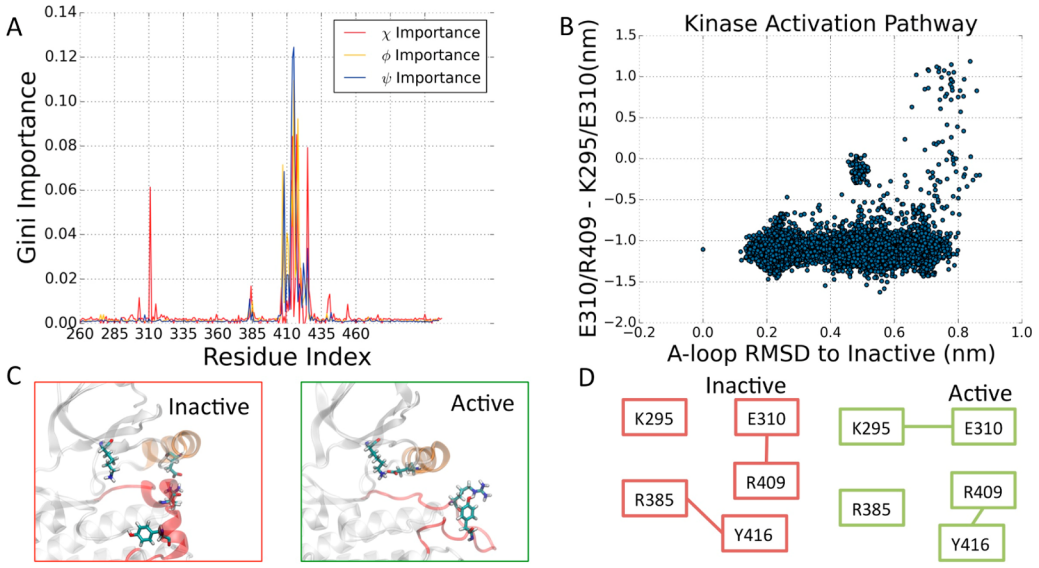


Figure 4. A) Gini importance of different dihedrals showing importance of the K295, E310, H384, and the A-loop (404−424). B) The c-Src activation pathways projected onto a two-dimensional reaction coordinate showing the sequence of events that needs to happen for the system to activate. Note how the A-loop (red) needs to first unfold, followed by the rotation of the C-Helix (orange). C) On an atomistic scale, the A-loop (red) unfolds followed by the twisting motion of the E310 to interact with the K295. D) The salt bridge switching mechanism involved in this system. For more details see the works of Shukla et al.[12] and Meng et al.[34,35]

sequential two step process in which the activation loop (A-loop, shown in red in Figure 4c) unfolds before the C-helix (shown in orange in Figure 4c) which then swings inward toward the core of the protein to form a critical Glu-Lys ion pair. Projecting the 20,000 frames onto these two degrees of freedom shows this two step process in Figure 4b. CB-FS selected the key dihedrals that are involved in this sequential activation mechanism (Figure 4a). For example, residues 410−420 are part of the activation loop (A-loop, shown in red in Figure 4c) that needs to unfold for the system to activate (Figure 4b). The CB-FS method also highlighted the importance of H384 that forms a part of the regulatory spine critical for catalysis and E310 which switches from interacting with R409 in the inactive state to K295 in the active state (Figure 4b-d).[12,35]

CONCLUSIONS

We present a clustering based feature selection approach for the analysis and interpretation of large scale simulations of biological systems. An information gain criterion is used to build decision trees and can also be employed to find important contacts, hydrogen bonds, salt bridges, etc. that uniquely identify the functional states of proteins. Since feature selection methodologies can ignore certain degrees of freedom when they correlate with other variables, it is desirable to use random forest classifiers, which can be built from bootstrapped samples of the training data. Human ubiquitin and Src kinase served as test cases and demonstrate the scalability of CB-FS, its ability to break down the steps within activation pathways, and the potential to provide key targets for mutation studies and protein engineering. Our work presents a step forward in the unbiased analysis of Markov models and related clustering methodologies of large scale molecular dynamics trajectories. Future work will focus on extensions to the metric space and on application of the method to systems with increasingly complex functional dynamics.

ASSOCIATED CONTENT

Supporting Information

Characterization of the symmetric interaction model, mathematical definitions and methods to calculate mutual information and Pearson correlation, links to data sets and scripts used to generate the alanine dipeptide landscape, and a description of an algorithm to convert any arbitrary tree into a binary tree. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pande@stanford.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

M.M.S. would like to acknowledge NSF-MCB-0954714 for funding. The authors would also like to thank Robert McGibbon and Bharat Ramsundar for their comments on the manuscript. D.S. acknowledges support from the Stanford School of Medicine fellowship made possible by a donation from the Li Ka Shing foundation. G.K. and V.S.P. acknowledge support from the Simbios NIH National Center on Biocomputing through the NIH Roadmap for Medical Research Grant U54 GM07297. G.K. and V.S.P. acknowledge support from a CMAD postdoctoral fellowship. The authors would also like to thank the two reviewers for helpful comments on the manuscript. The Src kinase database is now freely available through the Stanford digital repository at http://purl.stanford.edu/cm993jk8755.

REFERENCES

(1) Lane, T.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. To Milliseconds and beyond: Challenges in the Simulation of Protein Folding. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58−65.

(2) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41*, 429−452.
(3) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52*, 99−105.
(4) Kohlhoff, K. J.; Shukla, D.; Lawrenz, M.; Bowman, G. R.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. Cloud-Based Simulations on Google Exacycle Reveal Ligand Modulation of GPCR Activation Pathways. *Nat. Chem.* **2014**, *6*, 15−21.
(5) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. Markov State Models Based on Milestoning. *J. Chem. Phys.* **2011**, *134*, 204105.
(6) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States. *J. Chem. Phys.* **2007**, *126*, 155102.
(7) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134*, 174105.
(8) E, W.; Vanden-Eijnden, E. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *Annu. Rev. Phys. Chem.* **2010**, *61*, 391−420.
(9) E, W.; Vanden-Eijnden, E. Towards a Theory of Transition Paths. *J. Stat. Phys.* **2006**, *123*, 503−523.
(10) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the Equilibrium Ensemble of Folding Pathways from Short off-Equilibrium Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011−19016.
(11) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **2009**, *7*, 1192−1219.
(12) Shukla, D.; Meng, Y.; Roux, B.; Pande, V. S. Activation Pathway of Src Kinase Reveals Intermediate States as Targets for Drug Design. *Nat. Commun.* **2014**, *5*, 3397.
(13) Best, R. B.; Hummer, G. Reaction Coordinates and Rates from Transition Paths. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6732−6737.
(14) Ma, A.; Dinner, A. R. Automatic Method for Identifying Reaction Coordinates in Complex Systems. *J. Phys. Chem. B* **2005**, *109*, 6769−6779.
(15) Moghaddam, B. Learning Gender with Support Faces. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 707−711.
(16) Menze, B. H.; Kelm, B. M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F. A. A Comparison of Random Forest and Its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data. *BMC Bioinf.* **2009**, *10*, 213.
(17) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(18) Changeux, J.-P. Allostery and the Monod-Wyman-Changeux Model after 50 Years. *Annu. Rev. Biophys.* **2012**, *41*, 103−133.
(19) Changeux, J.-P.; Edelstein, S. J. Allosteric Mechanisms of Signal Transduction. *Science* **2005**, *308*, 1424−1428.
(20) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134*, 174105.
(21) Williams, J. C.; Wierenga, R. K.; Saraste, M. Insights into Src Kinase Functions: Structural Comparisons. *Trends Biochem. Sci.* **1998**, *23*, 179−184.
(22) Tan, P.-N. Classification: Basic Concepts, Decision Trees and Model Evaluation. *Introd. Data Min.* **2006**, 145−205.
(23) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.
(24) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139*, 015102.
(25) Pedregosa, F.; Varoquaux, G. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn.* **2011**, *12*, 2825−2830.
(26) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.
(27) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14* (33−38), 27−28.
(28) Lecker, S. H.; Goldberg, A. L.; Mitch, W. E. Protein Degradation by the Ubiquitin-Proteasome Pathway in Normal and Disease States. *J. Am. Soc. Nephrol.* **2006**, *17*, 1807−1819.
(29) Sundd, M.; Iverson, N.; Ibarra-Molero, B.; Sanchez-Ruiz, J. M.; Robertson, A. D. Electrostatic Interactions in Ubiquitin: Stabilization of Carboxylates by Lysine Amino Groups. *Biochemistry* **2002**, *41*, 7586−7596.
(30) Mcgibbon, R. T.; Ramsundar, B.; Kiss, G.; Sultan, M. M.; Pande, V. S. Understanding Protein Dynamics with L 1-Regularized Reversible Hidden Markov Models. *J. Mach. Learn. Res.* **2014**, *32*, 1197−1205.
(31) Wickliffe, K.; Lorenz, S.; Wemmer, D.; Kuriyan, J.; Rape, M. The Mechanism of Linkage-Specific Ubiquitin Chain Elongation by a Single-Subunit E2. *Cell* **2011**, *144*, 769−781.
(32) Bremm, A.; Komander, D. Emerging Roles for Lys11-Linked Polyubiquitin in Cellular Regulation. *Trends Biochem. Sci.* **2011**, *36*, 355−363.
(33) Hantschel, O.; Superti-Furga, G. Regulation of the c-Abl and Bcr-Abl Tyrosine Kinases. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 33−44.
(34) Ozkirimli, E.; Yadav, S.; Miller, W.; Post, C. An Electrostatic Network and Long-range Regulation of Src Kinases. *Protein Sci.* **2008**, *295*, 1871−1880.
(35) Yang, S.; Roux, B. Src Kinase Conformational Activation: Thermodynamics, Pathways, and Mechanisms. *PLoS Comput. Biol.* **2008**, *4*, e1000047.
(36) Shirts, M.; Pande, V. S. COMPUTING: Screen Savers of the World Unite! *Science* **2000**, *290*, 1903−1904.
(37) Xu, W.; Doshi, A.; Lei, M.; Eck, M. J.; Harrison, S. C. Crystal Structures of c-Src Reveal Features of Its Autoinhibitory Mechanism. *Mol. Cell* **1999**, *3*, 629−638.